ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αποτελεσματικές Τεχνικές Ταιριάσματος στον χώρο εργασίας

Κεφαλά Ευαγγελία

Επιβλέπων καθηγητής: Τζήμας Γιάννης

Πάτρα – Σεπτέμβριος 2021

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Πάτρα, Ημερομηνία

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Ονοματεπώνυμο, Υπογραφή
2. Ονοματεπώνυμο, Υπογραφή
3. Ονοματεπώνυμο, Υπογραφή

Αφιέρωση

Στην οικογένειά μου, ιδιαίτερα στην μητέρα μου αλλά και σε όλους όσοι με βοήθησαν να φτάσω ως εδώ.

Περιεχόμενα

Περίληψη

Το Διαδίκτυο έχει εξελιχθεί στο κύριο μέσο για τις διαδικασίες πρόσληψης και απασχόλησης. Πλέον, ο καθένας που αναζητά ή προσφέρει εργασία είναι πιθανότερο να στραφεί στο διαδίκτυο παρά με οποιοδήποτε άλλο τρόπο. Όμως, η ροή πληροφοριών στη διαδικτυακή αγορά εργασίας απέχει πολύ από τη βέλτιστη. Πολλές διαδικτυακές πλατφόρμες εργασίας ξεκίνησαν χωρίζοντας την ηλεκτρονική αγορά εργασίας σε πολλά μικρά κομμάτια κάνοντάς την δύσκολη για έναν αιτούντα να έχει πλήρη γνώση όλων των σχετικών ανοιχτών θέσεων.

Πληθώρα εργαλείων ηλεκτρονικής πρόσληψης έχουν κάνει την εμφάνισή τους τα τελευταία χρόνια, χωρίζονται και αναπτύσσονται από την σκοπιά του εκάστοτε ερευνητή και με την πάροδο του χρόνου τείνουν, προσθέτοντας πολλές παραμέτρους που ουδέποτε είχαν νόημα τα περασμένα χρόνια, να πλησιάζουν αρκετά στο πραγματικό στόχο.

Στην παρούσα διπλωματική αναλύονται Online recruitment systems τα οποία στηρίζονται στην αντιστοίχιση βιογραφικών και αγγελιών εργασίας μέσω της ανάλυσης λέξεων, της σημασιολογικής αντιστοίχισης λέξεων από διαφορετικά πεδία, της σχετικότητας λέξεων ή φράσεων ή ακόμη και προηγμένων τεχνολογιών μέσω χρήσης AI ώστε να καταφέρουν να ταιριάξουν όσο το δυνατό περισσότερα βιογραφικά σημειώματα στις καταλληλότερες θέσεις εργασίας μειώνοντας αντίστοιχα τον αριθμό των βιογραφικών εκείνων που κατά τον παραδοσιακό τρόπο θα απορρίπτονταν ως λανθασμένα λόγω ελλείψεων στην σύνταξη.

Abstract

As long as Internet has become the main tool for recruitment and employment processes, anyone who is looking for or offering a job is more likely to turn his research to the web than in any other way. But the flow of information in the online job market is far from optimal. Many online job platforms started by splitting the online job market into many small pieces making it difficult for an applicant to have full view of all the relevant vacancies.

A variety of e-learning tools have emerged in recent years, are separated and developed from the point of view of each researcher and over time tend, adding many parameters that never made sense in recent years, to be quite close to the real goal.

This essay analyzes Online recruitment systems that rely on matching CVs and job postings through word analysis, semantic matching, relevance matching, or even advanced technologies through the use of AI to match as much as possible more CVs in the most suitable jobs, thus reducing the number of CVs that would traditionally be rejected as incorrect due to pension deficiencies.

Εισαγωγή

Από καταβολής κόσμου, η ανάγκη για επιβίωση ήταν καθοριστικός παράγοντας για την ζωή ενός ατόμου. Η ανθρωπότητα έχει περάσει από διάφορες περιόδους, με σημαντικούς διαφορετικούς σταθμούς - κοινός παράγοντας παραμένει η επιβίωση, που κάθε φορά παίρνει διαστάσεις και ερμηνεύεται ανάλογα με την εποχή. Τις τελευταίες δεκαετίες, η επιβίωση προέρχεται μέσα από την εργασία, την ανθρώπινη δηλαδή δραστηριότητα που αποσκοπεί στην παραγωγή και προσφορά ενός προϊόντος (υλικού ή πνευματικού). Είναι η ενασχόληση με μία υποχρέωση έως ότου ληφθεί κάποιο μετρήσιμο αποτέλεσμα, μία χρονοβόρα διαδικασία που θα πρέπει να αμείβεται ειδικά όταν γίνεται για λογαριασμό άλλου. Εργοδότης είναι αυτός που προσφέρει εργασία, εργαζόμενος εκείνος που την φέρνει εις πέρας ενώ το σύνολο των εργαζομένων μιας χώρας αποτελεί την εργατική τάξη.

Η εργατική τάξη δραστηριοποιείται για βιοπορισμό, δηλαδή για την εξασφάλιση εισοδήματος που επιτρέπει την επιβίωση, ανάπτυξη και ευτυχία. Το πρότυπο εργασίας είναι μοντέλο που παρουσιάζει την ιδανική ποιότητα εργασίας, αποδοτικότητα, συμπεριφορά και ηθική, με τα οποία ένας εργαζόμενος θα εξασφαλίσει τις μέγιστες απολαβές. Η επαγγελματική ένταξη των νέων αποφοίτων διαφορετικών βαθμίδων, ειδών και κατευθύνσεων εκπαίδευσης και τα χαρακτηριστικά της περιόδου μετάβασής τους από την εκπαίδευση στην αμειβόμενη εργασία έχουν απασχολήσει τη διεθνή έρευνα από τα τέλη της δεκαετίας του 1960. Σε χώρες πρωτοπόρες σε αυτού του είδους την έρευνα, όπως η Γαλλία, η παρακολούθηση της επαγγελματικής διαδρομής των νέων μετά την αποφοίτηση ήταν άμεσα συνδεδεμένη με τον σχεδιασμό της εκπαιδευτικής πολιτικής. Σύμφωνα με τους σχεδιαστές της, το εκπαιδευτικό σύστημα έπρεπε να τροφοδοτεί την οικονομία με κατάλληλο σε ποσότητα και ειδικεύσεις εργατικό δυναμικό και η έρευνα – κατ' ανάθεση από το κράτος – βοηθούσε τον σχεδιασμό. Η κρίση της απασχόλησης στις δεκαετίες του 1970, 1980 και 1990 και τα υψηλά ποσοστά ανεργίας στους νέους έστρεψαν την έρευνα σε άλλα ερωτήματα και την υπέταξαν στις προτεραιότητες της πολιτικής αντιμετώπισης της ανεργίας [1].

Την σημερινή εποχή, το εργατικό δυναμικό αυξάνεται ολοένα και πιο πολύ και απαρτίζεται από εργαζομένους με υψηλό μορφωτικό επίπεδο, με σημαντικές ικανότητες και δεξιότητες, αλλά και από επίδοξους επιστήμονες διαφόρων ειδικοτήτων. Η απορρόφηση εργασίας εξαιτίας της οικονομικής κρίσης που μαστίζει την χώρα αλλά και τον δυτικό κόσμο από το 2008, κινείται σε χαμηλά επίπεδα και ενώ υπάρχει διαθέσιμο εργατικό δυναμικό η εύρεση εργασίας είναι από δύσκολη έως ακατόρθωτη. Ο εργαζόμενος ενώ διαθέτει την γνώση για να την υλοποιήσει και την όρεξη να προσφέρει και να γίνει και αυτός κομμάτι της παραγωγικής διαδικασίας παραμένει άνεργος είτε γιατί με το καθεστώς που επικρατεί στην εύρεση εργασίας η αντιστοίχισή του σε μια κατάλληλη για αυτόν θέση είναι σχεδόν ακατόρθωτη είτε γιατί δεν υπάρχουν προσφερόμενες θέσεις κατάλληλες για τις δεξιότητες του συγκεκριμένου ατόμου.

Η κατάλληλη θέση ή πιο σωστά η αντιστοίχιση εργαζομένου σε μια προσφερόμενη θέση οφείλεται κυρίως στην προσεκτική σύνταξη (ακριβής προσδιορισμός των χαρακτηριστικών μας) και δευτερευόντως στην παρουσίαση ενός καλού βιογραφικού σημειώματος. Η αποτύπωση της γνώσης, των δεξιοτήτων και της προηγούμενης

εργασιακής εμπειρίας γραμμένα είτε σε φυσική γλώσσα είτε ακολουθώντας κάποιο πρότυπο, αποτελεί το εισιτήριο για την πολυπόθητη συνέντευξη εργασίας καθώς απουσιάζει παντελώς η προσωπική επαφή, δηλαδή από την στιγμή που θα αιτηθεί κάποιος μια θέση έως την στιγμή που εργοδότης και υποψήφιος εργαζόμενος συναντηθούν δια ζώσης, η διαδικασία είναι απρόσωπη. Στην καλύτερη των περιπτώσεων, ο εργαζόμενος μέσα από την αναζήτηση και την αναμονή (που μπορεί να φτάσει έως και χρόνια) βρίσκει τελικά αυτό που του ταιριάζει. Υπάρχουν όμως και περιπτώσεις που μπορεί να παρουσιαστεί ένα άριστο ή έστω ενδιαφέρων βιογραφικό και εξαιτίας απρόβλεπτων παραγόντων, κριτήρια που δεν σχετίζονται πολλές φορές με την μόρφωση ή τις δεξιότητες αλλά με γεωγραφικά, πολιτισμικά, φύλου, οικογενειακών υποχρεώσεων, κοινωνικά κ.α. αντίστοιχα, να αποτελέσουν τροχοπέδη στην διαδικασία επιλογής και στην τελική απόρριψή του. Τις περισσότερες φορές τα κριτήρια αυτά δεν είναι ξεκάθαρα ή έχουν παραληφθεί κατά την δημοσίευση της εκάστοτε θέσης χάνοντας ο υποψήφιος έτσι πολύτιμο χρόνο.

Οι εργοδότες από την άλλη πλευρά ψάχνουν να προσθέσουν στο δυναμικό τους τους υποψήφιους εκείνους που είναι ικανοί να μεγιστοποιήσουν τα κέρδη της επιχείρησής τους. Για τον λόγο αυτό, είτε στις δημοσιεύσεις τους αυξάνουν τις δυνατότητες – δεξιότητες που θέλουν να κατέχει ο υποψήφιος ώστε να αποθαρρύνονται οι αδύναμοι, είτε αξιολογούν πιθανούς υποψηφίους αναζητώντας τους μέσα από το ίδιο το ανθρώπινο δυναμικό τους (πιθανή συγγένεια, κοινωνικές συναναστροφές) θεωρώντας πως έτσι κερδίζουν χρόνο βρίσκοντας από τους καλύτερους τους καλύτερους. Είναι προφανές πως αν ένας εργοδότης στραφεί προς αυτήν την κατεύθυνση και δεν δημοσιοποιήσει την προσφερόμενη θέση, αυτομάτως αποκλείει ένα μεγάλο κομμάτι υποψηφίων. Για τον λόγο αυτό, πολλοί αποστέλλουν το βιογραφικό τους στις επιχειρήσεις σε κενό χρόνο, ελπίζοντας σε πιθανή μελλοντική συνεργασία. Η αξιολόγηση των βιογραφικών με τις παραδοσιακές μεθόδους αναζήτησης εργαζομένων προϋποθέτει την ύπαρξη σχετικού τμήματος και τα στάδια που θα ακολουθήσει το γραφείο ανθρώπινου δυναμικού για να συλλέξει τα βιογραφικά είναι:

1. Προσδιορισμός δεξιοτήτων κενής θέσης εργασίας
2. Σύνταξη αγγελίας
3. Δημοσιοποίηση αγγελίας
4. Παραλαβή βιογραφικών σημειωμάτων υποψηφίων εργαζομένων
5. Αξιολόγηση αυτών
6. Επικοινωνία με τους τελικούς υποψηφίους για συνέντευξη

Για τον κάθε εργοδότη – επιχείρηση η στρατολόγηση και εκπαίδευση του νέου προσωπικού είναι χρονοβόρα, κοστοβόρα και δύσκολη διαδικασία και ενέχει τον κίνδυνο της λύσης της εργασιακής σχέσης σχετικά νωρίς.

Τα τελευταία χρόνια, και ενώ το διαδίκτυο έχει γίνει αναπόσπαστο κομμάτι της καθημερινότητάς μας, έχει ήδη εξελιχθεί και στο κύριο μέσο για τις διαδικασίες πρόσληψης και απασχόλησης αφού έχουν κάνει την εμφάνισή τους πλατφόρμες αναζήτησης εργασίας. Αυτές οι πλατφόρμες είναι προσβάσιμες από όσους χρησιμοποιούν το διαδίκτυο, προσφέροντας έτσι την άνεση για εύρεση εργασίας από τον υπολογιστή του σπιτιού τους, εξασφαλίζοντας την διαφάνεια των προσφερόμενων θέσεων, δηλαδή ο κάθε υποψήφιος να έχει ακριβώς τις ίδιες πιθανότητες να επιλεγεί

ανάμεσα σε χιλιάδες άλλους υποψηφίους. Πολλοί εργοδότες έχουν στραφεί στην δημοσιοποίηση των αγγελιών τους μέσω κάποιας τέτοιας διαδικτυακής πλατφόρμας ώστε να προσεγγίσουν όλο και περισσότερους υποψηφίους δίνοντάς τους την εξασφάλιση από την μεριά τους πως η διαδικασία είναι αξιοκρατική. Λόγω του κόστους δημοσιοποίησης των αγγελιών αυτών, οι επιχειρήσεις συχνά επιλέγουν ορισμένες πλατφόρμες που θα συνεργάζονται με αποτέλεσμα αυτό να οδηγεί στην διαίρεση της διαδικτυακής αγοράς εργασίας σε υποομάδες και να καθιστά σχεδόν αδύνατο στον εκάστοτε αιτούντα να έχει πλήρη εικόνα όλων των σχετικών ανοιχτών θέσεων [2].

Ένα άλλο ενδιαφέρον κομμάτι της διαδικτυακής εύρεσης εργασίας είναι ο μοντέρνος τρόπος κοινωνικής δικτύωσης των ατόμων, τα λεγόμενα social media. Είναι πλατφόρμες (Facebook, Twitter, Instagram, LinkedIn®) που μπορεί ο οποιοσδήποτε να γίνει εύκολα μέλος δημιουργώντας έναν λογαριασμό (είτε με τα προσωπικά του στοιχεία είτε με κάποιο προσωνύμιο), διατίθενται δωρεάν, παρέχουν άμεση πρόσβαση 24/7 μέσα από το κινητό τηλέφωνο, το tablet ή τον υπολογιστή και παρέχουν διάφορες δυνατότητες, μια εκ των οποίων πλέον είναι και η εύρεση εργασίας. Αυτό εξασφαλίζεται μέσα από τον τρόπο που λειτουργούν οι πλατφόρμες αυτές και στηρίζονται στους «φίλους» και στις διασυνδέσεις που έχει ο κάθε χρήστης. Πιο συγκεκριμένα εξαιτίας του μηδενικού κόστους δημιουργίας ενός λογαριασμού, ολοένα και περισσότεροι χρήστες γίνονται μέλη σε μία ή και περισσότερες πλατφόρμες, είτε είναι απλοί χρήστες είτε επιχειρήσεις. Σκοπός και των δύο είναι να αυξήσουν τους διαδικτυακούς τους φίλους και μέσα από αυτό να έχουν την μέγιστη προβολή. Η δημοσιοποίηση μιας προσφερόμενης θέσης εργασίας από πλευράς επιχειρήσεων μπορεί να γίνει γνωστή μέσα σε λίγα μόνο λεπτά «ανεβάζοντας» την στον λογαριασμό της και όσοι ακολουθούν είτε αναδημοσιεύουν το εν λόγω μήνυμα στον δικό τους λογαριασμό ενημερώνοντας έτσι τους διασυνδεδεμένους φίλους τους είτε το αποστέλλουν με προσωπικό μήνυμα στον φίλο που αναζητά εργασία. Από την σκοπιά των επιχειρήσεων, η αύξηση των ακολούθων – άρα της προβολής - συνεπάγεται αύξηση των μετοχών τους από πιθανή αύξηση των πωλήσεων (ενδεχομένως και δημιουργία νέων θέσεων εργασίας) ενώ αν αναφερόμαστε για απλούς χρήστες είναι αντίθετα τα αποτελέσματα καθώς όσο μεγαλύτερη είναι η προβολή και αλληλεπίδραση στον χώρο τόσο περισσότερο εκτίθεται η προσωπικότητα του ατόμου. Ένα καθαρά αντιπροσωπευτικό παράδειγμα κοινωνικής πλατφόρμας εύρεσης εργασίας θα μπορούσε να θεωρηθεί το LinkedIn, με πάνω από 300 εκατομμύρια εγγεγραμμένους χρήστες (https://el.wikipedia.org/wiki/LinkedIn) παγκοσμίως και λειτουργίες όπως:

1. τη δημιουργία και την παρουσίαση του προφίλ των μελών,
2. τη δυνατότητα λεπτομερούς παρουσίασης της εργασιακής εμπειρίας και ανάλυση του εκπαιδευτικού υπόβαθρου,
3. τη δυνατότητα καταχώρησης προσωπικών πληροφοριών και ενδιαφερόντων,
4. τη δικτύωση και την αλληλεπίδραση με τα συνδεδεμένα μέλη,
5. τη δυνατότητα παροχής συστάσεων από συναδέλφους, συνεργάτες και γενικότερα μέλη του δικτύου με τα οποία είναι κανείς συνδεδεμένος
6. τη δημοσίευση και το διαμοιρασμό αναρτήσεων, παρουσιάσεων κλπ.

Στηρίζεται καθαρά στην διασύνδεση των μελών (όπως όλα τα social media) και όχι στην αντιστοίχιση υποψηφίων εργαζομένων με προσφερόμενη θέση εργασίας.

Αδιαμφισβήτητα, λόγω της ταχείας ανάπτυξης των διαδικτυακών αγορών εργασίας (οποιασδήποτε μορφής) οι παραδοσιακές μέθοδοι πρόσληψης καθίστανται ανεπαρκείς. Αυτό συμβαίνει, όπως ήδη επισημάνθηκε προηγουμένως, επειδή οι εργοδότες λαμβάνουν συχνά έναν τεράστιο αριθμό βιογραφικών που είναι δύσκολο πλέον να επεξεργαστούν και να αναλυθούν χειροκίνητα. Για την αντιμετώπιση αυτού του ζητήματος, έχουν προταθεί διάφορα πιο αυτοματοποιημένα συστήματα στρατολόγησης προσωπικού σε επιχειρήσεις (Online Recruitment Systems). Πολλές επιχειρήσεις αρχίζουν να στρέφονται σε αυτά αφού προσφέρουν ένα σημαντικό κέρδος εξαιτίας της αποτελεσματικότητας τους, επιταχύνοντας τη διαδικασία της πρόσληψης και εξοικονομώντας χρόνο σε όλους τους εμπλεκομένους. Με άλλα λόγια, πρόκειται για ένα σύστημα σύστασης εργασίας που συνιστά στις επιχειρήσεις τους πιο κατάλληλους υποψηφίους για μια συγκεκριμένη θέση ενώ από την άλλη πλευρά προτείνει τις κατάλληλες θέσεις εργασίας στους υποψηφίους, ταιριάζοντας δηλαδή με ακρίβεια βιογραφικά σε υπάρχουσες προσφερόμενες θέσεις. Είναι δύσκολο να γίνει από έναν υποψήφιο η διαδικασία αυτή, να αναλύσει όλες εκείνες τις ανοιχτές θέσεις και στη συνέχεια να επιλέξει από αυτές εκείνες που ταιριάζουν πραγματικά στα χαρακτηριστικά του. Σε πολλές περιπτώσεις μπορεί να μην γνωρίζει και ο ίδιος ότι διαθέτει όλες τις απαιτούμενες ικανότητες για συγκεκριμένη εργασία, ή να μην γνωρίζει καν την ύπαρξη της εργασίας αυτής. Η αυτόματη αντιστοίχιση μεταξύ των δεξιοτήτων και των ικανοτήτων που απαιτεί μια επιχείρηση για μια κάποια θέση και εκείνων που διαθέτει ο αιτών, βοηθά στη μείωση των σφαλμάτων που μπορεί να προκύψουν όταν η διαδικασία γίνεται από το εκάστοτε γραφείο ανθρώπινου δυναμικού με τον παραδοσιακό τρόπο λαμβάνοντας υπόψη πάντα το ανθρώπινο λάθος. Σ' αυτό το σημείο θα πρέπει να αποσαφηνιστεί ότι τα Online Recruitment Systems βοηθούν και δεν αντικαθιστούν τον ανθρώπινο παράγοντα στη διαδικασία λήψης αποφάσεων [3].

Τελευταίες μελέτες στον χώρο κατατάσσουν τα Online Recruitment Systems σε κατηγορίες ανάλογα με τον τρόπο που είναι δομημένα τα δεδομένα που χρησιμοποιούν ή την ταξινόμηση του εκάστοτε συστήματος που θα χρησιμοποιήσουν. Όπως αναφέρεται σε άρθρο από το 2016 της Mihaela-Irina ENĂCHESCU[4], τα συστήματα χωρίζονται σε τέσσερις κύριες κατηγορίες:

- Collaborative Filtering (CF) - χρησιμοποιεί τη μέθοδο συσχέτισης μεταξύ χρηστών για να προβλέψει τις προτιμήσεις ενός νέου χρήστη με βάση τις προτιμήσεις παρόμοιων χρηστών.
- Content-Based Filtering (CBF) - συνιστά στοιχεία που έχουν παρόμοιο περιεχόμενο με αυτό που έχει δει ή επιλέξει ο χρήστης στο παρελθόν.
- Knowledge-Based Approach - κάνει προτάσεις βάσει συμπερασμάτων σχετικά με τις ανάγκες και τις προτιμήσεις του χρήστη.
- Hybrid Approach - συνδυάζει μία ή περισσότερες από τις ήδη αναφερθείσες μεθόδους για καλύτερη απόδοση.

Η ανάπτυξη ενός ισχυρού e-recruitment system δεν είναι καθόλου εύκολο έργο. Υπήρχαν πάντα διαφορετικά προβλήματα που έπρεπε να λυθούν. Το πρώτο πρόβλημα που αντιμετώπισαν τα συστήματα αντιστοίχισης ήταν η ποσότητα των ημιδομημένων δεδομένων (semi-structured data). Η συμπλήρωση ενός βιογραφικού σημειώματος σε ηλεκτρονική πλατφόρμα δημιουργεί μια αποθήκη δεδομένων που μέσω αυτής θα γίνει η

αντιστοίχιση. Εάν ο χρήστης έπρεπε να συμπληρώσει με ελεύθερο κείμενο κάποιο πεδίο αλλά παρέλειψε να το συμπληρώσει αποτελεί ημιδεδομένο καθώς και το κενό έχει το ίδιο βάρος με τα συμπληρωμένα πεδία. Έτσι προτάθηκε η χρήση των δομημένων συνάφειας (SRM) ως λύση για τα πεδία που λείπουν και αποτελεί ένα πρώτο παράδειγμα προσέγγισης CF. Περαιτέρω μελέτες σε αυτό τον τομέα προτείνουν την ανάπτυξη ενός αυτοματοποιημένου συστήματος που βασίζεται στην εποπτευόμενη εκμάθηση μιας μηχανής, τεχνική που χρησιμοποιεί την προσέγγιση φιλτραρίσματος βάσει περιεχομένου. Στόχος είναι να προτείνει θέσεις εργασίας σε χρήστες που έχουν διαθέσιμες πληροφορίες σχετικά με τις προηγούμενες εργασιακές θέσεις ενώ χρησιμοποιήθηκαν και πληροφορίες από τα διαθέσιμα προφίλ των υποψηφίων στα social media με τις ανάλογες τεχνικές.

Τέλος, κάνει λόγο για τα οφέλη από την χρήση ενός υβριδικού συστήματος ότι είναι περισσότερα, μιλώντας για ένα σύστημα αναζήτησης θέσεων εργασίας που βασίζεται στη συλλογή πληροφοριών από διαφορετικές πηγές. Το μοντέλο δεδομένων βασίζεται στις συνδέσεις των οντοτήτων που δημιουργούνται χρησιμοποιώντας σχέσεις που βασίζονται στο περιεχόμενο και τις αλληλεπιδράσεις. Οι σχέσεις που βασίζονται στο περιεχόμενο υποθέτουν ότι δημιουργείται μια σχέση μεταξύ των αιτούντων και των ατόμων που αναζητούν εργασία με δύο τρόπους όπως:

1. ταίριασμα προφίλ _προσδιορίζει τους αιτούντες με τα ίδια χαρακτηριστικά όπως αυτά που αναφέρονται στην περιγραφή της εργασίας, και
2. ομοιότητα προφίλ _συνδέει τους αιτούντες εργασία με όλους τους χρήστες από μια ομάδα που έχουν ήδη σχέση με έναν αιτούντα από την ίδια ομάδα,

ενώ οι σχέσεις που βασίζονται στην αλληλεπίδραση στηρίζονται στη συλλογή πληροφοριών από προηγούμενες αλληλεπιδράσεις του αιτούντος με το σύστημα.
Οι αλγόριθμοι που χρησιμοποιήθηκαν χωρίζονται βάσει των δεδομένων που αξιοποιήθηκαν ως εξής:

1. αλγόριθμοι που χρησιμοποιούν αξιολόγηση συμπεριφοράς στην κοινότητα,
2. αλγόριθμοι βάσει περιεχομένου για την εξαγωγή των χαρακτηριστικών,
3. αλγόριθμοι που βασίζονται στη γνώση για περαιτέρω διερεύνηση των διαθέσιμων γενικών γνώσεων μέσα στο σύστημα.

Οι ερευνητές των μοντέλων αυτών επέλεξαν να χρησιμοποιήσουν την υβριδική προσέγγιση προτάσεων γιατί θεωρούν πως κάθε χρήστης του συστήματος είναι μοναδικός και διαφορετικοί αλγόριθμοι μπορεί να είναι κατάλληλοι για διαφορετικούς χρήστες. Πολλές μελέτες έχουν προκύψει από αυτή την προσέγγιση και εστιάζουν περισσότερο στην προσωπικότητα του αιτούντος σε ένα Online Recruitment System και ισχυρίζονται ότι η προσωπικότητα θα μπορούσε να είναι ένα μετρήσιμο χαρακτηριστικό και τα στοιχεία αυτά να αντλούνται από την κοινωνική του δραστηριότητα. Ο αλγόριθμος θα λαμβάνει υπόψη του την διεύθυνση URL από το web κατά την δημιουργία του λογαριασμού του στα social media και εν συνεχεία, η γλωσσική ανάλυση θα εφαρμόζεται στις αναρτήσεις του για την εξαγωγή χαρακτηριστικών γνωρισμάτων προσωπικότητας.

Ο παρακάτω πίνακας αποτυπώνει επιγραμματικά τι θα πρέπει να ληφθεί υπόψη σε ένα online recruitment system πριν σχεδιαστεί.

Λειτουργικότητα	Περιγραφή
Δημιουργία εταιρικού προφίλ	Η κάθε εταιρία οφείλει να παρέχει πληροφορίες τέτοιες ώστε να είναι πιο αναγνωρίσιμη έναντι των υπολοίπων.
Δημοσίευση προσφερόμενης θέσης	Η κάθε εταιρία θα πρέπει να έχει τη δυνατότητα να δημοσιεύει ανά πάσα στιγμή προσφερόμενη θέση εργασίας, που να περιέχει το σύνολο των απαιτήσεων που πρέπει να κατέχει ο υποψήφιος.
Δημιουργία προφίλ υποψηφίου	Κάθε υποψήφιος που σκοπεύει να αιτηθεί σε μια θέση θα πρέπει να παρουσιάσει ένα σύνολο προσωπικών λεπτομερειών αλλά και πληροφοριών ώστε να αποδεικνύει έτσι την μοναδικότητά του.
Λίστα όλων των υποψηφίων που ταιριάζουν στην προσφερόμενη θέση	Η εταιρία θα πρέπει να μπορεί να ζητήσει για οποιαδήποτε προσφερόμενη θέση μια λίστα με τους πιο κατάλληλους για αυτήν την θέση υποψηφίους.
Λίστα με όλες τις δημοσιευμένες θέσεις εργασίας που ταιριάζουν στο προφίλ του υποψηφίου	Αφού συμπληρώσει το προφίλ του, ο υποψήφιος θα πρέπει να έχει πρόσβαση σε μια λίστα με όλες τις αναρτημένες θέσεις εργασίας, ανάλογα της συμβατότητάς του με τις δεξιότητες και τις ικανότητές του.
Λίστα με όλες τις προσφερόμενες θέσεις μιας εταιρίας	Η εταιρία θα πρέπει να προβάλλει όλες τις προσφερόμενες θέσεις εργασίας που έχει δημοσιεύσει.
Ενημέρωση του προφίλ του υποψηφίου	Ο υποψήφιος έχει τη δυνατότητα να προσθέτει συνεχώς νέες πληροφορίες στο προφίλ του ή να αλλάζει τις υπάρχουσες λεπτομέρειες. Μετά την ολοκλήρωση της ενημέρωσης, οι προτεινόμενες προσφορές που ταιριάζουν με το προφίλ του θα πρέπει επίσης να ενημερώνονται αντίστοιχα.
Ενημέρωση των δημοσιευμένων θέσεων εργασίας	Θα πρέπει να παρέχει στην εταιρεία τη δυνατότητα να διατηρεί τις καταχωρημένες αγγελίες εργασίας, ενώ μετά την εκτέλεση των αλλαγών, η λίστα με τους κατάλληλους υποψηφίους θα πρέπει να λαμβάνει υπόψη την τελευταία έκδοση της ενημερωμένης θέσης εργασίας.
Ενημέρωση του προφίλ της εταιρίας	Η εταιρεία θα πρέπει να μπορεί να ενημερώσει τα στοιχεία της ή να προσθέσει νέες πληροφορίες σχετικά με τη δραστηριότητά της.

Πίνακας 1: Βήματα δημιουργίας Online Recruitment System

Μία άλλη ενδιαφέρουσα προσέγγιση για τα online recruitment systems προέχεται από έρευνα του 2015 των Aseel B. Kmail, Mohammed Maree, Mohammed Belkhatir και Saadat M. Alhashmi [5], όπου ένα e-recruitment system χωρίζεται επίσης σε τέσσερις κατηγορίες προσανατολισμένες στην σημασία των λέξεων, έχοντας στόχο να εντοπίζουν και να εξάγουν λίστες υποψηφίων εκμεταλλεύοντας πολλαπλούς σημασιολογικούς πόρους σε μια προσπάθεια ανάδειξης και καταγραφής των σημασιολογικών πτυχών τόσο των θέσεων εργασίας όσο και των υποψηφίων. Κάνοντας χρήση στατιστικών μεθόδων, οι λίστες αυτές βελτιώνονται καθιστώντας τες πιο ακριβείς με πιο σχετικές εγγραφές απομακρύνοντας τις μη σχετικές ή εκείνες που λόγω κακής διάρθρωσης ή κενών πείων κατά την συμπλήρωση και υποβολή δεν αναγνωρίστηκαν από τους σχετικούς αλγορίθμους.

1. **Traditional Keyword – based Techniques**: Αυτές οι τεχνικές στηρίζονται κυρίως στην ακριβή αντιστοίχιση λέξεων-κλειδιών που εξάγονται από τις θέσεις εργασίας και διαθέσιμων βιογραφικών υποψηφίων.
2. **Relevance-based Models:** Δομημένα από ταξινομημένα έγγραφα, συσχετίζονται λίστες θέσεων εργασίας με περιγραφές υποψηφίων. Τα μοντέλα αυτά χρησιμοποιούνται για να αντισταθμίσουν τις παραλλαγές του λεξιλογίου ανάμεσα στα βιογραφικά σημειώματα και στις θέσεις εργασίας.
3. **Semantics-based Approaches**: Δίνεται έμφαση στην σημασιολογία, όχι απλώς στην συντακτική σύγκριση λέξεων που επιτυγχάνεται μέσω κοινού λεξιλογίου ώστε να περιγραφούν επακριβώς οι διαθέσιμες θέσεις εργασίας και τα βιογραφικά για να γίνει με μεγαλύτερη ακρίβεια η εκάστοτε αντιστοίχιση.
4. **Machine Learning Techniques**: Ένας αριθμός αλγορίθμων μηχανικής μάθησης αξιοποιείται στον τομέα διαδικτυακών προσλήψεων για ανάλυση δεδομένων και εξαγωγή πληροφοριών. Αυτοί οι αλγόριθμοι περιλαμβάνουν νευρωνικά δίκτυα, ομαδοποίηση, δέντρα αποφάσεων, ή τεχνικές ταξινόμησης Support Vector Machine (SVM) για την επισήμανση χαρακτηριστικών τόσο των υποψηφίων όσο και των θέσεων εργασίας προς αντιστοίχιση.

Στα επόμενα κεφάλαια της παρούσας εργασίας, θα αναλυθούν εκτενώς οι κατηγορίες αυτές καθώς επίσης θα παρουσιαστούν και θα αναλυθούν εξίσου ορισμένα από τα αντιπροσωπευτικά τους μοντέλα.

1. Traditional Keyword – based Techniques

1.1 NLP

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι η μηχανογραφημένη προσέγγιση για την ανάλυση κειμένου που βασίζεται τόσο σε ένα σύνολο θεωριών όσο και σε ένα σύνολο τεχνολογιών. Ένας κατάλληλος ορισμός για να το περιγράψει είναι ο εξής:

Η Επεξεργασία Φυσικής Γλώσσας είναι ένα θεωρητικά υποκινούμενο εύρος υπολογιστικών τεχνικών για την ανάλυση και την αναπαράσταση φυσικών κειμένων σε ένα ή περισσότερα επίπεδα γλωσσικής ανάλυσης με σκοπό την επίτευξη ανθρώπινης γλωσσικής επεξεργασίας για μια σειρά από εργασίες ή εφαρμογές.

Αρχικά, η ακριβής έννοια του «εύρος υπολογιστικών τεχνικών» είναι απαραίτητη επειδή υπάρχουν πολλές μέθοδοι ή τεχνικές που μπορούμε να χρησιμοποιήσουμε για την ανάλυσης της γλώσσας. Τα «φυσικά κείμενα» μπορούν να είναι γραμμένα σε οποιαδήποτε γλώσσα, με οποιοδήποτε τρόπο, μπορούν να είναι προφορικά ή γραπτά. Η μόνη απαίτηση είναι να είναι σε γλώσσα που χρησιμοποιούν για επικοινωνία οι άνθρωποι. Το κείμενο που αναλύεται δεν πρέπει να γράφεται ειδικά για τον σκοπό αυτό, αλλά να είναι ένα σύνηθες κείμενο το οποίο έχει συλλεχθεί από πραγματική χρήση.

Τα «επίπεδα γλωσσικής ανάλυσης» αναφέρονται στο γεγονός ότι υπάρχουν πολλοί τύποι επεξεργασίας γλωσσών που λειτουργούν όταν οι άνθρωποι παράγουν ή κατανοούν τη γλώσσα. Πιστεύεται δε, πως κάθε άνθρωπος κάνει χρήση όλων αυτών των επίπεδων αφού κάθε επίπεδο μεταφέρει διαφορετικούς τύπους νοήματος. Ωστόσο, τα συστήματα εφαρμογών NLP διαφέρουν μεταξύ τους επειδή χρησιμοποιούν διαφορετικά επίπεδα ή συνδυασμούς επιπέδων γλωσσικής ανάλυσης. Δημιουργούνται εύλογες απορίες λοιπόν ως προς το τι είναι πραγματικά ένα σύστημα NLP, επειδή μπορεί να χρησιμοποιεί οποιοδήποτε υποσύνολο αυτών των επιπέδων ανάλυσης. Η βασική διαφορά τους είναι αν το σύστημα χρησιμοποιεί «αδύναμο» NLP ή «ισχυρό» NLP. Η «επεξεργασία ανθρώπινων γλωσσών» δείχνει πως το NLP αποτελεί κλάδο της Τεχνητής Νοημοσύνης (AI), παρόλο που το NLP εξαρτάται από πλήθος άλλων κλάδων, δεδομένου ότι το NLP προσπαθεί για την ανθρώπινη απόδοση.

Η φράση «για μια σειρά εργασιών ή εφαρμογών» δείχνει πως το NLP δεν θεωρείται στόχος αυτοτελώς, ενδεχομένως ίσως μόνο από τους ερευνητές της τεχνητής νοημοσύνης. Για άλλους όμως αποτελεί το μέσο για επίτευξη του στόχου που είναι να ολοκληρώσει την ανθρώπινη επεξεργασία γλωσσών. Η επιλογή της λέξης «επεξεργασία» δεν έχει λεχθεί τυχαία και δεν πρέπει να αντικατασταθεί με την λέξη «κατανόηση». Ο κλάδος του NLP αναφέρθηκε ως «Κατανόηση Φυσικής Γλώσσας (NLU)» όταν έκανε την εμφάνισή του το AI, και ενώ έμοιαζε με εκείνου του NLU, στην πραγματικότητα ακόμη δεν μπορούν να ομαδοποιηθούν[6].

Ένα σύστημα NLP λοιπόν είναι ικανό να παραφράζει κείμενα, να μεταφράζει κείμενα σε διαφορετική γλώσσα, να μπορεί να δίνει απαντήσεις σε ερωτήματα σχετικά με το περιεχόμενο ενός κειμένου όμως δεν έχει καταφέρει ακόμη να μπορεί αντλήσει συμπεράσματα από κείμενο. Δηλαδή να παρέχει ακριβείς και πλήρεις πληροφορίες ως

απόκριση στην αναζήτηση πληροφοριών από ένα χρήστη. Να δώσει το πραγματικό νόημα και την πρόθεση του ερωτήματος του χρήστη και το οποίο να μπορεί να εκφραστεί όπως ακριβώς θα εκφραζόταν στην καθημερινή γλώσσα.

Όπως οι περισσότεροι σύγχρονοι κλάδοι, η προέλευση του NLP είναι πράγματι ανάμεικτη και εξακολουθεί να επηρεάζεται από διαφορετικές ομάδες και εκείνες με την σειρά τους από άλλους κλάδους. Το κλειδί μεταξύ των συντελεστών αυτών είναι:

- Γλωσσολογία: επικεντρώνεται στα επίσημα, δομικά μοντέλα της γλώσσας και στην ανακάλυψη γλωσσικών καθολικών - στην πραγματικότητα ο τομέας του NLP αρχικά αναφέρεται ως Υπολογιστική Γλωσσολογία
- Πληροφορική: ασχολείται με την ανάπτυξη εσωτερικών αναπαραστάσεων δεδομένων και την αποτελεσματική επεξεργασία αυτών των δομών, και
- Γνωστική Ψυχολογία: εξετάζει τη χρήση της γλώσσας ως παράθυρο στις ανθρώπινες γνωστικές διαδικασίες και έχει ως στόχο τη μοντελοποίηση της χρήσης της γλώσσας με ψυχολογικά εύλογο τρόπο.

Ενώ ολόκληρο το πεδίο αναφέρεται ως Επεξεργασία Φυσικής Γλώσσας, στην πραγματικότητα η επεξεργασία γλωσσών και η παραγωγή γλωσσών είναι εντελώς διαφορετικά πράγματα. Το πρώτο αναφέρεται στην ανάλυση της γλώσσας με σκοπό την παραγωγή μιας ουσιαστικής αναπαράστασης (παίζει το ρόλο του αναγνώστη/ακροατή), ενώ η δεύτερη αναφέρεται στην παραγωγή της γλώσσας από μια αναπαράσταση (παίζει το ρόλο του συγγραφέα/ομιλητή).

Επίπεδα Επεξεργασίας Φυσικής Γλώσσας

Αν θέλαμε να εξηγήσουμε τι πραγματικά συμβαίνει σε ένα σύστημα επεξεργασίας φυσικής γλώσσας το πιο σωστό θα ήταν μέσω της προσέγγισης «επίπεδα γλώσσας», καθώς παρουσιάζεται ως σύγχρονο μοντέλο γλώσσας, διακρίνεται από το προηγούμενο διαδοχικό μοντέλο και υποθέτει ότι τα επίπεδα επεξεργασίας της ανθρώπινης γλώσσας ακολουθούν το ένα το άλλο με αυστηρά διαδοχικό τρόπο. Η ψυχογλωσσολογική έρευνα δείχνει ότι η επεξεργασία γλωσσών είναι πιο δυναμική, καθώς τα επίπεδα μπορούν να αλληλοεπιδράσουν σε διάφορες εντολές. Η ενδοσκόπηση δείχνει ότι χρησιμοποιούμε συχνά πληροφορίες από το θεωρητικά υψηλότερο επίπεδο επεξεργασίας για να βοηθήσουμε ένα χαμηλότερο επίπεδο ανάλυσης. Για παράδειγμα, η γνώση ότι διαβάστηκε ένα έγγραφο, θα χρησιμοποιηθεί όταν βρεθεί μια συγκεκριμένη λέξη που έχει αρκετές πιθανές συνδέσεις ή έννοιες με το έγγραφο που είχε διαβαστεί. Άρα το νόημα μεταδίδεται από κάθε επίπεδο γλώσσας και επειδή έχει αποδειχθεί ότι οι άνθρωποι χρησιμοποιούν όλα τα επίπεδα γλώσσας για να φτάσουν στην κατανόηση, ένα σύστημα NLP είναι ικανό μόνο όταν θα κάνει χρήση, αν όχι σε όλα, αλλά σε όσα περισσότερα επίπεδα γλώσσας μπορεί.

Προσεγγίσεις στην επεξεργασία φυσικής γλώσσας

Οι προσεγγίσεις επεξεργασίας φυσικής γλώσσας εμπίπτουν σε τέσσερις κατηγορίες: συμβολικές, στατιστικές, συνδετικές και υβριδικές.

