Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων

ΠΜΣ Ψηφιακά Συστήματα και Υπηρεσίες

Κατεύθυνση: Δικτυοκεντρικά Πληροφοριακά Συστήματα

«Συναισθηματική Ανάλυση Δεδομένων Κοινωνικών Δικτύων με χρήση του ConceptNet και του SentiWordNet»

“Sentiment Analysis of Social-Network Data using ConceptNet and SentiWordNet”

Μεταπτυχιακή Φοιτήτρια: Μπρίλη Δέσποινα
Α.Μ.: ΜΕ13053

Επιβλέπων Καθηγητής: Δουλκερίδης Χ.

Πειραιάς, 2016

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής εργασίας μου, Επίκουρο Καθηγητή κ. Χρήστο Δουλκερίδη, για την πολύτιμη βοήθεια, την καθοδήγησή του, και την αμέριστη κατανόηση του κατά τη διάρκεια εκπόνησης της εργασίας μου. Επίσης, είμαι ευγνώμων στην μεταπτυχιακή φοιτήτρια κα. Μαρία Καρανάσου για τις κατευθυντήριες γραμμές που μου υπέδειξε, για τις σημαντικές συμβουλές της και για την βοήθεια που μου προσέφερε σε όσα προβλήματα αντιμετώπισα.

Μπρίλη Δέσποινα

Περίληψη

*Η συγκεκριμένη διπλωματική εργασία εκπονήθηκε από την μεταπτυχιακή φοιτήτρια Δέσποινα Μπρίλη (ΜΕ13053), στα πλαίσια της ολοκλήρωσης του Προγράμματος Μεταπτυχιακών Σπουδών «Ψηφιακά Συστήματα και Υπηρεσιες», με κατεύθυνση «Δικτυοκεντρικά Συστήματα», του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς. Υπεύθυνος επιβλέπων καθηγητής της εν λόγω εργασίας είναι ο επίκουρος καθηγητής κ. Χρήστος Δουλκερίδης. Ως θέμα της εργασίας τέθηκε ο τίτλος «*Συναισθηματική Ανάλυση Δεδομένων Κοινωνικών Δικτύων με χρήση του ConceptNet και του SentiWordNet*».*

Η εν λόγω διπλωματική εργασία αποτελείται από θεωρητικό και πρακτικό τμήμα. Σκοπός της είναι ο εμπλουτισμός του Conceptnet μέσω ενός sentiment lexicon και συγκεκριμένα του SentiWordnet. Στη συνέχεια, και αφού πραγματοποιηθεί η φάση της προεπεξεργασίας των δεδομένων, που έχουν συλλεχθεί από το Twitter, το κάθε tweet αξιολογείται με δεδομένο τόσο το SentiWordnet, όσο και το εμπλουτισμένο Conceptnet.

Τελικός στόχος της εργασίας είναι να αποδείξει, εάν ο συνδυασμός του Conceptnet με το SentiWordnet είναι πιο αποδοτικός από το απλό SentiWordnet, συγκριτικά με το Initial Score του αρχικού labelled dataset.

Abstract

This Master Thesis prepared by the post-graduate student Despoina Brilli (ME13053), in the context of the completion of the Master Program "Digital Systems and Services" in the area of "Network-Oriented Information Systems", at the department of Digital Systems, in University of Piraeus. Responsible supervisor of this work is the Assistant Professor. Christos Doulkeridis. As a matter of this work was the title " Sentiment Analysis of Social-Network Data using ConceptNet and SentiWordNet".

This thesis consists of a theoretical and a practical part. Its purpose is the enrichment of Conceptnet through a sentiment lexicon, which is named as SentiWordnet. Then, after realized the phase of preprocessing of data, which had been collected from Twitter platform later, each tweet of the corpus evaluated, due to SentiWordnet and enriched Conceptnet.

The final goal of this project is to demonstrate that the combination of Conceptnet and SentiWordnet is more efficient than the simple SentiWordnet, in comparison with the Initial Score original labelled dataset.

Περιεχόμενα

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

Κεφάλαιο 1: Εισαγωγή

Τα τελευταία χρόνια η ραγδαία ανάπτυξη του διαδικτύου δημιούργησε τεράστιες απαιτήσεις στην αποθήκευση, στη διανομή και στη διαχείριση των δεδομένων και έδωσε με τη σειρά της σημαντική ώθηση στα συστήματα επεξεργασίας και ανάλυσης των δεδομένων αυτών.

Μια σειρά από νέες ανερχόμενες τεχνολογίες φαίνεται να αλλάζουν τον τρόπο που δουλεύουμε σήμερα με τους υπολογιστές παρέχοντας στο ευρύ κοινό υπηρεσίες που μέχρι πρότινος ήταν διαθέσιμες σε μικρές ομάδες ατόμων. Οι υπηρεσίες αυτές δίνουν την δυνατότητα καταγραφής και ανάλυσης της πληροφορίας. Πως όμως μπορεί να πραγματοποιηθεί η καταγραφή και ή ανάλυση;

Ο βασικότερος τρόπος ανάλυσης και καταγραφής πληροφοριών είναι η ανθρώπινη γλώσσα. Ο ανθρώπινος φυσικός λόγος, ή όπως αλλιώς ονομάζεται φυσική γλώσσα (natural language) αποτελούσε τον βασικότερο κώδικα επικοινωνίας , ακόμα και πριν την χρήση της γραφής. Επιπλέον παρουσιάζεται σαν ένα από τα βασικά στοιχεία συνδέσμων μεταξύ των ατόμων μίας ομάδας. Η γλώσσα, είτε φυσική είτε γραπτή, διέπεται από κάποιους κανόνες. Κατά την πάροδο των χρόνων παρατηρήθηκε ότι, οι κανόνες αυτοί αλλάζουν, προσαρμόζονται και εξελίσσονται ανάλογα με τα δεδομένα της εποχής. Κατά συνέπεια η γλώσσα αποκτά μία δυναμική μορφή.

Με την εφεύρεση του ηλεκτρονικού υπολογιστή, άρχισαν να προκύπτουν νέες ανάγκες σχετικά με τον τρόπο επικοινωνίας του ανθρώπου με την μηχανή, αλλά και των μηχανών μεταξύ τους. Έτσι δημιουργήθηκε κάποια τεχνητή γλώσσα, που αποτελούσε το μέσο για την επικοινωνία του ανθρώπου με την μηχανή. Η γλώσσα αυτή κατείχε αυστηρή γραμματική και συντακτικό. Σε αντίθεση με την φυσική γλώσσα, σε περίπτωση που υπήρχαν αποκλίσεις από τους κανόνες γραμματικής και συντακτικού το μηχάνημα πιθανόν να ήταν ανίκανο να λειτουργήσει. Συνεπώς, υπάρχει ένα σημαντικό χάσμα ανάμεσα στην φυσική και την τεχνητή γλώσσα, η ελάττωση του οποίου θα ήταν ιδιαιτέρως ωφέλιμη. Ο βασικότερος λόγος για την ελάττωση του χάσματος είναι ότι το μεγαλύτερο μέρος της ανθρώπινης γνώσης είναι γραμμένο σε φυσική γλώσσα, έτσι το γεγονός της μετάφρασης της σε τεχνητή γλώσσα και η κατανόηση της από το μηχάνημα, αποτελεί ένα δύσκολο εγχείρημα.

Επιπλέον, η εξάπλωση του παγκόσμιου ιστού (World Wide Web ή Internet) είχε σαν αρχικό στόχο την διευκόλυνση της επικοινωνίας μεταξύ των χρηστών. Στις μέρες μας, η εξάπλωση του παγκοσμίου ιστού, οι διαστάσεις του, η συνεχής και αλλεπάλληλη επικοινωνία μεταξύ των χρηστών, παράγουν τεράστιο όγκο πληροφορίας, η διαχείριση του οποίου είναι ιδιαιτέρως δύσκολη. Για να πραγματοποιήσει ο χρήστης έστω μία απλή αναζήτηση στο Web είναι απαραίτητο ένα μηχάνημα να μπορέσει να καταλάβει την απλή φυσική ανθρώπινη γλώσσα. Πως μπορεί όμως να πραγματοποιηθεί κάτι τέτοιο αφού ένας υπολογιστής καταλαβαίνει μόνο τυχαίες ακολουθίες από μηδέν (0) και ένα (1);

Το χάσμα αυτό έρχεται να το γεφυρώσει ο τομέας της επεξεργασίας φυσικής γλώσσας (Natural Language Processing). Το ντεμπούτο της επιστήμης των υπολογιστών, σχεδόν συνάδει με την εμφάνιση της επεξεργασίας της φυσικής γλώσσας. Κύριος σκοπός του NLP είναι η κατανόηση της ανθρώπινης γλώσσας από πλευράς των υπολογιστών, έτσι ώστε να μπορέσουν να εξάγουν πληροφορία. Για τον σκοπό αυτό, έχει δημιουργηθεί μία σειρά εφαρμογών, όπως είναι η ορθογραφική και η συντακτική διόρθωση, ή η αυτόματη μετάφραση κειμένων. Άλλες ελαφρώς πιο εξεζητημένες εφαρμογές είναι η κατηγοριοποίηση και το φιλτράρισμα κειμένων καθώς και ο εντοπισμός συναισθήματος.

Ο εντοπισμός συναισθήματος – συναισθηματική ανάλυση (sentiment analysis) έχει σαν στόχο την εξαγωγή της υποκειμενικής πληροφορίας. Τέτοιου είδους πληροφορία καταγράφεται για εταιρίες, brand names, πολιτικά κόμματα ή δημοσκοπήσεις. Γενικότερα, θα μπορούσαμε να πούμε ότι ο σκοπός της συναισθηματικής ανάλυσης είναι ο εντοπισμός του συναισθήματος ή της οπτικής ενός συγγραφέα για κάποιο θέμα ή η άποψη μίας ομάδας ατόμων για κάποιο θέμα.

1.1. Παρούσα Κατάσταση

Στις μέρες μας, η συναισθηματική ανάλυση αποτελεί αναπόσπαστο κομμάτι της κοινωνικής ακρόασης. Η συνήθης προσέγγιση ακολουθεί την παρακάτω γραμμή:
"If a piece of content has more positive keywords than negative keyword, it's positive content; if it has more negative keywords than positive keywords, it's negative content."

Τα Συστήματα Ανάκτησης Πληροφοριών (Information Retrieval Systems), όπως οι εφαρμογές αναζήτησης, δεν μπορούν να ανιχνεύσουν το συναίσθημα, μπορούν απλά να εντοπίσουν αποσπασματικά, γνώση μέσω λέξεων - κλειδιών (keywords) [120]. Η συναισθηματική ανάλυση παρουσιάζεται ως νέα πηγή γνώσης με την ανάπτυξη του World Wide Web. Οι εταιρίες επιθυμούν την διαμόρφωση συναισθηματικής ανάλυσης (marketing analysis), έτσι ώστε να κατανοήσουν την συμπεριφορά των πελατών. Η συναισθηματική ανάλυση από την πλευρά της τους διευκολύνει να μειώσουν το κόστος, αφού δεν χρειάζονται επιπλέον προσωπικό για να κατανοήσουν τους πελάτες τους, αλλά είναι κάτι που γίνεται αυτόματα. Επιπλέον, η αυτόματη ανίχνευση συναισθήματος υποκαθιστά την ύπαρξη ερευνών και ερωτηματολογίων, τα οποία ρωτάνε άμεσα την γνώμη των πελατών [121].

Από αυτό το είδος «επανάστασης», ωφελούνται επίσης οι άνθρωποι - πελάτες, προσωπικά, αφού και αυτοί με τη σειρά τους επιθυμούν να έχουν προτάσεις για διάφορα αντικείμενα, ή υπηρεσίες από άλλους ανθρώπους. Γνωστό είναι ότι ένα καλό μέτρο της ποιότητας είναι η άφθονη ποσότητα καλών κριτικών. Οι άνθρωποι έχουν μία ροπή, οτιδήποτε και αν αναζητούν, οτιδήποτε είδους και αν αυτό είναι, προϊόν ή υπηρεσία, ή event ή πολιτικό γεγονός, ή ταινία, να βασίζονται σε θετικές κριτικές ώστε να το επιλέξουν. Όπως έγραψε και ο Liu [122]: «Όταν ένας άνθρωπος πρέπει να πάρει μία απόφαση, συνήθως ρωτάει την άποψη από φίλους ή την οικογένεια του». «Πλέον, αν κάποιος επιθυμεί να αγοράσει ένα προϊόν, δεν είναι υποχρεωμένος/η να περιοριστεί στο φιλικό ή οικογενειακό του/της περιβάλλον, καθώς υπάρχουν πολλές κριτικές προϊόντων στο Web, οι οποίες παρουσιάζουν την γνώμη ενός συνόλου χρηστών σχετικά με το προϊόν»

Γενικότερα, οι άνθρωποι χρησιμοποιούν το διαδίκτυο για να εκθέσουν τις γνώμες τους. Ειδικότερα, χρησιμοποιούνται forum, blogs και group συζητήσεων. Τα social networks, όπως το Facebook, ή το Twitter παρέχουν πολλές γνώμες και όχι μόνο ενός συγκεκριμένου είδους. Όμως υπάρχουν κάποια social networks, όπως το Foursquare που αναφέρονται σε έναν συγκεκριμένο τομέα και παρουσιάζουν ένα σύνολο από σχετικά feedback.

1.2. Σκοπός

Ο σκοπός της συγκεκριμένης εργασίας είναι η ανασκόπηση της βιβλιογραφίας, η οποία είναι σχετική με τα Big Data και στη συνέχεια με την συναισθηματική ανάλυση. Επιπλέον, θα δημιουργηθεί ένα σύστημα για την επέκταση του Conceptnet, μέσω κάποιου lexicon. Στη συνέχεια, με δεδομένο ένα labelled dataset θα πραγματοποιηθεί προεπεξεργασία και αξιολόγηση των tweets με το SentiWordnet καθώς και με το εμπλουτισμένο Conceptnet.

Τελικός στόχος της εργασίας είναι να αποδείξει, εάν ο συνδυασμός του Conceptnet με το SentiWordnet είναι πιο αποδοτικός από το απλό SentiWordnet, συγκριτικά με το Initial Score του αρχικού labelled dataset.

1.3. Αντικειμενικοί Στόχοι

Για την αντιμετώπιση της παραπάνω ερευνητικής πρόκλησης, έχουν τεθεί οι εξής στόχοι:

- **Αντικειμενικός Στόχος 1 - Ανασκόπηση βιβλιογραφίας:** Ο αρχικός στόχος της παρούσας εργασίας είναι να αναφερθούν και να αναλυθούν οι βασικές έννοιες που αφορούν τον κλάδο των big data και του sentiment analysis, σύμφωνα με την υπάρχουσα βιβλιογραφία.
- **Αντικειμενικός Στόχος 2 - Ανεύρεση Μηχανισμών:** Στόχος είναι η ανεύρεση αλγορίθμων και μηχανισμών για την βελτιστοποίηση της διαχείρισης και απόδοσης των δεδομένων.
- **Αντικειμενικός Στόχος 3 - Αξιολόγηση:** Έλεγχος και αξιολόγηση των μηχανισμών.

1.4. Δομή Εργασίας

Η δομή της παρούσας εργασίας ακολουθεί τη μεθοδολογία όπως έχει περιγραφεί από τους Phillips and Pugh [123]. Η δομή της συγκεκριμένης εργασίας είναι η παρακάτω:

- **Κεφάλαιο 1 - Εισαγωγή:** Στο εν λόγω κεφάλαιο πραγματοποιείται ο ορισμός του προβλήματος στο οποίο εστιάζει η συγκεκριμένη εργασία. Για την επίλυση του προβλήματος αυτού τίθενται αντικειμενικοί στόχοι και ένας συγκεκριμένος σκοπός.

- **Κεφάλαιο 2 – Big Data και NoSQL βάσεις δεδομένων:** Το συγκεκριμένο κεφάλαιο αρχικά αναφέρεται στα μεγάλα δεδομένα (Big Data), στις διαστάσεις και στις βασικές πτυχές τους. Ο όγκος των δεδομένων είναι τεράστιος και δύσκολα διαχειρίσιμος. Για την διαχείριση συνήθως χρησιμοποιούνται NoSQL βάσεις δεδομένων όπως η MongoDB, αναφορά στις οποίες πραγματοποιείται επίσης στο εν λόγω κεφάλαιο.
- **Κεφάλαιο 3 – Μηχανική μάθηση:** Σε αυτό το κεφάλαιο αναφερόμαστε στην μηχανική μάθηση. Έτσι παρουσιάζονται αναλυτικά οι φάσεις της, η αλληλεπίδραση με την στατιστική καθώς και ένας απλός ταξινομητής, ο Naïve Bayes Classifier.
- **Κεφάλαιο 4 – Συναισθηματική ανάλυση:** Στο εν λόγω κεφάλαιο, πραγματοποιείται διεξοδική αναφορά στην έννοια της συναισθηματικής ανάλυσης. Δηλαδή αρχικά, παρουσιάζεται ιστορική αναδρομή. Στη συνέχεια αναλύονται τα επίπεδα και οι τεχνικές της και πραγματοποιείται μία μικρή αναφορά σε sentiment lexicons και κάποιες εφαρμογές.
- **Κεφάλαιο 5 – Conceptnet και Sentiwordnet:** Το κεφάλαιο αυτό ασχολείται κυρίως με δύο εργαλεία, δύο λεξικά, το Conceptnet και το SentiWordnet. Αρχικά, αναφέρεται στο Conceptnet και τα πλεονεκτήματα του, καθώς και στην κατανόηση του περιεχομένου του. Στη συνέχεια πραγματοποιείται αντίστοιχη ανάλυση του SentiWordnet
- **Κεφάλαιο 6 – Το Σύστημα - Αποτελέσματα:** Στο εν λόγω κεφάλαιο, πραγματοποιείται η υλοποίηση του συστήματος. Αρχικά, παρουσιάζεται η εγκατάσταση του Conceptnet, στη συνέχεια υπάρχει ένας αλγόριθμος ο οποίος

επεκτείνει το Conceptnet μέσω του Sentiwordnet. Τέλος, με την βοήθεια των δύο παραπάνω λεξικών, δίνεται μία συναισθηματική σφραγίδα στο αρχικό μας corpus και παρουσιάζεται το τελικό αποτέλεσμα του αλγορίθμου, δηλαδή ποια από τις δύο περιπτώσεις δούλεψε καλύτερα, συγκριτικά με το initial score.

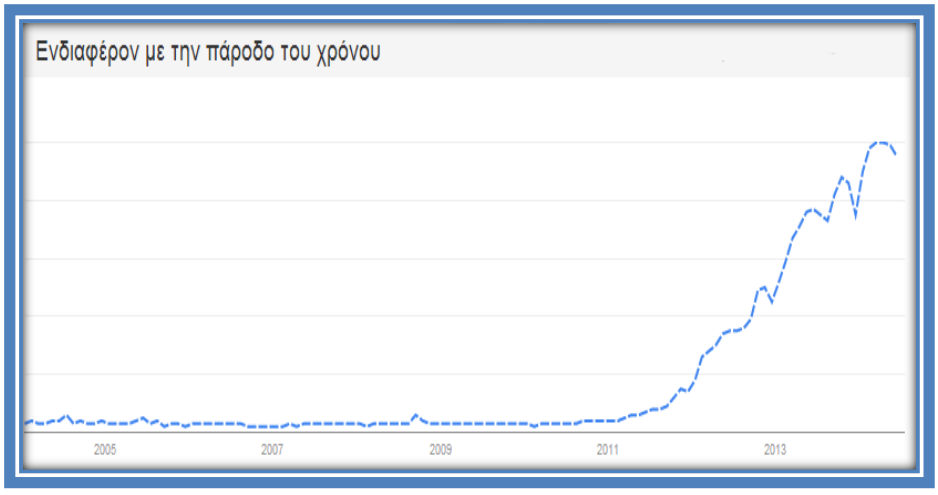
Κεφάλαιο 2: Big Data και NoSQL

2.1. Εισαγωγή

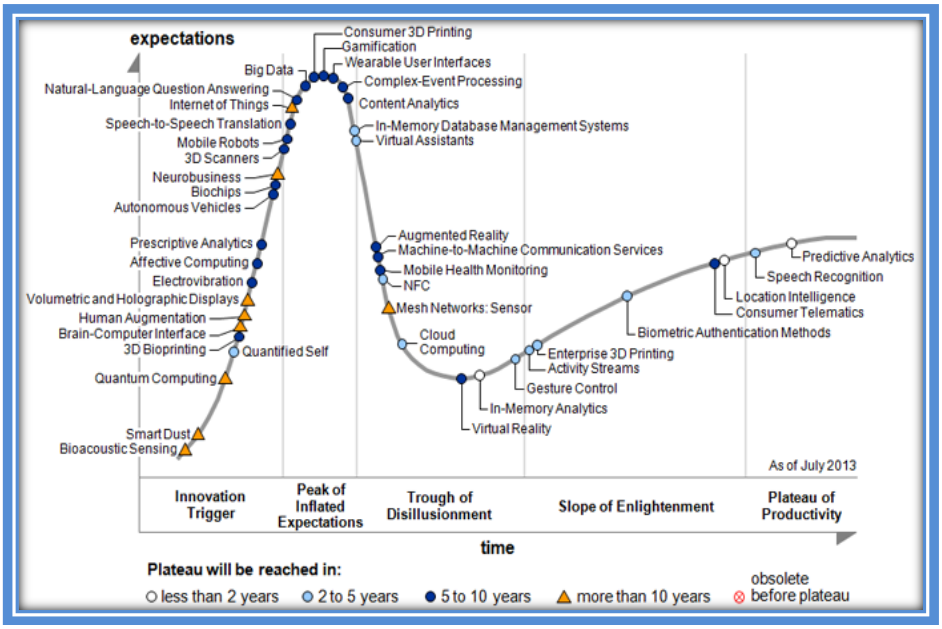
Ζούμε στην εποχή των “big data”. Τα δεδομένα έχουν μετατραπεί πλέον σε μια μορφή πρώτης ύλης, μια πηγή τεράστιας οικονομικής και κοινωνικής αξίας. Η ανάπτυξη σε τομείς όπως η εξόρυξη δεδομένων και η ανάλυση τους (analytics), καθώς και οι τεράστιες υπολογιστικές και αποθηκευτικές ικανότητες των σύγχρονων υπολογιστικών συστημάτων, κάνουν διαθέσιμο σε οργανισμούς και ιδιώτες, ένα τεράστιο όγκο πληροφοριών. Επιπρόσθετα, η σημασία των δεδομένων αυξάνεται καθώς αυξάνονται οι χρήστες και οι συσκευές που είναι διασυνδεδεμένες μέσω ψηφιακών επικοινωνιακών δικτύων και παράγουν, διαμοιράζονται και έχουν πρόσβαση σε δεδομένα.

2.2. Παρούσα κατάσταση – Εισαγωγή στα Big Data

Δεν υπάρχει ένας ακριβής ορισμός για την έννοια Big Data (μέγα δεδομένα). Ο όρος χρησιμοποιείται αρκετά συχνά από εταιρείες και γίνεται όλο και δημοφιλέστερος όπως φαίνεται στην Εικόνα 2.1 με το εργαλείο της Google’s : «Google Trend». Για να γίνει πιο κατανοητό το αποτέλεσμα της Εικόνας 2.1, παρατίθεται η Εικόνα 2.2 ώστε να αποκτήσουμε μια πληρέστερη εικόνα από το δημοφιλές Gartner’s Hype Cycle για τις αναδυόμενες τεχνολογίες. Το 2012, σύμφωνα με το Gartner, τα Big Data μετακινήθηκαν από το στάδιο "Technology Trigger" και μεταφέρθηκαν στο στάδιο "peak of inflated expectations".



Εικόνα 2.1 Google trends Big Data



Εικόνα 2.2 Gartner's Hype Cycle, [103]

Χωρίς καμία αμφιβολία τα Big Data, θα είναι ένας από τους δημοφιλέστερους όρους τα επόμενα χρόνια. Με μία πρώτη ματιά τα Big Data αναφέρονται σε κάτι μεγάλο και γεμάτο πληροφορίες αλλά ας γίνουμε λίγο πιο ακριβείς όσο αναφορά την έννοια αυτή [104].

Κατά κάποιο τρόπο όλοι μας, χειριζόμαστε καθημερινά (big) δεδομένα. Οι χρήστες δημιουργούν περιεχόμενα όπως blog, posts, tweets, φωτογραφίες και βίντεο στα social networks. Οι Servers συνεχώς καταγράφουν μηνύματα με τις ενέργειες τους και το τι κάνουν. Οι επιστήμονες πραγματοποιούν λεπτομερές μετρήσεις, οι εταιρείες καταγράφουν πληροφορίες σχετικά με τις πωλήσεις, τους προμηθευτές, τους πελάτες κτλ. Το 2010 πάνω από 4 δισεκατομμύρια κόσμος (το 60% του παγκόσμιου πληθυσμού) χρησιμοποιούσαν κινητά τηλέφωνα και 12% από αυτούς διέθεταν smartphones, όπου κάθε χρόνο υπολογίζεται ότι αυξάνονται κατά 20%. Πάνω από 30 εκατομμύρια δικτυωμένοι sensor nodes είναι πλέον στην μεταφορά, την αυτοκινητοβιομηχανία, τη βιομηχανία, τις επιχειρήσεις κοινής ωφέλειας και του λιανικού εμπορίου. Ο αριθμός αυτών των αισθητήρων αυξάνεται με ρυθμό 30% το χρόνο. Μελέτη έδειξε πως κάθε χρόνο ο όγκος των δεδομένων θα αυξάνεται κατά 40% μέχρι το 2020 [104].

Αυτή η έκρηξη δεδομένων έχει επηρεάσει τις επιχειρήσεις. Οι παραδοσιακές βάσεις δεδομένων (RDBMS), πλέον φανερώνουν πως έχουν φτάσει τα όρια τους, δε μπορούν ανταπεξέλθουν στις απαιτήσεις και έχουν αποτύχει στη διαχείριση των big

data. Το πρόβλημα είναι απλό: ενώ η χωρητικότητα των σκληρών δίσκων έχει αυξηθεί μαζικά τα τελευταία χρόνια, η ταχύτητα πρόσβασης (ο ρυθμός με τον οποίο τα δεδομένα μπορούν να διαβαστούν από το δίσκο) δεν είναι σε θέση να συμβαδίσει. Το 1990 ένας τυπικός δίσκος μπορούσε να αποθηκεύσει περίπου στα 100 MB δεδομένων έχοντας ταχύτητα μεταφοράς στα 4.4MB/s. Επομένως ήταν εφικτό να διαβαστεί ολόκληρος ο δίσκος περίπου στα 5 λεπτά. Τα τελευταία 20 χρόνια, οι δίσκοι που διαθέτουν ένα terabyte είναι ο κανόνας και με ταχύτητα μεταφοράς περίπου στα 100MB/s. Συνεπώς απαιτείται πάνω από δυόμιση ώρες για να προσπελαστούν και να διαβαστούν όλα τα δεδομένα από τον δίσκο (πόσο μάλλον να πραγματοποιηθεί κάποια εγγραφή στο δίσκο, που απαιτεί περισσότερο χρόνο). Ο προφανής τύπος για να μειωθεί το κόστος του χρόνου είναι να διαιρέσουμε την αρχική συλλογή σε μικρότερα κομμάτια και να πραγματοποιείται η ανάγνωση των δεδομένων από πολλαπλούς δίσκους ταυτόχρονα [105].

Για την αντιμετώπιση των προκλήσεων των Big Data, ένα νέο είδος τεχνολογιών έχει προκύψει. Οι περισσότερες από αυτές τις τεχνολογίες είναι κατανεμημένες σε πολλά μηχανήματα και έχουν ομαδοποιηθεί υπό τον όρο «NoSQL». NoSQL είναι στην πραγματικότητα ένα αρκτικόλεξο που έχει προκύψει από τον όρο «Not Only SQL». Σε ορισμένες περιπτώσεις, αυτές οι νέες τεχνολογίες είναι πιο περίπλοκες από ό,τι οι παραδοσιακές βάσεις δεδομένων, και σε άλλες είναι απλούστερες. Δεν υπάρχει καμία λύση που να ταιριάζει σε όλες τις περιπτώσεις. Αυτά τα νέα συστήματα μπορούν να αναβαθμίσουν σε πολύ μεγαλύτερα σύνολα δεδομένων, αλλά η χρησιμοποίηση των συστημάτων αυτών απαιτούν μια σειρά από νέες τεχνολογίες. Πολλές από αυτές τις τεχνολογίες άρχισαν για πρώτη φορά από δύο μεγάλες επιχειρήσεις: τη Google και τη Amazon. Η πιο δημοφιλής τεχνολογία είναι πιθανώς το πλαίσιο υπολογισμού MapReduce που εισήχθη από την Google το 2004. Η Amazon δημιούργησε ένα καινοτόμο κατανεμημένο σύστημα που ονομάζεται Dynamo και ουσιαστικά μπορεί να χαρακτηριστεί ως μια αποθήκη που περιέχει σετ από κλειδία-τιμές. Η open source κοινότητα ανταποκρίθηκε δημιουργώντας το Hadoop (ελεύθερη υλοποίηση του MapReduce), HBase, MongoDB, Cassandra, RabbitMQ και αναρίθμητα άλλα έργα [106].

Τα Big Data είναι μεγάλη πρόκληση όχι μόνο για τους μηχανικούς λογισμικού αλλά και για τις κυβερνήσεις, τους οικονομολόγους και γενικά σε κάθε τομέα. Υπάρχουν αρκετές ενδείξεις πως τα Big Data μπορούν να διαδραματίσουν σημαντικό

ρόλο όχι μόνο προς όφελος του ιδιωτικού τομέα αλλά και της εθνικής οικονομίας και των πολιτών. Χαρακτηριστικά μελέτες από το McKinsey Global Institute έχουν δείξει πως αν το ιατρικό σύστημα των Ηνωμένων Πολιτειών μπορούσε να αξιοποιήσει τα big data δημιουργικά και αποτελεσματικά, ώστε να συμβάλλουν ποιότητα και την απόδοση τους τότε το κέρδος των Ηνωμένων Πολιτειών να φτάσει και τα 300 δισεκατομμύρια δολλάρια το χρόνο. Ένα άλλο παράδειγμα δείχνει πως στον ιδιωτικό τομέα, ένας έμπορος λιανικής πωλήσεως που χρησιμοποιεί τα big data στο έπακρο έχει τη δυνατότητα να αυξήσει το περιθώριο λειτουργίας του κατά περισσότερο από 60 τοις εκατό [100].

2.3. Ορισμός των Big Data

Ένας ορισμός που θα μπορούσε να χρησιμοποιηθεί για τα Big Data είναι αυτός που αναφέρεται στο επιστημονικό άρθρο από το McKinsey Global Institute τον Μάιο του 2011: «Ως Big Data θεωρούνται τα σύνολα των δεδομένων που το μέγεθος τους είναι πέρα από την ικανότητα των τυπικών εργαλείων λογισμικού βάσεων δεδομένων να ανταπεξέλθουν, ώστε να συλλέξουν, αποθηκεύσουν, να διαχειριστούν και να αναλύσουν.» [99].

Ένας άλλος ενδιαφέρον ορισμός που μπορεί να χρησιμοποιηθεί για τα Big Data, πιο οικονομικός, είναι αυτός από το IDC :

«Οι Big data τεχνολογίες περιγράφουν μια γενιά τεχνολογιών και αρχιτεκτονικών, σχεδιασμένες ώστε να αποσπάσουν γνώση, σε σχέση με την οικονομία, από πολύ μεγάλους όγκους δεδομένων παρέχοντας τη δυνατότητα υψηλής ταχύτητας (high-velocity), ανακάλυψης (discovery), ή / και της ανάλυσης (analysis)» [100].

Σύμφωνα με τον Ed Dumbill επικεφαλή στο Strata O'Reilly Conference, τα Big Data μπορούν να περιγραφούν ως «τα δεδομένα που υπερβαίνουν την ικανότητα επεξεργασίας των συμβατικών συστημάτων βάσεων δεδομένων. Τα δεδομένα είναι πολύ μεγάλα, κινούνται πάρα πολύ γρήγορα, ή δεν ταιριάζουν με τις αρχιτεκτονικές της βάσης δεδομένων μας. Για να αποκτήσουν αξία από αυτά τα δεδομένα, θα πρέπει να βρούμε έναν εναλλακτικό τρόπο για να το επεξεργαστούμε.» [101].

Επιπλέον, στο Gartner’s IT Glosarry τα Big Data ορίζονται ως μεγάλου όγκου, ταχύτητας και ποικιλίας πληροφορίες που απαιτούν αποδοτικές, καινοτόμες

μορφές επεξεργασίας πληροφορίες για τη καλύτερη εποπτεία και λήψη αποφάσεων [102].

Σε ένα απλούστερο ορισμό θεωρούμε τα Big Data πως είναι μια έκφραση που αποτελείται από διαφορετικά πολύ μεγάλα σύνολα δεδομένων, εξαιρετικά πολύπλοκα, αδόμητα, που οργανώνονται, αποθηκεύονται και υφίστανται επεξεργασία σύμφωνα με ειδικές μεθόδους και τεχνικές που χρησιμοποιούνται για τις επιχειρηματικές διαδικασίες. Υπάρχουν πολλοί ορισμοί για τα Big Data σε όλο τον κόσμο, αλλά θεωρούμε ότι ο πιο σημαντικός είναι αυτός που κάθε ηγέτης δίνει στα δεδομένα μιας εταιρείας του. Ο τρόπος που τα Big Data ορίζονται έχει επιπτώσεις στη στρατηγική μιας επιχείρησης. Κάθε ηγέτης πρέπει να ορίσει την έννοια, προκειμένου να φέρει το ανταγωνιστικό πλεονέκτημα για την εταιρεία.

