Ευθυγράμμιση προτάσεων φυσικής γλώσσας με εξαγωγή ομοιότητας λέξεων από γράφο συνωνύμων

Ανδρέας Γρίβας 21005

Τμήμα Πληροφορικής και Τηλεματικής
Χαροκόπειο Πανεπιστήμιο

Πτυχιακή εργασία προπτυχιακού επιπέδου

Αθήνα, 27 Οκτωβρίου 2014

Στους γονείς μου που με στήριξαν αμέριστα ακόμα και όταν δεν συμφωνούσαν με τις επιλογές μου

Σύνοψη

Ευθυγράμμιση προτάσεων φυσικής γλώσσας με εξαγωγή ομοιότητας λέξεων από γράφο συνωνύμων

Ανδρέας Γρίβας

Επιβλέπων:

Δημοσθένης Αναγνωστόπουλος, Καθηγητής

Μέλη τριμελούς επιτροπής :

Μάρα Νικολαΐδου, Καθηγήτρια

Ουρανία Χατζή, Μεταδιδακτορική ερευνητής

Σύνοψη

Η ευθυγράμμιση προτάσεων φυσικής γλώσσας αποτελεί ένα σημαντικό βήμα για να μπορέσει κανείς να συνεχίσει στην εύρεση συνεπαγωγών και αντιθέσεων σε κείμενο. Στόχος της είναι, σε δύο κείμενα τα οποία βρίσκονται στην ίδια γλώσσα, να βρεθούν τα υποσύνολα των λέξεων που ταιριάζουν νοηματικά. Έτσι με μια περαιτέρω ανάλυση θα μπορεί κανείς να συμπεράνει αν τα δύο κείμενα μιλούν για το ίδιο θέμα.

Η προσέγγιση της πτυχιακής για την ευθυγράμμιση προτάσεων βασίζεται στην απεικόνιση λέξεων ως διανύσματα, τα οποία κατασκευάζονται κατά την διάσχιση γράφου που περιέχει πληροφορία συνωνύμων και ορισμών. Φιλοδοξία της είναι να εκμεταλλευτεί την αθροισιμότητα των διανυσμάτων για να είναι εύκολα επεκτάσιμη σε φράσεις καθώς και να πιάνει τις μεταβολές του νοήματος που εισάγουν κατά τη διαδικασία αυτή τα συμφραζόμενα.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τους επιβλέποντες μου κ. Δημοσθένη Αναγνωστόπουλο, κα. Μάρα Νικολαΐδου και κα. Ουρανία Χατζή για την κατανόηση και την καθοδήγηση τους παρόλη την απόκληση από το αρχικό θέμα που είχε συζητηθεί και τον ιδιόρυθμό μου επικοινωνιακό χαρακτήρα.

Επίσης είμαι ευγνώμων στον (κ.)[1] Γιώργο Γιαννακόπουλο τον κ. Ηρακλή Βαρλάμη και τον κ. Δημήτριο Μιχαήλ που άκουσαν τους προβληματισμούς μου και μοιράστηκαν την γνώση τους.

Ευχαριστώ τους Αλέξανδρο Ελευθερίου, Εμμανουήλ Μαρκόνη, Παναγιώτη Ευστρατιάδη, Ιωάννη Κατάκη, Γιώργο Παναγόπουλο και Νικόλαο Ζορμπά για την υποστήριξη, την συμμετοχή στα πειράματα, και την ανοχή που έδειξαν όλο αυτόν τον καιρό στα κακά ανέκδοτα!

Τέλος, ευχαριστώ τον δημιουργό του Vim Bram Moolenaar, και του TEX Donald E. Knuth καθώς και όσους έχουν συνεισφέρει σε αυτά με τα χρόνια, όπως και τον δημιουργό αυτού του όμορφου template M. Imran, που έκαναν ευκολότερη και ομορφότερη την συγγραφή του κειμένου της πτυχιακής.

[1] if (reader.getName()=='Γιώργος Γιαννακόπουλος') parentheses.deleteContent();

Παραδοχές

Για να είναι δυνατή η κατανόηση του κειμένου, η παραδοχή είναι πως ο αναγνώστης έχει γνώσεις ενός τριτοετούς φοιτητή Πληροφορικής. Για παράδειγμα, τι είναι γράφος, *xml* και κανονικές εκφράσεις.