- Συμβολική προσέγγιση: Οι συμβολικές προσεγγίσεις εκτελούν ανάλυση σε βάθος των γλωσσικών φαινομένων και βασίζονται σε αναπαράσταση των γεγονότων μέσω κατανοητών σχεδίων αναπαράστασης της γνώσης και συναφών αλγορίθμων.
- Στατιστική προσέγγιση: Οι στατιστικές προσεγγίσεις χρησιμοποιούν διάφορες μαθηματικές τεχνικές και συχνά χρησιμοποιούν μεγάλα κείμενα για την ανάπτυξη γενικευμένων μοντέλων γλωσσικών φαινομένων, βασίζονται σε πραγματικά παραδείγματα αυτών και παρέχονται από αυτά τα κείμενα χωρίς να προσθέτουν σημαντική γλωσσική γνώση.
- Συνδετική προσέγγιση: Οι συνδετικές προσεγγίσεις λειτουργούν παρόμοια με τις στατιστικές προσεγγίσεις αλλά αναπτύσσουν γενικευμένα μοντέλα από παραδείγματα γλωσσικών φαινομένων. Αυτό που τις διαχωρίζει είναι ότι τα συνδετικά μοντέλα συνδυάζουν τη στατιστική μάθηση με διάφορες θεωρίες αναπαράστασης - έτσι οι παραστατικές συνδέσεις επιτρέπουν μετασχηματισμό, συμπεράσματα και χειρισμό λογικών τύπων. Επιπλέον, στα συστήματα σύνδεσης, τα γλωσσικά μοντέλα είναι πιο δύσκολο να παρατηρηθούν λόγω του γεγονότος ότι οι αρχιτεκτονικές σύνδεσης είναι λιγότερο περιορισμένες από τις στατιστικές.

Συμβολικές και στατιστικές προσεγγίσεις συνυπάρχουν από την εμφάνιση του πεδίου αυτού. Το έργο Connectionist NLP εμφανίστηκε για πρώτη φορά στη δεκαετία του 1960. Για μεγάλο χρονικό διάστημα, οι συμβολικές προσεγγίσεις κυριάρχησαν στο πεδίο. Στη δεκαετία του 1980, οι στατιστικές προσεγγίσεις έγιναν πιο δημοφιλείς ως αποτέλεσμα της διαθεσιμότητας κρίσιμων υπολογιστικών πόρων και της ανάγκης αντιμετώπισης ευρέων, πραγματικών πλαισίων. Οι προσεγγίσεις του Connectionist ανέκαμψαν επίσης από προηγούμενες κριτικές αποδεικνύοντας τη χρησιμότητα των νευρωνικών δικτύων στο NLP. Αυτή η ενότητα εξετάζει καθεμία από αυτές τις προσεγγίσεις όσον αφορά τα θεμέλια τους, τις τυπικές τεχνικές, τις διαφορές στην επεξεργασία και τις πτυχές του συστήματος, την ανθεκτικότητα, την ευελιξία και την καταλληλότητα τους για διάφορου τύπου εργασίες.

Η έρευνα με τη χρήση αυτών των διαφορετικών προσεγγίσεων ακολουθεί ένα γενικό σύνολο βημάτων, δηλαδή τη συλλογή δεδομένων, την ανάλυση δεδομένων, δημιουργία μοντέλων, την κατασκευή κανόνων δεδομένων και την εφαρμογή κανόνων - δεδομένων στο σύστημα. Το στάδιο συλλογής δεδομένων είναι κρίσιμο και για τις τρεις προσεγγίσεις αν και οι στατιστικές και οι συνδετικές προσεγγίσεις συνήθως απαιτούν πολύ περισσότερα δεδομένα από τις συμβολικές προσεγγίσεις. Στο στάδιο της ανάλυσης δεδομένων - δημιουργίας μοντέλων, οι συμβολικές προσεγγίσεις βασίζονται στην ανθρώπινη ανάλυση των δεδομένων προκειμένου να σχηματίσουν μια θεωρία, ενώ οι στατιστικές προσεγγίσεις ορίζουν χειροκίνητα ένα στατιστικό μοντέλο που είναι μια κατά προσέγγιση γενίκευση των συλλεγόμενων δεδομένων. Οι προσεγγίσεις Connectionist δημιουργούν ένα μοντέλο σύνδεσης από τα δεδομένα.

Στο στάδιο κατασκευής κανόνα - δεδομένων, οι χειροκίνητες προσπάθειες είναι τυπικές για συμβολικές προσεγγίσεις και η θεωρία που διαμορφώθηκε στο προηγούμενο βήμα μπορεί να εξελιχθεί όταν συναντώνται νέες περιπτώσεις. Αντίθετα, οι στατιστικές και οι συνδετικές προσεγγίσεις χρησιμοποιούν το στατιστικό ή το συνδετικό μοντέλο ως καθοδήγηση και δημιουργούν κανόνες ή στοιχεία δεδομένων αυτόματα, συνήθως σε σχετικά μεγάλη ποσότητα. Μετά τη δημιουργία κανόνων ή στοιχείων δεδομένων, όλες οι προσεγγίσεις εφαρμόζονται αυτόματα σε συγκεκριμένες εργασίες στο σύστημα.

Πτυχές συστήματος

Από πλευράς συστήματος, εννοούμε πηγή δεδομένων μια θεωρία ή ένα μοντέλο που σχηματίζεται από ανάλυση δεδομένων, κανόνες και βάση για αξιολόγηση.

- Δεδομένα: οι συμβολικές προσεγγίσεις χρησιμοποιούν ανθρώπινα ενδοσκοπικά δεδομένα, τα οποία συνήθως δεν είναι άμεσα ορατά.
- Θεωρία ή μοντέλο που βασίζεται στην ανάλυση δεδομένων. Ως αποτέλεσμα της ανάλυσης δεδομένων, διαμορφώνεται μια θεωρία για συμβολικές προσεγγίσεις, ενώ ένα παραμετρικό μοντέλο σχηματίζεται για στατιστικές προσεγγίσεις και ένα μοντέλο σύνδεσης σχηματίζεται για συνδετικές προσεγγίσεις.
- Κανόνες: Για συμβολικές προσεγγίσεις, το στάδιο κατασκευής κανόνα συνήθως οδηγεί σε κανόνες με λεπτομερή κριτήρια. Στις στατιστικές προσεγγίσεις, τα κριτήρια εφαρμογής κανόνα είναι συνήθως στο αρχικό επίπεδο ή απλά δεν καθορίζονται ενώ στις προσεγγιστικές οι μεμονωμένοι κανόνες συνήθως δεν μπορούν να αναγνωριστούν.
- Βάση αξιολόγησης: Η αξιολόγηση των συμβολικών συστημάτων βασίζεται συνήθως σε διαισθητικές κρίσεις μη σχετιζόμενων θεμάτων. Αντίθετα, η βάση για την αξιολόγηση των στατιστικών και των συστημάτων σύνδεσης είναι συνήθως με τη μορφή βαθμολογιών που υπολογίζονται από κάποια λειτουργία αξιολόγησης. Ωστόσο, εάν όλες οι προσεγγίσεις χρησιμοποιούνται για την ίδια εργασία, τότε τα αποτελέσματα της εργασίας μπορούν να αξιολογηθούν ποσοτικά και ποιοτικά και να συγκριθούν.
-Ανθεκτικότητα: Τα συμβολικά συστήματα μπορεί να είναι εύθραυστα όταν παρουσιάζονται με ασυνήθιστη είσοδο. Για να αντιμετωπιστεί το ασυνήθιστο, μπορεί να προβλέψουν κάνοντας τη γραμματική πιο γενική. Σε σύγκριση με τα συμβολικά συστήματα, τα στατιστικά συστήματα μπορεί να είναι πιο ανθεκτικά έναντι της ασυνήθιστης εισαγωγής με την προϋπόθεση ότι τα δεδομένα εκπαίδευσης είναι επαρκή, κάτι που μπορεί να είναι δύσκολο να διασφαλιστεί. Τα συστήματα σύνδεσης μπορεί επίσης να είναι ανθεκτικά έναντι σφάλματων, επειδή η γνώση σε τέτοια συστήματα αποθηκεύεται σε όλο το δίκτυο. Όταν παρουσιάζεται θορυβώδης είσοδος, υποβαθμίζονται σταδιακά.
-Ευελιξία: Μιας και τα συμβολικά μοντέλα κατασκευάζονται με ανθρώπινη ανάλυση καλά διατυπωμένων παραδειγμάτων, τα συμβολικά συστήματα ενδέχεται να μην έχουν την ευελιξία να προσαρμόζονται δυναμικά. Αντίθετα, τα στατιστικά συστήματα επιτρέπουν ευρεία κάλυψη και μπορεί να είναι καλύτερα σε θέση να αντιμετωπίσουν ένα μακροσκελές κείμενο για αποτελεσματικότερο χειρισμό της υπό εξέταση εργασίας. Τα συστήματα Connectionist επιδεικνύουν ευελιξία αποκτώντας δυναμικά κατάλληλη συμπεριφορά βάσει της δεδομένης εισόδου. Για παράδειγμα, τα βάρη ενός δικτύου

σύνδεσης μπορούν να προσαρμοστούν σε πραγματικό χρόνο για τη βελτίωση της απόδοσης. Ωστόσο, τέτοια συστήματα μπορεί να έχουν δυσκολία με την αναπαράσταση δομών που απαιτούνται για τον χειρισμό σύνθετων εννοιολογικών σχέσεων, περιορίζοντας έτσι τις ικανότητές τους να χειρίζονται υψηλού επιπέδου NLP.
-Κατάλληλες εργασίες: Οι συμβολικές προσεγγίσεις φαίνεται να είναι κατάλληλες για φαινόμενα που εμφανίζουν αναγνωρίσιμη γλωσσική συμπεριφορά. Μπορούν να χρησιμοποιηθούν για τη μοντελοποίηση φαινομένων σε όλα τα διάφορα γλωσσικά επίπεδα που περιγράφονται σε προηγούμενες ενότητες. Οι στατιστικές προσεγγίσεις έχουν αποδειχθεί αποτελεσματικές στη μοντελοποίηση γλωσσικών φαινομένων με βάση τη συχνή χρήση της γλώσσας, όπως αντικατοπτρίζεται στο κείμενο των εταιρειών. Γλωσσικά φαινόμενα που δεν είναι καλά κατανοητά ή δεν εμφανίζουν σαφή κανονικότητα είναι υποψήφια για στατιστικές προσεγγίσεις. Παρόμοια με τις στατιστικές προσεγγίσεις, οι συνδετικές προσεγγίσεις μπορούν επίσης να αντιμετωπίσουν γλωσσικά φαινόμενα που δεν είναι καλά κατανοητά.

Συνοψίζοντας, οι συμβολικές, στατιστικές και συνδετικές προσεγγίσεις παρουσιάζουν διαφορετικά χαρακτηριστικά, επομένως διαφορετικό πρόβλημα αντιμετωπίζεται από διαφορετική προσέγγιση ενώ οι ερευνητές έχουν αρχίσει να αναπτύσσουν υβριδικές τεχνικές που χρησιμοποιούν τα πλεονεκτήματα κάθε προσέγγισης σε μια προσπάθεια αντιμετώπισης προβλημάτων NLP πιο αποτελεσματικά και με πιο ευέλικτο τρόπο.

1.2 NLP και Αγορά Εργασίας – σύστημα αντιστοίχισης spaCy

Σύμφωνα με όλα όσα προαναφέρθηκαν, το NLP εστιάζει στην ερμηνεία, στην ανάλυση και στον χειρισμό των δεδομένων της φυσικής γλώσσας με την χρήση κατάλληλων υπολογιστικών εργαλείων και μεθόδων, ενώ επίσης αποτελεί κομμάτι του AI.

Όσον αφορά την αγορά εργασίας και την αντιστοίχιση που θα έκανε ένα τέτοιο σύστημα σε έναν υποψήφιο εργαζόμενο στην κατάλληλη γι' αυτόν θέση εργασίας, μάλλον θα κατέληγε να ειπωθεί πως είναι ένας μη αποδοτικός τρόπος προσέγγισης. Γιατί πρέπει μέσα από ένα κείμενο λέξεων να γίνει ακριβής προσδιορισμός και σύνδεση των λέξεων του βιογραφικού του υποψηφίου με αυτές της προσφερόμενης θέσης. Άρα υπάρχει χαμηλή ακρίβεια σε μεγάλου όγκου δεδομένα καθώς τα αποτελέσματα που θα επιστρέψει ενδεχομένως να μην είναι σχετικά.

Καθώς όμως οι επιχειρήσεις πρέπει να ταιριάξουν χιλιάδες υποψηφίους με τις λίστες των διαθέσιμων θέσεων από τα χιλιάδες βιογραφικά σημειώματα που λαμβάνουν, και προφανώς δεν μπορούν να το κάνουν χειροκίνητα, δημιουργήθηκε το spaCy[7], μια μοντέρνα βιβλιοθήκη Python με βάση το Natural Language Processing, που βοηθά να συμπεριλάβει τους επικρατέστερους υποψηφίους για τις λίστες με τις διαθέσιμες θέσεις και προσπαθεί να αποδείξει ότι τελικά ακόμα και αυτός ο τρόπος μπορεί να αποδειχθεί αποδοτικός.

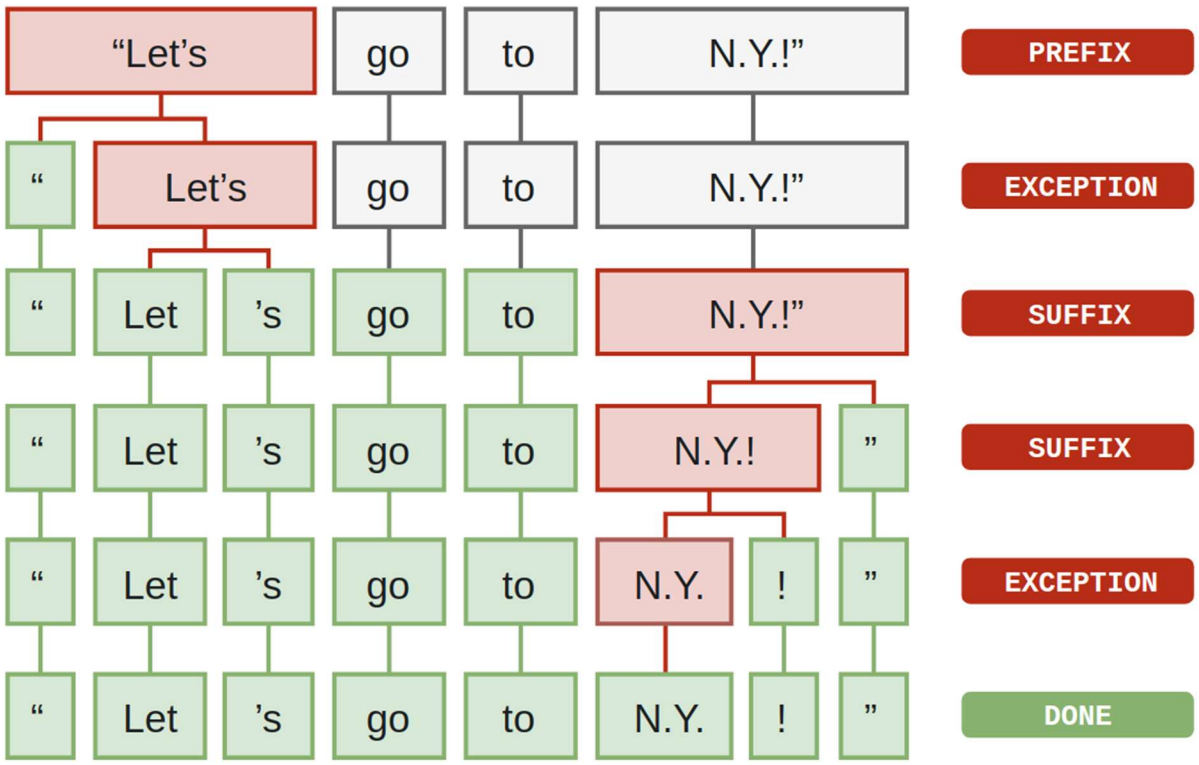
Το spaCy έχει σχεδιαστεί για να βοηθήσει τις επιχειρήσεις να κάνουν πραγματική δουλειά – να συγκεντρώσουν από τα χιλιάδες βιογραφικά πραγματικές πληροφορίες ώστε να επιτύχουν το στόχο τους. Με τις κατάλληλες βιβλιοθήκες που διαθέτει αξιοποιεί στο έπακρο τον διαθέσιμο χρόνο. Τα κύρια χαρακτηριστικά του spaCy είναι[8]:

- Ασφαλής μέθοδος τμηματοποίησης κειμένου σε λέξεις(tokenization)
- Μορφή εξαγωγής σημαντικών πληροφοριών από το κείμενο (NER – Name edentity recognition)
- Υποστηρίζει πάνω από 63 γλώσσες
- Υποστηρίζει 46 στατιστικά μοντέλα για 16 γλώσσες
- Προκαθορισμένα διανύσματα λέξεων
- Κορυφαία ταχύτητα
- Εύκολη ενσωμάτωση του deep-learning
- Επισημαίνει τα μέρη του λόγου
- Τεχνολογία labelled dependency
- Τμηματοποίηση προτάσεων βάσει σύνταξης
- Ενσωματωμένος οπτικοποιητής για σύνταξη και NER
- Αντιστοίχιση συμβολοσειρών με κατακερματισμό
- Εξαγωγή σε numpy πίνακες δεδομένων
- Αποτελεσματική δυαδική σειριοποίηση
- Ισχυρό, αυστηρό, ακριβές

Τμηματοποίηση (Tokenization)

Είναι η διαδικασία τμηματοποίησης κειμένων σε λέξεις, δευτερεύουσες λέξεις, σημεία στίξης κ.τ.λ. Αυτά τα μικρότερα τμήματα ονομάζονται tokens και συνήθως θεωρείται το δομικό στοιχείο της γλώσσας. Για κάθε γλώσσα υπάρχουν ιδιαίτεροι κανόνες, καθώς λέξεις που μπαίνουν στην αρχή ή στο τέλος μιας πρότασης ή ανάλογα τελικώς πως θα χρησιμοποιηθεί μέσα σε μια πρόταση θα δώσει διαφορετικό νόημα.

Έστω η πρόταση: "Let's go to N.Y.!", με την τμηματοποίηση παραστατικά θα έμοιαζε όπως στην εικόνα που ακολουθεί:



Εικ. 1 Τμηματοποίηση πρότασης

Επισήμανση των μερών του λόγου

Επισημαίνει τα μέρη του λόγου ή την γραμματική επισήμανση, δηλαδή αναγνωρίζει την προέλευση μιας λέξης από τον ορισμό και το περιεχόμενό της και την κατηγοριοποιεί στις γραμματικές της ιδιότητες (ουσιαστικό, ρήμα, αντικείμενο κ.τ.λ.) Παρόλο που μοιάζει απλό θεωρητικά, είναι αρκετά δύσκολο να επισημανθούν σωστά λέξεις που γράφονται και διαβάζονται με τον ίδιο τρόπο αλλά έχουν άλλη σημασία όταν συμμετέχουν στην πρόταση σαν ρήμα ή ουσιαστικό. Το μοντέλο SpaCy χρησιμοποιεί ένα στατιστικό μοντέλο που κάνει μια πρόβλεψη ανάλογα με τι ταιριάζει περισσότερο στο νόημα του κειμένου, κωδικοποιώντας όλες τις συμβολοσειρές ώστε να μειώνει την χρήση μνήμης κάνοντάς το πιο αποτελεσματικό.

Named Entity Recognition

Το NER είναι μια μορφή εξαγωγής πληροφοριών που εντοπίζει σημαντικές πληροφορίες (οντότητες) σε κείμενο. Αυτές οι οντότητες διαφέρουν από κάθε άτομο ανάλογα την διεύθυνση IP τους, από ποια χώρα προέρχονται ή ποιος είναι ο αριθμός τηλεφώνου τους. Το spaCy χρησιμοποιεί το σύστημα NER για να εξαγάγει δεξιότητες από λίστες θέσεων εργασίας και από βιογραφικά σημειώματα αντίστοιχα.

Πιο συγκεκριμένα, με βάση τον παρακάτω κώδικα σε Python αρχικά εισάγεται το αρχείο που επιλέχθηκε για να εξαχθούν τα δεδομένα (pdf ή word):

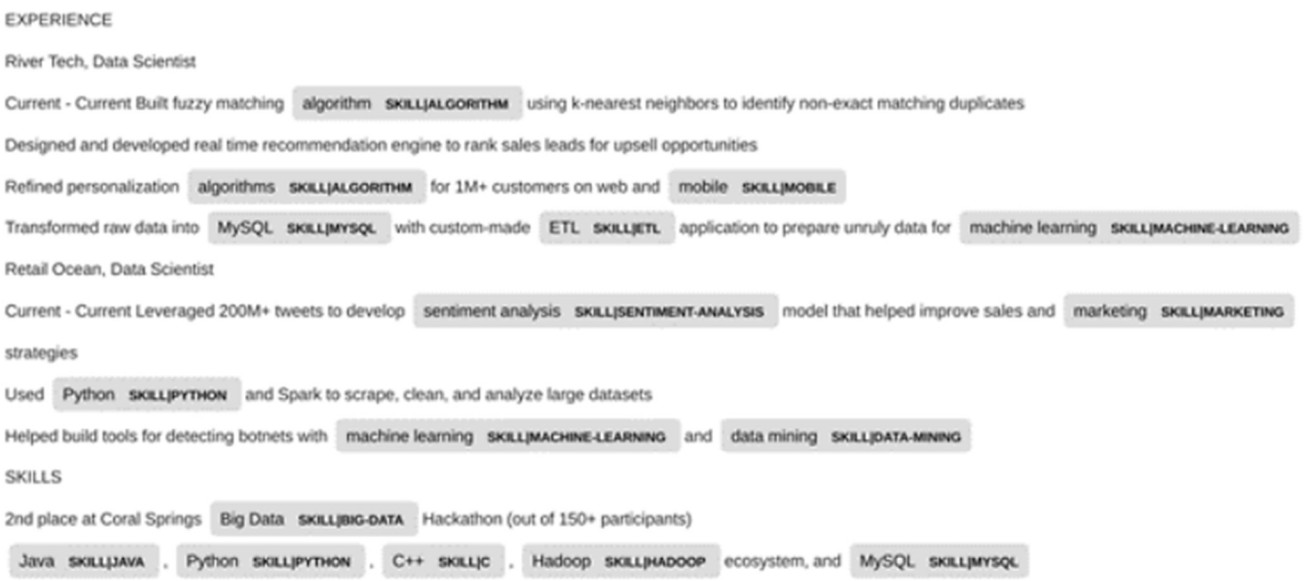
```
import textract
import PyPDF2

def extract_text_from_pdf(file):
    '''Opens and reads in a PDF file from path'''

    fileReader = PyPDF2.PdfFileReader(open(file,'rb'))
    page_count = fileReader.getNumPages()
    text = [fileReader.getPage(i).extractText() for i in range(page_count)]

    return str(text).replace("\\n", "")

def extract_text_from_word(filepath):
    '''Opens en reads in a .doc or .docx file from path'''

    txt = textract.process(filepath).decode('utf-8')

    return txt.replace('\n', ' ').replace('\t', ' ')
```

Διαβάζει τα βιογραφικά σε μορφή PDF ή Word, και αντίστοιχα ακολουθεί η διαδικασία της τμηματοποίησης ως εξής:

```
import nl_core_news_sm

# Load pre-trained Dutch language model
nlp = nl_core_news_sm.load()

# File Extension. set as 'pdf' or as 'doc(x)'
extension = 'pdf'

def create_tokenized_texts_list(extension):
    '''Create two lists, one with the names of the candidate and one with the tokenized
       resume texts extracted from either a .pdf or .doc'''
    resume_texts, resume_names = [], []

    # Loop over the contents of the directory containing the resumes, filtering by .pdf or .doc(x)
    for resume in list(filter(lambda x: extension in x, os.listdir(PROJECT_DIR + '/CV'))):
        if extension == 'pdf':
            # Read in every resume with pdf extension in the directory
            resume_texts.append(nlp(extract_text_from_pdf(PROJECT_DIR + '/CV/' + resume)))
        elif 'doc' in extension:
            # Read in every resume with .doc or .docx extension in the directory
            resume_texts.append(nlp(extract_text_from_word(PROJECT_DIR + '/CV/' + resume)))

        resume_names.append(resume.split('_')[0].capitalize())

    return resume_texts, resume_names
```

Αυτή η συνάρτηση θα διαβάσει όλα τα PDF ή Word από έναν καθορισμένο κατάλογο, θα εξαγάγει τα υποψήφια ονόματα από το όνομα του αρχείου και θα συμπληρώσει τα κείμενα πριν τα προσθέσει σε μια λίστα. Έπειτα από αυτό, το SpaCy αναλύει τα κείμενα, θα αναζητήσει τα μοτίβα όπως καθορίζονται στο αρχείο και θα επισημάνει αυτά τα μοτίβα σύμφωνα με την τιμή ετικέτας(label). Επίσης με έναν οπτικοποιητή δείχνει τι έχει επισημάνει το σύστημα NER. Για να γίνει κατανοητό, έχει εισαχθεί ένα παράδειγμα βιογραφικού και ακολουθεί το στιγμιότυπο οθόνης της εξόδου NER.

```
from spacy.pipeline import EntityRuler
from spacy import displacy
import jsonlines

# Create list with entity labels from jsonl file
with jsonlines.open(PROJECT_DIR + "data/skill_patterns.jsonl") as f:
    created_entities = [line['label'].upper() for line in f.iter()]

def add_newruler_to_pipeline(skill_pattern_path):
    '''Reads in all created patterns from a JSONL file and adds it to the pipeline after PARSER and before NER'''

    new_ruler = EntityRuler(nlp).from_disk(skill_pattern_path)
    nlp.add_pipe(new_ruler, after='parser')


def visualize_entity_ruler(entity_list, doc):
    '''Visualize the Skill entities of a doc'''

    options = {"ents": entity_list}
    displacy.render(doc, style='ent', options=options)

visualize_entity_ruler(created_entities, doc)
```



Έπειτα από αυτό, είναι σε θέση να εξάγει τις δεξιότητες από ένα βιογραφικό, κι είναι δυνατό να πραγματοποιηθεί η αντιστοίχιση αυτών των δειξιοτήτων για πολλά βιογραφικά αλλά και λίστες εργασίας. Για το επόμενο παράδειγμα χρησιμοποιήθηκε μια τυχαία λίστα εργασιών Data Science από το Indeed και έχει όπως εξής:

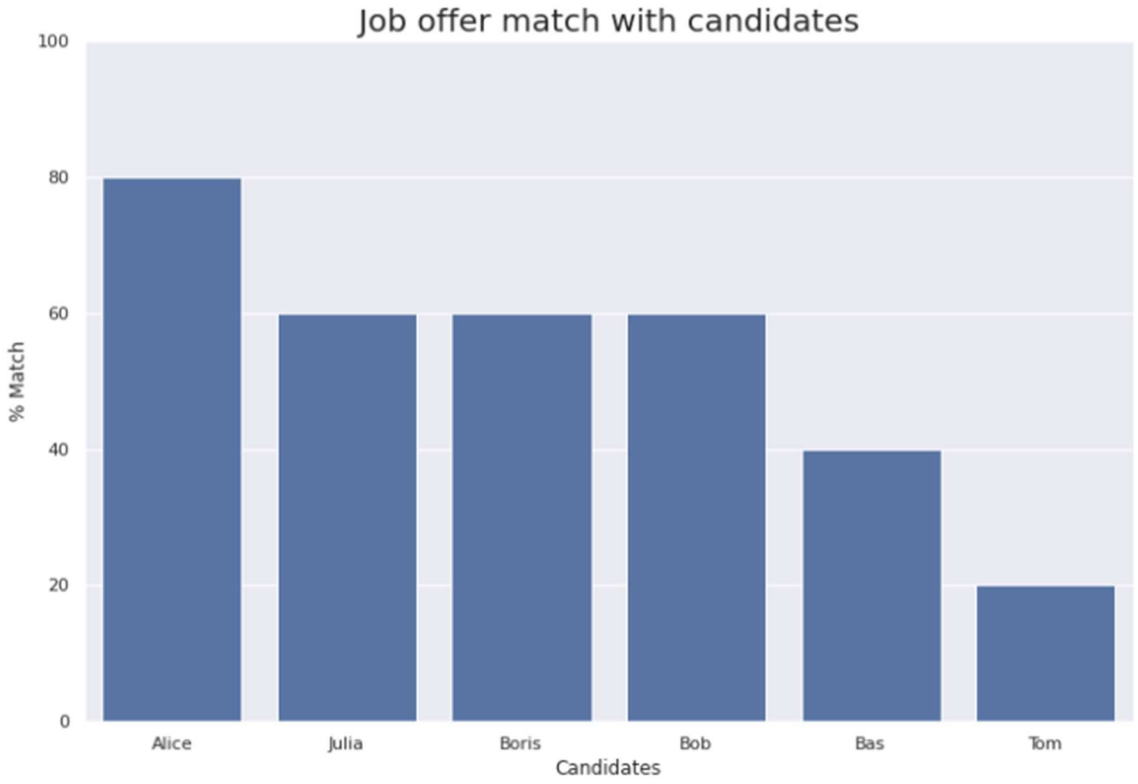
```
def create_skillset_dict(resume_names, resume_texts):
    '''Create a dictionary containing a set of the extracted skills. Name is key, matching skillset is value'''
    skillsets = [create_skill_set(resume_text) for resume_text in resume_texts]

    return dict(zip(resume_names, skillsets))

def match_skills(vacature_set, cv_set, resume_name):
    '''Get intersection of resume skills and job offer skills and return match percentage'''

    if len(vacature_set) < 1:
        print('could not extract skills from job offer text')
    else:
        pct_match = round(len(vacature_set.intersection(cv_set[resume_name])) / len(vacature_set) * 100, 0)
        print(resume_name + " has a {}% skill match on this job offer".format(pct_match))
        print('Required skills: {} '.format(vacature_set))
        print('Matched skills: {} \n'.format(vacature_set.intersection(skillset_dict[resume_name])))

        return (resume_name, pct_match)

add_newruler_to_pipeline(skill_pattern_path)

resume_texts, resume_names = create_tokenized_texts_list(extension)

skillset_dict = create_skillset_dict(resume_names, resume_texts)

# example of job offer text (string). Can input your own.
vacature_text = vacatures_df[vacatures_df['soort_vacature'] == 'Data Scientist'].skills.iloc[13]

# Create a set of the skills extracted from the job offer text
vacature_skillset = create_skill_set(nlp(vacature_text))

# Create a list with tuple pairs containing the names of the candidates and their match percentage
match_pairs = [match_skills(vacature_skillset, skillset_dict, name) for name in skillset_dict.keys()]
```

Αρχικά, δημιουργείται ένα λεξικό που περιέχει το όνομα του υποψηφίου ως κλειδί και τις δεξιότητές τους ως values. Κάνοντας χρήση των Python Sets, μπορούμε εύκολα να αναγνωρίσουμε τη διασταύρωση μεταξύ των δεξιοτήτων των υποψηφίων και των απαιτούμενων δεξιοτήτων που εξάγονται από τη λίστα θέσεων εργασίας. Ένα παράδειγμα με 6 βιογραφικά δίνει τα εξής:

```
Julia has a 60.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'DATA-SCIENCE', 'PYTHON', 'AZURE'}

Bas has a 40.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'DATA-SCIENCE', 'PYTHON'}

Boris has a 60.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'DATA-SCIENCE', 'PYTHON', 'AZURE'}

Bob has a 60.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}

Alice has a 80.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON', 'AZURE'}

Tom has a 20.0% skill match on this job offer
Required skills: {'AZURE', 'MICROSOFT-SQL-SERVER', 'DATA-SCIENCE', 'NATURAL-LANGUAGE-PROCESSING', 'PYTHON'}
Matched skills: {'PYTHON'}
```

Οπτικοποιώντας τα αποτελέσματα χρησιμοποιώντας τον παρακάτω κώδικα σε ένα γράφημα με μπάρες μας δίνει μια καλύτερη σύνοψη των αποτελεσμάτων:

```
# Sort tuples from high to low on match percentage
match_pairs.sort(key=lambda tup: tup[1], reverse=True)

# Unpack tuples
names, pct = zip(*match_pairs)

# Plotting
sns.set_theme(style='darkgrid')

fig, ax = plt.subplots(figsize=(12,8))
ax.set_title('Job offer match with candidates', fontsize=20)
ax.set(xlabel='Candidates', ylabel='% Match')
ax.set(ylim=(0, 100))


sns.set(font_scale=1.5)
sns.barplot(x=list(names), y=list(pct), color='b')

plt.show()
```



Εικ. 2 Αποτελέσματα αντιστοίχισης βιογραφικών με προσφερόμενες θέσεις εργασίας

Η Alice έχει 80% πιθανότητες να ταιριάξει το βιογραφικό της με κάποια πιθανή θέση εργασίας, η Julia, ο Boris και ο Bob από 60%, ο Bas 40% ενώ ο Tom μόλις 20%.

Παρουσιάστηκε αναλυτικά μια πιθανή λύση σε ένα πρόβλημα που πολλά τμήματα Ανθρώπινου Δυναμικού αντιμετωπίζουν. Παρόλο που αυτή η μέθοδος ανάλυσης και αντιστοίχισης των βιογραφικών είναι βελτιωμένη και οι δεξιότητες δεν είναι ο μόνος τρόπος αξιολόγησης των πιθανών υποψηφίων, μπορεί να εξοικονομήσει αρκετό χρόνο χρησιμοποιώντας αυτό το σενάριο για να δημιουργηθεί μια λίστα ορισμένων υποψηφίων.

2. Relevance-based Models

2.1 Τι είναι τα Relevance-based Models

Τα Relevance-based models επινοήθηκαν από τους Stephen E. Robertson και Karen Spärck Jones ως πλαίσιο για πιθανά μελλοντικά μοντέλα. Είναι μια τυποποίηση ανάκτησης πληροφοριών, χρήσιμη για την απόκτηση λειτουργιών κατάταξης που χρησιμοποιούνται από μηχανές αναζήτησης και μηχανές αναζήτησης Ιστού, προκειμένου να ταξινομήσουν τα αντίστοιχα έγγραφα ανάλογα της συνάφειά τους με ένα δεδομένο ερώτημα αναζήτησης.

Είναι ένα θεωρητικό μοντέλο που εκτιμά την πιθανότητα ότι ένα έγγραφο dj είναι σχετικό με ένα ερώτημα q. Το μοντέλο υποθέτει ότι αυτή η πιθανότητα συνάφειας εξαρτάται από το ερώτημα και τις παραστάσεις των εγγράφων. Επιπλέον, υποθέτει ότι υπάρχει ένα τμήμα όλων των εγγράφων που προτιμά ο χρήστης και αποτελεί το σύνολο απαντήσεων για το ερώτημα q. Ένα τέτοιο ιδανικό σύνολο απαντήσεων ονομάζεται R και πρέπει να μεγιστοποιήσει τη συνολική πιθανότητα συνάφειας με αυτόν τον χρήστη. Η πρόβλεψη είναι ότι τα έγγραφα σε αυτό το σύνολο R σχετίζονται με το ερώτημα, ενώ τα έγγραφα που δεν υπάρχουν στο σύνολο δεν είναι σχετικά.

Μια αρχική προσέγγιση για το πρόβλημα που δημιουργείται από τα μη σχετικά έγγραφα τα οποία μπορεί να απορρίπτονται λανθασμένα, είναι τα Structured Relevance Models (SRM). Στο άρθρο από το 2007 των Xing Yi – James Allan και W. Bruce Croft [9], περιγράφονται 3 μοντέλα συσχέτισης που βασίζονται στο ότι κάποια στοιχεία για ένα πεδίο μπορούν να συσχετιστούν με το περιεχόμενο ενός άλλου πεδίου μέσα σε ένα σύνολο δεδομένων δίνοντας την ευκαιρία να χρησιμοποιηθούν δεδομένα που έχουν μια κάποια σχέση ενώ κατά κανόνα δεν απαντούν πλήρως στα ερωτήματα ή δεν αποδίδουν το μέγιστο στην αντιστοίχιση. Δηλαδή αν σε μια δημοσίευση αγγελίας στον τίτλο αναφέρεται «Διαχειριστής Βάσης Δεδομένων», και οι υποψήφιοι που θα αιτηθούν αναφέρουν στις δεξιότητες την γνώση «SQL server»ή «MySQL» αυτό και μόνο είναι αρκετό ώστε από τον τίτλο να γίνει αντιστοίχιση με την συγκεκριμένη δεξιότητα που αναφέρεται σε άλλο πεδίο.

2.2 Structured Relevance Models

Οι εργοδότες που συνεργάζονται με ιστοσελίδες αναζήτησης εργασίας, υποβάλλουν την αγγελία για την προσφερόμενη θέση εργασίας μαζί με λοιπά στοιχεία που χρειάζεται να έχει ο υποψήφιος εργαζόμενος. Εκείνοι που αναζητούν εργασία υποβάλλουν στις έτοιμες φόρμες των ιστοσελίδων αυτών τα στοιχεία τους. Αυτές οι φόρμες αποτελούνται από πεδία ερωτήσεων ανοιχτού τύπου (ελεύθερο κείμενο), ή κλειστού τύπου (επιλογή προκαθορισμένων απαντήσεων). Είναι κατανοητό πως απαντώντας σε όλες τις ερωτήσεις, δημιουργείται ένα online νέο βιογραφικό σημείωμα που στόχο έχει

να αποδώσει στο μέγιστο βαθμό και να αναδείξει τις δεξιότητες και ενδεχομένως αν ζητούνται στοιχεία από τον χαρακτήρα του υποψηφίου.

Απαντώντας σε όλες τις ερωτήσεις, ανοιχτού-κλειστού τύπου, δημιουργούνται τα δεδομένα των υποψηφίων στην βάση δεδομένων των ιστοσελίδων. Σε αυτή την περίπτωση αναφερόμαστε σε δομημένα δεδομένα όπως λέγονται τα οποία είναι και εύκολα να επεξεργαστούν. Εάν ο υποψήφιος δεν απαντήσει σε όλες τις ερωτήσεις, αφήνοντας δηλαδή πεδία μη συμπληρωμένα, τότε δημιουργούνται τα λεγόμενα ημι-δεδομένα.

Σε ένα ιδεατό σύστημα, τα βιογραφικά θα ήταν τέλεια συμπληρωμένα και οι προσφερόμενες θέσεις εργασίας θα ήταν απολύτως κατανοητές και πλήρως κατατοπιστικές και η τέλεια αντιστοίχιση θα γινόταν χωρίς κανένα απολύτως πρόβλημα. Όταν όμως χρησιμοποιείται ένα σχεσιακό σύστημα αντιστοίχισης και τα δεδομένα που χρησιμοποιούνται έχουν σοβαρές ελλείψεις, τότε τα αποτελέσματα δεν είναι τα αναμενόμενα. Στα σχεσιακά μοντέλα λόγω των αναπάντητων ερωτήσεων δημιουργούνται κενά στα δεδομένα που συλλέγονται με αποτέλεσμα λέξεις κλειδιά για την αντιστοίχιση να λείπουν και έτσι να απορρίπτονται άσκοπα ή ακόμα και σε συμπληρωμένες ερωτήσεις ανοιχτού τύπου να έχουν χρησιμοποιηθεί λέξεις ή εκφράσεις που δεν παραπέμπουν στην προκαθορισμένη αντιστοίχιση που έχει ορισθεί και έτσι ομοίως να απορρίπτονται αφού δεν υπάρχει σύνδεση ώστε να αντιστοιχιστούν. Ενα σύνολο ημι-δομημένων δεδομένων βιογραφικών σημειωμάτων και ένα σύνολο από ημι-δομημένα δεδομένα αγγελιών εργασίας συσχετίζοντας τα θα αποδώσουν μια λίστα από σχετικού περιεχομένου βιογραφικά σημειώματα για κάθε υπάρχουσα ή νέα αγγελία εργασίας ή και αντίστοιχα μια λίστα από σχετικές αγγελίες εργασίας για κάθε υπάρχων βιογραφικό σημείωμα.

Αναλυτική περιγραφή

1. sLM (simple language modeling): εκτελείται αγνοώντας την δομή των δεδομένων, χωρίζει δηλαδή τη δομή των βιογραφικών και των αγγελιών διαιρώντας το ελεύθερο κείμενο από το κείμενο των υπολοίπων πεδίων, τα βιογραφικά χρησιμοποιούνται ως τα δεδομένα του μοντέλου ενώ οι αγγελίες ως τα ερωτήματα αυτού. Αναμένεται βέβαια να είναι αναποτελεσματικό καθώς δεν μπορεί να γεφυρώσει το λεξιλόγιο ανάμεσα στα βιογραφικά σημειώματα και στις αγγελίες εργασίας.
2. tRM (true relevance model): αποτελεί παραλλαγή των Relevance Models, εκτελείται αγνοώντας και αυτό την δομή αλλά αξιοποιεί τα ζεύγη που έχουν κριθεί σε ένα είδος εποπτευόμενης επέκτασης ερωτήματος (supervised query expansion) και χωρίζεται σε τρία στάδια:

a. Η απλοποιημένη εγγραφή της αγγελίας χρησιμοποιείται ως ερώτημα για το μοντέλο και όχι το σύνολο των ομοειδών αγγελιών όπως έχουν αρχικά συνταχθεί
b. Το σύνολο των ομοιειδών βιογραφικών σημειωμάτων που ταιριάζουν κατά την κρίση του χρήστη στο ερώτημα χρησιμοποιούνται για να δημιουργηθεί ένα κοινό λεξιλόγιο για την συσχέτιση
c. Εκτελείται το σχεσιακό μοντέλο με την απλοποιημένη λίστα βιογραφικών σημειωμάτων και την αρχική συνολική λίστα που ταιριάζουν με το ερώτημα μας

3. SRM: πρόκειται για ένα μοντέλο που χρησιμοποιεί σχεσιακές πληροφορίες αντίστοιχα με το tRM, όμως εκείνο που το διαχωρίζει είναι πως χρησιμοποιεί την δομή των πεδίων καθώς και την αλληλεξάρτησή τους. Ακολουθεί λοιπόν ακριβώς τα ίδια τρία βήματα του tRM αλλά διαφέρει καθώς λειτουργεί πολύ διαφορετικά λόγω των πολλαπλών πεδίων που χρησιμοποιεί.