2.4. Διαστάσεις και βασικές πτυχές εννοιών

Σύμφωνα με το Mike 2.0, το πρότυπο ανοιχτού κώδικα για τη διαχείριση των πληροφοριών, τα Big Data ορίζονται από το μέγεθός τους, το οποίο περιλαμβάνει μία μεγάλη, σύνθετη και ανεξάρτητη συλλογή συνόλων δεδομένων, με τη δυνατότητα η κάθεμια να αλληλεπιδρά με την άλλη. Επιπλέον, μια σημαντική πτυχή των Big Data είναι το γεγονός ότι δεν μπορούν να αντιμετωπιστούν με τυποποιημένες τεχνικές διαχείρισης δεδομένων λόγω της ανακολουθίας και της μη προβλεψιμότητας των πιθανών συνδυασμών [107].

Κατά την άποψη της IBM τα Big Data έχουν τέσσερις πτυχές:

- Όγκος: ο όγκος (volume) αναφέρεται στην ποσότητα των δεδομένων που συλλέγονται από μια εταιρεία. Αυτά τα στοιχεία πρέπει να χρησιμοποιηθούν περαιτέρω για να ληφθούν σημαντικές γνώσεις.
- Ταχύτητα: η ταχύτητα (velocity) αναφέρεται στον χρόνο στον οποίο μπορεί να γίνει η επεξεργασία των μαζικών δεδομένων. Ορισμένες δραστηριότητες είναι πολύ σημαντικές και πρέπει να ληφθούν άμεσες απαντήσεις, αυτός είναι ο λόγος που η γρήγορη επεξεργασία μεγιστοποιεί την απόδοση
- Ποικιλία: η ποικιλία (variety) Αναφέρεται στο είδος των δεδομένων που μπορεί να περιλαμβάνουν τα Big Data. Αυτά τα δεδομένα μπορούν να διαμορφωθούν, και ας είναι αδόμητα [108].

- Ειλικρίνεια: η ειλικρίνεια (veracity) αναφέρεται στο βαθμό στον οποίο ένας ηγέτης εμπιστεύεται το χρησιμοποιημένη πληροφορία προκειμένου να λάβει μια απόφαση. Έτσι, η λήψη των σωστών συσχετίσεων στο Big Data είναι πολύ σημαντική για το μέλλον των επιχειρήσεων [109].

2.5. Εισαγωγή στις NoSQL Βάσεις Δεδομένων

Όπως αναφέρθηκε σε προηγούμενη ενότητα, τα συστήματα NoSQL μπορεί να είναι πιο περίπλοκα από τις παραδοσιακές βάσεις δεδομένων, και σε άλλες περιπτώσεις μπορεί να είναι απλούστερα. Τα συστήματα αυτά μπορεί να επεκταθούν σε πολύ μεγαλύτερο σύνολο δεδομένων, αλλά πρέπει να γίνουν κάποιοι συμβιβασμοί. Στις NoSQL βάσεις δεδομένων το μοντέλο δεδομένων είναι γενικά πολύ πιο περιορισμένο από ό, τι έχουμε συνηθίσει σε μία παραδοσιακή βάση δεδομένων SQL.

NoSQL (Όχι μόνο SQL) είναι ένας όρος που χρησιμοποιείται για να ορίσει με κάποιο τρόπο τα συστήματα διαχείρισης βάσεων δεδομένων που διαφέρουν από το κλασικό σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS). Αυτές οι αποθήκες δεδομένων δεν απαιτούν προκαθορισμένα σχήματα πινάκων, συνήθως αποφεύγονται να εντάσσονται πράξεις ζεύξεων (Join), δεν επιχειρούν να παρέχουν ιδιότητες ACID (αναφερόμαστε εκτενέστερα στην Ενότητα 2.6) και γενικά χρησιμοποιούν τυπική οριζόντια κλιμάκωση (Scaling horizontally).

Μια σημαντική πτυχή του προηγούμενου ορισμού είναι η έννοια της επεκτασιμότητας. Η επεκτασιμότητα μπορεί να περιγράφει ως μια επιθυμητή ιδιότητα ενός συστήματος, ενός δικτύου ή μιας διαδικασίας, η οποία δείχνει την ικανότητά του να χειριστεί είτε κάποια αυξανόμενη ποσότητα της εργασίας με αποδοτικό τρόπο ή να είναι άμεσα διευρυμένη [110].

Η επεκτασιμότητα μπορεί να επιτευχθεί σε δύο διαστάσεις :

- Scale up : κάθετη κλιμάκωση (αύξηση της μνήμης RAM σε έναν υπάρχοντα κόμβο (node)).
- Scale out: οριζόντια κλιμάκωση (πρόσθεση ενός κόμβου στο σύμπλεγμα (cluster)).

Οι NoSQL βάσεις δεδομένων γενικά χρησιμοποιούν οριζόντια κλιμάκωση, εάν η βάση δεδομένων χρειάζεται περισσότερους πόρους (αποθήκευση, υπολογιστική ισχύ,

μνήμη, κλπ.), έναν ή περισσότερους κόμβους θα πρέπει να προστεθούν εύκολα στο σύμπλεγμα. Ωστόσο, η προσθήκη περισσότερων μηχανημάτων αυξάνει τον κίνδυνο των αποτυχιών (σε θέματα δικτύου, ανεπαρκής χώρος στο δίσκο ή μνήμη, προβλήματα υλικού, κλπ.). Ως εκ τούτου, μια βασική ιδιότητα ενός κατανεμημένου συστήματος είναι η ανεκτικότητα. Η ανοχή σε σφάλματα είναι η ιδιότητα που επιτρέπει σε ένα σύστημα να συνεχίσει να λειτουργεί σωστά σε περίπτωση αποτυχίας ορισμένων των συστατικά του [111].

Οι NoSQL βάσεις δεδομένων είναι πολύ διαφορετικές στο μοντέλο δεδομένων τους, αλλά και στην αρχιτεκτονική τους. Μήπως αυτό σημαίνει ότι δεν πρέπει πλέον να χρησιμοποιούνται οι σχεσιακές βάσεις δεδομένων; Φυσικά και όχι, δεν είναι δυνατόν να έχουμε ένα μέγεθος και τρόπο επίλυσης που να ταιριάζει σε όλες τις λύσεις. Οι RDBMS είναι πολύ ισχυρό εργαλείο σε ένα εύλογο ποσό δεδομένων και καλό για τις πράξεις τύπου Join. Ως εκ τούτου, θα ήταν πολύ ενδιαφέρον να αναρωτηθούμε τι θα μπορούσε να συμβεί συγκεκριμένα αν προσπαθήσουμε να ασχοληθούμε με ένα τεράστιο μέγεθος δεδομένων με ένα κλασικό RDBMS

2.6. Ιδιότητες ACID

Συναλλαγή (transaction) ονομάζεται κάθε σειρά ενεργειών, όπου κάθε ενέργεια διαβάζει ή γράφει αντικείμενα σε μια βάση δεδομένων. Η εφαρμογή των ιδιοτήτων ACID σε κάθε transaction εγγυάται την αξιοπιστία των σχεσιακών βάσεων δεδομένων. Ακολουθεί η ανάλυση των ιδιοτήτων ACID [112]:

- Ατομικότητα (atomicity): εξασφαλίζεται ότι είτε θα πραγματοποιηθούν όλες οι πράξεις ενός transaction είτε ότι θα αποτύχουν όλες, αφήνοντας τη βάση ανεπηρέαστη. Σε περίπτωση κατάρρευσης του συστήματος κατά τη διάρκεια ενός transaction θα αναιρεθούν ό,τι αλλαγές έχουν γίνει στη βάση και αυτή θα έρθει στη μορφή που ήταν πριν ξεκινήσει η εκτέλεση του.
- Συνέπεια (consistency): Εξασφαλίζεται ότι πριν και μετά την εκτέλεση ενός transaction, οι κανόνες και περιορισμοί που διέπουν τη βάση δεδομένων πληρούνται.
- Απομόνωση (isolation): Κάθε transaction θεωρεί πως είναι το μοναδικό που τρέχει στο σύστημα. Τα ενδιάμεσα αποτελέσματα ενός transaction δεν επηρεάζουν άλλα transaction που ενδεχομένως έχουν πρόσβαση στα ίδια

δεδομένα. Το τελικό αποτέλεσμα της βάσης είναι το ίδιο, αν όλα τα transactions έτρεχαν σειριακά.

- Μονιμότητα (durability): Οι αλλαγές που προκαλεί στη βάση ένα transaction που επιτυγχάνει θα παραμείνουν ακόμα και μετά από κατάρρευση του συστήματος.

2.7. Ανάγκη ανάπτυξης NoSQL Βάσεων Δεδομένων

Το σχεσιακό μοντέλο είναι γενικά ευέλικτο, διαθέτει πολλά εργαλεία διαχείρισης των δεδομένων, ενημέρωσης της βάσης, update, εκτέλεσης συναλλαγών με αυτήν, transactions, και παροχής πληροφοριών στο χρήστη μέσω υποβολής ερωτημάτων. Αναπαριστά τα δεδομένα με δισδιάστατους πίνακες και τις σχέσεις μεταξύ τους μέσω κλειδιών. Είναι μια εύκολη στην κατανόηση δομή και καλύπτει τις ανάγκες των περισσότερων εφαρμογών.

Ωστόσο όλο και περισσότερο έχει παρατηρηθεί η ανάγκη για ταυτόχρονη αποθήκευση και διαχείριση μεγάλου όγκου κατανεμημένων δεδομένων, τα οποία εμφανίζουν πολυμορφία. Παρουσιάζεται συχνά η ανάγκη μεταφοράς των δεδομένων πάνω από ένα δίκτυο και η άντληση γνώσης από συστήματα Peer-to-Peer, P2P. Τέλος απαιτείται η χρήση των υπηρεσιών πάνω από διαφορετικές πλατφόρμες, σε ποικίλες εφαρμογές, το λογισμικό και τα εργαλεία των οποίων μπορεί να διαφέρουν. Για αυτού του είδους τις εφαρμογές οι βάσεις δεδομένων που υλοποιούν το σχεσιακό μοντέλο δεν είναι ιδανικές, καθώς τα ερωτήματα που τίθενται στη βάση και η εκτέλεση των JOINS προς απάντησή τους, απαιτούν την επεξεργασία μεγάλου όγκου δεδομένων και τη μεταφορά τους πάνω από το δίκτυο, προκαλώντας μεγάλο latency και αυξημένο χρόνο απόκρισης των υπηρεσιών.

Για την υλοποίηση σύνθετων, κατανεμημένων υπηρεσιών δεν αρκούν παλαιότερα μοντέλα ανάπτυξης, όπως το μοντέλο καταρράκτη, αλλά συντομότερα, διαδραστικότερα μοντέλα, όπου η ανάπτυξη και σχεδίαση είναι πιο κοντά στην παραγωγή των τελικών προϊόντων. Ακόμα οι διακυμάνσεις στις απαιτήσεις των χρηστών έκανε επιτακτική την ανάγκη ανάπτυξης μηχανισμών γρήγορου scaleup και scaledown στους πόρους που χρησιμοποιεί η εφαρμογή ανάλογα με τις ανάγκες κάθε χρονική στιγμή. Οι σχεσιακές βάσεις δεν έχουν δημιουργηθεί για τόσο ευέλικτες

περιπτώσεις. Αναπτύσσονται σε γνωστούς εκ των προτέρων πόρους, με γνωστό schema και δεν επιτρέπουν πολυμορφία.

Για τους παραπάνω λόγους και για εφαρμογές που απαιτούν διαχείριση ολοένα αυξανόμενου όγκου δεδομένων, σε πραγματικό χρόνο ή σε στατιστικές μελέτες, ήταν απαραίτητη η δημιουργία ενός άλλου τρόπου αναπαράστασης και διαχείρισης των δεδομένων. Αναπτύχθηκαν οι NoSQL βάσεις δεδομένων, όπου τα δεδομένα δεν είναι απαραίτητα δομημένα σε πίνακες, όπως στο σχεσιακό μοντέλο, αλλά με διάφορους τρόπους όπως θα δούμε παρακάτω. Αυτό δίνει μεγάλη ελευθερία στους χρήστες που θέλουν να αναπτύξουν υπηρεσίες καθώς μπορούν να επιλέξουν το μοντέλο που ταιριάζει καλύτερα στις ανάγκες της εφαρμογής τους. Δεν υπάρχει συγκεκριμένη γλώσσα προγραμματισμού για το χειρισμό τους. Κάθε NoSQL βάση παρέχει εργαλεία για την διαχείριση των δεδομένων καθώς και δική της γλώσσα υποβολής ερωτημάτων, query language [113].

Οι NoSQL βάσεις παρέχουν επεκτασιμότητα, πολυμορφία και αύξηση στις επιδόσεις των εφαρμογών. Εφαρμόζονται σε clusters, υποστηρίζουν διαχείριση κατανεμημένων δεδομένων και παρέχουν μηχανισμούς ανάκτησης των δεδομένων μετά από καταστροφή, failure tolerant. Για την παρουσίαση των NoSQL βάσεων είναι απαραίτητη η αναφορά των χαρακτηριστικών που πρέπει να τις διακρίνουν και των κανόνων που πρέπει να ακολουθούν προκειμένου να αποτελούν αξιόπιστα εργαλεία αποθήκευσης και διαχείρισης δεδομένων. Στην επόμενες ενότητα δίνεται μια προεπισκόπηση των ιδιοτήτων που πρέπει να διακρίνουν τις βάσεις δεδομένων, καθώς και οι μηχανισμοί που χρησιμοποιούνται για την εκτέλεση των transactions.

2.8. MongoDB

Η MongoDB είναι μία cross-platform document-oriented βάση δεδομένων. Θεωρείται ως NoSQL βάση δεδομένων και αποφεύγει την παραδοσιακή σχεσιακή δομή μίας τυπικής βάσης δεδομένων πίνακα. Έτσι, αξιοποιεί έγγραφα τύπου JSON και δυναμικά σχήματα (BSON μορφή), καθιστώντας την ενσωμάτωση των δεδομένων σε ορισμένους τύπους εφαρμογών ευκολότερη και ταχύτερη. Η MongoDB κυκλοφόρησε κάτω από ένα συνδυασμό του GNU Affero General Public License και με άδεια χρήσης Apache. Αποτελεί ελεύθερο λογισμικό ανοικτού πηγαίου κώδικα.

Για πρώτη φορά αναπτύχθηκε από την εταιρεία λογισμικού MongoDB Inc. τον Οκτώβριο του 2007 ως συστατικό μιας σχεδιασμένης πλατφόρμας. Η εταιρεία στράφηκε στην ανάπτυξη ενός open source μοντέλου το 2009, με την MongoDB να προσφέρει εμπορική υποστήριξη και άλλες υπηρεσίες [114]. Από τότε και στο εξής, η MongoDB έχει υιοθετηθεί ως backend λογισμικό από πληθώρα ιστοσελίδων και υπηρεσιών, συμπεριλαμβανομένων των Craigslist, eBay, και Foursquare μεταξύ άλλων. Από τον Ιούλιο του 2015, η MongoDB είναι το τέταρτο πιο δημοφιλές σύστημα διαχείρισης βάσεων δεδομένων, και το πιο δημοφιλές για την αποθήκευση εγγράφων.

Μερικά από τα χαρακτηριστικά της είναι τα ακόλουθα:

- **Document-Oriented**

Αντί να λάβει ως δεδομένο μία επιχειρηματική δραστηριότητα και να τη χωρίσει σε πολλαπλές σχεσιακές δομές, η MongoDB μπορεί να αποθηκεύσει την επιχειρηματική δραστηριότητα στον ελάχιστο αριθμό εγγράφων. Για παράδειγμα, αντί να αποθηκευτεί ο τίτλος και οι πληροφορίες του συντάκτη σε δύο διακριτές σχεσιακές δομές, τίτλος και συγγραφέας, μπορούν όλες οι σχετικές πληροφορίες να αποθηκευτούν σε ένα ενιαίο έγγραφο που ονομάζεται Document [115].

- **Ad hoc ερωτήματα**

Η MongoDB υποστηρίζει αναζήτηση με βάση το πεδίο, επερωτήσεις εύρους καθώς και αναζητήσεις κανονικών εκφράσεων. Τα queries μπορούν να επιστρέψουν συγκεκριμένα πεδία εγγράφων και επίσης περιλαμβάνουν ορισμένες συναρτήσεις της JavaScript.

- **Indexing**

Οποιοδήποτε πεδίο σε ένα MongoDB έγγραφο μπορεί να αναπροσαρμοστεί (οι δείκτες της MongoDB είναι εννοιολογικά όμοιοι με εκείνους των RDBMSes). Δευτεροβάθμιοι δείκτες είναι επίσης διαθέσιμοι.

- **Αντιγραφής**

Η MongoDB προσφέρει υψηλή διαθεσιμότητα σε replica sets [115]. Ένα replica set αποτελείται από δύο ή περισσότερα αντίγραφα δεδομένων. Κάθε μέλος ενός

replica set μπορεί να ενεργεί με την ιδιότητα του πρωτογενούς ή δευτερογενούς αντίγραφου ανά πάσα στιγμή. Το κύριο αντίγραφο εκτελεί όλες τις εγγραφές και τις αναγνώσεις από προεπιλογή. Τα δευτερογενή αντίγραφα διατηρούν ένα αντίγραφο των δεδομένων της πρωτογενούς χρήσης ενσωματωμένων στην αντιγραφή. Όταν ένα κύριο αντίγραφο αποτύχει, το αντίγραφο ρυθμίζεται αυτόματα να πραγματοποιεί μια εκλογική διαδικασία για να προσδιοριστεί ποια δευτερεύουσα ενέργεια πρέπει να γίνει. Δευτερεύοντα μπορεί επίσης να εκτελέσει τις λειτουργίες ανάγνωσης, αλλά τα δεδομένα είναι τελικά συνεπής από προεπιλογή.

- **Εξισορρόπηση φορτίου**

Ο χρήστης επιλέγει ένα κλειδί [7], το οποίο καθορίζει το πώς θα κατανεμηθούν τα δεδομένα σε μια συλλογή. Τα δεδομένα χωρίζονται σε σειρές (με βάση το κλειδί) και διανέμεται σε πολλαπλά θραύσματα. (Ένα θραύσμα είναι ένα master με έναν ή περισσότερους σκλάβους). Η MongoDB μπορεί να τρέξει πολλούς διακομιστές, εξισορροπώντας το φορτίο και / ή αντιγραφή δεδομένων για να κρατήσει το σύστημα σε λειτουργία σε περίπτωση αστοχίας υλικού. Αυτόματη διαμόρφωση είναι εύκολο να αναπτυχθεί, και νέες μηχανές μπορούν να προστεθούν σε μια βάση δεδομένων.

Κεφάλαιο 3: Μηχανική Μάθηση

3.1. Εισαγωγή

Η ανακάλυψη γνώσης από βάσεις δεδομένων (Knowledge Discovery in Database –KDD) αναφέρεται στη διεργασία εξόρυξης γνώσης από μεγάλες αποθήκες δεδομένων. Ο όρος εξόρυξη δεδομένων χρησιμοποιείται ως συνώνυμο της ανακάλυψης γνώσης από βάσεις δεδομένων, καθώς επίσης και για την αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και την εξαγωγή της γνώσης από διάφορα σύνολα δεδομένων. Πολλοί ερευνητές, θεωρούν τον όρο εξόρυξη δεδομένων μη αντιπροσωπευτικό της διαδικασίας στην οποία αναφέρεται, υποστηρίζοντας ότι ο όρος εξόρυξη δεδομένων θα ήταν μία πιο κατάλληλη περιγραφή. Εν τούτοις ένας τέτοιος όρος μπορεί να μην δίνει έμφαση στην ανάλυση και την εξαγωγή προτύπων από μεγάλα σύνολα δεδομένων. Ο όρος εξόρυξη δεδομένων (Data Mining) είναι αυτός που έχει επικρατήσει και χαρακτηρίζει την διαδικασία της εύρεσης δομών γνώσης οι οποίες περιγράφουν με ακρίβεια μεγάλα σύνολα πρωτόγεννων δεδομένων. Οι δομές αυτές, αναδεικνύουν γνώση (συσχετίσεις ή κανόνες) που είναι κρυμμένη μέσα στα δεδομένα και δεν μπορούν να εξαχθούν από τον άνθρωπο-χρηστή της βάσης δεδομένων με «γυμνό μάτι». Οι προκύπτουσες δομές είναι πλούσιες σε σημασιολογία και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτόγεννων δεδομένων [16].

3.2. Τεχνητή νοημοσύνη

Η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI) αποτελεί ένα από τα πιο σύγχρονα ερευνητικά πεδία. Τυπικά, έκανε την εμφάνιση της το 1956 από την πρωτοβουλία κάποιων επιφανών επιστημόνων, όπως ο John McCarthy, ο Marvin Minsky, ο Claude Shannon κ.α..

Ο Douglas Hofstadter προτείνει ότι η νοημοσύνη περιλαμβάνει τα παρακάτω στοιχεία. Να:

- Ανταποκρίνεσαι σε καταστάσεις με ελαστικότητα (όχι μηχανική συμπεριφορά).
- Κατανοείς τα ασαφή ή αντιφατικά μηνύματα από τα συμφραζόμενα.

- Αναγνωρίζεις και να ιεραρχείς τα διάφορα δεδομένα με βάση τη σπουδαιότητα τους.
- Βρίσκεις ομοιότητες μεταξύ καταστάσεων οι οποίες μοιάζουν διαφορετικές.
- Βρίσκεις διαφορές μεταξύ καταστάσεων οι οποίες μοιάζουν παρόμοιες.

Η τεχνητή νοημοσύνη έχει πολλούς και διάφορους ορισμούς, κάποιοι από τους οποίους παρουσιάζονται παρακάτω:

Ορισμός 3.1: Η Τεχνητή Νοημοσύνη είναι ο τομέας της επιστήμης των υπολογιστών, που ασχολείται με τη σχεδίαση ευφυών (νοημόνων) υπολογιστικών συστημάτων, δηλαδή συστημάτων που επιδεικνύουν χαρακτηριστικά που σχετίζουμε με τη νοημοσύνη στην ανθρώπινη συμπεριφορά [9].

Ορισμός 3.2: Η Τεχνητή Νοημοσύνη είναι ο τομέας της Επιστήμης των Υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση προγραμμάτων τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας, κλπ. [10]

Ορισμός 3.3: Η προσπάθεια να κατασκευάσουμε υπολογιστές με διανοητική ικανότητα με την πλήρη και κυριολεκτική έννοια του όρου [11].

Ορισμός 3.4: Η μελέτη του πως να κάνουμε τους υπολογιστές να εκπονούν διαδικασίες στις οποίες αυτήν τη στιγμή οι άνθρωποι είναι καλύτεροι [12]

Ορισμός3.5: Ο τομέας της επιστήμης των υπολογιστών που ασχολείται με την αυτοματοποίηση της ευφυούς συμπεριφοράς [13].

Στις μέρες μας, η πρόοδος της Τεχνητής Νοημοσύνης επιταχύνεται από την αλματώδης εξέλιξη των υπολογιστικών συστημάτων, και δημιουργεί συνεχώς νέες απαιτήσεις για τον τρόπο που αυτά πρέπει να επιλύουν προβλήματα.

3.3. Μηχανική Μάθηση

Ακόμα και στις πιο κοινές περιπτώσεις, η πρόβλεψη ενός αποτελέσματος, καθώς και η απόκτηση γνώσης, δεν μπορούν να πραγματοποιηθούν από ακατέργαστα δεδομένα. Για παράδειγμα, για τον εντοπισμό κακόβουλων μηνυμάτων ηλεκτρονικού ταχυδρομείου, η εμφάνιση μίας μεμονωμένης λέξης, μπορεί να μην είναι ιδιάζουσας

σημασίας. Όμως, η εμφάνιση λέξεων, οι οποίες χρησιμοποιούνται συνήθως ως ένα σύνολο, συναρτήσει του μήκους του κειμένου, καθώς και άλλων παραγόντων, μπορούν να προσδιορίσουν πιο συγκεκριμένα, εάν ένα μήνυμα είναι ή δεν είναι κακόβουλο. Συνεπώς, η μηχανική μάθηση (machine learning) μετατρέπει τα δεδομένα (data) σε πληροφορία.

Κύριο αντικείμενο της μηχανικής μάθησης είναι η μελέτη υπολογιστικών αλγορίθμων για την εκπαίδευση και εκτέλεση διεργασιών. Η έρευνα που έχει ήδη πραγματοποιηθεί, βασίζεται πάντα σε κάποιο είδος των παρατηρήσεων ή δεδομένων, όπως στην άμεση εμπειρία, ή εντολή. Έτσι, η μηχανική μάθηση βοηθάει στο να αντιδράσουμε καλύτερα σε μια κατάσταση στο μέλλον, βασιζόμενοι στην εμπειρία του παρελθόντος.

Η μηχανική μάθηση είναι μία βασική υποκατηγορία της τεχνητής νοημοσύνης. Παρ' όλα αυτά, σε γενικές γραμμές αλληλεπιδρά και με άλλους τομείς. Οι τομείς αυτοί μπορεί να είναι η στατιστική , αλλά και τα μαθηματικά, η φυσική, οι θεωρητικής επιστήμης των υπολογιστών[4]. Η εφαρμογή της πραγματοποιείται σε πολλούς τομείς, ακόμα και πολιτικούς, και σε ποικίλα προβλήματα. Οποιοδήποτε πεδίο, που χρειάζεται να ερμηνεύσει ή να χρησιμοποιήσει δεδομένα, μπορεί να ωφεληθεί από τις τεχνικές τις μηχανικής μάθησης.

Ιδιαίτερη έμφαση δίδεται στις αυτοματοποιημένες μεθόδους. Με άλλα λόγια, στόχος της μηχανικής μάθησης είναι να δημιουργηθούν αλγόριθμοι οι οποίοι θα αυτοματοποιούν την μάθηση, χωρίς ανθρώπινη παρέμβαση ή βοήθεια. Έτσι λοιπόν, η μηχανική μάθηση, μπορεί να θεωρηθεί ως "προγραμματισμός μέσω παραδειγμάτων". Συνεπώς, αντί να δημιουργηθεί ένα πρόγραμμα το οποίο θα επιλύει απευθείας το πρόβλημα, ο ίδιος ο υπολογιστής θα είναι εκείνος που θα λύνει το πρόβλημα μέσα από παραδείγματα που του παρέχουμε.

3.4. Ορισμοί Μηχανικής Μάθησης

Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται μοντέλο (model). Επιπλέον, έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται πρότυπα (patterns). Πέρα όμως από τον άνθρωπο, και ένα υπολογιστικό σύστημα

είναι σε θέση να εκπαιδευτεί και να μάθει. Αυτού του είδους η μάθηση είναι η μηχανική. Ως μηχανική μάθηση ορίζεται:

Ορισμός 3.5: "Η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων μέσω ενός υπολογιστικού συστήματος, ονομάζεται μηχανική μάθηση"[5]

Άλλοι ορισμοί της μηχανικής μάθησης είναι οι παρακάτω:

Ορισμός 3.6: "... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης" [8]. □

Ορισμός 3.7: "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E" [6]. □

Ορισμός 3.8: "Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον" [7].

3.5. Είδη Μηχανικής Μάθησης

Πολλές είναι οι τεχνικές μηχανικής μάθησης που έχουν αναπτυχθεί και η χρήση τους γίνεται ανάλογα με το είδος του προβλήματος. Όλες αυτές οι τεχνικές ωστόσο αντιστοιχούν σε ένα από τα δύο παρακάτω είδη [5]:

- επιβλεπόμενη μάθηση (supervised learning) ή μάθηση με παραδείγματα (learning from examples): σε αυτό το είδος υπάρχει ένα σύνολο δεδομένων, μέσω του οποίου το σύστημα πρέπει να μελετήσει μία έννοια ή μία συνάρτηση, η οποία αποτελεί ουσιαστικά, την περιγραφή ενός μοντέλου. □ □
- μη επιβλεπόμενη μάθηση (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation): □ σε αυτό το είδος μάθησης το σύστημα είναι υπεύθυνο να δημιουργήσει πρότυπα, μέσα από την ανακάλυψη συσχετίσεων ή ομάδων μεταξύ των δεδομένων ενός συνόλου. Ωστόσο, τα πρότυπα αυτά δεν είναι εξαρχής γνωστό εάν υπάρχουν, πόσα και ποια είναι**.**

3.6. Παραδείγματα προβλημάτων Μηχανικής Μάθησης

Πολλά είναι τα προβλήματα τα οποία μπορούν να βρουν λύση μέσω της μηχανικής μάθησης. Μερικά παραδείγματα τέτοιων προβλημάτων είναι τα ακόλουθα [4]:

• ανίχνευση προσώπου: στην προκειμένη φάση ο αλγόριθμος βρίσκει πρόσωπα σε εικόνες

• φιλτράρισμα ανεπιθύμητων μηνυμάτων: εντοπίζονται και κατηγοριοποιούνται τα μηνύματα e-mail ως spam ή μη-spam

• εύρεση θέματος : κατηγοριοποίηση ειδησεογραφικών άρθρων ανάλογα με το περιεχόμενο τους σε διάφορες κατηγορίες όπως πολιτική, αθλητικά, ψυχαγωγία, κ.λπ.

• ιατρική διάγνωση: διάγνωση ενός ασθενούς ως πάσχοντα ή μη πάσχοντα από κάποια ασθένεια

• κατηγοριοποίηση των πελατών: πρόβλεψη αν ένας πελάτης θα ανταποκριθεί σε κάποιο συγκεκριμένο προϊόν

• πρόγνωση του καιρού: πρόβλεψη για την βελτίωση ή επιδείνωση του καιρού τις επόμενες μέρες.

3.7. Μηχανική Μάθηση και Στατιστική

Η μηχανική μάθηση χρησιμοποιεί στατιστική. Για τους περισσότερους ανθρώπους, η στατιστική είναι ένα αντικείμενο που χρησιμοποιείται από εταιρείες που ψεύδονται σχετικά με την ποιότητα και την αποτελεσματικότητα των προϊόντων τους. Στη μηχανική μάθηση συνηθίζεται να λύνεται ένα ντετερμινιστικό πρόβλημα, και η προτεινόμενη λύση επιλύει το πρόβλημα ανά πάσα ώρα και στιγμή. Σε περίπτωση που ζητηθεί να δημιουργηθεί λογισμικό για ένα μηχάνημα αυτόματης πώλησης, θα πρέπει να είναι σε λειτουργία συνέχεια, ανεξάρτητα από την κατάθεση χρημάτων, ή από το πάτημα κουμπιών. Υπάρχουν πολλά προβλήματα που η λύση δεν είναι ντετερμινιστική. Δηλαδή, δεν γνωρίζουμε αρκετά σχετικά με το πρόβλημα ή δεν υπάρχει αρκετή υπολογιστική ισχύς για να διαμορφωθεί σωστά το πρόβλημα. Σε τέτοια προβλήματα είναι απαραίτητη η χρήση στατιστικής.

Στις κοινωνικές επιστήμες, το να είναι κάποιος σωστός το 60% του χρόνου θεωρείται επιτυχία. Εάν μπορεί να προβλεφθεί ο τρόπος που θα συμπεριφερθεί ένας άνθρωπος, το 60% του χρόνου, τότε αυτό θεωρείται επιτυχία. Πώς μπορεί να γίνει αυτό; Μήπως θα έπρεπε η πρόβλεψη να είναι συνέχεια σωστή; Σε περίπτωση που η πρόβλεψη δεν είναι πάντα σωστή, αυτό σημαίνει ότι κάνουμε λάθος;

Ας δούμε ένα παράδειγμα στο οποίο παρουσιάζεται η προβληματική κατάσταση της δυσκολίας πλήρους μοντελοποίησης ενός προβλήματος. Μήπως οι άνθρωποι δεν συμπεριφέρονται σωστά, ώστε να μεγιστοποιήσουν την προσωπική τους ευτυχία; Πως μπορεί κάποιος να προβλέψει το αποτέλεσμα των γεγονότων που αφορούν τους ανθρώπους με βάση αυτή την παραδοχή; Ίσως, είναι δύσκολο να οριστεί τι κάνει ευτυχισμένο τον κάθε άνθρωπο, επειδή αυτό το στοιχείο πιθανόν να διαφέρει σημαντικά από το ένα άτομο στο άλλο. Έτσι, ακόμη και αν οι υποθέσεις μπορεί να είναι σωστές σχετικά με το στοιχείο που μεγιστοποιεί την προσωπική ευτυχία κάθε ανθρώπου, ο ορισμός της ευτυχίας είναι πολύ περίπλοκο να μοντελοποιηθεί. Εκτός της ανθρώπινης συμπεριφοράς, υπάρχουν και πολλά άλλα παραδείγματα που δεν μπορούν να μοντελοποιηθούν ντετερμινιστικά προς το παρόν. Για τη μοντελοποίηση των προβλημάτων αυτών είναι απαραίτητη η χρήση εργαλείων της στατιστικής.