Έγινε προσπάθεια να αποδοθούν πιστά οι μεταφράσεις αγγλικών όρων σε ελληνικές, ωστόσο όταν κάτι τέτοιο δεν ήταν εφικτό, ή ήταν αστείο για παράδειγμα : bag of words -> σακούλα λέξεων , θα αναγράφεται ο αγγλικός όρος σε αγκύλες [] , π.χ. [for example].

Σε μερικές περιπτώσεις, όταν ο ελληνικός όρος χρησιμοποιείται στην βιβλιογραφία, θα συνοδεύεται από τον Αγγλικό για λόγους πληρότητας σε παρενθέσεις (). Για παράδειγμα: κανονικές εκφράσεις (regular expressions)

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κατάλογος αλγορίθμων

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Το θέμα το οποίο αναλύθηκε αρχικά στα πλαίσια της παρούσας πτυχιακής ήταν η αυτοματοποιημένη μέτρηση αξιοπιστίας των μέσων μαζικής ενημέρωσης. Πρόκειται για ένα πολύ ενδιαφέρον πρόβλημα το οποίο για να το επεξεργαστεί κανείς πιο βαθιά από το να ψάχνει απλά αναδημοσιεύσεις ίδιου περιεχομένου στον ιστό, αναγκάζεται να παραδεχτεί πως χρειάζεται να "μάθει" στον υπολογιστή πως να αναλύει τον ανθρώπινο λόγο - την φυσική μας γλώσσα - για να μπορέσει να αντιμετωπίσει ένα τέτοιο ζήτημα ρεαλιστικά.

Η επεξεργασία της φυσικής γλώσσας ως κλάδος της πληροφορικής έχει προοδεύσει αρκετά τα τελευταία 10 με 20 χρόνια. Με την υιοθέτηση και την εδραίωση της χρήσης τεχνικών μηχανικής μάθησης για την εκμετάλλευση της σωρείας δεδομένων φυσικής γλώσσας που υπάρχει στο διαδίκτυο, η επεξεργασία της φυσικής γλώσσας επωφελείται πλέον από τις προόδους της, καθώς ο τομέας αυτός χαίρει μεγάλης προσοχής αυτό το διάστημα από την ερευνητική και την ακαδημαϊκή κοινότητα και δέχεται αρκετή χρηματοδότηση από οργανισμούς και επιχειρήσεις που φιλοδοξούν να καινοτομήσουν.

Ωστόσο, παρά την ακμή και την δραστηριότητα που υπάρχει στον χώρο, οι μέθοδοι επεξεργασίας και ανάλυσης φυσικής γλώσσας με χρήση ηλεκτρονικού υπολογιστή δεν παύουν ακόμα να είναι πιο κοντά στην αναγνώριση μοτίβων σε δεδομένα παρά στην ανάλυση και την κατανόηση της γλώσσας και την εξαγωγή συμπερασμάτων.

Ένα βασικό πρόβλημα κατανόησης γλώσσας στο οποίο μπορούμε να ανάγουμε την διαπίστωση ή μη της αξιοπιστίας μιας πηγής είναι η εύρεση αντιθέσεων σε κείμενο. Αν μπορούσαμε να πούμε με βεβαιότητα πως δύο κείμενα είναι ασυμβίβαστα, θα μπορούσαμε στην συνέχεια να ισχυριστούμε πως τουλάχιστον ένας από τους δύο συγγραφείς μας παραπληροφορεί.

Οι συγκρούσεις μπορούν να είναι πολλών μορφών. Για παράδειγμα, μπορεί να αναφέρονται διαφορετικές ημερομηνίες ή διαφορετικά αριθμητικά στοιχεία σε δύο κείμενα που μιλούν για το ίδιο γεγονός. Μπορεί να είναι λογικά ασύμβατα, δηλαδή το ένα γεγονός να αποκλείει την δυνατότητα να έχει συμβεί και το δεύτερο στο ίδιο χρονικό πλαίσιο. Φυσικά, το χαρακτηριστικό

που συμβαίνει πιο συχνά από όλα, να γίνεται λόγος για το ίδιο θέμα αλλά η κάθε πηγή να εστιάζει αλλού και να κάνει παραλήψεις ανάλογα με τα συμφέροντά της.

Ακόμα και η εύρεση αντιθέσεων σε κείμενα ωστόσο, προϋποθέτει πως μπορεί κανείς να διακρίνει αν δύο κείμενα μιλούν για το ίδιο θέμα, και ποια κομμάτια από κάθε κείμενο σχετίζονται. Συνεπώς, πρέπει κανείς να μπορεί να συγκρίνει λέξεις και φράσεις με βάση το νόημά τους και όχι με βάση τα γράμματα που τις αποτελούν.