Κάθε πεδίο από μια ορισμένη αγγελία εργασίας ορίζεται ως ερώτημα του μοντέλου και εκτελείται για κάθε αντίστοιχο πεδίο των ημι-δομημένων δεδομένων του συνόλου των αγγελιών συγχωνεύοντας το συγκεκριμένο πεδίο της αγγελίας χρησιμοποιώντας εντροπία με βάρη, διατηρώντας εκείνες μόνο που έλαβαν την μεγαλύτερη βαθμολόγηση. Όπως συμβαίνει και στο tRM, το SRM αποδίδει ένα σύνολο αγγελιών εργασίας που είναι όμοια με το ερώτημα του μοντέλου. Σε αντίθεση με το tRM, το SRM χρησιμοποιεί για αυτό το αποτέλεσμα συγκεκριμένα κομμάτια των αγγελιών εργασίας.

Στο δεύτερο βήμα αυτού του μοντέλου, χρησιμοποιούνται βιογραφικά σημειώματα τα οποία έχουν σχέση με το ερώτημα του μοντέλου και των υπολοίπων αγγελιών εργασίας δημιουργώντας έτσι ένα σχεσιακό μοντέλο, όμως δημιουργεί και επιπλέον «μοντέλα» για κάθε πεδίο του κάθε βιογραφικού σημειώματος.

Στο τρίτο βήμα, το SRM εκτελείται για κάθε επιπλέον μοντέλο του προηγούμενου βήματος χρησιμοποιώντας τα πεδία του μοντέλου ως ερωτήματα τα οποία έχουν βάρη, στην συνέχεια βαθμολογείται κάθε βιογραφικό σημείωμα και ανάλογα το σύνολο των βαθμών κατατάσσεται σαν συναφές ή μη.

Δεδομένα - αποτελέσματα - σύγκριση

Κάθε βιογραφικό σημείωμα ή αγγελία εργασίας αποτελεί σε όλα τα μοντέλα εγγραφή, στην οποία κάποια πεδία δεν είναι συμπληρωμένα ή είναι κενά, κάποια έχουν κείμενο σε φυσική γλώσσα, κάποια ενδεχομένως αριθμούς. Συνολικά, χρησιμοποιήθηκαν 1.276.573 βιογραφικά σημειώματα τα οποία διασπάστηκαν σε 90 πεδία, 12 εξ αυτών αποτελούν πεδία κειμένου. Όσον αφορά τις αγγελίες εργασίας, χρησιμοποιήθηκαν 206.393 οι οποίες διασπάστηκαν σε 20 πεδία, 9 εξ αυτών αποτελούν πεδία κειμένου,

ενώ χρησιμοποιήθηκαν και 1.820.420 ζευγάρια αγγελιών – βιογραφικών σημειωμάτων σχολιασμένα και δοσμένα από πράκτορες εργασίας.

Από τα πειράματα που αρχικά διενεργήθηκαν, συλλέχθηκαν 300 αγγελίες εργασίας που είχαν συσχετισμένα 60 – 80 βιογραφικά σημειώματα. Αυτό το δείγμα διαιρέθηκε σε 2 μέρη, το ένα αποτέλεσε το δείγμα για την εκπαίδευση του μοντέλου και το άλλο για την εκτέλεσή του. Αντίστοιχα, διαιρέθηκε ισόποσα το σύνολο των βιογραφικών σημειωμάτων και χρησιμοποιήθηκε με τον ίδιο τρόπο. Το κομμάτι που αφορά στην εκπαίδευση του μοντέλου, χρησιμοποιήθηκε και για να δημιουργηθεί το σχεσιακό μοντέλο αναζητώντας βιογραφικά στόχου στο κομμάτι της εκτέλεσης. Όταν ενσωματώθηκε η δομή για το μοντέλο SRM, χρησιμοποιήθηκε ο τίτλος και το κεντρικό πεδίο τόσο των αγγελιών εργασίας όσο και των βιογραφικών σημειωμάτων

Ο Πίνακας 2 παρουσιάζει την εκτέλεση του SRM μοντέλου έναντι των υπολοίπων δύο. Συσχετίστηκαν 150 αγγελίες για την εκπαίδευση του μοντέλου με το κομμάτι των βιογραφικών σημειωμάτων που κρατήθηκε για εκπαίδευση. Το άνω μισό μέρος του πίνακα παρουσιάζει την ακρίβεια σε σταθερές ανακλήσεις ενώ το υπόλοιπο μισό του πίνακα σε ακρίβεια διαφορετικών ranks. Η στήλη % παρουσιάζει την συσχέτιση μεταξύ SRM & tRM μοντέλων ενώ η τελευταία στήλη παρουσιάζει τον αριθμό των διασταυρώσεων που το SRM ξεπέρασε το tRM.

Εκτέλεση SRM έναντι των sLM - tRM					
	sLM	tRM	SRM	% σχετική διαφορά μεταξύ SRM - tRM	Αριθμός διασταυρώσεων
Επίπεδα ανάκλησης	242	1134	1255	10,67	74/116
Ενδιάμεσες ανακλήσεις - Ακρίβεια					
0,00	0,0299	0,2707	0,3133	15,70	78/120
0,10	0,0043	0,1547	0,1836	18,60	72/100
0,20	0,0019	0,1263	0,1439	13,90	47/63
0,30	0,0009	0,0839	0,0942	12,20	25/35
0,40	0,0003	0,0580	0,0625	7,90	18/24
Μέση ακρίβεια	0,0018	0,0638	0,0726	13,89	101/147
Ακρίβεια σε δεδομένα					
5 έγγραφα	0,0093	0,1627	0,1947	19,7	23 / 33
10 έγγραφα	0,0073	0,146	0,174	19,2	31 / 41
15 έγγραφα	0,0067	0,1289	0,1462	3,4	34 / 51
20 έγγραφα	0,007	0,1113	0,128	15	40 / 58
30 έγγραφα	0,0053	0,0876	0,1036	18,3	48 / 61

Σχετική ακρίβεια	0,0055	0,0824	0,0963	16,95	52/68

Πίνακας 2 Αποτελέσματα Σύγκρισης των μοντέλων

Συνολικά τα αποτελέσματα των μοντέλων δείχνουν πως ένα κλασσικό μοντέλο συσχέτισης όπως το sLM αποδίδει ελάχιστα αποδεικνύοντας πως από ένα κείμενο με το συγκεκριμένο τύπο μοντέλου δεν μπορεί να αντιστοιχήσει βιογραφικά σημειώματα με αγγελίες εργασίας. Τα σχεσιακά μοντέλα επιτυγχάνουν καλύτερη εκτέλεση μέσω της ανατροφοδότησης. Παρόλα αυτά, όταν η δομή παρέχεται το SRM ξεπερνά το tRM κατά 14%. Λόγω της συνολικής πολυπλοκότητας της συσχέτισης των βιογραφικών σημειωμάτων με τις αγγελίες εργασίας και των ημι-δομημένων δεδομένων, λιγότερο από το 20% (35 συνδεδεμένα βιογραφικά σημειώματα με αγγελία εργασίας από τα 60-80 του δείγματος) αποδείχθηκαν σχετικά με το μοντέλο SRM. Για να διερευνηθεί περαιτέρω, τα 150 του δείγματος των αγγελιών εργασίας χωρίστηκαν σε 3 ομάδες με βάση το 10 στην ακρίβεια. Αποδείχθηκε πως για κάποιες αγγελίες, τα μοντέλα συνάφειας βρίσκουν περισσότερα από 5 συνδυασμένα βιογραφικά σημειώματα στα 10 κορυφαία της λίστας.

2.3 Relevance based Models _ Word Embedding

Πιο πρόσφατες μελέτες για τα Relevance – based models αλλά και για τα NLP, θέλουν μέσω πιο εξελιγμένων αλγορίθμων (word embendding algorithms) να γίνεται το ταίριασμα των λέξεων όσο το δυνατό πιο κοντά στο πραγματικό, δεν στηρίζονται στο ταίριασμα μέσω της σύνταξης ή της σημασίας που έχουν οι λέξεις, αλλά ταίριασμα λέξεων που υπάρχει πραγματική αλληλεξάρτηση μεταξύ τους.

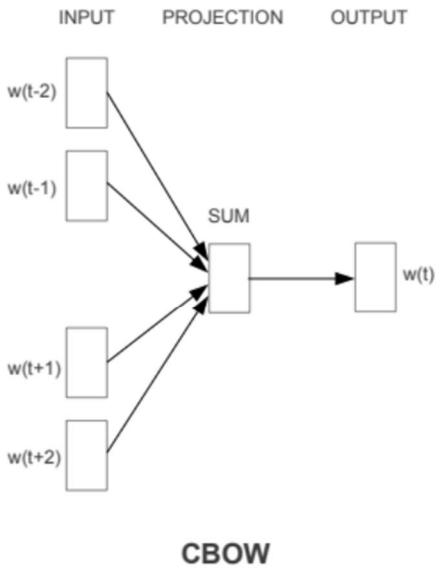
Σε άρθρο του 2017 των Hamed Zamani & W. Bruce Croft [10], αναλύονται κείμενα ψάχνοντας συνάφεια μεταξύ δεδομένων λέξεων ενός κειμένου ή ακόμη και σε ολόκληρο κείμενο. Το κίνητρο λοιπόν γι’ αυτή την έρευνα είναι να αναπτυχθεί ένα μη επιβλεπόμενο Relevance-based Model αναπαράστασης λέξεων, στο οποίο η εκμάθησή του στηρίζεται σε πληροφορίες σχετικές με το έγγραφο.

Παρόλο που η ανάλυση των παρακάτω μοντέλων δεν αναφέρεται ρητά σε ταίριασμα βιογραφικών σημειωμάτων και αγγελιών εργασίας, παρουσιάζεται μια πιο γενική εικόνα ταιριάσματος λέξεων από κείμενα, ωστόσο θα μπορούσε να ειπωθεί πως μπορούν να προβλεφθούν οι λέξεις που σχετίζονται με μια συγκεκριμένη ανάγκη πληροφοριών, δηλαδή την ανάγκη ταιριάσματος βιογραφικού σημειώματος και αγγελίας εργασίας και να δοθούν αξιοσημείωτα αποτελέσματα που αγγίζουν το τέλειο δυνατό αποτέλεσμα.

Αναλύονται λοιπόν δύο μοντέλα (word2vec, GloVe) με διαφορετικές αντικειμενικές συναρτήσεις, το ένα μαθαίνοντας μια κατανομή συνάφειας στο σύνολο λεξιλογίου για κάθε ερώτημα και το άλλο ταξινόμησης κάθε όρου που ανήκει στη σχετική ή μη σχετική τάξη για κάθε ερώτημα.

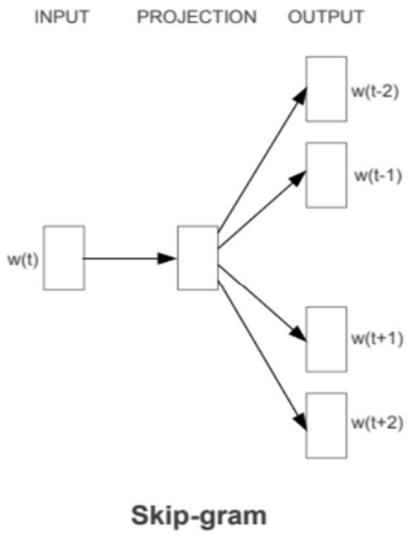
Word2vec

Δημοφιλή μοντέλα για την εκμάθηση της αναπαράστασης λέξεων είναι τα νευρωνικά μοντέλα που βασίζονται σε δίκτυο. Το μοντέλο word2vec που προτάθηκε από τους Mikolovetal είναι ένα μοντέλο ενσωμάτωσης που μαθαίνει διανύσματα λέξεων μέσω ενός νευρωνικού δικτύου με ένα μόνο κρυφό επίπεδο. Το CBOW[11] και το skip-gram[12] είναι δύο υλοποιήσεις του μοντέλου word2vec. Η αρχιτεκτονική το μοντέλου CBOW (Continuous bag of words) προσπαθεί να προβλέψει την τρέχουσα λέξη-στόχο (την κεντρική λέξη) με βάση τις λέξεις που βρίσκονται γύρω της.



The CBOW model architecture (Source: https://arxiv.org/pdf/1301.3781.pdf Mikolov el al.)

Εικ. 3 Αρχιτεκτονική μοντέλου CBOW



The Skip-gram model architecture (Source: https://arxiv.org/pdf/1301.3781.pdf Mikolov el al.)

Εικ. 4 Αρχιτεκτονική μοντέλου Skip-gram

Όσον αφορά το Skip-gram, συνήθως προσπαθεί να επιτύχει το αντίστροφο από αυτό που κάνει το μοντέλο CBOW. Προσπαθεί να προβλέψει τις λέξεις περιβάλλοντος – δηλαδή γύρω από λέξεις, με δεδομένη μια λέξη-στόχο - την κεντρική λέξη.

GloVe

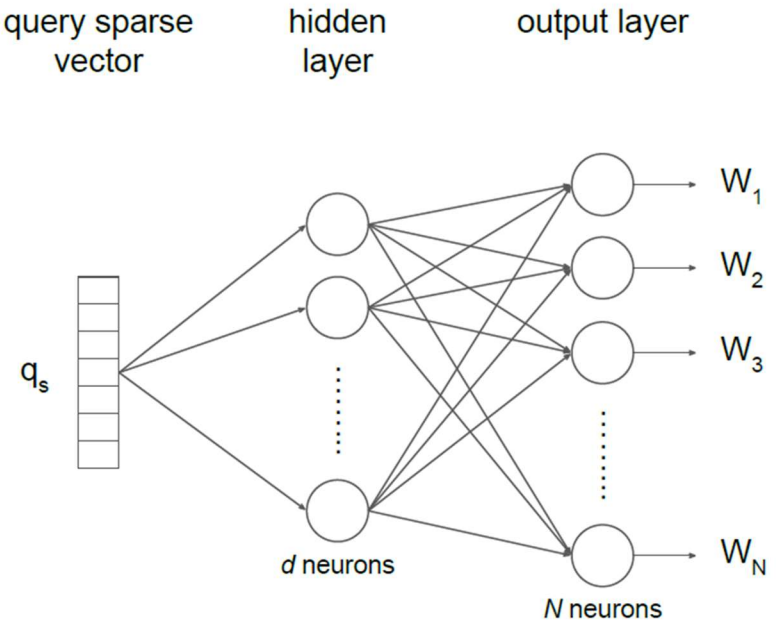
Ένα άλλο γνωστό μοντέλο που εκπαιδεύει διανύσματα ή λέξεις από τις πληροφορίες για τη συνύπαρξή τους, δηλαδή πόσο συχνά εμφανίζονται μαζί σε μεγάλα κείμενα, είναι το Global Vectors (GloVe)[13]. Ενώ το word2vec είναι ένα μοντέλο πρόβλεψης - ένα νευρωνικό δίκτυο τροφοδοσίας που εκπαιδεύει διανύσματα για τη βελτίωση της ικανότητας πρόβλεψης, το GloVe είναι ένα μοντέλο βασισμένο σε μετρήσεις. Σε γενικές γραμμές, τα μοντέλα που βασίζονται σε μετρήσεις εκπαιδεύουν διανύσματα κάνοντας μείωση διαστάσεων σε έναν πίνακα. Αρχικά κατασκευάζουν έναν μεγάλο πίνακα πληροφοριών συνύπαρξης, ο οποίος περιέχει τις πληροφορίες σχετικά με το πόσο συχνά κάθε «λέξη» (αποθηκεύεται σε σειρές), εμφανίζεται σε κάποιο «πλαίσιο» (οι στήλες). Ο αριθμός των «περιεχομένων» πρέπει να είναι μεγάλος, δεδομένου ότι είναι ουσιαστικά συνδυαστικό σε μέγεθος. Στη συνέχεια παραγοντοποιείται αυτός ο πίνακας για να αποδώσουν έναν πίνακα χαμηλότερων διαστάσεων λέξεων και χαρακτηριστικών, όπου κάθε σειρά αποδίδει μια διανυσματική αναπαράσταση για κάθε λέξη. Επιτυγχάνεται ελαχιστοποιώντας μια «απώλεια ανοικοδόμησης» που αναζητά αναπαραστάσεις χαμηλής διάστασης και μπορούν να εξηγήσουν τη διακύμανση των δεδομένων υψηλής διάστασης.

Ο στόχος που εξετάζεται αναφέρεται στην ανάπτυξη μοντέλων συνάφειας μιας τελευταίας τεχνολογίας ψευδο-προσέγγισης ανατροφοδότησης σχετικότητας. Τα μοντέλα συνάφειας προσπαθούν να βελτιστοποιήσουν αυτόν τον στόχο δεδομένου ενός συνόλου σχετικών εγγράφων για ένα δεδομένο ερώτημα ως ένδειξη της ανάγκης ενημέρωσης του χρήστη. Ελλείψει πληροφοριών σχετικότητας, τα κορυφαία έγγραφα που ανακτώνται σε απάντηση στο ερώτημα θεωρείται ότι είναι σχετικά. Επομένως, τα μοντέλα συνάφειας και, γενικά, όλα τα μοντέλα ψευδο-συνάφειας, χρησιμοποιούν μια διαδικτυακή αναζήτηση για να λάβουν τα δεδομένα που θα χρειαστούν για εκπαίδευση: ανάκτηση εγγράφων για το ερώτημα και στη συνέχεια χρήση των εγγράφων που ανακτήθηκαν για να εκτιμηθεί η κατανομή της συνάφειας.

Word Embedding – δεδομένα, αλγόριθμοι, εκπαίδευση

Ως δεδομένα για την εκτέλεση των μοντέλων, συλλέχθηκαν τα κορυφαία έγγραφα που ανακτήθηκαν από εκατομμύρια ερωτήσεις ως training set και ενσωματώνοντας διανύσματα για κάθε όρο, προκειμένου να προβλεφθούν οι λέξεις που παρατηρούνται σ' αυτά τα έγγραφα για κάθε ερώτημα. Θα αναλυθούν δύο μοντέλα ενσωμάτωσης λέξεων που βασίζονται στη συνάφεια, το πρώτο το μοντέλο μεγιστοποίησης πιθανότητας συνάφειας (RLM), στοχεύει στη μοντελοποίηση της κατανομής συνάφειας έναντι των όρων λεξιλογίου για κάθε ερώτημα, ενώ το δεύτερο το μοντέλο μεταγενέστερης εκτίμησης συνάφειας (RPE), ταξινομεί κάθε όρο ως σχετικό ή μη με κάθε ερώτημα. Παρέχονται αλγόριθμοι e-learning για την εκπαίδευση αυτών των μοντέλων σε μεγάλες ποσότητες δεδομένων εκπαίδευσης (deep learning), ενώ δεν ελέγχονται τα δεδομένα εκπαίδευσης αλλά δημιουργούνται αυτόματα.

Για να αξιολογηθούν τα μοντέλα, πραγματοποιήθηκαν δύο σειρές εξωγενών αξιολογήσεων. Το πρώτο σύνολο, εστιάζει στην επέκταση ερωτήματος για ανάκτηση ad-hoc, εξετάζονται τέσσερις συλλογές TREC, συμπεριλαμβανομένων δύο συλλογών news wire (AP και Robust) και δύο συλλογών ιστού μεγάλης κλίμακας (GOV2 και ClueWeb09 - Cat. B). Το δεύτερο σύνολο πειραμάτων, εστιάζει στην εργασία ταξινόμησης ερωτημά-των χρησιμοποιώντας το σύνολο δεδομένων KDD Cup 2005.



Εικ. 5 Αρχιτεκτονική ενός relevance –based word embedding

Τα σχόλια μέσω ανατροφοδότησης σχετικά με τη συνάφεια έχουν αποδειχθεί εξαιρετικά αποτελεσματικά στη βελτίωση της απόδοσης ανάκτησης. Σε σχέση με τα σχόλια, ένα σύνολο σχετικών εγγράφων σε ένα δεδομένο ερώτημα λαμβάνεται υπόψη

για την εκτίμηση μοντέλων ερωτημάτων. Δεδομένου ότι τα ρητά σήματα συνάφειας για ένα δεδομένο ερώτημα δεν είναι πάντα διαθέσιμα, τα ψευδο-σχετικά σχόλια (PRF) υποθέτουν ότι τα κορυφαία έγγραφα που ανακτήθηκαν ως απόκριση στο δεδομένο ερώτημα σχετίζονται με το ερώτημα και χρησιμοποιούνται για να εκτιμήσουν τα μοντέλα ερωτήσεων. Η αποτελεσματικότητα του PRF σε διάφορα σενάρια ανάκτησης δείχνει ότι από τα κορυφαία έγγραφα που ανακτήθηκαν μπορούν να ληφθούν χρήσιμες πληροφορίες. Πρέπει να σημειωθεί ότι υπάρχει σημαντική διαφορά μεταξύ του PRF και των προτεινόμενων μοντέλων: Στο PRF, το μοντέλο ανατροφοδότησης εκτιμάται από τα κορυφαία έγγραφα που ανακτήθηκαν με βάση ένα δεδομένο ερώτημα σε μια διαδικτυακή αναζήτηση, ανακτά τα έγγραφα για το αρχικό ερώτημα και στη συνέχεια γίνεται η εκτίμηση χρησιμοποιώντας τα έγγραφα αυτά που ανακτήθηκαν.

Η εκπαίδευση του μοντέλου γίνεται offline, καθώς οδηγεί σε σημαντικές βελτιώσεις στην αποτελεσματικότητα, επειδή δεν απαιτείται επιπλέον εκτέλεση ανάκτησης στην επιλογή. Για να εκπαιδευτεί ένα μοντέλο σε μια αναζήτηση, επιλέγεται ένα διάνυσμα μεγάλου μήκους για κάθε όρο λεξιλογίου και εκτιμάται αυτό το διάνυσμα με βάση τις πληροφορίες που εξήχθησαν από τα κορυφαία έγγραφα που ανακτήθηκαν για μεγάλο αριθμό εκπαιδευτικών ερωτημάτων.

Word Embedding – αποτελέσματα και σύγκριση μεταξύ των μοντέλων, αξιολόγηση μέσω Query Expansion

Για να αξιολόγηση των μοντέλων, εξετάστηκαν οι ακόλουθες γραμμές βάσης:

1. την τυπική εκτίμηση μέγιστης πιθανότητας (MLE) του μοντέλου ερωτημάτων χωρίς επέκταση ερωτήματος,
2. δύο σύνολα διανυσμάτων ενσωμάτωσης από εκπαίδευση μέσω word2vec model 6
3. δύο σύνολα διανύσματος ενσωμάτωσης από εκπαίδευση μέσω του μοντέλου GloVe.

Για την αποτελεσματικότητα χρησιμοποιήθηκαν τρεις τυπικές μετρήσεις αξιολόγησης: μέση ακρίβεια (MAP) των κορυφαίων 1000 εγγράφων, ακρίβεια των κορυφαίων 20 ανακτημένων εγγράφων (P @ 20) και κανονικοποιημένο μειωμένο σωρευτικό κέρδος υπολογίστηκε για τα κορυφαία 20 ανακτημένα έγγραφα (nDCG @ 20). Οι στατιστικές σημαντικές διαφορές των τιμών MAP, P @ 20 και nDCG @ 20 με βάση τη δοκιμή t-tailed two-tailed υπολογίζονται σε επίπεδο εμπιστοσύνης 95% (δηλαδή, τιμή p <0,05).
Τα αποτελέσματα φαίνονται στον Πίνακα 3:

Collection	Metric	MLE	word2vec		GloVe		Rel.-based Embedding	
			external	target	external	target	RLM	RPE
AP	MAP	0.2197	0.2399	0.2420	0.2319	0.2389	$\mathbf{0.2580}^{01234}$	0.2543^{01234}
	P@20	0.3503	0.3688	0.3738	0.3581	0.3631	$\mathbf{0.3886}^{01234}$	0.3812^{034}
	NDCG@20	0.3924	0.4030	0.4181	0.4025	0.4098	$\mathbf{0.4242}^{01234}$	0.4226^{01234}
Robust	MAP	0.2149	0.2218	0.2215	0.2209	0.2172	$\mathbf{0.2450}^{01234}$	0.2372^{01234}
	P@20	0.3319	0.3357	0.3337	0.3345	0.3281	$\mathbf{0.3476}^{01234}$	0.3409^{024}
	NDCG@20	0.3863	0.3918	0.3881	0.3918	0.3844	$\mathbf{0.3982}^{01234}$	0.3955^{0}
GOV2	MAP	0.2702	0.2740	0.2723	0.2718	0.2709	$\mathbf{0.2867}^{01234}$	0.2855^{01234}
	P@20	0.5132	0.5257	0.5172	0.5186	0.5128	$\mathbf{0.5367}^{01234}$	0.5358^{01234}
	NDCG@20	0.4482	0.4571	0.4509	0.4539	0.4485	$\mathbf{0.4576}^{0234}$	0.4557^{024}
ClueWeb	MAP	0.1028	0.1033	0.1033	0.1029	0.1026	$\mathbf{0.1066}^{01234}$	0.1031
	P@20	0.3025	0.3040	0.3053	0.3033	0.3048	**0.3073**	0.3030
	NDCG@20	0.2237	0.2235	0.2252	0.2244	0.2244	$\mathbf{0.2273}^{01}$	0.2241

Πίνακας 3 Αξιολόγηση relevance-based word embeddings με query expansion

* AP – Robust: test homogeenous collections consists thousands of new articles
GOV2 – ClueWeb: large scale web collections containing heterogeneous documents
MLE – maximum likelihood estimation
MAP – mean average precision

Σύμφωνα με τον πίνακα 3, όλα τα μοντέλα query expansion ξεπερνούν τη βασική γραμμή MLE σε όλες σχεδόν τις περιπτώσεις, γεγονός που δείχνει την αποτελεσματικότητα της χρήσης παραστάσεων λέξεων υψηλής διάστασης για επέκταση ερωτήματος. Σύμφωνα με τα αποτελέσματα, αν και το word2vec έχει ελαφρώς καλύτερη απόδοση από το GloVe, δεν παρατηρούνται σημαντικές διαφορές μεταξύ των επιδόσεών τους. Και και τα δύο μοντέλα ενσωμάτωσης βάσει συνάφειας ξεπερνούν όλες τις βασικές γραμμές σε όλες τις περιπτώσεις, γεγονός που δείχνει τη σημασία της συνεκτίμησης της εκπαίδευσης διανυσμάτων ενσωμάτωσης. Οι βελτιώσεις αυτές είναι στατιστικά σημαντικές σε σύγκριση με όλες τις βασικές γραμμές. Το μοντέλο μεγιστοποίησης πιθανότητας συνάφειας (RLM) λειτουργεί καλύτερα από το μοντέλο μεταγενέστερης εκτίμησης συνάφειας (RPE) σε όλες τις περιπτώσεις και ο λόγος σχετίζεται με την αντικειμενική λειτουργία τους. Το RLM μαθαίνει την κατανομή συνάφειας για όλους τους όρους, ενώ το RPE μαθαίνει την πιθανότητα ταξινόμησης ως συνάφεια με όρους λεξιλογίου.

Στο επόμενο σύνολο πειραμάτων, εξετάζονται οι μέθοδοι που χρησιμοποιούν τα κορυφαία έγγραφα που ανακτήθηκαν για επέκταση ερωτήματος: το Relevance-based Models RM3 ως ένα προηγμένο μοντέλο ανατροφοδότησης ψευδο-σχετικότητας και το Local embedding έχοντας ως γενική ιδέα κατάρτιση μοντέλων ενσωμάτωσης λέξεων στα κορυφαία έγγραφα που ανακτήθηκαν ως απάντηση σε ένα δεδομένο ερώτημα. Παρόμοιος, χρησιμοποιείται το μοντέλο word2vec για να εκπαιδευθούν διανύσματα ενσωμάτωσης λέξεων σε 1000 κορυφαία έγγραφα.

Τα αποτελέσματα αναφέρονται στον Πίνακα 4

Collection	Metric	RM3	Local Emb.	ERM	
				Local	RLM
AP	MAP	0.2927	0.2412	0.3047	$\mathbf{0.3119}^{12}$
	P@20	0.4034	0.3742	0.4105	$\mathbf{0.4233}^{12}$
	NDCG@20	0.4368	0.4173	0.4411	$\mathbf{0.4495}^{123}$
Robust	MAP	0.2593	0.2235	0.2643	$\mathbf{0.2761}^{123}$
	P@20	0.3486	0.3366	0.3498	$\mathbf{0.3605}^{123}$
	NDCG@20	0.4011	0.3868	0.4080	$\mathbf{0.4173}^{123}$
GOV2	MAP	0.2863	0.2748	0.2924	$\mathbf{0.2986}^{123}$
	P@20	0.5318	0.5271	0.5379	$\mathbf{0.5417}^{12}$
	NDCG@20	0.4503	0.4576	0.4584	$\mathbf{0.4603}^{123}$
ClueWeb	MAP	0.1079	0.1041	0.1094	$\mathbf{0.1121}^{12}$
	P@20	0.3111	0.3062	0.3145	**0.3168**
	NDCG@20	0.2309	0.2261	0.2328	$\mathbf{0.2360}^{2}$

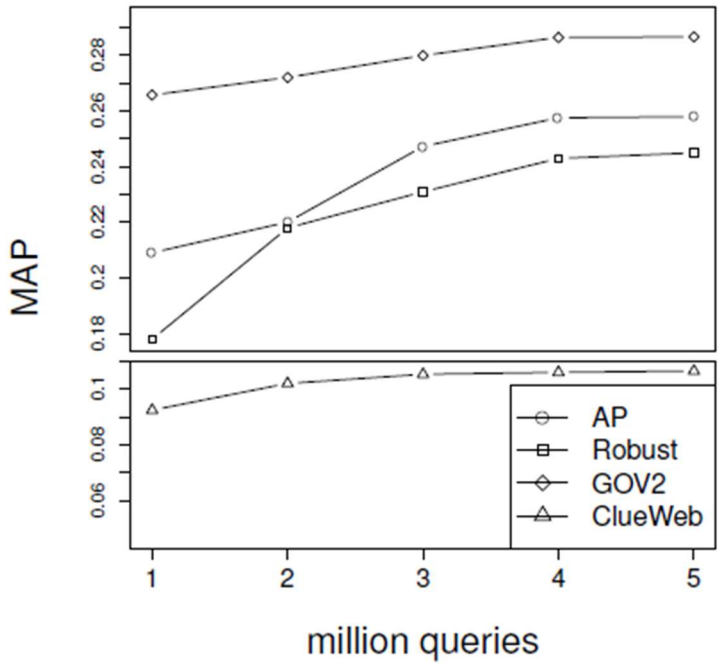
Πίνακας 4 Αξιολόγηση relevance-based word embedding spseudo – relevance feedback

Το ERM αναφέρεται στο Relevance-based Models που βασίζεται στην ενσωμάτωση προκειμένου να χρησιμοποιηθεί η σημασιολογική ομοιότητα που εκτιμάται με βάση τη λέξη στα διανύσματα ενσωμάτωσης σε ένα σενάριο ψευδο-σχετικότητας. Σύμφωνα με τον Πίνακα 4, το μοντέλο ERM που χρησιμοποιεί την ενσωμάτωση λέξεων βάσει συνάφειας (RLM10) ξεπερνά όλες τις άλλες μεθόδους. Αυτές οι βελτιώσεις είναι στατιστικά σημαντικές στις περισσότερες περιπτώσεις. Συγκρίνοντας τα αποτελέσματα που λαμβάνονται από την τοπική ενσωμάτωση και εκείνα που αναφέρονται στον Πίνακα 3, μπορεί να παρατηρηθεί ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των αποτελεσμάτων για την τοπική ενσωμάτωση και του word2vec, όταν τα διανύσματα ενσωμάτωσης εκπαιδεύονται στα κορυφαία έγγραφα της συλλογής στόχων, στην επιλογή μας.

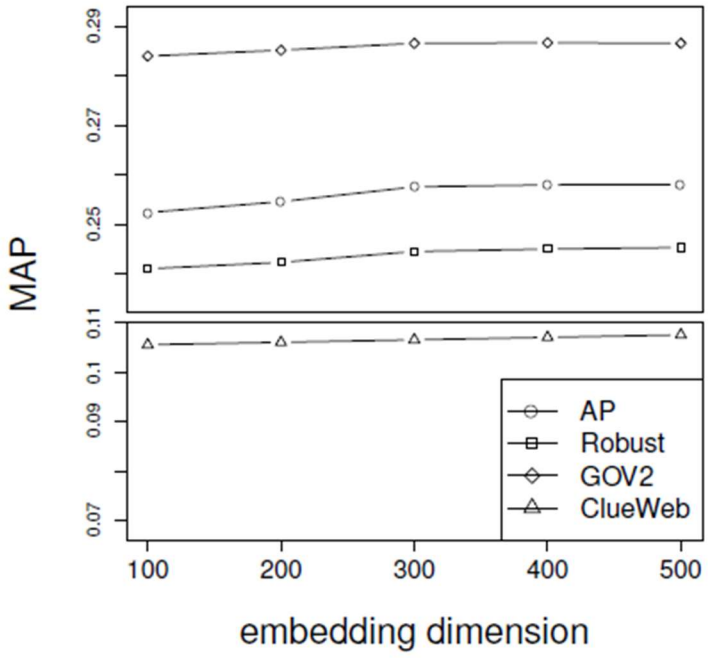
Μελετήθηκε επίσης η απόδοση του RLM ως μοντέλο ενσωμάτωσης λέξεων με την καλύτερη απόδοση για query expansion σε σχέση με τη διάσταση ενσωμάτωσης. Τα αποτελέσματα (εικ.7) δείχνουν την απόδοση επέκτασης ερωτήματος να βελτιώνεται γενικά όσο αυξάνεται η διάσταση ενσωμάτωσης. Οι επιδόσεις γίνονται σταθερές όταν η διάσταση είναι μεγαλύτερη από 300. Το πείραμά υποδηλώνει ότι 400 διαστάσεις θα ήταν αρκετές για το μοντέλο ενσωμάτωσης που βασίζεται στη συνάφεια.

Λόγω του μεγάλου αριθμού παραμέτρων στα νευρωνικά δίκτυα, μπορεί αν δοθεί πληθώρα εκπαιδευτικών δεδομένων να επιτύχουν καλή απόδοση. Στο επόμενο σύνολο πειραμάτων, μελετήθηκε πόσα δεδομένα εκπαίδευσης χρειάζονται για την εκπαίδευση

του καλύτερου μοντέλου. Τα αποτελέσματα παρουσιάζονται στην εικ.6 και δείχνουν πως αυξάνοντας τον αριθμό των ερωτημάτων εκπαίδευσης από ένα εκατομμύριο σε τέσσερα εκατομμύρια ερωτήματα, η απόδοση αυξάνεται σημαντικά και γίνεται πιο σταθερή μετά από τέσσερα εκατομμύρια ερωτήματα.



Εικ. 7 Βαθμός ευαισθησίας του RLM στο διαστατικό των embedding vectorsμε όρους μέτρησης MAP



Εικ. 6 Εκτέλεση του RLMστα διαφορετικά training queries με όρους μέτρησης MAP

Αξιολόγηση μέσω Classification

Η αξιολόγηση μέσω Classification, περιλαμβάνει αξιολόγηση των προτεινόμενων μοντέλων ενσωμάτωσης στο πλαίσιο της ταξινόμησης ερωτημάτων. Κάθε ερώτημα εκχωρείται σε έναν αριθμό ετικετών-labels (κατηγορίες) που έχουν οριστεί και λίγα ερωτήματα εκπαίδευσης είναι διαθέσιμα για κάθε ετικέτα. Αυτό είναι ένα εποπτευόμενο έργο ταξινόμησης πολλαπλών ετικετών με ελάχιστα δεδομένα εκπαίδευσης.

Το σύνολο των δεδομένων περιέχει 800 ερωτήματα ιστού που υποβάλλονται από πραγματικούς χρήστες που συλλέγονται τυχαία από τα αρχεία καταγραφής αναζήτησης. Τα ερωτήματα δεν περιέχουν "ανεπιθύμητο" κείμενο ή μη αγγλικούς όρους. Τα ερωτήματα επισημάνθηκαν από τρεις ανθρώπινους συντάκτες. Προετοιμάστηκαν 67 κατηγορίες και επιλέχθηκαν έως 5 ετικέτες για κάθε ερώτημα από κάθε συντάκτη.

Για τις μετρήσεις αξιολόγησης, χρησιμοποιήθηκαν η ακρίβεια και το F1- measure. Δεδομένου ότι οι ετικέτες που έχουν εκχωρηθεί από ανθρώπινους συντάκτες διαφέρουν σε ορισμένες περιπτώσεις, πρέπει να ληφθούν υπόψη όλα τα σύνολα ετικετών. Αυτές οι μετρήσεις υπολογίζονται με τον ίδιο τρόπο για την αξιολόγηση των υποβληθέντων διαδρομών του KDD Cup 2005. Οι στατιστικά σημαντικές διαφορές προσδιορίζονται με τη χρήση της δίπλευρης ζεύγους t-test που υπολογίζεται σε επίπεδο εμπιστοσύνης 95% (p – τιμή <0,05).

Πραγματοποιήθηκε 5 φορές πολλαπλή επικύρωση στα ερωτήματα και τα αποτελέσματα είναι ο μέσος όρος αυτών που αποκτήθηκαν στις δοκιμαστικές πτυχές. Για να ταξινομηθεί κάθε ερώτημα, χρησιμοποιήθηκε η απλή προσέγγιση που βασίζεται στο kNN. Υπολογίζεται η πιθανότητα κάθε κατηγορίας / ετικέτας σε κάθε ερώτημα q και, στη συνέχεια, επιλέγονται οι κορυφαίες κατηγορίες t με τις υψηλότερες πιθανότητες.
Στα πειράματα ταξινόμησης ερωτημάτων, εκπαιδεύτηκε η ενσωμάτωση λέξεων βάσει συνάφειας χρησιμοποιώντας το Robust ως συλλογή.

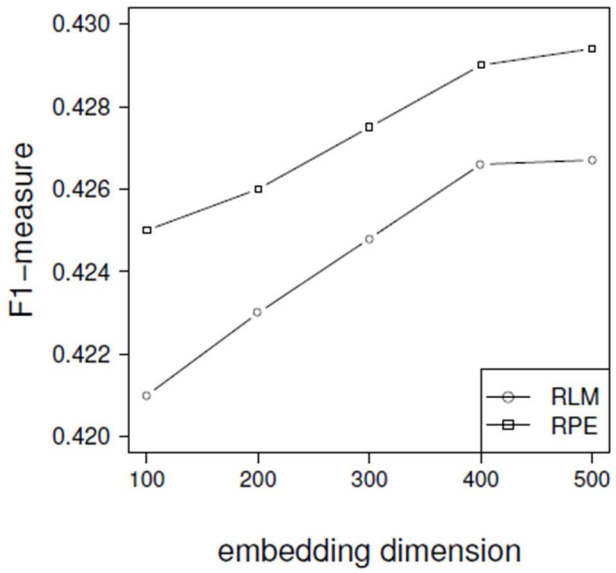
Τα αποτελέσματα φαίνονται παρακάτω:

Method	Precision	F1-measure
word2vec	0.3712	0.4008
GloVe	0.3643	0.3912
Rel.-based Embedding - RLM	0.3943^{12}	0.4267^{12}
Rel.-based Embedding - RPE	$\mathbf{0.3961}^{12}$	$\mathbf{0.4294}^{12}$

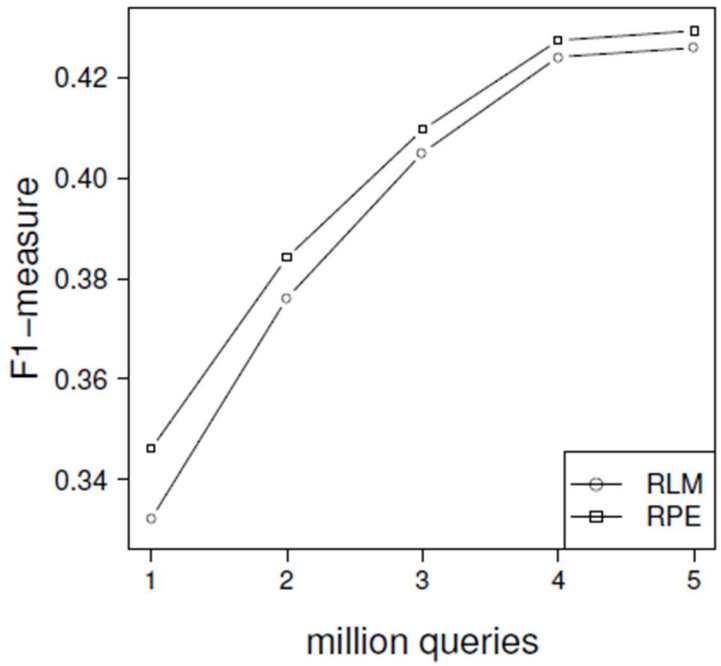
Πίνακας 5 Αξιολόγηση embedding algorithms μέσω classification

Τα μοντέλα συγκρίνονται με τις μεθόδους word2vec και GloVe που έχουν εκπαιδευτεί όπως περιγράφηκαν παραπάνω στα πειράματα query expansion. Τα αποτελέσματα αναφέρονται στον Πίνακα 5, όπου τα μοντέλα ενσωμάτωσης με βάση τη συνάφεια ξεπερνούν σημαντικά τις βασικές τιμές και από τις δύο μετρήσεις. Μια ενδιαφέρουσα παρατήρηση εδώ είναι ότι σε αντίθεση με τα πειράματα query expansion, το RPE εκτελεί καλύτερα από το RLM την ταξινόμηση ερωτημάτων. Ο λόγος είναι ότι κατά την επέκταση του ερωτήματος το βάρος κάθε όρου λαμβάνεται υπόψη προκειμένου να δημιουργηθεί το μοντέλο expand query language model. Επομένως, εκτός από τη σειρά των όρων, τα βάρη τους πρέπει επίσης να είναι αποτελεσματικά για τη βελτίωση της απόδοσης ανάκτησης με query expansion. Στην ταξινόμηση ερωτημάτων, εκχωρούνται μόνο λίγες κατηγορίες σε κάθε ερώτημα, και επομένως όσο η σειρά των κατηγοριών είναι σωστή, οι τιμές ομοιότητας μεταξύ των ερωτημάτων και των κατηγοριών δεν εμφανίζονται.