3.8. Αισθητήρες και καταιγισμός δεδομένων

Υπάρχει μία τεράστια ποσότητα δεδομένων που δημιουργήθηκε από ανθρώπους στο World Wide Web, αλλά πρόσφατα όλο και περισσότερες μη ανθρώπινες πηγές δεδομένων κάνουν την εμφάνιση τους στο διαδίκτυο. Η τεχνολογία πίσω από τους αισθητήρες δεν είναι νέα, καινοτομία όμως αποτελεί η σύνδεσή τους με το διαδίκτυο.

Το ακόλουθο είναι ένα παράδειγμα με άφθονα ελεύθερα δεδομένα, στο οποίο γίνεται εμφανής η ανάγκη ταξινόμησης των δεδομένων [123]. Το 1989, ο σεισμός Loma Prieta έπληξε τη βόρεια Καλιφόρνια, σκοτώνοντας 63 άτομα, τραυματίζοντας 3757 και αφήνοντας χιλιάδες άστεγους. Ένας σεισμός παρόμοιου μεγέθους έπληξε την Αϊτή, το 2010, σκοτώνοντας περισσότερους από 230.000 ανθρώπους. Λίγο μετά το σεισμό της Loma Prieta, δημοσιεύτηκε μια μελέτη που χρησιμοποιούσε μετρήσεις

μαγνητικών πεδίων χαμηλής συχνότητας, και ισχυρίζονται ότι προβλέπουν το σεισμό [1]. Μια σειρά από μελέτες που ακολούθησαν έδειξαν ότι η αρχική μελέτη ήταν λανθασμένη για διάφορους λόγους [2,3]. Ας υποθέσουμε ότι θέλουμε να επαναλάβουμε αυτή μελέτη και να ψάξουμε τρόπους πρόβλεψης των σεισμών, έτσι ώστε να μπορούμε να αποφύγουμε τις τρομερές συνέπειες και να πετύχουμε την καλύτερη κατανόηση του πλανήτη μας. Ποιος θα ήταν ο καλύτερος τρόπος για να πραγματοποιηθεί αυτή η μελέτη; Θα μπορούσαν υποθετικά να αγοραστούν μαγνητόμετρα και κομμάτια γης επάνω στα οποία θα τοποθετούνταν. Θα μπορούσε επίσης να ζητηθεί από τις εκάστοτε κυβερνήσεις να παρέχουν χρήματα και επίσης να διαθέσουν δωρεάν κάποια κτήματα στα οποία θα τοποθετούνταν τα μαγνητόμετρα. Όμως, πάντοτε θα υπάρχει μία αβεβαιότητα σχετικά με τα αποτελέσματα των μαγνητόμετρων και αν αυτά έχουν παραποιηθεί από ποικίλους παράγοντες. Επίσης, σημαντικό είναι να βρεθεί ένας τρόπος συλλογής των αποτελεσμάτων από τους ανθρώπους. Στην συνέχεια παρουσιάζεται μία λύση χαμηλού κόστους.

Τα κινητά τηλέφωνα ή τα smartphones είναι συσκευές με μαγνητόμετρα τριών αξόνων. Τα smartphones επίσης αποτελούντα από λειτουργικά συστήματα και οι χρήστες τους έχουν τη δυνατότητα να δημιουργήσουν και να εκτελέσουν τα δικά τους προγράμματα. Συνεπώς, με λίγες γραμμές κώδικα μπορούν να πάρουν εκατοντάδες ενδείξεις από μαγνητόμετρα ανά δευτερόλεπτο. Σε περίπτωση που οι άνθρωποι πειστούν να εγκαταστήσουν και να εκτελέσουν ένα πρόγραμμα αυτού του είδους, μπορεί να αποκτηθεί μεγάλη ποσότητα δεδομένων από τα εν λόγω μαγνητόμετρα, με πολλή μικρή επένδυση. Εκτός από τα μαγνητόμετρα, τα smartphones να έχουν ένα μεγάλο αριθμό άλλων αισθητήρων, συμπεριλαμβανομένων των γυροσκοπίων yawrate, επιταχυνσιόμετρα τριών αξόνων, αισθητήρες θερμοκρασίας, και δέκτες GPS, τα οποία μπορούν να χρησιμοποιηθούν για να υποστηρίξουν πρωτογενείς μετρήσεις. Μέσω του mobile computing και των δεδομένων που παράγονται από αισθητήρες είναι εμφανές ότι μελλοντικά θα παράγονται όλο και πιο πολλά δεδομένα.

3.9. Μελλοντική Σημασία Μηχανικής Μάθησης

Από το δεύτερο μισό του εικοστού αιώνα το μεγαλύτερο ποσοστό του εργατικού δυναμικού το οποίο απασχολείται στις ανεπτυγμένες και όχι στις

αναπτυσσόμενες χώρες, έχει μετακινηθεί από την χειρωνακτική εργασία στην εργασία η οποία βασίζεται στην γνώση. Η εργασία που βασίζεται στη γνώση κάνει αναφορά σε ορισμούς όπως «μεγιστοποίηση κέρδών», «ελαχιστοποίηση κινδύνου» και «στρατηγικό μάρκετινγκ». Το εύρος των πληροφοριών που έχουν αποκτηθεί με τη συμβολή του Internet και η δυσκολία στη διαχείριση αυτού του πλήθους μετατρέπει το είδος της εργασίας αυτής σε ιδιαιτέρως δύσκολο. Ο Hal Varian, επικεφαλής οικονομολόγος της Google, δήλωσε:

"Θα συνεχίσω να λέω ότι η πιο προκλητική δουλειά στα επόμενα δέκα χρόνια θα είναι οι στατιστικολόγοι. Οι άνθρωποι νομίζουν ότι αστειεύομαι, αλλά ποιος θα μάντευε ότι οι μηχανικοί ηλεκτρονικών υπολογιστών θα ήταν η πιο δημοφιλής δουλειά της δεκαετίας του 1990; Η ικανότητα κατοχής δεδομένων - τα οποία μπορούμε να τα καταλάβουμε , να τα επεξεργαστούμε, να κερδίσουμε αξία από αυτά, να τα απεικονίσουμε, να επικοινωνήσουμε μαζί τους - πρόκειται να είναι ένα εξαιρετικά σημαντικό προσόν κατά τις επόμενες δεκαετίες, όχι μόνο σε επαγγελματικό επίπεδο αλλά ακόμη και στο μορφωτικό επίπεδο, τόσο για τα παιδιά του δημοτικού σχολείου , όσο και για του γυμνασίου και του κολεγίου. Διότι, τώρα υπάρχουν πραγματικά ελεύθερα και πανταχού παρόντα δεδομένων. Έτσι, κάποιοι συμπληρωματικοί παράγοντες αποτελούν την ικανότητα να κατανοήσουμε τα δεδομένα καθώς και την αξία αυτών . Νομίζω ότι οι στατιστικολόγοι αποτελούν ναι μεν μέρος της προσπάθειας κατανόησης των δεδομένων, αλλά είναι απλά ένα μέρος. Θα είναι επίσης σε θέση, να απεικονίσουν τα δεδομένα, να τα κοινοποιήσουν, και να τα αξιοποιήσουν αποτελεσματικά. Πιστεύω όμως ότι αυτές οι δεξιότητες – της πρόσβασης, της κατανόησης και της επικοινωνίας - θα είναι εξαιρετικά σημαντικές. Οι διευθυντές πρέπει να είναι σε θέση να έχουν πρόσβαση και να κατανοούν τα δεδομένα μόνοι τους." —*McKinsey Quarterly,* January 2009

Με τόσο μεγάλο μέρος της οικονομικής δραστηριότητας να εξαρτάται από πληροφορίες, δεν πρέπει να χαθούμε στα δεδομένα. Η μηχανική μάθηση θα βοηθήσει να έχουμε πρόσβαση στα δεδομένα και να εξορύξουμε πληροφορία [123].

3.10. Διαδικασία Κατηγοριοποίησης

Για τον εντοπισμό του συναισθήματος μέσα σε κείμενα, παραγράφους, προτάσεις ή φράσεις, χρησιμοποιούνται τεχνικές, οι οποίες έχουν σαν βάση τους την μηχανική μάθηση και κατά συνέπεια την τεχνητή νοημοσύνη. Η επιλογή της κατάλληλης τεχνικής ποικίλει ανάλογα με την μορφή του προβλήματος .

Η επιβλεπόμενη μάθηση βασίζεται στην έννοια της κατηγοριοποίησης (classification) των δεδομένων εισόδου. Η κατηγοριοποίηση αποτελεί μία βασική εργασία στη διαδικασία εξόρυξης γνώσης, που έχει ως στόχο την ανάθεση ενός στοιχείου σε ένα προκαθορισμένο σύνολο κατηγοριών (classes). Η κατηγοριοποίηση λοιπόν μπορεί να περιγραφεί ως μία λειτουργία που αντιστοιχεί (κατηγοριοποιεί) το στοιχείο σε μία από τις διαφορετικές κατηγορίες που έχουν προκαθοριστεί [14].

Απαραίτητα στοιχεία της κατηγοριοποίησης είναι ένα καλά καθορισμένο σύνολο κατηγοριών, καθώς και ένα σύνολο από προκατηγοριοποιημένα (pre-classified) παραδείγματα. Συνεπώς, στόχος της διαδικασίας αυτής είναι η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων, που η κατηγοριοποίηση του τους είναι άγνωστη.
Συγκεκριμένα, η διαδικασία της κατηγοριοποίησης δεδομένων αποτελείται από τα παρακάτω 2 βήματα:

1. **Εκμάθηση (Learning)**

Σε αυτό το βήμα δημιουργείται ένα μοντέλο (model), περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Ο αλγόριθμος κατηγοριοποίησης αναλύει τα δεδομένα εκπαίδευσης (training data) , έτσι ώστε να κατασκευαστεί στη συνέχεια το μοντέλο. Τα στοιχεία που αποτελούν το σύνολο κατάρτισης επιλέγονται τυχαία από έναν πληθυσμό δεδομένων και ανήκουν σε μία από τις προκαθορισμένες κατηγορίες. Το μοντέλο που ορίζεται, είναι γνωστό και ως κατηγοριοποιητής (classifier), και αναπαριστάται με την μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων απόφασης (decision trees), ή μαθηματικών τύπων (mathematical formulas) [15].

2. Κατηγοριοποίηση (Classification)

Η κατηγοριοποίηση (classification) αποτελεί μία από τις βασικές εργασίες (tasks) εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένο) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαρίστανται γενικά από εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποια από τις προκαθορισμένες κατηγορίες.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιεί δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί (αναθέτει σε κάποια από τις κατηγορίες) [14].

Στις περισσότερες περιπτώσεις υπάρχει ένας περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε ένα στοιχείο που ανήκει στο σύνολο των δεδομένων στην κατάλληλη κατηγορία. Για το σκοπό αυτό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες . Η πρώτη χρησιμοποιεί δένδρα αποφάσεων (Decision Trees) και η δεύτερη Νευρωνικά δίκτυα (Neural Networks) [16, 17]. Και οι δύο στηρίζονται στην ιδέα της εκπαίδευσης (training), με την βοήθεια ενός υποσυνόλου δεδομένων που ονομάζονται σύνολο εκπαίδευσης (training set). Το υποσύνολο αυτό επιλέγεται σαν αντιπροσωπευτικό δείγμα του συνολικού όγκου δεδομένων. Έτσι, όταν προκύψει ένα νέο στοιχείο μπορεί εύκολα να κατηγοριοποιηθεί. Για τη διαδικασία αυτή χρησιμοποιούνται είτε τεχνικές βασισμένες σε νευρωνικά δίκτυα είτε συμβολικές τεχνικές.

Στην προκειμένη στιγμή, χρησιμοποιούνται τα δοκιμαστικά δεδομένα (test data), έτσι ώστε να προβλέψουν την ακρίβεια (accuracy) του μοντέλου. Οι μέθοδοι που χρησιμοποιούνται για να υπολογισθεί η ακρίβεια του κατηγοριοποιητή (classifier) είναι διάφορες. Η επιλογή των δεδομένων εκπαίδευσης γίνεται τυχαία και τα δεδομένα αυτά είναι ανεξάρτητα. Το μοντέλο κατηγοριοποιεί κάθε ένα από τα δοκιμαστικά παραδείγματα (training samples). Στη συνέχεια, η κατηγορία που ανήκουν τα δεδομένα με βάση το σύνολο δοκιμαστικών δεδομένων συγκρίνεται με την πρόβλεψη που έκανε το μοντέλο για την κατηγορία. Η ακρίβεια του μοντέλου σε

ένα καθορισμένο σύνολο δεδομένων δοκιμής είναι το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά από το υπό εκπαίδευση μοντέλο[16].

Σε περίπτωση που η ακρίβεια του μοντέλου θεωρείται ως αποδεκτή, τότε το μοντέλο μπορεί να χρησιμοποιηθεί έτσι ώστε να κατηγοριοποιεί μελλοντικά δείγματα δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη[16].

3.11. Εκπαίδευση Συστήματος

Όπως αναφέρθηκε και παραπάνω, διακρίνονται δύο διαφορετικές φάσεις. Οι φάσεις αυτές είναι η εκμάθηση και η κατηγοριοποίηση. Αυτές οι φάσεις περιέχουν κάποια βήματα, τα οποία θα παρουσιάσουμε αναλυτικά παρακάτω [125].

3.11.1. Σύνολο εκπαίδευσης (Training set)

Το σύνολο εκπαίδευσης (training set) είναι ένα σύνολο το οποίο δίδεται σαν είσοδος στον ταξινομητή, για παράδειγμα ένα σύνολο από Tweets.. το σύνολο αυτό από τα Tweets, ονομάζεται corpus. Έτσι είναι σημαντικό να υπάρχει ένα αντιπροσωπευτικό σύνολο εκπαίδευσης, δηλαδή ένα σύνολο το οποίο υπακούει στο νόμο του Zipf (Zipf'w law). Σύμφωνα με τον Zipf, δεδομένου ενός corpus, η συχνότητα μίας λέξης είναι αντιστρόφως ανάλογη της τάξης (rank) που κατέχει στον πίνακα συχνοτήτων [124].

Σε κάθε ένα από τα Tweets δίδεται μία αξία. Για παράδειγμα, κάθε tweet που ανήκει στο corpus, χαρακτηρίζεται ως αρνητικό (negative), ουδέτερο (neutral), ή θετικό (positive). Αυτόν τον χαρακτηρισμό ίσως το συναντήσουμε με τις τιμές 0, 2, 4 ή -1, 0,1 αντίστοιχα. Η αξία σε κάθε tweet μπορεί να δοθεί χειρωνακτικά, όμως, λόγω του ιδιαιτέρως μεγάλου όγκου των δεδομένων, η εν λόγω διαδικασία είναι σχεδόν ανέφικτη. Επίσης, ο χαρακτηρισμός μίας λέξης ως αρνητικής ή θετικής, είναι εντελώς υποκειμενικός. Για αυτό το σκοπό, δημιουργήθηκαν lexicons, τα οποία περιέχουν την συναισθηματική αξία κάθε λέξης, όπως για παράδειγμα το SentiWordNet .

3.11.2. Προεπεξεργασία

Κάθε tweet περιλαμβάνει λιγότερους ή ίσους με 140 χαρακτήρες. Οι χαρακτήρες αυτοί για τον υπολογιστή δεν είναι τίποτα παραπάνω από ακολουθίες συμβόλων. Δηλαδή δεν μπορεί ο υπολογιστής να καταλάβει τις λέξεις ή την αξία τους. Επιπλέον, πέρα από λέξεις μέσα στα tweets, περιέχονται hashtags, emoticons και αναφορές σε άλλους χρήστες.

Πριν από την εκπαίδευση του ταξινομητή είναι απαραίτητο να καθαριστούν τα tweets από τον θόρυβο με την βοήθεια κανονικών εκφράσεων (regular expressions). Έπειτα, χρησιμοποιούμε το απλό κείμενο που προκύπτει και πραγματοποιούμε την διαδικασία του “tokenization”. Token ουσιαστικά είναι μία μονάδα ή διαφορετικά μία μεταβλητή. Ένα token ορίζεται ως μία μεταβλητή από συνεχόμενους χαρακτήρες ή νούμερα. Τα διαφορετικά token χωρίζονται από κενά (whitespaces) ή σημεία στίξης. Πολλές είναι οι NLP τεχνικές που μπορούν να χρησιμοποιηθούν μετά την παραπάνω διαδικασία.

3.11.3. Features

Τα features είναι χαρακτηριστικά στα οποία βασίζεται ένας ταξινομητής για να πάρει μία απόφαση, δηλαδή είναι τα χαρακτηριστικά εκείνα που βοηθούν να αναγνωριστεί μία οντότητα. Σε ταξινομητές όπως ο Naive Bayes, τα features πρέπει να έχουν δυαδική μορφή, ώστε να κρίνεται αν ένα tweet έχει ή δεν έχει το συγκεκριμένο feature. Έστω ότι φτιάχνεται μία λίστα με λέξεις από ένα σύνολο εκπαίδευσης. Ο ταξινομητής θα ελέγξει ποιες από όλες τις πιθανές λέξεις του συνόλου εκπαίδευσης εμφανίζονται ή όχι μέσα σε ένα συγκεκριμένο tweet (unigram feature). Το Unigram μοντέλο έχει ως στόχο την ανάκτηση πληροφοριών. Ο κάθε όρος – λέξη που υπάρχει μέσα σε ένα συγκεκριμένο περιεχόμενο εξετάζεται χωριστά. Έτσι, πρακτικά θα λέγαμε ότι ο υπολογισμός της πιθανότητας σε στα unigrams μπορεί να γίνει ως εξής:

$$P_{uni}(t_1t_2t_3) = P(t_1)P(t_2)P(t_3)$$

Σε αυτό το μοντέλο η πιθανότητα ύπαρξης μίας λέξης εξαρτάται από την ίδια τη λέξη. Δηλαδή, ποια είναι η πιθανότητα εμφάνισης της λέξης αυτής μέσα στο

συγκεκριμένο κείμενο. Το σύνολο των πιθανοτήτων όλων των λέξεων πρέπει να ισούται με 1 [126].

Σε ένα σύνολο από tweets ο αριθμός των πιθανών features είναι πάρα πολύ μεγάλος. Συνεπώς, είναι απαραίτητο να επιλεγούν με κάποιο μέτρο ποιότητας τα features που περιγράφουν καλύτερα το κείμενο.

.

3.11.4. Κατηγοριοποιητής - Naïve Bayes Classifier

Ο Naïve Bayes Classifier αποτελεί μία μορφή απλού ταξινομητή που βασίζεται στον κανόνα του Bayes, σε συνδυασμό με την υπόθεση της ανεξαρτησίας

Το μοντέλο αυτό περιλαμβάνει μία παραδοχή ανεξαρτησίας των δεδομένων υπό όρους. Δίνεται μία κατηγορία (θετική ή αρνητική), οι λέξεις είναι υπό όρους ανεξάρτητες η μία από την άλλη. Η παραδοχή αυτή δεν επηρεάζει την ακρίβεια της ταξινόμησης κειμένου ιδιαιτέρως, αντιθέτως βελτιώνει σημαντικά την ταχύτητα των ταξηνομητών [17]. Στόχος του αλγορίθμου είναι να επιλεγεί μία ετικέτα για μία είσοδο (prosterior probability), όταν η εκ των προτέρων πιθανότητα (prior probability) εμφάνισης της ετικέτας αυτής είναι γνωστή. Από τη συνεισφορά του κάθε feature, υπολογίζεται η πιθανότητα μία είσοδος να λάβει μία συγκεκριμένη ετικέτα ή άλλη.

Σε πιο μαθηματικούς όρους, στόχος είναι να κατηγοριοποιηθεί ένα δείγμα X, σε μία από τις δεδομένες κατηγορίες C_1, C_2,...C_n,, χρησιμοποιώντας ένα μοντέλο πιθανότητας, το οποίο ορίζεται σύμφωνα με την θεωρία του Bayes [18]. Επιπλέον, υποθέτουμε ότι υπάρχει ένα σύνολο από στοιχεία – εγγραφές X, που ανήκουν σε μία κλάση Ci. Συναρτήσει των παραπάνω ορισμών αλλά και λαμβάνοντας υπόψη την θεωρία του Bayes, ορίζεται η εκ των υστέρων πιθανότητα (posterior probability):

Όπου:

$$p(X/Ci) \in [0,1]$$

Η εκ των προτέρων πιθανότητα (prior probability) είναι αυτή που χαρακτηρίζει κάθε κατηγορία της κλάσης Ci.

Ο διασημότερος και απλούστερος κατηγοριοποιητής είναι ο Naïve Bayes Classifier. Με την παραδοχή της υπο συνθήκης ανεξαρτησίας (conditional independence) μπορούν να απλοποιηθούν οι υπολογισμοί [16].

Έστω λοιπόν, ότι έχουμε ένα σύνολο δεδομένων S και έστω ότι κάθε δείγμα δεδομένων αντιπροσωπεύεται από το n-διάστατο χαρακτηριστικό διάνυσμα X, όπου $X = (x_1, x_2,.....x_n)$. Το διάνυσμα αυτό απεικονίζει τις μετρήσεις που πραγματοποιούνται στο παραπάνω διάνυσμα για τα n γνωρίσματα $A_1, A_2.....A_n$.

Υποθέτουμε ότι υπάρχει ένας αριθμός m κατηγοριών $C_1, C_2,….C_m$. Ο κατηγοριοποιητής θα προβλέψει ότι το X, ανήκει στην κατηγορία, που έχει την υψηλότερη εκ των υστέρων πιθανότητα δεδομένου του X. Με μαθηματικά σύμβολα θα λέγαμε ότι, ο Naïve Bayes Classifier υπολογίζει τις υπό συνθήκη πιθανότητες της κατηγορίας C, λαμβάνοντας ως δεδομένο την υπό-συνθήκη ανεξαρτησία. Έτσι τελικώς υποθέτουμε ότι [16]:

$$p(Ci\,|X) > p(Ci\,|X) for\ 1 \leq m, j \neq i$$

Στόχος του εν λόγω κατηγοριοποιητή είναι η μεγιστοποίηση της εκ των υστέρων υπόθεσης (posterior hypothesis). Πραγματοποιώντας την υπόθεση της υπό συνθήκη ανεξαρτησίας (conditional independence) υπολογίζει τις υπό συνθήκη πιθανότητες της κάθε κατηγορίας. Έπειτα, υποθέτουμε ότι:

$$p(X|C_i) = p(x_1|C_i)\ldots p(x_n|C_i)$$

Όλες οι παραπάνω πιθανότητες, δηλαδή $p(x_j|C_i)$, μπορούν να υπολογιστούν από τα δεδομένα εκπαίδευσης.

Ο Naïve Bayesian κατηγοριοποιητής αποτελεί μία πολύ αποδοτική τεχνική. Παραδοσιακά, οι Bayesian classifiers συγκριτικά με άλλους, κατέχουν ελάχιστο ποσοστό σφάλματος. Πρακτικά όμως το γεγονός αυτό είναι εν μέρει φαινομενικό, λόγο των υποθέσεων που απαιτούνται, όπως για παράδειγμα η υπό συνθήκη ανεξαρτησία. Πέρα όμως από το ποσοστό λάθους, οι κατηγοριοποιητές αυτοί ομοιάζουν με εκείνους που βασίζονται στα νευρωνικά δίκτυα.

3.12. Πλεονεκτήματα Naïve Bayes Classifier

Ο συγκεκριμένος classifier εμφανίζει μία σειρά από πλεονεκτήματα, τα οποία παρουσιάζονται παρακάτω [90]:

- Εύκολη εκπαίδευση και εφαρμογή. Από τη φύση το είναι ιδιαίτερα ευέλικτος. Επιπλέον, η διαδικασία της μάθησης από ταξινομητή είναι γρήγορη σε σύγκριση με άλλες μεθόδους μάθησης, όπως η λογιστική παλινδρόμηση.
- Γρηγορότερη Σύγκλιση. Έστω ότι, εφαρμοστεί η υπό όρους ανεξαρτησία του Naïve Bayes τότε ο ταξινομητής θα συγκλίνει πολύ πιο γρήγορα από τις άλλες εκπαιδευτικές μεθόδους.
- Καλή απόδοση. Έστω ότι η υπό όρους ανεξαρτησία δεν εφαρμοστεί, ωστόσο ο ταξινομητής συνεχίζει να αποδίδει εξίσου καλά.
- Ακρίβεια πρόβλεψης ή Απόρριψη. Ο ταξινομητής έχει τη δυνατότητα κατανομής πιθανότητας, δίνοντας έτσι την ακρίβεια της πρόβλεψης. Εάν η ακρίβεια δεν είναι αποδεκτή, τότε η πρόβλεψη μπορεί να απορριφθεί ή να αγνοηθεί.
- Αντιστάθμιση Ανισορροπίας Πρόβλεψης. Μπορεί να αντισταθμιστεί η ανισορροπία που παρουσιάζεται στις κλάσεις, όπου μία ή περισσότερες από τις περιπτώσεις συμβαίνουν πολύ σπάνια (1/1000). Αυτή λοιπόν η ανισορροπία που παρατηρείται στις κλάσεις πιθανόν να οδηγήσει σε λανθασμένες και εσφαλμένες προβλέψεις και ονομάζεται εκφυλισμένη λύση. Αυτό το πρόβλημα μπορεί να ξεπεραστεί με τη χρήση ενός ισορροπημένου συνόλου εκπαίδευσης.
- Ακρίβεια λειτουργίας. Λειτουργεί με εξαιρετική ακρίβεια, όταν είναι μεγάλο το σύνολο δεδομένων εκπαίδευσης.
- Γραμμική Πολυπλοκότητα Χρόνου Εκπαίδευσης. Η πολυπλοκότητα του χρόνου εκπαίδευσης του συγκεκριμένου ταξινομητή είναι γραμμική με τον αριθμό των δεδομένων εκπαίδευσης και η πολυπλοκότητα του χώρου είναι επίσης γραμμική με τον αριθμό των features. Συνεπώς, η συγκεκριμένη τεχνική μάθησης είναι αποδοτική, τόσο ως προς τον χρόνο, όσο και ως προς την αποθήκευση δεδομένων.

3.13. Χρήσεις Naïve Bayes Classifier

Οι χρήσεις του Naïve Bayes Classifier ποικίλουν και παρουσιάζονται αναλυτικότερα στη συνέχεια [91]:

- **Ταξινόμηση κειμένου (Text classification)**

Η Bayesian ταξινόμηση χρησιμοποιείται ως μέθοδος πιθανολογικής εκμάθησης (Naive Bayes ταξινόμηση κειμένου). Ο Naive Bayes ταξινομητής είναι μεταξύ των πιο επιτυχημένων και γνωστών αλγορίθμων μάθησης, που έχουν σαν στόχο την ταξινόμηση κειμένου.

- **Φιλτράρισμα Ανεπιθύμητης Αλληλογραφίας (Spam filtering)**

Το spam filtering αποτελεί την πιο γνωστή χρήση του Naive Bayesian ταξινομητή κειμένου. Κάνει χρήση ενός απλοϊκού ταξινομητή Bayes για τον εντοπισμό spam e-mails. Το Bayesian φιλτράρισμα ανεπιθύμητων μηνυμάτων έχει γίνει ένας δημοφιλής μηχανισμός για τη διάκριση παράνομων spam e-mails από το νόμιμα e-mails. Πολλοί σύγχρονοι πελάτες ηλεκτρονικού ταχυδρομείου εφαρμόζουν το Bayesian φιλτράρισμα ανεπιθύμητων μηνυμάτων.

- **Υβριδικό σύστημα συστάσεων (Hybrid Recommender System)**

Το σύστημα αυτό χρησιμοποιεί τον απλοϊκό ταξινομητή Bayes και το συνεργατικό φιλτράρισμα (Naive Bayes Classifier and Collaborative Filtering).
Εφαρμόζει τεχνικές μηχανικής μάθησης και εξόρυξης δεδομένων για το φιλτράρισμα αόρατων πληροφοριών και μπορεί να προβλέψει αν ένας χρήστης θα μπορούσε να εμφανίσει προτίμιση σε έναν συγκεκριμένο πόρο.

Κεφάλαιο 4: Συναισθηματική Ανάλυση

4.1. Εισαγωγή

Η άποψη των υπολοίπων αποτελούσε και συνεχίζει να αποτελεί έναν σημαντικό παράγοντα σχετικά με την λήψη αποφάσεων. Πριν από την ανάπτυξη του διαδικτύου, ο κόσμος ζητούσε την άποψη γνωστών και φίλων σχετικά με διάφορες αγορές που επρόκειτο να πραγματοποιήσει, ή αναρωτιόταν για το τι θα ψηφίσει ο συνάνθρωπος του στις εκλογές. Με την ανάπτυξη και την ευρεία χρήση του Internet, δόθηκε η δυνατότητα να συλλέγονται οι απόψεις όχι μόνο λίγων γνωστών, αλλά μίας ευρείας γκάμας ανθρώπων διαφόρων ηλικιών, κοινωνικών καταστάσεων, ακόμα και εθνικοτήτων. Με την χρήση του διαδικτύου όλο και περισσότεροι άνθρωποι διαθέτουν τις απόψεις τους δημόσια, σε αγνώστους.

Σύμφωνα με δύο έρευνες οι οποίες πραγματοποιήθηκαν πρόσφατα [19, 20]:

- το 81% των χρηστών του Internet (ή το 60% των Αμερικανών) έχουν κάνει έρευνα στο διαδίκτυο για ένα προϊόν τουλάχιστον μία φορά,
- το 20% (15% του συνόλου των Αμερικανών) έκαναν αυτή την πράξη μία τυπική ημέρα,
- από τους αναγνώστες των κριτικών ένα ποσοστό μεταξύ 73% και 87% ανέφερε ότι οι εκθέσεις που διάβασαν είχαν θετική επίδραση στην έρευνα τους,
- οι καταναλωτές αναφέρουν ότι είναι πρόθυμοι να πληρώσουν από 20% έως 99% περισσότερο για κάποιο προϊόν το οποίο έχει λάβει βαθμολογία 5 αστέρων έναντι ενός άλλου που έχει πάρει 4 αστέρια,
- το 32% των χρηστών έχουν δώσει μία βαθμολογία για ένα προϊόν, υπηρεσία ή πρόσωπο μέσω ενός online συστήματος αξιολόγησης, και το 30% (συμπεριλαμβανομένου του 18% των online ηλικιωμένων πολιτών) έχουν δημοσιεύσει ένα online σχόλιο ή κριτική για ένα προϊόν ή μια υπηρεσία

Πέρα από την κατανάλωση των αγαθών ένας άλλος τομέας για τον οποίο οι χρήστες σπεύδουν σε αναζήτηση στο διαδίκτυο είναι η πολιτική ενημέρωση. Για παράδειγμα, σε μια έρευνα πάνω από 2500 Αμερικανούς ενήλικες, οι Rainie και

Horrigan [21] κατέληξαν στο συμπέρασμα ότι το 31% των Αμερικανών - πάνω από 60 εκατομμύρια άνθρωποι - που ήταν το 2006 οι χρήστες του Διαδικτύου, συγκέντρωσαν πληροφορίες σχετικά με τις εκλογές του ιδίου έτους και αντάλλαξαν απόψεις μέσω του ηλεκτρονικού ταχυδρομείου. Πιο συγκεκριμένα,

- το 28% των χρηστών δήλωσε ότι ένας σημαντικός λόγος για τέτοιου είδους δραστηριότητες ήταν να ενημερωθεί για τις επικείμενες επιλογές εντός του κοινωνικού τους περίγυρου και το 34% για τις επιλογές εκτός αυτού
- 28% δήλωσε ότι τα περισσότερα από τα sites που χρησιμοποίησαν μοιράζονται την άποψή τους, από την άλλη το 29% δήλωσε ότι τα περισσότερα από τα sites που χρησιμοποιούν αμφισβητούν την άποψή τους,
- το 8% δημοσίευσε το δικό του πολιτικό σχόλιο real-time.

Ο Horrigan [20] αναφέρει ότι ενώ η πλειοψηφία των Αμερικανών χρηστών του Διαδικτύου αναφέρουν θετικές εμπειρίες κατά τη διάρκεια της έρευνας σε πραγματικό χρόνο. Όμως την ίδια στιγμή το 58% των χρηστών αναφέρουν, ότι ψάχνοντας πληροφορίες σε πραγματικό χρόνο, παρατήρησαν ελλείψεις και σύγχυση. Συνεπώς, εμφανίζεται μία σαφής ανάγκη για δημιουργία συστημάτων που θα παρέχουν καλύτερη πληροφόρηση σε πραγματικό χρόνο.

Το ενδιαφέρον που δείχνουν οι μεμονωμένοι χρήστες σε άντληση απόψεων σε πραγματικό χρόνο σχετικά με τα προϊόντα και τις υπηρεσίες, και η πιθανή επίδραση αυτών των γνωμοδοτήσεων, είναι κάτι που οι πωλητές των ειδών δίνουν όλο και περισσότερη προσοχή [22]. Παρακάτω παρατίθεται ένα απόσπασμα από τους Zabin and Jefferies [23]:

“Με την έκρηξη του Web 2.0 εργαλείων, όπως τα blogs, φόρουμ συζητήσεων, peer-to-peer δίκτυα, και διάφορα άλλα είδη των μέσων κοινωνικής δικτύωσης, οι καταναλωτές έχουν στη διάθεσή τους μεγάλη δύναμη και μπορούν να μοιραστούν τις εμπειρίες και τις απόψεις, θετικές ή αρνητικές, για οποιοδήποτε προϊόν ή υπηρεσία.... Οι εταιρείες μπορούν να ανταποκριθούν στις γνώμες των καταναλωτών που δημιουργούνται μέσω των κοινωνικών δικτύων και ανάλυση των μέσων ενημέρωσης, και να αναλύσουν τις γνώμες αυτές λαμβάνοντας μηνύματα σχετικά με την πολιτική τους, την ανάπτυξη προϊόντων, και άλλων δραστηριοτήτων ανάλογα.”

