

**ΣΥΣΤΗΜΑ ΟΠΤΙΚΟΠΟΙΗΣΗΣ ΣΥΣΤΑΣΕΩΝ  
ΚΑΙ ΕΞΑΓΩΓΗΣ ΠΤΥΧΩΝ ΑΠΟ ΕΠΕΞΕΡΓΑΣΜΕΝΑ  
ΣΧΟΛΙΑ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ**

**ΠΡΩΙΟΣ ΔΗΜΗΤΡΙΟΣ ΑΜ: 21034  
ΦΟΙΤΗΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ  
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**



**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:  
ΜΙΧΑΗΛ ΔΗΜΗΤΡΙΟΣ  
ΤΣΕΡΠΕΣ ΚΩΝΣΤΑΝΤΙΝΟΣ**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ  
ΒΑΡΛΑΜΗΣ ΗΡΑΚΛΗΣ**

**ΑΜ:21034  
ΠΡΩΙΟΣ ΔΗΜΗΤΡΙΟΣ**

# ΠΕΡΙΕΧΟΜΕΝΑ

|                                                                                                 |    |
|-------------------------------------------------------------------------------------------------|----|
| <a href="#">ΠΕΡΙΕΧΟΜΕΝΑ</a>                                                                     | 1  |
| <a href="#">ΠΕΡΙΛΗΨΗ</a>                                                                        | 2  |
| <a href="#">I. ΕΙΣΑΓΩΓΗ</a>                                                                     | 4  |
| <a href="#">Α. Αναζήτηση προϊόντων στο διαδίκτυο</a>                                            | 4  |
| <a href="#">Β. Στοχευμένη διαφήμιση και η δυνατότητα εμπλουτισμού του μέσου</a>                 | 5  |
| <a href="#">Γ. Προτεινόμενη λύση</a>                                                            | 5  |
| <a href="#">Δ. Συσταδοποίηση βάσει των πτυχών που εμπεριέχονται στα σχόλια</a>                  | 6  |
| <a href="#">Ε. Περιγραφή λειτουργίας πρωτοτύπου</a>                                             | 6  |
| <a href="#">II. ΒΙΒΛΙΟΓΡΑΦΙΑ</a>                                                                | 8  |
| <a href="#">Α. Εξαγωγή πληροφορίας από κείμενο</a>                                              | 8  |
| <a href="#">Β. Λογισμικό που χρησιμοποιήθηκε για την συντακτική ανάλυση της φυσικής γλώσσας</a> | 11 |
| <a href="#">Γ. Ρόλος της στοχευμένης διαφήμισης στα κοινωνικά δίκτυα</a>                        | 15 |
| <a href="#">Δ. Προσεγγίσεις στοχευμένης διαφήμισης χρησιμοποιώντας συστήματα συστάσεων</a>      | 15 |
| <a href="#">III. ΠΡΟΤΕΙΝΟΜΕΝΟ ΜΟΝΤΕΛΟ</a>                                                       | 17 |
| <a href="#">Α. Διαδικασία του συστήματος προτάσεων</a>                                          | 19 |
| <a href="#">Β. Εξόρυξη πτυχής και γνώμης</a>                                                    | 23 |
| <a href="#">Γ. Συνάθροιση απόψεων για μια πτυχή</a>                                             | 24 |
| <a href="#">Δ. Μηχανή Συστάσεων</a>                                                             | 24 |
| <a href="#">Ε. Μοντέλο λεκτικής ανάλυσης</a>                                                    | 25 |
| <a href="#">IV. ΕΦΑΡΜΟΓΗ ΛΟΓΙΣΜΙΚΟΥ</a>                                                         | 28 |
| <a href="#">Α. Συνοπτική περιγραφή της εφαρμογής</a>                                            | 28 |
| <a href="#">Β. Περιγραφή συνόλου δεδομένων (dataset)</a>                                        | 30 |

|                                                                                                                                  |    |
|----------------------------------------------------------------------------------------------------------------------------------|----|
| <u>Γ. Front-End Application</u>                                                                                                  | 34 |
| <u>Δ. Back-End Application</u>                                                                                                   | 35 |
| <u>Python CoreNLP wrapper</u>                                                                                                    | 36 |
| <u>Σύνδεση με MongoDB</u>                                                                                                        | 37 |
| source <a href="https://api.mongodb.com/python/current/api/index.html">https://api.mongodb.com/python/current/api/index.html</a> | 37 |
| <u>Ε. Αρχιτεκτονική της εφαρμογής</u>                                                                                            | 40 |
| <u>ΣΤ. Σενάρια Χρήσης της εφαρμογής (Use Cases)</u>                                                                              | 41 |
| <u>Υ. Συμπεράσματα και μελλοντικές επεκτάσεις</u>                                                                                | 42 |
| <u>ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ</u>                                                                                                   | 43 |
| <u>ΠΑΡΑΡΤΗΜΑ Α</u>                                                                                                               | 49 |

# ΠΕΡΙΛΗΨΗ

Οι διαδικτυακές διαφημίσεις είναι μια σημαντική πηγή κέρδους και έλξης πελατών για επιχειρήσεις που βασίζονται στο διαδίκτυο. Μέσω μιας πετυχημένης διαφημιστικής εκστρατείας, χρήστες και επιχειρήσεις μπορούν να ωφεληθούν καθώς οι χρήστες αναμένεται να ανταποκριθούν θετικά σε ειδικές προσφορές και να ικανοποιήσουν τις ανάγκες τους με το κατάλληλο προϊόν, ενώ οι επιχειρήσεις είναι σε θέση να προσεγγίσουν καταναλωτές για προϊόντα ή υπηρεσίες που παρέχουν. Η εξαγωγή προτιμήσεων χρηστών των μέσων κοινωνικής δικτύωσης, μέσω σχολίων ή κριτικών προϊόντων μπορεί να αναδειχθεί σε πολύτιμο εργαλείο που θα μπορούσε: α) να συντελέσει στη βελτίωση της διαδικασίας αναζήτησης προϊόντων και β) να μετατραπεί σε μέσο διερεύνησης των αναγκών του αγοραστικού κοινού για τις επιχειρήσεις.

Σε αυτή την εργασία, προτείνεται μία μεθοδολογία για την ανάλυση περιεχομένου από κριτικές (σχόλια), και ένα πρόγραμμα που υλοποιεί αυτήν τη μεθοδολογία και εξετάζει πτυχές του κάθε προϊόντος όπως αυτές σχολιάστηκαν από χρήστες στο κοινωνικό δίκτυο Yelp. Η πρότυπη εφαρμογή που δημιουργήθηκε στα πλαίσια της εργασίας παρέχει ένα γραφικό περιβάλλον που μπορεί να χρησιμοποιηθεί με δύο διαφορετικές προοπτικές: α) της εύρεσης του κατάλληλου προϊόντος για τον χρήστη και β) τη γνωμοδότηση βάσει του κοινού για την ίδια την επιχείρηση αλλά και για το σύνολο του ανταγωνισμού μιας κατηγορίας, χρησιμοποιώντας μία διαδραστική αναπαράσταση γράφου.

Ο γράφος για τις επιχειρήσεις, αποκαλύπτει τα αδύνατα και ισχυρά σημεία της, ενώ ο γράφος που απευθύνεται στον τελικό χρήστη χρησιμοποιείται για να πραγματοποιεί προτάσεις επιχειρήσεων που έχουν βαθμολογηθεί για τα χαρακτηριστικά τα οποία αναζητεί ο χρήστης. Στη πρώτη περίπτωση, ο γράφος βασίζεται στη βαθμολογία που συναθροίζουν οι θετικά ή αρνητικά φορτισμένες λέξεις που περιέχονται σε ένα κριτικό κείμενο και στις επιμέρους απόψεις που σχολιάζονται σε αυτό. Ενώ στη δεύτερη βασίζεται στη συνάθροιση όλων κριτικών για μία επιχείρηση. Αυτή η γραφική απεικόνιση βασίζεται στην εκτίμηση ενός σχολίου ή μιας κριτικής για το οποίο εξάγεται μια βαθμολογία, η οποία εν τέλει συνδέεται με ένα χρήστη, τόσο για το προϊόν προς βαθμολόγηση όσο και για τα χαρακτηριστικά για τα οποία βαθμολογήθηκε.

Η πρότυπη εφαρμογή χρησιμοποιεί κριτικές που απευθύνονται σε προϊόντα επιχειρήσεων όπως αυτές παρέχονται από χρήστες (σε μορφή γραπτών σχολίων). Για την ανάλυση αυτών των κριτικών γίνεται χρήση μηχανισμών επεξεργασίας φυσικής γλώσσας με σκοπό το προσδιορισμό των απόψεων χρηστών για συγκεκριμένες πτυχές ενός αντικείμενου (προϊόν ή υπηρεσία). Αυτές οι πτυχές καθώς και η γνώμη (sentiment) που συνοδεύει την κάθεμία, μπορούν να χρησιμοποιηθούν για τη δημιουργία προφίλ χρηστών και επιχειρήσεων, που μπορούν στη συνέχεια να χρησιμοποιηθούν για την πραγματοποίηση συστάσεων (recommendation system) ή ως εργαλείο ανάλυσης της αγοράς στην οποία μετέχει μία επιχείρηση. Η εφαρμογή μπορεί να χρησιμοποιηθεί ως βάση για τη δημιουργία ενός ισχυρού εργαλείου και την εκτέλεση μιας επιτυχούς εκστρατείας μάρκετινγκ μέσω των διαδικτυακών μέσων.

Τέλος, περιγράφεται με λεπτομέρεια η υλοποίηση της εφαρμογής, ο τρόπος που χρησιμοποίησε δεδομένα του κοινωνικού δικτύου Yelp και πως θα μπορούσε να εφαρμοστεί ως υπηρεσία σε υπάρχοντα συστήματα με δομή κοινωνικών δικτύων είτε ως εφαρμογή μέσα στο οικοσύστημα των περιηγητών διαδικτύου (browser plugins). Είναι σημαντικό να αναφερθεί ότι η μέθοδος που περιγράφεται σε αυτήν την εργασία και ένα πρώιμο πρωτότυπο της εφαρμογής παρουσιάστηκε στην ημερίδα Social Media and Risk (SOMERIS 2015)<sup>1</sup>

---

<sup>1</sup> <http://www.lix.polytechnique.fr/dascim/someris2015/>

που οργανώθηκε στα πλαίσια του διεθνούς συνεδρίου *Advances in Social Network Analysis and Mining (ASONAM 2015)* σε συνεργασία με τον επιβλέποντα καθηγητή κ. Ηρακλή Βαρλάμη και την καθηγήτρια του πανεπιστημίου του San Jose State University, Μαγδαληνή Ειρηνάκη.

**Λέξεις-κλειδιά:** advertising, personalization, targeted ads, recommender, Yelp

# I. ΕΙΣΑΓΩΓΗ

## A. Αναζήτηση προϊόντων στο διαδίκτυο

Η εξάπλωση των υπηρεσιών κοινωνικής δικτύωσης δημιούργησε την ανάγκη σε όλο και περισσότερες επιχειρήσεις να πραγματοποιούν πωλήσεις ή να διαφημίζουν προϊόντα και υπηρεσίες τους στο διαδίκτυο. Η διαφήμιση αποτελεί την κύρια δραστηριότητα πολλών από τις επιχειρήσεις κολοσσούς του διαδικτύου (Google, Facebook). Αλλά καθώς η εικόνα που παρουσιάζουν οι επιχειρήσεις μέσω των διαφημίσεων δεν ανταποκρίνεται πάντα στις καταναλωτικές προσδοκίες, καθώς οι υποψήφιοι πελάτες δείχνουν περισσότερη εμπιστοσύνη στην γνώμη άλλων χρηστών ενός προϊόντος. Αυτό αποτελεί ένα είδος παραλληλισμού της 'δια στόματος' (word of mouth) διάδοσης της πληροφορίας, η οποία υπάρχει άφθονη στο διαδίκτυο στη μορφή κριτικών (reviews) σε διάφορα κοινωνικά δίκτυα, e-shops, forums κλπ. Η απλή κατάταξη προϊόντων βάσει κόστους, χαρακτηριστικών ή βαθμολογίας χρηστών δεν επαρκεί για τη λήψη μιας τεκμηριωμένης απόφασης. Έτσι είναι σύνθηρες το φαινόμενο, χρήστες να καλούνται να διερευνήσουν σε βάθος ένα προϊόν ή υπηρεσία μέσα από την ανάγνωση κριτικών, ελπίζοντας να βρουν πληροφορίες για τις διαφορές πτυχές ενός προϊόντος ή υπηρεσίας που τους ενδιαφέρουν, μία διαδικασία που μπορεί να αποδειχθεί ιδιαίτερα χρονοβόρα και αμφίβολη ως προς το αποτέλεσμα καθώς είναι αδύνατη η συνολική η εικόνα όλων των χρηστών με ακρίβεια.

Οι περισσότερες από τις ιστοσελίδες που παρέχουν κριτικές προϊόντων, επιχειρήσεων ή υπηρεσιών, συνήθως απεικονίζουν αξιολογήσεις σε αντίστροφη χρονολογική σειρά, ξεκινώντας από τη πιο πρόσφατη, είτε με τη χρήση ευρετικών μετρικών (heuristic metrics), για παράδειγμα κατατάσσοντας υψηλότερα κριτικές του "σούπερ χρηστών", δηλαδή χρήστες με πολλά σχόλια που χαρακτηρίζονται ως «χρήσιμοι» για την κοινότητα με βάση της βαθμολογίας των ίδιων των σχολίων ως χρήσιμα από άλλους χρήστες της ίδιας διαδικτυακής κοινότητας.

Αλλά ακόμα και αυτές οι μέθοδοι μέτρησης της αξιοπιστίας των σχολίων από τους χρήστες, δεν επαρκούν, καθώς παρότι αποτελούν κοινά αποδεκτές γνώμες, δεν δύνανται να αντικατοπτρίσουν τα ακριβή χαρακτηριστικά που ενδιαφέρουν κάθε χρήστη, με αποτέλεσμα οι χρήστες να πρέπει να διαβάσουν ή τουλάχιστον να "σαρώσουν" δεκάδες σχόλια για προϊόντα, επιχειρήσεις και υπηρεσίες με τη προοπτική να βρουν επαρκή πληροφορία για πτυχές που τους ενδιαφέρουν.

## B. Στοχευμένη διαφήμιση και η δυνατότητα εμπλουτισμού του μέσου

Ο όρος "στοχευμένη διαφήμιση", χρησιμοποιείται προκειμένου να περιγραφεί η χρήση των πιο πολλά υποσχόμενων χαρακτηριστικών ενός προϊόντος που θα προσελκύσουν ένα συγκεκριμένο καταναλωτή. Αυτά τα χαρακτηριστικά μπορούν να εξαχθούν από περιεχόμενο που ο ίδιος ο χρήστης παρέχει. Η ομαδοποίηση χρηστών ως προς ένα σύνολο χαρακτηριστικών επιτρέπει σε μια επιχείρηση να πραγματοποιήσει εκστρατείες προσωπικών/στοχευμένων διαφημίσεων, προωθώντας προϊόντα σε επιλεγμένους καταναλωτές βάσει δημογραφικών γνωρισμάτων, ψυχογραφικών χαρακτηριστικών κλπ. συνδυαστικά με τη γενικότερη συμπεριφορά ενός χρήστη (π.χ. ιστορικό αγορών, πλοήγηση στον ιστότοπο, cookies κλπ).

Λόγω της ανάγκης για στοχευμένη διαφήμιση, η αναγνώριση των πτυχών των διάφορων προϊόντων αναγκάζεται σε ρόλο κλειδί για τη κατανόηση της αγοράς, καθώς επιτρέπει σε μια επιχείρηση να κατανοήσει

τα δυνατά και αδύνατα σημεία της και να ανιχνεύσει αντίστοιχα των ανταγωνιστών της αναγνωρίζοντας έτσι τις απειλές και τις ευκαιρίες που υπάρχουν σε μία αγορά. Αποτελεί εργαλείο δηλαδή για μια πληρέστερη εικόνα του μοντέλου SWOT (Δυνάμεις-Strengths, Αδυναμίες-Weaknesses, Ευκαιρίες-Opportunities, Απειλές-Threats). Αυτό επιτρέπει τη λήψη αποφάσεων βάσει δεδομένων (data-driven decisions) και τη βελτιστοποίηση της θέσης των προϊόντων μιας επιχείρησης. Αυτές οι πτυχές μπορούν να εξαχθούν από κριτικές χρηστών βλέποντας τι περιγράφει ο κάθε χρήστης στη κριτική εξατομικευμένα. Γνώση που μπορεί να αποκομισθεί από τις κριτικές του καταναλωτικού κοινού είναι:

- i) ο τρόπος που ομαδοποιούνται οι καταναλωτές,
- ii) ποια είναι τα συμφέροντα και οι προτιμήσεις της κάθε ομάδας που απαρτίζει το καταναλωτικό κοινό
- iii) ποιος είναι ο συσχετισμός γνώμης έναντι των ανταγωνιστικών προϊόντων,
- iv) ποια είναι τα πλεονεκτήματα τους και μειονεκτήματα των προϊόντων βάσει της γνώμης των τελικών χρηστών.

Η χρήση των παραπάνω πληροφοριών επιτρέπει στις εταιρείες να επικεντρωθούν σε συγκεκριμένες ομάδες χρηστών, δημιουργώντας στοχευμένα προϊόντα και προωθητικές διαφημίσεις επισημαίνοντας τις πτυχές ενός προϊόντος που αφορούν ένα συγκεκριμένο χρήστη. Έτσι στον κάθε χρήστη μπορεί να παρουσιάζεται ένα προϊόν υπό το πρίσμα των πτυχών που τον ενδιαφέρουν, αυτές οι πτυχές μπορεί να ναι κριτήρια που χρησιμοποιούνται ευρέως (π.χ. χαμηλή τιμή), είτε πολύ πιο συγκεκριμένες γνωρίζοντας περισσότερες λεπτομέρειες για τον χρήστη.

## Γ. Προτεινόμενη λύση

Το μοντέλο που προτείνεται στην παρούσα εργασία στοχεύει στην ανάλυση των κριτικών για ένα προϊόν ή υπηρεσία, στην εξαγωγή των πτυχών καθώς και τον εντοπισμό της χροιάς του συναισθήματος (sentiment) με την οποία μια πτυχή αναφέρεται (θετική/αρνητική), βάσει τεχνικών εξόρυξης γνώμης, επεξεργασίας φυσική γλώσσας και εν τέλει μηχανισμών συσταδοποίησης των χρηστών. Γίνεται μια προσπάθεια να υποβοηθηθούν και τα δύο υποκείμενα που εμπλέκονται στο μηχανισμό συστάσεων, δηλαδή ο **επιχειρηματικός κόσμος** και οι **τελικοί χρήστες**. Στη παρούσα εργασία γίνεται η προσπάθεια ενδυνάμωσης του μέσου της στοχευμένης διαφήμισης με τη χρήση δεδομένων και προτιμήσεων ενός χρήστη βάσει περιεχομένου που παρέχει ο ίδιος ή βάσει της ομαδοποίησης του με άλλους χρήστες. Αυτές οι προτιμήσεις μπορούν να συνδυαστούν ακόμα και με πιο “παραδοσιακές” ενδείξεις, όπως τα δημογραφικά στοιχεία, η τοποθεσία του χρήστη, κλπ. Τέτοια συστήματα συνδυασμού προτιμήσεων με άλλες ενδείξεις έχουν περιγραφεί στη βιβλιογραφία Bao, J. et al (2012), des Vignes, G.V (2014), Rambocas, M et al (2013).

Διαφορετικοί χρήστες εκτιμούν διαφορετικές πτυχές για το ίδιο προϊόν. Για παράδειγμα, όταν ψάχνουν για μια ψηφιακή φωτογραφική μηχανή, κάποιος μπορεί να ενδιαφέρεται για την τιμή και το μέγεθος, ενώ ένας άλλος χρήστης μπορεί να εκτιμά περισσότερο την ευκολία χρήσης. Ομοίως, όταν ψάχνουν για ένα καλό ιταλικό εστιατόριο, ένας χρήστης θα μπορούσε να αποτιμήσει την ατμόσφαιρα και τη λίστα κρασιών ενός τόπου, ενώ κάποιος άλλος μπορεί να προτιμά εστιατόρια που είναι φιλικά προς οικογένειες με μικρά παιδιά.

Στην ιδανική περίπτωση, οι επιχειρήσεις θα ήθελαν να στοχεύσουν σε πελάτες που εκτιμούν τις ιδιότητες των προϊόντων στις οποίες υπερέχουν έναντι του ανταγωνισμού. Κάτι τέτοιο μπορεί να πραγματοποιηθεί αναδεικνύοντας τις κατάλληλες πτυχές όπως αυτές κυκλοφορούν μέσω των σχολίων στο διαδίκτυο από τον κάθε χρήστη. Την ίδια στιγμή οι καταναλωτές μπορούν να επωφεληθούν με προϊόντα τα οποία ταιριάζουν απόλυτα στα κριτήριά τους κερδίζοντας πολύτιμο χρόνο από την αναζήτηση.

#### **Δ. Συσταδοποίηση βάσει των πτυχών που εμπεριέχονται στα σχόλια**

Όπως αναφέρθηκε, πληροφορίες των διαφόρων πτυχών προϊόντος μπορούν να εξαχθούν από τις παρατηρήσεις και τα σχόλια που αφήνουν οι χρήστες στο διαδίκτυο. Χρησιμοποιώντας αλγορίθμους εξαγωγής γνώμης από κείμενα, οι πτυχές της κάθε κριτικής, καθώς και η πολικότητα που τους αντιστοιχεί μπορούν να εντοπιστούν και να χρησιμοποιηθούν για το σχηματισμό του προφίλ χρηστών και ενός αντικειμένου (π.χ. επιχειρήσεων).