Στο επόμενο σύνολο πειραμάτων, μελετήθηκε η απόδοση των μοντέλων ενσωμάτωσης λέξεων που σχετίζονται με τη συνάφεια σε σχέση με τη διάσταση ενσωμάτωσης. Τα αποτελέσματα παρουσιάζονται στην εικόνα 8. Σύμφωνα με αυτό, η απόδοση γενικά βελτιώνεται αυξάνοντας τη διάσταση ενσωμάτωσης και γίνεται σταθερή όταν η διάσταση είναι μεγαλύτερη από 400. Αυτό είναι παρόμοιο με την παρατήρησή στα query expansion. Μελετήθηκε επίσης ο αριθμός των δεδομένων που απαιτούνται για την εκπαίδευση των μοντέλων, και χρειάζονται τουλάχιστον 4 εκατομμύρια ερωτήματα για να φανούν οι ακριβείς ενσωματώσεις λέξεων που βασίζονται στη συνάφεια. Στην εικόνα 9 φαίνεται ότι το RLM χρειάζεται περισσότερα δεδομένα εκπαίδευσης σε σύγκριση με το RPE για να έχει καλή απόδοση, διότι αυξάνοντας τον όγκο των δεδομένων εκπαίδευσης, οι καμπύλες μάθησης αυτών των δύο μοντέλων πλησιάζουν.



Εικ. 8 Βαθμός ευαισθησίας των relevance-based embedding μοντέλων σε embedding dimensions με όρους μέτρησης f1-measure



Εικ. 9 Εκτέλεση του Relevance-based embedding model στα διαφορετικά training queries με όρους μέτρησης F1-measure

Αναπτύχθηκαν δύο μοντέλα που βασίζονται σε νευρωνικά δίκτυα για εκμάθηση ενσωματώσεων λέξεων αλλά και στη συνάφεια. Το πρώτο μοντέλο, το μοντέλο μεγιστοποίησης πιθανότητας συνάφειας, στοχεύει στην εκτίμηση της πιθανότητας κάθε λέξης σε κατανομή συνάφειας για κάθε ερώτημα, ενώ το δεύτερο, το μοντέλο μεταγενέστερης εκτίμησης συνάφειας, ταξινομεί κάθε όρο και τους κατατάσσει σε σχετικούς ή μη για κάθε ερώτημα.

3. Semantics-based approaches

3.1 Τι είναι το Semantic Web

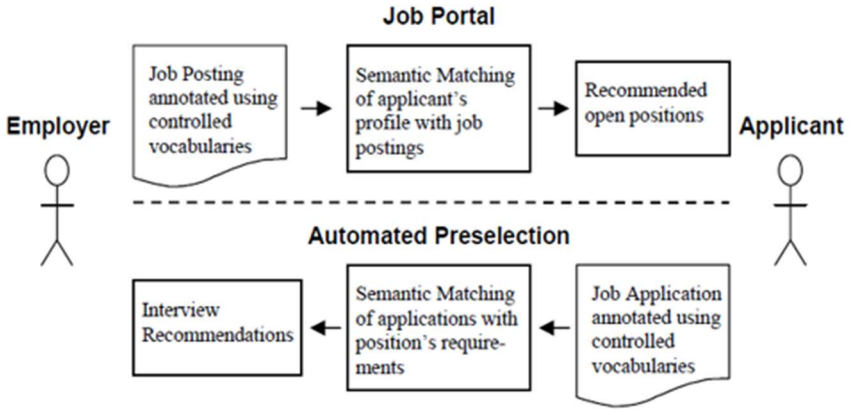
Η βασική ιδέα του Semantic Web θα μπορούσε να ορισθεί πως είναι η επέκταση του Ιστού, εκτός δηλαδή των κλασικών σελίδων HTML, με δεδομένα που είναι κατανοητά από μηχανή, οι πληροφορίες έχουν σαφώς καθορισμένη σημασία, χρησιμοποιείται λεξιλόγιο με συγκεκριμένο υπόβαθρο, και έχει ως στόχο να καταστήσει τη μηχανή δεδομένων του Διαδικτύου αναγνώσιμη, δηλαδή να επιτραπεί μια καλύτερη, ευκολότερη και πιο άμεση συνεργασία ανάμεσα στους υπολογιστές και τους ανθρώπους. Στόχος δηλαδή είναι η χρήση του Διαδικτύου σαν ένα παγκόσμιο κατανεμημένο σύστημα γνώσεων που να μπορεί να αξιοποιηθεί από εφαρμογές για την αυτόματη εκτέλεση εργασιών.
Οι βασικές τεχνολογίες για την υλοποίηση αυτού είναι:
• Ομοιόμορφα Αναγνωριστικά Πόρων – Uniform Resource Identifiers (URIs) ως παγκόσμιος μηχανισμός αναγνώρισης πόρων και όρων που χρησιμοποιούνται για την περιγραφή τους,
• Πλαίσιο Περιγραφής Πόρων – Resource Description Framework (RDF) ως βασικό μοντέλο δεδομένων μαζί με XML-based serialisation syntax για τη δημοσίευση δεδομένων στον Ιστό,
• το Web Ontology Language (OWL) που επεκτείνει το RDF με όρους και έννοιες για εκφραστική αναπαράσταση γνώσης.

Τα Semantic Web Technologies, και συγκεκριμένα οι οντολογίες (οντολογίες: δομημένα πλαίσια για την οργάνωση πληροφορίας), επιτρέπουν την μοντελοποίηση και τον συλλογισμό σχετικά με τις έννοιες των πληροφοριών που ανταλλάσσονται μεταξύ των συνιστωσών. Τα Semantic Web Services δείχνουν πώς οι οντολογίες μπορούν να βοηθήσουν στην αυτοματοποίηση της σύνθεσης των Υπηρεσιών Ιστού και να διευκολύνουν τη διαμεσολάβηση μεταξύ τους. Ωστόσο, η διαμεσολάβηση βασίζεται συχνά στον ορισμό νέων οντολογιών και στη χρήση τους για συμπεράσματα «μεταφράσεων» που είναι απαραίτητα για τη διασφάλιση της ουσιαστικής ανταλλαγής πληροφοριών μεταξύ των στοιχείων.

Η χρήση των Semantic Web Technologies στον τομέα των διαδικτυακών προσλήψεων θα μπορούσε ουσιαστικά να αυξήσει τη διαφάνεια της αγοράς, να μειώσει το κόστος συναλλαγής για τους εργοδότες και να αλλάξει τα επιχειρηματικά μοντέλα. Η εκμετάλλευση των οντολογιών στον τομέα των προσλήψεων βοηθά στη χρήση κοινών λεξιλογίων για την περιγραφή θέσεων εργασίας και βιογραφικών σημειωμάτων.

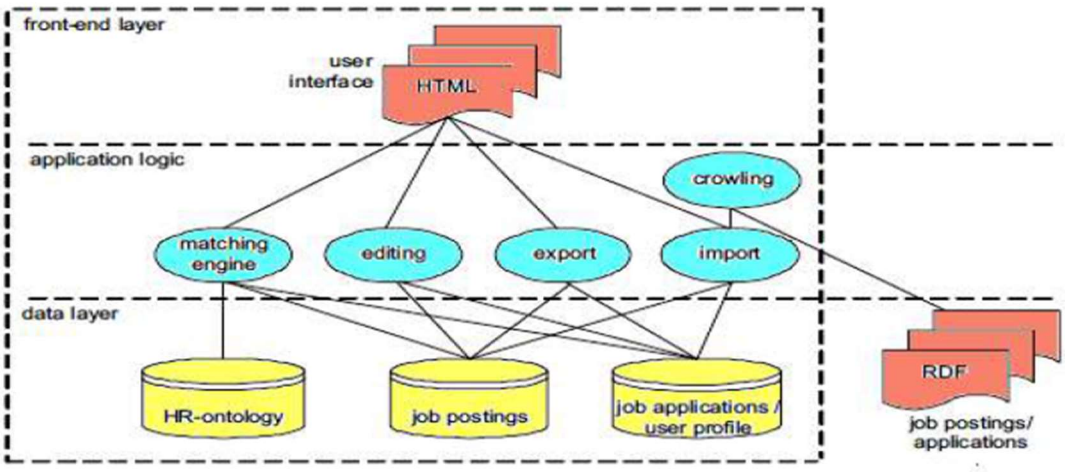
Υπάρχουν πολλοί τρόποι για να δημοσιευθεί μια αγγελία εργασίας στο διαδίκτυο όπως επίσης και τα μέρη όπου ένας υποψήφιος μπορεί να αναζητήσει τις αγγελίες αυτές. Με την χρήση των Semantic Web Technologies ένας εργοδότης δημοσιεύει μια θέση εργασίας μόνο στον ιστότοπο της επιχείρησής του και αυτή θα ανιχνευθεί από διάφορες

πύλες εργασίας. Έτσι, οι θέσεις εργασίας θα φτάσουν σε περισσότερους πιθανούς αιτούντες, γεγονός που θα έχει ως αποτέλεσμα μεγαλύτερη διαφάνεια στην αγορά. Επιπλέον, τόσο οι εργοδότες όσο και οι αιτούντες θα επωφεληθούν από αυτή την τεχνική. Οι αιτούντες μπορούν να επαναχρησιμοποιήσουν τα σημασιολογικά προφίλ τους, δηλαδή να τα στείλουν σε διαφορετικούς εργοδότες αντί να συμπληρώσουν διαφορετικές φόρμες ιστού. Οι εργοδότες θα επωφεληθούν επίσης από την αυτοματοποιημένη αντιστοίχιση μεταξύ των ικανοτήτων – δεξιοτήτων που απαιτεί η θέση εργασίας και των αιτήσεων, η οποία θα παρέχει μια κατάταξη λίστας με τους καλύτερους υποψηφίους που ταιριάζουν και έτσι θα μειώσει το διοικητικό κόστος[2].



Εικ. 10 Διαδικασία Αντιστοίχισης Αγγελιών Εργασίας με Βιογραφικά Σημειώματα, με την χρήση Semantic Web Technologies.

Μια τυπική αρχιτεκτονική μοντέλου αντιστοίχισης με εφαρμογή Semantic Web Technologies στην αγορά εργασίας που χρησιμοποιείται από διάφορους ιστοτόπους, έχει την παρακάτω μορφή.



Εικ. 11 Αρχιτεκτονική μοντέλου αντιστοίχισης

3.2 Χρήση Οντολογιών

Η χρήση οντολογιών για την ηλεκτρονική πρόσληψη καθίσταται σημαντικό έργο για την αντιστοίχιση των θέσεων εργασίας και των αιτούντων σε Semantic Web Technologies χρησιμοποιώντας τεχνικές αντιστοίχισης οντολογιών. Όπως αναφέρεται σε άρθρο του 2012 των Fuad Mire Hassan, Imran Ghani και Muhammad Faheem Abdirahman Ali Hajji [23], ο Παγκόσμιος Ιστός είναι ένα ταχέως αναπτυσσόμενο μέσο πληροφόρησης και εξυπηρέτησης, όπου τα άτομα που το χρησιμοποιούν μπορούν να μοιραστούν εφαρμογές και κατ' επέκταση πληροφορίες. Οι οντολογίες έπαιξαν σημαντικό ρόλο στην υποστήριξη του μηχανισμού ανταλλαγής πληροφοριών χρησιμοποιώντας σημασιολογική και εκτεταμένη συντακτική διαλειτουργικότητα του διαδικτύου. Η αντιστοίχιση οντολογίας είναι απαραίτητη στην ανταλλαγή πληροφοριών μεταξύ εφαρμογών του Semantic Web, όπως δημοσιεύσεις θέσεων εργασίας και αίτηση σε τομείς eRecruitment γενικότερα.

Έρευνες οντολογίας έχουν διεξαχθεί σε πολλούς διαφορετικούς ερευνητικούς τομείς την τελευταία δεκαετία για πολλούς σκοπούς. Στα περισσότερα πεδία, ωστόσο, έχουν προταθεί και εφαρμοστεί χρήσιμες οντολογικές λύσεις, ενώ παράλληλα έχει κερδίσει πολλή προσοχή των ερευνητών. Ωστόσο, υπάρχουν δύο διαφορετικοί τύποι οντολογίας:

1. Το Domain Ontologies παρέχει ένα σύνολο δομημένων εννοιών για τον προσδιορισμό συγκεκριμένου τομέα. Εφαρμόζεται σε τομείς όπως θέσεις εργασίας, ιατρική, γεωργία, αυτοκίνητα και σε πολλούς άλλους σχετικούς τομείς.
2. Οι Theory Ontologies παρέχουν ένα σύνολο όρων για την περιγραφή ορισμένων πτυχών του κόσμου, που μπορεί να είναι χρόνος, χώρος ή σχέδια. Η theory ontology είναι πιθανότατα μικρότερη και πιο αφηρημένη έννοια από την domain ontology.

Επομένως, μπορούμε να ορίσουμε την οντολογία ως ένα σύνολο δομημένων εννοιών ή όρων και σχέσεων μεταξύ τους σε έναν δεδομένο καθορισμένο τομέα. Η ανάπτυξη των τεχνολογιών επικοινωνίας και πληροφοριών έχει καταστήσει προσβάσιμη μια μεγάλη ποσότητα πληροφοριών. Ωστόσο, η πρόκληση της διαχείρισης της ετερογένειας μεταξύ διαφορετικών πηγών πληροφοριών αυξάνεται. Αυτό επιτυγχάνεται φυσικά σε δύο στάδια, όπως: πρώτον, αντιστοίχιση οντοτήτων για τον καθορισμό ευθυγράμμισης, δηλαδή, ένα σύνολο αντιστοιχιών και δεύτερον ερμηνεία ευθυγράμμισης σύμφωνα με τις ανάγκες της εφαρμογής, όπως απάντηση ερωτήματος ή μετάφραση δεδομένων. Το πρόβλημα της σημασιολογικής ετερογένειας μπορεί να επιλυθεί με τη χρήση οντολογικής αντιστοίχισης που βρίσκει αντιστοιχίες σημασιολογικά συναφών οντοτήτων μεταξύ διαφορετικών πηγών οντολογιών.

Από την άλλη πλευρά, υπάρχουν πολλές οντολογίες σε διάφορες εφαρμογές που δεν μπορούν να λειτουργήσουν. Αυτό είναι επειδή μπορεί να υπάρχουν οντότητες που έχουν διαφορετικά ονόματα σε πολλές οντολογίες, οι οποίες ενδέχεται να χρησιμοποιούν ανόμοιες γλώσσες για να αντιμετωπίσουν το ίδιο σύνολο όρων. Ωστόσο, η αντιστοίχιση οντολογίας είναι μια μέθοδος που χρησιμοποιείται για την χαρτογράφηση των

διαφορετικών όρων μεταξύ των οντολογιών. Η αντιστοίχιση οντολογίας είναι μια τεχνική που χρησιμοποιείται για τον εντοπισμό των σχέσεων (ισοδυναμίας) μεταξύ οντοτήτων δεδομένων οντολογιών. Κάθε οντολογία έχει έναν αριθμό οντοτήτων: τάξεις, ιδιότητες, κανόνες.
Το πρώτο βήμα για την ανάπτυξη ενός Semantic-based e-Recruitment [4] είναι η δημιουργία οντολογίας ανθρώπινου δυναμικού (HR-Ontology) στην οποία βασίζεται η βασική ιδέα:

- Αιτών: ο υποψήφιος για τη δουλειά,
- Εργοδότης: ο οργανισμός που προσέφερε τη δουλειά,
- Περιγραφή εργασίας: η εργασία που προσφέρει ο εργοδότης,
- Προφίλ: οι πληροφορίες προσόντων και εμπειρίας του αιτούντος.

Κατά τη δημιουργία του HR-Ontology, χρησιμοποιούνται για να ενσωματώσουν ορισμένα υπάρχοντα πρότυπα και ταξινομήσεις που περιέχουν σαφείς και καλά αναγνωρισμένες περιγραφές επαγγελματικών τίτλων που σχετίζονται. Ωστόσο, για να μειωθεί το κόστος δημοσίευσης θέσεων εργασίας και να μειωθεί ο αριθμός των αιτούντων, υπάρχουν οι τεχνικές αντιστοίχισης της σημασιολογικής οντολογίας στις διαδικασίες πρόσληψης. Η αντιστοίχιση οντολογίας είναι μια προσέγγιση για την εύρεση των σχέσεων μεταξύ των αντικειμένων δύο ή περισσότερων διαφορετικών οντολογιών χρησιμοποιώντας τις τεχνικές αντιστοίχισης οντολογίας.

Οι τεχνικές αντιστοίχισης μπορούν να χωριστούν στις παρακάτω κατηγορίες, όπως φαίνεται στον πίνακα:

Terminological	Υπολογίζει και ταιριάζει την ομοιότητα μεταξύ κειμένων χρησιμοποιώντας ονόματα, ετικέτες ή ορισμένες συμβολοσειρές.	Χρήσιμο όταν η αντίστοιχη-σύγκριση της οντολογίας βασίζεται σε συμβολοσειρές κειμένου ή στη γλώσσα.
Structural	Συγκρίνει τις περιγραφές οντοτήτων για κάθε οντολογία (εσωτερική δομή) ή τις αντιστοιχίες που μπορεί να έχει κάθε οντότητα με άλλες (εξωτερική δομή).	Πολύτιμη τεχνική για την αξιολόγηση και την ευθυγράμμιση της εσωτερικής δομής των οντοτήτων ή των αντιστοιχιών των οντοτήτων μπορεί να έχουν προς άλλες οντότητες εξωτερικής δομής.
Extensional (Instance)	Συγκρίνει την παρουσία / επέκταση ή το μήκος των κατηγοριών οντολογιών: με άλλους όρους, κατηγορίες κατηγοριών ή αντικείμενα	Χρήσιμο εάν οι πληροφορίες των οντοτήτων που συγκρίνονται είναι περιορισμένες και χρειάζονται για να συγκρίνουν πρόσθετα δεδομένα ή να υποστηρίξουν άλλες τεχνικές αντιστοίχισης για τον εντοπισμό όπου συχνά εμφανίζονται εσφαλμένες ή παραπλανητικές αλληλογραφίες

Πίνακας 6 Τεχνικές Αντιστοίχισης

3.3 Semantic-based e-Recruitment Systems

Οι πλατφόρμες ηλεκτρονικής πρόσληψης έχουν αποδειχθεί πιο αποτελεσματικές από τις παραδοσιακές μεθόδους πρόσληψης, καθώς παρέχουν στις επιχειρήσεις - οργανισμούς μεγάλη γεωγραφική προσέγγιση και εξοικονομούν χρόνο, κόστος και προσπάθεια που απαιτείται για την πρόσληψη του εκάστοτε υποψήφιου υπαλλήλου. Είναι η γενική πρακτική της χρήσης διαφορετικών πόρων που βασίζονται στο Web για αναζήτηση, αντιστοίχιση, έλεγχο, κριτική, συνέντευξη και πρόσληψη νέων αιτούντων (άτομα που αναζητούν εργασία) με αποτελεσματικό τρόπο. Καθώς τόσο οι εργοδότες όσο και οι αιτούντες έχουν στραφεί στη χρήση διαδικτυακών πυλών απασχόλησης, οι εργοδότες άρχισαν να λαμβάνουν μεγάλο αριθμό βιογραφικών που συνήθως μεταφορτώνονται ως μη δομημένα ηλεκτρονικά έγγραφα σε διαφορετικές μορφές όπως .pdf, .doc ή .rtf.

Για να βοηθήσουν τους εργοδότες να βρουν τον σωστό υποψήφιο από μια πληθώρα βιογραφικών, οι ερευνητές έχουν προτείνει αρκετές λύσεις που εκμεταλλεύονται τεχνικές επεξεργασίας κειμένου και σημασιολογίας. Παρόλο που αυτές οι προσεγγίσεις έχουν αποδειχθεί ότι βοηθούν τους εργοδότες να ελέγχουν τα βιογραφικά, εξακολουθούν να υστερούν από λόγους χαμηλής ακρίβειας όταν ταιριάζουν τα βιογραφικά με τις σχετικές θέσεις εργασίας τους. Για παράδειγμα, τα συστήματα που χρησιμοποιούν επεξεργασία κειμένου και επισήμανση δεξιοτήτων αλληλεπικαλύπτονται, αποτυγχάνουν να καλύψουν το κενό δεξιοτήτων - δηλαδή την ακριβή ανίχνευση και εξαγωγή δεξιοτήτων σε βιογραφικά υποψηφίων και θέσεις εργασίας - και να εντοπίσουν και να εξαγάγουν τις κρυφές σημασιολογικές διαστάσεις που κωδικοποιούνται στα βιογραφικά των υποψηφίων. Κατά συνέπεια, τα αποτελέσματα που προκύπτουν από αυτές τις τεχνικές δεν είναι ικανοποιητικά για τους εργοδότες, καθώς πολλά από τα βιογραφικά μπορούν να θεωρηθούν ψευδώς θετικά (όταν θεωρούν ότι τα μη σχετικά βιογραφικά είναι σχετικά με μια δεδομένη θέση εργασίας) ή ψευδώς αρνητικά (όταν τα βιογραφικά που σχετίζονται με μια δεδομένη θέση εργασίας δεν ανακτήθηκαν). Για να ξεπεράσουν τα μειονεκτήματα των παραδοσιακών τεχνικών αλληλοεπικάλυψης κειμένου και δεξιοτήτων, οι ερευνητές προτείνουν τη χρήση μηχανικής μάθησης και αλγορίθμων εξαγωγής χαρακτηριστικών, κατηγοριών και ταξινομήσεων, καθώς και λεξικών και γνώσεων.

Ταξινόμηση των συστημάτων ηλεκτρονικής πρόσληψης

Τα κριτήρια κατηγοριοποίησης των υπαρχόντων συστημάτων ηλεκτρονικής πρόσληψης [14] παρουσιάζονται αναλυτικά παρακάτω με μια συγκριτική ανάλυση μεταξύ τους:

- Στόχος του συστήματος: τα αναθεωρημένα συστήματα έχουν δύο βασικούς στόχους. Στοχεύουν είτε να βρουν μια αυστηρή αντιστοιχία μεταξύ των θέσεων εργασίας και των βιογραφικών (δηλ. Μοντέλο Boolean) ή εστιάζουν στην κατάταξη των βιογραφικών των αιτούντων ανάλογα με τη συνάφεια τους σε μια συγκεκριμένη θέση εργασίας. Στο πλαίσιο του δεύτερου τύπου συστήματος, οι εργοδότες μπορούν να εντοπίσουν εάν ένας υποψήφιος είναι αναλόγως των προσόντων για μια δεδομένη προσφορά εργασίας.
- Τεχνικές - προσεγγίσεις υλοποίησης: για να ταξινομήσουμε τα συστήματα ηλεκτρονικής πρόσληψης, εξετάζουμε επίσης στις τεχνικές - προσεγγίσεις που χρησιμοποιούνται από κάθε σύστημα. Αυτές οι τεχνικές περιλαμβάνουν διαλογή βάσει λέξεων-κλειδιών, σημασιολογία και μεθόδους βασισμένες στην επαγγελματική κατηγορία, αλγόριθμους μηχανικής μάθησης και συνδυασμό όλων αυτών των προσεγγίσεων.
- Τύπος εισόδου: τα συστήματα ηλεκτρονικής πρόσληψης δέχονται διαφορετικούς τύπους εισροών. Η είσοδος (βιογραφικά και θέσεις εργασίας) μπορεί να έχει τη μορφή δομημένων (χρησιμοποιώντας φόρμες), ημιδομημένων (χρησιμοποιώντας παραγόμενο έγγραφο xml), ή μη δομημένων (σε μορφή .pdf ή .doc) εγγράφων.
- Τύπος εξόδου: ένα άλλο σημαντικό κριτήριο που εξετάζεται για την κατηγοριοποίηση των συστημάτων είναι ο τύπος εξόδου που παράγει κάθε σύστημα κ μπορεί να ανήκει σε μία από τις δύο κατηγορίες. Στην πρώτη κατηγορία, τα παραγόμενα αποτελέσματα χαρακτηρίζονται από τη συνάφεια ή μη με μια δεδομένη θέση εργασίας. Τα συστήματα της δεύτερης κατηγορίας επεκτείνουν αυτήν την προσέγγιση παράγοντας ταξινομημένα αποτελέσματα. Σε αυτό το πλαίσιο, τέτοια συστήματα δεν φιλτράρουν μόνο ένα δεδομένο σύνολο βιογραφικών (δηλαδή ταιριάζουν ή δεν ταιριάζουν), αλλά συνιστούν επίσης βιογραφικά υψηλής βαθμολογίας στις σχετικές θέσεις εργασίας τους.
- Μέθοδος δοκιμών και αξιολόγησης: Έχουν πραγματοποιηθεί διαφορετικοί μηχανισμοί αξιολόγησης για τον έλεγχο και την αξιολόγηση της αποτελεσματικότητας των προτεινόμενων συστημάτων πρόσληψης και για να διαπιστωθεί εάν τα αποτελέσματα που επιστρέφονται (βιογραφικά) από κάθε σύστημα είναι αληθινά θετικά (δηλαδή σχετίζονται με μια δεδομένη θέση εργασίας και ανακτήθηκαν από το σύστημα). Για να γίνει αυτό, οι ερευνητές έχουν πραγματοποιήσει πειράματα χρησιμοποιώντας σενάρια προσλήψεων σε πραγματικό κόσμο και χειροποίητα συνθετικά σύνολα δεδομένων, ενώ άλλοι έχουν εφαρμόσει πρωτότυπα συστήματος όπου εξέτασαν τη συνολική αποτελεσματικότητα των εφαρμοζόμενων τεχνικών. Η αξιολόγηση των τεχνικών και των προσεγγίσεων που χρησιμοποιούνται στα συστήματα ηλεκτρονικής πρόσληψης έχει μεγάλο ενδιαφέρον, καθώς μπορούν να υιοθετηθούν με επιτυχία σε πρακτικά περιβάλλοντα και να έχουν τη θετική τους επίδραση στα μοντέλα των εταιρειών που τις υιοθετούν.

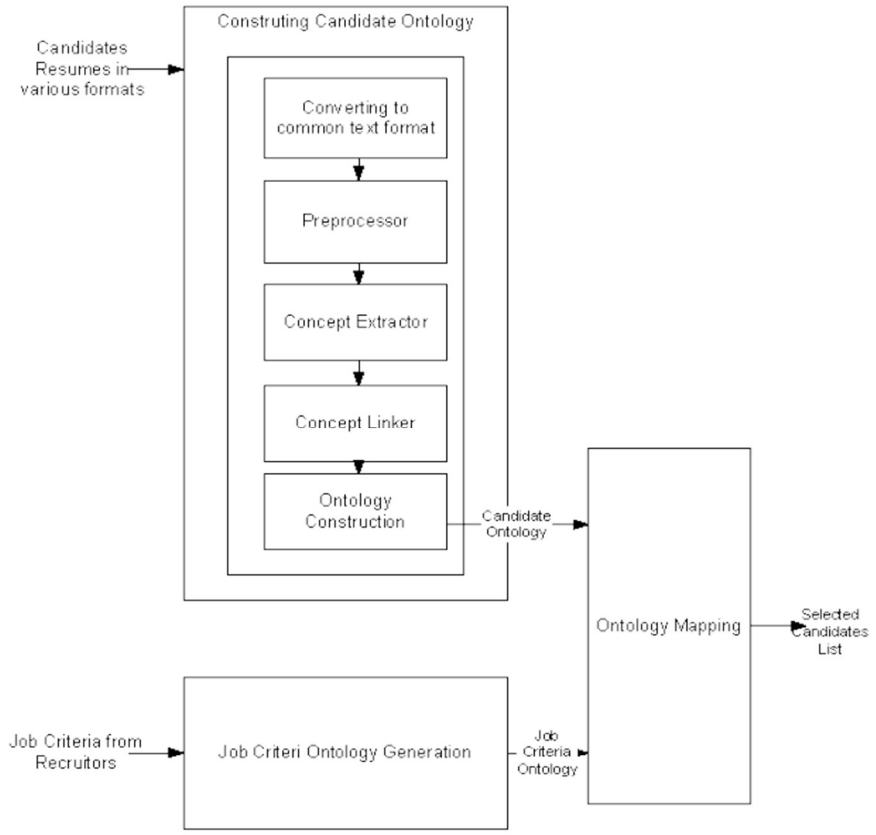
3.4 Αντιπροσωπευτικά Μοντέλα e-Recruitment on Semantic Web Technologies

Τα εργαλεία ηλεκτρονικής πρόσληψης έχουν εξαπλωθεί σημαντικά τα τελευταία χρόνια. Πολλές επιχειρήσεις στράφηκαν προς αυτή την κατεύθυνση καθώς οι εξελίξεις για αυξανόμενη ζήτηση στο θέμα της εργασίας οδήγησαν την ανάπτυξη τέτοιων πλατφορμών. Κατά συνέπεια, περισσότερες διαθέσιμες θέσεις εργασίας και προφίλ διαθέσιμων υποψηφίων γίνονται προσβάσιμες στο Διαδίκτυο. Παρόλο οι διαθέσιμες πληροφορίες ψηφιακής μορφής υποψηφίων θα έδιναν την δυνατότητα να βελτιωθεί η αντιστοίχιση εργαζομένου – διαθέσιμης θέσης εργασίας, αυτό το δυναμικό είναι σε μεγάλο βαθμό αχρησιμοποίητο, δεδομένου ότι η λειτουργικότητα αναζήτησης περιορίζεται κυρίως boolean αναζήτηση λέξεων-κλειδιών. Οι πρακτικές που ακολουθούνταν μέχρι πρότινος καθώς και οι θεωρητικές εκτιμήσεις δείχνουν ότι αυτός ο τύπος αναζήτησης είναι ακατάλληλος για την επίτευξη μιας καλής προσαρμογής μεταξύ των απαιτήσεων της εργασίας που πρέπει να καλυφθεί και των ικανοτήτων των υποψηφίων που βρέθηκαν.

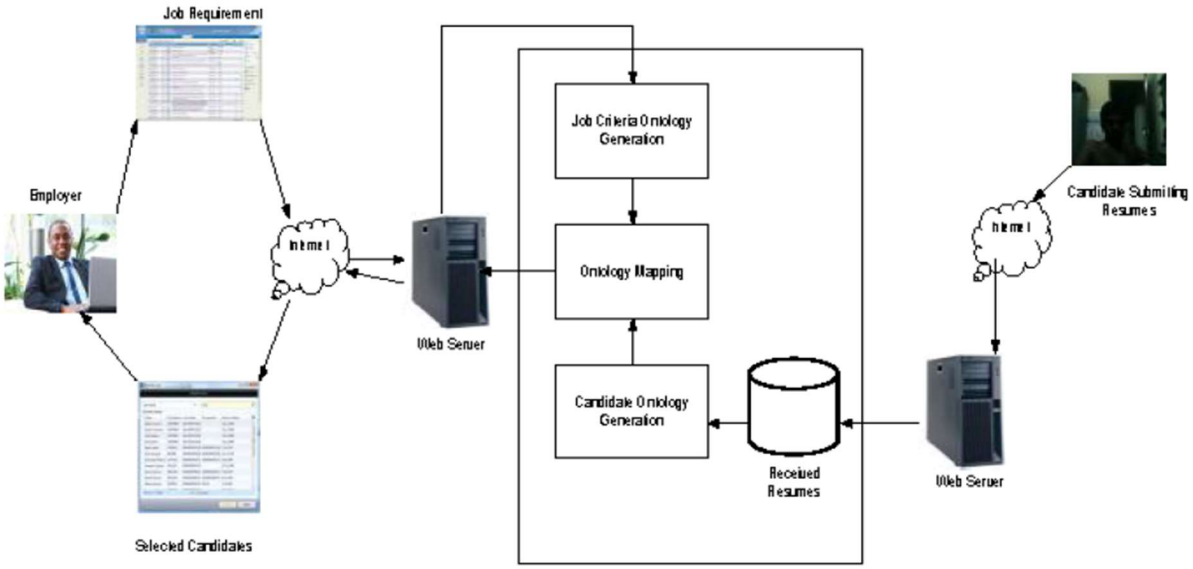
Ακολουθεί η ανάλυση αντιπροσωπευτικών μοντέλων όσον αφορά την προσέγγιση των semantic based, με πρώτο το –state of the art– μοντέλο EXPERT, σε μια προσπάθεια να δοθεί περαιτέρω επεξήγηση στο πως λειτουργούν τα συστήματα αυτά και τα ακριβή αποτελέσματα που μπορούν να αποφέρουν.

EXPERT

Το EXPERT [19], είναι ένα ευφυές εργαλείο για τον έλεγχο των υποψηφίων μιας θέσης εργασίας χρησιμοποιώντας οντολογία χαρτογράφησης (ontology mapping). Μπορεί με την μέθοδο της σημασιολογίας να εξετάσει τους υποψηφίους βάζοντας την ανάλογη βαθμολογία, κρίνει δηλαδή αν είναι κατάλληλοι των απαιτήσεων μιας θέσης, λύνοντας έτσι το πρόβλημα της μεγάλης εισροής βιογραφικών σημειωμάτων που αιτούνται για μια θέση εργασίας χωρίς να οδηγήσει σε λάθη όπως γινόταν με τις παλαιότερες μεθόδους όπου ο έλεγχος των βιογραφικών βασιζόταν σε υποκειμενικά κριτήρια από τους εκάστοτε ελεγκτές. Το EXPERT χωρίζεται σε τρεις φάσεις κατά την διαλογή υποψηφίων στις προσλήψεις. Στην πρώτη φάση, γίνεται η συλλογή των βιογραφικών των υποψηφίων και κατασκευάζει οντολογικό έγγραφο για τα χαρακτηριστικά τους, όπως προσωπικές πληροφορίες, τρέχουσα απασχόληση, προηγούμενες εργασίες, εκπαίδευση, δεξιότητες, ενδιαφέροντα και στόχοι. Οι νέες θέσεις εργασίας και οι απαιτήσεις αντιπροσωπεύονται ως οντολογία στη δεύτερη φάση, ενώ στην τρίτη φάση, το EXPERT χαρτογραφεί την οντολογία των απαιτήσεων εργασίας στο υποψήφιο οντολογικό έγγραφο και ανακτά τους επιλέξιμους υποψηφίους.



Εικ. 12 Στάδια μοντέλου Expert



Εικ. 13 Αρχιτεκτονική Expert

Τα στάδια του μοντέλου έχουν όπως παρακάτω:

Στάδιο 1: δημιουργία οντολογίας υποψηφίου

Τα βιογραφικά που υποβάλλονται από τους υποψηφίους, συλλέγονται ανάλογα με την μορφή που αποστέλλονται και αποθηκεύονται. Η προσέγγιση «έννοιας σύνδεσης» είναι αυτή που χρησιμοποιείται για να συλλεχθούν όλες οι σχετικές πληροφορίες από τα βιογραφικά. Το σχήμα της δομημένης οντολογίας δίνεται στην εικόνα 13.
Πρώτα, η είσοδος από την αποθήκευση δεδομένων επικυρώνεται και μετατρέπεται σε κοινή μηχανή που μπορεί να χρησιμοποιηθεί. Το concept extractor περιλαμβάνει tokenising, POST tagging του προεπεξεργασμένου κειμένου. Στη συνέχεια αναγνωρίζει τις ονομαστικές οντότητες και τις ομαλοποιεί. Έπειτα, το κανονικοποιημένο κείμενο αναλύεται για να εξαγάγει σημασιολογικά κατηγοριοποιημένες λεπτομέρειες. Το Concept linker ανακαλύπτει τη σχέση μεταξύ των εννοιών που εξήχθησαν από το concept extractor. Έτσι, όταν ολοκληρωθούν τα παραπάνω, κατασκευάζεται η οντολογία.

Στάδιο 2: κατασκευή οντολογικών κριτηρίων εργασίας

Σε αυτή τη φάση, κατασκευάζεται η οντολογία κριτηρίων εργασίας για το άνοιγμα - απαίτηση εργασίας. Κάποια απαίτηση μπορεί να έχει πολλαπλές τιμές, κάποια θα είναι υποχρεωτική και κάποια προαιρετική. Για παράδειγμα, η απαίτηση πιστοποίησης θα έχει την τιμή BE ή MCA ή MSc. Για κάποιες ορισμένες απαιτήσεις ως υποχρεωτικές, δίνεται το ανάλογο βάρος. Το βάρος κυμαίνεται από 0-1 με 1 την τιμή του υποχρεωτικού. Η άλλη υποδηλώνει το επίπεδο σπουδαιότητας αυτής της απαίτησης.

Στάδιο 3: χαρτογράφηση οντολογίας υποψηφίων & οντολογίας κριτηρίων εργασίας

Χρησιμοποιείται η προσέγγιση «άμεσης χαρτογράφησης οντολογίας» για να

$$M(i_1, i_2) = \frac{\sum_{k=1}^{n} Sim(p_k^{i1}, p_k^{i2}) * W_k^{i2}}{\sum_{k=1}^{n} W_k^{i2}}$$

χαρτογραφηθεί η οντολογία κριτηρίων εργασίας με όλες τις περιπτώσεις στην οντολογία των υποψηφίων. Η αντιστοίχιση μεταξύ της οντολογίας κριτηρίων εργασίας (i2) και ενός instance στην οντολογία των υποψηφίων (i1) υπολογίζεται ως εξής:

Όπου: p_k^{i1} είναι k-th ιδιότητα για την οντολογία i1 και p_k^{i2} αντίστοιχα για την οντολογία i2, ενώ W_k^{i2} είναι το βάρος που δίνεται από τον εργαζόμενο for the ιδιότητα p_k^{i2} .

Η συνάρτηση ομοιότητας Sim (p1, p2) ορίζεται ως ακολούθως:

$$Sim(p1, p2) = \begin{cases} 1, & if\ similarity\ of\ p1\ and\ p2 \geq t \\ 0, & otherwise \end{cases}$$

Κανόνας 1: εάν τα p1 και p2 είναι ιδιότητες με πολλαπλές τιμές, η ομοιότητα υπολογίζεται με αντιστοίχιση υποσυνόλου. Για παράδειγμα, η ιδιότητα πιστοποίησης στο i2 μπορεί να έχει πολλαπλές τιμές όπως BE ή MCA ή MSc οι οποίες θα αντιστοιχιστούν είτε σε BE είτε MCA ή MSc στην αντίστοιχη ιδιότητα στο i1.
Κανόνας 2: εάν τα p1 και p2 είναι ιδιότητες με ακέραιες τιμές, η ομοιότητα υπολογίζεται ως εξής: εάν η τιμή του p1 είναι μεγαλύτερη ή ίση με τις τιμές του p2, τότε η ομοιότητα είναι 1 διαφορετικά 0. Για παράδειγμα, εάν η ιδιότητα του έτους εμπειρίας στο i2 είναι 4, η συνάρτηση ομοιότητας θα επιστρέψει 1 για όλα τα i1, στο οποίο το έτος εμπειρίας η ιδιότηται2 έχει τιμή 4, διαφορετικά θα επιστρέψει το 0.
Το M (i1, i2) είναι μια τιμή στην περιοχή [0, 1] με 1 για ακριβή χαρτογράφηση και 0 για μη χαρτογράφηση. Η τιμή χαρτογράφησης υπολογίζεται για όλους τους υποψηφίους και επιστρέφει για όλους τους υποψηφίους με τιμή M ως τους επιλέξιμος υποψηφίους.
Για την λειτουργία του EXPERT χρησιμοποιήθηκαν περισσότερα από 800 βιογραφικά σημειώματα. Για να απεικονιστούν οι έννοιες, εξετάστηκε μέρος 10 τυχαίων βιογραφικών. Οι πιο σημαντικές ιδιότητες από αυτά τα δέκα βιογραφικά δίδονται στον παρακάτω πίνακα.

ID	*Qualification*	*CGPA*	*Total experience*	*Skills*
S1	MCA	6	3	Java, C++
S2	BE	7.5	4.5	Java, .Net, PHP
S3	MSc	6.5	7	.Net, C, C++
S4	MSc	8	7	.Net
S5	BE	9	4	Perl, Python, C++
S6	BE	8	5	Python, Java
S7	MCA	7	6	C++, Java
S8	BE	6.5	5	PHP, Perl
S9	MCA	8.5	5.5	Java, Perl, C++
S10	BE	9.25	2	Java, jScript, .Net

Πίνακας 7 Ιδιότητες 10 Βιογραφικών Σημειωμάτων

Attribute	*Expected value*	*Weight*
Qualification	BE/MCA	1
CGPA	7 and more	0.5
Experience	4 and more	0.5
Skill	Java	1

Πίνακας 9 Κριτήρια Θέσης Εργασίας

ID	*M (i1, i2)*
S1	0.67
S2	1.0
S3	0.17
S4	0.33
S5	0.67
S6	1.0
S7	1.0
S8	0.5
S9	1.0
S10	0.83

Πίνακας 8 Υπολογιζόμενα M-values – τιμή χαρτογράφησης

Ο Πίνακας 8 δείχνει την υπολογισμένη τιμή M για κάθε υποψήφιο. Με αυτόν τον τρόπο, η επιλογή των κατάλληλων υποψηφίων με ακριβές ταίριασμα θα ήταν αυτοί με το ID: S2, S6, S7 και S9 καθώς έχουν M-τιμή 1. Όταν η τιμή είναι μικρότερη 1, θα είναι επιλαχόντες.