Ωστόσο, οι αναλυτές της βιομηχανίας συμειώνουν ότι η μόχλευση των νέων μέσων με στόχο την ανίχνευση της εικόνας ενός προϊόντος ως προς τους

καταναλωτές απαιτεί νέες τεχνολογίες. Παρακάτω παρατίθεται ένα απόσπασμα σχετικά με αυτές τις απαιτήσεις [24]:

"Οι έμποροι χρειάζεται πάντα να κινητοποιούν τα μέσα ενημέρωσης, έτσι ώστε να λαμβάνουν πληροφορίες σχετικά με τα προϊόντα τους- είτε πρόκειται για δραστηριότητες δημοσίων σχέσεων, είτε για ανταγωνιστική νοημοσύνη. Ο κατακερματισμό των μέσων ενημέρωσης καθώς και η αλλαγή της συμπεριφοράς των καταναλωτών έχουν παραλύσει τις παραδοσιακές μεθόδους παρακολούθησης. Η Technorati εκτιμά ότι 75.000 νέα blogs δημιουργούνται καθημερινά, μαζί με 1,2 εκατομμύρια νέες δημοσιεύσεις. Επίσης, πολλές είναι οι συζητήσεις ανταλλαγής απόψεων των καταναλωτών σχετικά με τα προϊόντα και τις υπηρεσίες."

Έτσι, ένα πρόσθετο κοινό για συστήματα τα οποία αναλύουν αυτόματα το συναίσθημα του καταναλωτή, είναι εταιρείες που έχουν στόχο να κατανοήσουν με ποιον τρόπο αντιμετωπίζονται τα προϊόντα τους από το ευρύ κοινό.

4.2. Ιστορική Αναδρομή

Το τελευταίο διάστημα η συναισθηματική ανάλυση και η εξόρυξη γνώσης γνωρίζουν ιδιαίτερη άνθιση σε αντίθεση με παλαιότερα. Στα πρωταρχικά έργα μελετήθηκαν κυρίως οι πεποιθήσεις των ανθρώπων [25,26]. Στις μετέπειτα εργασίες το ενδιαφέρον επικεντρώθηκε κυρίως στην ερμηνεία της μεταφοράς, της αφήγησης, της άποψης και της επιρροής του κειμένου , και σε άλλους συναφείς τομείς [27, 28, 29, 30, 31, 32, 33, 34, 35]

Το έτος 2001 φαίνεται ότι είναι αυτό που σηματοδοτεί την έναρξη της ευρύτερης ευαισθητοποίησης σχετικά με τα ερευνητικά θέματα και τις ευκαιρίες που προσφέρει η συναισθηματική ανάλυση και η εξαγωγή απόψεων (opinion mining) [36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47 , 48]. Από τότε και στο εξής, υπήρξαν κυριολεκτικά εκατοντάδες εργασίες που δημοσιεύθηκαν σχετικά με αυτό το θέμα.

Οι παράγοντες που συνέβαλλαν την έναρξη αυτών των ερευνών είναι [49]:

- η ανάπτυξη των μεθόδων μηχανικής μάθησης για την επεξεργασία της φυσικής γλώσσας και την ανάκτηση πληροφοριών,

- η διαθεσιμότητα των συνόλων δεδομένων (datasets), μέσω των οποίων οι αλγόριθμοι μηχανικής μάθησης θα μπορούσαν να εκπαιδευτούν. Η διαθεσιμότητα αυτή οφείλεται στην άνθιση του διαδικτύου και ειδικότερα, στην ανάπτυξη ιστοσελίδων μέσω των οποίων οι χρήστες μπορούν να εκθέσουν την άποψη τους σχετικά με διάφορα ζητήματα, τέλος
- η υλοποίηση πνευματικών προκλήσεων, εμπορικών και κατασκοπευτικών εφαρμογών που προσέφερε το συγκεκριμένο πεδίο ερευνών.

4.3. Ορολογία

Η εξαγωγή απόψεων, η συναισθηματική ανάλυση και η υποκειμενική ανάλυση, ασχολούνται αντίστοιχα με την γνώμη, το συναίσθημα και την υποκειμενικότητα. Το 1994, ο Wiebe [51], επηρεασμένος από τα γραπτά του Μπάνφιλντ [50], εστίασε στην ιδέα της υποκειμενικότητας, που ορίστηκε από τους Quirk et al. [52], και επισήμανε ότι δεν υπάρχουν ανοιχτά πεδία στον τομέα της αντικειμενικής θεώρησης ή της επαλήθευσης. Ένα κανονικό παράδειγμα αυτής της έρευνας, που περιγράφεται συνήθως ως ένας τύπος της υποκειμενικής ανάλυσης είναι η αναγνώριση της γνωμο-κεντρικής γλώσσας, έτσι ώστε να το διακρίνεται από την αντικειμενική γλώσσα.

Ο όρος εξόρυξη γνώμης εμφανίζεται σε μία εργασία των Dave et al.. [53], που δημοσιεύτηκε στα πρακτικά του συνεδρίου WWW 2003. Σύμφωνα με τους Dave et al. [53], το ιδανικό εργαλείο εξόρυξης γνώμης «θα επεξεργαστεί ένα σύνολο αποτελεσμάτων αναζήτησης για ένα δεδομένο αντικείμενο, δημιουργώντας μία λίστα χαρακτηριστικών του προϊόντος (ποιότητα, τα χαρακτηριστικά, κλπ) και τη συγκέντρωση γνωμών για κάθε ένα από αυτά τα χαρακτηριστικά (φτωχός, καλός, κ.α)".

Η ιστορία της συναισθηματικής ανάλυσης παραλληλίζεται με εκείνη της εξόρυξης γνώμης. Ο όρος συναίσθημα (sentiment) που χρησιμοποιείται αναφορικά με την αυτόματη ανάλυση του κειμένου και την άντληση προσοδοφόρων κριτικών, παρουσιάστηκε για πρώτη φορά το 2001 σε έρευνες των Das και Chen [37] και Tong [45], οι οποίοι έδειξαν ιδιαίτερο ενδιαφέρον στο να αναλύσουν το συναίσθημα της αγοράς. Στη συνέχεια, παρόμοιες έρευνες πραγματοποιήθηκαν από τους Turney [46]

και Pang et al. [43], που δημοσιεύτηκαν στα πρακτικά της ετήσιας συνάντησης της Ένωσης Υπολογιστικής Γλωσσολογίας (ACL) και της ετήσιας διάσκεψης για εμπειρικές μεθόδους στην επεξεργασία φυσικής γλώσσας (EMNLP) αντίστοιχα. Επιπλέον, οι Nasukawa και Yi [42] το 2003 πραγματοποίησαν έρευνα με τίτλο "Sentiment analysis: Capturing favorability using natural language processing", και το ίδιο έτος οι Yi et al. [54] δημοσίευσαν έρευνα που ονομάστηκε ""Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques". Όλα αυτά τα στοιχεία μαζί εξηγούν την άνθιση της "συναισθηματικής ανάλυσης" και την εστίαση στο NLP. Ένας σημαντικός αριθμός των ερευνών αναφέρουν ότι η "συναισθηματική ανάλυση " εστιάζει σε συγκεκριμένες εφαρμογές ταξινόμησης (classification) κριτικών ως προς την πολικότητα (θετική ή αρνητική). Αυτό το γεγονός , φαίνεται πως έχει ωθήσει κάποιους ερευνητές να προτείνουν ότι ο όρος "συναισθηματική ανάλυση" αναφέρεται ειδικά στην πολικότητα. Παρόλα αυτά, στις μέρες μας οι έρευνες γενικεύονται, παρουσιάζοντας μία ευρύτερη έννοια του όρου, σχετικά με την υπολογιστική επεξεργασία των απόψεων, του συναισθήματος και της υποκειμενικότητας ενός κειμένου[49].

Έτσι, με την ευρεία έννοια η "συναισθηματική ανάλυση" και η "εξόρυξη γνώμης" ανήκουν στο ίδιο πεδίο ερευνών (το οποίο μπορεί να θεωρηθεί υποκατηγορία της ανάλυσης της υποκειμενικότητας). Τέλος καλό θα ήταν να ορίσουμε τις έννοιες [49] γνώμη (opinion), οπτική (view), πεποίθηση (belief), κοσμοθεωρία (conviction), πειθώ (persuasion), συναίσθημα (sentiment):

Ορισμός 4.1: Η γνώμη (opinion) αφορά ένα συγκεντρωτικό συμπέρασμα, το οποίο έχει ως στόχο να αμφισβητεί κάθε εμπειρογνώμονα, ο οποίος έχει διαφορετική άποψη.

Ορισμός 4.2: Η οπτική (view) αναφέρεται σε μία υποκειμενική γνώμη

Ορισμός 4.3: Η πεποίθηση (belief) συχνά συνεπάγεται την εκ προθέσεως αποδοχή και την σύμφωνη γνώμη.

Ορισμός 4.4: Η κοσμοθεωρία (conviction) αναφέρεται στην σταθερή και σοβαρή πεποίθηση (πχ. Στην πεποίθηση ότι η ζωή ενός ζώου είναι το ίδιο σημαντική με αυτή του ανθρώπου)

Ορισμός 4.5: Η πειθώ (persuasion) προτείνει μία πεποίθηση η οποία βασίζεται στην θεωρία της αξιοπιστίας μέσω αποδεικτικών στοιχείων

Ορισμός 4.6: Το συναίσθημα (sentiment) προτείνει μία πάγια άποψη, που αντικατοπτρίζει το πως αισθάνεται κάποιος, το πως νιώθει.

Στη συνέχεια αναφέρουμε τον ορισμό της συναισθηματικής ανάλυσης:

Ορισμός 4.7: Συναισθηματική ανάλυση (sentiment analysis) είναι η διαδικασία κατά την οποία ορίζεται πότε ένα κομμάτι ενός εγγράφου είναι θετικό αρνητικό ή ουδέτερο [127].

4.4. Επίπεδα συναισθηματικής ανάλυσης

Η συναισθηματική ανάλυση έχει μελετηθεί κυρίως σε τρία επίπεδα [57]:

- **Επίπεδο κειμένου (Document level):** Το ζητούμενο σε αυτό το επίπεδο είναι να ξεκαθαριστεί εάν το συναίσθημα που εκφράζεται σε ένα συγκεκριμένο κείμενο είναι θετικό ή αρνητικό [43, 46]. Για παράδειγμα, δεδομένης μίας έκθεσης κριτικών για ένα προϊόν, το σύστημα ορίζει πότε η κριτική εκφράζει μία θετική ή μία αρνητική άποψη για το προϊόν. Το επίπεδο αυτό είναι γνωστό ως συναισθηματική κατηγοριοποίηση επιπέδου κειμένου (document-level sentiment analysis) Αυτό το επίπεδο ανάλυσης ισχυρίζεται ότι κάθε έγγραφο εκφράζει μία άποψη για μία διαφορετική εγγραφή. Άρα, αυτό το επίπεδο δεν είναι χρήσιμο για έγγραφα τα οποία παρουσιάζουν ή συγκρίνουν πολλαπλές οντότητες.
- **Επίπεδο πρότασης (Sentence level):** Το βασικό θέμα αυτού του επιπέδου είναι οι προτάσεις και ορίζει πότε κάθε πρόταση εκφράζει αρνητική, θετική, ή ουδέτερη άποψη. Συνήθως, ουδέτερη σημαίνει καθόλου άποψη. Αυτό το επίπεδο ανάλυσης σχετίζεται στενά με την υποκειμενική κατηγοριοποίηση (subjectivity classification) [34], μέσω της οποίας διακρίνονται
    - οι αντικειμενικές προτάσεις (objective sentences), δηλαδή αυτές που εκφράζουν πραγματική πληροφορία , από τις
    - υποκειμενικές προτάσεις (subjective sentences), που εκφράζουν υποκειμενικές οπτικές και απόψεις.
- **Επίπεδο οντότητας και χαρακτηριστικών (Entity and Aspect level):** Τόσο το επίπεδο κειμένου όσο και το επίπεδο πρότασης δεν ανακαλύπτουν ακριβώς

τι είναι αυτό που αρέσει και τι αυτό που δεν αρέσει στους χρήστες. Το επίπεδο άποψης πραγματοποιεί λεπτομερέστερη ανάλυση. Παλαιότερα, το επίπεδο αυτό, ονομαζόταν επίπεδο χαρακτηριστικών (feature level) [55]. Αντί να επικεντρώνουμε το ενδιαφέρον σε γλωσσικές δομές, όπως κείμενα, παραγράφους, προτάσεις ή φράσεις), το επίπεδο αυτό, επικεντρώνεται απευθείας στην άποψη. Η βασική ιδέα λοιπόν είναι ότι μία άποψη αποτελείται από ένα συναίσθημα, θετικό ή αρνητικό, και έναν στόχο (του συναισθήματος). Η ανακάλυψη του στόχου μιας άποψης, βοηθάει στην καλύτερη κατανόηση του πεδίου της συναισθηματικής ανάλυσης. Για παράδειγμα, είναι εύκολα κατανοητό ότι η πρόταση "παρόλο που η εξυπηρέτηση δεν ήταν άριστη, το συγκεκριμένο εστιατόριο που αρέσει πολύ", έχει θετικό τόνο, ωστόσο δεν μπορούμε να πούμε ότι οι οντότητες της πρότασης είναι απόλυτα θετικές. Στην πραγματικότητα, η πρόταση είναι θετική σχετικά με την οντότητα εστιατόριο, στην οποία δίδεται ιδιαίτερη έμφαση και αρνητική σχετικά με την εξυπηρέτηση (οντότητα στην οποία δεν δίδεται έμφαση). Σε πολλές εφαρμογές, οι στόχοι μίας άποψης-κριτικής περιγράφονται από οντότητες και/ ή τα διαφορετικά χαρακτηριστικά (aspects). Αυτό το επίπεδο ανάλυσης έχει κύριο μέλημα του να ανακαλύψει συναισθήματα στις οντότητες και στις απόψεις. Για παράδειγμα, η πρόταση "Η ποιότητα των κλήσεων του iPhone είναι καλή, αλλά η ζωή της μπαταρίας του είναι μικρή" αναφέρεται σε δύο χαρακτηριστικά, την ποιότητα των κλήσεων και την ζωή της μπαταρίας, ενώ η οντότητα είναι το iPhone. Το συναίσθημα για την ποιότητα των κλήσεων είναι θετικό ενώ για την ζωή της μπαταρίας αρνητικό. Η ποιότητα κλήσεων και η ζωή της μπαταρίας του iPhone είναι οι στόχοι των απόψεων. Βασιζόμενοι σε αυτό το επίπεδο της ανάλυσης, μπορεί να δημιουργηθεί μία δομημένη ανασκόπηση των απόψεων που είναι σχετικές με την οντότητα και με τα χαρακτηριστικά της, η οποία μετατρέπει μη δομημένο κείμενο σε δομημένα δεδομένα και μπορεί να χρησιμοποιηθεί σε όλες τις μορφές της ποιοτικής και της ποσοτικής ανάλυσης. Τόσο το επίπεδο κειμένου, όσο και το επίπεδο πρότασης, παρουσιάζουν μεγάλες προκλήσεις. Το επίπεδο των χαρακτηριστικών εμφανίζει όμως ακόμη περισσότερες προκλήσεις.

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι υπάρχουν δύο είδη απόψεων [56]:

- Οι κανονικές απόψεις (regular opinions): εκφράζουν ένα συναίσθημα, μόνο για μία συγκεκριμένη οντότητα ή ένα συγκεκριμένο χαρακτηριστικό. Για παράδειγμα, "Η Coca-cola έχει ωραία γεύση", όπου εμφανίζει θετικό συναίσθημα για την οντότητα Coca-Cola και για το χαρακτηριστικό γεύση.
- Οι συγκριτικές απόψεις (comparative opinions): εκφράζουν συναίσθημα για περισσότερες από μία οντότητες, βασιζόμενοι σε κάποια κοινά τους χαρακτηριστικά. Για παράδειγμα, "Η Coca Cola έχει καλύτερη γεύση από την Pepsi", όπου συγκρίνονται οι οντότητες Coca Cola και Pepsi σχετικά με το κοινό χαρακτηριστικό της γεύσης, και εκφράζει προτίμηση στην Coca Cola.

4.5. Τεχνικές συναισθηματικής ανάλυσης

Οι τεχνικές που χρησιμοποιούνται από τα μοντέλα συναισθηματικής ανάλυσης μπορεί να κατηγοριοποιηθούν σε προσεγγίσεις μηχανής μάθησης [93] και σημασιολογικού προσανατολισμού [94]. Ακολούθως, δύο τύποι τεχνικών έχουν περιγραφεί ως προς την προσέγγιση του σημασιολογικού προσανατολισμού με στόχο την συναισθηματική ανάλυση, δηλαδή, με βάση το corpus και το λεξικό. Στην προσέγγιση ως προς το corpus, η αξία της πολικότητα υπολογίζεται με βάση τις cooccurrences συναρτήσει άλλων θετικών ή αρνητικών λέξεων στο σώμα. Υπάρχουν διάφορες μέθοδοι που αναφέρονται στη βιβλιογραφία για τον προσδιορισμό της αξίας της πολικότητας, ενώ, με βάση τις προσεγγίσεις λεξικού χρησιμοποιούνται τα προ αναπτυγμένα λεξικά πολικότητας όπως τα SentiWordNet [3], WordNet, SenticNet [5], και ούτω καθεξής. Αυτές οι μέθοδοι που ονομάζεται επίσης lexicons ή προσεγγίσεις βασισμένες στην γνώση.

4.6. Sentiment Lexicons

Ο πιο βασικός παράγοντας των συναισθημάτων είναι αδιαμφισβήτητα οι συναισθηματικές λέξεις (sentiment words), οι οποίες επίσης αναφέρονται και ως λέξεις άποψης (opinion words). Αυτά τα δύο είδη λέξεων χρησιμοποιούνται συχνά για να εκφράσουν θετικά ή αρνητικά συναισθήματα. Όμως πέρα από τις κοινές λέξεις υπάρχουν και φράσεις ή ιδιωματισμοί. Οι συναισθηματικές λέξεις (sentiment words)

και οι φράσεις, κάποιες φορές, αποτελούν εργαλείο της συναισθηματικής ανάλυσης. Μία λίστα από συναισθηματικές λέξεις και φράσεις ονομάζεται συναισθηματικό λεξιλόγιο (sentiment lexicon). Κατά την πάροδο των χρόνων έχουν δημιουργηθεί διάφοροι αλγόριθμοι που συντάσσουν τέτοια λεξιλόγια [57]. Αν και τα συναισθηματικά λεξιλόγια είναι απαραίτητα ωστόσο δεν είναι επαρκή και το πρόβλημα είναι πολύ πιο πολύπλοκο. Στη συνέχεια αναφέρουμε διάφορα προβλήματα των lexicons:

- Μια θετική ή αρνητική συναισθηματική λέξη μπορεί να έχει αντίθετους προσανατολισμούς σε διαφορετικά πεδία εφαρμογής.
- Μία πρόταση η οποία περιέχει συναισθηματικές λέξεις, μπορεί να μην εκφράζει κανένα συναίσθημα. Αυτό το φαινόμενο πραγματοποιείται συχνά σε διάφορους τύπους προτάσεων. Οι ερωτηματικές προτάσεις ή οι υποθετικές προτάσεις είναι δύο σημαντικού τύποι προτάσεων, αλλά όχι οι μοναδικοί που δεν εκφράζουν συναίσθημα.
- Οι σαρκαστικές προτάσεις (sarcastic sentences) είναι δύσκολα διαχειρίσιμες, είτε έχουν είτε δεν έχουν συναισθηματικές λέξεις. Ο σαρκασμός δεν είναι τόσο συχνός σε κριτικές προϊόντων ή υπηρεσιών, αλλά είναι πολύ κοινός σε πολιτικές συζητήσεις.
- Πολλές προτάσεις, οι οποίες δεν περιέχουν συναισθηματικές λέξεις μπορούν να εκφράζουν άποψη-γνώμη. Πολλές από αυτές τις προτάσεις είναι αντικειμενικές και χρησιμοποιούνται για να εκφράσουν πραγματική πληροφορία. Ένα τέτοιο παράδειγμα είναι “Το πλυντήριο χρησιμοποιεί πολύ νερό”, που εμφανίζει έναν αρνητικό τόνο για την οντότητα πλυντήριο.

4.7. Προσεγγίσεις Sentiment Lexicons

Τα Sentiment lexicons αποτελούν τις πιο βασικές πηγές για τους περισσότερους αλγορίθμους συναισθηματικής ανάλυσης. Στη συνέχεια, παρατίθενται τρεις διαφορετικές προσεγγίσεις τέτοιων λεξιλογίων [58].

1. **Χειροκίνητη προσέγγιση (manual approach):** η προσέγγιση αυτή δεν είναι εφικτή σε γενικές γραμμές, αφού κάθε κείμενο χρειάζεται το δικό του λεξικό,

γεγονός το οποίο είναι απαγορευτικό και η διαδικασία είναι ιδιαιτέρως επίπονη.

2. **Λεξικό-κεντρική προσέγγιση (dictionary -based approach):** Η προσέγγιση αυτή ξεκινά από ένα μικρό σετ από λέξεις ρίζες (seed words). Αυτό το σετ λέξεων στη συνέχεια επεκτείνεται με την χρήση lexicons, τα οποία περιέχουν συνώνυμα και αντώνυμα, όπως το Wordnet. Βασικό μειονέκτημα, της dictionary-based προσέγγισης αποτελεί το γεγονός ότι, τέτοιου είδους λεξικά, δεν λαμβάνουν υπόψιν τις ιδιαιτερότητες που μπορεί να έχει ένα συγκεκριμένο δείγμα δεδομένων.

Σε περίπτωση που θέλουμε να επικεντρωθούμε σε κάποιο δείγμα, μπορούμε να χρησιμοποιήσουμε κάποιον από τους πολλούς corpus-based αλγορίθμους [59]. Αυτού του είδους οι αλγόριθμοι επιτρέπουν να προσδιοριστούν πρόσθετα επίθετα που έχουν μια συνεπή πολικότητα ως προς το σύνολο των επιθέτων – σπόρων. Ένα σύνολο από συνδέσμους (AND, OR, NEITHER-NOR, EITHER-OR) χρησιμοποιείται για να βρεθούν επίθετα τα οποία συνδέονται με άλλα επίθετα με γνωστή πολικότητα. Για παράδειγμα, η πρόταση "Ο σκύλος είναι έξυπνος και φιλικός" αν γνωρίζουμε ότι η λέξη "έξυπνος" έχει θετική πολικότητα, τότε με την χρήση του συνδέσμου "και" συμπεραίνουμε ότι η λέξη "φιλικός" έχει επίσης θετική πολικότητα. Για να μειώσουν τον θόρυβο, αυτού του είδους οι αλγόριθμοι δημιουργούν έναν γράφο, και έπειτα πραγματοποιούν διαδικασία της συσταδοποίησης, κατά την οποία διαμορφώνονται θετικής ή αρνητικής πολικότητας συστάδες.

3. **Double propagation for simultaneous acquisition of a domain-specific sentiment lexicon approach:** Αυτή η προσέγγιση χρησιμοποιεί τον minipar μεταφραστή [60], ο οποίος μεταφράζει τις προτάσεις ενός συνόλου και εντοπίζει παρόμοιες γνώμες και συναισθηματικές εκφράσεις. Ο αλγόριθμος ξεκινά με ένα σύνολο "ρίζα" από συναισθηματικές εκφράσεις . Χρησιμοποιεί ένα σύνολο από προκαθορισμένους κανόνες ανεξαρτησίας και τον minipar parser για να εντοπίσει τις απόψεις που σχετίζονται με τις συναισθηματικές εκφράσεις. Στη συνέχεια, χρησιμοποιεί τις γνώμες αυτές για να εντοπίσει επιμέρους συναισθηματικές εκφράσεις και επαναλαμβάνει τη διαδικασία. Αυτή η διαδικασία σταματάει όταν δεν υπάρχει καμία άλλη γνώμη ή συναισθηματική έκφραση που μπορεί να προστεθεί. Για παράδειγμα, στην

πρόταση "η Μαρινέλλα πραγματοποίησε μία εκπληκτική εμφάνιση", το εκπληκτική είναι επίθετο, ενώ το εμφάνιση είναι ουσιαστικό. Έχοντας ως δεδομένο ότι το επίθετο "εκπληκτική" αποτελεί μία συναισθηματική έκφραση και με την ισχύ του κανόνα ότι "το ουσιαστικό που προσδιορίζεται από μία συναισθηματική έκφραση αποτελεί μία άποψη", συμπεραίνουμε ότι και η λέξη "εμφάνιση", αποτελεί μία άποψη. Αν από την άλλη έχουμε ως δεδομένο ότι η εμφάνιση αποτελεί μία άποψη, μπορούμε να συμπεράνουμε ότι το επίθετο "εκπληκτική" είναι μία συναισθηματική έκφραση. Ο αλγόριθμος χρησιμοποιεί διάφορους επιπρόσθετους περιορισμούς για να μειώσει την επίδραση του θορύβου.

4.8. Εφαρμογές

Οι πιο κοινές εφαρμογές συναισθηματικής ανάλυσης έχουν σαν θέμα τις κριτικές των χρηστών προϊόντων και υπηρεσιών. Πολλοί είναι οι ιστότοποι εκείνοι που παρέχουν αυτοματοποιημένες συνόψεις κριτικών σχετικών με προϊόντα. Ένα τέτοιο παράδειγμα είναι το "Google Product Search". Παρακάτω παρουσιάζονται κάποια είδη εφαρμογών [49].

1. **Applications to Review-Related Websites**

Αρχικά, πρέπει να αναφέρουμε ότι υπάρχουν ιστότοποι οι οποίοι ζητούν από τους χρήστες πληροφορίες και κριτικές. Τα θέματα που διαπραγματεύονται όμως δεν περιορίζονται αυστηρά και μόνο σε κριτικές προϊόντων, αλλά περιλαμβάνουν και απόψεις σχετικές με υπηρεσίες στις οποίες παρέχονται, πολιτικά ζητήματα κ.α.. Επίσης, υπάρχουν εφαρμογές, που έχουν σαν στόχο την προσέλκυση πελατών. Σε αυτού του είδους τις εφαρμογές, το πιο σημαντικό πρόβλημα που αντιμετωπίζεται είναι η σύνοψη των απόψεων των πελατών. Επίσης, ένα άλλο πολύ σημαντικό θέμα αποτελεί το πως ένας χρήστης που επιθυμεί να κάνει την κριτική του μπορεί να την διορθώσει σε περίπτωση λάθους. Έτσι, υπάρχουν σημαντικές ενδείξεις ότι οι κριτικές των χρηστών ίσως είναι μεροληπτικές, ή ίσως χρειάζονται διορθώσεις. Οι αυτοματοποιημένοι κατηγοριοποιητές, επιτρέπουν τέτοιες ενημερώσεις - διορθώσεις.

2. **Applications as a Sub-Component Technology**

Τα συστήματα της συναισθηματικής ανάλυσης και της εξόρυξης απόψεων έχουν έναν σημαντικό ρόλο αφού παρέχουν τεχνολογίες για κάποια άλλα συστήματα [61].

Έτσι τα συστήματα συναισθηματικής ανάλυσης, μπορεί να έχουν σαν στόχο την ενίσχυση των συστημάτων παροχής συστάσεων [62, 63]. Δηλαδή, ένα σύστημα παροχής συστάσεων να μην προτείνει αντικείμενα ή υπηρεσίες για τα οποία έχει λάβει αρνητικές πληροφορίες.

Ένας άλλος τρόπος υποκειμενικής ανίχνευσης ή κατηγοριοποίησης είναι η ανίχνευση εναυσμάτων (flames) σε μηνύματα ηλεκτρονικού ταχυδρομείου ή άλλους τύπους επικοινωνίας [61] .

Επιπλέον, στα online συστήματα που περιέχουν διαφημίσεις ως sidebars, είναι πολύ βοηθητικό να εντοπιστούν ιστοσελίδες που περιέχουν συναισθηματικό περιεχόμενο ακατάλληλο για την τοποθέτηση διαφημίσεων . Για τα πιο προσεγμένα συστήματα, θα ήταν χρήσιμο να υπάρχουν διαφημίσεις προϊόντων, εάν και εφόσον έχει διαγνωστεί γι αυτά θετικό συναίσθημα, και ίσως να ήταν ακόμα πιο σημαντικό να αποκλείονται οι διαφημίσεις των προϊόντων εκείνων για τα οποία έχει ανιχνευθεί αρνητική τάση [64].

Είναι πλέον αποδεκτό ότι η εξόρυξη πληροφορίας μπορεί να βελτιωθεί με την χρήση υποκειμενικών προτάσεων και δη με την ανίχνευση αρνητικών πληροφοριών στο εσωτερικό των προτάσεων αυτών [65].

Οι απαντήσεις σε ερωτήσεις είναι ένας άλλος τομέας στον οποίο η συναισθηματική ανάλυση μπορεί να αποδειχθεί χρήσιμη [66, 67, 68]. Οι Lita et al. [68] προτείνουν ότι οι ερωτήσεις, που παρέχουν απαντήσεις , οι οποίες περιλαμβάνουν κυρίως πληροφορίες, για το πως μία οντότητα κρίνεται από το ευρύ κοινό, μπορούν να παρέχουν πολλή καλή πληροφόρηση στον χρήστη.

Επιπλέον, σημαντική είναι και η ανάλυση των παραπομπών, όπου για παράδειγμα, κάποιος μπορεί να εμφανίζει το ενδιαφέρον του για τον λόγο που ένας άλλος συγγραφέας προτείνει κάποιο κομμάτι δουλειάς, που ο ίδιος απορρίπτει [69]. Παρομοίως, μία προσπάθεια επιχειρεί να χρησιμοποιήσει το σημασιολογικό προσανατολισμό για να εξορύξει την αξία σύμφωνα με τον κριτή[70].

Συμπερασματικά, η υπολογιστική επεξεργασία του συναισθήματος έχει κίνητρο εν μέρει από την επιθυμία για βελτίωση της αλληλεπίδρασης ανθρώπου-υπολογιστή [40, 71].

3. **Applications in Business and Government Intelligence**

Οι τομείς της εξόρυξης γνώμης και ανάλυσης συναισθήματος είναι κατάλληλοι για διάφορους τύπους εφαρμογών νοημοσύνης. Πράγματι, η επιχειρηματική ευφυΐα φαίνεται να είναι ένας από τους κύριους παράγοντες πίσω από το συνεργατικό ενδιαφέρον αυτού του τομέα.

Σύμφωνα με ένα σενάριο των Lee [73], ένας μεγάλος κατασκευαστής υπολογιστών, απογοητευμένος από τις απρόσμενα χαμηλές πωλήσεις, βρίσκεται αντιμέτωπος με το εξής ερώτημα: «Γιατί οι πελάτες δεν αγοράζουν τον φορητό υπολογιστή της εταιρείας μας;». Συγκεκριμένα στοιχεία, όπως το βάρος του φορητού υπολογιστή ή την τιμή του μοντέλου ενός ανταγωνιστή είναι απολύτως σχετικά με το θέμα. Για να απαντηθεί λοιπόν αυτό το ερώτημα πρέπει να επικεντρωθούν περισσότερο στις προσωπικές απόψεις των ανθρώπων σχετικά με τα εν λόγω χαρακτηριστικά. Επιπλέον, υποκειμενικές κρίσεις αναφορικά με άυλες ιδιότητες - π.χ., «ο σχεδιασμός είναι όμορφος" ή "η εξυπηρέτηση πελατών είναι καλή» - ή ακόμα και λανθασμένες αντιλήψεις - π.χ., "τα ενημερωμένα προγράμματα οδηγιών της συσκευής δεν είναι διαθέσιμα", πρέπει να λαμβάνονται επίσης υπόψη.

Οι τεχνολογίες ανάλυσης συναισθήματος για την εξαγωγή γνώμης από αδόμητα, συγγραφικά έγγραφα, αποτελούν εξαιρετικά εργαλεία για το χειρισμό πολλών εργασιών των επιχειρήσεων Συνεχίζοντας με το παραπάνω σενάριο θα ήταν δύσκολο να προσπαθήσει να ερευνήσει άμεσα τους λόγους για τους οποίους δεν επιλέγουν οι πελάτες το συγκεκριμένο προϊόν. Αντ 'αυτού, θα μπορούσε να χρησιμοποιήσει ένα σύστημα το οποίο:

a. θα εντοπίζει αξιολογήσεις ή άλλες ανάλογες εκφράσεις γνώμης στο Διαδίκτυο - ομάδες συζήτησης, ατομική blogs και sites συσσωμάτωση
b. θα δημιουργεί συνοπτικές εκδόσεις των ατομικών αξιολογήσεων ή μιας σύνοψης της συνολικού συναισθήματος των χρηστών.

Το γεγονός αυτό θα βοηθήσει τον αναλυτή από το να διαβάσει πιθανώς δεκάδες ή ακόμα και εκατοντάδες εκδόσεις των ίδιων των καταγγελιών.