Χρησιμοποιώντας αυτά τα προφίλ, μπορούμε στη συνέχεια να ταιριάξουμε χρήστες σε επιχειρήσεις, με τη χρήση του περιεχομένου των αξιολογήσεων τους (προτάσεις με βάση το περιεχόμενο), την ομοιότητα μεταξύ τους (φιλτράρισμα βάσει της ομαδοποίησης χρηστών και αντικειμένων), καθώς και αλγορίθμους συσταδοποίησης με σκοπό τον εντοπισμό ομάδων χρηστών για χαρακτηριστικά που ενδιαφέρονται ή επιχειρήσεων ως προς τα χαρακτηριστικά με τα οποία τα προϊόντα τους αναφέρονται. Η διαφορά των δύο τελευταίων κριτηρίων “ταιριάσματος” είναι ότι στη πρώτη περίπτωση γίνεται απλό ταίριασμα σε επίπεδο πτυχών προϊόντων, ενώ η σύνδεση στη δεύτερη περίπτωση γίνεται βάσει αλγορίθμων συσταδοποίησης.

Για παράδειγμα στη πρώτη περίπτωση κριτήριο η σύνδεση μπορεί να γίνει επειδή και οι 2 χρήστες σχολίασαν τη πτυχή «οίνος» στη κατηγορία αντικειμένου «ιταλικό εστιατόριο», ενώ στη δεύτερη μια σύνδεση μπορεί να γίνει επειδή ο αλγόριθμος συσταδοποίησης τους ομαδοποίησε με αυτό το τρόπο, πιθανόν λόγω ενός συνδυασμού χαρακτηριστικών (πχ. Παρόμοια ηλικία, περιοχή, πολιτιστικά χαρακτηριστικά κλπ), αυτή η σύνδεση δεν είναι προφανής εκ πρώτης όψης αλλά έγινε λόγω εξαγωγής γνώσης από δεδομένα.

#### **Ε. Περιγραφή λειτουργίας πρωτοτύπου**

Βάσει της παραπάνω περιγραφής συσταδοποίησης δημιουργήθηκε ένα σύστημα, το οποίο είχε ως απώτερο στόχο να παρέχεται ως εργαλείο για το σχεδιασμό και την εκτέλεση μιας επιτυχούς εκστρατείας μάρκετινγκ. Η πρότυπη υλοποίηση έγινε για την επίδειξη των δυνατοτήτων αυτού του θεωρητικού συστήματος. Στα παραδείγματα που δοκιμάσαμε το σύστημα, χρησιμοποιώντας ένα σύνολο δεδομένων από το κοινωνικό δίκτυο Yelp, καταφέραμε να εντοπίσουμε συγκεκριμένα στοιχεία/ πτυχές στις διάφορες επιχειρήσεις αυτοματοποιώντας την αναζήτηση τέτοιων πληροφοριών.

Πιο συγκεκριμένα, στα πειράματά μας έγινε χρήση μικρών σε έκταση σχολίων (tips) που παρέχονται από τους χρήστες του Yelp ως κριτικές προς τις επιχειρήσεις και όχι των αναλυτικών κριτικών που κάνουν οι χρήστες για μια επιχείρηση. Τα tips είναι σύντομα κείμενα με σχόλια όπου αναφέρονται οι καλές ή κακές πτυχές μιας επιχείρησης με τη χρήση επιθέτων που εκθέτουν θετική ή αρνητική γνώμη. Για παράδειγμα, σε μια τέτοια μίνι- κριτική για ένα εστιατόριο όπως "εξαιρετικό ιταλικό φαγητό, μέτριο κρασί ... :(" είναι σχετικά εύκολο να εξαχθεί ως πληροφορία ότι το "ιταλικό φαγητό" είναι η καλή πλευρά του εστιατορίου, ενώ το "κρασί" είναι η κακή πλευρά. Σε στην περίπτωση αυτή, θα είμαστε σε θέση να προτείνουμε αυτό το εστιατόριο σε έναν άλλο χρήστη, ο οποίος ενδιαφέρεται πρωτίστως για ιταλικό φαγητό. Όπως θα ήταν επίσης χρήσιμο να προωθηθεί αυτή η πληροφορία τόσο στην ίδια την εταιρεία, όσο και στους ανταγωνιστές της προκειμένου να βελτιωθεί η ποικιλία κρασιών και στους ανταγωνιστές τους, προκειμένου να αναδείξουν την ανωτερότητα τους στην πτυχή κρασί σε αυτόν το χρήστη εφόσον δεν ήταν ικανοποιημένος σε αυτό το σενάριο.

Η συνολική διαδικασία που ακολουθήθηκε στην εργασία αποτελεί εφαρμογή των βασικών αρχών εξόρυξης άποψης για τις πτυχές μιας κριτικής και επεκτείνεται στη δημιουργία εξατομικευμένων προτάσεων που θα μπορούσαν να χρησιμοποιηθούν σε διαφημιστικές εκστρατείες. Τα αποτελέσματα της εξόρυξης

παρουσιάζονται με τη βοήθεια ενός γραφικού περιβάλλοντος, το οποίο έχει ως στόχο να εξυπηρετήσει τόσο τους χρήστες όσο και τις επιχειρήσεις.

Το υπόλοιπο της εργασίας οργανώνεται ως εξής: στο δεύτερο τμήμα περιγράφεται η συναφής βιβλιογραφία πάνω στον τομέα της στοχευμένης διαφήμισης και οι σύγχρονες προσεγγίσεις πάνω στο θέμα. Ταυτόχρονα γίνεται επισκόπηση των τεχνικών εξόρυξης γνώμης από κριτικές και των εφαρμογών που υλοποιούν τέτοιου είδους ανάλυση και παρουσιάζονται τα βασικά βήματα για την υλοποίηση μιας τέτοιας διαδικασίας. Στο τρίτο τμήμα, περιγράφεται το προτεινόμενο μοντέλο και η διαδικασία για τη συλλογή, ανάλυση και αξιοποίησης των κριτικών των χρηστών προκειμένου να βελτιστοποιηθεί η στόχευση της διαφήμισης σε συγκεκριμένους χρήστες. Στο τέταρτο τμήμα, παρέχονται λεπτομέρειες της εφαρμογής του πρωτοτύπου τόσο του μέρους της επεξεργασίας των δεδομένων όσο και της γραφικής διεπαφής (οπτικοποίησης των δεδομένων σε μορφή γράφου) και παρουσιάζονται κάποια σενάρια χρήσης της εφαρμογής αυτής. Ενώ τέλος, στο πέμπτο τμήμα συνοψίζονται επόμενα βήματα που θα μπορούσαν να επεκτείνουν τη παρούσα εφαρμογή.

## II. ΒΙΒΛΙΟΓΡΑΦΙΑ

### A. Εξαγωγή πληροφορίας από κείμενο

Υπάρχουν πολλές ενδιαφέρουσες δημοσιεύσεις που εστιάζουν στην εξαγωγή πληροφορίας από μια κριτική σε μορφή γραπτού κειμένου, η εργασία του Liu, B., (2012). καθώς και η εργασία των Pontiki et al. (2015)] συγκεκριμένα εστιάζουν στον εντοπισμό πτυχών και την πολικότητα τους (θετική/αρνητική) χρησιμοποιώντας αυτά τα αποτελέσματα ως ένα εργαλείο στον σημασιολογικό προσανατολισμό μιας κριτικής. Αυτή η διαδικασία μπορεί να πραγματοποιηθεί σε επίπεδο λέξεων μιας κριτικής Ding, X., et al (2008), Eiriraki, M. et al (2012), ή ακόμα και σε επίπεδο φράσεων Pontiki, M. et al (2014). Η εξαγόμενη πληροφορία μπορεί στη συνέχεια να χρησιμοποιηθεί για την παροχή διαισθητικών οπτικοποιήσεων των δεδομένων.

Πολλές εφαρμογές που εστιάζουν στην ανάλυση κριτικών ξενοδοχείων , χαρακτηριστικά οι Bjørkelund, (2012), καθώς και οι Wu, et al (2010), αναγνωρίζουν τη χρηστικότητα των χαρακτηριστικών / στοιχείων ως καταλυτικό παράγοντα στη λήψη αποφάσεων των χρηστών και εισάγουν οπτικοποιήσεις της συμπυκνωμένης αυτής πληροφορίας (infographics) των κριτικών των πελατών. Ωστόσο, δεν επωφελούνται αυτής της δυνατότητας για τη πραγματοποίηση προτάσεων σε χρήστες ή για την ανάλυση των επιχειρηματικών στρατηγικών μιας επιχείρησης. Σε πιο πρόσφατη βιβλιογραφία, ερευνητές στο τομέα της ανάλυσης κοινωνικών δικτύων και μηχανών συστάσεων έχουν αρχίσει να πραγματεύονται με τις διάφορες πτυχές που υπάρχουν σε μια κριτική και τις αντίστοιχες βαθμολογίες γνώμης για καθεμία από αυτές τις πτυχές κατά τη διαδικασία της σύστασης.

Οι συγγραφείς McAuley, και Leskovec (2013), με σκοπό τον εντοπισμό θεματικών ενός κειμένου δημιούργησαν το μοντέλο HFT (Hidden Factors as Topics) που βασίζεται στο συνδυασμό των ακόλουθων τεχνικών: α) Latent-Factor Recommender Systems (σχήμα 1) ενός μοντέλου πρόβλεψης τη γνώμη ενός χρήστη για ένα αντικείμενο  $rec(u,i)$  βάσει της τυχαιότητας  $\alpha$ , τους παράγοντες επιρροής του αντικειμένου προς σχολιασμό  $\beta_i$  και του ίδιου χρήστη  $\beta_u$  και τέλος των παραγόντων  $K$ -διαστάσεων  $\gamma_u$  του χρήστη και  $\gamma_i$  του αντικειμένου,

β) LDA (Latent Dirichlet Allocation) (σχήμα 2) ένα πιθανοτικό μοντέλο που εντοπίζει τη πιθανότητα αναφοράς  $\Theta_{d,k}$  ενός κειμένου  $d$  στο θέμα  $k$  από ένα σύνολο πιθανών θεμάτων. Δεδομένου ότι κάθε θέμα είναι συνδεδεμένο με μία κατανομή λέξεων  $\Phi_k$  που αντιπροσωπεύει τη πιθανότητα μία λέξη να χρησιμοποιείται σε ένα θέμα απ'όλα τα πιθανά θέματα και δεδομένης της πιθανότητας να βρεθεί ένα θέμα σ'ένα κείμενο  $\Theta_{z,d,j}$  και της πιθανότητας να βρεθεί μία λέξη σ'ένα θέμα.

$$rec(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i,$$

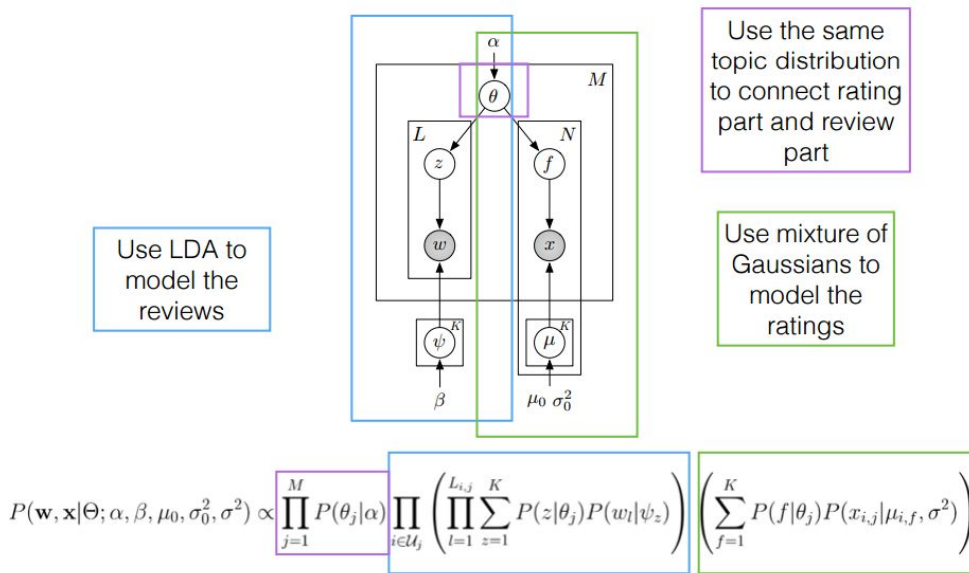
σχήμα 1

$$p(T|\theta, \phi, z) = \prod_{d \in T} \prod_{j=1}^{N_d} \theta_{z,d,j} \phi_{z,d,j} w_{d,j},$$

σχήμα 2

Το μοντέλο που δημιουργείται από το συνδυασμό των δύο προαναφερόμενων ονομάζεται HFT (Hidden Factors as Topics) και αποπειράται να ανακαλύψει την πιθανότητα κάθε πτυχής να συνδέεται με τους κρυφούς παράγοντες προϊόντων και χρηστών  $\gamma_u$  και  $\gamma_{i,l}$

Οι Lyu et al (2014) αναφέρονται στο πρόβλημα “cold-start”, δηλαδή την ανάγκη για συστάσεις χωρίς προηγούμενη γνώση για τον χρήστη και τις προτιμήσεις του και για τον λόγο αυτό προσπαθούν να εμπλουτίσουν τα συστήματα συστάσεων μέσω των κριτικών. Για το σκοπό αυτό δημιούργησαν το μοντέλο επεξεργασίας κριτικών RMR (Ratings Meet Reviews,) χρησιμοποιώντας την τεχνική LDA η οποία αναφέρθηκε παραπάνω για μοντελοποίηση των κριτικών. Χρησιμοποίησαν γκαουσιανές συναρτήσεις για την μοντελοποίηση των βαθμολογιών και συνδύασαν τα δύο εξαγόμενα μοντέλα κριτικών και βαθμολογιών στην ίδια θεματική κατανομή. (σχήμα 3)



Σχήμα 3

Σε ένα πίνακα σύγκρισης των μοντέλων που περιγράφηκαν στις δύο αυτές δημοσιεύσεις οι τελευταίοι βρέθηκαν να έχουν μεγαλύτερη ακρίβεια στην εύρεση της θεματικής των σχολίων (σύγκριση d-RMR, e-HFT) (Σχήμα 4):

| Dataset                     | a                   | b            | c                   | d                   | e                   | Improvement of RMR versus |        |        |
|-----------------------------|---------------------|--------------|---------------------|---------------------|---------------------|---------------------------|--------|--------|
|                             | MF                  | LDAMF        | CTR                 | HFT                 | RMR                 | min(a,b)                  | c      | d      |
| Arts                        | 1.565 (0.04)        | 1.575 (0.04) | 1.471 (0.04)        | 1.390 (0.04)        | <b>1.371 (0.04)</b> | 14.15%                    | 7.29%  | 1.39%  |
| Jewelry                     | 1.257 (0.03)        | 1.279 (0.03) | 1.206 (0.03)        | 1.177 (0.02)        | <b>1.160 (0.02)</b> | 8.36%                     | 3.97%  | 1.47%  |
| Industrial Scientific       | 0.461 (0.02)        | 0.462 (0.02) | 0.382 (0.02)        | <b>0.359 (0.02)</b> | 0.362 (0.02)        | 27.35%                    | 5.52%  | -0.83% |
| Watches                     | 1.535 (0.03)        | 1.518 (0.03) | 1.491 (0.03)        | 1.488 (0.03)        | <b>1.458 (0.02)</b> | 4.12%                     | 2.26%  | 2.06%  |
| Cell Phones and Accessories | 2.230 (0.04)        | 2.308 (0.04) | 2.177 (0.04)        | 2.135 (0.03)        | <b>2.085 (0.03)</b> | 6.95%                     | 4.41%  | 2.40%  |
| Musical Instruments         | 1.506 (0.02)        | 1.520 (0.02) | 1.422 (0.02)        | 1.395 (0.02)        | <b>1.374 (0.02)</b> | 9.61%                     | 3.49%  | 1.53%  |
| Software                    | 2.409 (0.02)        | 2.214 (0.02) | 2.254 (0.02)        | 2.219 (0.02)        | <b>2.173 (0.02)</b> | 1.89%                     | 3.73%  | 2.12%  |
| Gourmet Foods               | 1.515 (0.01)        | 1.491 (0.01) | 1.482 (0.01)        | <b>1.457 (0.01)</b> | 1.465 (0.01)        | 1.77%                     | 1.16%  | -0.55% |
| Office Products             | 1.814 (0.01)        | 1.796 (0.01) | 1.733 (0.01)        | 1.669 (0.01)        | <b>1.638 (0.01)</b> | 9.65%                     | 5.80%  | 1.89%  |
| Automotive                  | 1.570 (0.01)        | 1.585 (0.01) | 1.492 (0.01)        | 1.432 (0.01)        | <b>1.403 (0.01)</b> | 11.90%                    | 6.34%  | 2.07%  |
| Patio                       | 1.771 (0.01)        | 1.793 (0.01) | 1.720 (0.01)        | 1.698 (0.01)        | <b>1.669 (0.01)</b> | 6.11%                     | 3.06%  | 1.74%  |
| Pet Supplies                | 1.700 (0.01)        | 1.700 (0.01) | 1.613 (0.01)        | 1.583 (0.01)        | <b>1.562 (0.01)</b> | 8.83%                     | 3.27%  | 1.34%  |
| Beauty                      | 1.399 (0.01)        | 1.414 (0.01) | 1.361 (0.01)        | 1.358 (0.01)        | <b>1.334 (0.01)</b> | 4.87%                     | 2.02%  | 1.80%  |
| Shoes                       | 0.305 (0.00)        | 0.335 (0.00) | 0.271 (0.00)        | <b>0.247 (0.00)</b> | 0.251 (0.00)        | 21.51%                    | 7.97%  | -1.59% |
| Kindle Store                | 1.553 (0.01)        | 1.561 (0.01) | 1.457 (0.01)        | 1.437 (0.01)        | <b>1.412 (0.01)</b> | 9.99%                     | 3.19%  | 1.77%  |
| Clothing and Accessories    | 0.393 (0.00)        | 0.406 (0.00) | 0.355 (0.00)        | 0.349 (0.00)        | <b>0.336 (0.00)</b> | 16.96%                    | 5.65%  | 3.87%  |
| Health                      | 1.615 (0.01)        | 1.608 (0.01) | 1.552 (0.01)        | 1.538 (0.01)        | <b>1.512 (0.01)</b> | 6.35%                     | 2.65%  | 1.72%  |
| Toys and Games              | 1.467 (0.01)        | 1.395 (0.01) | 1.389 (0.01)        | <b>1.370 (0.01)</b> | 1.372 (0.01)        | 1.68%                     | 1.24%  | -0.15% |
| Tools and Home Improvement  | 1.600 (0.01)        | 1.610 (0.01) | 1.513 (0.01)        | 1.510 (0.01)        | <b>1.491 (0.01)</b> | 7.31%                     | 1.48%  | 1.27%  |
| Sports and Outdoors         | 1.219 (0.01)        | 1.223 (0.01) | 1.150 (0.01)        | 1.138 (0.01)        | <b>1.129 (0.01)</b> | 7.97%                     | 1.86%  | 0.80%  |
| Video Games                 | 1.610 (0.01)        | 1.608 (0.01) | 1.572 (0.01)        | 1.528 (0.01)        | <b>1.510 (0.01)</b> | 6.49%                     | 4.11%  | 1.19%  |
| Home and Kitchen            | 1.628 (0.05)        | 1.610 (0.05) | 1.577 (0.05)        | 1.531 (0.04)        | <b>1.501 (0.04)</b> | 7.26%                     | 5.06%  | 2.00%  |
| Amazon Instant Video        | 1.330 (0.01)        | 1.328 (0.01) | 1.291 (0.01)        | <b>1.260 (0.01)</b> | 1.270 (0.01)        | 4.57%                     | 1.65%  | -0.79% |
| Electronics                 | 1.828 (0.00)        | 1.823 (0.00) | 1.764 (0.00)        | <b>1.722 (0.00)</b> | <b>1.722 (0.00)</b> | 5.87%                     | 2.44%  | 0.00%  |
| Music                       | <b>0.956 (0.00)</b> | 0.958 (0.00) | 0.959 (0.00)        | 0.980 (0.00)        | 0.959 (0.00)        | -0.31%                    | 0.00%  | 2.19%  |
| Movies and TV               | 1.119 (0.00)        | 1.117 (0.00) | <b>1.114 (0.00)</b> | 1.119 (0.00)        | 1.120 (0.00)        | -0.27%                    | -0.54% | -0.09% |
| Books                       | 1.107 (0.00)        | 1.109 (0.00) | <b>1.106 (0.00)</b> | 1.138 (0.00)        | 1.113 (0.00)        | -0.54%                    | -0.63% | 2.25%  |
| Average on all datasets     |                     |              |                     |                     |                     | 7.79%                     | 3.28%  | 1.22%  |

Σχήμα 4. Συγκριτικός πίνακας μεθόδων εντοπισμού πτυχών.