Το εγγενές πρόβλημα με την σύγκριση οντοτήτων στον υπολογιστή σε σχέση με τον τρόπο που κάνει συγκρίσεις ένας άνθρωπος, είναι πως οι μεταβλητές που επιλέγονται για να γίνουν οι συγκρίσεις είναι αυστηρά ορισμένες.

Με τον όρο αυστηρά ορισμένες εννοείται πως η σύγκριση οντοτήτων σε υπολογιστή καταλήγει τελικά σε επίπεδο μηχανής, με τον έναν ή με τον άλλο τρόπο, να ανάγεται πάντα σε σύγκριση byte. Αυτό μπορεί να είναι πολύ χρήσιμο για τον υπολογισμό μαθηματικών εκφράσεων, όμως ο άνθρωπος επεξεργάζεται και διακρίνει ομοιότητες ανάμεσα σε οντότητες που είναι πολύ πιο γενικές, περιέχουν ασάφειες και οι ομοιότητες που περιέχουν πολλές φορές είναι μη προφανείς. Ομοιότητες όπως η ομοιότητα λέξεων, εικόνων και ήχων. Ομοιότητες τέτοιου είδους είναι πολύ χρήσιμο να γίνονται "αντιληπτές" αυτοματοποιημένα με τη χρήση ηλεκτρονικού υπολογιστή. Ωστόσο, δεν είναι εύκολη η αναγωγή των αφηρημένων χαρακτηριστικών που διακρίνει ο άνθρωπος σε χειροπιαστούς κανόνες σύγκρισης byte που μπορεί να εφαρμόσει ένας υπολογιστής.

Στην βιβλιογραφία έχει γίνει πολλή δουλειά πάνω στην αναζήτηση αντιθέσεων και μετρικών ομοιότητας ανάμεσα σε λέξεις, κείμενα και συγγραφείς. Ωστόσο, αν και τα προβλήματα αυτά έχουν αναλυθεί από πολλές διαφορετικές σκοπιές, οι λύσεις που παρέχονται είναι πολύ εξειδικευμένες στο υποπρόβλημα το οποίο προσπαθούν να λύσουν και τα δεδομένα στα οποία εφαρμόζονται. Κατά αυτόν τον τρόπο η αποτελεσματική επίλυση των παραπάνω προβλήματων για ένα μεγάλο έυρος δεδομένων και κυρίως η δημιουργία εφαρμογών που βασίζονται στην λύση τους αποτελούν δύσκολες εργασίες.

Συνεπώς, θεωρήθηκε μάταια η κατασκευή κώδικα που απλά θα εφάρμοζε κανόνες για την αναζήτηση αντιθέσεων ή την ομοιότητα λέξεων. Έγινε προσπάθεια να εστιαστεί περισσότερο ένας πιο ελαστικός τρόπος εύρεσης ομοιότητας ανάμεσα σε λέξεις, ο οποίος να μπορεί να γενικευτεί και να είναι βασισμένος όσο γίνεται σε σχέσεις που μπορούν να απεικονιστούν σε υπολογιστή και περιέχουν για εμάς νόημα. Σκοπός ήταν στην συνέχεια να επεκταθεί αυτή η έννοια της ομοιότητας για φράσεις και συνεπώς να μπορεί να γίνει έλεγχος αν δύο κείμενα μιλούν για το ίδιο θέμα και ποια μέρη τους σχετίζονται.

Η ιδέα είναι πως αν ρωτήσει κανείς έναν άνθρωπο τι σημαίνει μια λέξη, θα προσπαθήσει να

μας την εξηγήσει με βάση άλλες λέξεις και έννοιες που γνωρίζουμε. Για αυτό άλλωστε ζωγραφίζουμε τους εξωγήινους πράσινους αλλά ανθρωπόμορφους, πράσινους γιατί έχουμε φαντασία και ανθρωπόμορφους γιατί δεν έχουμε. Περιοριζόμαστε από το κοινό περιβάλλον στο οποίο πρέπει να αναφερθούμε για να μπορέσουμε να επικοινωνήσουμε.