Εκτός από τη διαχείριση της φήμης και των δημοσίων σχέσεων, θα μπορούσε κανείς να ελπίζει ότι ίσως με την εξόρυξη των δημοσίων απόψεων, θα μπορούσε να πραγματοποιηθεί πρόβλεψη σχετική με την εξέλιξη των πωλήσεων ή άλλων σχετικών δεδομένων [74].

4. **Applications Across Different Domains**

Όπως είναι γνωστό, οι γνώμες έχουν ιδιαίτερα σημαντικό ρόλο πολιτική. Κάποια έρευνα έχει επικεντρωθεί στην κατανόηση του τι οι ψηφοφόροι σκέφτονται [75, 76,77, 78, 79], ενώ άλλα έργα έχουν ως μακροπρόθεσμο στόχο τη διευκρίνιση των θέσεων των πολιτικών, δηλαδή ποιες είναι εκείνες οι απόψεις που υποστηρίζουν ή απορρίπτουν όπως αυτό της δημόσιας υποστήριξης στοιχεία ή να απορρίπτουν, έτσι ώστε να ενισχυθεί η ποιότητα των πληροφοριών, στις οποίες έχουν πρόσβαση οι ψηφοφόροι [80, 81, 82].

Η συναισθηματική ανάλυση έχει προταθεί σαν το κλειδί της εφαρμογής της τεχνολογίας στο eRulemaking, επιτρέποντας την αυτόματη ανάλυση των απόψεων που οι άνθρωποι υποβάλλουν σχετικά με την επικείμενη πολιτική ή τις κυβερνητικές προτάσεις [83, 84, 85].

Η αλληλεπίδραση με την κοινωνιολογία παρουσιάζεται εξαιρετικά γόνιμη. Για παράδειγμα, το ζήτημα του πώς οι ιδέες και οι καινοτομίες διαχέονται [86] περιλαμβάνει την ερώτηση του ποιος είναι θετικά ή αρνητικά διακείμενος ως προς ποιον, και ως εκ τούτου ποιοι θα είναι περισσότερο ή λιγότερο δεκτικοί σε νέες πληροφορίες, οι οποίες θα προέρχονται από μία συγκεκριμένη πηγή. Για να πάρουμε ένα άλλο παράδειγμα: η θεωρία του διαρθρωτικού ισοζυγίου βασικά ασχολείται με την πολικότητα των "δεσμών" μεταξύ των ανθρώπων [87] και πώς αυτοί επηρεάζουν τη συνοχή μίας ομάδας. Αυτές οι ιδέες έχουν αρχίσει να εφαρμόζεται από την ανάλυση των online μέσων [88, 89].

4.9. Online πηγές γνώσεις

Ως on-line πηγές γνώσης (Online knowledge sources) στην επεξεργασία φυσικής γλώσσας αναφέρονται οι βάσεις δεδομένων που είναι διαθέσιμες στο κοινό μέσω του Διαδικτύου και περιέχουν γλωσσικές γνώσεις . Συγκεκριμένα, αυτές οι βάσεις περιέχουν πληροφορίες σχετικά με τις λέξεις, τις έννοιες, ή τις φράσεις, καθώς και συνδέσεις – σχέσεις μεταξύ τους. Το είδος της σύνδεσης – σχέσης μπορεί να είναι διαφορετικής φύσης, όπως commonsense knowledge, ή λεξιλογικές σχέσεις. Οι περισσότερες βάσεις δεδομένων έχουν επικεντρωθεί σε ένα από αυτά τα πεδία. Για

την παροχή πληροφορίας με έναν κατάλληλο και αποτελεσματικό τρόπο, έχουν αναπτυχθεί διάφορα σχήματα αναπαράστασης. Έτσι, στα σημασιολογικά δίκτυα, οι λέξεις ή οι έννοιες αναπαριστώνται ως οι κόμβοι ενός γραφήματος και οι σχέσεις – συνδέσεις αναπαριστώνται από δεσμούς οι οποίοι φέρουν ένα όνομα [97]. Επίσης, υπάρχουν και τα "σημειωμένα" λεξικά, όπου οι ιδιότητες ενός όρου αποθηκεύονται ως ετικέτες. Αξιοσημείωτο είναι ότι τα λεξικά συνήθως δεν περιέχουν σχέσεις μεταξύ των όρων. Μία από τις πιο γνωστές online πηγές δεδομένων είναι το Conceptnet το οποίο παρουσιάζεται αναλυτικότερα στο κεφάλαιο 5.

Κεφάλαιο 5: Conceptnet και Sentiwordnet

5.1. Εισαγωγή

Για να πραγματοποιηθεί η συναισθηματική ανάλυση ενός corpus μπορούν να χρησιμοποιηθούν διάφορα εργαλεία. Πολλές έρευνες, για την αξιολόγηση του ενός συγκεκριμένου corpus βασίζονται σε συναισθηματικά λεξικά (sentiment lexicons), όπως αναφέρθηκε στο κεφάλαιο 4. Όμως, πολλές είναι οι έρευνες που αναφέρουν ότι τα συναισθηματικά λεξικά συνεργάζονται πολύ καλά με σημασιολογικά δίκτυα, όπως το Conceptnet. Ιδιαίτερο ενδιαφέρον θα παρουσίαζε να εξετάσουμε το Conceptnet και το πως αυτό εναρμονίζεται με κάποιο sentiment lexicon και συγκεκριμένα το SentiWordnet. Στη συνέχεια, παρουσιάζονται πληροφορίες σχετικά με τα εν λόγω εργαλεία.

5.2. Conceptnet

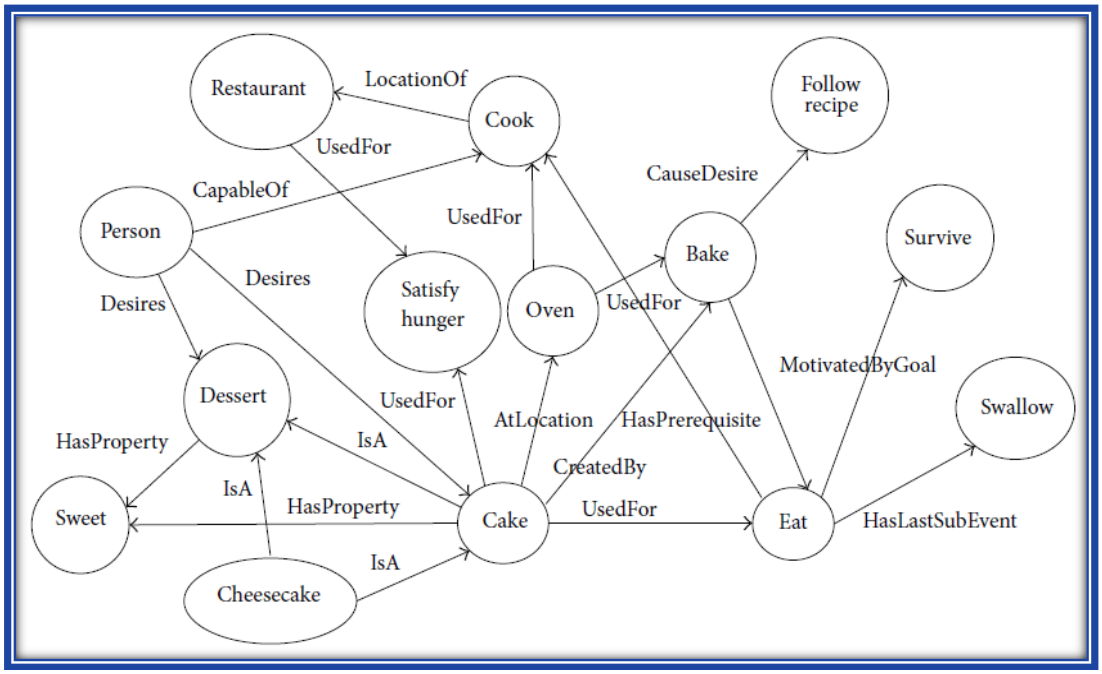
Το ConceptNet είναι ένα μεγάλο σημασιολογικό δίκτυο που αποτελείται από μεγάλο αριθμό εννοιών κοινής λογικής [11, 12] σε μορφή αναγνώσιμη από μηχάνημα. Οι έννοιες κοινής λογικής που εμφανίζονται στο ConceptNet υπάρχουν στο διαδίκτυο από απλούς χρήστες. Επιπλέον, η γνώση προστίθεται από μη εξειδικευμένους χρήστες μέσω μίας διεπαφής απόκτησης δεδομένων. Το Conceptnet είναι η μεγαλύτερη μηχανή που βασίζεται σε γνώση η οποία αποκτάται μέσω κοινής λογικής και αποτελείται από περισσότερες από 250.000 σχέσεις. Αξιοσημείωτο είναι πως το conceptnet αποτελεί τη μεγαλύτερη διαθέσιμη πηγή γνώσης η οποία δύναται να χρησιμοποιηθεί ως δεδομένο για την άντληση συμπερασμάτων από το κείμενο.

Όπως υποδηλώνει το όνομα, ConceptNet είναι ένα δίκτυο εννοιών. Έτσι, αποτελείται από κόμβους (έννοιες) που συνδέονται μέσω σχέσεων. Μερικές από τις σχέσεις μεταξύ των εννοιών του ConceptNet είναι το isA, EffectOf, CapableOf, MadeOf, DesireOf, και ούτω καθεξής [12]. Το ConceptNet ορίζεται από πέντε ιδιότητες, οι οποίες είναι:

- η γλώσσα,
- η σχέση,
- η πρώτη έννοια ,
- η δεύτερη έννοια, και

- η συχνότητα.

Η πρώτη και η δεύτερη έννοια ενώνονται με μία σχέση. Η συχνότητα αντιπροσωπεύει το πόσο συχνά οι έννοιες χρησιμοποιούνται με μία συγκεκριμένη σχέση, για παράδειγμα, οι σχέσεις , Restaurant UsedFor eat, Restaurant IsA place, και ούτω καθεξής. Το σημασιολογικό γράφημα του ConceptNet αντιπροσωπεύει τις πληροφορίες του corpus σαν ένα κατευθυνόμενο γράφημα, στο οποίο οι κόμβοι είναι έννοιες και οι ακμές είναι ο λόγος για τον οποίο οι δύο αυτές έννοιες εννόνονται. Για παράδειγμα, με δεδομένες τις δύο έννοιες «person» και «cook», η σχέση μεταξύ τους είναι το CapableOf? Δηλαδή, ένα άτομο έχει την ικανότητα του μαγειρέματος, όπως φαίνεται στην Εικόνα 5.1.



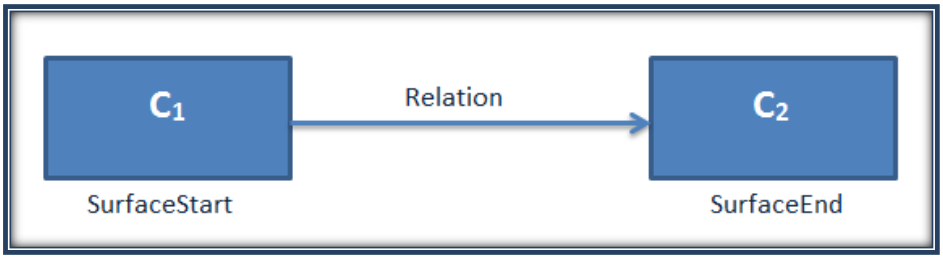
Εικόνα 5. 1 Παράδειγμα οντότητας Conceptnet [128]

Όταν αναφερόμαστε σε μία έννοια μπορεί να εννοούμε μία συγκεκριμένη λέξη, όμως κάποιες φορές η έννοια αποτελείται από έναν ακαθόριστο αριθμό λέξεων, κάτι που περιπλέκει ελαφρώς την κατάσταση. Για να ξεπεραστεί το παραπάνω πρόβλημα διαμορφώνονται άκρως συγκεκριμένες έννοιες. Οι έννοιες αποθηκεύονται σε κανονικοποιημένη μορφή, η οποία στοχεύει στην αγνόηση αμελητέων συντακτικών αλλαγών, που δεν επηρεάζουν την σημασία της εκάστοτε έννοιας. Η διαδικασία ομαλοποίησης μίας έννοιας περιέχει τα παρακάτω βήματα [98]:

1. Αφαίρεση των σημείων στίξης

2. Αφαίρεση των stop words
3. Χρήση του Porter stemmer για κάθε λέξη
4. Τοποθέτηση των στελεχών (stems) σε αλφαβητική σειρά, έτσι ώστε η σειρά στην οποία έχουν τοποθετηθεί να μην έχει καμία απολύτως σημασία.

Για να υπάρξει πλήρης κατανόηση του ConceptNet, θα περιοριστούμε μόνο σε δύο έννοιες (concepts). Έτσι, δεδομένων δύο εννοιών των C_1 και C_2 αλλά και μίας σχέσης (relation) προκύπτει η εικόνα 5.2



Εικόνα 5.2 Conceptnet Relation

Στο σχήμα 5.2 υπάρχουν δύο έννοιες, οι C_1 και C_2 οι οποίες ενώνονται με μία σχέση την relation. Στην προκείμενη εργασία έχουμε επιλέξει να αναλύσουμε συγκεκριμένες σχέσεις, οι οποίες αναφέρονται αναλυτικά στο κεφάλαιο 6.4.1. Στα αρχεία του ConceptNet, η έννοια C_1 απεικονίζεται ως SurfaceStart, ενώ η C_2 ως SurfaceEnd.

5.3. Πλεονεκτήματα του Conceptnet

Τα πλεονεκτήματα της χρήσης του Conceptnet ως λεξικής πηγής είναι ποικίλα [92]:

- Αριθμός και είδος των σημασιολογικών σχέσεων (IsA, HasA, Desires, UsedFor, HasProperty, LocationOf, DefinedAs, αιτίες και πολλά άλλα)
- Μέγεθος , πεδίο και εύρος γνώσης
- Το πεδίο εφαρμογής του ConceptNet είναι η γενική παγκόσμια γνώση και δεν περιορίζεται σε έναν συγκεκριμένο τομέα.

- Συλλογή γνώσεων σε πολλές γλώσσες (Παραδοσιακά Κινέζικα, Πορτογαλικά, Κορεατικά, Ιαπωνικά, Ολλανδικά)
- Οι έννοιες στο ConceptNet έχουν άτυπο χαρακτήρα
- Περιέχονται περιττές έννοιες και πολλοί τρόποι ώστε να εκφραστεί η ίδια έννοια (A camera is used for: photography, record images, making fotos, take photographs, take pictures)
- Δεν απαιτείται προ-σχολιασμένο σύνολο δεδομένων εκπαίδευσης ή δημιουργία χειρωνακτικής λίστας.
- Έργα για οποιοδήποτε μέρος του λόγου / λεξιλογική κατηγορία (επίθετο, ουσιαστικό, ρήμα, επίρρημα).

5.4. Commonsense Knowledge

Από τα διάφορα είδη σημασιολογικής γνώσης που ερευνώνται, αναμφισβήτητα η πιο γενική και ευρεία εφαρμογή του είδους είναι η γνώση για τον καθημερινό κόσμο, στην οποία έχουν πρόσβαση όλοι οι άνθρωποι - αυτό που είναι ευρέως ονομάζεται ως «commonsense knowledge». Σε έναν κοινό άνθρωπο, ο όρος «κοινή λογική» θεωρείται συνώνυμη με την «καλή κρίση», ωστόσο στο AI χρησιμοποιείται ως τεχνική γνώση και αναφέρεται στα εκατομμύρια των βασικών γεγονότων, στα οποία έχουν πρόσβαση οι περισσότεροι άνθρωποι.

Ένα λεμόνι είναι ξινό. Για να ανοίξετε μια πόρτα, θα πρέπει συνήθως να στρέψετε αρχικά το πόμολο. Εάν ξεχάσετε τα γενέθλια κάποιου, μπορεί να είναι δυσαρεστημένος μαζί σας. Η γνώση κοινής λογικής (commonsense knowledge), καλύπτει ένα τεράστιο μέρος της ανθρώπινης εμπειρίας, που περιλαμβάνει τη γνώση σχετικά με τις σωματικές, κοινωνικές και ψυχολογικές πτυχές της τυπικής καθημερινής ζωής. Επειδή είναι δεδομένο ότι κάθε άτομο διαθέτει κοινή λογική, αυτού του είδους η γνώση συνήθως παραλείπεται από κάποια έγγραφα, όπως κείμενο. Μια πλήρης κατανόηση του οποιουδήποτε κειμένου, απαιτεί σε μεγάλο βαθμό την κοινή λογική, την οποία διαθέτουν μόνο οι άνθρωποι. Σκοπός μας είναι να βρούμε τρόπους για την παροχή κοινής λογικής σε μηχανήματα.

5.5. Κατανόηση κειμένου

Οι υπολογιστές λοιπόν δεν διαθέτουν κοινή λογική, κάτι που έχει σαν αποτέλεσμα να μην μπορούν να αντλήσουν σωστά συμπεράσματα από τις πληροφορίες που παρέχει ένα κείμενο. Ένας υπολογιστής μπορεί να παίξει σκάκι αρκετά καλά, αλλά δεν μπορεί να καταλάβει ούτε μία απλή ιστορία για παιδιά. Ένας ταξινομητής, ο οποίος βασίζεται στη στατιστική, μπορεί να ταξινομήσει ένα e-mail, ως «flame», αλλά δεν μπορεί να εξηγήσει γιατί ο συγγραφέας είναι θυμωμένος (οι περισσότερες στατιστικοί ταξινομητές χρησιμοποιούν χαρακτηριστικά μεγάλων διαστάσεων) . Έχοντας την πρόταση, «έφαγα μερικά τσιπς με το γεύμα μου», με βάση την ανθρώπινη κοινή λογική γίνεται κατανοητό ότι ένα σύστημα είναι πιθανό να μην γνωρίζει ότι η λέξη «τσιπς» μάλλον παραπέμπει σε «πατατάκια», και όχι σε «τσιπ υπολογιστών».

Καθώς η εύρεση λέξεων κλειδιών, η ανάλυση της γλώσσας, και οι στατιστικές μέθοδοι έχουν βοηθήσει στην ανάλυση κειμένου, ωστόσο, υπάρχει ένα υποκατάστατο που αφορά την πληρότητα και την ευρωστία της ερμηνείας που δίνεται στης μεγάλης κλίμακας κοινή λογική. Χωρίς κοινή λογική, ένας αναγνώστης-υπολογιστής μπορεί να είναι σε θέση να μαντέψει ότι η φράση «Είχα μια απαίσια ημέρα" είναι αρνητικό, με τον εντοπισμό της λέξης-κλειδί «απαίσια». Με δεδομένη τη φράση «εγώ απολύθηκα σήμερα», ο αναγνώστης-υπολογιστής δεν ξέρει τι ακριβώς σκέφτεται ο συγγραφέας.

Σε αντίθεση, η κοινή λογική με βάση τις γνώσεις θα πρέπει να είναι σε θέση να αιτιολογήσει σχετικά με την ψυχολογική κατάσταση του ατόμου που απολύθηκε. Ίσως ξέρει κάποια πράγματα για την απόλυση. Μερικές φορές οι άνθρωποι απολύονται, επειδή είναι ανίκανοι. Μια πιθανή συνέπεια της απόλυσης είναι ο πρώην εργαζόμενος να έχει έλλειψη χρημάτων. Οι άνθρωποι χρειάζονται χρήματα για να πληρώσουν για τα τρόφιμα ή τη στέγη. Ακόμη και αν η βάση γνώσεων δεν έχει άμεση συναισθηματική γνώση για την κατάσταση της απόλυσης, ωστόσο μέσω του δικτύου των σχετικών γνώσεων θα πρέπει να είναι σε θέση να αντιλαμβάνονται ότι η κατάσταση αυτή, συνήθως φέρει πολλές αρνητικές συνέπειες, όπως φόβο, θυμό και θλίψη. Η γνώση κοινής λογικής είναι αναιρέσιμη, πράγμα που σημαίνει ότι συχνά είναι απλά μια προεπιλεγμένη υπόθεση για την τυπική περίπτωση (οι άνθρωποι

μπορεί να αισθάνονται ευτυχείς εάν απολυθούν από μια δουλειά που δεν τους αρέσει). Παρ 'όλα αυτά, αυτό το είδος της γνώσης θέτει ένα κρίσιμο θεμέλιο χωρίς το οποίο μία ερμηνεία πιθανόν να μην υπάρχει.

5.6. Conceptnet ως μηχανή περιεχομένου

Το ConceptNet φιλοδοξεί να συλλάβει γενικού σκοπού κόσμο-σημασιολογική γνώση (world-semantic knowledge). Παράλληλα οι ποιοτικές διαφορές στις αναπαραστάσεις των γνώσεων το καθιστά κατάλληλο για διάφορους σκοπούς.

Ο ερευνητής Gelernter [116] χαρακτηρίζει την ανθρώπινη συλλογιστική ως ένα "πέρασμα" μέσα από το φάσμα της ψυχικής εστίασης. Όταν η ψυχική εστίαση είναι υψηλή, το αποτέλεσμα είναι η λογική και η ορθολογική σκέψη. Η παραδοσιακή AI βαπτίζει μόνο μία πτυχή αυτού του φάσματος ως «αιτιολογία». Ωστόσο, ο Gelernter βιάζεται να επισημάνει ότι ένα μεγάλο μέρος, αν όχι η συντριπτική πλειοψηφία της ανθρώπινης κατανόησης συμβαίνει σε ένα μεσαίο ή χαμηλό επίπεδο, όπου το crisp deduction ανταλλάσσεται με την αντίληψη, με δημιουργική αναλογία και στο χαμηλότερο επίπεδο με απλή διασύνδεση. Ακόμα και αν είμαστε σκεπτικοί ως προς την ψυχολογία του Gelernter, η κατανόηση του κειμένου είναι ιδιάζουσας σημασίας. Χωρίς την κατανόηση του περιεχομένου μίας φράσης ή μιας ιστορίας, δεν θα είμαστε σε θέση να επιλέξουμε τις ερμηνείες των διφορούμενων λέξεων και περιγραφές που θα αφορούν άλλους ανθρώπους. Χωρίς ένα πλαίσιο εμπιστοσύνης, δεν θα είμαστε σε θέση να κατανοήσουμε πολλά παραδείγματα σαρκασμού, ειρωνείας, ή υπερβολής. Χωρίς να συνδυάσουμε κομμάτια της ιστορίας όλα μαζί σε ένα εννοιολογικό πλαίσιο δεν θα είμαστε σε θέση να κατανοήσουμε ένα βιβλίο. Όπως οι άνθρωποι χρειάζονται αυτό το είδος εννοιολογικού μηχανισμού για να διαβάσουν, έτσι και οι υπολογιστές απαιτούν κάποιον εννοιολογικό μηχανισμό για να διαχειριστούν έξυπνα την πληροφορία που υπάρχει μέσα στα κείμενα. Επανάσταση ως προς την διαχείριση των πληροφοριών ενός κειμένου θα αποτελούσε, εάν οι υπολογιστές μπορούσαν να διδαχθούν τον τρόπο να κατανοήσουν καλύτερα το περιεχόμενο του κειμένου. Η γενική αντίληψη είναι ότι το ConceptNet σημειώνει πρόοδο ως προς αυτή την κατεύθυνση.

Το Conceptnet είναι ένα σημασιολογικό δίκτυο, και αποτελεί ένα δεκτικό

πλαίσιο με μεθόδους φιλικές ως προς την ενίσχυση του σκεπτικού, όπως το "spreading activation" [117] και "graph traversal". Δεδομένων των κόμβων του ConceptNet και της σχεσιακής οντολογίας, μπορούν να επιτευχθούν αρκετά καλά συμπεράσματα με βάση τα συμφραζόμενα και απαιτούνται μόνο απλές βελτιώσεις στο "spreading activation".

Οι context-based μέθοδοι για εξαγωγή συμπεράσματος, επιτρέπουν στο Conceptnet να εκτελέσει ενδιαφέρουσες εργασίες όπως οι ακόλουθες [129]:

- Έχουμε σαν δεδομένο μία ιστορία που περιγράφει γεγονότα τα οποία συμβαίνουν σε καθημερινή βάση. Σε αυτή την ιστορία κάποια γεγονότα είναι πιθανό να συμβούν, επίσης είναι εύκολο να προβλεφθεί η διάθεση που επικρατεί μέσα στην ιστορία , και τι επίπτωση μπορεί να έχουν τα επόμενα γεγονότα
- Έχουμε σαν δεδομένο ένα query αναζήτησης (υποθέτουμε ότι οι όροι είναι όροι κοινής λογικής). Κάθε ένας από αυτούς τους όρους μπορεί να έχει πολλές σημασίες, ποια όμως σημασία είναι η πιο πιθανή; (η αποσαφήνιση πραγματοποιείται από τα συμφραζόμενα)
- έστω ότι παρουσιάζεται μια νέα έννοια σε μια ιστορία, ποιες είναι οι γνωστές έννοιες που μοιάζουν περισσότερο ή κατά προσέγγιση με τη νέα έννοια;» (analogy-making).

Δύο βασικοί λόγοι για τους οποίους ConceptNet είναι έμπειρο στη διαχείριση δεδομένων:

- οι επενδύσεις που έχει κάνει στην σχεσιακή γνώση, και
- οι επενδύσεις στη φυσική γλώσσα παράστασης της γνώσης.

Το ConceptNet επενδύει σημαντικά στην πραγματοποίηση συσχετίσεων μεταξύ των εννοιών, ακόμα και αν η αξία τους δεν είναι εμφανής. Από τα 1,6 εκατομμύρια γεγονότα που διασυνδέονται με τις έννοιες στο ConceptNet, περίπου 1.250.000 είναι αφιερωμένο στο να κάνει γενικές συνδέσεις μεταξύ των εννοιών. Αυτό το είδος της γνώσης περιγράφεται ως k-lines, το οποίο ο Minsky [118] περιγράφει ως πρωταρχικό μηχανισμό για το περιεχόμενο και τη μνήμη. Η K-line γνώση του ConceptNet αυξάνει τη συνδεσιμότητα του σημασιολογικού δικτύου, και

την καθιστά πιθανή έτσι ώστε να αντιστοιχιστούν οι έννοιες που περιγράφονται μέσα σε ένα κείμενο με αυτές που υπάρχουν στο ConceptNet.

Η φυσική γλώσσα παράστασης γνώσης ευνοεί το contextual reasoning. Σε αντίθεση με τα λογικά σύμβολα, τα οποία δεν έχουν a priori ερμηνεία, οι λέξεις τοποθετούνται πάντα σε δεσμούς και έχουν κάποια πιθανή σημασία. Αυτές οι λέξεις μεταφέρουν νοήματα, κάτι που δεν είναι αρνητικό ειδικά στο παιχνίδι περιεχομένου. Θέτοντας κόμβους στο ConceptNet είναι πιθανό να εξάγουμε λεξιλογικές ιεραρχίες, έτσι ώστε να κάνουμε ευέλικτη τη σημασία των κόμβων. Για παράδειγμα, οι κόμβοι "buy foods" και "purchase groceries" θεωρηθούν μέλη της ίδιας κατηγορίας, αν θεωρηθεί ότι το "buy" και το "purchase " είναι συνώνυμες λέξεις και ότι το "groceries" είναι ένα στιγμιότυπο του "food".

Ένα κρίσιμο ζήτημα που συχνά επιβάλλεται από τη φυσική γλώσσα είναι ότι υπάρχουν πολλοί ασαφείς τρόποι για να καθοριστεί η ίδια έννοια. Υποστηρίζουμε ότι αυτές οι «ασαφείς» έννοιες μπορούν να συμβιβαστούν με τις γλωσσικές γνώσεις, και υπάρχει επίσης η αξία για τη διατήρηση διαφορετικών τρόπων επεξήγησης της ίδιας έννοιας (π.χ. "car" και "automobile" είναι σχεδόν το ίδιο, αλλά μπορεί να σημαίνει κάτι διαφορετικό σύμφωνα με τα συμφραζόμενα) [119].

5.7. Ιστορία του Conceptnet

Μέχρι πρόσφατα, ο μόνος τρόπος για να δομηθεί μία βάση γνώσεων κοινής λογικής, θα ήταν η ιδιαιτέρως δαπανηρή διαδικασία, της πρόσληψης πολλών μηχανικών γνώσης. Ωστόσο, εμπνευσμένοι από την επιτυχία των συνεργατικών έργων στο Web, στραφήκαμε σε εθελοντές, οι οποίοι θα αναλάβουν την επίλυση του προβλήματος δημιουργίας μίας commonsense base. Το 2000, δημιουργήθηκε το Open Mind Common Sense (OMCS) Web Site [6] και αποτελούσε μια συλλογή από 30 διαφορετικές δραστηριότητες, κάθε μία από τις οποίες αφορά ένα διαφορετικό είδος ισχυρισμών κοινής λογικής όπως, περιγραφές τυπικών καταστάσεων, ιστορίες που περιγράφουν τις συνήθεις δραστηριότητες και δράσεις, και ούτω καθεξής. Από τότε και στο εξής η ιστοσελίδα έχει συγκεντρώσει πάνω από 700.000 φράσεις γνώσης κοινής λογικής και περισσότερους από 14 000 συνεργάτες σε όλο τον κόσμο, πολλοί από τους οποίους δεν έχουν ειδική εκπαίδευση στην επιστήμη των υπολογιστών. Στις

μέρες μας, το OMCS σώμα αποτελείται από ένα τεράστιο φάσμα διαφορετικών τύπων γνώση κοινής λογικής, που εκφράζεται μέσω της φυσικής γλώσσας [129].

Η πρώτη εφαρμογή του OMCS corpus σε κάποιο έργο, κάνει χρήση των OMCS προτάσεων, συντάσσοντας κανόνες για άντληση γνώσης από ένα σημασιολογικό δίκτυο. Το ARIA photo retrieval system είναι ένα εύρωστο σύστημα κοινωνικού συμπερασμού (CRIS) και είχε σαν βασικό στόχο να εξάγει ταξινομική, χωρική, λειτουργική, αιτιώδη και συναισθηματική γνώση από το OMCS. Χρησιμοποιεί το spreading activation για να βελτιώσει τον τρόπο ανάκτησης πληροφοριών. Σε αυτό το σημείο πρέπει να αναφέρουμε ότι CRIS, ήταν ο πρώτος προάγγελος του ConceptNet.

Η καινοτομία του CRIS για την ανάκτηση πληροφοριών πρότεινε μια νέα προσέγγιση για την δόμηση μιας βάσης γνώσεων κοινής λογικής. Το OMCS ενθαρρύνει τους ανθρώπους να παρέχουν πληροφορία με σαφήνεια και σε φυσική γλώσσα. Από αυτές τις ημί-δομημένες αγγλικές προτάσεις, είμαστε σε θέση αντλήσουμε γνώση σε πιο υπολογίσιμες παραστάσεις. Εξειδικεύοντας τη CRIS, δημιουργήθηκε ένα σημασιολογικό δίκτυο που ονομάζεται OMCSNet. Το OMCSNet προέκειψε με τη συστηματική αναδιατύπωση όλων των ημι-δομημένων φράσεων του OMCS σε ένα σημασιολογικό δίκτυο που περιέχει 280.000 άκρα και 80.000 κόμβους. Επιπλέον έχει αναπτυχθεί ένα API για OMCSNet, το οποίο υποστηρίζει τρεις βασικές λειτουργίες, την FindPathsBetween Nodes (node1, Node2), την GetContext (node), και την Get AnalogousConcepts (nodes). Το πακέτο του OMCSNet χρησιμοποιήθηκε από τις πρώτες εφαρμογές για την κατασκευή πολλών νέων εφαρμογών, όπως ένός δυναμικά παραγόμενου βιβλίου φράσεων ξένης γλώσσας που ονομάζεται GloBuddy (μια νεότερη έκδοση συζητείται από Lieberman et al [11]).

Τέλος, το OMCSNet υιοθετήθηκε ευρέως από προπτυχιακούς και μεταπτυχιακούς φοιτητές και επεδίωκε να κάνει εργασίες για το MIT Media Lab seminar, το επονομαζόμενο ως Common Sense Reasoning for Interactive Applications. Μέσω της χρήσης του OMCSNet, οι φοιτητές ήταν σε θέση να κατασκευάσουν μια διαφοροποιημένη συλλογή από ενδιαφέρουσες εφαρμογές, όπως μια AI-έκδοση του παιχνιδιού Taboo, ένας οικονομικός σύμβουλος κοινής λογικής, ένα περιβάλλον που δημιουργεί αυτόματα παιχνίδια [13]. Από αυτές τις πρώτες

εφαρμογές, παρατηρήσαμε ότι η ενσωμάτωση της επεξεργασίας φυσικής γλώσσας και του OMCSNet παρέμειναν ένα εμπόδιο της μηχανικής [129].

5.8. Δόμηση του Conceptnet

Το ConceptNet παράγεται από μια αυτόματη διαδικασία, και εφαρμόζει για πρώτη φορά ένα σύνολο κανόνων εξόρυξης στις ημι-δομημένες αγγλικές προτάσεις του OMCS corpus. Στη συνέχεια εφαρμόζει ένα επιπρόσθετο σύνολο των διαδικασιών «χαλάρωσης» (δηλαδή τη συμπλήρωση και την εξομάλυνση πάνω από τα κενά δικτύου) για να βελτιστοποιήσει τη συνδεσιμότητα του σημασιολογικού δικτύου [129].