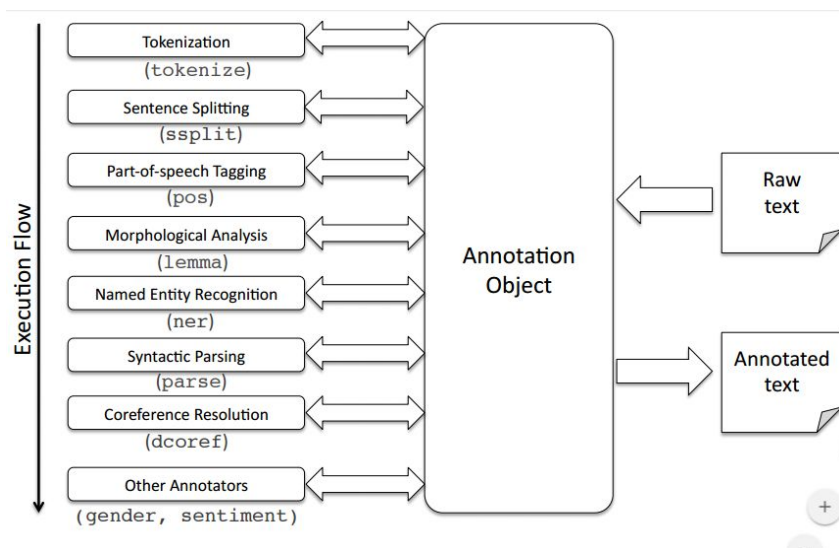
Σε μια άλλη δημοσίευση των Diao et al(2014) χρησιμοποιούνται ξεχωριστές βαθμολογίες για διαφορετικές πτυχές σχολίων και συνδυάζονται με τις βαθμολογίες των χρηστών και άλλα μετα-δεδομένα σε μια υβριδική προσέγγιση (συνδυασμός πληροφορίας που εξάγεται από περιεχόμενο με συνεργατικό φιλτράρισμα), με σκοπό τις συστάσεις ταινιών. Σε άλλες εργασίες Suresh, et al (2014) και Roohi, et al (2014) χρησιμοποιούνται βαθμολογίες του συναισθήματος της κάθε πτυχής ενός σχολίου για τη παρουσίαση ενός υποσυνόλου κριτικών με σκοπό να βρεθούν συστάσεις που είναι πιο πιθανό να βοηθήσουν το χρήστη.

## B. Λογισμικό που χρησιμοποιήθηκε για την συντακτική ανάλυση της φυσικής γλώσσας

Για την επίλυση μερικών εκ των εγγενών δυσκολιών που υπάρχουν στο τομέα της ανάλυσης της φυσικής γλώσσας, θεωρήθηκε σκόπιμο να ερευνηθούν τα εργαλεία που επιλύουν σε μεγάλο βαθμό προβλήματα όπως η αναγνώριση λεκτικών συνόλων σε μια πρόταση και η συντακτική ανάλυση ενός κειμένου.

Στα πλαίσια του υλοποιημένου μοντέλου της εργασίας χρησιμοποιήθηκε ένα python πρόγραμμα που έκανε χρήση του Stanford CoreNLP εργαλείου, για τις ανάγκες της λεκτικής επεξεργασίας. Το Stanford CoreNLP (σχήμα), είναι ένα Java πρόγραμμα που λειτουργεί ως ένα εργαλείο υποβοήθησης της διαδικασίας σχολιασμού κειμένων (annotation pipeline framework) και παρέχει το μεγαλύτερο μέρος τα ευρέως αποδεκτά εργαλεία του πυρήνα του κλάδου της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), δηλαδή σύμφωνα με το σχήμα 5 της σχετικής βιβλιογραφικής αναφορά “The Stanford CoreNLP Natural Language Processing Toolkit” των Christopher et al (2014) :

- τη σημειογραφία (tokenization)
- το χωρισμό των προτάσεων (sentence splitting)
- αναγνώριση του μέρους του λόγου της κάθε λέξης (POS tagging)
- τη μορφολογική ανάλυση του κειμένου (morphological analysis)
- την αναγνώριση της ονοματικής οντότητας (named entity resolution)
- συντακτική ανάλυση (Syntactic parsing)
- την ανάλυση έμμεσων αναφορών (coreference analysis)
- άλλα χαρακτηριστικά όπου αυτά είναι δυνατόν να αναγνωριστούν πχ. φύλο, συναίσθημα κλπ.



Σχήμα 5. Βήματα ανάλυσης κειμένου από το πρόγραμμα CoreNLP

Τα κίνητρα δημιουργίας του προγράμματος ήταν:

- Να είναι σε θέση γρήγορα και ανώδυνα να αναλύσει τη γλωσσική σημειογραφία για ένα κείμενο.
- Να είναι σε θέση να αποκρύψει τις διαφορές μεταξύ των διάφορων κατασκευαστικών στοιχείων πίσω από ένα κοινό API
- Να έχει μια ελάχιστη λειτουργικότητα εντοπισμού σημασιολογίας
- Να παρέχει ένα ελαφρύ πλαίσιο, χρησιμοποιώντας Java αντικείμενα (όχι κάτι “βαρύτερο”, όπως XML ή UIMA’s Common Analysis System (CAS) objects αντικείμενα). (Manning, et al 2014)

Για την περαιτέρω ανάλυση του εξαγόμενου αποτελέσματος με σκοπό τον εντοπισμό συναισθήματος στο κείμενο χρησιμοποιούνται με συνδυαστικό τρόπο το εξαγόμενο αποτέλεσμα της παραπάνω λεκτικής επεξεργασίας με θεματικά λεξικά της μεθόδου SO-CAL, ( Semantic Orientation CALculator).

Πρόκειται για λεξικά που περιέχουν έναν βαθμό για κάθε λέξη που αναδεικνύει τη πολικότητα της ανάλογα με το αν είναι θετικός ή αρνητικός και την ένταση της πολικότητας που προκύπτει από την απόλυτη τιμή του αριθμού αυτού. Σύμφωνα με τους Taboada et al (2011), αυτή μέθοδος χρησιμοποιεί δύο παραδοχές. Αρχικά ότι οι μεμονωμένες λέξεις έχουν ό,τι αναφέρεται ως προηγούμενη πολικότητα βάσει λεξικού, ο σημασιολογικός προσανατολισμός γίνεται ανεξάρτητα από το πλαίσιο που εμφανίζεται η λέξη. Και δεύτερον ότι το εν λόγω σημασιολογική προσανατολισμός μπορεί να εκφραστεί ως μία διακριτή αριθμητική τιμή. Σύμφωνα με τη συγκριτική ανάλυση των Collomb, A. et al (2014) τα μοντέλα που βασίζονται στα λεξικά επιτυγχάνουν καλύτερα αποτελέσματα, αλλά οι άλλες μέθοδοι δεν έχουν εξερευνηθεί σε μεγάλο βαθμό και δεν περιείχαν πάντα συγκριτικά αποτελέσματα με ανθρώπινα χαρακτηρισμένες κριτικές. Ακολούθηθηκε η μέθοδος ανάλυσης της δομής της πρότασης και η ανάλυσή της βάσει λεξικών. Αυτή η μέθοδος σύμφωνα με τον Khan et al. (2014) είναι δυνατόν να εφαρμοστεί ανεξαρτήτως τομέα και αποτελείται από τρία μέρη:

- Το χαρακτηρισμό με POS κάθε λέξης μιας πρότασης
- Την εύρεση όλων των λέξεων που υπάρχουν στα βαθμολογημένα λεξικά
- Τον συσχετισμό τους με όλες τις πιθανές πτυχές, δηλαδή τα ουσιαστικά ενός λεκτικού συνόλου.

**Οπτικοποίηση δεδομένων και οι γράφοι ως μέσα παροχής πληροφοριών (Data-visualization, Graph Inforgraphics)**

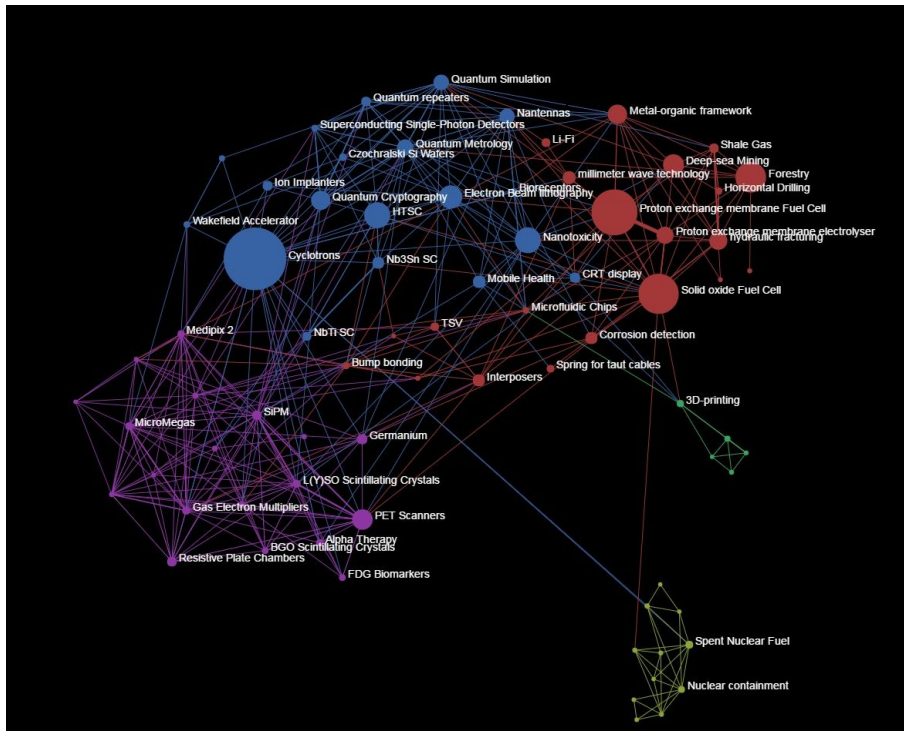
Οι οπτικοποιήσεις της πληροφορίας (πληροφοριογραφήματα-infographics) έχουν αρχίσει να χρησιμοποιούνται με την έλευση του Web 2.0, καθώς οι δικτυακές κοινότητες δημιούργησαν έφορο έδαφος για συλλογή και ανάλυση δεδομένων και τη διεξαγωγή μελετών στατιστικής φύσης, όπως αναφέρουν και οι Fazio et al (2015). Οι γράφοι και οι οπτικοποιήσεις γενικότερα έχουν το συγκριτικό πλεονέκτημα της συμπύκνωσης πληροφοριών σε εικόνα και δεδομένα που είναι πολλά για να είναι επεξεργάσιμα από τον άνθρωπο σε ρεαλιστικό χρόνο και μπορούν να αποδώσουν το νόημα σε πολύ λιγότερο χρόνο. Τα πληροφοριογραφήματα λαμβάνουν συνήθως, πληροφορίες γραπτής μορφής και τις παρουσιάζουν με ένα οπτικό τρόπο. Πρόκειται για ένα μέσο συνοπτικό, εύπεπτο, και αισθητικά ελκυστικό, ενσωματώνοντας έξυπνα οπτικά στοιχεία για να επισημανθούν βασικές πληροφορίες, Bellato, 2013. Στα πλαίσια της οπτικοποίησης της πληροφορίας χρησιμοποιούνται συχνά οπτικοποιήσεις που χρησιμοποιούν γράφους.

Στα μαθηματικά, ένας γράφος είναι μια αφηρημένη αναπαράσταση ενός συνόλου στοιχείων, όπου μερικά ζευγάρια (συνήθως) στοιχείων συνδέονται μεταξύ τους με δεσμούς. Τα διασυνδεδεμένα στοιχεία αναπαριστώνται με μαθηματικές έννοιες οι οποίες ονομάζονται κορυφές ενώ οι δεσμοί που συνδέουν τα ζευγάρια των κορυφών ονομάζονται ακμές. Συνήθως, ένα γράφημα απεικονίζεται σε διαγραμματική μορφή ως ένα σύνολο κουκκίδων για τις κορυφές, ενωμένα μεταξύ τους με γραμμές ή καμπύλες για τις ακμές. Τα γραφήματα είναι ένα από τα αντικείμενα μελέτης στα διακριτά μαθηματικά. Οι κορυφές καλούνται επίσης κόμβοι ή σημεία, και οι ακμές ονομάζονται επίσης γραμμές. Τα γραφήματα είναι το βασικό θέμα μελέτης από τη θεωρία γραφημάτων. Η λέξη «γράφημα» χρησιμοποιήθηκε για πρώτη φορά με αυτή την έννοια από τον James Joseph Sylvester το 1878, (Berge, C., 1958).

Με την αύξηση της υπολογιστικής ισχύος δημιουργήθηκε πληθώρα προγραμματιστικών εργαλείων και ψηφιακών βιβλιοθηκών (framework-libraries), με σκοπό την οπτικοποίηση της πληροφορίας. Όπως αναφέρεται και από τους δημιουργούς του Escher, ενός συστήματος που επιτρέπει την οπτικοποίηση στον client (King, 2015). Τα πλεονεκτήματα των web εφαρμογών περιλαμβάνουν την ταχεία χρήση του συστήματος από τον χρήστη (δεν χρειάζεται να κατεβάσει μια εφαρμογή ή browser plug-in), τη συμβατότητα με τις διαφορετικές τεχνολογίες υλικού (π.χ. κινητές συσκευές), την ευέλικτη ανταλλαγή, τη συνεργασία καθώς και την εύκολη ανάπτυξη εφαρμογών.

Βελτιώσεις απόδοσης και δημιουργία νέων βιβλιοθηκών καθιστούν client-based εργαλεία συγκρίσιμα με τα εργαλεία desktop σε πολλές περιπτώσεις (πχ Google Docs/maps, Slack Collaboration Tool). Εμπλουτίζοντας το HTML έγγραφο που λαμβάνεται από το διακομιστή με υπολογισμούς που λαμβάνουν χώρα στον δέκτη με τη χρήση της javascript, δίνονται μεγάλες δυνατότητες αφού αυτές οι εφαρμογές τρέχουν ανεξαρτήτως λειτουργικού συστήματος, κάνοντας πιο ελαφρύ ταυτόχρονα το φόρτο των υπολογισμών που λαμβάνουν χώρα στον server.

Έτσι υπάρχει μια αύξηση στις εφαρμογές που χρησιμοποιούν javascript λόγω των εγγενών αυτών διευκολύνσεων που προσφέρουν οι browsers. Ενώ διάφορες τεχνολογίες χρησιμοποιούνταν ευρέως για την υποστήριξη δημιουργίας γραφικών στο οικοσύστημα των browsers η εξειδίκευση έγινε τα τελευταία χρόνια με τεχνολογίες όπως βιβλιοθήκες ανοικτού κώδικα που είναι διαθέσιμες στην κοινότητα των προγραμματιστών είναι οι d3.js, sigma.js κ.α. (<https://d3js.org/>, <http://sigma.js.org/>). Οι οποίες χρησιμοποιήθηκαν και μελετήθηκαν εκτεταμένα για τη δημιουργία του πρωτοτύπου της εφαρμογής. (Εικόνες 6 και 7)



Εικόνα 6. Στιγμιότυπο χρήσης της sigma.js βιβλιοθήκης για τις ανάγκες ανάλυσης δημοσιεύσεων και πατέντων ομαδοποιημένες κατά τεχνολογίες Agocs, et al (2014)



Εικόνα 7. Δείγμα από την έκθεση του δημιουργού της βιβλιοθήκης d3.js, Bostock, et al (2011)

## MVC web application models

Σε αυτή τη παράγραφο περιγράφεται η αρχιτεκτονική των MVC μοντέλων δικτυακών εφαρμογών της οποίας η χρήση στα πλαίσια της πτυχιακής αναλύεται στο κεφάλαιο στη παράγραφο Ε του Κεφαλαίου IV. Εφαρμογή Λογισμικού. Σύμφωνα με τους Sridaran και Padmavathi (2010), το πρότυπο σχέδιο διαδικτυακών εφαρμογών Model / View / Controller (MVC) χωρίζει ένα περιβάλλον εφαρμογής σε τρία συστατικά για να χειριστεί τις αλληλεπιδράσεις με τον χρήστη, τους υπολογισμούς και τα εξαγόμενα δεδομένα αντίστοιχα. Αυτός ο διαχωρισμός ευνοεί σημαντικά την επαναχρησιμοποίηση (reusability). Οι Web εφαρμογές έχουν αναγκάσει τους σχεδιαστές να αποφασίσουν με ποιο τρόπο οι συνιστώσες του σχεδίου πρέπει να κατανέμονται μεταξύ του διακομιστή και του πελάτη πριν αρχίσει η φάση του σχεδιασμού. Για κάθε ταχέως αναπτυσσόμενη διαδικτυακή εφαρμογή, είναι πολύ δύσκολο να ενσωματωθούν μελλοντικές αλλαγές στις πολιτικές που σχετίζονται με τον διαχωρισμό (partitioning) σε κάθε επίπεδο της εφαρμογής. Μία λύση που προτείνει η σχεδιαστική αυτή τεχνική για το συγκεκριμένο πρόβλημα είναι να επαναλαμβάνει το μοντέλο (Model) και τον ελεγκτή (Controller) συστατικά τόσο σε εξυπηρετητή (server) όσο και του πελάτη (client).

Αν και υπάρχουν διάφορες βελτιώσεις της MVC αρχιτεκτονικής και της δημιουργίας νέων συστημάτων, υπάρχει ένα μεγάλο οικοσύστημα εργαλείων υποστήριξης του συγκεκριμένου μοντέλου που επιτρέπουν την γρήγορη ανάπτυξη πρωτοτύπων καλύπτοντας ταυτόχρονα απαιτήσεις που θα επιτρέψουν στο σύστημα να επεκταθεί στο μέλλον. Για αυτό έπειτα ενδελεχή σύγκριση ενός υποσυνόλου από αυτά τα εργαλεία, Εικόνα 8, (Plekhanova, 2009) έγινε η επιλογή του Django Web framework.

| Evaluation Criteria           | Weight | Django |                | Ruby on Rails |                | CakePHP |                |
|-------------------------------|--------|--------|----------------|---------------|----------------|---------|----------------|
|                               |        | Rating | Weighted Score | Rating        | Weighted Score | Rating  | Weighted Score |
| User interface development    | 0.2    | 4      | 0.8            | 3             | 0.6            | 3       | 0.6            |
| Maintainability               | 0.15   | 4      | 0.6            | 3             | 0.45           | 3       | 0.45           |
| Data management and migration | 0.2    | 4      | 0.8            | 5             | 1              | 2       | 0.4            |
| Testability                   | 0.15   | 3      | 0.45           | 4             | 0.6            | 2       | 0.3            |
| Popularity                    | 0.1    | 5      | 0.5            | 3             | 0.3            | 5       | 0.5            |
| Community and Maturity        | 0.1    | 5      | 0.5            | 5             | 0.5            | 3       | 0.3            |
| Marketability                 | 0.1    | 4      | 0.4            | 4             | 0.4            | 4       | 0.4            |
| Total:                        | 1      |        | 4.05           |               | 3.85           |         | 2.95           |

Εικόνα 8. Συγκριτικό σχήμα των MVC frameworks

## Γ. Ρόλος της στοχευμένης διαφήμισης στα κοινωνικά δίκτυα

Οι επικερδείς αγορές λειτουργούν κατά κύριο λόγο βάσει της αρχής ότι οι επενδυτές τους έχουν πρόσβαση στις "καλύτερες" πληροφορίες κατά τη λήψη μιας απόφασης αγοράς. Οι στοχευμένες- αυτοματοποιημένες διαφημίσεις πετύχουν ως στόχο να παρέχουν μια προσωποποιημένη επισκόπηση του κάθε καταναλωτή εξάγοντας πληροφορίες για αυτό από όλα τα ψηφιακά κανάλια μέσω μαζικής ενημέρωσης. Το κοινωνικό οικοσύστημα των μέσων ενημέρωσης προσφέρει τεράστια αξία για τους διαφημιστές ως ένα κομμάτι της Προγραμματικής διαφήμισης και παρέχει μία ακόμα πιο πετυχημένη και προσωπική διαφημιστική λύση, όπου η επιρροή ενός φίλου ή άλλη αξιόπιστη πηγή, καθοδηγεί το χρήστη κατά μήκος μιας διαδρομής προς την κατανάλωση νέων μορφών αγαθών (Guy, 2015).

Με την εισαγωγή πιο στοχευμένων μορφών διαφημίσεων π.χ. διαφημίσεις καταναλωτικές ομάδες με ιδιαιτερότητες, διαφημίσεις πολλαπλών προϊόντων ίδιας κατηγορίας, δυναμικές διαφημίσεις (custom audiences, multi-product ads, dynamic product ads), οι πλατφόρμες των κοινωνικών μέσων μαζικής ενημέρωσης αγκαλιάζουν την προσέγγιση αυτή σε πραγματικό χρόνο και επιτρέπουν την πραγματοποίηση τέτοιων προσφορών. Με την αυξανόμενη ανάγκη για μεγαλύτερη παροχή περιεχομένου, έχουμε αρχίσει να βλέπουμε μια αλλαγή στην χρήση της αυτοματοποιημένης-στοχευμένης διαφήμιση σε όλα τα είδη των μέσων μαζικής ενημέρωσης (Facebook, Google, Amazon), Dawson, και Lamb, 2016.