Αναλυτικά η λίστα με τα βιογραφικά με τιμή 1 έχει ως εξής:

S.No.	Name of the Candidate	Gender	Age	Qualification	CGPA	Skills	Total Experience	M Value
1	Dhilip Kumaran S	M	30	MCA	6	Java, C++	3	1
2	Kamaraj K	M	30	BE	7.5	Java, .NET, PHP	4.5	1
3	Ramya E	F	28	M.Sc	6.5	.Net, C, C++	7	1
4	Kirthiga M	F	24	M.Sc	8	.NET	7	1
5	Rajam L	F	26	BE	9	Perl, Pythan, C++	4	1
6	Gokul Raj D	M	25	BE	8	Pythan, Java	5	1
7	Aperna K	F	26	MCA	7	C++, Java	6	1
8	Arun N	M	26	BE	6.5	PHP, Perl	5	1
9	Sujitha S	F	27	MCA	8.5	Java, Perl, C++	5.5	1
10	Rubella Mary T	F	27	BE	8.5	Java, Perl, C++	4	1
11	Sabari A	M	28	MCA	9	C++, Java	4	1

Εικ. 14 Λίστα των βιογραφικών με M-value1

Αναλύθηκε η δομή και η λειτουργία του EXPERT, τα αποτελέσματα και η ακρίβεια που προσφέρει το σύστημά αυτό. Η πιο σημαντική φάση του EXPERT είναι η λειτουργία της χαρτογράφησης, η οποία δίνει την τιμή χαρτογράφησης (M-value) για κάθε υποψήφιο. Παλαιότερες έρευνες που διενεργήθηκαν από διάφορους ερευνητές δίνουν τα εξής στοιχεία: από τους Ross και Young (2005) δείχνει ότι μια αντικειμενική δήλωση είναι πολύ σημαντική για το φιλτράρισμα των υποψηφίων. Οι Coleetal. (2007)

κατέληξαν στο συμπέρασμα ότι η εκπαίδευση θα αποφασίσει την καταλληλότητα του υποψηφίου για μια συγκεκριμένη εργασία. Ανέφεραν επίσης τη σημασία της φήμης των πανεπιστημίων που παρακολούθησαν καθώς και τη διάρκεια σπουδών του υποψηφίου. Οι Roth και Bobko (2000) τόνισαν ότι ο μέσος βαθμός αντικατοπτρίζει τη νοημοσύνη των υποψηφίων. Επεσήμαναν επίσης τη σημασία της χρήσης της εργασιακής εμπειρίας και των επιτευγμάτων εργασίας για το φιλτράρισμα των υποψηφίων. Οι Coleetal. (2007) διαπίστωσαν ότι οι εξωσχολικές δραστηριότητες είναι ένα άλλο σημαντικό βιογραφικό πεδίο για να φιλτράρεται ο υποψήφιος. Όμως, χρησιμοποιώντας ένα ή δύο πεδία βιογραφικού για το φιλτράρισμα του υποψηφίου θα καταλήγει με περισσότερους ανεπιθύμητους υποψηφίων ή επιλαχόντες. Λαμβάνοντας υπόψη όλα αυτά τα ευρήματα των ερευνών, οι ερευνητές του EXPERT, κατέληξαν στο συμπέρασμα πως θα υπολογίζονται όλα τα πιθανά πεδία του βιογραφικού με την αντίστοιχη βαρύτητα, όπως απαιτεί η εκάστοτε θέση εργασίας για να φιλτράρονται οι υποψήφιοι και να εμφανίσει μόνο τους επιθυμητούς υποψηφίους.

Lo-MATCH

Ένα διαδικτυακό εργαλείο που υλοποιείται στο πλαίσιο του matching[20], η πλατφόρμα αυτή βασίζεται σε semantic web technologies για την αντιμετώπιση ζητημάτων ετερογένειας στις περιγραφές των προσόντων των βιογραφικών των υποψηφίων και των αναγκών της αγοράς εργασίας λόγω της χρήσης μη κοινόχρηστων λεξιλογίων. Επιπλέον, κάνει χρήση μιας τεχνικής οπτικοποίησης που βασίζεται σε ετικέτες (cloud tags) για να απεικονίσει γρήγορα πτυχές που πρέπει να λαμβάνονται υπόψη στις φάσεις κινητικότητας και αναζήτησης εργασίας. Συγκεκριμένα, οι ιδιότητες cloud tag, όπως το μέγεθος της γραμματοσειράς και η απόσταση από το κέντρο του cloud, χρησιμοποιούνται για να παρέχουν μια άμεση επισκόπηση των κύριων χαρακτηριστικών ενός δεδομένου προσόντος σε σχέση με συγκεκριμένες ανάγκες του υποψηφίου, καθώς και για να επισημάνουν τις βασικές διαθέσεις του υποψηφίου σε σχέση με μια συγκεκριμένη προσφορά εργασίας.

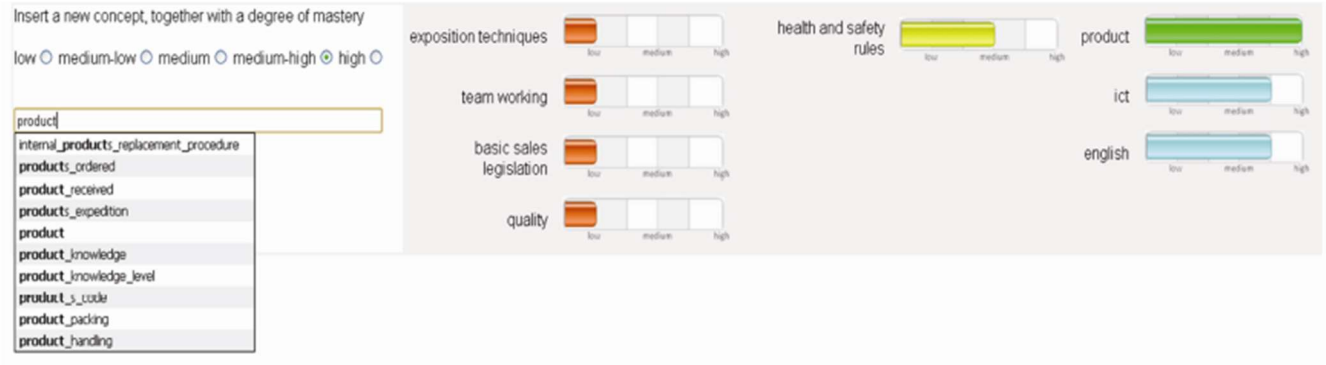
Αντιστοίχιση και tag cloud

Αυτή η πλατφόρμα σχεδιάστηκε προκειμένου να βοηθηθούν οι αιτούντες εργασία με τον εντοπισμό της εκπαίδευσης / κατάρτισης ή των ευκαιριών εργασίας να ικανοποιήσουν καλύτερα τις ανάγκες ή τις προσδοκίες τους (όσον αφορά τις γνώσεις, τις δεξιότητες και τις ικανότητες που λείπουν ή ταιριάζουν αντίστοιχα), καθώς και να υποστηριχθούν οι επιχειρήσεις στην επιλογή των σωστών υποψηφίων για μια δεδομένη θέση εργασίας. Το προτεινόμενο αυτό διαδικτυακό εργαλείο εκμεταλλεύεται μια οντολογία, μια ρητή δηλαδή προδιαγραφή, για να περιγράψει τα προσόντα του βιογραφικού σημειώματος και τα επαγγελματικά κομμάτια του προφίλ (που εκφράζουν τις ανάγκες της αγοράς εργασίας) που συλλέγονται στη «βάση γνώση» του Lo-MATCH. Για να καταστούν τα παραπάνω στοιχεία συγκρίσιμα, οι περιγραφές έχουν δομηθεί σε όρους μαθησιακών αποτελεσμάτων σύμφωνα με τις ενδείξεις του EQF και κάθε μαθησιακό αποτέλεσμα έχει σχολιαστεί -επισημανθεί με μία ή περισσότερες ετικέτες, με χειροκίνητο ή ημιαυτόματο τρόπο κάνοντας αναφορά σε έννοιες που ορίζονται στην οντολογία. Στην οντολογία, οι έννοιες συνδέονται με άλλες έννοιες μέσω σχέσεων, οι οποίες ανήκουν κυρίως στην κατηγορία «υπαναχώρησης». Οι σχέσεις υπαγωγής συμβάλλουν στη δημιουργία της συνολικής ιεραρχίας εννοιών - ετικετών (ταξινόμηση) που επιτρέπει στην πλατφόρμα να χειρίζεται τα μαθησιακά αποτελέσματα που εκφράζονται σε διαφορετικά επίπεδα λεπτομερειών, βελτιώνοντας έτσι τα αποτελέσματα σύγκρισης. Οι συλλεγόμενες πληροφορίες αξιοποιούνται για να αντληθεί μια ιδιότητα - τίτλου που βασίζεται στο cloud (όταν λαμβάνεται υπόψη η μαθησιακή διάσταση) ή των χαρακτηριστικών-απαιτήσεων της θέσης του αιτούντος εργασία (όταν λαμβάνονται υπόψη οι τομείς που αναζητούν εργασία / πρόσληψη).

Κατασκευή της «Βάσης Γνώσης»

Για να διευκολυνθεί η εισαγωγή σχετικών πληροφοριών στη βάση γνώσεων, δημιουργήθηκε μια ημιαυτόματη διαδικασία χαρτογράφησης. Για την λειτουργία ενός τέτοιου εργαλείου όταν ο χρήστης καθορίζει ένα νέο μαθησιακό αποτέλεσμα (π.χ., ως μέρος ενός τίτλου / βιογραφικού ή ενός επαγγελματικού προφίλ), το σύστημα εντοπίζει αυτόματα και τον προτείνει στις σχετικές έννοιες που θα μπορούσαν να συνδέονται με κάθε λέξη στο νέο στοιχείο. Για την εκτέλεση αυτής της εργασίας, το εργαλείο εκμεταλλεύεται το αποθετήριο Wordnet που συλλέγει λεξικές και σημασιολογικές σχέσεις μεταξύ όρων. Όταν ο χρήστης επιλέγει μια συγκεκριμένη ιδιότητα για σχολιασμό μιας δεδομένης λέξης του επιλεγμένου μαθησιακού αποτελέσματος, καταγράφονται επίσης οι λεξικές - σημασιολογικές σχέσεις στην οντολογία (μαζί με σχετικές έννοιες). Όταν μια συγκεκριμένη λέξη δεν βρίσκεται στο αποθετήριο, ο χρήστης μπορεί να καθορίσει έναν άλλο όρο που μπορεί να θεωρηθεί σχετικός με τον αρχικό (π.χ., θα μπορούσε να είναι πιο γενικός, πιο συγκεκριμένος, μπορεί να μοιράζεται τον ίδιο ορισμό κ.λπ.). Έννοιες που σχετίζονται με τον νέο όρο μπορούν στη συνέχεια να χρησιμοποιηθούν για να σχολιάσουν την επιλεγμένη λέξη μαθησιακού αποτελέσματος.

Με αυτόν τον τρόπο, το αρχικό πεδίο της οντολογίας επεκτείνεται και οι νέοι σχολιασμοί θα μπορούσαν ενδεχομένως να βασίζονται σε ένα πιο πλήρες σύνολο εννοιών και σχέσεων. Για κάθε έννοια που συνδέεται με ένα δεδομένο μαθησιακό αποτέλεσμα, πρέπει να παρέχεται μια τιμή σπουδαιότητας (όταν ο χρήστης θεωρεί ορισμένες έννοιες πιο σημαντικές από άλλες). Η διαφορά μεταξύ βαθμού γνώσης και σπουδαιότητας συνδέεται με το συγκεκριμένο είδος τελικού χρήστη στην πλατφόρμα: στην πραγματικότητα ένας υποψήφιος για εργασία εισάγει ένα προσόν (ή βιογραφικό σημείωμα) στη βάση γνώσεων, αντίστοιχα θα πρέπει να καθορίσει και ένα βαθμό γνώσης, ενώ όταν μια εταιρεία εισάγει τις απαιτήσεις της, πρέπει να καθορίσει έναν βαθμό σπουδαιότητας. Έτσι, ο βαθμός γνώσης αναφέρεται στην προοπτική προσφοράς εργασίας της διαδικασίας αντιστοίχισης, ενώ ο βαθμός σπουδαιότητας σχετίζεται με την πλευρά της ζήτησης. Η παρακάτω εικόνα δείχνει τη γραφική διεπαφή που επιτρέπει στις εταιρείες να προσδιορίσουν νέες απαιτήσεις.



Εικ. 15 Εισαγωγή νέας εγγραφής στο σύνολο των απαιτήσεων της εταιρείας

Δημιουργία Tag-Cloud

Οι έννοιες που είναι αποθηκευμένες στη βάση γνώσεων και ο βαθμός σπουδαιότητάς-δεξιότητάς τους χρησιμοποιούνται για να δημιουργήσουν την αναπαράσταση που βασίζεται στο tag-cloud:

- α) των χαρακτηριστικών του προσόντος που ικανοποιούν καλύτερα τις απαιτήσεις,
- β) τα χαρακτηριστικά του βιογραφικού ενός αιτούντος εργασίας που ταιριάζει καλύτερα στις απαιτήσεις της εταιρείας,
- γ) τις κύριες πτυχές του προφίλ εργασίας μιας εταιρείας που θα μπορούσαν να αξιοποιηθούν καλύτερα στις ικανότητες του υποψηφίου για την θέση εργασίας.

Στην παρούσα υλοποίηση η σημασία i μιας έννοιας αντιπροσωπεύεται μέσω του μεγέθους της γραμματοσειράς, με μεγαλύτερα γράμματα να επισημαίνονται οι πιο σχετικές έννοιες, ο βαθμός γνώσης m συνδέεται με την απόσταση από το κέντρο του

cloud, ενώ για αιτούντες με πλήρη γνώση των ζητηθέντων θεμάτων, δημιουργείται ένα συμπαγές tag-cloud. Αυτή η αναπαράσταση επιτρέπει την ταυτόχρονη εμφάνιση και των δύο διαστάσεων του προβλήματος αντιστοίχισης, δηλαδή των απαιτήσεων του αιτούντα που σχετίζονται με ένα συγκεκριμένο προσόν και των αναγκών της εταιρείας με τα χαρακτηριστικά του ατόμου που αναζητά εργασία. Όταν εστιάζουμε στην οπτική γωνία ενός υποψηφίου που αναζητά ένα προσόν ικανό να καλύψει μια θέση εργασίας ή με την προοπτική ενός εργοδότη που ψάχνει έναν εργαζόμενο να προσλάβει, το μέγεθος της γραμματοσειράς που χρησιμοποιείται για τη σχεδίαση των ετικετών καθορίζεται με ταξινόμηση. Οι ανάγκες ταξινομούνται σε φθίνουσα σειρά με βάση τη σημασία i και με βάση το σχετικό βάρος μιας δεδομένης έννοιας, υπολογίζοντας πάντα το πλήρες σύνολο απαιτήσεων. Τότε, οι συντεταγμένες υπολογίζονται ως x = r cos (θ) και y = r sin (θ), το r ορίζεται ως R (1 - m + D) / D, όπου: το R είναι η μέγιστη ακτίνα του σύννεφου, το m είναι το βαθμός γνώσης, D είναι ο αριθμός πιθανών τιμών της κλίμακας βαθμολόγησης των i και m, και θ είναι τυχαία γωνία.

Στο παράδειγμα του παρακάτω πίνακα, παρουσιάζονται οι απαιτήσεις μιας θέσης εργασίας και τα προσόντα δύο πιθανών αιτούντων. Συγκεκριμένα, εάν οι τιμές από 1 (χαμηλή) έως 5 (υψηλή) χρησιμοποιούνται για τη μέτρηση i και m (δηλαδή, D = 5), οι έννοιες των προϊόντων και των τεχνικών πώλησης θα αντιπροσωπεύουν το 20% των γνώσεων που ζητά η εταιρεία. Οι εσωτερικές διαδικασίες και πολιτικές και οι κανόνες υγείας και ασφάλειας θα αντιπροσωπεύουν το 12%, ενώ στις υπόλοιπες έννοιες θα δοθεί το υπόλοιπο 4%. Το μέγεθος της γραμματοσειράς καθορίζεται αποδίδοντας μια διαφορετική τιμή στα διάφορα ποσοστά εύρους, π.χ. μέγεθος γραμματοσειράς 10 για τιμές μεταξύ μηδέν και 5%. Στη συνέχεια, υποθέτοντας για παράδειγμα R = 500 και επιλέγοντας τυχαία γωνία θ = 335 °, η ετικέτα ICT που προσδιορίστηκε για τον πρώτο αιτούντα θα τοποθετηθεί στα x = 181 και y = −84 (υποθέτοντας το κέντρο του cloud σε x = 0 και y = 0).

Knowledge element (concept)	First applicant	Second applicant	Company
Product	high	high	high
Selling techniques	-	-	high
Negotiation techniques	-	high	-
Customer identification techniques	-	high	-
Internal procedures and policies	low	medium-high	medium
Health and safety rules	medium	low	medium
ICT	medium-high	low	low

English	medium-high	low	low
Exposition techniques	low	medium-high	low
Organization techniques	-	-	low
Team working	low	medium-high	low
Basic sales legislation	low	low	low
Inventory techniques	-	-	low
Quality	low	medium-low	low
Analysis techniques	-	-	low

Πίνακας 10 Βαθμός γνωστικών αντικειμένων δύο υποψηφίων

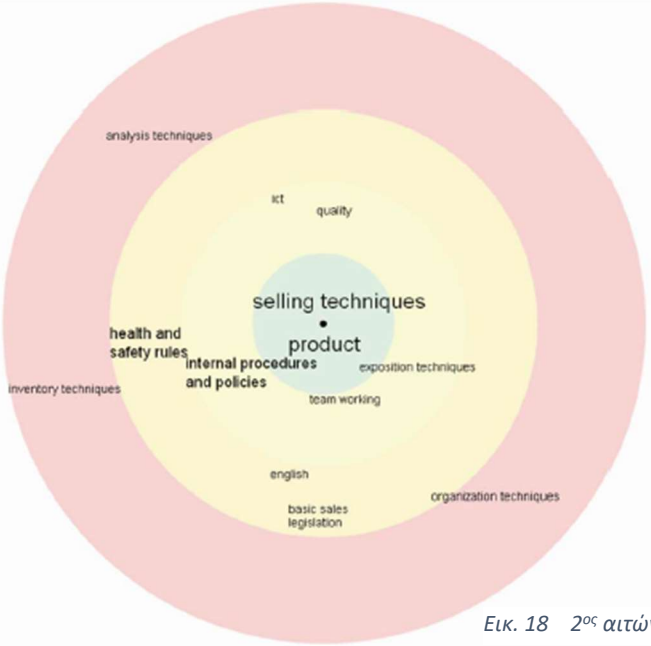
Οι εικόνες 17 & 18 δείχνουν τα tag-clouds για τα προγράμματα σπουδών των δύο αιτούντων, με βάση την ταξινόμηση που αναφέρεται στην εικόνα 16: δεδομένου ότι η εταιρεία αναγνώρισε ως κρίσιμη πτυχή τη γνώση των τεχνικών προϊόντων και πωλήσεων, σχετικές ετικέτες σχεδιάζονται με μια μεγάλη γραμματοσειρά, ακολουθούμενη από τη γνώση των εσωτερικών διαδικασιών και πολιτικών, καθώς και των κανόνων υγείας και ασφάλειας, και από τα υπόλοιπα στοιχεία γνώσης.

Ο πρώτος υποψήφιος έχει υψηλή γνώση του προϊόντος, μέτρια προς υψηλή γνώση Αγγλικών, μέτρια γνώση των κανόνων υγείας και ασφάλειας. Ωστόσο, έχει χαμηλή ή μηδενική γνώση για άλλες πτυχές της εργασίας. Έτσι, μόνο τέσσερα στοιχεία έλκονται κοντά στο κέντρο του νέφους, ενώ λείπουν στοιχεία γνώσης, όπως οι τεχνικές πώλησης, που τοποθετούνται στην εξωτερική περιοχή. Ο δεύτερος υποψήφιος είχε ήδη κάποια εμπειρία στον τομέα. Στην πραγματικότητα, αυτός είναι που δείχνει υψηλή γνώση του προϊόντος, τεχνικές διαπραγμάτευσης και τεχνικές αναγνώρισης πελατών, μέτρια προς υψηλή γνώση πολλών άλλων πτυχών και χαμηλή γνώση των υπόλοιπων στοιχείων. Δεδομένου ότι, σύμφωνα με την οντολογία, οι τεχνικές διαπραγμάτευσης και ταυτοποίησης πελάτη εντάσσονται στην έννοια των τεχνικών πωλήσεων, κατέχει επίσης σημαντική γνώση των τεχνικών πώλησης. Ως εκ τούτου, το προϊόν, οι τεχνικές πώλησης και οι εσωτερικές διαδικασίες και ετικέτες πολιτικών εμφανίζονται στην κεντρική περιοχή, καθιστώντας έτσι τον δεύτερο υποψήφιο τον καλύτερο (ή, τουλάχιστον, έναν καλό) υποψήφιο για τη συγκεκριμένη θέση εργασίας.

Τα παραπάνω παραδείγματα αναλύουν τα αποτελέσματα αντιστοίχισης από την άποψη της εταιρείας. Η διεπαφή που έχει σχεδιαστεί για αυτόν τον σκοπό απεικονίζεται στην εικόνα 19 και βασίζεται στο παραπάνω παράδειγμα. Στην αριστερή πλευρά, ένα tagcloud δείχνει κατά πόσο οι έννοιες που εκφράζονται στο βιογραφικό του δεύτερου υποψηφίου είναι σαφείς με την περιγραφή των απαιτήσεων των εργοδοτών. Σε αυτήν

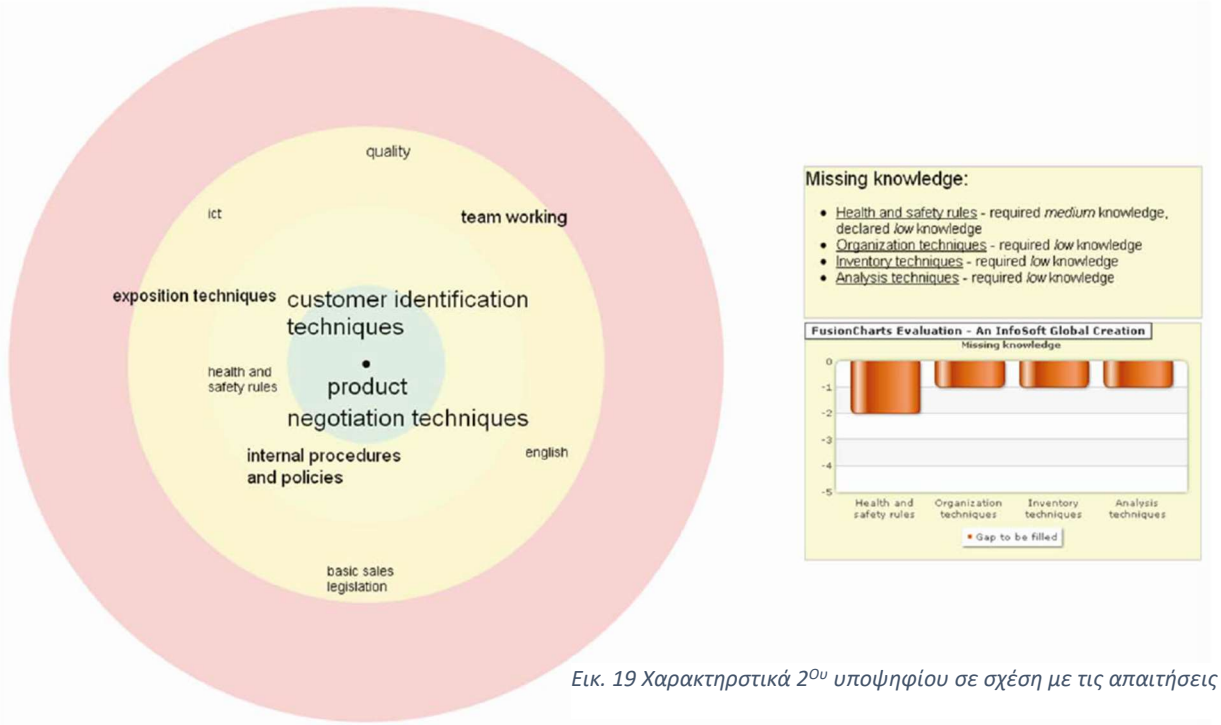
την περίπτωση, προκειμένου να εστιαστεί στον υποψήφιο, το tagcloud δημιουργείται αναστρέφοντας το i και το m (δηλαδή, συνδέοντας το μέγεθος της γραμματοσειράς και την απόσταση από το κέντρο του cloud με το βαθμό σπουδαιότητας και το βαθμό γνώσης, αντίστοιχα). Στη δεξιά πλευρά, εμφανίζονται υποδείξεις σχετικά με αυτές τις πτυχές που πρέπει να εξεταστούν από τον ίδιο τον υποψήφιο ώστε να αυξήσει τις ευκαιρίες πρόσληψης του από τη συγκεκριμένη εταιρεία. Ο υποψήφιος θα μπορούσε στη συνέχεια να εκμεταλλευτεί την επινοημένη πλατφόρμα για να βρει ένα προσόν (ή μέρος αυτής) παρέχοντας τις γνώσεις που λείπουν. Σε αυτήν την περίπτωση, το σύστημα θα καταγράψει αυτόματα τις απαιτήσεις του μαζί με το απαιτούμενο επίπεδο σπουδαιότητας, και θα κάνει την αντιστοίχιση με μια εισαγωγή ζήτησης και όχι με μια περιγραφή προσφοράς.

Εικ. 17 1ος αιτών



Εικ. 18 2ος αιτών

Εικ. 16 Τμήμα της οντολογίας των γνώσεων και ικανοτήτων των 2 υποψηφίων



Εικ. 19 Χαρακτηριστικά 2ου υποψηφίου σε σχέση με τις απαιτήσεις της εταιρίας.

Συμπερασματικά, αυτή η εφαρμογή βασίζεται στα tag-clouds και υποστηρίζει τη σύγκριση προσόντων με την αντιστοίχιση εργασίας, βασίζεται επίσης σε μια βάση γνώσεων που περιλαμβάνει προσόντα, βιογραφικό σημείωμα και προφίλ εργασίας που εκφράζονται σε γνώσεις, δεξιότητες και ικανότητες. Οι πληροφορίες που είναι αποθηκευμένες στη βάση γνώσεων έχουν σχολιασθεί με την εκμετάλλευση μιας οντολογίας που βασίζεται αρχικά στη βάση δεδομένων του Wordnet, η οποία επεκτάθηκε αργότερα από τους χρήστες, όπου απαιτείται. Αυτοί που θα μπορούσαν να εκμεταλλευτούν το εργαλείο αυτό είναι άτομα που αναζητούν εργασία ή οι ίδιες οι επιχειρήσεις. Οι υποψήφιοι θα μπορούσαν να χρησιμοποιήσουν το εργαλείο για να βρουν τις προσφορές εργασίας που θα ταιριάξουν καλύτερα στις ικανότητές τους, ενώ οι επιχειρήσεις θα μπορούσαν να πάρουν μια άμεση επισκόπηση της εμπειρίας των υποψηφίων που υποβάλλουν αίτηση για μια συγκεκριμένη θέση εργασίας. Αφενός, χάρη στη χρήση μιας ομοιόμορφης σημειογραφίας για την περιγραφή πτυχών που μπορεί να εκφράζονται με διαφορετικούς όρους και σε διαφορετικά επίπεδα λεπτομερειών από διάφορους παράγοντες, το προτεινόμενο εργαλείο στοχεύει στην υπέρβαση λεξικών και σημασιολογικών εμποδίων μεταξύ εκπαίδευσης, κατάρτισης, προσφορών εργασίας αλλά και απαιτήσεων.

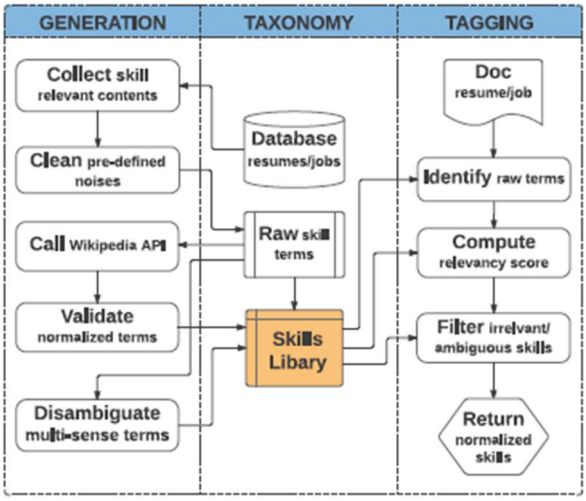
SKILL

Το SKILL [21] είναι ένα σύστημα Named Entity Normalization (NEN) για επαγγελματικές δεξιότητες. Αποτελείται από tagger skills, που χρησιμοποιούν τις ιδιότητες σημασιολογικών διανυσμάτων-λέξεων για την αναγνώριση και ομαλοποίηση σχετικών δεξιοτήτων, και από ένα skill entity που αποσαφηνίζει τη σωστή έννοια μιας αναγνωρισμένης ικανότητας, αξιοποιώντας τον MarkovChain Monte Carlo (MCMC) αλγόριθμο. Η αξιολόγηση Data driven χρησιμοποιώντας έρευνες τελικών χρηστών δείχνει ότι το SKILL επιτυγχάνει 90% ακρίβεια και 73% ανάκληση για επισημάνσεις δεξιοτήτων.

Ένα αποτελεσματικό σύστημα δεξιοτήτων θα πρέπει να μπορεί να κάνει:

- Να αναγνωρίζει την οντότητα δεξιοτήτων τόσο από τις θέσεις εργασίας όσο και από τα βιογραφικά. Αυτές οι πηγές είναι ημι-δομημένες και μπορεί να περιέχουν διαφορετικούς βαθμούς θορύβου.
- Να αντιμετωπίζει παραλλαγές ονόματος. Η οντότητα δεξιοτήτων «Τεχνητή Νοημοσύνη» μπορεί να έχει γραφεί στον πληθυντικό αριθμό, με πεζά ή κεφαλαία, μπορεί να είναι γραμμένο με το ακρωνύμιο του - μορφή AI ή να περιέχει τυπογραφικά λάθη.
- Να αξιοποιεί το σημασιολογικό πλαίσιο για να αναγνωρίζει μη καθορισμένες οντότητες δεξιοτήτων. Μια δημοσίευση στατιστικής εργασίας περιέχει ανάλυση συσχέτισης, οντότητες δεξιοτήτων, παλινδρόμησης πολλαπλών παραλλαγών, καθώς και ορισμένους άλλους όρους όπως (απαιτεί) ένας Διδακτορικός ή Μεταπτυχιακός τίτλος, αλλά δεν περιέχει δοκιμή λογιστικής παλινδρόμησης και υπόθεσης. Αυτές οι απροσδιόριστες δεξιότητες πρέπει να αναγνωρίζονται με λογικά επίπεδα εμπιστοσύνης.
- Να μειώνει τα ψεύτικα θετικά στις δεξιότητες προσθήκης ετικετών με πολλαπλές αισθήσεις (π.χ., ο διακομιστής διαθέτει γεύσεις σερβιρίσματος φαγητού και τεχνολογίας πληροφοριών).

Μερικές από αυτές τις προκλήσεις, η 1 και η 3, είναι μοναδικές στον τομέα των προσλήψεων, ενώ άλλες προκλήσεις, η 2 και η 4, υπάρχουν ήδη σε συστήματα NEN. Το σύστημα SKILL που θα περιγραφεί στοχεύει στην αντιμετώπιση όλων αυτών των προκλήσεων. Η αρχιτεκτονική του συστήματός φαίνεται στο παρακάτω σχήμα. Στα αριστερά εμφανίζεται η δημιουργία ταξινόμησης δεξιοτήτων. Ενώ στα δεξιά εμφανίζεται η βιβλιοθήκη δεξιοτήτων αφού ολοκληρωθεί η λειτουργία του συστήματος καθαρά για λειτουργίες αναγνώρισης.



Εικ. 20 Αρχιτεκτονική Συστήματος

Ταξινόμηση

Η ταξινόμηση στην πλατφόρμα αυτή ακολουθεί τα παρακάτω βήματα:

Συλλογή Για να δημιουργηθούν οι δεξιότητες των υποψηφίων, συλλέγεται το σχετικό περιεχόμενο με τις δεξιότητες από περισσότερα από 60 εκατομμύρια βιογραφικά υποψηφίων και 1,6 εκατομμύρια δημοσιεύσεις θέσεων εργασίας διαθέσιμες στο διαδικτυακό ιστότοπο CB. Η επιλεγμένη ενότητα μπορεί να είναι Δεξιότητες, Τεχνικές, Τεχνική Επάρκεια αν μιλάμε για βιογραφικό ή η ενότητα Απαιτήσεις στην αγγελία εργασίας.

Καθαρισμός Διαχωρίζεται το κείμενο από τα σημεία στίξης και, στη συνέχεια, αφαιρείται «ο θόρυβος», εάν υπάρχει. Το προκαθορισμένο λεξικό θορύβου περιέχει λέξεις-κλειδιά, ονόματα χωρών και πόλεων, επιρρήματα, επίθετα και άλλους προκαθορισμένους όρους. Στόχος είναι να απορριφθούν οι εξαιρετικά κοινές λέξεις που συμβάλλουν ελάχιστα και δεν έχουν καμία σημασιολογική αξία στην οικοδόμηση της ταξινόμησης δεξιοτήτων.

Χρήση Wiki Api Αφού συγκεντρωθούν οι ακατέργαστοι όροι, γίνεται χρήση του Wikipedia API7 για ομαλοποίηση των δεδομένων και να αποφευχθεί ο διπλασιασμός. Γίνεται πρώτα μια ανοιχτή ενέργεια αναζήτησης χρησιμοποιώντας συγκεκριμένες φράσεις ως ερώτημα εισαγωγής ακολουθούμενη από μια ενέργεια ερωτήματος για συσχετισμένα έγγραφα της Wikipedia, εάν υπάρχουν, και συλλέγονται ετικέτες κατηγορίας και ανακατευθύνσεις.

Επικύρωση Ο στόχος αυτού του βήματος είναι να διατηρήσει τις φράσεις που σχετίζονται άμεσα με τις επαγγελματικές δεξιότητες. Βασιζόνται σε λέξεις-κλειδιά από το σύστημα τυπικής επαγγελματικής ταξινόμησης (SOC) 8 για την επικύρωση των επιστρεφόμενων ετικετών κατηγορίας Wikipedia, επομένως λαμβάνεται μια απόφαση σχετικά με την πιστοποίηση των δεξιοτήτων. Συνολικά, ένα ερώτημα εισαγωγής θα

θεωρηθεί μορφή επιδεξιότητας εάν οι προκύπτουσες ετικέτες κατηγορίας τίτλου εγγράφου Wikipedia περάσουν τον έλεγχο λέξεων-κλειδιών SOC.
Αποσαφήνιση Ο στόχος εδώ είναι να αντιμετωπιστεί το πρόβλημα του Word Sense Disambiguation (WSD). Για παράδειγμα, μια φράση δεξιοτήτων συνδέεται με πολλαπλά πιστοποιημένα έγγραφα της Wikipedia, γι᾽αυτό και οι πολλαπλές κανονικοποιημένες δεξιότητες. Η αρχική προσέγγιση για το WSD χρησιμοποιούσε το Google Search API9. Για παράδειγμα, δεδομένης μιας φράσης δεξιοτήτων με πολλαπλή έννοια, επιλέγεται αυτή με την υψηλότερη κατάταξη στην Αναζήτηση Google (κατά συνάφεια). Αυτή η προσέγγιση, ωστόσο, δείχνει την προφανή αδυναμία του να μην λαμβάνεται υπόψη το σημασιολογικό πλαίσιο στο οποίο ανήκει η ικανότητα, και οδηγεί έτσι να αναπτυχθεί μια πιο ισχυρή προσέγγιση ως προς την εργασία WSD.
Βιβλιοθήκη δεξιοτήτων Στην τρέχουσα ταξινόμηση, υπάρχουν 39K επιφανειακές φόρμες που αντιστοιχίζονται σε 26K οντότητες κανονικοποιημένων δεξιοτήτων. Κάθε οντότητα δεξιοτήτων περιέχει έναν μοναδικό κωδικό ταυτοποίησης (αναγνωριστικό δεξιοτήτωνID), τον ακατέργαστο όρο του (αρχική μορφή), τον κανονικοποιημένο όρο του, το διάνυσμα ή την σχετική μορφή, το αντίστοιχο διάνυσμα συνημιτόνων και έναν τύπο δεξιοτήτων.

Η παρακάτω εικόνα απεικονίζει μια τυπική οντότητα δεξιοτήτων στη βιβλιοθήκη δεξιοτήτων του SKILL.

Skill ID KS4401T642KKKL4FQJMF

Raw Term restaurant

Normalized Term Restaurant Operation

RelatedSF Vector {hospitality, food services, customer service, clean room, retail stores, public utilities, vacuum, . . . }

Cosine Vector {0.774, 0.727, 0.719, 0.709, 0.707, 0.668, 0.661, . . . }

Skill Type Hard Skill

Εικ. 21 Μορφή εισηγμένης οντότητας

Ετικέτες δεξιοτήτων

Μετά το στάδιο της ταξινόμησης ακολουθεί το στάδιο των ετικετών και έχει ως εξής:

Προσδιορισμός Προσδιορίζονται οι αρχικές δεξιότητες από ένα δεδομένο έγγραφο εισαγωγής με άμεση αντιστοίχιση με την ταξινόμηση. Διαχωρίζεται το κείμενο εισαγωγής σε διακριτικά uni-gram, συναρμολογείται σεη-gram και στη συνέχεια αντιστοιχίζεται με την υπάρχουσα ταξινόμηση και αποθηκεύεται.
Υπολογισμός Υπολογίζεται η βαθμολογία συνάφειας για κάθε αντίστοιχη φόρμα. Για παράδειγμα, δεδομένης μιας μορφής στόχου, η βαθμολογία συνάφειας αντιστοιχεί στο ποσοστό των σχετικών μορφών από το word2vec όλων των μορφών του εγγράφου εισαγωγής. Είναι σημαντικό να σημειωθεί ότι η αρχική μέθοδος βαθμολόγησης είναι η ίδια, ανεξάρτητα από τη σχέση τους με τη μορφή-στόχο.
Φίλτρο Αναγνωρίζεται οποιαδήποτε μορφή με κανονικοποιημένη βαθμολογία σχετικότητας μεγαλύτερη ή ίση του 2% που ισοδυναμεί με 70% κανονικοποιημένη βαθμολογία. Αυτή η τιμή κατωφλίου επιλέγεται εμπειρικά. Κανονικά, δίνοντας μια μορφή που έχει αντιστοιχηθεί, αναμένουμε μόνο να δοθεί μια κανονικοποιημένη οντότητα δεξιοτήτων στη βιβλιοθήκη δεξιοτήτων. Σε πολλές περιπτώσεις, μια ασαφής μορφή προκαλεί την επισήμανση πολλών δεξιοτήτων. Προτείνεται για την μείωση τέτοιων σφαλμάτων μια προσέγγιση κατωφλίου με βάση τη σχετικότητα που θα τιμωρεί τη συνύπαρξη πολλών δεξιοτήτων, με αποτέλεσμα μια πιο ακριβή επισήμανση για διφορούμενες δεξιότητες.
Επιστροφή Αφού οι φόρμες εξάγονται τελικώς με επιτυχία από το έγγραφο εισαγωγής και επικυρώνονται έναντι της βιβλιοθήκης δεξιοτήτων, θα επιστραφεί μια σειρά από εξαγόμενες οντότητες δεξιοτήτων. Κάθε οντότητα δεξιοτήτων που επέστρεψε περιέχει έναν μοναδικό κωδικό αναγνώρισης, τον ακατέργαστο όρο του (ή την αρχική του μορφή), τον κανονικοποιημένο όρο του, τη βαθμολογία εμπιστοσύνης (ή τη βαθμολογία συνάφειας) και τον τύπο του.

Raw Term	Normalized Term	Relevancy Score	Type
accounting	Accounting	.95	Hard Skill
financial	Finance	.92	Hard Skill
forensic accounting	Forensic Accounting	.90	Hard Skill
budgets	Budgeting	.87	Hard Skill
CPA	Certifiedt Public Accountant	.87	Certification
forecasts	Forecasting	.85	Hard Skill
Excel	Microsoft Excel	.85	Hard Skill

Εικ. 22 Λίστα ταξινομημένων δεξιοτήτων

Job Description We are currently seeking a financial accountant for the Honolulu area. The ideal candidate would collect and analyze financial data to assist in making financial decisions. **Financial Accountant Job Duties:**

- Collect, analyze, and summarize data and trends.
- Prepare monthly, quarterly, and annual statements based of reporting.
- Provide financial advice while complying with federal ANS state laws.
- Must have current knowledge of financial regulations.
- Monitoring budgets, developing forecasts, and investigating variances.