5.8.1. Φάση Εξόρυξης

Περίπου πενήντα κανόνες εξόρυξης χρησιμοποιούνται για τη χαρτογράφηση αγγλικών OMCS προτάσεων στους δυαδικούς ισχυρισμούς του ConceptNet. Η χαρτογράφηση διευκολύνεται από το γεγονός ότι το OMCS website προκαλεί γνώση με έναν ημι-δομημένο τρόπο, ζητώντας από τους χρήστες να συμπληρώσουν τα κενά στα templates (π.χ. The effect of [falling off a bike] is [you get hurt]’). Οι προτάσεις για τις οποίες δεν υπάρχουν κατάλληλοι τύποι σχέσεων μπορούν να συσχετιστούν με k-line σχέσεις, αν και εφόσον περιέχουν σημασιολογικά γόνιμους όρους. Οι κανόνες εξόρυξης είναι πρότυπα κανονικών εκφράσεων που έχουν δημιουργηθεί για να αξιοποιήσουν την ήδη ημι-δομημένη φύση των OMCS προτάσεων. Επιπλέον, κάθε πρόταση δίνει μια επιφάνεια ανάλυσης σε MontyLingua, έτσι ώστε να μπορούν να εκτελεστούν οι συντακτικοί και σημασιολογικοί περιορισμοί στους κόμβους.

Ως αποτέλεσμα, οι κόμβοι του ConceptNet έχουν καθορισμένη συντακτική δομή, διευκολύνοντας την υπολογισημότητα τους. Κάθε κόμβος είναι ένα “πακέτο” διατυπωμένο στα αγγλικά, το οποίο αποτελείται από συνδυασμούς τεσσάρων συντακτικών μορφών. Οι συντακτικές αυτές μορφές είναι οι παρακάτω:

- ρήματα (π.χ. “buy”, “not eat”, “drive”),
- ονοματικές φράσεις (π.χ. “red car”, “laptop computer”),
- εμπρόθετες φράσεις (π.χ. “in restaurant”, “at work”), και

- φράσεις που περιέχουν επιθετικό προσδιορισμό (π.χ. “very sour”, “red”)

Η σειρά με την οποία εμφανίζονται οι συντακτικές μορφές έχει επίσης ιδιαίτερη σημασία, καθώς τα ρήματα πρέπει να προηγούνται και στη συνέχεια να ακολουθούν οι ονοματικές φράσεις, καθώς και οι φράσεις που περιέχουν επιθετικό προσδιορισμό, οι οποίες με την σειρά τους πρέπει να προηγούνται των εμπρόθετων φράσεων.

5.8.2. Φάση Κανονικοποίησης

Κατά την φάση της κανονικοποίησης (normalized phase), οι κόμβοι έχουν κανονικοποιηθεί. Η ορθογραφία των κόμβων διορθώνεται από ένα σύστημα ελέγχου της ορθογραφίας, χωρίς επίβλεψη. Οι συντακτικές μορφές (π.χ. ρήματα, ονοματικές φράσεις, εμπρόθετες φράσεις και φράσεις που περιέχουν επιθετικούς προσδιορισμούς) έχουν απαλλαχθεί από determiners (π.χ. “the” και “a”), από modals, και άλλα σημασιολογικά χαρακτηριστικά. Οι λέξεις επίσης απαλλάσσονται από τους χρόνους (π.χ. “is/are/were”→ “be”) και από πληθυντικό αριθμό (π.χ. “apples”→”apple.

5.8.3. Φάση Χαλάρωσης

Πέρα από τη φάση της εξόρυξης η οποία παράγει μία λίστα κανονικοποιημένων συνδέσμων, υπάρχει το επίπεδο της χαλάρωσης (relaxation) μέσα στο δίκτυο, το οποίο βοηθάει στην εξομάλυνση και στη βελτίωση της συνδεσιμότητας του δικτύου, παρακάμπτοντας τα σημασιολογικά κενά. Πρώτον, οι διπλοί σύνδεσμοι συγχωνεύονται και προστίθεται ένα επιπλέον πεδίο μεταδεδομένων, που ονομάζεται “συχνότητα”. Το πεδίο αυτό προστίθεται σε κάθε κατηγόρημα-σχέση για να παρακολουθείτε πόσες φορές ειπώθηκε κάτι. Δεύτερον, ιεραρχική σχέση του “IsA” χρησιμοποιείται για να εξάγουν γνώση με φορά από τα παιδί κόμβο ως προς τον μητρικό κόμβο. Ένα παράδειγμα αυτού δίδεται στην εικόνα 5.3.

```
[(IsA 'apple' 'fruit');
 (IsA 'banana' 'fruit');
 (IsA 'peach' 'fruit')]

AND

[(PropertyOf 'apple' 'sweet');
 (PropertyOf 'banana' 'sweet');
 (PropertyOf 'peach' 'sweet')]

IMPLIES

 (PropertyOf 'fruit' 'sweet')
```

Εικόνα 5. 3 Εξαγωγή γνώσης από παιδί προς μητρικό κόμβο [129]

Επιπλέον, οι θεματικές και λεξιλογικές γενικεύσεις που παράγονται, σχετίζουν την ειδική γνώση με την γενική γνώση, και αυτό εμπίπτει στην σχέση του τύπου SuperThematicKLine. Σε αυτό το σημείο χρησιμοποιούνται τα WordNet και FrameNet [15] synonym sets και οι ταξική ιεραρχίες. Δύο παραδείγματα αυτών των γενικεύσεων δίνονται στην εικόνα 5.3.

```
(SuperThematicKLine 'buy food' 'buy')
(SuperThematicKLine 'purchase food' 'buy')
```

Εικόνα 5.4 Σχέση τύπου SuperThematicKLine [129]

Επιπρόσθετα, όταν οι κόμβοι ονοματικών φράσεων περιέχουν επίθετικούς τροποποιητές,, οι επιθετικοί τροποποιητές μπορούν να αφαιρεθούν και να χρησιμοποιηθεί η σχέση PropertyOf, όπως φαίνεται στο ακόλουθο παράδειγμα:

```
[(IsA 'apple' 'red round object');
 (IsA 'apple' 'red fruit')]

IMPLIES

(PropertyOf 'apple' 'red')
```

Εικόνα 5.5 Σχέση τύπου PropertyOf [129]

Στη συνέχεια, οι αποκλίσεις του λεξιλογίου και οι μορφολογικές παραλλαγές θα απαλειφθούν. Έτσι, διαφορές στο λεξιλόγιο όπως “bike” και “bicycle” έχουν

γεφυρωθεί. Μορφολογικές παραλλαγές, όπως “relax” και “relaxation”, (δράση εναντίον κατάστασης) ή “sad”και “sadness” (επίθετο και ονομαστική έκφραση) έχουν επίσης παρακαμφθεί με την προσθήκη του SuperThematicKLine.

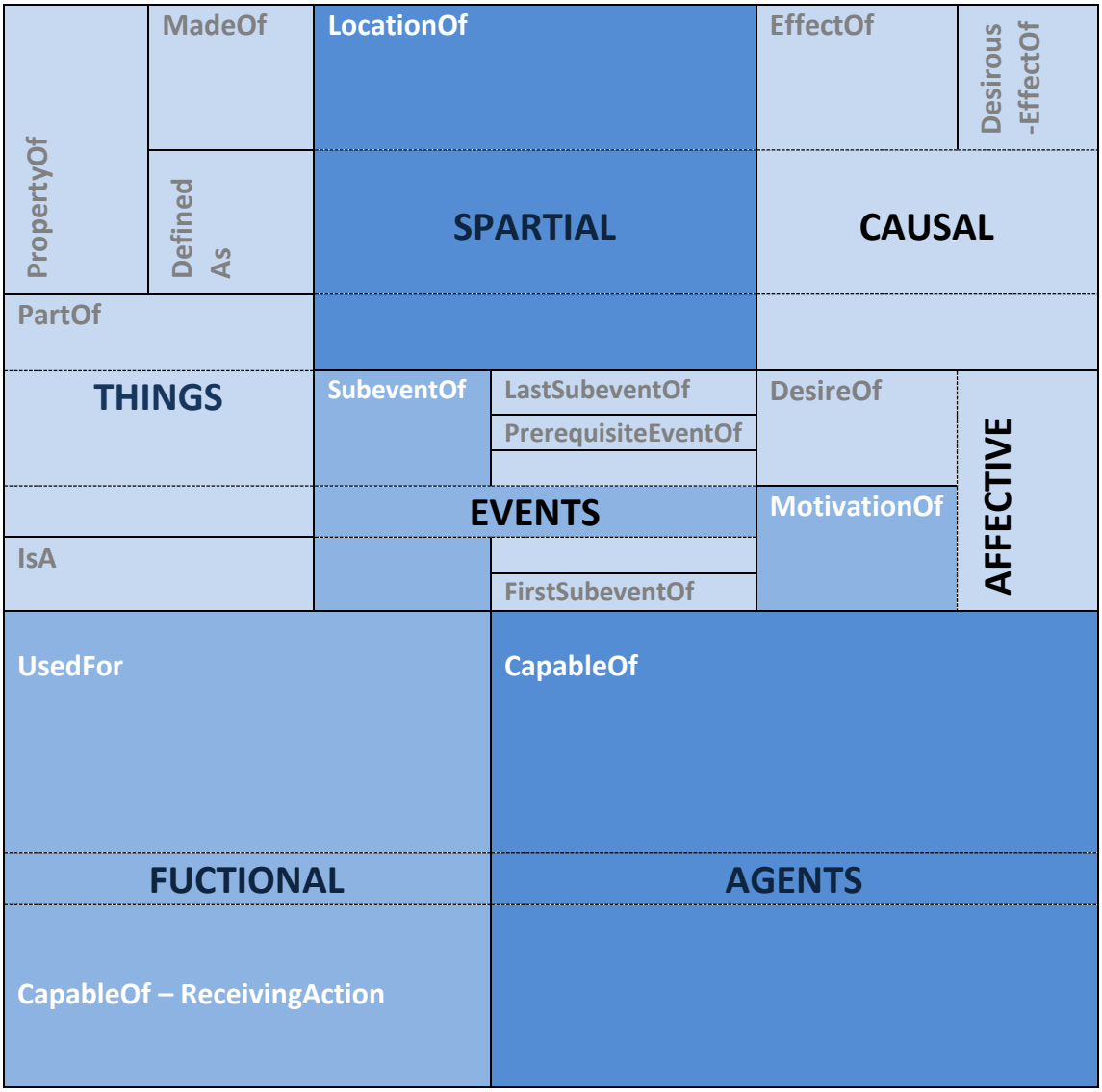
Για να εξάγουμε την γνώση που παράγεται από τις παραπάνω πρόσθετες γενικεύσεις, ένα πεδίο μεταδεδομένων που ονομάζεται «inferred_frequency» προστίθεται σε κάθε κατηγόρημα-σχέση. Οι Οι διαδικασίες εξαγωγής συμπερασμάτων του ConceptNet, βελτιώνουν την τεκμαιρόμενη γνώση, ως κατώτερη uttered-knowledge, αλλά παρ 'όλα αυτά τη χρησιμοποιούν. Όλη η επιπρόσθετη γνώση που εξάγεται από αυτή τη φάση χαλάρωσης, θεωρητικά, θα μπορούσε να διαμορφωθεί στο χρόνο εκτέλεσης του συμπεράσματος. Το παραπάνω γεγονός, θα εξοικονομούσε υπολογιστικές δαπάνες με τη χρήση τεχνικών φυσικού language processing.

5.9. Δόμηση της Conceptnet knowledge base

Η ConceptNet knowledge base σχηματίζεται από συνδυασμό 1,6 εκατ ισχυρισμών (assertions) (όπου τα 1,25 εκατομμύρια από τα οποία είναι klines) σε ένα σημασιολογικό δίκτυο το οποίο περιέχει πάνω των 300.000 κόμβους. Η παρούσα σχεσιακή οντολογία αποτελείται από είκοσι τύπους - σχέσεις. Ο πίνακας 5.1 είναι ένα treemap της σχεσιακής οντολογίας του ConceptNet, που δείχνει τις σχετικές ποσότητες των γνώσεων που εμπίπτουν στο πλαίσιο κάθε τύπου σχέσης.

Συγκεκριμένα, στον πίνακα 5.1 οι τύποι σχέσεων έχουν κατηγοριοποιηθεί σε διάφορες θεματικές περιοχές και τα σχετικά μεγέθη των ορθογωνίων είναι προσεγγιστικά ανάλογα, ως προς τον αριθμό των ισχυρισμών που ανήκουν σε κάθε τύπο σχέσης.

Στην συνέχεια, παρουσιάζεται ο Πίνακας 5.2, που δίνει ένα συγκεκριμένο παράδειγμα για κάθε τύπο σχέσης. Συγκεκριμένα, στον πίνακα αυτόν παρουσιάζονται με παραδείγματα από πραγματικά δεδομένα ConceptNet, οι είκοσι τύποι σχέσεων. Οι σχέσεις ομαδοποιούνται σε διάφορες θεματικές ενότητες. Η παράμετρος “f” μετρά τον αριθμό των φορών που ένα γεγονός ειπώθηκε στο OMCS corpus. Η παράμετρος “i” μετράει πόσες φορές ένας ισχυρισμός προκύπτει κατά τη διάρκεια της “relaxation” φάσης.



Πίνακας 5. 1 Κατηγοριοποίηση σχέσεων σε θεματικές περιοχές [129]

K-LINES (1.25 million assertions) (ConceptuallyRelatedTo 'bad breath' 'mint' '*f*=4;*i*=0;') (ThematicKLine 'wedding dress' 'veil' '*f*=9;*i*=0;') (SuperThematicKLine 'western civilisation' 'civilisation' '*f*=0;*i*=12;')
THINGS (52 000 assertions) (IsA 'horse' 'mammal' 'f=17;i=3;') (PropertyOf 'fire' 'dangerous' 'f=17;i=1;') (PartOf 'butterfly' 'wing' 'f=5;i=1;') (MadeOf 'bacon' 'pig' 'f=3;i=0;') (DefinedAs 'meat' 'flesh of animal' 'f=2;i=1;')
AGENTS (104 000 assertions) (CapableOf 'dentist' 'pull tooth' 'f=4;i=0;')
EVENTS (38 000 assertions) (PrerequisiteEventOf 'read letter' 'open envelope' '*f*=2;*i*=0;') (FirstSubeventOf 'start fire' 'light match' '*f*=2;*i*=3;') (SubeventOf 'play sport' 'score goal' '*f*=2;*i*=0;') (LastSubeventOf 'attend classical concert' 'applaud' '*f*=2;*i*=1;')
SPATIAL (36 000 assertions) (LocationOf 'army' 'in war' 'f=3;i=0;')
CAUSAL (17 000 assertions) (EffectOf 'view video' 'entertainment' '*f*=2;*i*=0;') (DesirousEffectOf 'sweat' 'take shower' '*f*=3;*i*=1;')
FUNCTIONAL (115 000 assertions) (UsedFor 'fireplace' 'burn wood' '*f*=1;*i* =2;') (CapableOfReceivingAction 'drink' 'serve' '*f* =0;*i* =14;')
AFFECTIVE (34 000 assertions) (MotivationOf 'play game' 'compete' 'f =3;i=0;') (DesireOf 'person' 'not be depressed' 'f=2;i=0;')

Πίνακας 5.2 Παραδείγματα σχέσεων Conceptnet [129]

Η σχεσιακή οντολογία του ConceptNet καθοριζόταν αρκετά οργανικά. Το αρχικό σώμα του OMCS δημιουργήθηκε σε μεγάλο βαθμό από τους χρήστες του γεμίζοντας τα κενά των προτύπων όπως “a hammer is for…”. Άλλα τμήματα του σώματος OMCS κάνουν δεκτές τις εισόδους ελεύθερης μορφής, αλλά περιορίζουν το μήκος της εισόδου, έτσι ώστε να ενθαρρύνεται η μεστή διατύπωση και η απλή σύνταξη.

5.10. Εξειδικευμένες γειτονίες

Δεδομένης μία «έννοιας», ποιες είναι οι υπόλοιπες έννοιες που μπορούν να συσχετιστούν με αυτήν; Το Conceptnet API παρέχει μία βασική λειτουργία που διευκολύνει αυτόν τον υπολογισμό, την αναφερόμενη ως GetContext(). Οι άνθρωποι είναι πολύ καλοί σε αυτή την μορφή context task, κάτι το οποίο δεν συμβαίνει και με τους υπολογιστές, οι οποίοι πάσχουν από έλλειψη προσοχής και σημασιολογικής ένωσης των εννοιών, όπως συμβαίνει κι στον ανθρώπινο εγκέφαλο. Ως ένα σημασιολογικό δίκτυο, του οποίου οι έννοιες ενώνονται μέσω ποικίλων διαστάσεων, το Conceptnet έχει τη δυνατότητα να ξεκινήσει να προσεγγίζει τις ανθρώπινες πιθανότητες που είναι σχετικές με το περιεχόμενο.

Από τεχνικής απόψεως, η γειτονία με έναν κόμβο μπορεί να βρεθεί από τον βασικό κόμβο (source node) προς τα έξω. Η συγγένεια κάθε συγκεκριμένου κόμβου δεν είναι απλά μια συνάρτηση της απόστασης του συνδέσμου από τον source code, αλλά επιπλέον συνυπολογίζει τον αριθμό και τις δυνάμεις όλων των μονοπατιών που συνδέουν δύο κόμβους.

5.10.1. Realm-filtering

Αναγνωρίζοντας ότι η σημασία του κάθε τύπου σχέσης διαφέρει ανάλογα με τον τομέα εργασίας στον οποίο αναφερόμαστε , θα εκχωρηθεί ένα διαφορετικό σύνολο των αριθμητικών βαρών για κάθε εργασία, ανάλογα με τον κάθε τύπο. Σύμφωνα με το ARIA Photo Agent, Liu et al [10] κάθε είδος σημασιολογικής σχέσης σταθμίζεται με βάση την αντίληψη της σημασίας της για την περιοχή ανάκτησης φωτογραφιών, και στη συνέχεια εκπαιδεύεται περαιτέρω με τα αριθμητικά βάρη κάθε

τύπου σχέσης σε ένα corpus. Σε αυτό το σημείο είναι επιθυμητό να απενεργοποιηθούν ορισμένοι τύποι σχέσεων. Με αυτόν τον τρόπο, μπορούμε να συλλέξουμε χρονικούς, χωρικούς, και άλλους γείτονες των εννοιών. Αυτή τη διαδικασία την ονομάζουμε φιλτράρισμα . Για παράδειγμα, αν χρησιμοποιήσουμε τις χρονικές, προς τα εμπρός εννοιολογικές επεκτάσεις, θα είναι εύκολο να προβλέψουμε τις επόμενες καταστάσεις, λαμβάνοντας υπόψη την τρέχουσα κατάσταση.

5.10.2. Topic generation

Η λειτουργία GetContext () είναι χρήσιμη για τη σημασιολογική επέκταση ενός query και δημιουργία ενός θέματος (topic generation). Κάποια νέα AI συστήματα έχουν κατασκευαστεί γύρω από αυτή την απλή λειτουργία. Για παράδειγμα, το Musa et al's GloBuddy system [16] είναι ένα δυναμικό ξενόγλωσσο βιβλίο φράσεων που χρησιμοποιεί την GetContext() λειτουργία του ConceptNet, με στόχο τη δημιουργία μιας συλλογής φράσεων, σε συνδυασμό με τις μεταφράσεις τους, για ένα συγκεκριμένο θέμα. Για παράδειγμα, εισάγοντας την έννοια "restaurant", θα επιστρέψει φράσεις όπως "order food" , "waiter" και "menu", καθώς επίσης και τη μετάφρασή τους στη γλώσσα-στόχο.

Ένας άλλος τρόπος για να χρησιμοποιηθεί η λειτουργία GetContext () είναι για την αναζήτηση της σημασιολογικής διασταύρωσης διαφόρων εννοιών. Αν εξάγουμε όλες τις έννοιες από ένα έγγραφο κειμένου και λάβουμε υπόψη τους τομείς τους, μπορούμε να επιτύχουμε το αντίστροφο του topic generation, το οποίο είναι θέμα gisting.

5.11. Analogy - making

Το analogy-making είναι μια διαδικασία αποσύνθεσης μιας ιδέας σε συστατικά στοιχεία και τα μέρη της, και, έπειτα, η ιδέα κατατάσσεται στον τομέα προορισμού που μοιράζεται ένα εξέχον υποσύνολο αυτών των συστατικών στοιχείων και των μερών. Επειδή η AI αναφέρεται συχνά στον επιχειρηματικό κλάδο, τοποθετώντας τις ιδέες σε «καλούπια» όπως σχήματα και πλαίσια [2], το analogy making είναι απαραίτητο να χρησιμοποιείται σε μεγάλο βαθμό. Αρχικά, μια βασική μορφή αναλογίας είναι εύκολο να υπολογιστεί. Ωστόσο, τα προγράμματα AI έχουν

ανάγκη μία «δεξαμενή» εννοιών και δομικών χαρακτηριστικών, τα οποία θα χρησιμοποιηθούν για την υποστήριξη commonsensical αναλογιών αποφάσεων. Πιστεύουμε ότι ConceptNet εξυπηρετεί κατά προσέγγιση, αυτή την ανάγκη.

Η structure-mapping θεωρία αναλογίας του Gentner δίνει έμφαση στις επίσημες, κοινές συντακτικές σχέσεις μεταξύ των εννοιών. Αντίθετα, κάποιο άλλο project δίνει έμφαση στις εννοιολογικές ομοιότητες και χρησιμοποιεί connectionist έννοιες για την εννοιολογική απόσταση. Η αναλογία στο Conceptnet μπορεί να εξαρτάται από ασθενείς σημασιολογικές σχέσεις (πχ. 'LocationOf', 'IsA') ή ισχυρές σημασιολογικές σχέσεις (πχ. 'PropertyOf', 'MotivationOf'). Η αναλογία στο Conceptnet έχει την ιδιότητα ότι οι συνδέσεις μεταξύ των κόμβων σταθμίζονται από την δύναμη ή την ασφάλεια ενός συγκεκριμένου ισχυρισμού.

Δύο κόμβοι στο Conceptnet είναι ανάλογοι αν το σύνολο των back-edges υπερκαλύπτεται. Για παράδειγμα, όσο το "apple" και το "cherry" μοιράζονται τα back-edges [(PropertyOf x 'red'); (PropertyOf x 'sweet'); (IsA x 'fruit')], θεωρούνται υπό μία έννοια ανάλογες έννοιες.

Η λειτουργία GetContext() μπορεί να είναι χρήσιμη για την εφαρμογή realm-filtering σε διαστάσεις κλίσης της λειτουργίας GetAnalogousConcepts(). Ίσως για παράδειγμα, να προτιμούμε να δώσουμε έμφαση στην λειτουργική ομοιότητα έναντι της συναισθηματικής ομοιότητας.

5.12. Projection

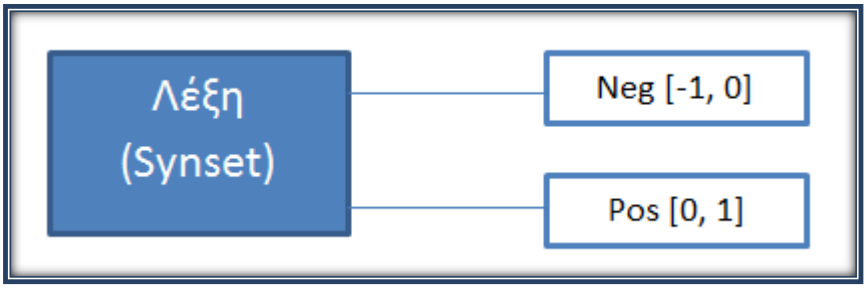
Ένας θεμελιώδης μηχανισμός συμπερασμού είναι το projection. Με αυτόν τον όρο αναφερόμαστε σε έναν γράφο διάσχισης από έναν κόμβο προέλευσης που ακολουθεί μία μορφή απλής μεταβατικής σχέσης. Έτσι, "'Los Angeles' is located in 'California', which is located in 'United States', which is located on 'Earth'" αποτελεί ένα παράδειγμα χωρικού projection, αφού το "LocationOf" είναι μία μεταβατική σχέση. Στο ConceptNet, τόσο ο τύπος της σχέσης που παρουσιάζει περιεχόμενο (πχ. LocationOf, IsA, PartOf, MadeOf, FirstSubeventOf, LastSubeventOf, SubeventOf), όσο και οι τύποι σχέσεων που παρουσιάζουν αποτέλεσμα (πχ. EffectOf, DesirousEffectOf) είναι μεταβατικοί και μπορούν να αξιοποιηθούν για το projection [129].

5.13. Wordnet και SentiWordNet

Το WordNet είναι μια μεγάλη λεξιλογική βάση δεδομένων. Τα γνωστικά συνώνυμα (synsets) είναι κατηγορίες που αποτελούνται από ουσιαστικά, ρήματα, επίθετα και επιρρήματα. Το καθένα από τα synsets εκφράζει μια ξεχωριστή ιδέα. Τα synsets διασυνδέονται μεταξύ τους μέσω εννοιολογικών-σημασιολογικών και λεξιλογικών σχέσεων. Το WordNet είναι ελεύθερο και δημόσια διαθέσιμο για download. Αποτελεί ένα χρήσιμο εργαλείο για την υπολογιστική γλωσσολογία και την επεξεργασία φυσικής γλώσσας [130].

Από την άλλη, το SentiWordNet αποτελεί έναν λεξιλογικό πόρο για opinion mining. Το SentiWordNet ουσιαστικά αντιστοιχίζει σε κάθε synset του WordNet τρία είδη συναισθημάτων, τα οποία είναι το θετικό, το αρνητικό κι το ουδέτερο [131]. Πρακτικά, το SentiWordNet είναι μία επέκταση του WordNet.

Συμπερασματικά, το SentiWordNet ουσιαστικά αποτελείται από λέξεις, όπου για κάθε λέξη υπάρχουν κάποια πεδία. Από αυτά τα πεδία, αυτά που μας αφορούν στην συγκεκριμένη εργασία είναι το Pos και το Neg, τα οποία αντιπροσωπεύουν την θετική και την αρνητική συναισθηματική αξία κάθε λέξης αντίστοιχα. Οι τιμές του πεδίου Pos κυμαίνονται από [0,1], ενώ οι τιμές του Neg από [-1, 0], όπως φαίνεται στο σχήμα 5.5.



Εικόνα 5..6 Synsets

5.14. Δημιουργία SentiWordNet

Η διαδικασία σύμφωνα με την οποία παράγεται το SentiWordNet αποτελείται από δύο στάδια [133]:

1. από ελαφριά επίβλεψη, βήμα μέσης επίβλεψης και
2. από τυχαίο βήμα (random-walk step)

Στη συνέχεια παρουσιάζονται αναλυτικά τα βήματα 1 και 2.

1. **Το βήμα μέσης επίβλεψης**

Το βήμα αυτό αποτελείται από τέσσερα επί μέρους βήματα. Τα βήματα αυτά είναι:

a. η επέκταση των ριζών,
b. η εκπαίδευση του αλγορίθμου που ταξινομεί τα synset
c. η ταξινόμηση των Synsets και
d. η συνδυαστική ταξινόμηση

Στο πρώτο (α) βήμα δύο ρίζες (seed) η μία αποτελούμενη από όλα τα synsets που περιλαμβάνουν 7 "paradigmatically positive" όρους, και η άλλη περιλαμβάνει όλα τα synsets αποτελούμενα από 7 "paradigmatically negative" όρους [132] επεκτείνονται με τη διασταύρωση ενός αριθμού των δυαδικών σχέσεων του WordNet ώστε να ληφθούν είτε για τη διατήρηση είτε για την αντιστροφή των Pos. και των Neg. ιδιοτήτων (δηλ. συνδέοντας τα synsets ενός δεδομένου προσανατολισμού με άλλα synsets είτε του ιδίου προσανατολισμού –για παράδειγμα την " also see" σχέση – η του αντίθετου προσανατολισμού – δηλαδή την "direct antonymy"σχέση), και με την προσθήκη synsets που προσεγγίζουν στην θετική ή στην αρνητική ομάδα ριζών. Αυτή η επέκταση μπορεί να διαμορφωθεί με ένα συγκεκριμένο "radius" δηλαδή με την χρήση k "radius" κάτι που σημαίνει ότι προσθέτουμε στην ομάδα των ριζών όλα τα synsets, τα οποία είναι εντός μίας απόστασης k από τα μέλη της αρχικής ομάδας ριζών.

Στο δεύτερο (b) υποβήμα, οι δύο ομάδες των synsets που δημιουργήθηκαν στο προηγούμενο βήμα χρησιμοποιούνται παράλληλα με άλλες ομάδες synsets, οι οποίες έχουν την Obj ιδιότητα σαν ομάδα εκπαίδευσης για την εκπαίδευση ενός τριαδικού ταξινομητή (δηλααδή ενός ο οποίος ταξινομεί ένα synset ως θετικό , αρνητικό ή ουδέτερο). Τα glosses των synsets χρησιμοποιούνται από το training module αντί για τα ίδια τα synsets, κάτι που σημαίνει ότι ο προκύπτων βασίζεται σε ένα gloss και όχι σε ένα synset.

Στο τρίτο βήμα (c) όλα τα synsets του WORDNET περιλαμβανομένων εκείνων που προστέθηκαν στην ομάδα των ριζών (Seeds) στις κατηγορίες "Pos", "Neg", και "Obj", μέσω του ταξινομητή που δημιουργήθηκε στο υποβήμα (b).

Στο τέταρτο βήμα (d) η τελική Pos αξία ενός δοθέντος Synset δημιουργείται σαν μια μέση Pos τιμή διά μέσου των 8 ταξινομητών.

2. Το τυχαίο βήμα

Το τυχαίο βήμα συνιστάται από παρατήρηση του WORDNET 3.0 με τη μορφή γραφήματος και με την εκτέλεση μιας διαδικασίας εικονοληπτικού τυχαίου βηματισμού στην οποίαν τα "Pos" και "Neg" τιμές έχουν εκείνες που καθορίσθηκαν στο προηγούμενο στάδιο και πιθανώς οι τιμές αυτές να αλλάζουν σε κάθε επανάληψη. Το βήμα του τυχαίου βηματισμού σταματά όταν η επαναληπτική διαδικασία συγκλίνει .

Το διάγραμμα που χρησιμοποιείται από το random-walk βήμα είναι το μόνο που καθορίζεται επακριβώς από το WordNet από την definiens-definiendum δυαδική σχέση. Με άλλα λόγια υποθέτουμε την ύπαρξη ενός δεσμού μεταξύ των Synset s1 και Synset s2, αν και μόνο το s1 (definiers) υπάρχει στο gloss των Synset s2 (definiendum) Το βασικό στοιχείο σε αυτό το σημείο είναι ότι, αν οι περισσότεροι από τους όρους οι οποίοι χρησιμοποιούνται για να ορίσουν ένα δεδομένο όρο είναι θετικοί (ή είναι αρνητικοί) τότε υπάρχει μία μεγάλη πιθανότητα ώστε ο όρος που καθορίζεται να είναι θετικός (ή αντίστοιχα αρνητικός) .

Σε αυτό το σημείο, θα πρέπει να σημειωθεί ότι στο regular WORDNET το definiendum είναι ένα Synset ενώ το definiens είναι ένας αμφίλογος όρος , ενώ το gloss είναι μία σειρά από μη αμφίλογων όρων. Όσον αφορά αυτό το τυχαίο βήμα, πρέπει τα glosses να είναι μη αμφιλεγόμενα ως προς το ίδιο το WORDNET, δηλαδή , πρέπει να ακολουθούν το WORDNET Synset. Δύο διαφορετικές διαδικασίες τυχαίου-βηματισμού εκτελούνται για τις θετικές και αρνητικές διαστάσεις, αντίστοιχα, δημιουργώντας δύο διαφορετικές βαθμίδες των WORDNET SYNSETS. Η πραγματική αριθμητική αξία προκύπτει από την διαδικασία τυχαίου βηματισμού και είναι αταίριαστη για να χρησιμοποιηθεί σαν τελικό Pος και Neg αποτέλεσμα. Ακόμη και το κορυφαίας αξιολόγησης θετικό Synset προκύπτει να είναι υπερβολικά ουδέτερο και μόνο ελαφρώς θετικό. Όπως έχει παρατηρηθεί το θετικό και αρνητικό αποτέλεσμα που προκύπτουν από ημιελεγχόμενο μαθησιακό βήμα και ακολουθούν έναν «νόμο ισχυρής κατανομής» δηλαδή, πολύ λίγα Synsets έχουν ένα πολύ υψηλό Pos και Neg βαθμό ενώ πάρα πολλά Synset είναι κυρίως ουδέτερα.