#### **Δ. Προσεγγίσεις στοχευμένης διαφήμισης χρησιμοποιώντας συστήματα συστάσεων**

Μέχρι στιγμής, δεδομένων των πληροφοριών που χρησιμοποιούν κατά τη πραγματοποίηση στοχευμένων διαφημίσεων εταιρείες όπως η Google ή το Facebook, μπορούμε να συμπεράνουμε ότι υπάρχουν πολλές διαφορετικές προσεγγίσεις στην διεξαγωγή μιας τέτοιας διαφημιστικής εκστρατείας, αλλά συνήθως λαμβάνονται υπόψη ξεχωριστά ή και συνδυαστικά μερικές από τις ακόλουθες πληροφορίες: τοποθεσία, δημογραφικά στοιχεία, ενδιαφέροντα, (on-line) δραστηριότητα, cookies (πληροφορία που βρίσκεται εάν ένας χρήστης έχει ήδη επισκεφθεί μια συγκεκριμένη ιστοσελίδα) ή ακόμα και βάση μετα-δεδομένων (συσκευή που χρησιμοποιείται, ιστοσελίδα που επισκέπτεστε κλπ. ).

Είναι σημαντικό να γίνει αναφορά στα συστήματα συστάσεων που διαδραματίζουν ολοένα και πιο σημαντικό ρόλο στην επιτυχία των ιστοσελίδων κοινωνικής δικτύωσης. Η πλειοψηφία της κυκλοφορίας σε ιστοσελίδες κοινωνικών δικτύων προκλήθηκαν από συστάσεις του ίδιου του δικτύου. Άρα είναι ασφαλές να συμπεράνουμε ότι η ποιότητα των συστάσεων αυτών είναι ζωτικής σημασίας για την προσέλκυση νέων χρηστών αλλά και τη διατήρηση των ήδη υπαρχόντων. Δύο από τις πιο εξέχουσες μορφές συστάσεων των συστημάτων κοινωνικής δικτύωσης είναι η εισήγηση του περιεχομένου κοινωνικών μέσων μαζικής ενημέρωσης και η σύσταση των ανθρώπων - χρηστών που θα πρέπει να επηρεάζουν το περιεχόμενο ενός χρήστη, Guy, I., 2015.

Αντί για τις εκτεταμένες αξιολογήσεις κειμένου, επικεντρωθήκαμε στα tips που παρέχονται από χρήστες του κοινωνικού δικτύου Yelp. Οι συγγραφείς Sklar, M. and Concerpcion, K.J., 2014, στο ερευνητικό τους έργο, επικεντρώθηκαν επίσης σε tips αντί για πλήρεις αξιολογήσεις, ωστόσο με διαφορετικό στόχο: να προτείνουν τις σωστές κριτικές-tips στους σωστούς ανθρώπους μέσω της πλατφόρμας Foursquare, λαμβάνοντας υπόψη το σύνολο των κριτικών που κάνανε οι ίδιοι οι χρήστες καθώς και τις προτιμήσεις των χρηστών και των κοινωνικών σχέσεων. Σε μια παρόμοια περίπτωση, η ανίχνευση του συναισθήματος πολικότητας για τις διάφορες πτυχές που υπάρχουν σε ένα κείμενο, έχει βρεθεί ότι βελτιώνει το περιεχόμενο για στοχευμένες διαφημίσεις, δεδομένου ότι οι χρήστες ανταποκρίθηκαν περισσότερο σε διαφημίσεις που ταιριάζουν στις θετικές πτυχές (αυτές που είναι στο προφίλ τους), αντί εκείνων που ταιριάζουν τυχαία σε πτυχές από όλο το κείμενο, Fan, & Chang, 2010.

Σε αυτή την εργασία, γίνεται μια προσπάθεια να εφαρμοστούν τεχνικές εξόρυξης συναισθήματος (opinion mining) από κείμενο οι οποίες θα μπορούσαν να εφαρμοστούν στο πεδίο των συστημάτων συστάσεων προσθέτοντας το στοιχείο του "word of mouth" που δεν έχει εφαρμοστεί προς το παρόν σε μεγάλο βαθμό και το αποτέλεσμα αυτής της επεξεργασμένης πληροφορίας να απεικονισθεί συμπιεσμένα σε έναν γράφο (infographic) που παρέχει τη δυνατότητα στο χρήστη αυτού του συστήματος (είτε χρήστης ως επιχείρηση είτε ως καταναλωτής) να αποκτήσει μια άποψη για μια κατηγορία προϊόντων και να αναζητήσει τις αναφερόμενες πτυχές όπως αυτές έχουν αναφερθεί από άλλους καταναλωτές.

## **Χρήση των εργαλείων των προαναφερόμενων κλάδων της Πληροφορικής στα πλαίσια της εργασίας**

Η εργασία που παρουσιάζεται εδώ στηρίζεται στην ίδια βάση με τις προαναφερθείσες δημοσιεύσεις. Δηλαδή εστιάζεται στον σχηματισμό προφίλ χρηστών και επιχειρήσεων. Το καθένα εξ αυτών συνδέει με πτυχές. Οι πτυχές αυτές εκπορεύονται από σχόλια που όπου βαθμολόγησαν οι χρήστες ή βαθμολογήθηκαν οι επιχειρήσεις αντίστοιχα για ένα προϊόν ή μια υπηρεσία τους.

Κάθε πτυχή συνδέεται με πολικότητα (θετική/αρνητική), καθώς και με μία βαθμολογία όπως αυτή ανιχνεύεται με τη χρήση λεξικών. Η σύγκριση με τα βαθμολογημένα λεξικά πραγματοποιείται βάσει της εξαγόμενης ανάλυσης φυσικής γλώσσας με τη χρήση του εργαλείου Stanford CoreNLP. Το τελικό αποτέλεσμα παρουσιάζεται χρησιμοποιώντας τεχνικές οπτικοποίησης σε που παρέχεται από τη βιβλιοθήκη d3.js και συνδέεται με τα επεξεργασμένα δεδομένου μέσω του MVC Django Web Framework.

Το προτεινόμενο εργαλείο μπορεί να χρησιμοποιηθεί τόσο ως ένα σύστημα συστάσεων για τον τελικό χρήστη / καταναλωτή, προτείνοντας προϊόντα και ταυτόχρονα αναδεικνύοντας τις πτυχές που για τις οποίες εκτιμάται ένα προϊόν περισσότερο από άλλους χρήστες και ως εργαλείο ανάλυσης της αγοράς για τις επιχειρήσεις, η οποία μπορεί να πάρει μια επισκόπηση των πλεονεκτημάτων και των μειονεκτημάτων των προϊόντων τους, αλλά και να συγκριθούν με τους ανταγωνιστές τους και να αναλύσουν περαιτέρω στις απόψεις των διάφορων ομάδων καταναλωτών, επιτρέποντας έτσι να διεκπεραιώσουν μια αρχική ανάλυση για την αγορά που στοχεύουν.

### III. ΠΡΟΤΕΙΝΟΜΕΝΟ ΜΟΝΤΕΛΟ

Η κύρια ιδέα είναι να βγει συμπυκνωμένα η πληροφορία σε γραφική απεικόνιση, από τις παρατηρήσεις και τα σχόλια που παρέχονται από τους χρήστες σε ιστοσελίδες με κριτικές προϊόντων/υπηρεσιών προκειμένου να εντοπιστούν σημαντικοί παράγοντες ενός προϊόντος που θα ενδιαφέρουν τον καταναλωτή κατά την αναζήτηση του και τις επιχειρήσεις στη διαδικασία λήψης επιχειρηματικών αποφάσεων. Η ανάλυση αυτού του περιεχομένου θα αποκαλύψει πτυχές που ενδιαφέροντος σε κάθε χρήστη και τις πτυχές επιδρούν θετικά ή αρνητικά στην αγορά μιας υπηρεσίας. Το σύστημα έχει ως απώτερο στόχο να αξιοποιήσει τις πληροφορίες αυτές και να καταφέρει να αναδείξει προς τον χρήστη τη βέλτιστη επιλογή για τα κριτήρια που θέτει.

Υποθέτουμε στη συνέχεια ότι ο “στόχος” της διαφήμισης είναι ένας χρήστης, και υποθέτουμε ότι ο χρήστης αυτός εξέφρασε την άποψή του σχετικά με ένα προϊόν μιας επιχείρησης. Για παράδειγμα δήλωσε ότι ένα εστιατόριο είχε ένα “άγευστο λευκό κρασί». Το σύστημα στη συγκεκριμένη περίπτωση θα παρέχει στο χρήστη περισσότερες πτυχές της ίδιας / ή άλλων επιχειρήσεων, χρησιμοποιώντας όλα τα άλλα σχόλια των χρηστών που βαθμολόγησαν το εστιατόριο, ή και ενός υποσυνόλου των χρηστών των οποίων τα προφίλ ταιριάζουν με τον χρήστη που πραγματοποιεί την αναζήτηση, κάτι που μπορεί να επιτευχθεί με την εφαρμογή τεχνικών συσταδοποίησης (clustering).

Μερικά παραδείγματα των ενεργειών που μπορεί να πραγματοποιήσει ο χρήστης στο σύστημα σύμφωνα με το προαναφερθέν σενάριο είναι.

1. Εφόσον το λευκό κρασί αυτού του εστιατορίου είναι σίγουρα μια κακή επιλογή, αφού βαθμολογήθηκε αρνητικά από περισσότερους χρήστες, αλλά το εστιατόριο έχει πολύ καλή “μπύρα”, ή πολύ καλή “επιλογή τυριών” σύμφωνα με άλλους χρήστες, άρα θα προταθούν αυτές η πτυχές για το ίδιο εστιατόριο δίνοντας του έτσι και μια δεύτερη ευκαιρία.
2. Θα του προτείνει άλλο εστιατόριο στην περιοχή που έχει και κατέχει υψηλή βαθμολογία για το “λευκό κρασί” του.
3. Αν ο χρήστης έχει σχολιάσει ή έχει πάει σε αρκετά ιταλικά εστιατόρια ή αν οι χρήστες της ίδια τεχνητής ομάδας των οποίων τα σχόλια έχει βρεθεί πως τον χαρακτηρίζουν θα του προταθούν κάποια καλά ιταλικά εστιατόρια σε εκείνη την περιοχή.

Το πρώτο παράδειγμα αποσκοπεί στο να πειστεί ο χρήστης να δώσει μια δεύτερη ευκαιρία την επιχείρηση, δίνοντας έμφαση στα “δυνατά” χαρακτηριστικά/πτυχές μιας επιχείρησης (στο συγκεκριμένο παράδειγμα, θετικά βαθμολογημένα πιάτα). Στα δύο τελευταία παραδείγματα των επιχειρήσεων θα παρέχονται εναλλακτικές επιλογές για το χρήστη, οι οποίες αναμένεται να παρέχουν μια πιο ευχάριστη εμπειρία για το συγκεκριμένο χαρακτηριστικό που αναζητά ο χρήστης.

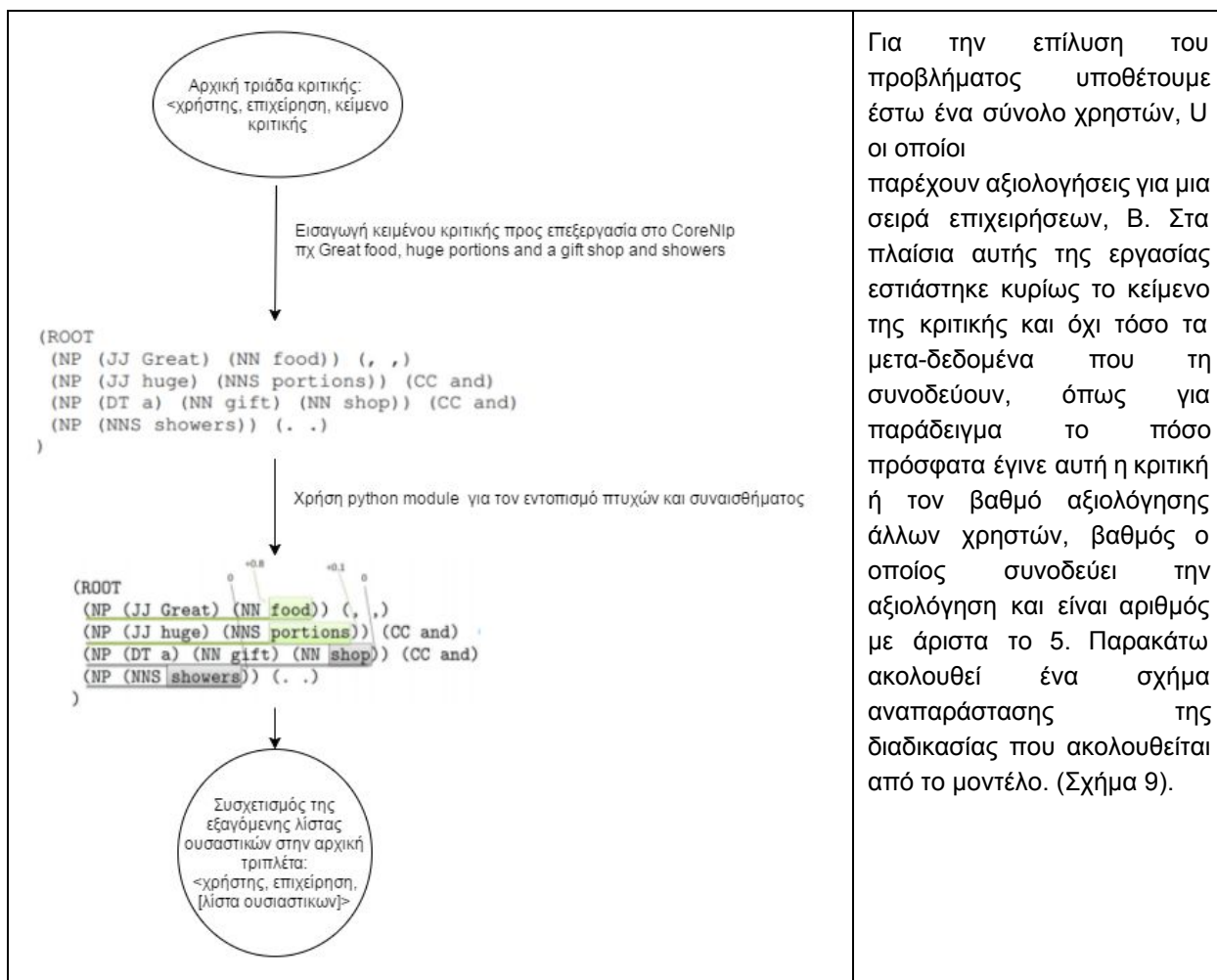
Ομοίως, αν υποθέσουμε ότι ο στόχος της υπηρεσίας είναι η επιχείρηση, τότε το σύστημα θα έχει ως σκοπό:

- Να παρέχει στους ιδιοκτήτες επιχειρήσεων τις πτυχές τις οποίες οι χρήστες σχολίασαν για την επιχείρησή τους. Αυτή η πληροφορία μπορεί να είναι πολύτιμη για τους ιδιοκτήτες των επιχειρήσεων, δεδομένου ότι είναι ο καλύτερος τρόπος να είναι ενήμεροι για τα πλεονεκτήματα και τις αδυναμίες τους. Αναλύοντας τα προφίλ των χρηστών επίσης είναι δυνατόν να γίνει ανάλυση της αγοράς στην οποία στοχεύουν και έτσι οι διαφημιστικές εκστρατείες τους μπορούν να στοχεύσουν στους κατάλληλους χρήστες με το παρουσιάζοντας τις πτυχές που του είναι ταιριαστές. Η

παραπάνω μέθοδος μπορεί να ενισχυθεί ακόμα περισσότερο με ανάλυση των δημογραφικών στοιχείων των χρηστών που θα επιτρέψει στους ιδιοκτήτες μιας επιχείρησης να τρέξουν εξατομικευμένες εκστρατείες ή να προσφέρουν κουπόνια / προσφορές σε συγκεκριμένες ομάδες χρηστών.

- Να παρέχει στους ιδιοκτήτες επιχειρήσεων μια σαφή εικόνα του ανταγωνισμού που επικρατεί για ένα προϊόν σε μια κατηγορία χρηστών. Χρησιμοποιώντας αυτή τη πληροφορία, οι ιδιοκτήτες των επιχειρήσεων θα είναι σε θέση να εντοπίζουν ευκαιρίες σε πτυχές που παρουσιάζουν μεγάλο ενδιαφέρον μεταξύ των ομάδων χρηστών στις οποίες στοχεύουν, αλλά βαθμολογούνται αρνητικά ή ακόμα και καθόλου στις κριτικές που λαμβάνουν από τους χρήστες των προϊόντων τους. Με αυτό το τρόπο μπορούν να διαφημίσουν τα δυνατά τους σημεία, να καταλάβουν τις ανάγκες των πελατών τους, και κατά συνέπεια να βελτιώσουν τις πτυχές των προϊόντων και των υπηρεσιών τους, ή να δουν πιο δυνητικά κομμάτι της αγοράς θα μπορούσαν να προσελκύσουν εφόσον. υπάρχει κάποια ανάγκη που δεν καλύπτεται και να δουν σε τι υστερούν του ανταγωνισμού.

#### A. Διαδικασία του συστήματος προτάσεων



Σχήμα 9. Σχήμα ροής διεργασιών ανάλυσης κειμένου.

Για την καλύτερη παρακολούθηση παραθέτουμε ένα πίνακα των συμβόλων που χρησιμοποιούνται στη συνέχεια

| Σύμβολο | Αγγλικός όρος                     | Ελληνικός όρος                            |
|---------|-----------------------------------|-------------------------------------------|
| r       | review                            | κριτική                                   |
| u       | user                              | χρήστης                                   |
| b       | business                          | επιχείρηση                                |
| a       | aspect                            | πτυχή                                     |
| d       | dominant aspect subset of aspects | σύνολο πτυχών                             |
| M       | set of all dominant aspects       | σύνολο κυρίαρχων πτυχών                   |
| K       | set of matched businesses to user | σύνολο ταιριαστών αντικειμένων σε χρήστες |

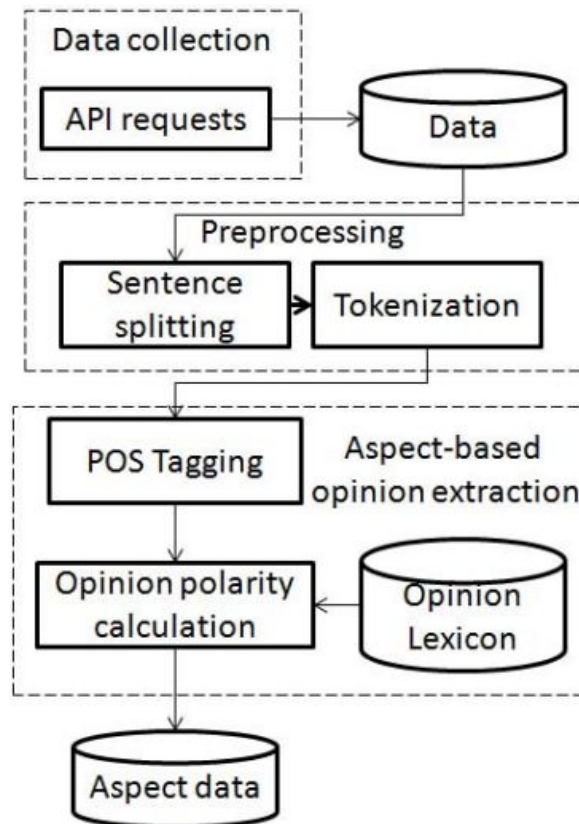
Κάθε αξιολόγηση  $r_k$ , από το σύνολο των κριτικών R ( $r_k \in R$ ) είναι ένα κείμενο που δημιουργείται έναν χρήστη και αναφέρεται σε μια επιχείρηση. Η περιγραφή αυτή της σχέσης μπορεί να οριστεί ως τριπλέτα:

$$r_k = \langle u_i, b_j, text_k : u_i \in U, b_j \in B$$

Ο όρος πτυχή στον κλάδο της εξόρυξης γνώμης συνήθως συνδέεται με χαρακτηριστικά του προς αξιολόγηση αντικειμένου και αντιστοιχεί σε ένα ουσιαστικό ή ρήμα ή ένα σύνολο ουσιαστικών και ρημάτων. Για παράδειγμα οι πτυχές που εξάγονται από το κείμενο της αξιολόγησης ενός εστιατορίου μπορεί να αποτελούν την “υπηρεσία σερβιρίσματος”, τη “ποιότητα φαγητού”, τη “γεύση”, κλπ. Συνήθως οι πτυχές αυτές συνοδεύονται από επίθετα ή επιρρήματα τα οποία βρίσκονται στην ίδια λεκτική φράση και είανια αυτά που δίνουν θετικό είτε αρνητικό προσανατολισμό σε μία ή περισσότερες πτυχές. πχ “αγενές προσωπικό”, “απαράδεκτη κουζίνα” κλπ.

Σύμφωνα με τη βιβλιογραφία, αυτό αντιστοιχίζεται με ένα ζεύγος <κύριος όρος, όρος τροποποιητής> όπου ο “κύριος όρος αναφέρεται σε μια “πτυχή ενώ ο “τροποποιητής” στη γνώμη που δίνεται, για μία πτυχή. π.χ. <υπηρεσία, ακατάστατη>, <φαγητό, κακό>. Με τη χρήση ενός λεξικού, λέξεις-πτυχές συγχέονται με γνώμες (sentiment) μέσω της αντιστοιχίας τους με λέξεις τροποποιητές οι οποίες δίνουν θετικές ή αρνητικές τιμές. Έτσι γίνεται δυνατό να βρεθεί η πολικότητα και η ένταση αυτής της πολικότητας για μια πτυχή μιας αξιολόγησης.