Job Requirements:

- Knowledge of SFAS rules.
- Proficient on Excel.
- CPA and/or forensic accounting experience preferred.
- Must be detail oriented and business knowledge

Pay: $65K-$85K, depending on experience

Εικ. 23 Δημοσίευση Αγγελίας Εργασίας

Ο αλγόριθμος της πλατφόρμας παρουσιάζεται παρακάτω – αλγόριθμος 1. Λαμβάνει ως εισαγωγή τη λίστα των φράσεων δεξιοτήτων της αρχικής εισαγωγής x, w, v, c όπου x είναι ένας ακατέργαστος όρος (ή επιφανειακή μορφή), w είναι ο κανονικοποιημένος όρος του (ή μια αίσθηση δεξιοτήτων που έχει) από την ταξινόμηση, v, c είναι τα διανύσματα των σχετικών αρχικών μορφών και τις αντίστοιχες αποστάσεις συνημιτόνου word2vec. Επιστρέφει μια λίστα με κανονικοποιημένα αντικείμενα δεξιοτήτων x, w, v, c, ξ όπου διατηρούνται μόνο οι πιο σχετικές δεξιότητες και ξ είναι η βαθμολογία συνάφειας σε σχέση με το περιεχόμενο εισαγωγής.
Στη γραμμή 2 του αλγόριθμου2, υπολογίζεται η βαθμολογία συνάφειας χρησιμοποιώντας τη συνάρτηση getRelevancyScore που περιγράφεται στον Αλγόριθμο 2. Οι πρώτες βαθμολογίες είναι εξαιρετικά χαμηλές με μέση διακύμανση 2% και 0,003%, καθιστώντας δύσκολη την κατάταξη της συνάφειας. Προς επίλυση αυτού, χρησιμοποιήθηκε η προσαρμογή Beta και εμφανίζεται στην γραμμή 3 του Αλγόριθμου 2. Οι παράμετροι για την κατανομή Beta (α = 0,1627, β = 6,2385) επιλέχθηκαν εμπειρικά έτσι ώστε οι τελικές βαθμολογίες συνάφειας να εκτείνονται πιο ομοιόμορφα στο επιθυμητό διάστημα [0,1]. Επιστρέφονται φόρμες με βαθμολογία 70% ή υψηλότερη (γραμμή 4 του Αλγόριθμου 2).

Algorithm 1 Skills Tagging

Input: SKILLS_{raws}: list of raw skill objects $\langle x, w, \mathrm{v}, \mathrm{c} \rangle$.
Output: SKILLS_{norms}: list of normalized skill objects $\langle x, w, \mathrm{v}, \mathrm{c}, \xi \rangle$

1: **for each** $\langle x, w, \mathrm{v}, \mathrm{c} \rangle \in \text{SKILLS}_{raws}$ **do**
2: $\quad \xi \Leftarrow \text{getRelevancyScore}(x, w, \mathrm{v}, \mathrm{c}, X)$ ▷ see Algorithm 2
3: $\quad$ **if** isMultiSensesSF(x) **then**
4: $\quad\quad w' \Leftarrow \text{filterWSD}(x, W, \Xi)$ ▷ see Algorithm 3
5: $\quad\quad \text{SKILLS}_{norms} \Leftarrow \langle x, w', \mathrm{v}, \mathrm{c}, \xi \rangle$
6: $\quad$ **else**
7: $\quad\quad \text{SKILLS}_{norms} \Leftarrow \langle x, w, \mathrm{v}, \mathrm{c}, \xi \rangle$
8: $\quad$ **end if**
9: **end for**
10: **return** SKILLS_{norms}

Εικ. 25 μορφή Αλγορίθμου 1

Algorithm 3 filterWSD: Skills entity disambiguation

Input: (x, W, Ξ) where x is an ambiguous surface form multiple senses, $W = \{w_i\}$ denotes the set of all possible senses with the set of relevancy scores $\Xi = \{\xi_i\}$.
Output: List of strongest senses $\{w_k\}$ that can be referred to by the given ambiguous surface form x based on relevancy scores.

1: **return** w_{max} s.t. $\xi_{max} = \max\{\xi_i\}_{i=1}$
2: **for each** sense $w_i \in W \setminus w_{max}$ **do**
3: $\quad$ **if** $\frac{\xi_i}{\xi_{max}} \geq 90\%$ **then**
4: $\quad\quad$ **return** w_i
5: $\quad$ **end if**
6: **end for**

Εικ. 26 μορφή Αλγορίθμου 3

Algorithm 2 getRelevancyScore: Relevancy score computation

Input: $(x, w, \mathrm{v}, \mathrm{c}, X)$ where X is the set of all candidate surface forms from the input text.
Output: Confidence score ξ of surface form x with respect to the input context.

1: (Legacy relevancy score:

$$\lambda(x) = \frac{\sum_{j; x_j \in X} I_{\mathbf{v}}(x_j)}{\sum_{j; x_j \in X} I_X(x_j)} \quad , \tag{1}$$

where $I_A(x)$ is the indicator function s.t. $I_A(x) = 1$ if $x \in A$, and 0 otherwise.)
2: New relevancy score:

$$\xi(x) = \frac{\sum_{j; v_j \in X} c_j}{\sum_{j; v_j \in \mathbf{v}} c_j} \quad , \tag{2}$$

where v_j and c_j denote the j component of v and c respectively.
3: $\xi \Leftarrow \text{fitBetaDist}(\xi)$
4: **if** $\xi \geq \alpha = 70\%$ **then**
5: $\quad$ **return** ξ
6: **end if**

Εικ. 24 μορφή Αλγορίθμου 2

Ο αλγόριθμος 1 συνεχίζεται στη γραμμή 3, όπου η συνάρτηση is Multi Senses SF εξετάζει την ασάφεια της φόρμας εισόδου, επικυρώνοντάς την στη λίστα WSD από τη βιβλιοθήκη δεξιοτήτων. Εάν αναγνωριστούν πολλαπλές αισθήσεις δεξιοτήτων, χρησιμοποιείται φίλτρο λειτουργίας WSD, στη γραμμή 4 του Αλγόριθμου 1, για να επιλεγούν οι κατάλληλες. Εδώ, η αίσθηση δεξιοτήτων με την υψηλότερη βαθμολογία συνάφειας επιστρέφεται μαζί με άλλες που είναι τουλάχιστον 90% καλές σχετικά (γραμμή 1-3 του Αλγόριθμου 3). Τέλος, στη γραμμή 10 του Αλγόριθμου 1 αποκτάται το σύνολο των κανονικοποιημένων δεξιοτήτων. Η βασική βελτίωση της βαθμολογίας συνάφειας περιγράφεται στον Αλγόριθμο 2. Δεδομένης μιας μορφής που ταιριάζει από ένα κείμενο εισαγωγής, η αρχική προσπάθεια, όπως ορίζεται στην εξίσωση (1), παίρνει απλώς την αναλογία του μεγέθους της διασταύρωσης των v και X στο μέγεθος του X.

Είναι σημαντικό να σημειωθεί ότι όλες οι μορφές στο word2vec των μορφών v αντιμετωπίζονται ως εξίσου σημαντικές. Αυτό οδηγεί σε ένα σημαντικό μειονέκτημα. Για παράδειγμα, εάν ένα βιογραφικό περιέχει τόσο C ++ όσο και Visual C ++, τότε η βαθμολογία συνάφειας της γλώσσας C θα πρέπει να είναι υψηλή, αντίθετα, αν η Hewlett- Packard Graphics Language (HPGL) και το Automata Theory εμφανίζονται αντ '

αυτού, τότε το σκορ θα πρέπει να είναι χαμηλότερο. Δυστυχώς, δεν υπάρχει διάκριση στη βαθμολογία συνάφειας μεταξύ αυτών των δύο περιπτώσεων βάσει της προσέγγισης συχνότητας. Για να αντιμετωπιστεί αυτό το μειονέκτημα, προτείνεται μια σταθμισμένη προσέγγιση σημασιολογικής συνάφειας. Όπως περιγράφεται στην εξίσωση (2), η νέα βαθμολογία συνάφειας λαμβάνει υπόψη το βάρος κάθε αντιστοιχισμένης μορφής. Αυτά τα βάρη είναι στην πραγματικότητα οι συγγενείς ομοιότητες του διανύσματος word2vec με τις σχετικές δεξιότητες. Ως εκ τούτου, για μια δεδομένη ταιριαστή μορφή, η εμφάνιση των στενά συνδεδεμένων μορφών της αυξάνει τη βαθμολογία συνάφειας ουσιαστικά, ενώ η εμφάνιση των χαλαρά σχετικών μορφών της δεν επηρεάζει πολύ τη βαθμολογία της σχετικότητας.

Για να μετρηθεί η ακρίβεια, ζητάμε από τους χρήστες να επικυρώσουν τις κορυφαίες 10 δεξιότητες ανά βιογραφικό που ταξινομούνται κατά βαθμολογία συνάφειας. Για να μετρηθεί το recall (ανάκληση), ζητείται από τον χρήστη να προσθέσει έως και 5 δεξιότητες που λείπουν από την λίστα. Τα αποτελέσματα δείχνουν ότι το τρέχον πλαίσιο προσθήκης δεξιοτήτων επιτυγχάνει ακρίβεια 90% και ανάκληση 73%, η οποία είναι καλύτερη από την ακρίβεια 82% και ανάκληση 70% σε σχέση με την παλιά έκδοση της πλατφόρμας. Επιπλέον, παρατηρείται μια ισχυρή συσχέτιση μεταξύ της συνάφειας - βαθμολογίας και ποσοστό έγκρισης χρήστη. Η εικόνα 28 δείχνει ότι όσο υψηλότερη είναι η βαθμολογία συνάφειας, τόσο υψηλότερη είναι η πιθανότητα έγκρισης από τους χρήστες.

Version	Precision	Recall
Old	82%	70%
Current	**90%**	**73%**

Εικ. 27

Relevancy Score	Approved Skills	Total Skills	Approval Rate
.95	130	149	.8725
.90	316	371	.8518
.85	455	546	.8333
.80	317	407	.7897
.75	109	146	.7466
.70	8	12	.6667

Εικ. 28

Συμπερασματικά, για το σύστημα SKILL, ο στόχος ήταν να δοθεί μια υπηρεσία προσθήκης ετικετών δεξιοτήτων ως μικροϋπηρεσία. Τα βασικά στοιχεία Έρευνας& Ανάπτυξης του συστήματος SKILL είναι παρόμοια στο σχεδιασμό με προηγούμενες προσπάθειες NEN, καθώς και συστήματα που δημιουργούν ταξινομήσεις από πηγές δεδομένων χρησιμοποιώντας βάσεις γνώσεων όπως η Wikipedia. Τόσο η φάση παραγωγής ταξινόμησης όσο και οι φάσεις WSD είναι διαδικασίες εκτός σύνδεσης που μπορούν να προγραμματιστούν να εκτελούνται κατ'απαίτηση. Η φάση δημιουργίας ταξινόμησης κάνει εκτεταμένη χρήση σεναρίων συλλογής, καθαρισμού και εξαγωγής δεδομένων που εκτελείται σε εκατομμύρια θέσεις εργασίας. Ο αλγόριθμος προσθήκης ετικετών δεξιοτήτων αναπτύχθηκε για να υποστηρίζει τις απαιτήσεις σχεδόν σε πραγματικό χρόνο μιας υπηρεσίας Ιστού.

4. Τεχνικές Machine learning

Από τον εικοστό αιώνα έχει ξεκινήσει μια παρατεταμένη περίοδος ταχείας κοινωνικοοικονομικής μεταμόρφωσης με την ανάπτυξη του αυτοματισμού, της τεχνολογίας και των ψηφιακών επικοινωνιών [26]. Η τεχνητή νοημοσύνη (AI) και η μηχανική μάθηση αποτελούν πλέον κομμάτι των διαδικτυακών αλληλεπιδράσεων, των επικοινωνιών και της εργασιακής ζωής. Η πρόοδος στην τεχνητή νοημοσύνη δεν ασχολείται αποκλειστικά με την ανάπτυξη «συνείδησης υπολογιστών». Αντίθετα, αναπτύσσεται πιο συχνά ως μια εξελιγμένη μορφή σχεδιασμού και μηχανικής με απτούς και λειτουργικά συγκεκριμένους στόχους. Η μηχανική μάθηση, που αποτελούσε κομμάτι του AI και τώρα είναι ένα αυτοτελές πεδίο, προτείνει ότι οι υπολογιστές μπορούν να αναπτυχθούν σε περιβάλλοντα επίλυσης προβλημάτων ενημερώνοντας τα γενικά συμπεράσματα από συγκεκριμένα σύνολα δεδομένων. Εάν αυτό ενσωματωθεί σε ένα σύστημα ανατροφοδότησης, τότε ισχυρά, εύλογα συμπεράσματα και συσχετισμοί μπορούν να μετρηθούν έναντι της απόδοσης.

Ο τομέας του ανθρώπινου δυναμικού (HR) αντιμετωπίζει παράλληλες μετατοπίσεις προς την πλήρη ψηφιοποίηση, με τις εξελίξεις στην τεχνητή νοημοσύνη να βελτιώνουν την ανάλυση ανθρώπων και να βελτιώνουν τις μεθόδους απόκρισης στις πολύπλοκες διαδικασίες πρόσληψης και διαχείρισης του σύγχρονου εργατικού δυναμικού.

Λόγω της καινοτομίας της μηχανικής μάθησης, η τεχνολογική αυτή αναδυόμενη τάση μπορεί να εφαρμοστεί και στην τεχνολογία Ανθρώπινου Δυναμικού, δηλαδή μπορεί να είναι πολύ αποτελεσματική και επωφελής στον τομέα των προσλήψεων. Διευκολύνει τη μείωση έως και εξάλειψη των χρονοβόρων δραστηριοτήτων, στον εξορθολογισμό και την αυτοματοποίηση του ελέγχου των βιογραφικών, την αντιστοίχιση των απαιτήσεων εργασίας και των διαθέσιμων δεξιοτήτων των υποψηφίων πιο αποτελεσματικά, γεγονός που επιτρέπει την έγκαιρη λήψη αποφάσεων από τους επαγγελματίες.

Η καινοτόμος τεχνολογία διευκολύνει τις μηχανές να μάθουν την ικανότητα της λήψης αποφάσεων, τη λογική σκέψη και να αντιδρούν συστηματικά για να αποκομίσουν τα οφέλη από την ανάπτυξη λογισμικού που βασίζεται σε AI στις προσλήψεις μετατρέποντας την πλειονότητα της διαδικασίας σε αυτοματοποιημένη διαδικασία που μπορεί να εκτελεστεί από μια μηχανή, ενδυναμώνεται με την ικανότητα να αναλύονται τα μεγάλα δεδομένα πιο γρήγορα και να προβλέπονται τα πιθανά αποτελέσματα [25]. Επί του παρόντος, οι λύσεις για τις χρονοβόρες προσλήψεις δίνονται μέσω των βάσεων δεδομένων, της βελτιωμένης εμπειρίας υποψηφίων, πρόσληψης μέσω προφίλ κοινωνικών μέσων δικτύωσης, συνέντευξης εργασίας μέσω βίντεο, εξέτασης έντυπων αιτήσεων βάσει AI. Οι εικονικοί βοηθοί είναι αρκετά ικανοί να συνδεθούν με υποψηφίους, να αποθηκεύσουν, να αξιολογήσουν τις αιτήσεις και να αλληλεπιδράσουν για να βοηθήσουν τους υπευθύνους των προσλήψεων να διαχειριστούν τη βάση δεδομένων των υποψηφίων για περαιτέρω αναφορά. Είναι σε θέση να ελέγξουν το ιστορικό περιήγησης των χρηστών από αναρτήσεις διαφημίσεων σχετικά με τις κενές θέσεις μεταξύ εταιρειών διαφόρων θέσεων εργασίας που δημοσιεύονται στους πιθανούς

σωστούς υποψηφίους στις σελίδες του ιστοτόπου. Τα Chatbots χρησιμοποιούνται ευρέως για την εξαγωγή βασικών πληροφοριών από τα βιογραφικά των υποψηφίων, τα οποία αντιστοιχούν αυτόματα στις απαιτήσεις με τις δεξιοτήτες των υποψηφίων. Αυτά είναι ένα μείγμα τεχνολογίας - επεξεργασίας τεχνητής νοημοσύνης και φυσικής γλώσσας για καλύτερη αλληλεπίδραση, όπως οι άνθρωποι που ανταποκρίνονται λογικά.

Σύμφωνα με το blog προσλήψεων Undercover Recruiter, το AI αναμένεται να αντικαταστήσει το 16% των θέσεων εργασίας HR μέσα στα επόμενα 10 χρόνια. (Forbes Coaches Council, 2018). Η πρόσληψη και η επιλογή αποτελούν μέρος της διαδικασίας διαχείρισης ανθρώπινου δυναμικού (HRM). Η διαδικασία επιλογής κατάλληλων υποψηφίων για συνέντευξη έχει γίνει ταχύτερη και πιο αποτελεσματική με την τεχνολογία AI φέρνοντας την επανάσταση στη σχέση μεταξύ αιτούντος και εργοδότη. Παρόλα αυτά, ανακύπτουν θετικά και αρνητικά θέματα από αυτή την εξέλιξη [28].

Πλεονεκτήματα και Μειονεκτήματα χρήσης AI στις προσλήψεις

Πλεονεκτήματα

Το AI μπορεί να διευκολύνει τον έλεγχο τεράστιων δεδομένων εφαρμογής με ταχύτερο και αποτελεσματικότερο τρόπο συγκριτικά με τους παραδοσιακούς τρόπους που εκτελούνται από ανθρώπους. Σημαντική συνεισφορά, καθώς η επεξεργασία και ο έλεγχος τεράστιου αριθμού εφαρμογών πρέπει να γίνει αποτελεσματικά, ώστε να μην χαθεί ο κατάλληλος υποψήφιος κατά τη διάρκεια της διαδικασίας. Ορίζοντας συγκεκριμένα χαρακτηριστικά ή φίλτρα, τα εργαλεία AI έχουν την ικανότητα να εκτελούν εργασίες γρηγορότερα και αυτό να εξοικονομεί χρόνο ειδικά όταν υπάρχει επείγουσα ανάγκη για πρόσληψη.

Αντίστοιχα πλεονέκτημα αποτελεί και για τον εντοπισμό υποψηφίων μέσω κοινωνικών μέσων, καθώς πολλοί είναι εκείνοι που αναζητούν εργασία τείνουν να εισάγουν το προφίλ τους στο διαδίκτυο (LinkedIn-Monster). Αυτό καθιστά ευκολότερη την όλη διαδικασία από τους διαχειριστές ανθρώπινου δυναμικού, καθώς όλα αυτά τα δεδομένα και το προφίλ είναι διαθέσιμα στο διαδίκτυο, και με την χρήση της τεχνολογίας AI να γίνεται σε μικρότερο χρόνο.

Επίσης, πολλές εταιρείες αναθέτουν τη διαδικασία πρόσληψης υπαλλήλων τους σε εταιρείες προσλήψεων. Αυτό μπορεί να εξοικονομήσει το λειτουργικό κόστος των μικρών εταιρειών καθώς δεν χρειάζεται να επενδύσουν στην συγκεκριμένη τεχνολογία αλλά να αναθέσουν σε εταιρείες προσλήψεων που χρησιμοποιούν αυτή την τεχνολογία να φέρουν εις πέρας την διαδικασία της αντιστοίχισης έτσι ώστε τα σωστά άτομα να ταιριάζουν με τις σωστές θέσεις εργασίας στις σωστές εταιρείες. Παράγοντες που χρησιμοποιούνται για τη δημιουργία αλγορίθμου αντιστοίχισης θέσεων εργασίας είναι το εκπαιδευτικό υπόβαθρο, το εύρος εργασίας και η εταιρική κουλτούρα. Σε αυτήν την εποχή του IOT (Internet of Things) οι άνθρωποι συνδέονται μέσω Διαδικτύου ανά πάσα στιγμή, οπουδήποτε χωρίς γεωγραφικούς περιορισμούς. Η διαδικασία πρόσληψης επίσης

δεν περιορίζεται σε τοπικούς αιτούντες αλλά και σε διεθνείς ή σε όλο τον κόσμο. Οι υποψήφιοι προς εργασία μπορεί να θέλουν να κάνουν περαιτέρω έρευνα σχετικά με την εταιρεία που ενδιαφέρονται να συνεργαστούν. Προηγουμένως, η επικοινωνία μέσω επιστολών και τηλεφωνικών κλήσεων ήταν αυτή που χρησιμοποιούνταν, και αργότερα μέσω emails κατά την πρώιμη εποχή του Διαδικτύου. Σήμερα, μια νέα τεχνολογία στο AI που ονομάζεται chatbots κάνει αυτήν την επικοινωνία πιο αποτελεσματική. Τα Chatbots επιτρέπουν την προσωπική αφοσίωση σε πραγματικό χρόνο με τους υποψηφίους. Παρόμοια με εικονικούς προσωπικούς βοηθούς όπως το Siri και το GoogleNow, το chatbot χρησιμοποιεί την επεξεργασία φυσικής γλώσσας για να κατανοήσει μηνύματα και απαντήσει σε αυτά. Υπάρχουν επίσης εταιρείες που χρησιμοποιούν αυτήν την τεχνολογία σε συνεντεύξεις για τη διαδικασία πρόσληψης. Οι εφαρμογές ηλεκτρονικού ταχυδρομείου, SMS, κοινωνικών μέσων και μηνυμάτων είναι κανάλια επικοινωνίας που μπορούν να χρησιμοποιήσουν την εφαρμογή chatbot για να διευκολύνουν την επικοινωνία μεταξύ αιτούντων και επιχειρήσεων χωρίς περιορισμό χρόνου και τοποθεσίας.

Οι βιντεοκλήσεις επίσης είναι ένα άλλο κομμάτι που συναντάται λόγω του IOT για συνεντεύξεις εργασίας, καθώς παλαιότερα οι υποψήφιοι ταξίδευαν υπομένοντας το κόστος ενός τέτοιου ταξιδιού αλλά και της αβεβαιότητας της κάλυψης της θέσης. Οι συνεντεύξεις μπορούν να γίνουν οπουδήποτε με καλή σύνδεση στο Διαδίκτυο και αυτό μπορεί να εξοικονομήσει χρόνο και χρήμα. Υπάρχει επίσης τεχνολογία AI που μπορεί να αναλύσει το μοτίβο γλώσσας του σώματος και την έκφραση του προσώπου κατά τη διάρκεια συνέντευξης μέσω βίντεο. Αυτό μπορεί να διευκολύνει τη διαδικασία πρόσληψης καθώς ο υποψήφιος θα γίνει πιο χαλαρός και ταυτόχρονα, ο ερευνητής μπορεί να αναλύσει τις εγγραφές της συνέντευξης αργότερα και αυτό μπορεί να οδηγήσει σε πιο ακριβή και αποτελεσματική διαδικασία λήψης αποφάσεων εξαλείφοντας έτσι τους ανθρώπινους προκατειλημμένους παράγοντες κατά τη διάρκεια συνεντεύξεων.

Τέλος, σημαντικό στοιχείο είναι ότι αυτή η διαδικασία πρόσληψης που διευκολύνεται από εργαλεία και λογισμικό AI θα εξοικονομήσει χρόνο και θα μειώσει το φόρτο εργασίας στα στελέχη και εργαζομένους του ανθρώπινου δυναμικού. Αυτό σημαίνει ότι μπορούν να επικεντρωθούν σε άλλα πράγματα και συμβάλλοντας έτσι στην αποτελεσματικότητα της εργασίας και στην απόδοση της οργάνωσης.

Μειονεκτήματα

Παρόλα τα πλεονεκτήματα της χρήσης AI στην πρόσληψη, ένα πράγμα που πρέπει να επισημανθεί είναι ότι το AI δεν κάνει θαύματα. Είναι ένα πρόγραμμα που βασίζεται σε δεδομένα και αλγόριθμους που έχουν γραφτεί από προγραμματιστές με σκοπό την αποτελεσματική εκτέλεση εργασιών.

Το κύριο θέμα για το AI είναι να κάνει αντιστοίχιση θέσεων εργασίας, έλεγχο και ανάλυση του μοτίβου των αιτούντων για εργασία μέσα από ένα τεράστιο αριθμό δεδομένων. Χωρίς αυτά τα δεδομένα, τα αποτελέσματα που δίνονται από τα εργαλεία υποβοηθούμενης από AI μπορεί να μην είναι όπως αναμένονται. Υπάρχει λοιπόν ένα ζήτημα αβεβαιότητας στη λήψη αποφάσεων, καθώς αυτό βασίζεται κυρίως σε δεδομένα και όχι σε ανθρώπινη προσέγγιση.

Υπάρχει επίσης ανησυχία ότι το AI μπορεί να «μάθει» να αντιγράφει την απόφαση της ανθρώπινης προκατάληψης στο μέλλον. Αυτό μπορεί να συμβεί λόγω της έννοιας της «μηχανικής μάθησης» στο AI όπου τα πρότυπα αποφάσεων που χρησιμοποιούνται από τον άνθρωπο θα αναλύονται και θα αναπαράγονται από καιρό σε καιρό με αποτέλεσμα τον τύπο της ανθρώπινης προκατάληψης.

Τελευταίο θέμα όσον αφορά τα μειονεκτήματα είναι ο σκεπτικισμός στην αποδοχή της τεχνολογίας AI. Αυτό έγκειται στην πτυχή της «ανθρώπινης προσέγγισης» κατά τη διάρκεια προσωπικών συνεντεύξεων που απουσιάζουν κατά τη διεξαγωγή τηλεοπτικών συνεντεύξεων με χρήση AI. Είναι δύσκολο να εξηγηθεί, όμως οι περισσότεροι ερευνητές μπορούν πάντα να βλέπουν «κάτι» ξεχωριστό από έναν υποψήφιο που δεν είναι ορατό σε βιογραφικά και φυσική εμφάνιση. Προς το παρόν, αυτό το είδος παρατήρησης δεν μπορεί να αναπαραχθεί από λογισμικό AI.

Συμπερασματικά, αναδύεται ο ρόλος της τεχνητής νοημοσύνης στις τρέχουσες τάσεις πρόσληψης. Αυτό συμβαίνει επειδή οι τρέχοντες υποψήφιοι που αναζητούν εργασία είναι ως επί το πλείστον άτομα μικρής ηλικίας όπου η τεχνολογία, και ειδικά το IoT και τα κοινωνικά μέσα είναι μέρος της καθημερινότητας των ατόμων αυτών. Υπάρχει επίσης ένας τεράστιος ανταγωνισμός όσον αφορά την κάλυψη θέσεων εργασίας, καθώς είναι αυξανόμενος ο αριθμός των άτομων που κατέχουν πτυχίο τριτοβάθμιας εκπαίδευσης τα τελευταία χρόνια. Βάσει αυτών των συνθηκών, κάνει το σύνολο των αιτούντων για εργασία να αυξάνεται με αποτέλεσμα τις ανάγκες τεχνολογικής βοήθειας για τον έλεγχο και την επιλογή των καλύτερων υποψηφίων που μπορούν να οδηγήσουν έναν οργανισμό να αδυνατεί να αξιοποιήσει πλήρως τις δυνατότητές του. Μπορεί να υπάρχει ανησυχία ότι το AI θα αναλάβει ορισμένες ανθρώπινες δουλειές στο μέλλον, αλλά με τον έναν ή με τον άλλο τρόπο, η τεχνολογία οδηγεί σε εφαρμογή AI και όποιος δεν είναι σε θέση να ακολουθήσει την εξέλιξη της τεχνολογίας θα υστερεί.

4.1 Αλγόριθμοι Ομαδοποίησης και Ταξινόμησης για Εξόρυξη Δεδομένων

Η επιλογή του σωστού ατόμου στην σωστή θέση εργασίας, όπως έχει αναφερθεί, είναι η πιο σημαντική πρόκληση στη διαχείριση του ανθρώπινου δυναμικού. Οι διάφορες παλαιότερες μέθοδοι επιλογής περιλαμβάνουν ανάλυση του εντύπου αίτησης, αυτοαξιολόγηση, τηλεφωνική εξέταση και δοκιμές ανάλογα με τις απαιτήσεις του κλάδου (όπως ικανότητα, τεχνικός, προγραμματισμός, προσωπικότητα, τεστ ενδιαφέροντος κ.λπ.). Σήμερα, με βάση τα όσα αναλύθηκαν παραπάνω όσον αφορά την τρέχουσα τεχνολογία του AI και της μηχανικής μάθησης, τα διάφορα βιογραφικά ομαδοποιούνται και ταξινομούνται ώστε να εξαχθούν τα δεδομένα και να προχωρήσει σε αντιστοίχιση. Οι συνηθέστεροι αλγόριθμοι αυτής της κατηγορίας είναι οι κάτωθι, σε άρθρο των N. Sivaram & K. Ramar (2010) [32].

Fuzzy C-means Clustering

Η ενσωμάτωση της ασαφούς λογικής με τις τεχνικές εξόρυξης δεδομένων έχει γίνει ένα από τα βασικά συστατικά του προγραμματισμού, στον χειρισμό και τις προκλήσεις που θέτουν οι μαζικές συλλογές φυσικών δεδομένων. Η κεντρική ιδέα στα fuzzy clustering είναι το μη μοναδικό κομμάτι των δεδομένων σε μια συλλογή συστάδων. Τα datapoints είναι εκχωρημένες τιμές των ιδιοτήτων των μελών για κάθε μία από τις ομάδες και οι αλγόριθμοι fuzzy clustering επιτρέπουν στα clusters να αναπτυχθούν μέσα στα φυσικά σχήματα. Σε ορισμένες περιπτώσεις η τιμή ιδιότητας μέλους μπορεί να είναι μηδέν που δείχνει ότι το datapoint δεν είναι μέλος του συμπλέγματος που βρίσκεται υπό εξέταση. Πολλές τεχνικές ομαδοποίησης δείχνουν δυσκολίες στο χειρισμό ακραίων ορίων αλλά οι αλγόριθμοι αυτοί τείνουν να τους δίνουν πολύ μικρό βαθμό συμμετοχής γύρω από τις συστάδες. Οι μη μηδενικές τιμές ιδιοτήτων, με max την τιμή 1, δείχνει τον βαθμό στον οποίο το datapoint αντιπροσωπεύει ένα cluster. Τα σημεία στο κέντρο του έχουν μέγιστες τιμές συμμετοχής και σταδιακά η ιδιότητα μέλους μειώνεται όταν κάποια απομακρύνεται από το κέντρο του cluster. Έτσι, το fuzzy clustering παρέχει μια ευέλικτη και στιβαρή μέθοδο για χειρισμό φυσικών δεδομένων με ασάφεια και αβεβαιότητα. Στα fuzzy clustering, κάθε datapoint έχει σχετικό βαθμό των μελών για κάθε cluster.

Ο αλγόριθμος ομαδοποίησης FuzzyC-means περιλαμβάνει δύο διαδικασίες, τον υπολογισμό του cluster center και την ανάθεση των points σε αυτά τα κέντρα χρησιμοποιώντας την μέθοδο της Ευκλείδειας απόστασης. Η διαδικασία συνεχίζεται μέχρι το cluster center να σταθεροποιηθεί. Ο αλγόριθμος ενσωματώνει το ασαφές σύνολο εννοιών της μερικής ιδιότητας μέλους και οι φόρμες υπερκαλύπτουν τα clusters για να το υποστηρίξουν.

Σε κάθε data item εκχωρείται τιμή ιδιότητας μέλους στο εύρος [0,1] για τα clusters. Ο βαθμός ασάφειας στα clusters υποδεικνύεται από την παράμετρο που ονομάζεται fuzzification(m). Όταν η τιμή του m ισούται με 1 ο αλγόριθμος λειτουργεί σαν ξεκάθαρος αλγόριθμος partisioning και για μεγαλύτερες τιμές αλληλεπικάλυπτει το

cluster. Η ιδιότητα μέλους κάθε στοιχείου δεδομένων υπολογίζεται χρησιμοποιώντας τον τύπο:

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}}$$

Όπου, $\mu_j \chi_i$ είναι χ_i τα μέλη στο j^{th} cluster:
dji – απόσταση xi στο κλάστερ cj
m – fuzzification parameter
p – αριθμός συγκεκριμένων clusters
dki – απόσταση των xi στα cluster ck

Το άθροισμα των μελών ενός datapoint σε όλες τις ομάδες πρέπει να ισούται με 1. Τα νέα cluster centers υπολογίζονται χρησιμοποιώντας την συνάρτηση:

$$c_j = \frac{\sum_i \mu_j(x_i)^{\bar{m}} x_i}{\sum_i \mu_j(x_i)^{\bar{m}}}$$

Ο αλγόριθμος ξεκινά επιλέγοντας τον αριθμό των συστάδων και την παράμετρο fuzzification. Επιλέγεται το κέντρο για όλες τις συστάδες τυχαία. Ο αλγόριθμος συνεχίζει να ενημερώνει το κέντρο των clusters μέχρι να σταθεροποιηθεί η τιμή.

K-means Clustering

Το K-means είναι ένας από τους πιο απλούς αλγορίθμους μη επιβλεπόμενης μάθησης για ομαδοποίηση προβλημάτων. Ο αλγόριθμος στοχεύει στη διαμόρφωση kclusters των n αντικειμένων, έτσι ώστε η προκύπτουσα ομοιότητα εντός cluster να είναι υψηλή, αλλά η ομοιότητα μεταξύ ομάδων χαμηλή. Ο αλγόριθμος επιλέγει τυχαία k από τα n αντικείμενα και ένα από αυτά εκχωρείται σε κάθε cluster για να αντιπροσωπεύει το μέσο όρο του clustermean ή το κέντρο. Για καθένα από τα υπόλοιπα αντικείμενα, ένα αντικείμενο αντιστοιχίζεται στο cluster με το οποίο είναι παρόμοιο, με βάση την απόσταση μεταξύ του αντικειμένου και του meancluster. Στη συνέχεια υπολογίζεται ο νέος μέσος όρος για

κάθε σύμπλεγμα και η διαδικασία επαναλαμβάνεται έως ότου συγκλίνει η συνάρτηση κριτηρίου. Χρησιμοποιείται ένα τετράγωνο σφάλμα και ορίζεται ως:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

1. Επιλέγει αυθαίρετα σημεία K στο διάστημα που αντιπροσωπεύει τα αντικείμενα που είναι συγκεντρωμένα. Αυτά τα σημεία αντιπροσωπεύουν την αρχικά κεντροειδή ομάδα.
2. Αντιστοιχεί κάθε εναπομείναν αντικείμενο στην ομάδα που έχει το πλησιέστερο κέντρο.
3. Όταν έχουν αντιστοιχιστεί όλα τα αντικείμενα, υπολογίζει ξανά τις θέσεις των K κεντροειδών.
4. Επαναλαμβάνει τα βήματα 2 και 3 έως ότου τα κεντροειδή δεν κινούνται πλέον.

Decision Trees

Το δέντρο αποφάσεων είναι μια δομή δέντρου, όπου οι εσωτερικοί κόμβοι δηλώνουν μια δοκιμή σε ένα χαρακτηριστικό, κάθε κλαδί αντιπροσωπεύει τα αποτελέσματα της δοκιμής και ο κόμβος φύλλων αντιπροσωπεύει τις ετικέτες κλάσης. Η επαγωγή του δέντρου αποφάσεων είναι η εκμάθηση των δέντρων αποφάσεων από εκπαιδευτικές πλειάδες με σήμανση τάξης.

Η κατασκευή δέντρων αποφάσεων είναι απλή και γρήγορη και δεν χρειάζεται ιδιαίτερη γνώση και ως εκ τούτου αποτελεί κατάλληλη λύση για την ανακάλυψη διερευνητικών γνώσεων. Γενικά, οι ταξινομητές δέντρων αποφάσεων είναι ακριβής, αλλά η επιτυχής χρήση τους εξαρτάται από τα διαθέσιμα δεδομένα. Τα δέντρα αποφάσεων χρησιμοποιούνται για την ταξινόμηση και οι κανόνες ταξινόμησης δημιουργούνται εύκολα από αυτά. Μια άγνωστη πλειάδα X μπορεί να ταξινομηθεί, δεδομένης της τιμής χαρακτηριστικού δοκιμάζοντας τις τιμές χαρακτηριστικών έναντι του δέντρου απόφασης. Ο γενικός αλγόριθμος αποφάσεων λαμβάνει το σύνολο δεδομένων εκπαίδευσης, τη λίστα χαρακτηριστικών και τη μέθοδο επιλογής χαρακτηριστικών ως εισαγωγή. Ο αλγόριθμος δημιουργεί έναν κόμβο και στη συνέχεια εφαρμόζει τη μέθοδο επιλογής χαρακτηριστικών για τον προσδιορισμό των καλύτερων κριτηρίων διαχωρισμού, ενώ ο κόμβος που δημιουργείται ονομάζεται από αυτό το χαρακτηριστικό. Το υποσύνολο των εκπαιδευτικών πλειάδων σχηματίζεται χρησιμοποιώντας το χαρακτηριστικό διαχωρισμού.

Ο αλγόριθμος καλείται αναδρομικά για κάθε υποσύνολο, έως ότου το υποσύνολο περιέχει πλειάδες της ίδιας κλάσης. Όταν το υποσύνολο περιέχει πλειάδες από την ίδια τάξη, ένα φύλλο επισυνάπτεται με μια ετικέτα της μεγαλύτερης τάξης από το training set της ρίζας. Τα ID3, C4.5 και CART που παρουσιάζονται στην συνέχεια υιοθετούν μια

άπληστη – οπισθοδρομική προσέγγιση, στην οποία τα δέντρα αποφάσεων κατασκευάζονται με αναδρομικό τρόπο, από πάνω προς τα κάτω, με «διαίρεση και κατάκτηση» τρόπο. Οι τρεις μέθοδοι ποικίλλουν στο κριτήριο διαχωρισμού που χρησιμοποιείται για την κατάτμηση των δεδομένων.

ID3 algorithm

Το ID3 είναι ένας επαναληπτικός αλγόριθμος που χρησιμοποιεί το κέρδος πληροφοριών ως κριτήριο διαχωρισμού για την κατασκευή του δέντρου. Για κάθε χαρακτηριστικό A, η μέθοδος υπολογίζει το κέρδος πληροφοριών ως τη διαφορά μεταξύ των πληροφοριών που απαιτούνται για την ταξινόμηση του συνόλου δεδομένων με βάση μόνο την αναλογία και τις πληροφορίες που απαιτούνται για την ταξινόμηση μετά την κατάτμηση στο A. Οι αναμενόμενες πληροφορίες που απαιτούνται για την ταξινόμηση μιας πλειάδας στην εκπαίδευση δίνεται από την παρακάτω σχέση:
όπου pi είναι η πιθανότητα μιας αυθαίρετη πλειάδα στο D που ανήκει στην κλάση Ci και

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

εκτιμάται ως ο λόγος του αριθμού των παρουσιών στην κλάση Ci σε D προς το συνολικό αριθμό παρουσιών σε D. Ο αριθμός των πληροφοριών που απαιτούνται ακόμη για την ταξινόμηση D, μετά τον διαχωρισμό τους χρησιμοποιώντας το A με v πιθανές τιμές υπολογίζεται χρησιμοποιώντας τον παρακάτω τύπο:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Το κέρδος των πληροφοριών που αποκτήθηκε διακλαδίζοντας το training set στο χαρακτηριστικό A δίνεται όπως στο:

$$Gain(A) = Info(D) - Info_A(D)$$

Ο αλγόριθμος εφαρμόζεται αναδρομικά για τα υποσύνολα έως ότου όλα τα μέλη του συνόλου ανήκουν στην ίδια τάξη.

C4.5 Algorithm

Ο αλγόριθμος C4.5 είναι διάδοχος του ID3 που χρησιμοποιεί την αναλογία κέρδους ως κριτήριο διαχωρισμού για τη διαμέριση του συνόλου δεδομένων. Ο αλγόριθμος εφαρμόζει ένα είδος ομαλοποίησης στην απόκτηση πληροφοριών χρησιμοποιώντας μια τιμή "splitinformation". Οι πληροφορίες διαχωρισμού για ένα χαρακτηριστικό A με τιμές v ορίζονται ως εξής:

$$split\inf(A) = -\sum_{i=1}^{v} \frac{|D_i|}{|D|} \times \log_2\left(\frac{|D_i|}{|D|}\right)$$

όπού | Di | είναι ο αριθμός παρουσιών στο training set D με τιμή i^{th}για τα χαρακτηριστικά A και | D | είναι ο συνολικός αριθμός παρουσιών στο training set. Ο λόγος κέρδους ορίζεται όπως παρακάτω και το χαρακτηριστικό με τη μέγιστη αναλογία κέρδους επιλέγεται ως χαρακτηριστικό διαχωρισμού.

$$Gainratio(A) = \frac{Gain(A)}{split\inf(A)}$$

CART Algorithm

Το CART είναι μια αναδρομική μέθοδος διαμέρισης που δημιουργεί δέντρα ταξινόμησης και παλινδρόμησης για την πρόβλεψη συνεχών εξαρτώμενων μεταβλητών και κατηγορηματικών μεταβλητών πρόβλεψης. Η θεμελιώδης ιδέα είναι να επιλεγεί κάθε διαίρεση ενός υποσυνόλου έτσι ώστε τα δεδομένα σε κάθε ένα από τα υποσύνολα «απογόνων» να είναι καθαρότερα από τα δεδομένα στο γονικό υποσύνολο. Ο δείκτης Gini χρησιμοποιείται για τη μέτρηση της καθαρότητας του D, του συνόλου των πλειάδων στο training set όπως αναφέρεται παρακάτω:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Όπου pi είναι η πιθανότητα ότι μια παρουσία στο D ανήκει στην κλάση Ci και εκτιμάται ότι χρησιμοποιεί :

$$p_i = \frac{|C_{i,D}|}{|D|}$$

| Ci, D | είναι ο αριθμός των παρουσιών στο D που ανήκουν στην κατηγορία Ci και | D | είναι ο συνολικός αριθμός παρουσιών στο trainingset. Ο δείκτης Gini χρησιμοποιεί δυαδικό διαχωρισμό για κάθε χαρακτηριστικό, για ένα διακριτό χαρακτηριστικό A με v

γνωστές διακριτές τιμές, P = {a1, a2, a3,…, av}, το καλύτερο δυαδικό split καθορίζεται εξετάζοντας όλα τα πιθανά υποσύνολα του P.
Ο δείκτης gini ενός δυαδικού διαχωρισμού στο A που χωρίζει το σετ εκπαίδευσης D σε D1 και D2 είναι:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

Για ένα χαρακτηριστικό διακριτής τιμής, το διαχωρισμό που δίνει τον ελάχιστο δείκτη gini επιλέγεται ως χαρακτηριστικό διαχωρισμού. Για ένα χαρακτηριστικό συνεχούς αποτίμησης, το σημείο που δίνει τον ελάχιστο δείκτη gini επιλέγεται ως το σημείο διαχωρισμού του χαρακτηριστικού. Το σύνολο πιθανών σημείων διαχωρισμού προσδιορίζεται με ταξινόμηση των τιμών και στη συνέχεια με τη λήψη του μέσου σημείου των παρακείμενων τιμών. Η χρήση του δείκτη gini υπολογίζεται για το χαρακτηριστικό, όπου το D1 είναι το σύνολο παρουσιών με τιμή A μικρότερη ή ίση με το σημείο διαχωρισμού και το D2 είναι το σύνολο παρουσιών με τιμή A μεγαλύτερη από το σημείο διαχωρισμού. Η μείωση των μη καθαρών που προκύπτει από ένα δυαδικό διαχωρισμό σε ένα διακριτό ή συνεχές αποτιμημένο χαρακτηριστικό δίνεται ως εξής:

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Το χαρακτηριστικό που μεγιστοποιεί τη μείωση των μη-καθαρών επιλέγεται ως χαρακτηριστικό διαχωρισμού.