Κεφάλαιο 6: Εφαρμογή και Αποτελέσματα

6.1. Εισαγωγή

Διάφορες έρευνες αναφέρονται στην χρήση του Conceptnet και κάποιου lexicon για την συναισθηματική ανάλυση ενός corpus, είτε αυτό έχει προκύψει από το Twitter, είτε από κάποιον ιστότοπο στον οποίο οι χρήστες του διαδικτύου παραθέτουν τις γνώμες τους ή βαθμολογούν ένα προϊόν ή μία υπηρεσία. Όμως, κατά πόσο συμφέρει να χρησιμοποιήσουμε το Conceptnet και όχι απευθείας το sentiment lexicon. Ποιες είναι οι διαφορές μεταξύ των δύο αυτών διαδικασιών; Στη συγκεκριμένη διπλωματική εργασία έχουμε ως στόχο να δημιουργήσουμε έναν crawler με την βοήθεια του οποίου θα εντοπιστούν στο ConceptNet λέξεις οι οποίες υπάρχουν σε ένα sentiment lexicon και συγκεκριμένα στο SentiWordnet. Δηλαδή, θα εντοπιστούν τα synsets του SentiWordNet μέσα στο ConceptNet. Στη συνέχεια, θα αξιολογηθεί ένα dataset, το οποίο έχει προκύψει από το Twitter, με βάση τα δύο αυτά εργαλεία, δηλαδή με βάση αντίστοιχα το SentiWordNet και το Annotated ConceptNet. Τέλος, θα αξιολογηθεί ποιο από τα δύο εργαλεία παρουσιάζουν καλύτερο αποτέλεσμα ως προς την απόδοση τους συγκριτικά με το initial score.

Στη συνέχεια του κεφαλαίου, παρουσιάζονται οι τεχνολογίες οι οποίες χρησιμοποιήθηκαν για την εκπόνηση της εν λόγω διπλωματικής εργασίας, η διαδικασία και βασικές εντολές του συστήματος.

6.2. Εργαλεία

Για να πραγματοποιηθεί η συγκεκριμένη διπλωματική εργασία αρχικά εγκαταστάθηκε η MongoDB, η οποία χρησιμοποιήθηκε για την διαχείριση του ConceptNet5, καθώς και η βιβλιοθήκη NLTK.

Σε αυτό το σημείο είναι απαραίτητο να αναφερθεί ότι, στην MongoDB δημιουργείται ένα νέο collection. Κάθε εγγραφή του collection αντιπροσωπεύει μία σχέση που βρέθηκε στο ConceptNet. Το νέο αυτό collection περιέχει τα πεδία: surfaceStart, surfaceEnd, relation, posStart, posEnd, negStart, negEnd, findEnd.

Διευκρινίζονται τα παρακάτω:

- Τα πεδία SurfaceStart, SurfaceEnd και Relation είναι πεδία του ConceptNet, τα οποία υπάρχουν στο αρχικό collection του ConceptNet.
- Τα πεδία posStart, posEnd, negStart, και negEnd, έχουν προκύψει με τον συνδυασμό στοιχείων από το Conceptnet και από το SentiWordNet. Συγκεκριμένα:
    - posStart: το Pos του SurfaceStart, σύμφωνα με το Sentiwordnet
    - negStart: το Neg του SurfaceStart, σύμφωνα με το Sentiwordnet
    - posEnd: το Pos του SurfaceEnd
    - negEnd: το Neg του SurfaceEnd
    - findEnd: πεδίο το οποίο παίρνει τιμές 0 ή 1. Η τιμή 1 σημαίνει ότι το surfaceEnd βρέθηκε ως SurfaceStart και πως τα πεδία του έγιναν update με τις αντίστοιχες τιμές που υπάρχουν στο SentiWordNet για την συγκεκριμένη λέξη. Κατά την περίπτωση που η τιμή είναι 0, δεν βρέθηκε η λέξη.

Τέλος, όσον αφορά το πλήθος των σχέσεων που υπάρχουν στο καινούριο collection είναι μικρότερο από το πλήθος των συγκεκριμένων σχέσεων που υπάρχουν στο αρχικό collection του ConceptNet, αφού γίνεται έλεγχος εάν το SurfaceStart του collection υπάρχει σαν λέξη στο SentiWordNet.

6.2.1. Επιλογή γλώσσας προγραμματισμού: Python

Η συγκεκριμένη διπλωματική εργασία υλοποιήθηκε σε γλώσσα προγραμματισμού Python. Η **Python** είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού, η οποία αναπτύσσεται ως ανοιχτό λογισμικό και δημιουργήθηκε το 1990 από τον Ολλανδό Guido van Rossum [134].

Ο βασικότερος λόγος για τον οποίο χρησιμοποιήθηκε η Python είναι ότι, είναι ότι υποστηρίζει πολλές βιβλιοθήκες. Μία βιβλιοθήκη της, η οποία είναι πολλή σημαντική στην επεξεργασία κειμένου είναι η NLTK (Natural Language Toolkit).

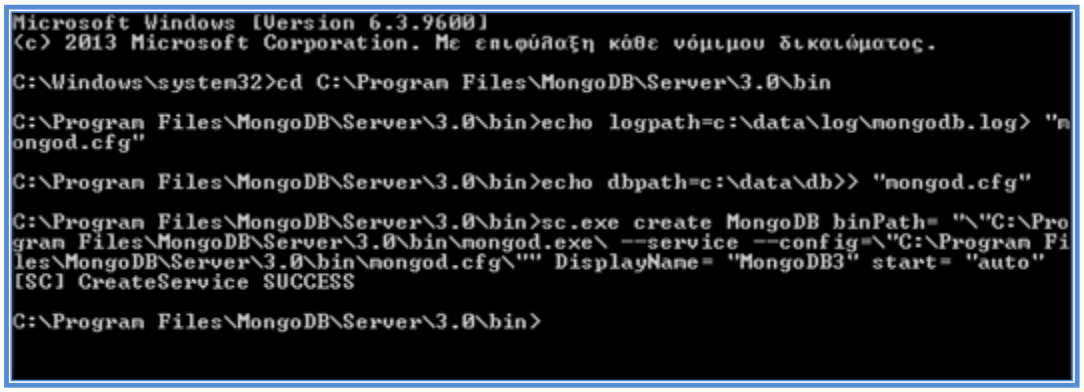
Επιπλέον, η γλώσσα αυτή υποστηρίζει την επεξεργασία αλφαριθμητικών (string), καθώς και πολλές μορφές προγραμματισμού όπως object-oriented, functional κ.α.. Τέλος, είναι μία απλή γλώσσα, αφού ο προγραμματιστής έχει τη δυνατότητα να εκφραστεί με πολύ λιγότερες γραμμές κώδικα, σε σύγκριση με άλλες γλώσσες οι οποίες χρησιμοποιούνται ευρέως.

6.2.2. Προαπαιτούμενα

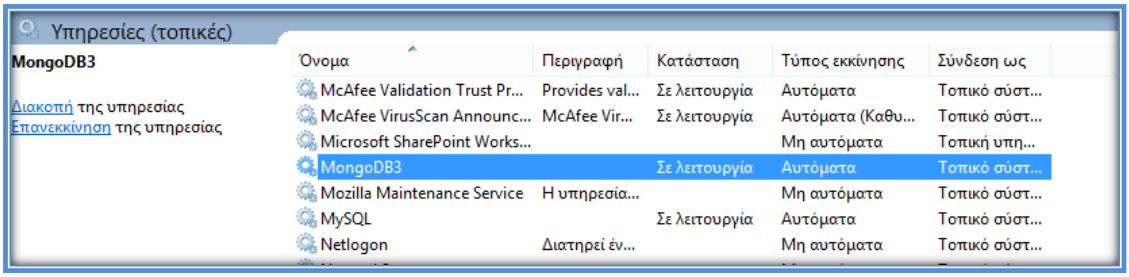
Για την εκπόνηση της συγκεκριμένης εργασίας και όπως φαίνεται και στο υποκεφάλαιο 6.3 υπάρχουν κάποια προαπαιτούμενα προγράμματα – συστήματα, τα οποία πρέπει να εγκατασταθούν. Πιο συγκεκριμένα, απαραίτητη κρίθηκε η χρήση της MongoDB και η εισαγωγή εντός αυτής των αρχείων του ConceptNet, καθώς και η εγκατάσταση της βιβλιοθήκης NLTK (Natural Language Tool-kit).

6.2.2.1 Διαχείριση ConceptNet

Για την διαχείριση και την εγκατάσταση του Conceptnet, χρησιμοποιείται η MongoDB, η οποία παρουσιάστηκε στο κεφάλαιο 2.8. Έτσι πραγματοποιήθηκε εγκατάσταση της MongoDB στο περιβάλλον των Windows 8. Στη συνέχεια, δημιουργήθηκε η MongoDB ως υπηρεσία, όπως φαίνεται στην εικόνα 6.1, και η ενεργοποίηση της υπηρεσίας αυτής όπως φαίνεται στην εικόνα 6.2.



Εικόνα 6.1 Δημιουργία υπηρεσίας MongoDB



Εικόνα 6.2 Ενεργοποίηση υπηρεσίας MongoDB

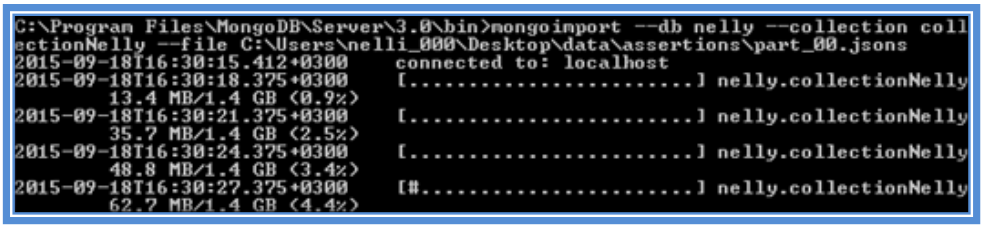
Έπειτα, πραγματοποιήθηκε εισαγωγή των flat json files του ConceptNet (τα οποία βρίσκονται στο url: http://conceptnet5.media.mit.edu/downloads/current/) στην MongoDB. Η εισαγωγή αυτών των αρχείων έγινε σε ένα σε ένα collection μίας συγκεκριμένης βάσης. Η εντολή που χρησιμοποιήθηκε γι' αυτό τον σκοπό είναι:

```
mongoimport --db databaseName --collection collectionName –file filename.json
```

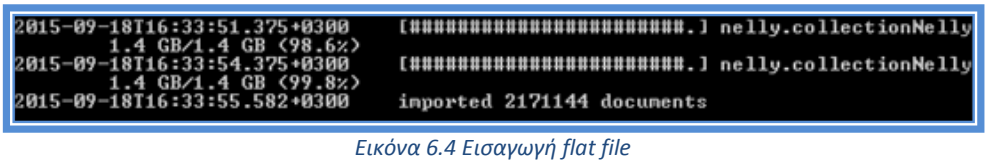
Όπου στην συγκεκριμένη εργασία:

- databaseName = nelly και
- collectionName = collectionNelly

Στην εικόνα 6.3 φαίνεται η αρχή της εισαγωγής των αρχείων του Conceptnet στην MongoDB και στην εικόνα 6.4 ο αριθμός των αρχείων που εισήχθησαν από ένα flat file.



Εικόνα 6.3 Εισαγωγή αρχείων Conceptnet



Εικόνα 6.4 Εισαγωγή flat file

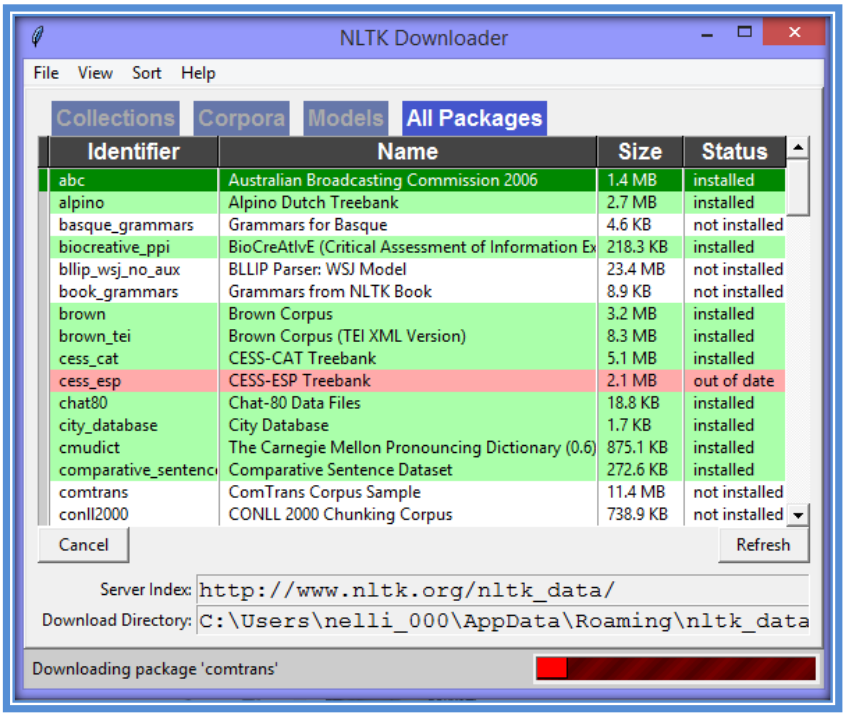
Τέλος, για την οπτικοποίηση της MongoDB, χρησιμοποιήθηκε το εργαλείο MongoVUE.

6.2.2.2 NLTK

Μέσα στα προαπαιτούμενα εντάσσεται και η βιβλιοθήκη NLTK. Έτσι, στο command prompt περιβάλλον της Python εισήχθηκαν οι εντολές:

```
import nltk
nltk.download()
```

Έτσι πραγματοποιείται download όλων των packages του NLTK. Στην εικόνα 6.5 φαίνεται ο NLTK Downloader.



Εικόνα 6. 5 NLTK Downloader

6.3 Dataset

Το dataset που χρησιμοποιήθηκε για την συγκεκριμένη εργασία είναι το semeval_task2_subtask2B.csv. Το αρχείο αυτό περιεχει 9045 γραμμές. Όλες αυτές οι γραμμές είναι tweets που προέρχονται από την πλατφόρμα του Twitter. Το dataset περιέχει τα εξής πεδία:

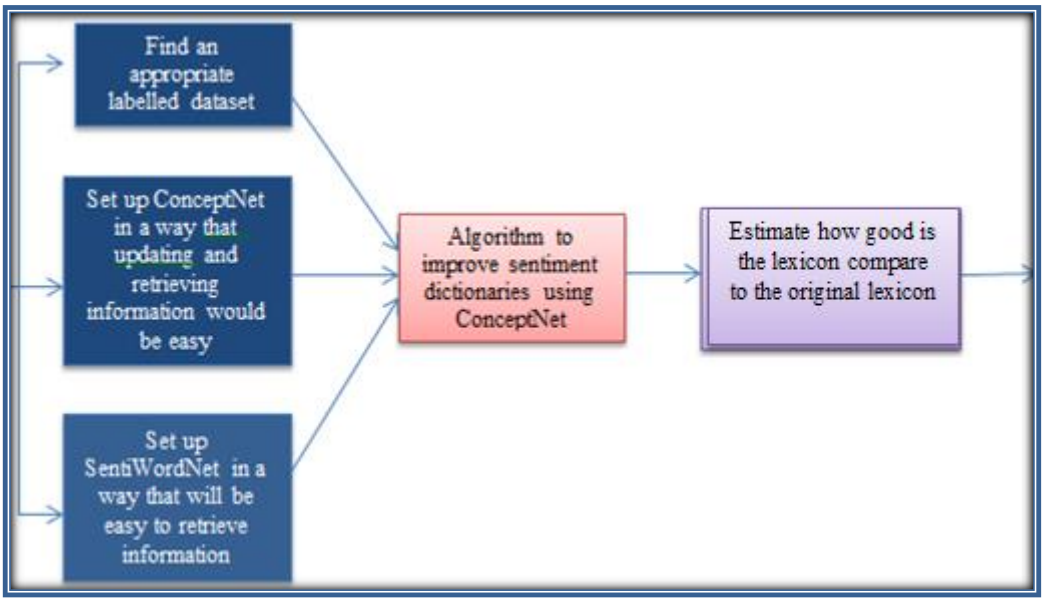
- Στήλη 1: id
- Στήλη 2: tweet
- Στήλη 3: Initial score

6.4 Βασική Διαδικασία

Η διαδικασία που ακολουθήθηκε για την εκπόνηση της συγκεκριμένης εργασίας είχε πέντε βασικά βήματα:

1. Εντοπισμός του κατάλληλου labelled dataset
2. Εγκατάσταση του ConceptNet με τρόπο κατά τον οποίο θα ήταν εύκολη η εξόρυξη και η ενημέρωση δεδομένων - πληροφοριών
3. Εγκατάσταση του SentiWordNet με τρόπο κατά τον οποίο θα ήταν εύκολη η εξόρυξη δεδομένων – πληροφοριών
4. Αλγόριθμος με στόχο την βελτίωση των sentiment dictionaries με την χρήση του ConceptNet
5. Υπολογισμός τις αποδοτικότητας του λεξικού που δημιουργήθηκε σε σύγκριση με το αυθεντικό λεξικό.

Τα παραπάνω βήματα αναπαριστώνται εικονικά στην Εικόνα 6.6.



Εικόνα 6. 6 Main Process

6.4.1 Εμπλουτισμός Conceptnet

Αρχικά, υπήρξε ενασχόληση με τον εμπλουτισμό του ConceptNet μέσω του SentiWordNet. Συγκεκριμένα, το Conceptnet εμπλουτίστηκε με κάποια επιπλέον πεδία, τα οποία προέκυψαν από το SentiWordnet. Τα πεδία αυτά είναι το Pos και το Neg, καθώς και το πεδίο findEnd, που προσδιορίζει αν η λέξη αυτή έχει βρεθεί στο SentiWordnet ή όχι.

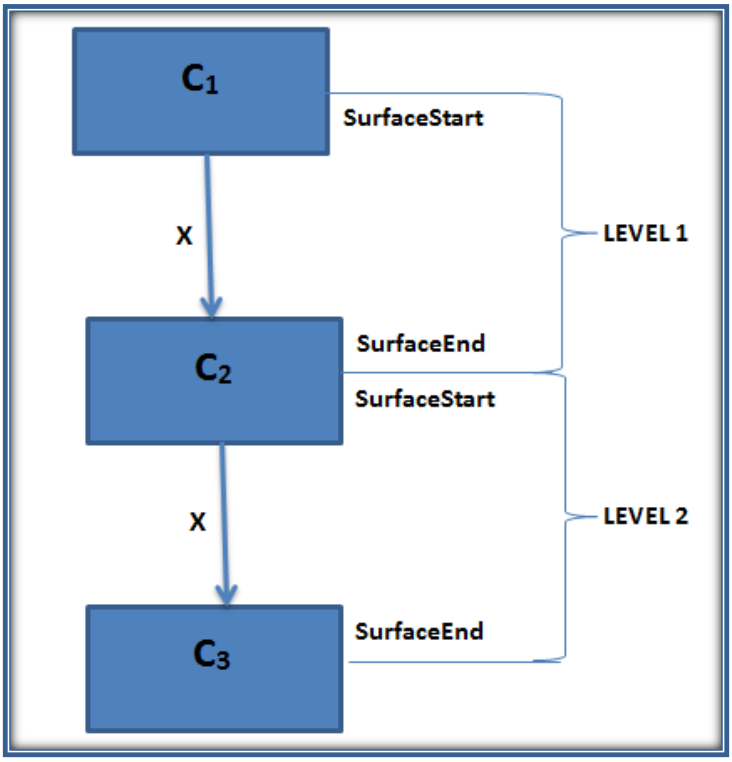
Επισημαίνεται ότι, στην εν λόγω διπλωματική εργασία ασχοληθήκαμε μόνο με το πρώτο επίπεδο του ConceptNet, αλλά και με συγκεκριμένες σχέσεις, όπως οι: RelatedTo, DerivedFrom, Synonym, IsA, HasA, EtymologicallyDerivedFrom, Antonym, PartOf και DefinedAs. Σε αυτό το σημείο πρέπει να διευκρινιστεί το τι ακριβώς εννοούμε λέγοντας πρώτο επίπεδο του Conceptnet.

Όπως αναφέρθηκε και παραπάνω το collection του ConceptNet αποτελείται από σχέσεις, οι οποίες συνδέουν δύο λέξεις (έννοιες). Έστω ότι έχουμε τις έννοιες C_1, C_2 και C_3. Η έννοια C_1 ενώνεται με την έννοια C_2 μέσω μίας σχέσης X. Η σχέση X ενώνει την έννοια C_2 με την C_3. Συνεπώς, η έννοια C_2 λειτουργεί και ως SurfaceStart και ως SurfaceEnd. Έτσι είναι η έννοια η οποία μας περνάει στο δεύτερο επίπεδο του Conceptnet.

Σύμφωνα με όλα τα παραπάνω, αρχικά ακολουθήθηκε η εξής διαδικασία:

1.1 Ξεκινάμε με μία λέξη – ρίζα από το Sentiwordnet
1.2 Εντοπίζω την λέξη αυτή στο Conceptnet
1.3 Ελέγχω αν το relation ανήκει στο set των relations που έχω ορίσει ότι θα μελετηθούν
1.4 Αν ναι, αποθηκεύω το αποτελέσματα
1.5 Επαναλαμβάνω την διαδικασία μέχρι να μην υπάρχει άλλη λέξη στο SentiWordNet

Λόγω όμως του τεράστιου όγκου των δεδομένων του Conceptnet (17366404 documents), ο αλγόριθμος αυτός ήταν ιδιαιτέρως αργός.



Εικόνα 6. 7 ConceptNet Levels

Συνεπώς, αναθεωρήθηκε η διαδικασία για να υπάρχει μικρότερη πολυπλοκότητα, έτσι ώστε ο αλγόριθμος να είναι πιο γρήγορος, με λιγότερους ελέγχους και να απασχολεί λιγότερους πόρους, δηλαδή να απασχολείται μικρότερο ποσοστό του δίσκου, της μνήμης και της CPU. Ως αποτέλεσμα, η τελική διαδικασία που ακολουθήθηκε για τον εμπλουτισμό του Conceptnet είναι η παρακάτω:

1.1 Ξεκινάμε με μία σχέση του Conceptnet

1.2 Ελέγχω αν το relation είναι ένα από τα relations που έχουμε επιλέξει να μελετήσουμε

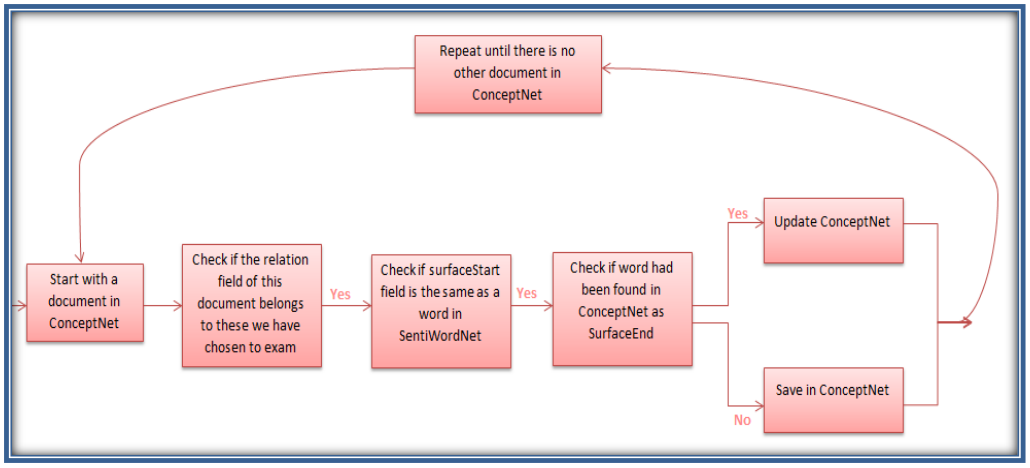
1.3 Αν ναι, ελέγχω αν το surfaceStart είναι ίδιο με κάποια λέξη του SentiWordnet

1.4 Εάν ναι, ελέγχω αν η λέξη αυτή είχε εντοπιστεί νωρίτερα ως SurfaceEnd

1.4.1 Εάν ναι, πραγματοποιείται update σε κάποια συγκεκριμένα πεδία (PosEnd, NegEnd, FindEnd), με τις τιμές Pos και Neg, που υπάρχουν στο SentiWordNet.

1.4.2 Εάν όχι, αποθηκεύω το αποτελέσματα.

1.5 Επαναλαμβάνω την διαδικασία μέχρι να μην υπάρχει άλλο document στο αρχικό collection του ConceptNet.



Εικόνα 6.8 Εμπλουτισμός ConceptNet

6.4.1.1 Ψευδοκώδικας

Για να υπάρξει καλύτερη κατανόηση της διαδικασίας που παρουσιάζεται στο κεφάλαιο 6.3.2, παρατίθεται το παρακάτω τμήμα ψευδοκώδικα.

ΠΡΟΓΡΑΜΜΑ Εμπλουτισμός ConceptNet μέσω SentiWordNet
ΔΕΔΟΜΕΝΑ document
ΑΡΧΗ
ΓΙΑ documents ΣΕ collection
ΑΝ relation == (RelatedTo ‘Η DerivedFrom ‘Η Synonym ‘Η IsA ‘Η HasA Ή EtymologicallyDerivedFrom ‘Η PartOf ‘Η DefinedAs ‘Η Antonym)
ΑΝ surfaceStart == word_in_Sentiwordnet
ΑΝ (word == surfaceEnd_in_new_collection **ΚΑΙ** findEnd ==0)
Ενημέρωσε PosEnd, NegEnd, FindEnd
ΑΛΛΙΩΣ
Break
Τέλος_αν
ΑΝ word =! surfaceEnd_in_new_collection
Βάλε νέο document στο new_collection
Τέλος_Αν
Τέλος_Αν
Τέλος_Αν

6.4.2 Ανάλυση – Αξιολόγηση Corpus

Έπειτα, ακολουθεί η διαδικασία με την οποία θα γίνει αξιολόγηση του αρχικού dataset, με δεδομένο τόσο το SentiWordNet, όσο και το εμπλουτισμένο Conceptnet. ‘Ετσι, για κάθε tweet του labelled dataset:

2.1 Πραγματοποιείται preprocessing

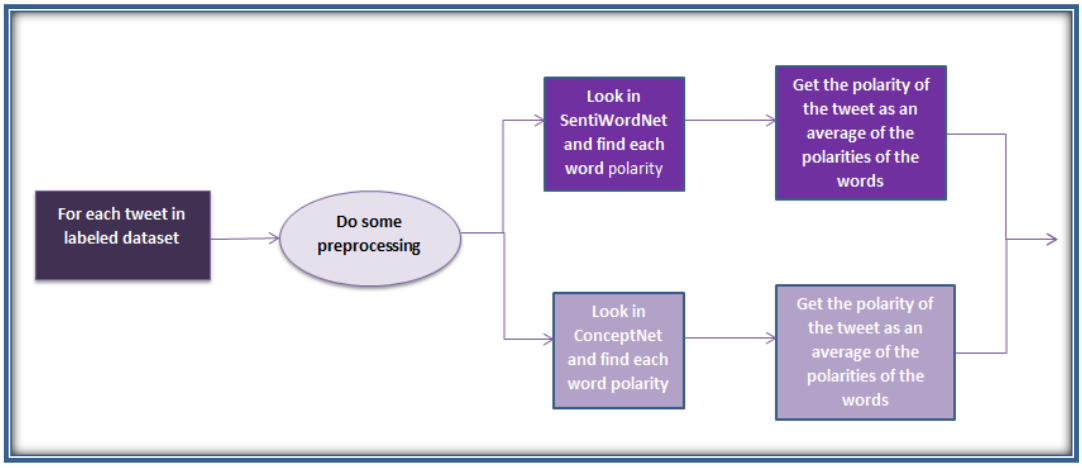
2.2 Ελέγχουμε το SentiwordNet, για την πολικότητα κάθε μίας από τις λέξεις του tweet (όπως αυτό έχει μετατραπεί μέσω της διαδικασίας του preprocessing – βλ. υποκεφάλαιο 5.2.1)

2.3 Η πολικότητα του κάθε tweet προκύπτει ως μέσος όρος της πολικότητας των επιμέρους λέξεων που βρέθηκαν στο SentiWordNet

2.4 Ελέγχουμε το Conceptnet, για την πολικότητα κάθε μίας από τις λέξεις του tweet (όπως αυτό έχει μετατραπεί μέσω της διαδικασίας του preprocessing– βλ. υποκεφάλαιο 5.2.1)

2.5 Η πολικότητα του κάθε tweet προκύπτει ως μέσος όρος της πολικότητας των επιμέρους λέξεων που βρέθηκαν στο ConceptNet

Στην εικόνα 6.9 παρουσιάζεται η παραπάνω διαδικασία.



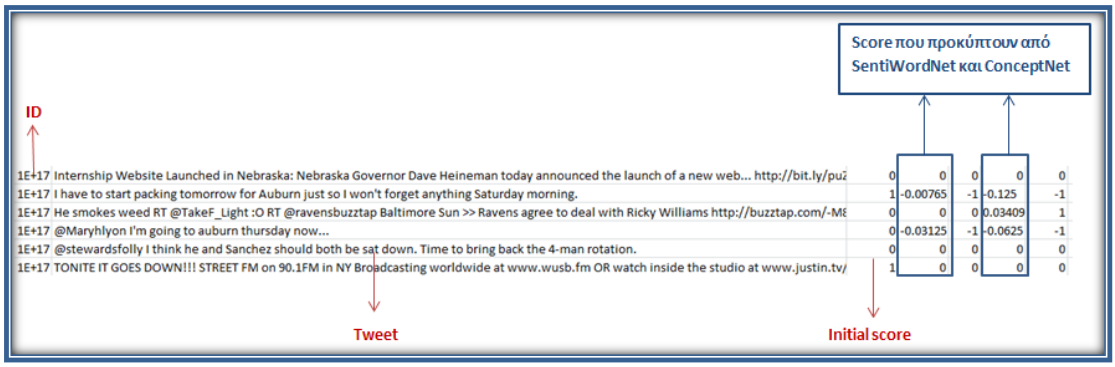
Εικόνα 6.9 Αξιολόγηση

Ως αποτέλεσμα της διαδικασίας ανάλυσης και αξιολόγησης προκύπτει ένα αρχείο .csv (semevalFinal) το οποίο έχει την εξής δομή:

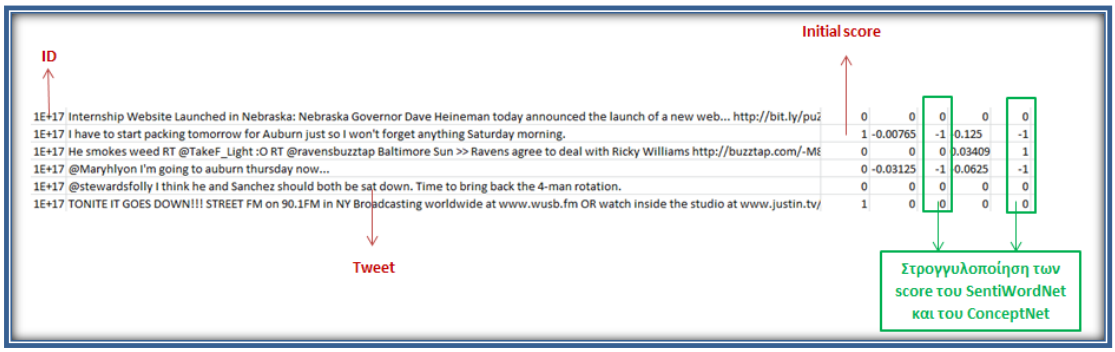
- Στήλη 1: id tweet
- Στήλη 2: κείμενο του tweet
- Στήλη 3: Score που προκύπτει μέσω του SentiWordNet

- Στήλη 4: Στρογγυλοποίηση του score της στήλης 3, ανάλογα με το αν αυτό είναι μεγαλύτερο, μικρότερο ή ίσο με το 0
- Στήλη 5: Score που προκύπτει μέσω του ConceptNet
- Στήλη 6: Στρογγυλοποίηση του score της στήλης 5, ανάλογα με το αν αυτό είναι μεγαλύτερο, μικρότερο ή ίσο με το 0

Στις εικόνες 6.10 και 6.11 παρουσιάζεται η μορφή αυτού του csv και εξηγούμε τι αναπαριστά η κάθε στήλη.



Εικόνα 6. 10 Score που προκύπτουν από αξιολόγηση μέσω SentiWordNet και ConceptNet



Εικόνα 6. 11 Στρογγυλοποίηση Score που προκύπτουν από αξιολόγηση μέσω SentiWordNet και ConceptNet

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι η στρογγυλοποίηση πραγματοποιείται ως εξής:

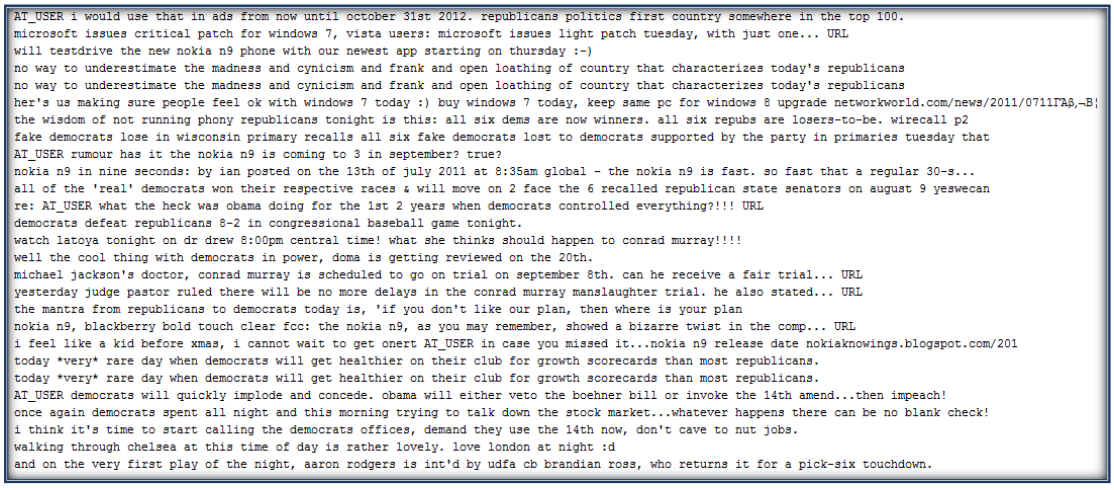
- -1: αν το score με το οποίο αξιολογήθηκε το tweet είναι μικρότερο του 0.
- +1: αν το score με το οποίο αξιολογήθηκε το tweet είναι μεγαλύτερο του 0.
- 0: αν το score με το οποίο αξιολογήθηκε το tweet είναι ίσο με το 0.