Η ανάλυση του κειμένου της αξιολόγησης οδηγεί στην ανίχνευση των πτυχών που συνδέονται με τις τιμές γνωμοδότησης. Η ροή της διαδικασίας για την εξαγωγή γνώμης βάσει των πτυχών που εμφανίζονται στο κείμενο, απεικονίζεται στο Σχήμα 10.



Σχήμα 10. Η εξαγωγής γνώμης βάσει των πτυχών μιας αξιολόγησης.

Το αποτέλεσμα αυτής της διαδικασίας είναι η αναπαράσταση των πτυχών μιας αξιολόγησης,  $r_k \in R$  στην μορφή τριπλέτας ως εξής:

$r_k = \langle u_i, b_j, \langle a_i, s_i, v_i \rangle \rangle$  όπου

- $a_i$  είναι ο όρος-πτυχή (π.χ. ένα ουσιαστικό ή ρήμα το οποίο εξάγεται από την αξιολόγηση και τελικά συνδέεται με μια θετική ή αρνητική γνωμοδότηση),
- $s_i$  είναι η σημασιολογική πολικότητα (δηλαδή  $\pm$ ) της γνωμοδότησης
- και  $v_i$  είναι η τιμή που σχετίζεται με αυτή τη πολικότητα.

Η αξία της πολικότητας αυτής αυξάνεται όταν ένας όρος που δηλώνει εξαιρετικά πολωμένη γνώμη εμφανίζεται στο ίδιο πλαίσιο με την πτυχή.

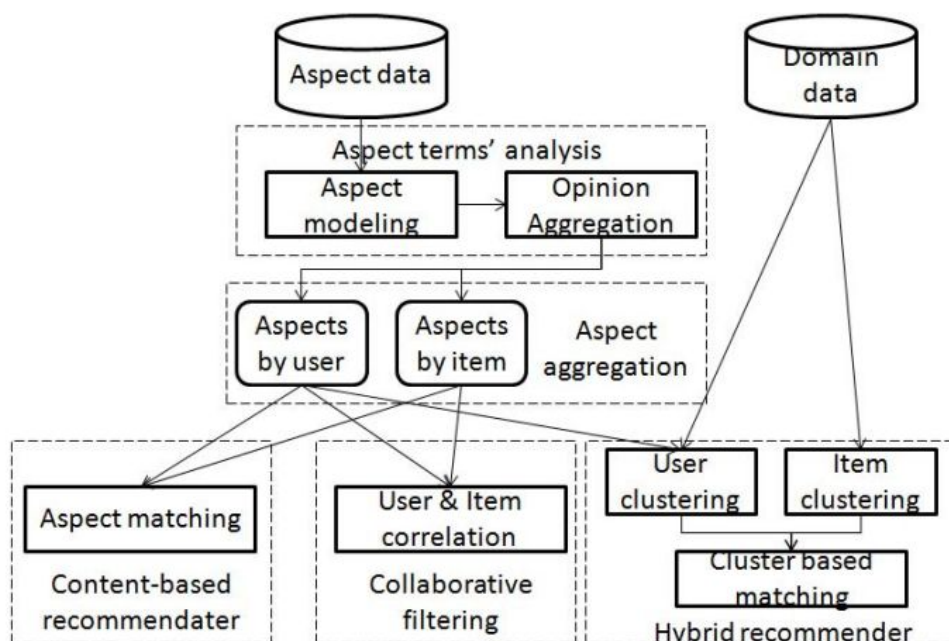
Η εκμετάλλευση δεδομένων που αναδεικνύουν γνώμη βάσει των πτυχών μιας αξιολόγησης, μπορεί να επιτευχθεί με την συνάθροιση και την ομαδοποίηση των πτυχών αυτών κατά χρήστη ή κατά αντικείμενου προς αξιολόγηση, Ricci, et al (2011). Εάν εξετάσουμε τις συστάσεις κάτω από το πρίσμα του συνεργατικού φιλτραρίσματος Ekstrand, (2011), τελικός μας στόχος είναι να συνδεθούν οι πτυχές και οι βαθμοί τους

(όπως αυτές παρέχονται από το χρήστη) με τα ενδιαφέροντα του κάθε χρήστη, και κατά συνέπεια να χρησιμοποιηθούν αυτά για τον υπολογισμό της ομοιότητας μεταξύ των χρηστών ή και αντικειμένων προς αξιολόγηση με τελικό στόχο να επιτευχθεί βελτίωση των συστάσεων.

Ομοίως, σύμφωνα με το φιλτράρισμα που βάσει περιεχομένου, οι πτυχές που παρέχεται από το χρήστη καθώς και οι πτυχές που συνδέονται με ένα αντικείμενο μπορούν να χρησιμοποιηθούν για την εύρεση των κατάλληλων λημμάτων. Αυτό μπορεί να πραγματοποιηθεί εύκολα χρησιμοποιώντας το κατάλληλο ερώτημα (query) σε μια βάση δεδομένων για το σύνολο των πτυχών που αντιστοιχούν σε κάθε μεμονωμένο χρήστη.

Τέλος, ακολουθώντας ένα υβριδικό μοντέλο στρατηγικής συστάσεων, είναι δυνατό να ομαδοποιηθούν χρήστες και αντικείμενα βάσει τις ομοιότητας των πτυχών ή και οποιασδήποτε συναφούς με το περιεχόμενο, πληροφορίας (π.χ. πληροφορίες γεωγραφικού εντοπισμού, δημογραφικά στοιχεία, ενδιαφέροντα όπως αυτά έχουν ρητά δηλωθεί από τον χρήστη κλπ.), και κατά συνέπεια, να πραγματοποιηθούν προτάσεις, χρησιμοποιώντας συνεργατικό (υπό την έννοια της χρήσης πολλαπλών χαρακτηριστικών) φιλτράρισμα μεταξύ των ομάδων.

Στο σχήμα 11 συνοψίζονται αυτές οι προσεγγίσεις, οι οποίες αναλύονται περαιτέρω στη συνέχεια.



Σχήμα 11. Η ανάλυση του αθροιστικού συνόλου των πτυχών και οι στρατηγικές σύστασης.

## B. Εξόρυξη πτυχής και γνώμης

Το πρώτο βήμα στην ανάλυση των πτυχών των αξιολογήσεων είναι να εντοπιστούν οι πιο ενδιαφέρουσες πτυχές και να ομαδοποιηθούν σε γενικότερες κατηγορίες, οι στα πλαίσια αυτή της εργασίας θα ονομάζονται "κυρίαρχες πτυχές" στη συνέχεια. Αυτό γίνεται συνήθως με τη χρήση τεχνικών θεματικής μοντελοποίησης

(e.g. Probabilistic Latent Semantic Analysis, Lu Y. 2011, Factorised Latent Dirichlet Allocation Moghaddam, S. 2013) ή ιεραρχικής ομαδοποίησης, Li X. et al (2012), Shepitsen A., (2008).

Το αποτέλεσμα αυτών των προσεγγίσεων είναι ένα υποσύνολο του αρχικού συνόλου πτυχών, όπου υπάρχει μια χαρτογράφηση των κυρίαρχων πτυχών με τις υπόλοιπες πτυχές και όρους που βρέθηκαν στις αξιολογήσεις. Ως εκ τούτου, από την ανάλυση του κειμένου μιας κριτικής  $r_k$  υπάρχει μια αντιστοίχιση με τις πτυχές της κριτικής αυτής στο σύνολο των κυρίαρχων πτυχών  $M$ .

Αυτή η σχέση μπορεί να συμπυκνωθεί στην ακόλουθη αναπαράσταση:

$$r_k = \langle u_i, b_j, \langle d_{am}, s_{im}, v_{im} \rangle \rangle \quad m=1..M$$

Όπου  $d_{am}$  είναι μία από το σύνολο  $M$  των κυρίαρχων πτυχών που βρέθηκαν στο προηγούμενο βήμα.

Σύμφωνα με την αναπαράσταση αυτή, μια κριτική  $r$  αντιστοιχίζεται με μια τριπλέτα  $\langle \text{user}, \text{business}, \text{text} \rangle$  το  $\text{text}$  είναι ένα σύνολο από κυρίαρχες πτυχές  $d$ ,  $s$  (πρόσημο +/-) και  $v$  που χαρακτηρίζει την τιμή του sentiment. Ο δείκτης  $a$  χρησιμοποιείται για να προσδιορίσει την πτυχή στο σύνολο των πτυχών  $D$  (επειδή μπορεί να έχει αναφερθεί από πολλούς χρήστες μια πτυχή).

## Γ. Συνάθροιση απόψεων για μια πτυχή

Το επόμενο βήμα είναι να συγκεντρωθούν οι πτυχές γνωμοδότησης για κάθε χρήστη, και για κάθε αντικείμενο προς αξιολόγηση, για να αποκτηθεί μια ολοκληρωμένη εικόνα των ενδιαφερόντων του χρήστη ή των ισχυρών/αδύναμων πτυχών του αντικειμένου προς αξιολόγηση.

Το αποτέλεσμα ενδιαφερόντων interests για ένα  $u_i$  χρήστη θα είναι ένα σύνολο από τριάδες:

$$\text{interests: } u_i = \langle d_{am}, s_{im}, v_{im} \rangle \quad m=1..M$$

Το αποτέλεσμα είναι το άθροισμα όλων των τιμών που παρέχονται από το χρήστη φράγμα  $u_i$  για την εμφάνισή της στα σχόλια και είναι  $s_{im}$  σημάδι του ποσού αυτού (+ ή -).

Ομοίως, ορίζουμε τις εντύπωση (impressions) για μία επιχείρηση  $b_j$ , ως:

$$\text{impressions: } b_j = \langle d_{am}, s_{jm}, v_{jm} \rangle \quad m=1..M$$

όπου  $v_{jm}$  είναι το άθροισμα όλων των τιμών που παρέχονται από όλους τους χρήστες για την εμφάνιση της πτυχής  $d_{am}$  στο αντικείμενο  $b_j$  και  $s_{jm}$  δηλαδή την ένδειξη του ποσού αυτού (+ ή -).

## Δ Μηχανή Συστάσεων

1) Ταίριασμα πτυχών: Με βάση τη συνολική γνώση των πτυχών στο σύνολο των κριτικών, είναι δυνατό γίνει μια απλή αναζήτηση για αντικείμενα που βάσει των αξιολογήσεων περιέχουν ακριβώς αυτές τις πτυχές που έχουν συσχετισθεί με το προφίλ του χρήστη. Για παράδειγμα, έστω ένας  $u_i$  χρήστης που είναι δυσαρεστημένος με την πτυχή  $d_{am}$  ενός αντικειμένου που αξιολόγησε όπως δηλώνεται από το αντίστοιχο συνδυασμό  $\langle \text{sign}, \text{value} \rangle$ .

(πχ <sign: -, vim: 1.5/2 --> μεγάλη τιμή).

Τότε μπορούμε να προτείνουμε το πιο ταιριαστό αντικείμενο  $b_k$  (όπου  $k$  μία από το σύνολο  $K$  ταιριαστών επιχειρήσεων) στο χρήστη, επειδή έχει αξιολογηθεί πολύ θετικά από άλλους χρήστες για το συγκεκριμένο χαρακτηριστικό (θα πρέπει δηλαδή να επιστραφεί ένα αντικείμενο με θετικό πρόσημο στο sign και υψηλή τιμή στο value δηλαδή βάσει του πεδίου  $v_{jm}$ ). Μιας τέτοια μορφής πρωταρχικής ανάλυσης μπορεί να πραγματοποιηθεί με ένα απλό ερώτημα (query) για τη συγκεκριμένη πτυχή στη βάση δεδομένων.

## 2) Συνεργατικό φιλτράρισμα βάσει της ομαδοποίησης πτυχών:

Η παραγοντοποίηση πινάκων είναι μια τεχνική που μπορεί να χρησιμοποιηθεί για την παροχή συστάσεων αντικειμένων (στη περίπτωση μας εστιατορίων) χρησιμοποιώντας το συνεργατικό φιλτράρισμα. Πολλοί αλγόριθμοι παραγοντοποίησης (όπως SVD, FunkSVD ή SVD++) έχουν γίνει πολύ δημοφιλείς μεταξύ των μηχανών συστάσεων για τα κοινωνικά δίκτυα, Yang, et al. (2011), Zhang, et al. (2014), Roshchina, et al (2015).

Όλες οι προτεινόμενες μεταβολές αποσυνθέτουν τον αρχικά διαμερισμένο πίνακα που περιέχει βαθμολογήσεις των στοιχείων (ή έμμεσα τις γνώμες των χρηστών) (Singular Value Decomposition - SVD) και στη συνέχεια αντιστοιχεί σε κάθε  $B_j$  στοιχείο με δύο πίνακες  $q_b$ ,  $y_b$  και κάθε  $U_i$  χρήστη με έναν πίνακα  $p_u$ . Οι συστάσεις για έναν χρήστη  $U_i$  βασίζονται στις προβλέψεις για όλες τις αντίστοιχες τιμές το χρήστη αυτόν.

Οι συγγραφείς Mirbakhsh, and Ling, (2013) χρησιμοποίησαν κατηγοριοποίηση (categorization όχι με την έννοια του classification), που επιτεύχθηκε μέσω τεχνικών συσταδοποίησης προκειμένου να βελτιωθούν οι επιδόσεις του αλγορίθμου SVD++. Η ομαδοποίηση των πτυχών μπορεί να επιτευχθεί με τη χρήση ιεραρχικών αλγορίθμων ομαδοποίησης, όπως εκείνες που εφαρμόστηκαν με επιτυχία στην παραγωγή ταξινόμησης Shepitsen, A., et al (2008).

Ο γενικός σκοπός του συνεργατικού φιλτραρίσματος στην περίπτωση της συγκεκριμένης μελέτης μπορεί να οριστεί ως εξής: "Λαμβάνοντας υπόψη το σύνολο χρηστών  $U$ , το σύνολο των μέτρων των επιχειρήσεων  $B$  και ένας πίνακας  $m \times n$  που περιέχει πληροφορία για το ποιοι χρήστες εξέφρασαν θετική γνώμη προς ποιες επιχειρήσεις (ή πιο απλοϊκά πώς οι χρήστες να αξιολόγησαν τις επιχειρήσεις), σκοπός είναι να προβλεφθεί η βαθμολογία ενός χρήστη  $U_i$  για μια επιχείρηση  $b_j$ ."

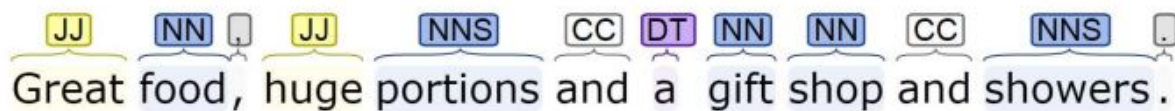
Όταν χρησιμοποιείται πληροφορία που έχει εξαχθεί από ανάλυση βάσει των πτυχών όπως αναφέρεται παραπάνω αντί ρητών αξιολογήσεων τότε θα έχουμε το σύνολο των χρηστών  $U$ , το σύνολο των μέτρων των επιχειρήσεων  $B$  και το σύνολο όλων των πτυχών  $A$  που αναφέρθηκαν στις αξιολογήσεις αυτές.

Κατά συνέπεια, οι διαστάσεις του πίνακα θα είναι  $m \times (n \cdot l)$  όπου  $l$  είναι ο αριθμός των πτυχών και το πρόβλημα θα είναι να προβλεφθεί η βαθμολογία ενός χρήστη  $u_i$  για την πτυχή  $A_k$  μιας επιχείρησης  $b_j$ . Προκειμένου να συστηθούν επιχειρήσεις στον χρήστη, συγκεντρώνονται οι προβλεπόμενες αξιολογήσεις για όλες τις πτυχές της κάθε επιχείρησης και κατατάσσονται οι επιχειρήσεις αναλόγως.

## Ε. Μοντέλο λεκτικής ανάλυσης

Το πρώτο βήμα του αγωγού επεξεργασίας μας ήταν να αναλύσει κάθε tip χρησιμοποιώντας τον core NLP Stanford parser και, κατά συνέπεια, να δημιουργήσει το αντίστοιχο κείμενο χωρισμένο σε φράσεις (lexical clauses) εμπλουτισμένο με τη πληροφορία για το τι μέρος του λόγου (POS) είναι η κάθε λέξη. Για παράδειγμα, το αποτέλεσμα που προκύπτει για το tip “Great food, huge portions and a gift shop and showers.”

Απεικονίζεται στα σχήματα 12. Σε μια ένθετη, δένδρική διάταξη, που μπορεί επίσης να απεικονιστεί στην εξής μορφή:



Σχήμα. 12. Ένα παράδειγμα μιας κριτικής (tip) με σημειολογία της ανάλυσης από τα μέρη του λόγου.

Το επόμενο βήμα για την επεξεργασία των tips εμπλουτισμένα πια με την επιπρόσθετη πληροφορία για το τι μέρος του λόγου είναι η κάθε λέξη, είναι να βρεθούν οι πτυχές καθώς και η πολικότητα της γνώμης που προσδίδεται σε αυτές. Σε αυτό το βήμα ελέγχθηκε αν τα ουσιαστικά, ρήματα και τα επίθετα του ενός μιας φράσης ήταν λέξεις που υπονοούσαν γνώμη θετική ή αρνητική βάσει της ύπαρξης τους σε ένα συνδιαστικό λεξικό. Το συνδιαστικό λεξικό βασίστηκε στα “AFFIN-111”<sup>6</sup> και “Harvard General Inquirer”<sup>7</sup> λεξικά, τα οποία βαθμολογούν λέξεις ως “γνωμοδότες” με δεκαδικό αριθμό που ανήκει στο σύνολο [-1.1]. Στα πλαίσια αυτής της έρευνας ελέγχθηκαν τα ουσιαστικά, τα επίθετα τα ρήματα και τα επιρρήματα για τυχόν προσδίδουσα πολικότητα. Εάν μία ή περισσότερες λέξεις όπως είχαν εντοπισθεί μέσα σε μία φράση είχαν πολικότητα, τότε το συνολικό άθροισμα των βαθμολογιών αυτών προστίθεται στο σκορ της φράσης αυτής.

Στο τελικό στάδιο, το σκορ πολικότητας από μία λέξη μέρος του λόγου (POS) μέσα στη ίδια φράση, συσχετίζεται με ένα ουσιαστικό ή ρήμα εντός του άκρου, καθώς αυτά τα μέρη του λόγου μπορούν να αποτελέσουν πτυχές. . Μια απεικόνιση των τιμών πολικότητας που αποδίδονται στις διάφορες πτυχές του προαναφερθέντος παραδείγματος υπάρχει στο σχήμα 13. Οι θετικά φορτισμένες πτυχές σημειώνονται με πράσινο χρώμα, ενώ τα αρνητικά και ουδέτερα αντίστοιχα με κόκκινο και γκρι. Η προκύπτουσα πληροφορία που αποτελείται από τις πτυχές και τις βαθμολογίες της πολικότητας των πτυχών, που υπονοούν γνώμη για κάθε tip, αποθηκεύεται μαζί με το αρχικό σύνολο δεδομένων των χρηστών, τις επιχειρήσεις (και τις κατηγορίες τους) και τα ίδια τα tips.



## IV. ΕΦΑΡΜΟΓΗ ΛΟΓΙΣΜΙΚΟΥ

### A. Συνοπτική περιγραφή της εφαρμογής

Η λειτουργία της εφαρμογής μπορεί να περιγραφεί ως εξής. Χρησιμοποιώντας ως αρχική πληροφορία κριτικές πελατών σε εστιατόρια και στη συνέχεια αναλύοντας τη κάθε κριτική προς την εύρεση πτυχών που περιέχουν sentiment, δημιουργήθηκε μια βάση δεδομένων που περιέχει την παραπάνω ανάλυση και τη παρουσιάζει με τη χρήση γράφων μέσω μιας web εφαρμογής. Η εφαρμογή παρέχεται στον ιστότοπο αποθετήριο ανοιχτού κώδικα github.

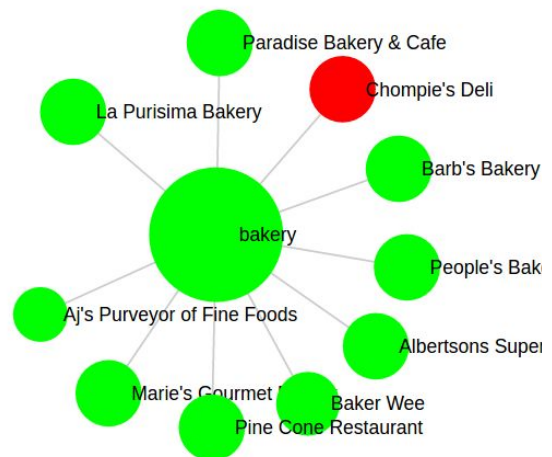
Η προεπεξεργασία των δεδομένων έγινε μέσω της εφαρμογής CoreNLP, ένα module ανοικτού κώδικα που υλοποιήθηκε από την ομάδα NLP (Natural Language Processing) του πανεπιστημίου Stanford. Η λειτουργία του περιγράφεται αναλυτικότερα παρακάτω αλλά μπορεί να συμπυκνωθεί αναφέροντας ότι αναλύει συντακτικά τις προτάσεις ως προς τα λεκτικά σύνολα και το μέρος του λόγου της κάθε λέξης για μια σειρά από γλώσσες. Στα πλαίσια αυτής της εργασίας επικεντρωθήκαμε στην αγγλική γλώσσα αλλά η λεκτική ανάλυση γίνεται για μια σειρά από γλώσσες (γαλλικά, αραβικά, γερμανικά). (Manning, et al 2014)

Η ενσωμάτωση δεδομένων από πρόσθετες πηγές είναι στις προτεραιότητες των μελλοντικών σχεδίων της εφαρμογής. Η πλήρης αξιοποίηση της αρχιτεκτονικής MVC (Models, Views, Controllers) στην οποία βασίζεται το Django (Python web framework) το οποίο αποτελεί και τη βάση του υλοποιημένου μοντέλου, θα επιτρέψει τον εμπλουτισμό της βάσης δεδομένων χάρη στο μηχανισμό Models.