Tree Pruning

Όταν δημιουργείται ένα δέντρο αποφάσεων, ορισμένοι κλάδοι ενδέχεται να αντικατοπτρίζουν ανωμαλίες στα training data λόγω θορύβου, που αφαιρείται με τις τεχνικές κλαδέματος - pruningτων δέντρων. Οι τεχνικές αυτές χρησιμοποιούν στατιστικά μέτρα για την αφαίρεση των λιγότερο αξιόπιστων κλάδων. Οι δύο προσεγγίσεις είναι αυτή πριν το κλάδεμα και αυτή μετά. Στην προσέγγιση προ-κλαδέματος το δέντρο κλαδεύεται αποφασίζοντας να μην χωρίσει περαιτέρω το υποσύνολο των πλειάδων κατάρτισης σε έναν δεδομένο κόμβο. Οι τεχνικές μετά-το-κλάδεμα αφαιρούν τα δέντρα από ένα πλήρως αναπτυγμένο δέντρο, αντικαθιστώντας ένα υπόδεντρο με ένα φύλλο που φέρει την ετικέτα ως την πιο συχνή κατηγορία σε αυτό.
Το CART χρησιμοποιεί αλγόριθμο κλαδέματος πολυπλοκότητας κόστους, μια προσέγγιση μετά τον κλάδεμα που υποθέτει ότι η μεροληψία στο σφάλμα επανακατάστασης ενός δέντρου αυξάνεται γραμμικά με τον αριθμό των κόμβων φύλλων. Η τεχνική κλαδέματος ξεκινά από το κάτω μέρος του δέντρου. Για κάθε εσωτερικό κόμβο, N, υπολογίζει την

πολυπλοκότητα κόστους του υποδέντρου στο N και την πολυπλοκότητα κόστους του υποδέντρου στο N εάν επρόκειτο να κλαδευτεί και συγκρίνονται οι δύο τιμές. Εάν το κλάδεμα του δευτερεύοντος δέντρου στον κόμβο N θα είχε ως αποτέλεσμα μικρότερη πολυπλοκότητα κόστους, τότε το υποδένδρο κλαδεύεται. Αυτή η τεχνική χρησιμοποιεί κλάδεμα σετ κλαδιών με ετικέτα κλάσης και χρησιμοποιείται για την εκτίμηση της πολυπλοκότητας του κόστους. Αυτό το σετ είναι ανεξάρτητο από το trainingset που χρησιμοποιείται για την κατασκευή του μη τεμαχισμένου δέντρου και από οποιοδήποτε σύνολο δοκιμών που χρησιμοποιείται για τον υπολογισμό της ακρίβειας. Ο αλγόριθμος δημιουργεί ένα σύνολο σταδιακά κλαδευμένων δέντρων. Γενικά, προτιμάται το μικρότερο δέντρο αποφάσεων που ελαχιστοποιεί την πολυπλοκότητα κόστους.
Το C4.5 χρησιμοποιεί pessimistic pruning παρόμοιο με τη μέθοδο πολυπλοκότητας κόστους, δηλαδή χρησιμοποιεί εκτιμήσεις ποσοστών σφάλματος για τη λήψη αποφάσεων σχετικά με το κλάδεμα υποδένδρων. Ωστόσο, η μέθοδος δεν χρησιμοποιεί setprun, αλλά εκτιμά τα ποσοστά σφάλματος χρησιμοποιώντας το trainingset.

Support Vector Machine

Οι μηχανές SVM, που προτάθηκαν από τον Vapnik έχουν χρησιμοποιηθεί με επιτυχία σε πολλά κομμάτια της μηχανικής μάθησης. Συγκεκριμένα, προσφέρουν μια καλή εκτίμηση της αρχής ελαχιστοποίησης του διαρθρωτικού κινδύνου. Οι κύριες ιδέες πίσω από αυτήν τη μέθοδο είναι:
- Τα δεδομένα χαρτογραφούνται σε χώρο υψηλής διαστάσεων μέσω μετασχηματισμού που βασίζεται σε γραμμικό, πολυωνυμικό ή πυρήνα gaussien.
- Οι τάξεις διαχωρίζονται (στον νέο χώρο) με γραμμικούς ταξινομητές, οι οποίοι μεγιστοποιούν το περιθώριο (απόσταση μεταξύ των τάξεων).
- Τα υπερπρογράμματα μπορούν να προσδιοριστούν από μερικούς αριθμούς σημείων: καθένα από αυτά ονομάζεται φορέας υποστήριξης.
Έτσι, η πολυπλοκότητα ενός ταξινομητή SVM εξαρτάται, όχι από τη διάσταση του χώρου δεδομένων, αλλά από τον αριθμό των διανυσμάτων υποστήριξης που απαιτούνται για την πραγματοποίηση του καλύτερου διαχωρισμού. Το SVM έχει ήδη εφαρμοστεί στον τομέα της ταξινόμησης του κειμένου σε πολλές μελέτες.

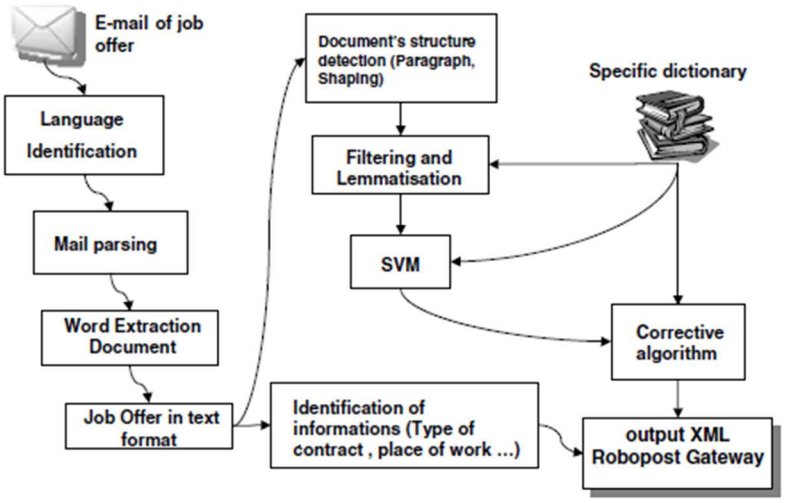
4.2 Αντιπροσωπευτικά Μοντέλα e-Recruitment με Machine Learning Techniques

Αναλύθηκε η δομή των συστημάτων που χρησιμοποιούν τεχνικές machine learning, τα υπέρ και τα κατά στην τάση για περισσότερη αυτοματοποίηση ώστε να γίνονται ολοένα και πιο αποτελεσματικά οι προσλήψεις. Στην σύγχρονη κοινωνία που συνεχώς αναπτύσσεται τα συστήματα θα πρέπει να εξελίσσονται και αυτά, λαμβάνοντας ολοένα και περισσότερες παραμέτρους υπόψη, πέρα από το εκπαιδευτικό υπόβαθρο που κάποτε ήταν το σήμα κατατεθέν. Το εκπαιδευτικό υπόβαθρο και η προηγούμενη εργασιακή εμπειρία παραμένουν στην υψηλότερη θέση των ικανοτήτων ενός υποψηφίου, όμως σημαντική θέση έχουν αρχίσει να λαμβάνουν παράμετροι όπως της ψυχοσύνθεσης του ατόμου ή στοιχεία της προσωπικότητάς του με βάση την παρουσία του στα social media. Ένα τέτοιο παράδειγμα θα αναλυθεί παρακάτω, όπως και τα πλέον αντιπροσωπευτικά μοντέλα e-recruitments που χρησιμοποιούν τεχνολογία machine learning.

E-Gen

Η εκθετική ανάπτυξη του Διαδικτύου επέτρεψε την ανάπτυξη μιας διαδικτυακής αγοράς αναζήτησης εργασίας με την μάζα των πληροφοριών που λαμβάνονται μέσω της απόκρισης των υποψηφίων να είναι τόσο μεγάλη που είναι δύσκολο για τις εταιρείες να διαχειριστούν. Είναι επομένως απαραίτητο να επεξεργάζονται αυτές οι πληροφορίες με αυτόματο ή υποβοηθούμενο τρόπο. Οι Laboratoire Informatique d'Avignon (LIA) και Aktor Interactive έχουν αναπτύξει το σύστημα E-Gen [24] προκειμένου να επιλύσουν αυτό το πρόβλημα, αποτελούμενο από δύο κύριες ενότητες:
1. Μια ενότητα για την εξαγωγή πληροφοριών από εταιρικά e-mail που αποστέλλονται σε αυτό βιογραφικά σημειώματα.
2. Μια ενότητα για την ανάλυση και τον υπολογισμό της κατάταξης της καταλληλότητας των υποψηφίων μέσω της συνοδευτικής επιστολής και του βιογραφικού σημειώματος.

Για να εξαχθούν χρήσιμες πληροφορίες, το σύστημα αναλύει το περιεχόμενο των e-mail που περιέχουν βιογραφικά σημειώματα. Σε αυτό το βήμα, υπάρχουν πολλές δυσκολίες και ενδιαφέροντα προβλήματα προς επίλυση που σχετίζονται με την Επεξεργασία Φυσικής Γλώσσας (NLP), για παράδειγμα, ότι οι δημοσιεύσεις θέσεων εργασίας γράφονται σε ελεύθερη μορφή, μη δομημένες, με ορισμένες ασάφειες, τυπογραφικά λάθη κ.λπ. αλλά αυτό αφορά μόνο τον χειρισμό των απαντήσεων και όχι την ένταξη των προσφορών εργασίας.



Εικ. 29 E-Gen overflow

Για να δημοσιευθεί on-line μια θέση εργασίας, απαιτούνται ορισμένες πληροφορίες από τους πίνακες εργασίας. Επομένως, πρέπει να βρεθούν αυτά τα πεδία στην ανάρτηση της θέσης εργασίας και να συμπεριληφθούν στο έγγραφο XML. Δημιουργούνται διαφορετικές λύσεις για να εντοπιστεί κάθε τύπος πληροφοριών:
- Μισθός: Δημιουργήθηκαν τακτικές εκφράσεις και κανόνες για τον εντοπισμό εκφράσεων όπως "Μισθός: από Χ σε Υ", "Μισθός: μεταξύ Χ και Υ" ή "Χ σταθερός μισθός με μπόνους" κ.λπ.
- Τόπος εργασίας: Δημιουργήθηκε πίνακας με πεδία περιοχής, πόλης και τμήματος για να βρεθεί η καταχώριση θέσης σε μια ανάρτηση εργασίας. Οι περισσότεροι από τους πίνακες εργασίας κατηγοριοποιούν τις θέσεις εργασίας ανάλογα με την περιοχή για να βοηθήσουν τους αιτούντες στην αναζήτηση τους.
- Εταιρεία: Για να μπορέσει να ενσωματώσει λογότυπα σε προσφορές εργασίας, μια λίστα πελατών προστέθηκε στο σύστημα για να εντοπίζει το όνομα της εταιρείας στις θέσεις εργασίας.
Άλλες πληροφορίες ανακτώνται με παρόμοιες διαδικασίες (σύμβαση, αναφορά, διάρκεια αποστολής κ.λπ.). Τέλος, αποστέλλεται μια αναφορά στον χρήστη για να δείξει τα πεδία που εντοπίστηκαν σωστά και τα πεδία που δεν βρέθηκαν (είτε λόγω σφάλματος εξαγωγής ή ελλείψεων πληροφοριών στην προσφορά εργασίας).
Ένα παράδειγμα καταχώρισης εργασίας παρουσιάζεται στον παρακάτω πίνακα. Η εξαγωγή της βάσης δεδομένων κατέστη δυνατή χωρίς χειροκίνητη κατηγοριοποίηση. Μια πρώτη ανάλυση αυτής δείχνει ότι οι προσφορές εργασίας αποτελούνται συχνά από παρόμοια τμήματα πληροφοριών που παραμένουν, ωστόσο, έντονα αδόμητα. Κάθε ανάρτηση εργασίας χωρίζεται σε τέσσερις κατηγορίες, ως εξής:
1. "Περιγραφή_εταιρίας": Σύντομη περίληψη της επιχείρησης που προσλαμβάνει.
2. "Τίτλος": πιθανώς ο τίτλος εργασίας.
3. "Αποστολή": μια σύντομη περιγραφή της φύσης της επιχείρησης.
4. "Προφίλ": απαιτούμενες δεξιότητες και γνώσεις για τη θέση.

This french firm, specialised in chemical analysis, is looking for:
PERSON IN CHARGE OF LABORATORY TRANSFER
South East
You will be in charge of regrouping the transfer activities of different analysis laboratories. You will analyse, conduct and implement the necessary phases of the project, respecting budgets and previously defined, dead lines.
Your solution will need to consider different parameters of the project (social, logistic, materials, data processing...) and integrate a roadmap (production, methods, accreditations, development, commercial...).
Being a post graduate in chemical engineering with a focus on environmental analytical chemistry, you have already led an activity transfer project.
Fluent English required. Please send your CV and cover letter indicating reference number VA 11/06 to beatrice.lardon@atalan.fr

Εικ. 31 Δημοσίευση Αγγελίας Εργασίας

Number of job postings	D=1000	
Number total of Segments	P=15621	
Number of Segments "Title"	1000	6.34%
Number of Segments "Description_of_the_company"	3966	25.38%
Number of Segments "Mission" description	4401	28.17%
Number of Segment "Profile" description	6263	40.09%

Εικ. 30 Στατιστικά των κατηγοριών της δημοσιευμένης αγγελίας

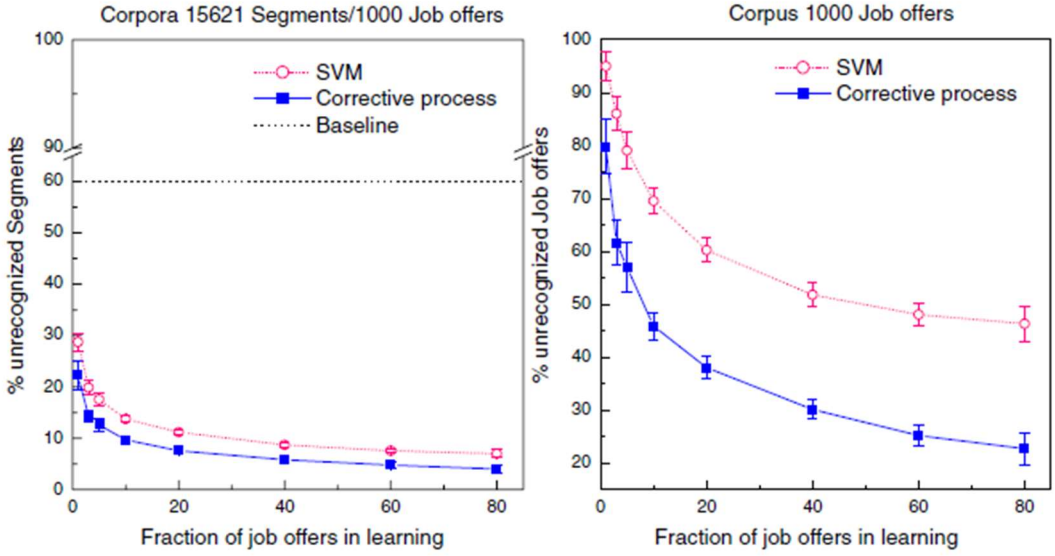
Βάση της εικόνας 31, μια προεπεξεργασία των πληροφοριών έχει ολοκληρωθεί για να αποκτήσει μια κατάλληλη αναπαράσταση στο Vector Space Model (VSM). Κυρίως θα ακολουθηθεί διαγραφή των ακόλουθων στοιχείων: ρήματα και λειτουργικές λέξεις (να είναι, να έχουν, να είναι σε θέση, να χρειάζονται, ...), κοινές εκφράσεις (για παράδειγμα, κάθε ένα, ...), αριθμοί (σε αριθμητική ή / και μορφή κειμένου) και σύμβολα όπως $, #, * κ.λπ. επειδή αυτοί οι όροι ενδέχεται να προκαλέσουν «θόρυβο» στην ταξινόμηση. Η επεξεργασία «Lemmatisation» έχει επίσης πραγματοποιηθεί για τη λήψη σημαντικής μείωσης του λεξικού. Αποτελείται από την εύρεση της ρίζας των ρήματος και τη μετατροπή πληθυντικών και / ή θηλυκών λέξεων σε αρσενική μοναδική μορφή. Αυτή η διαδικασία επιτρέπει τη μείωση της διαστατικότητας που δημιουργεί σοβαρά προβλήματα αναπαράστασης των τεράστιων διαστάσεων στους πίνακες. Χρησιμοποιούνται επίσης άλλοι μηχανισμοί μείωσης του λεξικού: οι σύνθετες λέξεις αναγνωρίζονται από ένα λεξικό και μετά μετατρέπονται σε έναν μοναδικό όρο. Όλες αυτές οι διαδικασίες επιτρέπουν να αποκτηθεί μια αναπαράσταση σε bag-of-words (πίνακας συχνοτήτων / απουσιών κειμένων τμημάτων (σειρές) και λεξιλόγιο όρων (στήλες).

Μηχανισμός Marcov

Τα προκαταρκτικά πειράματα δείχνουν ότι η κατηγοριοποίηση τμημάτων χωρίς τοποθέτηση τμήματος μιας θέσης εργασίας δεν είναι αρκετή, και μπορεί να είναι πηγή

σφαλμάτων. Η εικόνα 32 δείχνει ότι το SVM παράγει μια καλή ταξινόμηση τμημάτων, αλλά οι θέσεις εργασίας (έγγραφα) σπάνια ταξινομούνται πλήρως. Επομένως, λόγω του τεράστιου αριθμού των περιπτώσεων, οι κανόνες δεν φαίνεται να είναι ο καλύτερος τρόπος επίλυσης του προβλήματος. Έτσι έχει εφαρμοστεί ένας μηχανισμός με 6 καταστάσεις ("Έναρξη" (0), "Τίτλος" (1), "Περιγραφή_ της_εταιρίας" (2), "Αποστολή" (3), "Προφίλ" (4) και "Τέλος" (5)).

$$M = \begin{array}{c} \\ \text{Start} \\ \text{Title} \\ \text{Description} \\ \text{Mission} \\ \text{Profil} \\ \text{End} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Start} & \text{Title} & \text{Description} & \text{Mission} & \text{Profil} & \text{End} \end{array} \\ \begin{pmatrix} 0 & 0,01 & 0,99 & 0 & 0 & 0 \\ 0 & 0,05 & 0,02 & 0,94 & 0 & 0 \\ 0 & 0,35 & 0,64 & 0,01 & 0 & 0 \\ 0 & 0 & 0 & 0,76 & 0,24 & 0 \\ 0 & 0 & 0 & 0 & 0,82 & 0,18 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$



Εικ. 32 Αποτελέσματα SVM και η σωστή διαδικασία για τα τμήματα στο αριστερό μέρος και των προσφερόμενων εργασιών στα δεξιά

Το MatrixM (ισοδ. 1) δείχνει τις τιμές των πιθανοτήτων:

Τα προκαταρκτικά αποτελέσματα που λαμβάνονται με τη μέθοδο SVM δείχνουν μια ταξινομημένη απόδοση των τμημάτων. Ωστόσο, κατά την ταξινόμηση μιας πλήρους θέσης εργασίας, ορισμένα τμήματα ταξινομήθηκαν λανθασμένα, χωρίς κανονική συμπεριφορά. Προκειμένου να αποφευχθεί αυτό το είδος σφάλματος, εφαρμόστηκε μια μετεπεξεργασία, με βάση τον αλγόριθμο Viterbi. Η ταξινόμηση SVM για κάθε τμήμα

παρέχει μια προβλεπόμενη κλάση, και επομένως για μια θέση εργασίας, έχουμε μια ακολουθία τάξης.

Δηλαδή η ακολουθία 0 → 2 → 2 → 1 → 3 → 3 → 4 → 5 , και μεταφράζεται σε: "Έναρξη" → "Περιγραφή της εταιρείας" → "Περιγραφή της εταιρείας" → "Τίτλος" → "Αποστολή" → "Αποστολή" → "Προφίλ" → Τέλος).

Ένας κλασικός αλγόριθμος Viterbi θα υπολογίσει την πιθανότητα της ακολουθίας. Εάν η ακολουθία δεν είναι πιθανή, ο αλγόριθμος Viterbi επιστρέφει 0. Όταν η ακολουθία έχει μηδενική πιθανότητα, η διορθωτική διαδικασία επιστρέφει την ακολουθία με ένα ελάχιστο σφάλμα και μέγιστη πιθανότητα (σε σύγκριση με την αρχική ακολουθία που δημιουργήθηκε από το SVM). Τα πρώτα αποτελέσματα ήταν ενδιαφέροντα αλλά περιελάμβαναν σημαντικό χρόνο επεξεργασίας. Έτσι εισήχθη μια βελτίωση χρησιμοποιώντας τον αλγόριθμο Branch and Bound για το κλάδεμα του δέντρου: μόλις βρεθεί μια αρχική λύση, το σφάλμα και η πιθανότητα συγκρίνονται κάθε φορά που γίνεται επεξεργασία μιας νέας ακολουθίας. Εάν και πάλι δεν βελτιωθεί, το τέλος της ακολουθίας δεν υπολογίζεται. Η χρήση αυτού του αλγορίθμου επιτρέπει να βρεθεί μια βέλτιστη λύση, αλλά όχι την καλύτερη στιγμή (έχει εκθετική πολυπλοκότητα). Σε δοκιμή, αυτή η στρατηγική υπολογίζει ακολουθίες ≤ 50 σύμβολα σε περίπου 4 δευτερόλεπτα.

```
Calcul next symbol()
Processes the current sequence (Viterbi): full sequence, their probability, and the
number of errors
if the error of current sequence > max error found then
    return current sequence
end
if symbol is the last of the sequence then
    if current error < max error then
        maxerror = currenterror;
    end
    return sequence;
end
else
    foreach symbol successor of the sequence do
        current sequence = Calcul next symbol()
        if current sequence is the best sequence then
            bestsequence = currentsequence;
            if current error < max error then
                maxerror = currenterror;
            end
        end
    end
end
```

Εικ. 34 Εκτέλεση Διαδικασίας με την χρήση αλγορίθμου Branch and Bound

Εικ. 33 Block boundary error

Τα δεδομένα που χρησιμοποιήθηκαν D = 1, 000 θέσεις εργασίας χωρισμένες σε P = 15.621 τμήματα. Κάθε δοκιμή πραγματοποιήθηκε 20 φορές με τυχαία κατανομή. Η εικόνα 4 δείχνει τη σύγκριση μεταξύ των αποτελεσμάτων που ελήφθησαν από SVMτο correctiveprocess. Τα αποτελέσματα είναι επιτυχημένα και δείχνουν ότι ακόμη και με ένα μικρό κλάσμα 20% του συνόλου, ο ταξινομητής SVM επιτυγχάνει χαμηλό ποσοστό εσφαλμένης ταξινόμησης (λιγότερο από 10% σφάλμα). Το corrective process (σταθερή γραμμή) δίνει πάντα καλύτερα αποτελέσματα από το SVM ανεξάρτητα από το κλάσμα των προτύπων. Η καμπύλη στα δεξιά συγκρίνει τα αποτελέσματα που λαμβάνονται από κάθε μέθοδο σύμφωνα με μη αναγνωρισμένες θέσεις εργασίας. Μπορούμε επίσης να δούμε μια σημαντική βελτίωση του αριθμού των θέσεων εργασίας που αναγνωρίζονται με το corrective process. Ο αλγόριθμος SVM φτάνει το μέγιστο ≈ 50% των μη αναγνωρισμένων θέσεων εργασίας, ενώ το corrective process δίνει το 20% των μη αναγνωρισμένων θέσεων εργασίας, οπότε μια βελτίωση μεγαλύτερη από 50% στη βαθμολογία SVM. Μια ανάλυση λανθασμένων ταξινομήσεων θέσεων εργασίας δείχνει ότι περίπου το 10% των θέσεων εργασίας περιέχει ένα ή δύο σφάλματα. Αυτά τα λανθασμένα ταξινομημένα τμήματα αντιστοιχούν γενικά στα όρια μεταξύ 2 διαφορετικών κατηγοριών, όπως φαίνεται στην εικόνα 34.

CaSMoS

Οι μηχανές αναζήτησης είναι γνωστό ότι είναι καλοί στην ανάκτηση ενός συνόλου σχετικών εγγράφων που ταιριάζουν με ένα δεδομένο ερώτημα από ένα τεράστιο σύνολο εγγράφων, οπότε για συσχετίσεις σε πραγματικό χρόνο μεγάλης κλίμακας δεδομένων είναι φυσική επιλογή τα προτεινόμενα στοιχεία να μοντελοποιηθούν ως δομημένα έγγραφα και να οριστεί μια συνάρτηση για την κατασκευή ερωτημάτων για την επιλογή υποψηφίων. Όμως, η αποτελεσματική χρήση μιας μηχανής αναζήτησης για την εκλογή ενός συνόλου υποψηφίων (με καλή ακρίβεια και ανάκληση, και ένα μικρό μέγεθος υποψηφίων) δεν είναι ασήμαντη. Το CaSMoS[23], είναι ένα πλαίσιο επιλογής υποψηφίων που χρησιμοποιεί το ερώτημα Weighted AND (WAND). Έχει σχεδιαστεί για την περικοπή άσχετων εγγράφων και την ανάκτηση εγγράφων που ενδέχεται να αποτελούν μέρος των κορυφαίων αποτελεσμάτων για το ερώτημα. Εφαρμόζεται ένας αλγόριθμος περιορισμένης επιλογής δυνατοτήτων για να μας αποδώσει θετικά βάρη για συνδυασμούς δυνατοτήτων που χρησιμοποιούνται ως μέρος του σταθμισμένου ερωτήματος επιλογής υποψηφίων. Έχει εφαρμοστεί και αναπτυχθεί αυτό το σύστημα για να εκτελείται σε πραγματικό χρόνο χρησιμοποιώντας την πλατφόρμα αναζήτησης Galene του LinkedIn. Η ανάπτυξη αυτού του συστήματος ως μέρος της μηχανής προτάσεων εργασίας του LinkedIn έχει οδηγήσει σε σημαντική μείωση του λανθάνοντος

χρόνου (έως και 25%) χωρίς να θυσιαστεί η ποιότητα των ανακτημένων αποτελεσμάτων, ανοίγοντας έτσι το δρόμο για πιο εξελιγμένα μοντέλα βαθμολογίας.

Τα εξατομικευμένα συστήματα αναζήτησης και προτάσεων αποτελούν τη ραχοκοκαλιά πολλών προϊόντων που απευθύνονται σε χρήστες του διαδικτύου όπως το LinkedIn. Οι υποκείμενες εργασίες αναζήτησης / προτάσεων μπορούν να μοντελοποιηθούν ως προβλήματα ανάκτησης πληροφοριών, όπου το ερώτημα αποτελείται από το περιβάλλον χρήστη, τα ενδιαφέροντα που εκφράζονται μέσω προφίλ χρήστη και μέσω λέξεων-κλειδιών αναζήτησης (παρέχονται από τον χρήστη), εάν υπάρχουν, και το έγγραφο που αποτελείται από εκατομμύρια αντικείμενα (π.χ., το σύνολο των θέσεων εργασίας σε περίπτωση εξατομικευμένης αναζήτησης εργασίας).
Από τη μία πλευρά, τα σχετικά αποτελέσματα πρέπει να είναι διαθέσιμα μόλις ένας νέος χρήστης δημιουργήσει ένα προφίλ ή ένας χρήστης αλλάξει το προφίλ. Από την άλλη πλευρά, το σύνολο των έγκυρων στοιχείων μπορεί να αλλάζει συχνά με την πάροδο του χρόνου (π.χ., λόγω της άφιξης νέων θέσεων εργασίας ή της λήξης των θέσεων εργασίας), και ως εκ τούτου η ανανέωση των αποτελεσμάτων είναι ζωτικής σημασίας (π.χ., οι χρήστες θα ήθελαν να δουν συστάσεις σχετικών θέσεων εργασίας μόλις αναρτηθούν).

Πλατφόρμα αναζήτησης Galene

Το Galene χρησιμοποιεί τους Lucene ανεστραμμένους πίνακες περιεχομένου στον πυρήνα του και παρέχει αρκετές λειτουργίες που είναι επιθυμητές για την εφαρμογή. Η πλατφόρμα Galene υποστηρίζει την περιοδική ανακατασκευή ολόκληρων δεδομένων ευρετηρίου εκτός σύνδεσης στο Hadoop και προσφέρει μια αρχιτεκτονική κατανεμημένων υπηρεσιών με διαχωρισμό ευρετηρίου και σπασμένα κομμάτια αντιγράφων. Το ευρετήριο όταν αναπτύσσεται χωρίς σύνδεση στην ηλεκτρονική υπηρεσία συμπληρώνεται από μια διαδικτυακή βαθμίδα ζωντανών ευρετηρίων, η οποία αντλεί έγγραφα από το Kafka. Το στατικό βασικό ευρετήριο και οι σειρές live index παρέχουν μια μοναδική δυναμική προβολή του ευρετηρίου αναζήτησης στους χρήστες της υπηρεσίας. Το Galene παρέχει μια πλούσια γλώσσα ερωτημάτων που υποστηρίζει σταθμισμένα ερωτήματα (WAND). Προσφέρει επίσης ένα ευέλικτο API που επιτρέπει σε μεμονωμένες εφαρμογές να καθορίζουν επεκτάσεις ερωτημάτων και λειτουργίες επανεγγραφής ερωτημάτων, καθώς και αυθαίρετες λειτουργίες βαθμολογίας για την κατάταξη των αντιστοιχισμένων εγγράφων. Αυτά τα API επιτρέπουν την εύκολη ενσωμάτωση των μηχανογραφημένων βαθμολογιών και των υποψήφιων μονάδων επιλογής στις αναπτύξεις του Galene, συμπεριλαμβανομένων εκείνων που τροφοδοτούν τις μεθοδολογίες που συζητούνται σε αυτό το έγγραφο.

Επισκόπηση Λειτουργίας Συστήματος

Όπως έχει συζητηθεί, για κάθε χρήστη, εκατομμύρια αντικείμενα θα μπορούσαν να αντιστοιχούν σε σχέση με κάθε λειτουργία (χρήστης, στοιχείο) που χρησιμοποιείται στο μοντέλο βαθμολογίας. Ωστόσο, τα πιο σχετικά αντικείμενα είναι πολύ πιθανό να ταιριάζουν σε πολλές λειτουργίες (χρήστης, στοιχείο). Επομένως, η απαίτηση ενός αντικειμένου να ταιριάζει σε πολλές δυνατότητες αφαιρεί τον αριθμό των αντικειμένων που θα σημειωθούν, διατηρώντας παράλληλα τα πιο σχετικά αντικείμενα. Επιπλέον, μπορεί να είναι επιθυμητό να δοθούν διαφορετικά βάρη σε διαφορετικούς συνδυασμούς δυνατοτήτων.

Σε ένα σενάριο εφαρμογής, ένας συνδυασμός {(user_title, job_title), (user_skills, job_skills)} είναι πιο σημαντικός από έναν συνδιασμό {(user_industry, job_industry), (user_seniority, job_seniority)}. Επιλέγεται το μοντέλο επιλογής ερώτησης WAND. Το F δηλώνει το σύνολο των χαρακτηριστικών (χρήστη, στοιχείο). Επίσης, 1≤ i ≥k, Ci είναι Boolean μεταβλητή που δηλώνει εάν ένα στοιχείο ταιριάζει σε συνδυασμό Fi υποσύνολο F των αντικειμένων (χρήστης, στοιχείο). Ομοίως, το Ci μπορεί να θεωρηθεί ως συνδυασμός σε σχέση με την Boolean έκδοση των αντίστοιχων δυνατοτήτων στο Fi. Κάθε ρήτρα Ci σχετίζεται με ένα θετικό βάρος wi. Ένα στοιχείο επιλέγεται εάν το άθροισμα των βαρών που σχετίζονται με ρήτρες που ισχύουν υπερβαίνει ένα όριο. Έτσι, καθορίζουμε το υποψήφιο μοντέλο επιλογής από την άποψη του WAND Boolean:

Για 1 δεδομένο στοιχείο WAND(C1;w1;… ;Ck;wk; θ) είναι αληθές μόνο αν

$$\sum_{1 \le i \le k} w_i \cdot x_i \ge \theta,$$

Δίνουμε μια εικόνα χρησιμοποιώντας ένα μοντέλο επιλογής υποψήφιου παιχνιδιού και το υποθετικό προφίλ χρήστη όπως παρακάτω. Ας υποθέσουμε ότι το μοντέλο έχει καθοριστεί με κατώφλι 0: 5 και σε όρους μόνο τεσσάρων ζευγών (όρος, βάρος)

1. ([user_title, job_title] ^ (user_skills, job_skills), 0,55)
2. ([user_title, job_title] ^ (user_position_summary, job_skills), 0,35)
3. ([user_industry, job_industry] ^ (user_position_summary, job_skills), 0,25)
4. ([user_industry, job_industry] ^ (user_seniority, job_seniority), 0,05)

Μια δημοσίευση εργασίας (στοιχείο) ικανοποιεί μια ρήτρα εάν οι υποκείμενες λειτουργίες θα είναι όλες μη μηδενικές. Για παράδειγμα, μια ανάρτηση εργασίας θα ικανοποιούσε την πρώτη ρήτρα εάν ο τίτλος της αντιστοιχεί στον τίτλο του χρήστη και υπάρχει τουλάχιστον μία κοινή ικανότητα μεταξύ της εργασίας και του χρήστη. Παρατηρείται ότι, για να επιλεγεί μια θέσης εργασίας, πρέπει είτε να ικανοποιεί την πρώτη ρήτρα είτε να ικανοποιεί τόσο τη δεύτερη όσο και την τρίτη ρήτρα. Για τον χρήστη αυτό (Εικόνα 7) μια ανάρτηση εργασίας για το "Software Engineer" (τίτλος) με το "Java" ως μία από τις δεξιότητες θα ικανοποιήσει την πρώτη πρόταση, και ως εκ τούτου θα επιλεγεί. Ομοίως, μια ανάρτηση θέσης εργασίας για το «Software Engineer» σε μια εταιρεία του κλάδου του Διαδικτύου που απαριθμεί τα «συστήματα σύστασης» ως μία από τις δεξιότητες θα ικανοποιήσει τόσο τη δεύτερη όσο και την τρίτη ρήτρα, και ως εκ τούτου θα επιλεγεί. Ωστόσο, μια ανάρτηση εργασίας για το "Product Manager" σε μια εταιρεία στον κλάδο του Διαδικτύου δεν θα επιλεγεί ακόμη και αν αναφέρει κάποια από τις δεξιότητες του χρήστη. Η τέταρτη ρήτρα δε, δεν επηρεάζει το εάν θα επιλεγεί μια θέσης εργασίας, δεδομένης της επιλογής του παραπάνω κατωφλίου.

User field	Value of field
Title	Software Engineer
Company	LinkedIn Corporation
Industry	Internet
Location	San Francisco Bay Area, CA, USA
Skills	C++, Java, Linux, Machine Learning
Position summary	recommendation systems, professional content

Εικ. 35 Προφίλ χρήστη

Μπορεί να γίνει χρήση του χειριστή ερωτημάτων WAND, το οποίο υποστηρίζεται από την πλατφόρμα αναζήτησης του LinkedIn Galene. Το Galene εφαρμόζει διάφορες τεχνικές βελτιστοποίησης ως μέρος της εκτέλεσης ερωτημάτων WAND. Για παράδειγμα, δεδομένου ότι τα βάρη που σχετίζονται με τους όρους είναι θετικά στο ερώτημα WAND, το σύστημα ανάκτησης μπορεί να επιλέξει ένα έγγραφο μόλις το άθροισμα των βαρών για τις ικανοποιημένες ρήτρες που έχουν αξιολογηθεί μέχρι στιγμής ξεπερνά το όριο. Σε αυτήν την περίπτωση, δεν χρειάζεται να αξιολογηθούν οι υπόλοιπες ρήτρες ερωτήματος. Αυτή η βελτιστοποίηση δεν θα ήταν δυνατή σε περίπτωση που επιτρέπονται βάρη αρνητικών ρητρών καθώς το σύστημα ανάκτησης θα έπρεπε τότε να αξιολογήσει όλες τις ρήτρες.

Εκπαίδευση μοντέλου επιλογής υποψηφίου με χρήση δεδομένων καταγραφής αλληλεπίδρασης χρήστη-στοιχείου

Στο σύστημα γίνεται η εκμάθηση εκτός σύνδεσης. Λαμβάνοντας υπόψη ένα μακρύ, περίπλοκο δομημένο ερώτημα, ο στόχος της επιλογής υποψηφίων μπορεί να θεωρηθεί ότι επιτυγχάνει διαχωρισμό μεταξύ στοιχείων που θα μπορούσαν ενδεχομένως να είναι στα κορυφαία αποτελέσματα και των στοιχείων που είναι πολύ απίθανο να είναι στα κορυφαία αποτελέσματα. Ως εκ τούτου, υιοθετούμε μια εποπτευόμενη προσέγγιση μηχανικής μάθησης για να εκπαιδεύσουμε το υποψήφιο μοντέλο επιλογής. Τα στοιχεία στα οποία βασίζεται παρουσιάζονται στον Αλγόριθμο 1 (Εικόνα 36).

Algorithm 1 Algorithm for Learning Candidate Selection Model

Input: Set F of (user, item) feature definitions; User-item interaction log data; Selection factor, T.

Output: Candidate selection query model, specified as WAND query predicate.

1: Generate training data (labeled (user, item) pairs) from user-item interaction log data.
2: Create the configuration set of possible conjunction clauses, by taking combinations of up to T (user, item) feature definitions.
3: Compute Boolean feature values for the (user, item) pairs present in the training data.
4: Generate Boolean feature vector of conjunction clauses, along with labels.
5: Learn weights for the WAND query.

Εικ. 36 Αλγόριθμος εκμάθησης μοντέλου

Δημιουργία δεδομένων εκπαίδευσης

Αυτό το στοιχείο εκχωρεί θετικές και αρνητικές ετικέτες σε ζεύγη (χρήστη, στοιχείο), με βάση τα δεδομένα καταγραφής αλληλεπίδρασης χρήστη-στοιχείου και χωρίζει αυτά τα δεδομένα σε σύνολα trainingset και testset. Τα δεδομένα καταγραφής περιέχουν συμβάντα αλληλεπιδράσεων χρηστών με την εφαρμογή προτάσεων. Τα συμβάντα θα μπορούσαν να αντιστοιχούν σε μια εμφάνιση στοιχείου (υποδεικνύοντας ότι το στοιχείο προτάθηκε στον χρήστη), καθώς και διαφορετικούς τύπους αλληλεπίδρασης, όπως κάνοντας κλικ (υποδεικνύοντας ότι ο χρήστης έκανε κλικ στο απόσπασμα αντικειμένου για να δεί περισσότερες λεπτομέρειες), εξοικονομώντας (υποδεικνύοντας ότι ο χρήστης αποθηκεύτηκε το στοιχείο για μελλοντική αναφορά) και άλλες αλληλεπιδράσεις για συγκεκριμένες εφαρμογές (π.χ., αίτηση για εργασία). Μπορούμε να συμπεράνουμε θετικές και αρνητικές ετικέτες για ζεύγη (χρήστη, στοιχείο) με διάφορους τρόπους, πιθανώς ανάλογα με τη συγκεκριμένη εφαρμογή. Για παράδειγμα, για έναν δεδομένο

χρήστη, τα στοιχεία στα οποία έγινε κλικ από τον χρήστη θα μπορούσαν να αντιμετωπίζονται ως θετικά παραδείγματα και τα στοιχεία που δεν εμφανίστηκαν ποτέ στον χρήστη θα μπορούσαν να αντιμετωπίζονται ως αρνητικά παραδείγματα.