6.4.2.1 Ανάλυση και αποτελέσματα βήματος 2.1

Όσον αφορά το βήμα 2.1, δηλαδή το preprocessing πραγματοποιούμε τα εξής για κάθε tweet. Αρχικά μετατρέπουμε:

- Όλους τους χαρακτήρες κάθε tweet σε πεζούς
- Ότι ξεκινάει με την ακολουθία χαρακτήρων "www." ή "http://" σε "URL"
- Τα username σε AT_USER
- Τα hashtags σε απλές λέξεις

Επιπλέον, αφαιρούμε τα περιττά κενά μέσα από κάθε tweet. Για του λόγου του αληθές, παρατίθεται η εικόνα 6.8 στην οποία όπως μπορεί εύκολα κάποιος να παρατηρήσει δεν υπάρχουν κεφαλαίοι χαρακτήρες, δεν υπάρχουν σύνδεσμοι και αντ' αυτών υπάρχει η λέξη "URL", απουσιάζουν τα usernames των χρηστών και αντ' αυτών υπάρχει το AT_USER και τέλος, δεν υπάρχουν hashtags.



Εικόνα 6. 12 Διαδικασία ProcessedTweet

Στη συνέχεια, πρέπει το κάθε tweet να μετατραπεί στην κατάλληλη μορφή, έτσι ώστε να είναι διαχειρίσιμο. Συνεπώς, παίρνουμε ως δεδομένο τα ProcessedTweet που έχουν προκύψει από την παραπάνω διαδικασία και χωρίζουμε το κάθε tweet σε λέξεις. Για κάθε μία από αυτές τις λέξεις:

- Αφαιρούμε τους χαρακτήρες που επαναλαμβάνονται
- Αφαιρούμε τα σημεία στίξης
- Ελέγχουμε αν οι λέξεις περιέχουν μόνο γράμματα ή αριθμούς
- Ελέγχουμε αν οι λέξεις είναι stopWords και τις αφαιρούμε.

Το αποτέλεσμα που προκύπτει από αυτή τη διαδικασία παρουσιάζεται στην εικόνα 6.13.

```
['underestimate', 'madness', 'cynicism', 'frank', 'loathing', 'country', 'characterizes', 'republicans']
['people', 'feel', 'ok', 'windows', 'buy', 'windows', 'pc', 'windows', 'upgrade']
['wisdom', 'running', 'phony', 'republicans', 'tonight', 'six', 'dems', 'winners', 'six', 'repubs', 'wirecall']
['fake', 'democrats', 'lose', 'wisconsin', 'primary', 'recalls', 'six', 'fake', 'democrats', 'lost', 'democrats', 'supported', 'party', 'primaries', 'tuesday']
['rumour', 'nokia', 'coming', 'september', 'true']
['nokia', 'nine', 'ian', 'posted', 'july', 'global', 'nokia', 'fast', 'fast', 'regular']
['real', 'democrats', 'won', 'respective', 'races', 'move', 'recalled', 'republican', 'senators', 'august', 'yeswecan']
['heck', 'obama', 'doing', 'democrats', 'controlled']
['democrats', 'defeat', 'republicans', 'congressional', 'baseball', 'game', 'tonight']
['watch', 'latoya', 'tonight', 'dr', 'drew', 'central', 'happen', 'conrad']
['cool', 'democrats', 'power', 'doma', 'getting', 'reviewed']
['michael', 'doctor', 'conrad', 'murray', 'scheduled', 'trial', 'september', 'receive', 'fair', 'trial']
['yesterday', 'judge', 'pastor', 'ruled', 'delays', 'conrad', 'murray', 'manslaughter', 'trial', 'stated']
['mantra', 'republicans', 'democrats', 'plan', 'plan']
['nokia', 'blackberry', 'bold', 'touch', 'nokia', 'remember', 'bizarre', 'twist', 'comp']
['feel', 'kid', 'xmas', 'wait', 'onert', 'missed', 'release', 'date']
['rare', 'day', 'democrats', 'healthier', 'club', 'growth', 'scorecards', 'republicans']
['rare', 'day', 'democrats', 'healthier', 'club', 'growth', 'scorecards', 'republicans']
['democrats', 'quickly', 'implode', 'concede', 'obama', 'veto', 'boehner', 'bill', 'invoke']
['democrats', 'spent', 'night', 'morning', 'trying', 'talk', 'stock', 'happens', 'blank']
['time', 'start', 'calling', 'democrats', 'offices', 'demand', 'cave', 'nut', 'jobs']
['walking', 'chelsea', 'time', 'day', 'lovely', 'love', 'london', 'night']
['play', 'night', 'aaron', 'rodgers', 'udfa', 'cb', 'brandian', 'ross', 'returns', 'touchdown']
['drove', 'bike', 'miles', 'jim', 'carrey']
['looking', 'temp', 'hotter', 'sun', 'goes', 'feel', 'vegas', 'adding', 'humidity']
['expect', 'arsenal', 'sign', 'jag', 'cahill', 'option', 'bolton', 'willing', 'talk', 'arsenal', 'monday', 'bid']
```

Εικόνα 6.13 GetFeatureVector

6.4.2.2 Ανάλυση και αποτελέσματα βημάτων 2.2 – 2.5

Στα βήματα 2.2 – 2.5, πραγματοποιούνται έτσι ώστε να αξιολογηθεί ένα tweet ως θετικό, αρνητικό ή ουδέτερο (να χαρακτηριστεί δηλαδή από μία ετικέτα). Η αξιολόγηση αυτή γίνεται με ένα score. Πως όμως προκύπτει αυτό το score;

Με δεδομένο το feature Vector, το οποίο έχει προκύψει από την διαδικασία του preprocessing, κάθε tweet του dataset πρέπει να αξιολογηθεί με την χρήση τόσο του SentiWordnet, όσο και του Conceptnet. Έτσι, για κάθε λέξη που υπάρχει μέσα στο Feature Vector, πραγματοποιείται αναζήτηση στο Sentiwordnet και στο Conceptnet, για να εντοπίσουμε αν αυτή η λέξη υπάρχει σε κάθε ένα από τα δύο αυτά εργαλεία. Για κάθε ένα από τα εργαλεία αυτά, υπάρχει ένας counter που δείχνει πόσες

λέξεις εντοπίστηκαν για κάθε Feature Vector. Οι counter ουσιαστικά αποτελούν τον παρανομαστή της σχέσης Score.

Το score αποτελεί τον μέσο όρο των scores όλων των λέξεων που περιέχει κάθε tweet (FeatureVector). Όσον αφορά το SentiWordNet, η πολικότητα ενός tweet προκύπτει ως το άθροισμα των θετικών και των αρνητικών score των λέξεων ενός tweet, για κάθε λέξη του tweet προς τον αριθμό των λέξεων με πολικότητα διάφορη του μηδενός. Η σχέση που δημιουργήθηκε είναι η ακόλουθη:

$$\text{senti\_total\_score} = \frac{\text{senti\_score}}{\text{senti\_word\_counter}}$$

Όσον αφορά το ConceptNet, η πολικότητα ενός tweet προκύπτει ως το άθροισμα των θετικών και των αρνητικών score των λέξεων ενός tweet, για κάθε λέξη του tweet προς τον αριθμό των λέξεων με πολικότητα διάφορη του μηδενός. Η σχέση που ισχύει είναι η παρακάτω:

$$\text{concept\_total\_score} = \frac{\text{concept\_score}}{\text{concept\_word\_counter}}$$

Για την παραπάνω σχέση ισχύουν τα εξής:

- Αν μία λέξη έχει βρεθεί ως SurfaceStart για να προκύψει το concept_score ισχύει η σχέση:

  $concept\_score = concept\_score + ["posStart"] - ["negStart"]$

  στην οποία "posStart" είναι το Pos του SurfaceStart, δηλαδή της πρώτης έννοιας και "negStart" είναι το Neg του surfaceStart στο ConceptNet (βλ. 5.13)

- Αν έχει βρεθεί ως SurfaceEnd ισχύει η σχέση:

  $concept\_score = concept\_score + ["posEnd"] - ["negEnd"]$

στην οποία "posEnd" είναι το Pos του SurfaceEnd και "negEnd" είναι το Neg του surfaceEnd στο ConceptNet.

6.5 Απόδοση εργαλείων

Για να υπολογίσουμε ποιο από τα δύο εργαλεία είναι πιο αποδοτικό ως προς αξιολόγηση των tweets του corpus, υπολογίζουμε μία τιμή. Ποια είναι όμως αυτή η τιμή;

Η τιμή αυτή είναι η απόλυτη τιμή της διαφοράς του initial score με το score που προέκυψε στο υποκεφάλαιο 6.3.3.2 . Ονομάζεται senti_results για το SentiWordNet και concept_results για το ConceptNet

- senti_results = $\sum$ |Initial_score – senti_score|
- concept_results = $\sum$ |Initial_score – concept_score|

Για να μετρήσουμε την απόδοση των εργαλείων πρέπει η δούμε ποιο από αυτά έχει την μικρότερη τιμή. Όποιο εργαλείο έχει μικρότερη τιμή είναι και αυτό που αποδίδει καλύτερα συγκριτικά με το άλλο. Έτσι αν το senti_results είναι μικρότερο από το concept_results τότε το SentiWordNet αποδίδει καλύτερα συγκριτικά με το ConceptNet, αλλιώς αυτό που αποδίδει καλύτερα είναι το Conceptnet. Αυτό παρουσιάζεται στο υποκεφάλαιο 6.5.

6.6 Αποτελέσματα

Αρχικά, ως αποτέλεσμα της διαδικασίας που περιγράφεται στο υποκεφάλαιο 6.4.3, προκύπτει ένα νέο collection της MongoDB, το οποίο περιέχει το annotated ConceptNet. Στις παρακάτω εικόνες φαίνονται ενδεικτικά αποτελέσματα για τις σχέσεις : RelatedTo, DerivedFrom, Synonym, IsA, HasA, EtymologicallyDerivedFrom, PartOf, DefinedAs και Antonym.

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
566175ba9b1f7...	0.125	0.125	/r/DerivedFrom	0	0.375	absens	0.375	absent
566175bd9b1f7...	0.125	0.75	/r/DerivedFrom	1	0	accident	0	accidental

Εικόνα 6. 14 DerivedFrom

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
5661bec19b1f7...	0.375	0	/r/IsA	1	0.125	medicine	0	abortifacient
5661bec29b1f7...	0	0	/r/IsA	1	0	person	0	abolitionist

Εικόνα 6.15 IsA

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
56626eb59b1f7...	0.75	0.625	/r/RelatedTo	1	0	none	0.125	dim
56626ebe9b1f7...	0	0.25	/r/RelatedTo	1	0.5	bestow	0	euro

Εικόνα 6. 16 RelatedTo

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
567d9b919b1f7...	0	0	/r/Synonym	0	0.125	barrier	0.125	plugged

Εικόνα 6.17 Synonym

5690e65e9b1f7...	0	0	/r/Etymologic...	0	0	ball	0	pushball
5690eab29b1f7...	0.875	0.875	/r/Etymologic...	0	0	Schlacke	0	schlock

Εικόνα 6.18 EtymologicallyDerivedFrom

5697e1b19b1f7...	0	0	/r/PartOf	0	0	california	0	whittier
5697e1b69b1f7...	0	0	/r/PartOf	1	0.125	ontario	0	windsor

Εικόνα 6. 19 PartOF

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
569a42129b1f7...	0	0	/r/HasProperty	0	0.625	difficult to attain	0.625	tranquility

Εικόνα 6. 20 HasPropertyOf

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
569ac26f9b1f7...	0	0	/r/DefinedAs	0	0.5	past tense of like	0.5	liked

Εικόνα 6. 21 DefinedAs

_id	negStart	negEnd	rel	findEnd	posEnd	surfaceEnd	posStart	surfaceStart
5698c79e9b1f7...	0.25	0	/r/Antonym	0	0.25	fading	0	unfading
5698dc3b9b1f7...	0.625	0	/r/Antonym	0	0.625	concentrated	0	diluted

Εικόνα 6.22 Antonym

Οι νέες λέξεις που βρίσκουμε μέσω της παραπάνω διαδικασίας ονομάζονται Annotated Words και ο αριθμός τους ανέρχεται στις 36388.

```
('Annotated words:', 36388)

Process finished with exit code 0
```

Εικόνα 6. 23 Annotated Words

Όσον αφορά τα senti_results και concept_results προκύπτει το αποτέλεσμα της εικόνας 6.24.

```
The final result for SentiWordNet is: 1925.71572882
The final result for ConceptNet is: 1721.85543786
```

Εικόνα 6. 24 Final Results

Όπως παρατηρούμε στην εικόνα 6.24, η τιμή senti_results για το SentiWordnet είναι μεγαλύτερη από την concept_results για το ConceptNet. Συνεπώς, το annotated ConceptNet που δημιουργήθηκε είναι πιο αποδοτικό συγκριτικά με το Sentiwordnet.

6.7 Συμπεράσματα

Στην προκειμένη διπλωματική εργασία με την χρήση δύο εργαλείων του ConceptNet και του SentiWordNet δημιουργήσαμε το annotated ConceptNet, το οποίο περιέχει σχέσεις και την πολικότητα κάθε λέξης μίας σχέσης. Στην συνέχεια, αξιολογήσαμε ένα corpus, που περιέχει δεδομένα από το Twitter, μέσω των δύο εργαλείων, δηλαδή του SentiWordNet και του annotated ConceptNet. Μέσω δύο τιμών που υπολογίσαμε του senti_results και του concept_results, καταλήξαμε στο συμπέρασμα ότι το annotated ConceptNet είναι πιο αποδοτικό για την αξιολόγηση του αρχικού Corpus, καθώς το άθροισμα των διαφορών των απόλυτων τιμών του initial score, με την πολικότητα είναι μικρότερο έναντι του απλού SentiWordNet.

Μελλοντικά, θα μπορούσαμε να χρησιμοποιήσουμε το annotated Conceptnet και να επεκτείνουμε την έρευνα μας. Θα μπορούσαμε δυνητικά να χρησιμοποιήσουμε έναν classifier, όπως είναι ο Naïve Bayesian Classifier, να τον εκπαιδεύουμε και στο τέλος για κάποιο ποσοστό του corpus να προβλέπουμε το αποτέλεσμα. Τέλος, θα μετρούσαμε πόσο αποδοτικός είναι αυτός ο Classifier με την χρήση του accuracy.

Παράρτημα

Όπως αναφέρθηκε εντός της εν λόγω διπλωματικής εργασίας χρησιμοποιήθηκαν κάποια stopwords, τα οποία παρατίθενται αναλυτικά στο συγκεκριμένο παράρτημα.

a	f	my	Such
about	face	myself	sure
above	faces	n	t
across	fact	necessary	take
after	facts	need	taken
again	far	needed	than
against	felt	needing	that
all	few	needs	the
almost	find	never	their
alone	finds	new	them
along	first	newer	then
already	for	newest	there
also	four	next	therefore
although	from	no	these
always	full	nobody	they
among	fully	non	thing
an	further	noone	things
and	furthered	not	think
another	furthering	nothing	thinks
any	furthers	now	this
anybody	g	nowhere	those
anyone	gave	number	though
anything	general	numbers	thought
anywhere	generally	o	thoughts
are	get	of	three

area
areas
around
as
ask
asked
asking
asks
at
away
b
back
backed
backing
backs
be
became
because
become
becomes
been
before
began
behind
being
beings
best
better
between
big
gets
give
given
gives
go
going
good
goods
got
great
greater
greatest
group
grouped
grouping
groups
h
had
has
have
having
he
her
here
herself
high
higher
highest
him
himself
off
often
old
older
oldest
on
once
one
only
open
opened
opening
opens
or
order
ordered
ordering
orders
other
others
our
out
over
p
part
parted
parting
parts
per
perhaps
through
thus
to
today
together
too
took
toward
turn
turned
turning
turns
two
u
under
until
up
upon
us
use
used
uses
v
very
w
want
wanted
wanting
wants
was

both	his	place	way
but	how	places	ways
by	however	point	we
c	i	pointed	well
came	if	pointing	wells
can	important	points	went
cannot	in	possible	were
case	interest	present	what
cases	interested	presented	when
certain	interesting	presenting	where
certainly	interests	presents	whether
clear	into	problem	which
clearly	is	problems	while
come	it	put	who
could	its	puts	whole
d	itself	q	whose
did	j	quite	why
differ	just	r	will
different	k	rather	with
differently	keep	really	within
do	keeps	right	without
does	kind	room	work
done	knew	rooms	worked
down	know	s	working
downed	known	said	works
downing	knows	same	would
downs	l	saw	x
during	large	say	y
e	largely	says	year
each	last	second	years

early
either
end
ended
ending
ends
enough
even
evenly
ever
every
everybody
everyone
everything
everywhere
might
more
most
mostly
mr
mrs
much
later
latest
least
less
let
lets
like
likely
long
longer
longest
m
made
make
making
man
many
may
me
member
members
men
seconds
see
seem
seemed
seeming
seems
sees
several
shall
she
should
show
showed
showing
shows
side
sides
since
small
smaller
smallest
so
yet
you
young
younger
youngest
your
yours
z
some
somebody
someone
something
somewhere
state
states
still
must

Πηγές

1. Fraser-Smith, A. C., Bernardi, A., McGill, P. R., Ladd, M. E., Helliwell, R. A., & Villard Jr, O. G. (1990). Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. *Geophys. Res. Lett*, *17*(9), 1465-1468.

2. Campbell, W. H. (2009). Natural magnetic disturbance fields, not precursors, preceding the Loma Prieta earthquake. *Journal of Geophysical Research: Space Physics*, *114*(A5).

3. Thomas, J. N., Love, J. J., & Johnston, M. J. (2009). On the reported magnetic precursor of the 1989 Loma Prieta earthquake. *Physics of the Earth and Planetary Interiors*, *173*(3), 207-215.

4. Syed, U., Bowling, M., & Schapire, R. E. (2008, July). Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning* (pp. 1032-1039). ACM.

5. Βλαχάβας, Ι., Κεφαλάς, Π., Βασιλειάδης, Ν., Κόκκορας, Φ., & Σακελλαρίου, Η. (2005). Τεχνητή Νοημοσύνη: Μηχανική Μάθηση, Β Έκδοση. *Εκδόσεις Γαρταγάνη, Θεσσαλονίκη*.

6. Mitchell, T. M. (1997). Machine learning. WCB.

7. Witten, I. H., & Frank, E. (2000). Data mining: practical machine learning tools and techniques with Java implementations. 2000

8. Langley, P., &Carbonell, J. G. (1987). Machine learning.In S. Shapiro (Ed.), Encyclopedia of artificial inteligence. New Work: Wiley

9. Avron, B., & Feigenbaum, E. A. (1982). The handbook of artificial intelligence. William Kaufmann, Inc.

10. Κουτσίκου, Π. (2012). Τεχνητή Νοημοσύνη, σύντομη προσέγγιση. Retrieved from: http://users.sch.gr/jenyk/index.php/artificialintelligence

11. Haugeland, J. (1985). Artificial Intelligence: The Very Idea. Cambridge, Massachusetts: Bradford.

12. Rich, E., & Knight, K. (1991). Artificial intelligence. *McGraw-Hill, New*.

13. Luger, G. F., & Stubblefield, W. A. (1993). Artificial intelligence: its roots and scope. *Artificial intelligence: structures and strategies for Complex Problem Solving*, 1-34.

14. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. MIT Press.

15. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

16. Χαλκίδη Μ., & Βαζιργιάννης, Μ. (2005). Εξόρυξη γνώσης από βάσεις δεδομένων και τον παγκόσμιο ιστό. ΕκδόσειςΤυπωτήτω.

17. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 142-150). Association for Computational Linguistics.

18. Cheeseman, P., Self, M., Kelly, J., & Stutz, J. (1996). Bayesian Classification Theory and result. *Advances in Knowledge Discovery and Data Mining.AAAI Press*.

19. comScore / The Kelsey group. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. Press Release. Retrieved from: http://www.comscore.com/press/release.asp?press=1928

20. Horrigan, J. A. (2008). Online shopping. *Pew Internet & American Life Project Report*, *36*.

21. Rainie, L., & Horrigan, J. (2007). Election 2006 online. *Pew Internet & American Life Project Report*.

22. Hoffman, T. (2008). Online reputation management is hot—but is it ethical. *Computerworld, February*, 1-4.

23. Zabin, J., & Jefferies, A. (2008). Social media monitoring and analysis: Generating consumer insights from online conversation. *Aberdeen Group Benchmark Report*, *37*(9).

24. Kim, P., Anderson, E., & Joseph, J. (2006). The forrester wave: Brand monitoring. *Cambridge: Forrester Research*.

25. Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems* (no. rr-150). Yale university new haven conn dept of computer science.

26. Wilks, Y., & Bien, J. (1983). Beliefs, Points of View, and Multiple Environments. *Cognitive Science*, *7*(2), 95-119.

27. Hearst, M. A. (1992). Direction-based text interpretation as an information access refinement. *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, 257-274.

28. A. Huettner and P. Subasic, "Fuzzy typing for document management," in *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pp. 26–27, 2000.

Huettner, A., & Subasic, P. (2000). Fuzzy typing for document management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, 26-27.

29. Kantrowitz, M. (2003). *U.S. Patent No. 6,622,140*. Washington, DC: U.S. Patent and Trademark Office.

30. Sack, W. (1994, October). On the computation of point of view. In *AAAI* (p. 1488).

31. Wiebe, J. M., & Bruce, R. F. (2001). Probabilistic Classifiers for Tracking Point of View. *Theory, Method, and Practice in Computer Content Analysis*, *16*, 125.

32. Wiebe, J. M. (1990). Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics-Volume 2* (pp. 401-406). Association for Computational Linguistics.

33. Wiebe, J. M. (1990). Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics-Volume 2* (pp. 401-406). Association for Computational Linguistics.

34. Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246-253). Association for Computational Linguistics.

35. Wiebe, J. M., & Rapaport, W. J. (1988, June). A computational theory of perspective and reference in narrative. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics* (pp. 131-138). Association for Computational Linguistics.

36. Cardie, C., Wiebe, J., Wilson, T., & Litman, D. J. (2003). Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. In *New directions in question answering* (pp. 20-27).

37. Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)*, (35), 43.

38. Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, 519-528. ACM.

39. Dini, L., & Mazzini, G. (2002). Opinion classification through information extraction. In *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, 299-310.

40. Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on*

Intelligent user interfaces, 125-132. ACM.

41. Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 341-349. ACM.

42. Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, 70-77. ACM.

43. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10*, 79-86. Association for Computational Linguistics.

44. Tateishi, K., Ishiguro, Y., & Fukushima, T. (2001). Opinion information retrieval from the internet. *Information Processing Society of Japan (IPSJ) SIG Notes*, *69*(7), 75-82.

45. Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* 1, 6.

46. Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424. Association for Computational Linguistics.

47. Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D. S., & Maybury, M. (2003). Recognizing and Organizing Opinions Expressed in the World Press. In *New Directions in Question Answering*, 12-19.

48. Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 129-136. Association for Computational Linguistics.

49. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

50. Mosher, H. F., & Banfield, A. (1984). Unspeakable Sentences: Narration and Representation in the Language of Fiction. Routledge and Kegan Paul.

51. Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20*(2), 233-287.

52. Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

53.Carenini, G., Ng, R. T., & Pauls, A. (2006). Interactive multimedia summaries of evaluative text. In *Proceedings of the 11th international conference on Intelligent user interfaces*, 124-131. ACM.

54. Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference*, 427-434). IEEE.

55. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177. ACM.

56. Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. In *AAAI*, 22, 1331-1336.

57. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

58. Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

59. Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, 174-181. Association for Computational Linguistics.

60. Lin, D. (2003). Dependency-based evaluation of MINIPAR. In *Treebanks*, 317-329. Springer Netherlands.

61. Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, 1058-1065.

62. Tatemura, J. (2000). Virtual reviewers for collaborative exploration of movie reviews. In *Proceedings of the 5th international conference on Intelligent user interfaces*, 272-275. ACM.

63. Terveen, L., Hill, W., Amento, B., McDonald, D., & Creter, J. (1997). PHOAKS: A system for sharing recommendations. *Communications of the ACM*, *40*(3), 59-62.

64. Jin, X., Li, Y., Mah, T., & Tong, J. (2007). Sensitive webpage classification for content advertising. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, 28-33. ACM.

65. Riloff, E., Wiebe, J., & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On*

Artificial Intelligence 20, (3), 1106. Menlo Park, CA.

66. Somasundaran, S., Wilson, T., Wiebe, J., & Stoyanov, V. (2007). QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Intl. Conference on Weblogs and Social*.

67. Stoyanov, V., Cardie, C., & Wiebe, J. (2005). Multi-perspective question answering using the OpQA corpus. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 923-930. Association for Computational Linguistics.

68. Lita, L. V., Schlaikjer, A. H., Hong, W., & Nyberg, E. (2005). Qualitative dimensions in question answering: Extending the definitional QA task. In *Proceedings of the national conference on artificial intelligence*, 20(4), 1616. Menlo Park, CA.

69. Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y., & McNaught, J. (2007, January). Mining opinion polarity relations of citations. In *International Workshop on Computational Semantics (IWCS)*, 366-371.

70. Taboada, M., Gillies, M. A., & McFetridge, P. (2006). Sentiment classification techniques for tracking literary reputation. In *LREC workshop: towards computational models of literary analysis*, 36-43.

71. Liscombe, J., Riccardi, G., & Hakkani-Tur, D. (2005). Using context to improve emotion detection in spoken dialog systems. *Interspeech*, 1845–1848.

72. Tokuhisa, R., & Terashima, R. (2009). Relationship between utterances and enthusiasm in non-task-oriented conversational dialogue. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 161-167. Association for Computational Linguistics.

73. Lee, L. (2003). " I'm sorry Dave, I'm afraid I can't do that": Linguistics, Statistics, and Natural Language Processing circa 2001. In *Computer Science: Reflections on the Field, Reflections from the Field*, (Committee on the Fundamentals of Computer Science: Challenges and Opportunities, Computer Science and Telecommunications Board, National Research Council, ed.), 111–118, The National Academies Press.

74. Mishne, G., & Glance, N. S. (2006). Predicting Movie Sales from Blogger Sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 155-158.

75. Efron, M. (2004). Cultural orientation: Classifying subjective documents by cocitation analysis. In *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*, 41-48.

76. Goldberg, A. B., Zhu, X., & Wright, S. J. (2007). Dissimilarity in graph-based semi-supervised classification. In *International Conference on Artificial Intelligence and Statistics*, 155-162.

77. Hopkins, D., & King, G. (2007). Extracting systematic social science meaning from text. Retrieved from: http://gking. harvard. edu/files/words. pdf, 20(07).

78. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(02), 311-331.

79. Mullen, T., & Malouf, R. (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 159-162.

80. Bansal, M., Cardie, C., & Lee, L. (2008). The Power of Negative Thinking: Exploiting Label Disagreement in the Min-cut Classification Framework. In *COLING (Posters)*, 15-18.

81. Greene, S. C. (2007). *Spin: Lexical semantics, transitivity, and the identification of implicit sentiment*. University of Maryland.

82. Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 327-335. Association for Computational Linguistics.

83. Cardie, C., Farina, C., & Bruce, T. (2006). Using natural language processing to improve erulemaking: project highlight. In *Proceedings of the 2006 international conference on Digital government research*, 177-178. Digital Government Society of North America.

84. Kwon, N., Shulman, S. W., & Hovy, E. (2006). Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research*, 157-166. Digital Government Society of North America.

85. Shulman, S., Hovy, E., Callan, J., & Zavestoski, S. (2005). Language processing technologies for electronic rulemaking: A project highlight. In *Proceedings of the 2005 national conference on Digital government research*, 87-88. Digital Government Society of North America.

86. Rogers, E. M. (1962). *Diffusion of innovations*. Free Press, New York.

87. Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological review*, *63*(5), 277.

88. Vincent, B., Xu, L., Chesley, P., & Srhari, R. K. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia*, 27–29, 20.

89. Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T., & Joshi, A. (2007). Modeling trust and influence in the blogosphere using link polarity. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

90. Mahendran, A., Duraiswamy, A., Reddy, A., & Gonsalves, C. (2013). Opinion mining for text classification. *International Journal of Scientific Engineering and Technology*, *2*(6), 589-594.

91. Naive-Bayes Classification Algorithm. Retrieved from: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf

92. Sureka, A., Goyal, V., Correa, D., & Mondal, A. (2009). Polarity classification of subjective words using common-sense knowledge-base. In *Rough sets, fuzzy sets, data mining and granular computing*, 486-493. Springer Berlin Heidelberg.

93. Agarwal, B., & Mittal, N. (2016). Semantic Orientation-Based Approach for Sentiment Analysis. In *Prominent Feature Extraction for Sentiment Analysis*, 77-88. Springer International Publishing.

94. Agarwal, B., & Mittal, N. (2016). Machine Learning Approach for Sentiment Analysis. In *Prominent Feature Extraction for Sentiment Analysis*, 21-45. Springer International Publishing.

95. Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (6), 417-422.

96. Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *FLAIRS conference*, 202-207.

97. Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. *International Edition*.

98. Havasi, C., Speer, R., & Alonso, J. (2007). ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, 27-29.

99. Vom Brocke, J., & Rosemann, M. (2010). *Handbook on business process management*. Heidelberg: Springer.

100. Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, *1142*, 1-12.

101. Dumhill, E. (2012). What is big data? Retrieved from: http://strata.oreilly.com/2012/01/what-is-big-data.html

102. Gartner. Big Data Definition. Retrieved from: http://www.gartner.com/it-glossary/big-data/

103. Gartner. Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between Humans and Machines. Retrieved from: http://www.gartner.com/newsroom/id/2575515

104. McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

105. White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media, Inc.

106. Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co.

107. MIKE 2.0. Big Data Definition. Retrieved from: http://mike2.openmethodology.org/wiki/Big_Data_Definition

108. Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*. McGraw Hill Professional.

109. Gartner. Big Data Definition. Retrieved from: http://www.gartner.com/it-glossary/big-data/

110. BenoîtPerroud (2013). Apache cassandra. Retrieved from: http://fr.slideshare.net/benoitperroud/cassandra-talk-jug-lausanne-20120614

111. Wikipedia (2013). Fault-tolerant system. Retrieved from: http://en.wikipedia.org/wiki/Fault-tolerant_system

112. Lämmel, R. (2008). Google's MapReduce programming model—Revisited. *Science of computer programming*, *70*(1), 1-30.

113. Bernstein, P. A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems*, 370. New York: Addison-Wesley.

114. Derrick Harris (2013). "10gen embraces what it created, becomes MongoDB Inc.".

115. MongoDB. "Introduction to Replication".MongoDB.
Chodorow, K. (2013). *MongoDB: the definitive guide*. O'Reilly Media, Inc.

116. Gelernter, D. (2010). *The muse in the machine: Computerizing the poetry of human thought*. Simon and Schuster.

117. Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

118. Minsky, M. (1988). *Society of mind*. Simon and Schuster.

119. Liu, H., & Singh, P. (2004). Commonsense reasoning in and over natural language. In *Knowledge-based intelligent information and engineering systems*, 293-

306. Springer Berlin Heidelberg.

120. Liu, B. (2008). Opinion mining and summarization-sentiment analysis. *Tutorial in the Proceedings of WWW*, *8*.

121. Liu, H., & Singh, P. (2002, July). MAKEBELIEVE: Using commonsense knowledge to generate stories. In *AAAI/IAAI*, 957-958.

122. Wang, A. (2002). *Turn-taking in a collaborative storytelling agent*. Doctoral dissertation, Masters Thesis, MIT Department of Electrical Engineering and Computer Science.

123. Harrington, P. (2012). *Machine learning in action*, 28-178. Manning.

124. Bogomolny, A., (2016). Zipf's Law, Benford's Law. *Interactive Mathematics Miscellany and Puzzles*. Retrieved from: http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml

125. Δήμας, Α. (2013). *Τεχνικές για την εξαγωγή γνώσης από την πλατφόρμα του Twitter*. Πανεπιστήμιο Πατρών.

126. Wikipedia. (2015). Language model. Retrieved from: http://en.wikipedia.org/wiki/Language_model#Unigram_models

127. Lexalytics. Sentiment Analysis. Retrieved from: https://www.lexalytics.com/technology/sentiment. 2016

128. Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*, *2015*, 30.

129. Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, *22*(4), 211-226.

130. Princeton University. (2015). Wordnet: A lexical database for English. Retrieved from: https://wordnet.princeton.edu/

131. SentiWordNet. (2010) Retrieved from: http://sentiwordnet.isti.cnr.it/

132. Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315-346.

133. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC* 10, 2200-2204.

134. Wikipedia. (2015). Python. Retrieved from: https://el.wikipedia.org/wiki/Python