Η χρήση της βάσης δεδομένων MongoDB έγινε επίσης σκόπιμα. Καθώς τα περισσότερα API παρόχων ή τα περισσότερα εργαλεία για crawling δεδομένων παρέχουν έγγραφα JSON μορφής τα οποία μπορούν να αποθηκευτούν ως έχουν στη βάση καθώς παρέχει τη δυνατότητα NoSQL δομής και τη δυνατότητα απευθείας αποθήκευσης εγγράφων (document-databases) και τη διατήρηση δυναμικού σχήματος στη βάση δεδομένων. Τα παραπάνω πλεονεκτήματα αναγνωρίστηκαν ως κρίσιμης σημασίας εφόσον κρίθηκε ότι το μελλοντικό σχήμα στα δεδομένα δεν είναι δυνατό να είναι αυστηρά στατικό και καθορισμένο αφού ο απώτερος σκοπός είναι να μη περιοριστούμε σε ένα τύπο δεδομένων αλλά να προστεθούν δεδομένα διαφορετικών πηγών.

Τέλος στο κομμάτι διεπαφής με το χρήστη χρησιμοποιούμε διακτυακό γράφο (network graph) μέσω της βιβλιοθήκης d3. Η χρήση των δυνατοτήτων που προσφέρουν τα visual analytics κρίνεται ιδιαίτερα χρήσιμη εφόσον μπορεί να συμπυκνωθεί και να μεταδοθεί πληροφορία με αρκετά εύκολο τρόπο.

Το χρώμα, το μέγεθος, το σχήμα των κόμβων καθώς αποστάσεις μεταξύ τους καθώς και οι αποστάσεις των ακμών μπορούν να χρησιμοποιηθούν ως διαφορετικοί τρόποι απεικόνισης πληροφορίας στην ίδια εικόνα. (σχήμα 14)



Σχήμα 14. Παράδειγμα παραγόμενου γράφου

Είναι σημαντικό να αναφερθεί πως παρότι το πρωτότυπο αυτό στηρίχθηκε σε κριτικές εστιατορίων είναι επεκτάσιμο και σε άλλες μορφές γραπτής κριτικής. Αν και σε τέτοια περίπτωση εφόσον τα δεδομένα ενσωματωθούν στην ίδια βάση δεδομένων είναι απαραίτητος ο διαχωρισμός τους ως προς την κατηγορία που αναφέρεται μια πτυχή. Καθώς η ίδια πτυχή σε διαφορετικές κατηγορίες ενδέχεται να έχει άλλο νόημα. Αλλά τα 4 βασικά σενάρια χρήσης που περιγράφηκαν είναι διαθέσιμα στην παρούσα εφαρμογή.

Ο στόχος της εφαρμογής είναι να παραχθεί η απεικόνιση ενός γραφήματος το οποίο θα επιτρέπει να προσαρμοστεί εύκολα στις διαφορετικές ανάγκες του τελικού χρήστη. Όπως αναφέρεται στο τμήμα III, (προτεινόμενο μοντέλο) στόχος της εφαρμογής είναι να μπορεί να προβεί σε συστάσεις τόσο προς τους χρήστες είτε αυτοί είναι πελάτες είτε ιδιοκτήτες επιχειρήσεων. Για αυτό το λόγο, υλοποιούμε διάφορα σενάρια που αποδεικνύουν τη χρηστικότητα του πολυ-εργαλείου.

Για τους ιδιοκτήτες επιχειρήσεων που επιθυμούν να βελτιώσουν τις υπηρεσίες τους και, συνεπώς, τη φήμη. Διαφημιστές που επιθυμούν να αυξήσουν την αφοσίωση των πελατών και της ανταπόκρισής τους στις εκστρατείες διαφήμισης τους και τέλος για τους χρήστες και εν δυνάμει καταναλωτές των προϊόντων να βρουν τα προϊόντα που χρησιμοποίησαν άλλοι χρήστες με ίδια ενδιαφέροντα.

Το γράφημα αυτό παράγεται ως αποτέλεσμα αναζήτησης σε aspects. Από το οποίο ο χρήστης μπορεί να αναζητήσει επιχειρήσεις στις οποίες έχουν γραφτεί σχόλια που περιέχουν το συγκεκριμένο aspect. Όλο το user interface παρέχεται μέσω του Web framework Django ενώ τα προεπεξεργασμένα δεδομένα βρίσκονται σε βάση δεδομένων MongoDB όπου είναι αποθηκευμένα όλα τα προεπεξεργασμένα δεδομένα.

Τα δεδομένα προεπεξεργάστηκαν με τη χρήση ενός wrapper υλοποιημένου σε python ο οποίος χρησιμοποιώντας το αρχείο jar που αποτελεί τον επεξεργαστή λεκτικής ανάλυσης Stanford Core NLP. Λειτουργεί ξεκινώντας έναν Server που περιέχει ένα API που δέχεται Unicode Strings και επιστρέφει την επεξεργασμένη ανάλυση που εξάγεται ως string. Το API καλείται με ένα python script και καλεί το λέξικο βάσει του οποίου ανιχνεύει ποια aspects περιγράφονται σε ένα επεξεργασμένο tip. Το κάθε ένα από τα αναφερόμενα components περιγράφονται αναλυτικά στη συνέχεια.

## B. Περιγραφή συνόλου δεδομένων (dataset)

Για τον έλεγχο της εφαρμογής των ιδεών που περιγράφηκαν παραπάνω υλοποιήθηκε ένα σύστημα χρησιμοποιώντας δεδομένα του κοινωνικού δικτύου Yelp το οποίο χρησιμοποιείται για την παροχή κριτικών και αξιολογήσεων διάφορων επιχειρήσεων. Οι αξιολογήσεις γίνονται από χρήστες της πλατφόρμας (crowd-sourcing feedback) χρησιμοποιώντας βαθμολόγηση (rating) καθώς και περιγραφή κειμένου (review). Το σύνολο δεδομένων που χρησιμοποιήθηκαν στην ανάλυση και οπτικοποίηση αποτελείται με περισσότερες από σαράντα χιλιάδες επιχειρήσεις και διακόσιες πενήντα χιλιάδες χρήστες που πραγματοποίησαν τετρακόσιες χιλιάδες κριτικές -tips, δηλαδή κείμενα μικρής έκτασης.

Για κάθε επιχείρηση περιγράφεται ένα σύνολο χαρακτηριστικών όπως: το όνομα της επιχείρησης, οι κατηγορίες στις οποίες ανήκει η συγκεκριμένη επιχείρηση, οι γεωγραφικές συντεταγμένες της επιχείρησης, η διεύθυνση, ωράρια λειτουργίας, ιδιαίτερα χαρακτηριστικά που, βαθμολογίες χρηστών κλπ.

Οι χρήστες του Yelp σχηματίζουν ένα κοινωνικό δίκτυο αφού εκτός από τα στοιχεία του προφίλ τους, έχουν και φίλους στα πλαίσια της πλατφόρμας Yelp. Οι γνώμες των χρηστών για τις επιχειρήσεις εκφράζονται μέσω των συμβουλών (tips) και των σχολίων (reviews). Οι "συμβουλές" είναι σύντομες, με μέγεθος ενός ή δύο προτάσεων κείμενα που παρέχονται από έναν χρήστη για μια επιχείρηση. Έχουν ημερομηνία συγγραφής και έναν αριθμό που αντιπροσωπεύει τον αριθμό των χρηστών στους οποίους άρεσε η συμβουλή. Από την άλλη μεριά οι κριτικές αποτελούν επίσης κομμάτια κειμένου, αλλά συνήθως είναι περισσότερο εκτενή σε μέγεθος. Έχουν επίσης ημερομηνία συγγραφής, αστέρια (αριθμητική βαθμολογία χρήστη), αριθμό ψήφων από άλλους χρήστες (ανάλογα με το πόσο διασκεδαστική, και χρήσιμη είναι η κριτική). Εκτός των συμβουλών και των σχολίων, το σύνολο δεδομένων περιέχει και check-in που συσχετίζουν τους χρήστες με τις επιχειρήσεις που έχουν επισκεφθεί.

Η ανάλυσή που πραγματοποιήθηκε επικεντρώθηκε μόνο στις πληροφορίες που προέρχονται από τα tips, αν και η προσέγγισή αυτή είναι εύκολο να επεκταθεί για να συμπεριλάβει τις πιο λεπτομερείς αξιολογήσεις. Το κίνητρο πίσω από αυτή την απόφαση είναι ότι τα tips:

- α) παρέχουν με ξεκάθαρο τρόπο συνήθως θετικές ή αρνητικές γνώμες και πτυχές,
- β) κρίνει μόνο μία ή δύο πτυχές της στοχευμένης επιχείρησης, εστιάζοντας στα πιο σημαντικά σημεία και
- γ) η επεξεργασία λόγω του περιορισμένου μεγέθους είναι ταχύτερη από ό, τι η επεξεργασία πλήρη σχολίων.

Κρατήσαμε επίσης πληροφορίες σχετικά με τις επιχειρήσεις και τις κατηγορίες τους και προσδιορίζονται οι χρήστες με user\_id τους στο Yelp. Στο πλαίσιο αυτής της εργασίας χρησιμοποιήθηκαν δεδομένα της ηλεκτρονικής πλατφόρμας Yelp. Από το yelp διοργανώθηκε ένας διαγωνισμός με σκοπό την επεξεργασία δεδομένων με σκοπό την απόκτηση γνώσης από τα reviews και έτσι διατέθηκαν δεδομένα της πλατφόρμας προς αξιοποίηση.

Αξιοποιώντας αυτά τα δεδομένα, στην παρούσα εργασία γίνεται προσπάθεια εντοπισμού κοινοτήτων (clustering) χρηστών βάσει παρόμοιων κριτικών που κάνουν αλλά και η δημιουργία ενός πίνακα ενδιαφερόντων με αυτόματο τρόπο, βάσει αναλυτών κειμένου (text analysis) και ανάλυση συναισθήματος (sentiment analysis) για τον χαρακτηρισμό ενός κειμένου αν είτε ως θετικό ή αρνητικό ως προς την γενική διάθεση. Αυτές οι κατηγορίες μπορούν να χρησιμοποιηθούν τόσο στον καλύτερο προσδιορισμό των κοινοτήτων όσο και σε ένα βασικό σύστημα προτάσεων στις ίδιες τις επιχειρήσεις για την ενσωμάτωση στοιχείων που αναζητούν οι πελάτες τους.

Συνολικά δίνονται 5 αρχεία που περιέχουν δεδομένα σε json μορφή (JavaScript Object Notation):

-yelp\_academic\_dataset\_business.json  
-yelp\_academic\_dataset\_checkin.json  
-yelp\_academic\_dataset\_review.json  
-yelp\_academic\_dataset\_tip.json  
-yelp\_academic\_dataset\_user.json

Στη συνέχεια δίνεται μια γενική περιγραφή του κάθε αρχείου και επεξήγηση των χαρακτηριστικών (attribute) σε κάθε αντικείμενο json.

#### 1.yelp\_academic\_dataset\_business.json

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Το αρχείο περιλαμβάνει πληροφορίες για 42153 επιχειρήσεις στις οποίες γίνανε αξιολογήσεις, πρόκειται για επιχειρήσεις που δεν ανήκουν σε κάποια συγκεκριμένη κατηγορία.</p> <p>-type είναι business για όλες τις επιχειρήσεις<br/>-business_id το μοναδικό αναγνωριστικό id της επιχείρησης<br/>-name: το όνομα της επιχείρησης,<br/>-neighborhood: όνομα της περιοχής<br/>-full_address: διεύθυνση,<br/>-city: πόλη,<br/>-state: πολιτεία,<br/>-latitude: γεωγραφικό μήκος,<br/>-longitude: γεωγραφικό πλάτος,<br/>-stars: μέση βαθμολογία<br/>-review_count: αριθμός κριτικών,<br/>-categories: κατηγορίες στις οποίες εμπίπτει η επιχείρηση<br/>-open: αν η επιχείρηση είναι ή όχι σε λειτουργία<br/>-hours: ώρες λειτουργίας<br/>-attributes: ιδιαίτερα χαρακτηριστικά που σχετίζονται με την επιχείρηση</p> | <p>Γενική μορφή</p> <pre>{<br/>  'type': 'business',<br/>  'business_id': (encrypted business id),<br/>  'name': (business name),<br/>  'neighborhoods': [(hood names)],<br/>  'full_address': (localized address),<br/>  'city': (city),<br/>  'state': (state),<br/>  'latitude': latitude,<br/>  'longitude': longitude,<br/>  'stars': (star rating, rounded to half-stars),<br/>  'review_count': review count,<br/>  'categories': [(localized category names)]<br/>  'open': True / False (corresponds to closed, not<br/>business hours),<br/>  'hours': {<br/>    (day_of_week): {<br/>      'open': (HH:MM),<br/>      'close': (HH:MM)<br/>    }, ...<br/>  },<br/>  'attributes': {<br/>    (attribute_name): (attribute_value),<br/>    ...<br/>  },<br/>}</pre> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 2.yelp\_academic\_dataset\_checkin.json

|                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>ο checkin είναι ουσιαστικά χρήστες που δίνουν το παρόν σε επιχειρήσεις. Στο αρχείο περιλαμβάνονται 31617 checkin για διαφορετικές επιχειρήσεις. Ακολουθεί αναλυτική περιγραφή των attributes:</p> <p>-'type': 'checkin', βρίσκεται σε όλα τα checkin αντικείμενα</p> <p>-'business_id', id της επιχείρησης την οποία αφορά το check-in</p> <p>-'checkin_info', αριθμός checkin ανά μέρα και ώρα</p> | <p>Γενική μορφή</p> <pre>check-in {   'type': 'checkin',   'business_id': (encrypted business id),   'checkin_info': {     '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),     '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),     ...     '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),     ...     '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)   }, # if there was no checkin for a hour-day block it will not be in the dict }</pre> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 3.yelp\_academic\_dataset\_review.json

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                       |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Το αρχείο περιλαμβάνει 1.125.458 κριτικές που έγιναν απο χρήστες για έπειτα από τη χρήση προϊόντος ή υπηρεσίας μιας επιχείρησης. Ακολουθεί αναλυτική περιγραφή των attributes:</p> <p>"votes": κατηγορία προς ψήφιση και αριθμός ψήφων</p> <p>"user_id": id χρήστη</p> <p>"review_id": id κριτικής</p> <p>"stars": βαθμολογία χρήστη,</p> <p>"date": ημερομηνία</p> <p>"text": το κείμενο της κριτικής</p> <p>"type": "review" χαρακτηρίζει ότι το αντικείμενο είναι κριτική</p> <p>"business_id": id επιχείρησης</p> | <p>Γενική μορφή</p> <pre>review {   'type': 'review',   'business_id': (encrypted business id),   'user_id': (encrypted user id),   'stars': (star rating, rounded to half-stars),   'text': (review text),   'date': (date, formatted like '2012-03-14'),   'votes': {(vote type): (count)}, }</pre> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 4.yelp\_academic\_dataset\_tip.json

|                                                                                                                                                                                                                                                                                                                                               |                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Το αρχείο περιλαμβάνει 403210 tips, μικρότερης μορφής κριτικές πιο απλής από τα reviews, τα attributes του αρχείου είναι:</p> <pre> "user_id": "Vefj29mjork1DLhALLNAsg", "text": "Great food, huge portions and a gift shop and showers.", "business_id": "JwUE5GmEO-sH1FuwJgKBIQ", "likes": 0, "date": "2012-05-16", "type": "tip" </pre> | <p>Γενική μορφή</p> <pre> tip {   'type': 'tip',   'text': (tip text),   'business_id': (encrypted business id),   'user_id': (encrypted user id),   'date': (date, formatted like '2012-03-14'),   'likes': (count), } </pre> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

#### 5.yelp\_academic\_dataset\_user.json

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Το αρχείο περιλαμβάνει 252898 διαφορετικούς χρήστες που πραγματοποίησαν reviews,tip ή checkin. Ακολουθεί η ανάλυση των attributes του αντικειμένου</p> <pre> -'type': 'user' χαρακτηρίζει ότι το αντικείμενο ανήκει στη κατηγορία user -'user_id': το id του χρήστη -'name': όνομα -'review_count': αριθμός reviews που έχει κάνει ο χρήστης -'average_stars': μέσος όρος βαθμολογιών που έχει πραγματοποιήσει ο χρήστης -'votes': χαρακτηρισμός των ψήφων που έχει κάνει από άλλους χρήστες -'friends': ids των φίλων-χρηστών -'elite': [(years_elite)], -'yelping_since': ημερομηνία έναρξης χρήσης της πλατφόρμας -'compliments': αριθμός και είδος κοπλιμέντου που έλαβε ο χρήστης -'fans': αριθμός θαυμαστών </pre> | <p>Γενική μορφή</p> <pre> user {   'type': 'user',   'user_id': (encrypted user id),   'name': (first name),   'review_count': (review count),   'average_stars': (floating point average, like 4.31),   'votes': {(vote type): (count)},   'friends': [(friend user_ids)],   'elite': [(years_elite)],   'yelping_since': (date, formatted like '2012-03'),   'compliments': {     (compliment_type):     (num_compliments_of_this_type),     ...   },   'fans': (num_fans), } </pre> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Γ. Front-End Application

Η δημιουργία του γραφήματος που απεικονίζει τη σχέση της πτυχής με τις διάφορες επιχειρήσεις γίνεται δυναμικά στον browser. Για την πραγματοποίηση του γίνεται χρήση της d3.js βιβλιοθήκης οπτικοποίησης. Αυτός ο γράφος χρησιμοποιείται ως ένα σύστημα συστάσεων για τη συμπύκνωση όλων των πιθανών επιλογών μέσω ενός infographic.

Το χρώμα χρησιμοποιείται για την ένδειξη πολικότητας του aspect για μια συγκεκριμένη επιχείρηση ενώ το μέγεθος για το πόσο έντονη είναι η πολικότητα αυτή. Η διαμόρφωση της τοπολογίας των κόμβων γίνεται χρησιμοποιώντας τον αλγόριθμο force Atlas ο οποίος υπολογίζει τις συντεταγμένες των κόμβων βάσει των μεγεθών των κόμβων και του βάρους των ακμών του γραφήματος.

Στο γράφο δίνεται επιπληρωμένο μεγαλύτερο μέγεθος στο κόμβο που απεικονίζει το όνομα της πτυχής προς αναζήτηση γιατί έτσι ο αλγόριθμος forceAtlas τον θέτει στη μέση του γράφου το οποίο είναι πιο κατατοπιστικό για το χρήστη του συστήματος.

Εν τέλει επιλέγοντας έναν κόμβο του συστήματος δίνεται η δυνατότητα αναγραφής των tips που αναφέρονται τη συγκεκριμένη πτυχή. Οι γράφοι ως σύστημα συστάσεων μπορούν να χρησιμοποιηθούν με τους ακόλουθους δύο τρόπους:

- Για τον χρήστη καταναλωτή ως μέσο εύρεσης της βέλτιστης, ή των πιθανών βέλτιστων μορφών μιας υπηρεσίας που μπορεί να παρέχει τη συγκεκριμένη πτυχή.
- Ενώ για μια επιχείρηση ως ένα σύστημα ανάλυσης των πτυχών με τις οποίες συνδέεται η ίδια. Καθώς και ως σύστημα ανάλυσης του ανταγωνισμού που υπάρχει στα πλαίσια αναζήτησης της συγκεκριμένης πτυχής.

Το γεγονός ότι παρέχεται ως φορητό web component η front-end εφαρμογή σημαίνει ότι υπάρχει η δυνατότητα χρήσης του συγκεκριμένου κώδικα ως Plugin, από εξωτερικές εφαρμογές για την αντίστοιχη παροχή υπηρεσιών συστάσεων στους χρήστες της, υπό την προϋπόθεση φυσικά ότι υπάρχει αντίστοιχο backend API (όχι απαραίτητα Django), και να παρέχει τα δεδομένα στη συγκεκριμένη μορφή JSON (data schema). Οδηγίες για τη χρήση της υπηρεσίας ως plugin υπάρχουν στο παράρτημα.