Υπολογισμός των τιμών χαρακτηριστικών Boolean

Δεδομένου του συνόλου ορισμών χαρακτηριστικών με την μέθοδο Boolean δημιουργούνται τα αντίστοιχα χαρακτηριστικά δυνατοτήτων για τα ζεύγη (χρήστη, στοιχείο) που υπάρχουν στα δεδομένα εκπαίδευσης. Στο παραπάνω παράδειγμα σύστασης εργασίας, η δυνατότητα (user_skills, job_skills) θα αξιολογηθεί ως αληθής εάν και μόνο εάν υπάρχει τουλάχιστον μία κοινή ικανότητα μεταξύ του χρήστη και της εργασίας. Για να ενεργοποιηθεί αυτός ο υπολογισμός, το σύστημα που δημιουργεί τα δεδομένα δομημένου πεδίου για τους χρήστες γράφει ένα αντίγραφο αυτών των δεδομένων στο HDFS, εκτός από τη συμπλήρωση-ενημέρωση του χώρου αποθήκευσης πεδίων χρήστη (χρησιμοποιείται από το ηλεκτρονικό σύστημα προτάσεων επεξεργασίας ερωτημάτων). Ομοίως, το σύστημα που δημιουργεί δεδομένα πεδίων στοιχείων γράφει ένα αντίγραφο των δεδομένων στο HDFS, εκτός από την ενημέρωση του αντίστοιχου ευρετηρίου αναζήτησης στο διαδίκτυο. Τα πεδία χρήστη μπορεί να περιλαμβάνουν τμήματα του προφίλ LinkedIn ενός χρήστη, όπως τον τρέχοντα τίτλο εργασίας του χρήστη, τις δεξιότητες και τους βασικούς όρους που εξάγονται από τα περιγραφικά πεδία όπως η τρέχουσα κατάσταση του χρήστη, η τρέχουσα περιγραφή εργασίας, και προηγούμενες περιγραφές θέσεων εργασίας. Στην περίπτωση της αίτησης σύστασης εργασίας, τα πεδία του αντικειμένου μπορεί να περιλαμβάνουν τον τίτλο εργασίας, τη λειτουργία, τις δεξιότητες που χρειάζονται και τους βασικούς όρους που εξάγονται από την περιγραφή της εργασίας.

Δημιουργία διανύσματος δυνατοτήτων Boolean συνδυαστικών ρητρών:

Αυτό το στοιχείο δημιουργεί το σύνολο δεδομένων που απαιτείται για τον αλγόριθμο μηχανικής μάθησης, με βάση τις ακόλουθες εισόδους: τα δεδομένα εκπαίδευσης, τις τιμές δυνατοτήτων Boolean και το σύνολο ρυθμίσεων των συνδυαστικών ρητρών. Κάθε σειρά αντιστοιχεί σε ένα ζευγάρι (χρήστης, στοιχείο). Οι στήλες αντιστοιχούν στις ρήτρες συνδυασμού των χαρακτηριστικών (χρήστης, στοιχείο), με την τελευταία στήλη να δείχνει την ετικέτα. Μια ρήτρα σύζευξης ορίζεται σε 1 (true) εάν και μόνο εάν ο χρήστης και το στοιχείο ταιριάζουν σε όλες τις δυνατότητες που υπάρχουν στην ενότητα.

Μοντέλο ερωτήματος Learning WAND:

Χρησιμοποιείται η πλατφόρμα μηχανικής εκμάθησης του LinkedIn για να μάθουμε τα βάρη και να καθορίσουμε το βέλτιστο όριο για το υποψήφιο μοντέλο επιλογής που εκφράζεται ως ερώτημα Boolean WAND Δεδομένου ότι τα βάρη που σχετίζονται με τις ρήτρες πρέπει να είναι θετικά στο ερώτημα WAND, δεν μπορούν να χρησιμοποιηθούν απευθείας τεχνικές μηχανικής μάθησης. Γι᾽αυτό θα χρησιμοποιηθεί η λογιστική παλινδρόμηση με θετικούς συντελεστές στους περιορισμούς.

Αλγόριθμος επιλογής περιορισμένης λειτουργίας:

Ο αλγόριθμος 2 εκπαιδεύει πρώτα το μοντέλο λογιστικής παλινδρόμησης με όλες τις λειτουργίες και, στη συνέχεια, κόβει επαναλαμβανόμενα χαρακτηριστικά με συντελεστή βάρους κάτω από ένα μικρό θετικό όριο. Έτσι, και τα δύο χαρακτηριστικά με βάρη αρνητικού συντελεστή και πολύ μικρά βάρη αφαιρούνται. Αυτή η διαδικασία τερματίζεται με την επίτευξη του επιθυμητού αριθμού ρητρών σύνδεσης.

Μη αρνητικός αλγόριθμος περιορισμένου συντελεστή περιορισμού:

Ο αλγόριθμος 3 πραγματοποιεί τροποποίηση στον αλγόριθμο καθόδου κλίσης, έτσι ώστε οι αρνητικοί συντελεστές να τίθενται στο μηδέν μετά από κάθε βήμα καθόδου κλίσης. Αυτή η διαδικασία επαναλαμβάνεται έως ότου είτε η διαδικασία εκπαίδευσης συγκλίνει είτε, ο αριθμός των επαναλήψεων φτάσει ένα όριο.

Παρατηρήθηκε ότι ο Αλγόριθμος 3 είχε παρόμοια απόδοση με τον Αλγόριθμο 2 όσον αφορά τις ποιοτικές μετρήσεις. Ως εκ τούτου, λαμβάνοντας υπόψη την πρόσθετη πολυπλοκότητα που σχετίζεται με την τροποποίηση των εσωτερικών της διαδικασίας εκπαίδευσης, χρησιμοποιήθηκε ο Αλγόριθμος 2 στο σύστημα. Χρησιμοποιήθηκε η καμπύλη ακριβείας ανάκλησης (PR) για να επιλέξει την παράμετρο κατωφλίου για το ερώτημα WAND. Η καμπύλη PR μπορεί να κατασκευαστεί μετρώντας την ακρίβεια και την ανάκληση σε σχέση με διαφορετικές επιλογές του κατωφλίου.

Algorithm 2 Constrained Feature Selection Algorithm

Input: Dataset comprising Boolean feature vector of conjunction clauses, along with labels.
Output: Candidate selection query model, specified as WAND query predicate.

1: Train the logistic regression model with all features (Boolean conjunction clauses).
2: **repeat**
3: Remove features whose coefficients are below a small (positive) threshold.
4: Retrain the model without the removed features.
5: **until** Desired number of conjunction clauses are left.

Εικ. 37 Αλγόριθμος Επιλογής Περιορισμένης

Algorithm 3 Non-negative Coefficient Constrained Boundary Algorithm

Input: Dataset comprising Boolean feature vector of conjunction clauses, along with labels.
Output: Candidate selection query model, specified as WAND query predicate.

1: **repeat**
2: Perform a step of the gradient descent algorithm, and update the coefficients of the features (Boolean conjunction clauses).
3: Assign negative coefficients to zero.
4: **until** The training procedure converges, or the number of iterations reaches a limit.

Εικ. 38 Μη αρνητικός αλγόριθμος περιορισμένου συντελεστή

Application ranking candidates

Πρόκειται για μια εφαρμογή εποπτευόμενης μάθησης αλγορίθμων σε e-recruitment συστήματα για την κατάταξη των υποψηφίων. Έχει εφαρμοστεί ένα ολοκληρωμένο σύστημα ηλεκτρονικών προσλήψεων προσανατολισμένο στην εταιρεία που αυτοματοποιεί την υποψήφια διαδικασία προελέγχου και κατάταξης. Στο προτεινόμενο σύστημα [29], η αξιολόγηση των αιτούντων βασίζεται σε ένα προκαθορισμένο σύνολο αντικειμενικών κριτηρίων, τα οποία εξάγονται απευθείας από το προφίλ LinkedIn του αιτούντος. Επιπλέον, τα χαρακτηριστικά προσωπικότητας του υποψηφίου, τα οποία εξάγονται αυτόματα από τα social media, λαμβάνονται υπόψη στην αξιολόγησή του. Στόχος είναι να περιοριστεί η διαδικασία της συνέντευξης, της έρευνας ιστορικού των αιτούντων και να γίνεται αποκλειστικά στους κορυφαίους υποψηφίους που προσδιορίζονται από το σύστημα, έτσι ώστε να αυξηθεί η αποτελεσματικότητα της διαδικασίας πρόσληψης. Το σύστημα έχει σχεδιαστεί με σκοπό να ενσωματωθεί στην υποδομή διαχείρισης ανθρώπινου δυναμικού των εταιρειών, βοηθώντας και όχι αντικαθιστώντας τους ειδικούς των προσλήψεων (recruiters) στη διαδικασία λήψης αποφάσεων.

Αρχιτεκτονική συστήματος

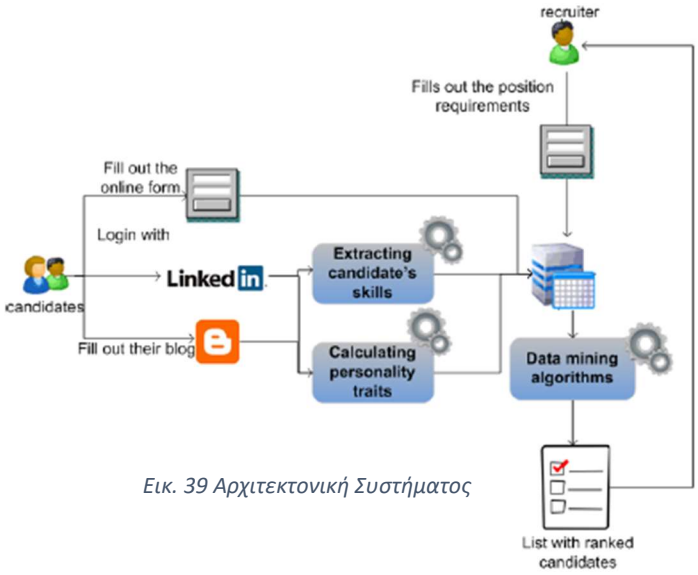
Το προτεινόμενο σύστημα ηλεκτρονικής πρόσληψης εφαρμόζει την αυτοματοποιημένη κατάταξη υποψηφίων με βάση ένα σύνολο αξιόπιστων κριτηρίων, τα οποία είναι εύκολα για τις εταιρείες να ενσωματωθούν στην υπάρχουσα υποδομή της διαχείρισης ανθρώπινου δυναμικού. Εστιάζει επίσης σε 4 συμπληρωματικά κριτήρια επιλογής: Εκπαίδευση (σε έτη επίσημης ακαδημαϊκής κατάρτισης), Εργασιακή Εμπειρία, Πίστη (μέσος αριθμός ετών που εργάζεται ανά εργασία) και Εξωστρέφεια.

Η αρχιτεκτονική του συστήματος, αποτελείται από τα ακόλουθα στοιχεία:

1. Ενότητα αίτησης εργασίας: Υλοποιεί τις φόρμες εισαγωγής που επιτρέπουν στους υποψηφίους να υποβάλουν αίτηση για θέση εργασίας. Στον υποψήφιο δίνεται η δυνατότητα να συνδεθεί στο σύστημά δίνοντας τα διαπιστευτήρια του λογαριασμού του στο LinkedIn, το οποίο επιτρέπει στο σύστημα να εξαγάγει αυτόματα όλα τα αντικειμενικά κριτήρια επιλογής απευθείας από το προφίλ του.
2. Ενότητα εξόρυξης στοιχείων προσωπικότητας: Εάν παρέχεται το URL του ιστολογίου του υποψηφίου, εφαρμόζει γλωσσική ανάλυση στις αναρτήσεις του ιστολογίου για να αντλήσει χαρακτηριστικά που αντικατοπτρίζουν την προσωπικότητα του.
3. Ενότητα βαθμολόγησης υποψηφίου: Συνδυάζει τα κριτήρια επιλογής του υποψηφίου για να αποκομίσει τη βαθμολογία συνάφειας για την υποψήφια θέση. Η συνάρτηση βαθμολόγησης παράγεται μέσω εποπτευόμενων αλγορίθμων μάθησης

Τα προσόντα κάθε αιτούντος, καθώς και η βαθμολογία του, αποθηκεύονται στη βάση δεδομένων του συστήματος. Στο τέλος της διαδικασίας πρόσληψης, οι κορυφαίοι υποψήφιοι καλούνται να συμμετάσχουν στη διαδικασία της συνέντευξης.

Κατά τη διάρκεια της διαδικασίας αίτησης, ο αιτών δεν απαιτείται να εισάγει με μη αυτόματο τρόπο πληροφορίες ή να συμμετάσχει σε χρονοβόρες δοκιμές προσωπικότητας. Έτσι, διατηρείται η φιλικότητα προς το χρήστη και η πρακτικότητα του συστήματος.



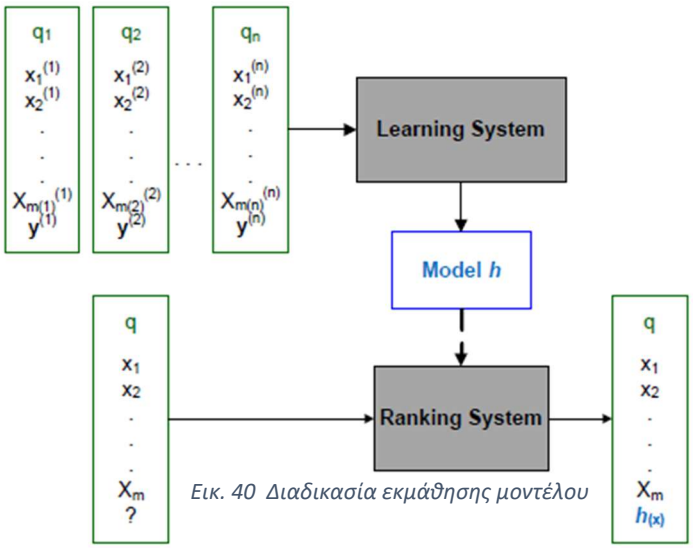
Εικ. 39 Αρχιτεκτονική Συστήματος

Κατάταξη υποψηφίων

Τα αυτοματοποιημένα συστήματα κατάταξης υποψηφίων, τα οποία έχουν προταθεί για την επιτάχυνση της διαδικασίας πρόσληψης ομοιάζουν της διαδικασίας λήψης αποφάσεων του τμήματος Ανθρώπινου Δυναμικού, με μια πιο προσεκτική παραμετροποίηση από τους ειδικούς των προσλήψεων του τμήματος, που είναι πολύπλοκη και επιρρεπής σε σφάλματα και πρέπει να επαναλαμβάνεται κάθε φορά που αλλάζουν τα κριτήρια επιλογής. Το σύστημα αυτό αξιοποιεί τους αλγόριθμους μηχανικής μάθησης για να δημιουργήσει αυτόματα τα μοντέλα κατάταξης των υποψηφίων. Αυτή η προσέγγιση απαιτεί επαρκή δεδομένα εκπαίδευσης ως εισαγωγή. Οι μέθοδοι που μαθαίνουν πώς να συνδυάζουν προκαθορισμένα χαρακτηριστικά για την κατάταξη μέσω εποπτευόμενων αλγορίθμων μάθησης ονομάζονται μέθοδοι «μάθηση-κατάταξη».

Στην παρακάτω εικόνα εμφανίζεται η τυπική διαδικασία «εκμάθηση-κατάταξης». Χρησιμοποιείται ένα training set που αποτελείται από προηγούμενες εφαρμογές,

αναπαριστάται από διανύσματα χαρακτηριστικών και δηλώνονται ως xi (k), μαζί με την κρίση ενός εμπειρογνώμονα για τη βαθμολογία συνάφειας του υποψηφίου, που δηλώνεται ως yi. Οι δυνατότητες του υποψηφίου μπορούν να αξιολογηθούν είτε σε αριθμητική κλίμακα (π.χ. έτη επαγγελματικής εμπειρίας) είτε με μεταβλητή Boolean, η οποία δείχνει εάν ο υποψήφιος αναφέρει μια συγκεκριμένη ικανότητα ή όχι στο προφίλ του. Το training set τροφοδοτείται σε έναν αλγόριθμο μάθησης που κατασκευάζει το μοντέλο κατάταξης, έτσι ώστε η έξοδος του να προβλέπει την κρίση του υπευθύνου πρόσληψης όταν δίνεται το διάνυσμα χαρακτηριστικών του υποψηφίου ως είσοδος.



Εικ. 40 Διαδικασία εκμάθησης μοντέλου

Εξόρυξη χαρακτηριστικών προσωπικότητας

Τα χαρακτηριστικά της προσωπικότητας των αιτούντων είναι κρίσιμα για την επιλογή τους σε πολλές θέσεις εργασίας, αλλά συνήθως δεν προσμετρώνται στα υπάρχοντα συστήματα ηλεκτρονικής πρόσληψης. Συνήθως, η προσωπικότητα των υποψηφίων αξιολογείται κατά τη διάρκεια της φάσης της συνέντευξης, η οποία προορίζεται για τους υποψηφίους που πέρασαν τη φάση πριν από τον έλεγχο. Ωστόσο, η συλλογή ορισμένων προκαταρκτικών δεδομένων για την προσωπικότητα του υποψηφίου στη φάση πριν από τον έλεγχο θεωρείται πολύτιμη και αυτές οι πληροφορίες λαμβάνονται συχνά μέσω αναζητήσεων στον ιστό. Στην εποχή του Web 2.0, υπάρχουν μεγάλα ποσά δεδομένων κειμένου για εκατομμύρια χρήστες ιστού, τα οποία έχουν αποδειχθεί αξιόπιστα προγνωστικά της προσωπικότητας του χρήστη. Το σύστημα αυτό αυτοματοποιεί το έργο της εξόρυξης της προσωπικότητας χρησιμοποιώντας ανάλυση κειμένου. Η ανάλυση κειμένου σε αυτά τα έργα πραγματοποιείται με το σύστημα LIWC (LinguisticInquiry and Word Count), το οποίο εξάγει γλωσσικά χαρακτηριστικά που λειτουργούν ως δείκτες της

προσωπικότητας του συγγραφέα. Το LIWC χρησιμοποιεί ένα λεξικό λέξεων που ταξινομούνται σε ορισμένες ψυχογλωσσολογικές σημασιολογικές και συντακτικές κατηγορίες λέξεων. Αναλύει γραπτά δείγματα κειμένου μετρώντας τις σχετικές συχνότητες των λέξεων που εμπίπτουν σε κάθε κατηγορία λέξεων. Επίσης, εστιάζει στο χαρακτηριστικό της προσωπικότητας εξωστρέφειας, λόγω της σημασίας του στην επιλογή υποψηφίων. Η εξωστρέφεια είναι ένα κρίσιμο χαρακτηριστικό της προσωπικότητας σε θέσεις που αλληλεπιδρούν με τους πελάτες, ενώ οι κοινωνικές δεξιότητες είναι σημαντικές για την ομαδική εργασία.

Κατά την εκτέλεση του πειράματος 100 αιτήσεων, βαθμολογήθηκε καθένας από αυτούς με προσωπικά ιστολόγια, τα οποία ήταν μέρος ενός σεναρίου πρόσληψης μεγάλης κλίμακας. Οι βαθμολογίες του recruiter χρησιμοποιήθηκαν για να εκπαιδεύσουν ένα μοντέλο παλινδρόμησης, το οποίο προβλέπει την υπερβολή των υποψηφίων από τις βαθμολογίες LIWC στις κατηγορίες {posemo, negemo, social}. Στη συνέχεια, επιλέχθηκε ένα μοντέλο γραμμικής παλινδρόμησης ως προγνωστικός δείκτης της βαθμολογίας εξωστρέφειας E, όπως προτείνεται, λόγω της καλής ακρίβειας και της χαμηλής πολυπλοκότητάς του. Η εξίσωση παρακάτω αντιστοιχεί στο γραμμικό μοντέλο που ελαχιστοποιεί το Σφάλμα μέσου τετραγώνου μεταξύ των πραγματικών τιμών που έχουν εκχωρηθεί από τον recruiter και των προβλεπόμενων αποτελεσμάτων από το μοντέλο:

$$E = S + 1.335 * P - 2.250 * N$$

όπου S είναι η συχνότητα των κοινωνικών λέξεων (όπως φίλος, φίλος, συνάδελφος) που δόθηκε από το LIWC, P η συχνότητα των θετικών συναισθημάτων και N η συχνότητα των αρνητικών συναισθημάτων.

Learning-rankΑλγόριθμοι

Καθώς η βαθμολογία συνάφειας είναι μια συνεχής μεταβλητή, το πρόβλημα κατάταξης των υποψηφίων μπορεί να μειωθεί σε ένα πρόβλημα παλινδρόμησης όπου η συνάρτηση βαθμολογίας υποψηφίων πρέπει να «μάθει» χρησιμοποιώντας εποπτευόμενες τεχνικές μάθησης για να εξάγει την τελική λίστα κατάταξης. Η συνάρτηση βαθμολογίας h (x) αντλεί το βαθμό συνάφειας του υποψηφίου yi από τις τιμές του διανύσματος χαρακτηριστικών του xi (τα xi είναι ένα σύνολο χαρακτηριστικών m {a1,…, am} που αντιστοιχούν στα κριτήρια επιλογής του υποψηφίου),και διακρίνονται σε συνεχή (χαρακτηριστικά ενός υποψηφίου βαθμολογημένα σε αριθμητική κλίμακα) ή Boolean μεταβλητές (έχει ή όχι την ικανότητα). Η πραγματική συνάρτηση βαθμολόγησης είναι συνήθως άγνωστη και μια προσέγγιση μαθαίνεται από το training set D (στο σύστημα το training set αποτελείται από ένα σύνολο N παραδειγμάτων προηγούμενης επιλογής υποψηφίων, που δίδονται ως είσοδος στο σύστημα).

$$D = \left\{ (x_i, y_i) \mid x_i \in R^m, y_i \in R \right\}_{i=1}^{N}.$$

Συλλογή Δεδομένων

Στο πείραμα που διενεργήθηκε τα δεδομένα που χρησιμοποιήθηκαν συνολικά ήταν 100 βιογραφικά σημειώματα με λογαριασμό LinkedIn και δεδομένα από προσωπικό ιστολόγιο, όπως ακριβώς απαιτούνται από το σύστημα. Οι αιτούντες επιλέχθηκαν τυχαία μέσω της GoogleAPI αναζήτησης ιστολογίου με τη μοναδική απαίτηση να έχουν ένα τεχνικό υπόβαθρο, όπως υποδεικνύεται από τα μεταδεδομένα του ιστολογίου (λίστα ενδιαφέροντος), καθώς και του προφίλ LinkedIn. Από πλευράς θέσεων εργασίας, τρεις εκπρόσωποι τεχνικών θέσεων ανακοινώθηκαν από μια ανώνυμη εταιρεία πληροφορικής με διαφορετικές απαιτήσεις (μια τεχνική πωλήσεων θέση, κατώτερη θέση προγραμματιστή και μία υψηλόβαθμη θέση προγραμματιστή). Η θέση μηχανικού πωλήσεων ευνοεί υψηλό βαθμό εξωστρέφειας, ενώ η εμπειρία είναι το πιο σημαντικό χαρακτηριστικό για ανώτερους προγραμματιστές. Οι προγραμματιστές Junior κυρίως κρίνοται από την πίστη (επειδή μια εταιρεία δεν θα επενδύσει σε εκπαίδευση ενός ατόμου που είναι επιρρεπές σε μεταβαλλόμενες θέσεις συχνά) καθώς και εκπαίδευση. Επιπλέον, κάθε θέση έχει το δικό του επιθυμητό σύνολο δεξιοτήτων, οι οποίες συνδυάζονται με το skillet που αναφέρεται από κάθε χρήστη στο προφίλ του στο LinkedIn. Συγκεκριμένα, η κατώτερη θέση απαιτεί δεξιότητες προγραμματισμού σε γλώσσες ανάπτυξης C ++ ή Java, ενώ οι υψηλόβαθμη θέση απαιτεί πενταετή εμπειρία στις τεχνολογίες J2EE.

Πειραματικά αποτελέσματα

Με βάση τα παραπάνω, ο κάθε υποψήφιος έχει υποβάλει αίτηση και για τις τρεις διαθέσιμες θέσεις εργασίας. Για κάθε θέση εργασίας, οι αιτούντες κατατάχθηκαν σύμφωνα με την καταλληλότητά τους για τη θέση εργασίας τόσο από το σύστημα (αυτοματοποιημένη κατάταξη) όσο και από έναν recruiter. Οι άνθρωποι που προσλήφθηκαν είχαν πρόσβαση στις ίδιες πληροφορίες με το σύστημα, δηλαδή στο ιστολόγιο του υποψηφίου και στο προφίλ LinkedIn.

Τα κριτήρια επιλογής είναι γνωστά στο σύστημα, η ερμηνεία των δεδομένων και η ακριβής διαδικασία λήψης αποφάσεων άγνωστη. Στο πρώτο πείραμα, χρησιμοποιείται το Weka για να αξιολογηθούν τα μοντέλα μάθησης προς κατάταξη. Συγκεκριμένα, δοκιμάζεται η συσχέτιση των αποτελεσμάτων από το σύστημα (δηλαδή, προβλέψεις μοντέλου) με τις πραγματικές βαθμολογίες που έχουν εκχωρηθεί από τους recruiters, χρησιμοποιώντας για τη μέτρηση συντελεστή συσχέτισης Pearson. Ο πίνακας 11 δείχνει τους συντελεστές συσχέτισης για 4 διαφορετικά μοντέλα μηχανικής εκμάθησης, με δύο μη γραμμικούς πυρήνες.

Όπως είναι φανερό, η συνοχή των βαθμολογιών του συστήματος εξαρτάται σε μεγάλο βαθμό από τη φύση των προσφερόμενων θέσεων. Για τη θέση πωλήσεων, η κρίση του recruiter κυριαρχείται από την εξαιρετικά υποκειμενική βαθμολογία αυξάνοντας έτσι την αβεβαιότητα της συνολικής βαθμολογίας συνάφειας. Ωστόσο, το

σύστημα μπόρεσε να επιτύχει έναν συντελεστή συσχέτισης έως 0,81, ανάλογα με το μοντέλο παλινδρόμησης που χρησιμοποιήθηκε. Από την άλλη πλευρά, η επιλογή των υποψηφίων κατώτερου προγραμματιστή βασίζεται σε πιο αντικειμενικά κριτήρια, όπως η πίστη και η εκπαίδευση, με αποτέλεσμα έναν ελαφρώς υψηλότερο συντελεστή συσχέτισης, έως 0,85. Τέλος, η θέση του ανώτερου προγραμματιστή παρουσίασε τη χαμηλότερη συνοχή, με συσχέτιση του Pearson έως 0,73. Αυτό μπορεί να αποδοθεί στην υψηλή πολυπλοκότητα της δημιουργίας ενός μοντέλου παλινδρόμησης για μια ανώτερη θέση, το οποίο συνήθως απαιτεί εμπειρία για συγκεκριμένο τομέα και συγκεκριμένα προσόντα.

Correlation coefficient	LR	M5' Tree	REP Tree	SVR, poly	SVR, PUK
Sales engineer	0.74	0.81	0.81	0.61	0.81
Junior programmer	0.79	0.85	0.84	0.81	0.84
Senior programmer	0.64	0.63	0.68	0.62	0.73

Πίνακας 11 Συντελεστές συσχέτισης διαφορετικών μοντέλων μηχανικής μάθησης

Το σκορ εξωστρέφειας προβλέπεται να εκπαιδεύσει ένα μοντέλο παλινδρόμησης στις βαθμολογίες εξωστρέφειας που αποδίδονται από τον recruiter σε καθέναν από τους 100 υποψηφίους. Ο πίνακας 12, δείχνει τους συντελεστές συσχέτισης Pearson και σχετικά σφάλματα μεταξύ των βαθμολογιών του συστήματος και του recruiter.

Correlation coefficient	LR	M5' Tree	REP Tree	SVR, poly	SVR, PUK
Pearson's Coefficient	0.63	0.63	0.65	0.28	0.65
Relative error	25.3%	25.3%	22.5%	57.4%	23.1%

Πίνακας 12 Συντελεστές συσχέτισης

Εν κατακλείδι, αυτό το σύστημα ηλεκτρονικής πρόσληψης εφαρμόστηκε πλήρως ως εφαρμογή διαδικτύου, στο περιβάλλον ανάπτυξης Microsoft .Net.

- Διαδικασία αίτησης εργασίας (από πλευράς χρήστη)

Οι υποψήφιοι έχουν τη δυνατότητα πρόσβασης χρησιμοποιώντας τα διαπιστευτήρια του λογαριασμού τους στο LinkedIn για να υποβάλουν αίτηση για μία ή περισσότερες από τις διαθέσιμες θέσεις εργασίας. Αυτό επιτρέπει στο σύστημα να εξαγάγει αυτόματα τα κριτήρια επιλογής που απαιτούνται για μια πρώτη επιλογή των υποψηφίων, έτσι ώστε η εμπειρία του χρήστη να βελτιστοποιηθεί. Μετά την επιτυχή πιστοποίηση χρήστη, ένα διακριτικό OAuth (πρωτόκολλο ελέγχου ταυτότητας) επιστρέφεται στο σύστημά, το οποίο επιτρέπει την ανάκτηση πληροφοριών από το ιδιωτικό προφίλ LinkedIn του υποψηφίου. Οι χρήστες χωρίς προφίλ LinkedIn έχουν τη δυνατότητα να εισάγουν τις απαιτούμενες πληροφορίες με μη αυτόματο τρόπο.
Στο πλαίσιο της διαδικασίας αίτησης εργασίας, ζητείται από τον υποψήφιο να συμπληρώσει το URI τροφοδοσίας του προσωπικού του ιστολογίου. Αυτό επιτρέπει στο σύστημά να συνδιοργανώνει το περιεχόμενο του ιστολογίου και να υπολογίσει τη βαθμολογία εξωστρέφειας με την τεχνική εξόρυξης προσωπικότητας. Οι αναρτήσεις ιστολογίου εισάγονται στο εργαλείο TreeTagger για λεξική ανάλυση. Στη συνέχεια, χρησιμοποιώντας το λεξικό LIWC, το σύστημά ταξινομεί την κανονική μορφή λέξεων που εξέρχονται από το TreeTagger σε μία από τις κατηγορίες ενδιαφερόντων λέξεων (δηλ. Θετικό συναίσθημα, αρνητικό συναίσθημα και κοινωνικές λέξεις) και υπολογίζει το LIWC αποτελέσματα. Τέλος, το σύστημα εκτιμά τη βαθμολογία υπερβολής του αιτούντος.

- Διαδικασία πρόσληψης (πλευρά προσλήψεων)

Μετά τον έλεγχο ταυτότητας με τα διαπιστευτήρια του λογαριασμού τους, οι recruiters έχουν πρόσβαση στη λειτουργική μονάδα προσλήψεων, η οποία τους δίνει δικαιώματα να δημοσιεύουν νέες θέσεις εργασίας και να αξιολογούν τους αιτούντες. Στο μενού «κατάταξη υποψηφίων», παρουσιάζεται στον εκάστοτε recruiter μια λίστα με όλες τις διαθέσιμες θέσεις εργασίας και τους υποψηφίους που έχουν υποβάλει αίτηση για καθεμία από αυτές. Κατόπιν αιτήματος του recruiter, το σύστημα εκτιμά τις βαθμολογίες συνάφειας των αιτούντων και τις κατατάσσει ανάλογα. Αυτό επιτυγχάνεται καλώντας τον αντίστοιχο ταξινομητή Weka, μέσω κλήσεων στο API. Ο recruiter μπορεί να τροποποιήσει την κατάταξη των υποψηφίων, αναθέτοντας τις δικές του βαθμολογίες συνάφειας στους υποψηφίους βελτιώνοντας έτσι τη μελλοντική απόδοση του συστήματος, καθώς οι προτάσεις του recruiter ενσωματώνονται στο εκπαιδευτικό σύνολο του συστήματος και έτσι το μοντέλο κατάταξης ενημερώνεται.

Επίλογος

Η ταχεία ανάπτυξη του Διαδικτύου προκάλεσε αντίστοιχη αύξηση του αριθμού των διαθέσιμων διαδικτυακών πληροφοριών που αύξησαν με την σειρά τους την ανάγκη επέκτασης της ικανότητας των χρηστών να διαχειρίζονται όλες αυτές τις πληροφορίες. Αυτό κεντρίζει το ενδιαφέρον ορισμένων ερευνητικών τομών ώστε να βρεθεί ένα σύστημα τέτοιο που θα διαχειρίζεται αυτή την υπερφόρτωση πληροφοριών.

Η διαδικτυακή (ηλεκτρονική) πλατφόρμα προσλήψεων είναι μια από τις πιο επιτυχημένες επιχειρηματικές αλλαγές, οι οποίες άλλαξαν τον τρόπο με τον οποίο οι εταιρείες προσλαμβάνουν υποψηφίους. Αυτές οι πλατφόρμες εξαπλώθηκαν τα τελευταία χρόνια επειδή η πρόσληψη του κατάλληλου ατόμου είναι μια πρόκληση που αντιμετωπίζουν οι περισσότερες επιχειρήσεις, καθώς η μη διαθεσιμότητα ορισμένων υποψηφίων σε ορισμένους τομείς δεξιοτήτων έχει αναγνωριστεί από καιρό ως σημαντικό εμπόδιο στην επιτυχημένη εύρεση υποψηφίων.

Για κάθε θέση εργασίας, χιλιάδες βιογραφικά παραλαμβάνονται από εταιρείες. Κατά συνέπεια, ένας τεράστιος όγκος θέσεων εργασίας και βιογραφικών υποψήφιων διατίθεται στο Διαδίκτυο. Αυτός ο τεράστιος όγκος πληροφοριών δίνει μια μεγάλη ευκαιρία για ενίσχυση της ποιότητας που ταιριάζει, αυξάνοντας την ανάγκη για δημιουργία και εφαρμογή τεχνολογιών recommender systems που μπορούν να βοηθήσουν στην διαχείριση αυτών των πληροφοριών αποτελεσματικά.

Επιγραμματικά, θετικό στοιχείο αυτών είναι η συμπλήρωση δωρεάν φόρμας καθώς αποτελεί ευέλικτη διαδικασία πρόσληψης σε σύγκριση με τη διαδικασία φυσικής πρόσληψης, συμβάλλουν σημαντικά στην μείωση κόστους διαφήμισης και μάρκετινγκ πολύ περισσότερο δε αφού δεν περιλαμβάνει μεσάζοντες ενώ μειώνεται ο και χρόνος ως προς την πρόσληψη του κατάλληλου υποψηφίου και η διαδικασία πρόσληψης γίνεται πιο αποτελεσματική και εύκολη στην καταγραφή των στοιχείων του αιτούντος. Επίσης, ο αιτών μπορεί να υποβάλει αίτηση σε οποιοδήποτε μέρος του κόσμου και να εξερευνήσει όλες τις διαθέσιμες κενές θέσεις εργασίας ανά πάσα στιγμή ενώ τόσο η εταιρεία προσλήψεων όσο και ο υποψήφιος για εργασία, θα έχουν το πλεονέκτημα να συναντηθούν μέσω Διαδικτύου και μέσω της πύλης εργασίας που αναπτύχθηκε.

Στον αντίποδα, η ανάκτηση δεδομένων στην τρέχουσα διαδικασία ηλεκτρονικής πρόσληψης βασίζεται στην ακριβή αντιστοίχιση της εργασίας και στα στοιχεία του αιτούντος ενώ οι εταιρείες δεν μπορούν να βασίζονται πλήρως στο Διαδίκτυο για την πρόσληψη κατάλληλων υποψηφίων καθώς η συνάντηση προσωπικά παραμένει μια σημαντική πτυχή για να γνωρίζει και να κατανοεί κάποιος τον υποψήφιο προς εργασία.

Παρόλα αυτά, έχουν διεξαχθεί πολλές έρευνες, το πεδίο είναι ακόμη πρόσφορο για περαιτέρω διερεύνηση αφού η σύσταση εργασίας εξακολουθεί να

είναι ένας απαιτητικός και αναπτυσσόμενος τομέας έρευνας ενώ η επιτυχία των τεχνολογιών εξατομίκευσης εξαρτάται σε μεγάλο βαθμό από την ύπαρξη ολοκληρωμένων προφίλ χρηστών που αποτυπώνουν ακριβώς τα ενδιαφέροντα των χρηστών και την τέλεια μέθοδο αντιστοίχισης.

Αναφορές

- [1] Μαρία Καραμεσίνη, Πάντειο Πανεπιστήμιο, 2006: *«Κοινωνική Συνοχή και Ανάπτυξη - Από την εκπαίδευση στην αμειβόμενη εργασία: Εμπειρική διερεύνηση της εργασιακής ένταξης των νέων στην Ελλάδα»*.
- [2] Christian Bizer, Ralf Heese, Malgorzata Mochol, Radoslaw Oldakowski, Robert Tolksdorf, Rainer Eckstein, 2005: *«The impact of semantic web technologies on job recruitment processes»*.
- [3] Simona Colucci, Tommaso Di Noia, Eugenio Di Sciascio, Francesco M. Donini, Marina Mongiello, Marco Mottola, Journal a/Universal Computer Science, 2003: *«A Formal Approach to Ontology-Based Semantic Match of Skills Descriptions»* .
- [4] Mihaela-Irina ENĂCHESCU, Bucharest University of Economic Studies, 2016: *«A Prototype for an e-Recruitment Platform using Semantic Web Technologies»*.
- [5] Aseel B. Kmail, Mohammed Maree, Mohammed Belkhatir, Saadat M. Alhashmi, IEEE 27th International Conference on Tools with Artificial Intelligence, 2015: *«An Automatic Online Recruitment System based on Exploiting Multiple Semantic Resources and Concept-relatedness Measures»*.
- [6] Elizabeth D. Liddy, Syracuse University, 2001: *«Natural language processing»*.
- [7] Ανακτήθηκε από: https://spacy.io/usage/models
- [8] Ανακτήθηκε από: https://alwaysbelearning.nl/matching-resumes-with-job-offers-using-spacy-a-natural-language-processing-nlp-library-in-python
- [9] Xing Yi, James Allan, W. Bruce Croft,University of Massachusetts, 2007: *«Matching Resumes and Jobs Based on Relevance Models»*.
- [10] Hamed Zamani, W. Bruce Croft, University of Massachusetts, 2017: *«Relevance-based Word Embedding»*.
- [11] Ανακτήθηκε από: https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html
- [12] Ανακτήθηκε από: https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-skip-gram.html
- [13] Ανακτήθηκε από: https://www.kdnuggets.com/2018/08/word-vectors-nlp-glove.html
- [14] Mohammed Maree, Aseel B. Kmail, Mohammed Belkhatir, Journal of Information Science, 2018: *«Analysis & Shortcomings of E-Recruitment Systems: Towards a Semantics-based Approach Addressing Knowledge Incompleteness and Limited Domain Coverage»*.

- [15] Hina H Soni, Dr. Priya R Swaminarayan, International Journal of Advanced Research in Computer Science, 2017: *«Study of Semantic Web Based E-Recruitment System: Review»*.
- [16] Nabeel Ahmed, Sharifullah Khan, Khalid Latif, International Conference on Frontiers of Information Technology, 2016: *«Job Description Ontology»*.
- [17] Yao Lu, Sandy El Helou, Denis Gillet, International World Wide Web Conference Committee, 2013: *«A recommender system for job seeking and recruiting website»*.
- [18] In Lee, Communications of the Acm, 2007: *«An architecture for a next-generation holistic e-recruiting system»*.
- [19] V. Senthil Kumaran, A. Sankar, Int. J. Metadata, Semantics and Ontologies, 2013: *«Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT)»*.
- [20] V. Gatteschi, F. Lamberti, A. Sanna, C. Demartini, 9th IEEE International Conference on Emerging eLearning Technologies and Applications, 2011: *«Using Tag Clouds to Support the Comparison of Qualifications, Résumés and Job Profiles»*.
- [21] Faizan Javed, Phuong Hoang, Thomas Mahoney, Matt McNair, Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications, 2017: *«Large-Scale Occupational Skills Normalization for Online Recruitment»*.
- [22] Fuad Mire Hassan, Imran Ghani, Muhammad Faheem, Abdirahman Ali Hajji, International Journal of Computer Applications, 2012: *«Ontology matching approaches for e-recruitment»*.
- [23] Fedor Borisyuk, Krishnaram Kenthapadi, David Stein, Bo Zhao, KDD'16, 2016: *«Casmos_A framework for learning candidate selection models over structured queries and document»*.
- [24] Rémy Kessler, Juan Manuel, Torres-Moreno, Marc El-Bèze, A. Gelbukh, A.F. Kuri Morales, MICAI, 2007: *«E-Gen Automatic Job Offer Processing System for Human Resources»*.
- [25] R. Vedapradha, Ravi Hariharan , Rajan Shivakami, Journal of Service Science and Management, 2019: *«Artificial Intelligence A Technological Prototype in Recruitment»*.
- [26] Fred Gulliford, Amy Parker Dixon, Emerald Publishing Limited, 2019: *«AI the HR revolution»*.
- [27] Luis Adrián, Cabrera-Diego, Marc El-Bèze, Juan-Manuel Torres-Moreno, Barthélémy Durette, Elsevier Ltd, 2019: *«Ranking résumés automatically using only résumés A method free of job offers»*.

- [28] Wan Mohd Rusydan, Wan Ibrahim, Roshidi Hassan, Asian Journal of Research in Business and Management, 2019: *«Recruitment trends in the era of Industry 4.0 using Artificial Intelligence: PRO AND CONS»*.
- [29] Evanthia Faliagka, Lazaros Iliadis, IoannisKarydis, Maria Rigou, Spyros Sioutas, Athanasios Tsakalidis, Giannis Tzimas, Springer Science+Business Media Dordrecht, 2014: *«Online consistent ranking on e-recruitment seeking the truth behind a well formed cv»*.
- [30] Shaha T. Al-Otaibi1, Mourad Ykhlef, International Journal of the Physical Sciences, 2012: *«A survey of job recommender systems»*.
- [31] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, Giannis Tzimas, The Seventh International Conference on Internet and Web Applications and Services, 2012: *«Application of machine learning algorithms to an online recruitment system»*.
- [32] N. Sivaram, K. Ramar, International Journal of Computer Applications, 2010: *«Applicability of clustering and classification algorithms for recruitment data mining»*.