Το data schema που μπορεί να εισαχθεί στο γράφο περιγράφεται συνοπτικά παρακάτω:

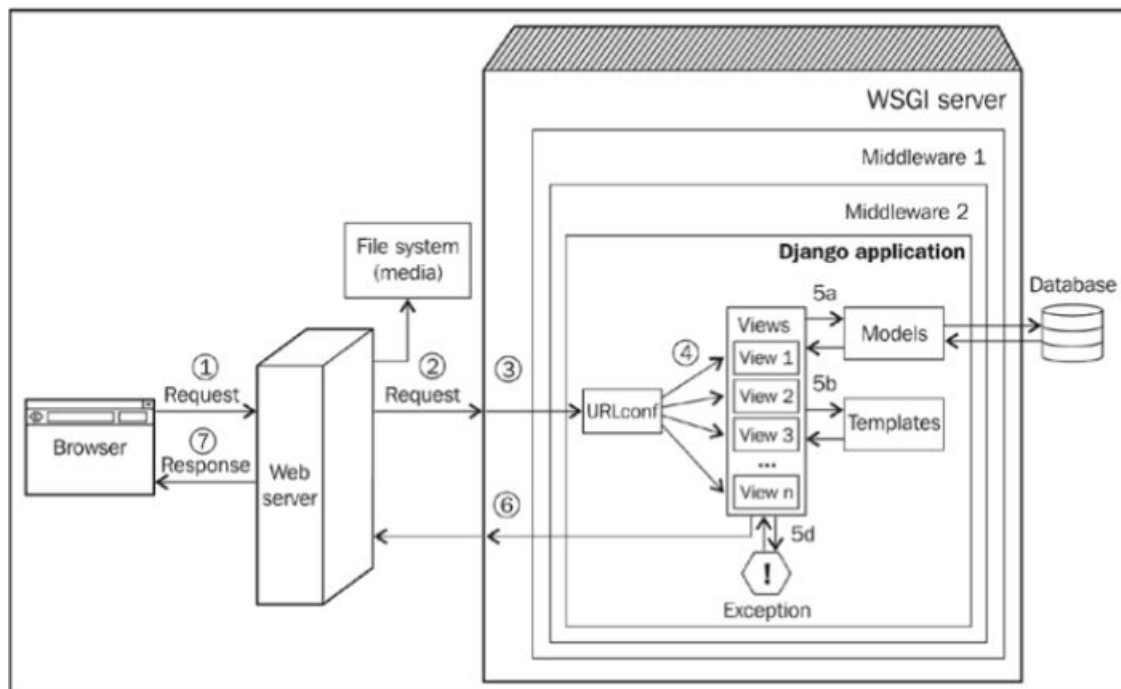
```
[
  nodes:
    [{
      "name": "displayed_name",
      "value": Integer //(depending on value appropriate color and size are displayed
    }],

  links:
    [{
      "source": Integer,//The Nth node of nodes array
      "target": Integer,//The Nth node of nodes array
    }]
]
```

## Δ. Back-End Application

Το Django πρόκειται για ένα web framework ανοιχτού κώδικα, γραμμένο σε Python, το οποίο ακολουθεί την (MVC- model view controller) αρχιτεκτονική σχεδιασμού web-εφαρμογών. Το λογισμικό υποστηρίζεται από το μη-κερδοσκοπικό ίδρυμα Django Software (DSF).

Ο κυρίαρχος στόχος του Django είναι η εύκολη δημιουργία πολύπλοκων web-εφαρμογών που βασίζονται στη χρήση βάσεων δεδομένων για το περιεχόμενό τους. Εστιάζει πολύ στην επαναχρησιμοποίηση components (reusability), στην εύκολη ενσωμάτωσή τους ("pluggability"), στη γρήγορη ανάπτυξη λογισμικού (rapid development), και στην αρχή DRY (don't repeat yourself). Χρησιμοποιείται αποκλειστικά η Python για τις ρυθμίσεις της εφαρμογής και την δημιουργία των μοντέλων των βάσεων δεδομένων. Παρέχει επίσης προαιρετικά μία διαδραστική διεπαφή χρήστη (GUI) τύπου CRUD (create, read, update and delete) που χρησιμοποιείται μέσω ελέγχου των templates από του web server.



Σχήμα 15. **Django Design Patterns** Ravindran 2015

Η κύρια εφαρμογή του Django αποτελείται από έναν σχεσιακό μηχανισμό αντιστοίχισης αντικειμένων (object-relational mapper ORM) που βρίσκεται ανάμεσα στα μοντέλα δεδομένων (data models που ορίζονται ως κλάσεις (με την ορολογία της αντικειμενοστρέφειας) και το μηχανισμό της σχεσιακής βάσης δεδομένων "Models. Εκτός αυτού παρέχει ένα σύστημα για την επεξεργασία και αποστολή αιτημάτων HTTP (HTTP requests ) τα οποία δύναται να εμπλουτιστούν μέσω των Views (web templating system ). Τέλος παρέχει έναν μηχανισμό αντιστοίχισης των URI με τα Views με το μηχανισμό των "Controllers".

Μερικά δευτερεύοντα στοιχεία που συνοδεύουν το κύριο μέρος του Django Web Framework:

- Έναν διακομιστή εστιασμένο κυρίως στην ανάπτυξη και τον έλεγχο του λογισμικού.
- Ένα σύστημα επικυρώσεων των διάφορων φορμών που αντιστοιχούν σε πεδία της βάσης δεδομένων

- Ένα σύστημα δημιουργίας των τελικών HTML εγγράφων
- Σύστημα caching για την ελάφρυνση του φόρτου του Διακομιστή
- Υποστήριξη ενδιάμεσων κλάσεων για συστήματα που θα λειτουργούν στο server
- Έναν εσωτερικό μηχανισμό επικοινωνίας μεταξύ των διαφορετικών modules της εφαρμογής
- παροχή μετάφρασης των django εφαρμογών σε διάφορες γλώσσες.
- Ένα σύστημα αποθήκευσης της κατάστασης των διάφορων στιγμιότυπων του Django σε (serialization) XML και JSON αναπαραστάσεις
- Δυνατότητα επέκτασης του μηχανισμού δημιουργίας templates
- Διεπαφή με το σύστημα testing της python

Η web εφαρμογή που υλοποιήθηκε, παρέχει τρεις κύριους μηχανισμούς οι οποίοι παίρνουν ως δεδομένο ότι τα προεπεξεργασμένα δεδομένα (με το τρόπο προεπεξεργασίας που περιγράφηκε παραπάνω) είναι αποθηκευμένα στη βάση δεδομένων MongoDB.

Γίνεται χρήση της αρχιτεκτονικής MVC που περιγράφηκε παραπάνω το σύστημα.

Συνδέεται με τη βάση δεδομένων (που τρέχει στο ίδιο μηχάνημα)

Παρέχει 2 View. Το πρώτο View που στέλνει την HTML σελίδα με τον Javascript κώδικα οπτικοποίησης του γράφου του aspect προς αναζήτηση. Και το δεύτερο που παρέχει το μηχανισμό αναζήτησης στη βάση και τη δημιουργία του αρχείου JSON το οποίο επιστρέφει σαν απάντηση αν η αναζήτηση στη βάση ήταν επιτυχής. Το template από όπου εξάγεται μετά από την επεξεργασία του Django template engine είναι αρκετά απλό. Αποτελείται από μια φόρμα αναζήτησης πτυχών και το χώρο της οπτικοποίησης ενός γράφου όλων των επιχειρήσεων που έχουν σχολιαστεί με τη πτυχή προς αναζήτηση από τη javascript εφαρμογή πελάτη στον server .

## Python CoreNLP wrapper

Για την αυτοματοποίηση της λεκτικής ανάλυσης χρησιμοποιήθηκε κώδικας ανοικτού λογισμικού από το παρακάτω αποθετήριο <https://github.com/dasmith/stanford-corenlp-python>.

Σύμφωνα με τη περιγραφή του αποτελεί έναν Python wrapper για το εργαλείο CoreNLP που υλοποιήθηκε από την ομάδα NLP του Πανεπιστημίου του Στάνφορντ και βασίζεται σε Java . Μπορεί είτε να εισαχθεί ως μονάδα ή να τρέξει ως διακομιστής JSON-RPC. Λόγω το ότι χρησιμοποιεί πολλά μεγάλα εκπαιδευμένα μοντέλα (απαιτεί τουλάχιστον 3GB RAM για 64-bit επεξεργαστές και συνήθως λίγα λεπτά χρόνο φόρτωσης), έχει πιο πολύ νόημα να τρέξει ως διακομιστής.

1. Η διασύνδεση Python με το Στάνφορντ CoreNLP παρέχει τα εξής εργαλεία:
  - a. χαρακτηρισμός φράσεων ως μέρους του λόγου (tagging POS),
  - b. ανάλυσης της φράσης της δομής (phase-structure parsing),
  - c. ανάλυση εξαρτήσεων, (dependency parsing)
  - d. αναγνώριση ονομάτων οντοτήτων, (named-entity recognition)
  - e. και έμμεση αναφορά σε προηγούμενους όρους (coreference).
2. Τρέχει ένα διακομιστή JSON-RPC που “τυλίγει” το διακομιστή Java και εξάγει JSON.

3. Έξοδος του συστήματος αποτελεί η ανάλυση του συντακτικού δέντρου το οποίο μπορεί να χρησιμοποιηθεί από το NLTK.

Για την εκτέλεση του module απαιτούνται τα παρακάτω πακέτα εγκατεστημένα `rexpact`, `jsonrpc` και `python-progressbar`. Τρέχει το jar αρχείο του `StanfordCoreNLP` σε ξεχωριστό thread και επικοινωνεί με το java module χρησιμοποιώντας τη διεπαφή γραμμής εντολών του. Ένα παράδειγμα εγκατάστασης και χρήσης του συστήματος της εφαρμογής δίνεται στο παράρτημα Α.

## Σύνδεση με MongoDB

Για τη πραγματοποίηση των queries στη βάση δεδομένων που γίνονται μέσω της Web εφαρμογής, χρησιμοποιείται το πρόγραμμα `PyMongo` για τη σύνδεση με τη βάση δεδομένων.

Το `PyMongo`<sup>2</sup> είναι μια διανομή για Python που παρέχει ένα εργαλείο API για πραγματοποίηση εργασιών στη MongoDB, και είναι ο συνιστώμενος τρόπος για τη κλήση της MongoDB σε προγράμματα Python. Η τεκμηρίωση αυτή επιχειρεί να εξηγήσει όλα όσα χρειάζεται να ξέρετε για να χρησιμοποιήσετε `PyMongo`.

Αφού γίνει `import` η συγκεκριμένη βιβλιοθήκη, και πραγματοποιηθεί η σύνδεση με τη βάση γίνονται δύο queries ένα που επιλέγει τα tips με βάση τα aspects, και ένα που συνδέει τα tips με τις επιχειρήσεις, χρησιμοποιώντας το `buisness_id` attribute.

Το MongoDB είναι ένα δωρεάν, ανοικτού κώδικα λογισμικό, συμβατό με όλες τις πλατφόρμες (λειτουργικά συστήματα) σύστημα βάσης δεδομένων προσανατολισμένο στη αποθήκευση και διαχείριση εγγράφων. Χαρακτηρίζεται ως βάση δεδομένων NoSQL, και αποφεύγει το παραδοσιακό μοντέλος βάσης σχεσιακής δομής, εστιάζοντας σε JSON έγγραφα με δυναμικό χαρακτήρα, κάνοντας την ενσωμάτωση των δεδομένων σε ορισμένους τύπους εφαρμογών ευκολότερη και ταχύτερη. Το MongoDB έχει αναπτυχθεί από την MongoDB Inc. και είναι δωρεάν και ανοιχτού κώδικα, που δημοσιεύθηκε κάτω από τις άδειες λογισμικού GNU Affero General Public License και Apache. Τον Ιούλιο του 2015, MongoDB είναι το τέταρτο πιο δημοφιλές σύστημα διαχείρισης βάσεων δεδομένων, και το πρώτο πιο δημοφιλές ανάμεσα στα συστήματα που προσανατολίζονται στην αποθήκευση εγγράφων.

### Δυνατότητες του συστήματος

#### Πραγματοποίηση Ad hoc ερωτημάτων (queries)

MongoDB υποστηρίζει ερωτήματα βάσει του πεδίου και ελυρους, ερωτήματα κανονικών εκφράσεων (regular expressions), τακτική αναζητήσεις έκφρασης. Ερωτήματα που μπορούν να επιστρέψουν συγκεκριμένους τομείς των εγγράφων και επιτρέπουν επίσης τη χρήση συναρτήσεων JavaScript.

#### Υποστήριξη ευρετηρίων (Indexing)

Οποιοδήποτε πεδίο σε ένα έγγραφο MongoDB μπορεί να χρησιμοποιηθεί σε ευρετήρια-συμπεριλαμβανομένων και αυτών που βρίσκονται μέσα σε λίστες ή ενσωματωμένα JSON έγγραφα (δείκτες

---

<sup>2</sup> <https://api.mongodb.com/python/current/api/index.html>

στη MongoDB είναι εννοιολογικά παρόμοιες με εκείνες RDBMSes). Πολλαπλοί δείκτες ανά ευρετήριο είναι επίσης διαθέσιμοι.

### Πολλαπλά αντίγραφα (Replication)

MongoDB παρέχει υψηλή διαθεσιμότητα σε σύνολα αντιγράφων της βάσης. Ένα σύνολο αντιγράφων αποτελείται από δύο ή περισσότερα αντίγραφα των δεδομένων. Κάθε αντίγραφο- μέλος μπορεί να ενεργεί στο ρόλο του πρωτογενούς ή δευτερογενούς αντιγράφου ανά πάσα στιγμή. Όλες οι εγγραφές και οι αναγνώσεις δεδομένων πραγματοποιούνται στο πρωτεύον αντίγραφο στις προεπιλεγμένες ρυθμίσεις. Τα δευτεροβάθμια αντίγραφα διατηρούν ένα αντίγραφο των δεδομένων του πρωτογενούς χρησιμοποιώντας τον ενσωματωμένο μηχανισμό αντιγραφής. Όταν ένα κύριο αντίγραφο δεν είναι διαθέσιμο, το σύστημα των συνόλων των αντιγράφων ρυθμίζεται αυτόματα και διεξάγει μια εκλογική διαδικασία για να καθορίσει ποιο δευτερεύων αντίγραφο πρέπει να γίνει το κύριο. Ένα δευτερεύων μπορεί προαιρετικά να εξυπηρετήσει τις λειτουργίες ανάγνωσης.

### Load balancing (εξισορρόπηση φορτίου)

MongoDB μπορεί να χρησιμοποιηθεί σε ευρεία κλίμακα χρησιμοποιώντας sharding. Ο χρήστης επιλέγει ένα κλειδί, το οποίο καθορίζει πώς θα διανεμηθούν τα δεδομένα σε μια συλλογή. Τα δεδομένα χωρίζονται σε σειρές (ranges) (με βάση το κλειδί) και κατανέμονται σε πολλά τεμάχια (sharding). (Ένα τεμάχιο είναι το κύριο με έναν ή περισσότερους “σκλάβους” τεμάχια.). Εναλλακτικά, το κλειδί- τεμάχιο μπορεί να κατακερματίζεται σε ένα ευρετήριο βάσει των τεμαχίων- που επιτρέπει την ομοιόμορφη κατανομή των δεδομένων.

Η βάση MongoDB μπορεί να τρέξει πάνω σε πολλούς διακομιστές, εξισορροπώντας το φορτίο και / ή να παρέχει μηχανισμό για αντιγραφή δεδομένων για να κρατήσει το σύστημα σε λειτουργία σε περίπτωση αστοχίας υλικού. Η MongoDB είναι εύκολο να εγκατασταθεί, και νέες μηχανές μπορούν να προστεθούν σε μια βάση δεδομένων που “τρέχει” ήδη.

### File storage (αποθήκευση αρχείων)

Η MongoDB μπορεί να χρησιμοποιηθεί ως ένα σύστημα αρχείων, χάριν στην εξισορρόπηση φορτίου και τα χαρακτηριστικά αντιγραφής δεδομένων σε πολλαπλά μηχανήματα για την αποθήκευση αρχείων.

Αυτή η λειτουργία, που ονομάζεται Σύστημα Grid αρχείων (GridFS)], περιλαμβάνεται με τους drivers της MongoDB και είναι διαθέσιμη για πολλές γλώσσες προγραμματισμού. Η MongoDB παρέχει μηχανισμούς για το χειρισμό αρχείων και περιεχομένου για τους προγραμματιστές. Το GridFS χρησιμοποιείται, για παράδειγμα, σε plugins για τους server nginx και το lighttpd. Αντί για την αποθήκευση ενός αρχείου σε ένα ενιαίο έγγραφο, το GridFS χωρίζει ένα αρχείο σε τμήματα, και αποθηκεύει κάθε ένα από αυτά τα κομμάτια ως ξεχωριστό έγγραφο.

Σε ένα σύστημα πολλαπλών μηχανημάτων MongoDB, τα αρχεία μπορεί να διανεμηθούν και να αντιγραφούν πολλές φορές μεταξύ των μηχανημάτων με διαφάνεια, δημιουργώντας έτσι αποτελεσματικά ένα εξισορροπημένο και ανεκτικό σε σφάλματα σύστημα.

### Aggregation (Συσσωμάτωση)

Ο αλγόριθμος MapReduce μπορεί να χρησιμοποιηθεί για την επεξεργασία των δεδομένων και την ομαδοποίηση διαδικασιών.

Το μοντέλο συνάθροισης επιτρέπει στους χρήστες να αποκτήσουν το είδος των αποτελεσμάτων για τα οποία χρησιμοποιείται η SQL χρησιμοποιεί το μηχανισμό GROUP BY. Μηχανισμοί συνάθροιση μπορεί να συνεργαστούν για να σχηματίσουν έναν αγωγό - όπως οι σωλήνες Unix. Το μοντέλο συνάθροισης περιλαμβάνει το χειριστή αναζήτησης "\$" που μπορούν να ενώσει έγγραφα από διαφορετικά έγγραφα.

#### Server-side εκτέλεση Javascript

Κώδικας JavaScript μπορεί να χρησιμοποιηθεί σε ερωτήματα, σε συναρτήσεις συνάθροισης (όπως MapReduce), και αποστέλλεται απευθείας στη βάση δεδομένων που πρόκειται να εκτελεστεί.

#### Capped Collections (συλλογές ανώτατου ορίου)

Η MongoDB υποστηρίζει συλλογές σταθερού μεγέθους που ονομάζεται Capped Collections. Αυτό το είδος της συλλογής διατηρεί τη πληροφορία της εκχώρησης δεδομένων στο σύστημα κατά την εισαγωγή και, μόλις επιτευχθεί το καθορισμένο μέγεθος, συμπεριφέρεται σαν κυκλική ουρά.

Η MongoDB χρησιμοποιήθηκε λόγω του ότι δύναται να αποθηκεύσει δεδομένα απευθείας σε JSON μορφή το οποίο είναι ιδιαίτερα χρήσιμο εφόσον εν τέλει το visualization component απαιτεί το ίδιο format.

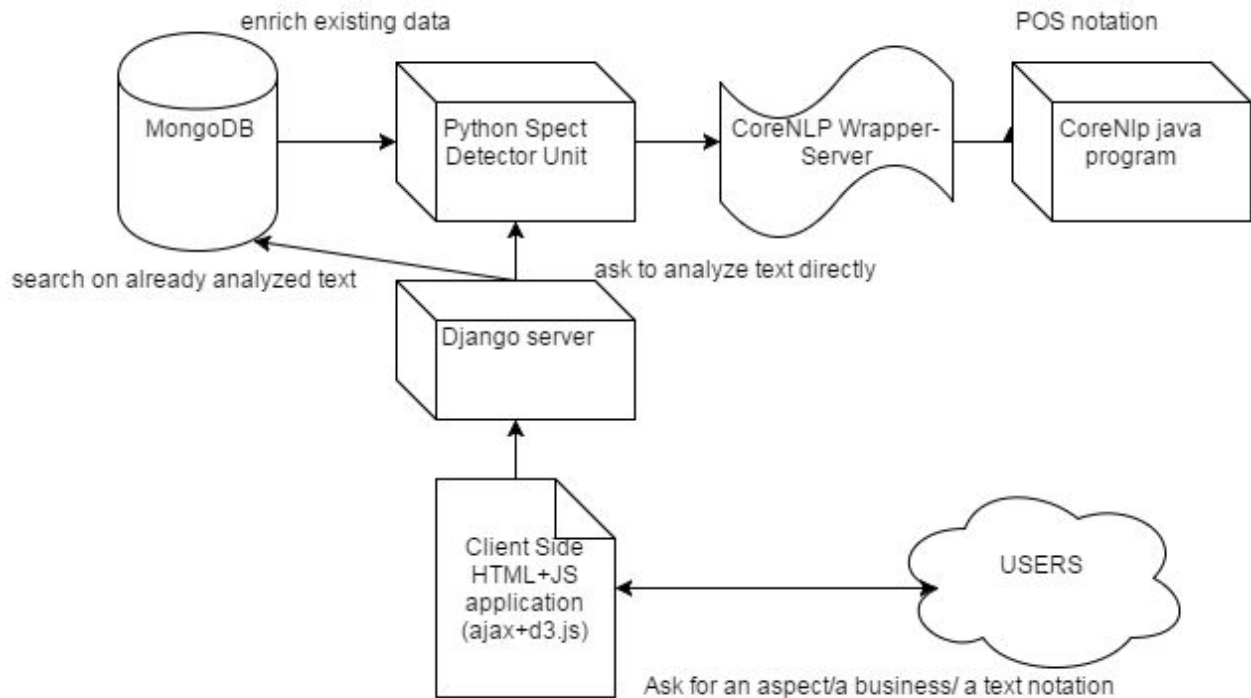
Αρχικά έγινε η εγκαταστημένη της βάσης δεδομένων και δημιουργήθηκε η yelp database. Στη συνέχεια έγινε χρήση ενός python script που για κάθε ένα από τα json files του dataset (έχοντας προεπεξεργαστεί και εμπλουτίσει το tips.json αρχείο), δημιουργεί ένα collection και κάνει import τα data.

Επιλέγοντας τη συγκεκριμένη βάση και συλλογή μπορούμε στη συνέχεια να χρησιμοποιήσουμε τα δεδομένα από την web εφαρμογή μας.

### **E. Αρχιτεκτονική της εφαρμογής**

Η εφαρμογή αποτελείται από τα κομμάτια που περιγράφηκαν παραπάνω και χρησιμοποιούνται για τη δημιουργία της web εφαρμογής. Στο παρακάτω σχήμα αναλύεται η αλληλεπίδραση μεταξύ τους.

Σύμφωνα με το σχήμα 16 τα κύρια μέρη της εφαρμογής και ο ρόλος τους είναι:



Σχήμα 16. Σχήμα αλληλεπίδρασης των διαφορετικών μερών της εφαρμογής

Ο ακόλουθος πίνακας συνοψίζει τα αρχιτεκτονικά στοιχεία που ενσωματώνονται στην εφαρμογή και δίνει μια συνοπτική περιγραφή.

| Component - Αρχιτεκτονικό στοιχείο                     | Περιγραφή                                                                                                                                                                                                                                                                                                                                                 |
|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MongoDB                                                | Βάση δεδομένων που αποθηκεύονται όλοι πίνακες (collections) και στην οποία πραγματοποιούνται ερωτήματα βάσει των πτυχών από τον server. Καθώς και εμπλουτίζονται δεδομένα μέσω του Python CoreNlp wrapper.                                                                                                                                                |
| Python CoreNlp / python wrapper/java Stanford Core NLP | Στήνει ένα server που τρέχει το jar Stanford CoreNlp. Λαμβάνει κείμενο και επιστρέφει το κείμενο συνοφασμένο με POS σημειογραφία. Στη συνέχεια χρησιμοποιείται κώδικας για τον εντοπισμό aspect και sentiment στο POS-annotated text. Το module αυτό καλείται για τον εμπλουτισμό δεδομένων (προεπεξεργασία) αλλά και για την απευθείας ανάλυση κειμένου. |
| Django Server                                          | Ο server παρέχει 3 end-points στον client:<br>1- σύσταση επιχειρήσεων βάσει πτυχών<br>2- ανάλυση πτυχών μιας επιχείρησης<br>3- ανάλυση κειμένου σε πραγματικό χρόνο                                                                                                                                                                                       |
| Front-end application                                  | Η εφαρμογή που τρέχει στον client αποτελείται από                                                                                                                                                                                                                                                                                                         |

|  |                                                                                                                                                                   |
|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | javascript που χρησιμοποιώντας τη βιβλιοθήκη d3.js δημιουργεί το γράφο των πτυχών/επιχειρήσεων αντίστοιχα και πραγματοποιεί την αποστολή των αιτημάτων μέσω ajax. |
|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|

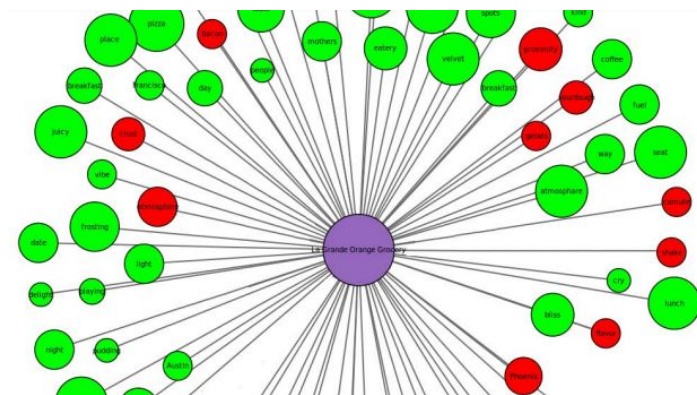
## ΣΤ. Σενάρια Χρήσης της εφαρμογής (Use Cases)

Στο πρώτο σενάριο ένας χρήστης της εφαρμογής ενδιαφέρεται για μία συγκεκριμένη πτυχή μιας κατηγορίας προϊόντων. Ψάχνοντας μία συγκεκριμένη πτυχή μπορεί να δει τις 10 καλύτερα βαθμολογημένους εφαρμογές βάσει της συγκεκριμένης πτυχής. Και στη συνέχεια επιλέγοντας μία να δει όλες τις πτυχές για τις οποίες έχει σχολιαστεί το προϊόν. (Σχήμα 17&18)

Το ίδιο σενάριο θα μπορούσε να εφαρμοστεί στη περίπτωση ήδη υπάρχοντων συστημάτων συστάσεων όπου όταν ένας χρήστης καταχωρεί μία βαθμολογία και ένα σχόλιο αρνητικής βαθμολόγησης να του συστήνονται επιχειρήσεις ή προϊόντα που περιλαμβάνουν τις ίδιες πτυχές αλλά με θετική βαθμολόγηση.



Σχήμα 17. Στιγμιότυπο προβολής δέκα επιχειρήσεων μιας κατηγορίας.



Σχήμα 18. Στιγμιότυπο προβολής των πτυχών με τα το συναίσθημα που συνδέει κάθε μία από αυτές.

Στο δεύτερο σενάριο ένας επιχειρηματίας ή αναλυτής μια συγκεκριμένης αγοράς θέλει να δει τον ανταγωνισμό σε μία ή περισσότερες πτυχές. Κάτι τέτοιο απαιτεί βέβαια συλλογή περισσότερων δεδομένων για κάθε αγορά ξεχωριστά που αποτελεί μία μεγάλη πρόκληση. Για τη πραγματοποίηση ενός τέτοιου φιλόδοξου στόχου θα μπορούσε να δημιουργηθεί μια ξεχωριστή πλατφόρμα καταχωρήσεις σχολίων ή και να υπάρξει δυνατότητα να τρέξει το σύστημα σαν plugin στον browser για τα σχόλια που υπάρχουν σε μια σελίδα.

## V. Συμπεράσματα και μελλοντικές επεκτάσεις

Στα πλαίσια αυτής της πτυχιακής μελετήθηκαν μέθοδοι ανάλυσης κειμένου κριτικών και δημιουργήθηκε ένα θεωρητικό μοντέλο που βάσει της ανάλυσης κριτικών από βάσεις με δομή κοινωνικών δικτύων πραγματοποιεί α) την σύσταση προϊόντων βάσει πτυχών προς το καταναλωτικό κοινό και β) ανάλυση της γνώμης αγοράς προς τις επιχειρήσεις.

Στη συνέχεια με τη χρήση του Stanford CoreNLP αναλύθηκαν κριτικές του κοινωνικού δικτύου Yelp ως προς την σηματοδότηση και εντοπισμό των λεκτικών συνόλων και τα μέρη του λόγου για τη κάθε κριτική. Αναλύοντας την εξαγόμενη πληροφορία με κώδικα python που δημιουργήθηκε στα πλαίσια αυτής της εργασίας, εντοπίστηκαν πτυχές μαζί με το συναίσθημα βάσει λεξικών φορτισμένων λέξεων που συνοδεύει τις πτυχές. Στο χρήστη παρουσιάζεται όλο αυτό το αποτέλεσμα με τη μορφή γραφημάτων μέσω μιας web εφαρμογής που δημιουργήθηκε με τη χρήση και τη μελέτη των: Django MVC web framework που βασίζεται στη γλώσσα προγραμματισμού python, της NoSql document-based βάσης δεδομένων MongoDB και της Javascript βιβλιοθήκης οπτικοποίησης d3.js .

Η προβληματική που τέθηκε αρχικά για το εντοπισμό των πτυχών των προϊόντων και τη σύσταση τους πραγματοποιήθηκε όπως επίσης και η χρήση των ίδιων πτυχών για την ανάλυση των προϊόντων της γνώμης του αγοραστικού κοινού.

Μερικές από τις πιθανές επεκτάσεις της πτυχιακής εργασίας που θα μπορούσαν να πραγματοποιηθούν αναλύονται παρακάτω. Στο επίπεδο της λεκτικής ανάλυσης της εφαρμογής ένας παράγοντας που θα μπορούσε να συντελέσει καταλυτικά είναι η βελτίωση του αλγορίθμου ανίχνευσης των πτυχών που υπάρχουν σε ένα συγκεκριμένο σχόλιο καθώς και υιοθέτηση ευρεστικών αλγορίθμων και η καταγραφή βελτιώσεων που θα μπορούσαν να συνεισφέρουν, όπως αυτοί που αναφέρθηκαν στο βιβλιογραφικό μέρος. Τα λεξικά θα μπορούσαν να εμπλουτιστούν για περισσότερες θεματικές λέξεις καθώς και λέξεις που χρησιμοποιούνται αποκλειστικά στα ηλεκτρονικά μέσα (πχ emoticons) που να είναι επίσης συνυφασμένες με θετική ή αρνητική βαθμολόγηση.

Ένας από τους άμεσους στόχους είναι επίσης η συσταδοποίηση των πτυχών σε ομάδες συνωνύμων.

Υπάρχουν εφαρμοσμένες λύσεις στο συγκεκριμένο πρόβλημα και θα ενεργήσει αποτελεσματικά στην αύξηση του πλήθους των πτυχών που παρουσιάζονται σε έναν γράφο.

Ο γενικός σκοπός της εφαρμογής είναι η αυτοματοποίηση ή τουλάχιστον η υποβοήθηση της αναζήτησης ενός προϊόντος που πληρεί τα κριτήρια ενός συνόλου πτυχών όπως αναφέρθηκαν από το καταναλωτικό κοινό. Κάτι τέτοιο θα μπορούσε όπως περιγράφηκε να χρησιμοποιηθεί αμφότερα από τελικούς χρήστες αλλά και επαγγελματίες αναλυτές. Είναι στα σχέδια η περαιτέρω επέκταση της εφαρμογής σε σενάρια χρήσης που θα υποστηρίζουν ολοκληρωμένες αναλύσεις κοινωνικών δικτύων με σκοπό την καλύτερα-στοχευμένη διαφήμιση προϊόντων, εργαλείου που θα επιτρέψει την πραγματοποίηση ενός ακριβέστερου επιχειρηματικού σχεδίου.

# ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

- ❖ Agocs, A., Tanev, H., Van der Goot, E., Le Goff, J.M., Skogstad, E. and Frangkiskos, S., 2013. *Collaboration Spotting* (No. AIDA-CONF-2013-006).
- ❖ Bao, J., Zheng, Y. and Mokbel, M.F., 2012, November. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (pp. 199-208). ACM.
- ❖ Bellato, N., 2013. Infographics: A visual link to learning. *ELearn*, 2013(12), p.1.
- ❖ Berge, C., 1958. Théorie des graphes et ces applications, collection universitaire de mathématiques. *Paris, Dunod*.
- ❖ Bjørkelund, E., Burnett, T.H. and Nørvåg, K., 2012, December. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services* (pp. 229-238). ACM.
- ❖ Bostock, M., Ogievetsky, V. and Heer, J., 2011. D<sup>3</sup> data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12), pp.2301-2309.
- ❖ Brody, S. and Elhadad, N., 2010, June. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 804-812). Association for Computational Linguistics.
- ❖ Collomb, A., Costea, C., Joyeux, D., Hasan, O. and Brunie, L., 2014. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation [PDF]. *month*.
- ❖ Dawson, P. and Lamb, M., 2016. Enhanced Success with Programmatic Social Advertising. In *Programmatic Advertising* (pp. 103-110). Springer International Publishing.
- ❖ des Vignes, G.V., Sentiment Analysis for Dynamic User Preference Inference in Spoken Dialogue Systems.
- ❖ Diao, Q., Qiu, M., Wu, C.Y., Smola, A.J., Jiang, J. and Wang, C., 2014, August. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 193-202). ACM.
- ❖ Ding, X., Liu, B. and Yu, P.S., 2008, February. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 231-240). ACM.
- ❖ Eirinaki, M., Pisal, S. and Singh, J., 2012. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), pp.1175-1184.
- ❖ Ekstrand, M.D., Riedl, J.T. and Konstan, J.A., 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2), pp.81-173.
- ❖ Fan, T.K. and Chang, C.H., 2010. Sentiment-oriented contextual advertising. *Knowledge and Information Systems*, 23(3), pp.321-344.
- ❖ Fazio, D., Scrivens, K. and Grazia, M., The potential of Web2.0 communities for statistics.
- ❖ Greenfeld, D. and Roy, A., 2013. Two Scoops of Django.
- ❖ Guy, I., 2015. Social recommender systems. In *Recommender Systems Handbook* (pp. 511-543). Springer US.

- ❖ Herman, I., Melançon, G. and Marshall, M.S., 2000. Graph visualization and navigation in information visualization: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1), pp.24-43.
- ❖ Khan, A., Baharudin, B. and Khan, K., 2011. Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure. In *Software Engineering and Computer Systems* (pp. 317-331). Springer Berlin Heidelberg.
- ❖ King, Z.A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N.E. and Palsson, B.O., 2015. Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput Biol*, 11(8), p.e1004321.
- ❖ Li, X., Wang, H., Yin, G., Wang, T., Yang, C., Yu, Y. and Tang, D., 2012. Inducing taxonomy from tags: An agglomerative hierarchical clustering framework. In *Advanced Data Mining and Applications* (pp. 64-77). Springer Berlin Heidelberg.
- ❖ Ling, G., Lyu, M.R. and King, I., 2014, October. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 105-112). ACM.
- ❖ Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), pp.1-167.
- ❖ Lu, Y., Zhai, C. and Sundaresan, N., 2009, April. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web* (pp. 131-140). ACM.
- ❖ Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)* (pp. 55-60).
- ❖ Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)* (pp. 55-60).
- ❖ McAuley, J. and Leskovec, J., 2013, October. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 165-172). ACM.
- ❖ Mirbakhsh, N. and Ling, C.X., 2013, October. Clustering-based factorized collaborative filtering. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 315-318). ACM.
- ❖ Moghaddam, S. and Ester, M., 2013, May. The flda model for aspect-based opinion mining: Addressing the cold start problem. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 909-918). International World Wide Web Conferences Steering Committee.
- ❖ Plekhanova, J., 2009. Evaluating web development frameworks: Django, Ruby on Rails and CakePHP. *Institute for Business and Information Technology*.
- ❖ Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I., 2015, June. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, Denver, Colorado (pp. 486-495).
- ❖ Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S., 2014, August. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 27-35).
- ❖ Proios, D., Eirinaki, M. and Varlamis, I., 2015, August. TipMe: Personalized advertising and aspect-based opinion mining for users and businesses. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 1489-1494). ACM.
- ❖ Rambocas, M. and Gama, J., 2013. *Marketing research: The role of sentiment analysis* (No. 489). Universidade do Porto, Faculdade de Economia do Porto.

- ❖ Ravindran, A., 2015. *Django Design Patterns and Best Practices*. Packt Publishing Ltd.
- ❖ Ricci, F., Rokach, L. and Shapira, B., 2011. *Introduction to recommender systems handbook* (pp. 1-35). Springer US.
- ❖ Roshchina, Alexandra, John Cardiff, and Paolo Rosso. "TWIN: Personality-based Intelligent Recommender System." *Journal of Intelligent & Fuzzy Systems* 28.5 (2015): 2059-2071.
- ❖ Shepitsen, A., Gemmell, J., Mobasher, B. and Burke, R., 2008, October. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems* (pp. 259-266). ACM.
- ❖ Sklar, M. and Concepcion, K.J., 2014. Timely Tip Selection for Foursquare Recommendations. In *RecSys Posters*.
- ❖ Sridaran, R., Padmavathi, G., Iyakutti, K. and Mani, M.N.S., 2010. SPIM Architecture for MVC based Web Applications. *arXiv preprint arXiv:1006.2702*.
- ❖ Suresh, V., Roohi, S. and Eirinaki, M., 2014, October. Aspect-based opinion mining and recommendation system for restaurant reviews. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 361-362). ACM.
- ❖ Suresh, V., Roohi, S., Eirinaki, M. and Varlamis, I., 2014, October. Using Social Data for Personalizing Review Rankings. In *RSWeb@ RecSys*.
- ❖ Taboada, Maite, et al. "Lexicon-based methods for sentiment analysis." *Computational Linguistics* 37.2 (2011): 267-307
- ❖ Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H. and Qu, H., 2010. OpinionSeer: interactive visualization of hotel customer feedback. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6), pp.1109-1118.
- ❖ Yang, Shuang-Hong, et al. "Like like alike: joint friendship and interest propagation in social networks." *Proceedings of the 20th international conference on World wide web*. ACM, 2011.
- ❖ Zhang, Jia, Zhenming Yuan, and Kai Yu. "Cross Media Recommendation in Digital Library." *The Emergence of Digital Libraries—Research and Practices*. Springer International Publishing, 2014. 208-217.



```
        u'Lemma': u'hello',
        u'NamedEntityTag': u'O',
        u'PartOfSpeech': u'UH'}]},
    [u'world',
     {u'CharacterOffsetBegin': u'6',
      u'CharacterOffsetEnd': u'11',
      u'Lemma': u'world',
      u'NamedEntityTag': u'O',
      u'PartOfSpeech': u'NN'}]},
    [u'!',
     {u'CharacterOffsetBegin': u'11',
      u'CharacterOffsetEnd': u'12',
      u'Lemma': u'!',
      u'NamedEntityTag': u'O',
      u'PartOfSpeech': u'.'}]]},
    {u'parsetree': u'(ROOT (S (NP (PRP It)) (VP (VBZ is) (ADJP (RB so) (JJ
beautiful)))) (. .)))',
     u'text': u'It is so beautiful.',
     u'tuples': [[u'nsubj', u'beautiful', u'It'],
                  [u'cop', u'beautiful', u'is'],
                  [u'advmod', u'beautiful', u'so'],
                  [u'root', u'ROOT', u'beautiful']],
     u'words': [[u'It',
                  {u'CharacterOffsetBegin': u'14',
                   u'CharacterOffsetEnd': u'16',
                   u'Lemma': u'it',
                   u'NamedEntityTag': u'O',
                   u'PartOfSpeech': u'PRP'}],
                  [u'is',
                   {u'CharacterOffsetBegin': u'17',
                    u'CharacterOffsetEnd': u'19',
                    u'Lemma': u'be',
                    u'NamedEntityTag': u'O',
                    u'PartOfSpeech': u'VBZ'}],
                  [u'so',
                   {u'CharacterOffsetBegin': u'20',
                    u'CharacterOffsetEnd': u'22',
```

```

        u'Lemma': u'so',
        u'NamedEntityTag': u'O',
        u'PartOfSpeech': u'RB'}]},
    [u'beautiful',
     {u'CharacterOffsetBegin': u'23',
      u'CharacterOffsetEnd': u'32',
      u'Lemma': u'beautiful',
      u'NamedEntityTag': u'O',
      u'PartOfSpeech': u'JJ'}]},
    [u'.',
     {u'CharacterOffsetBegin': u'32',
      u'CharacterOffsetEnd': u'33',
      u'Lemma': u'.',
      u'NamedEntityTag': u'O',
      u'PartOfSpeech': u'.'}]]}],
u'coref': [[[[u'It', 1, 0, 0, 1], [u'Hello world', 0, 1, 0, 2]]]]}

```

Αυτό το JSON αποτελεί ένα python dictionary που περιέχει διάφορα κλειδιά για τη συγκεκριμένη πρόταση προς επεξεργασία. Το κλειδί sentences λεξικού περιέχει μια λίστα λεξικών για κάθε πρόταση, που με τη σειρά της περιέχει τα κλειδιά parsetree, text, tuples που περιέχουν τις εξαρτήσεις, ενώ το κλειδί words, που περιέχει πληροφορίες σχετικά με το τι μέρος του λόγου είναι η κάθε εντοπισμένη οντότητα. Το παραπάνω εξαγόμενο json χρησιμοποιείται στην συνέχεια για την ανίχνευση των aspects βάσει του συνδιασμένου βαθμολογημένου λεξικού αφού “ενωθεί” με το αρχικό tip από στο οποίο υπάρχει η αρχική πρόταση.

### Python Script προεπεξεργασίας δεδομένων (client.py)

Το PythonTreeParser.py αρχείο εκτελεί τις εξείς εργασίες.

- Αρχικά διαβάζει και στέλνει τα tips για επεξεργασία στο module Python CoreNLP wrapper, λειτουργώντας ως client στην προκειμένη περίπτωση.
- Στη συνέχεια “γράφει” το αποτέλεσμα στο tip και όλα μαζί πάλι σε ένα αρχείο.

### Python Script προεπεξεργασίας δεδομένων (TreeParser.py)

- Το treeParser αρχείο παίρνει το επεξεργασμένο αρχείο από τον Client
- Στη συνέχεια για κάθε λεκτικό σύνολο του tip αναγνωρίζει τα ουσιαστικά ως υποψήφια aspects, και τα επίθετα, τα ρήματα και τα επιρρήματα ως υποψήφια sentiment words.

- Ελέγχουμε ποια από τα υποψήφια sentiment words υπάρχουν στο λεξικό μας. Και κρατάμε μόνο αυτά. Είναι σημαντικό να σημειωθεί ότι λέξεις που δεν υπάρχουν στο λεξικό δεν λαμβάνονται υπόψη ως προς το sentiment.
- Για κάθε λεκτικό σύνολο υπάρχει ένα ουσιαστικό άρα μπορούμε να αναθέσουμε όλα τα sentiment words σ' αυτό ως τη μοναδική πτυχή.
- Αν υπάρχουν ένα ή περισσότερα ουσιαστικά παίρνουμε το συνολικό άθροισμα των sentiment words που υπάρχουν σε μια πρόταση.
- Τέλος αναθέτουμε το σύνολο των aspects που βρέθηκαν στο tip, ώστε να μπορεί να διασυνδεθεί με τον χρήστη και την επιχείριση μέσω των αναγνωριστικών primary\_key id