Η κατανόηση λέξεων μέσω άλλων λέξεων βέβαια είναι ο εσπευσμένος τρόπος να μάθει κανείς μια έννοια. Ο άνθρωπος έχει την ικανότητα να μαθαίνει και με την μακροχρόνια ενεργητική έκθεση στο περιβάλλον του. Κατά την διαδικασία αυτή ο άνθρωπος επεξεργάζεται την πληροφορία που αντιλαμβάνεται και βγάζει ορισμένα συμπεράσματα. Με βάση αυτά και με βάση την επιρροή του περιβάλλοντος, συνδέει την φυσική γλώσσα με αισθήσεις, εικόνες, ήχους, αναμνήσεις και συναισθήματα λόγω της ανάγκης για επικοινωνία με άλλους.

Συνεπώς, το σύνολο όλων αυτών των συνδέσεων που κάνει κανείς ενώ σκέφτεται μια λέξη είναι το σύνολο των εννοιών που περιέχει η λέξη για αυτόν. Οι συνδέσεις μπορούν να είναι διαφορετικές για κάθε άνθρωπο στον βαθμό που επιτρέπει το κοινό περιβάλλον, καθώς το περιβάλλον και οι συζητήσεις με όσους βρίσκονται σε αυτό έχουν καταλυτικό ρόλο στην εύρεση ισορροπίας στο σύστημα.

Με βάση τις παραπάνω παρατηρήσεις - που δεν είναι αναγκαστικά ορθές αλλά εξηγούν τον χαρακτήρα και το σκεπτικό της προσέγγισης - έγινε προσπάθεια η μέθοδος αναζήτησης ομοιότητας να μην εστιάζει τόσο στην ίδια την λέξη, αλλά στη σχέση της με άλλες. Έτσι δημιουργήθηκε αρχικά ένας γράφος λέξεων στον οποίο κόμβοι ήταν όλες οι λέξεις που έχει ένα λεξικό, και υπήρχε συνδέση από την πρώτη στην δεύτερη λέξη, αν η δεύτερη υπήρχε στον ορισμό της πρώτης. Ένας τρόπος που μπορεί να φανταστεί κανείς την διαδικασία είναι σε μια κάτοψη του γράφου. Όταν εξετάζεται μια λέξη, χρωματίζεται μια άμεση γειτονιά της με τις λέξεις που υπήρχαν στον ορισμό της. Το σύνολο των χρωματισμένων κόμβων ως κατάσταση εκφράζουν - αποτελούν την έννοια της λέξης που εξετάζουμε.

Για την σύγκριση συνεπώς δύο λέξεων μπορούμε να φανταστούμε πως έχουμε δύο κατόψεις γράφων, έναν για κάθε λέξη. Αρχικά σημειώνουμε τις δύο λέξεις που θέλουμε να συγκρίνουμε. Στην συνέχεια χρωματίζουμε τα συνώνυμα τους. Για να συγκρίνουμε την ομοιότητα μπορούμε να φανταστούμε πως βάζουμε την μια κάτοψη πάνω στην άλλη και εξετάζουμε την επικάλυψη τους.

Μια προσέγγιση για επέκταση σε επίπεδο φράσεων θα ήταν να προβάλλουμε σε έναν γράφο τους γράφους - κατόψεις των λέξεων της φράσης αθροιστικά. Έτσι, πάλι θα έχουμε δύο κατόψεις, μια για κάθε φράση. Με αυτόν τον τρόπο οι επικαλύψεις εννοιών που υπήρχαν σε παραπάνω από μια λέξεις μπορούμε να φανταστούμε πως θα διακρίνονται με πιο πυκνό χρώμα, ενώ τα

κομμάτια του γράφου που εκφράζουν έννοιες που δεν εκφράστηκαν από πολλές λέξεις, θα είναι πιο αχνές. Έτσι ίσως θα μπορούσε να μπεί στο σχέδιο και η αποσαφήνιση εννοιών, καθώς έννοιες της λέξης που δεν εκφράστηκαν στη φράση θα παραμείνουν αχνές, ενώ αυτές που εκφράστηκαν και σε άλλες λέξεις της φράσεις θα είναι πιο έντονες.

Σκοπός της εισαγωγής αυτής της πιο αφηρημένης ομοιότητας λέξεων και φράσεων τελικά είναι, δοθέντος δύο κειμένων, η εύρεση κομματιών ή λέξεων που μπορούν να αντιστοιχιστούν νοηματικά. Απώτερος σκοπός είναι να αποφασιστεί αν δύο κείμενα αναφέρονται στο ίδιο θέμα ή γεγονός.

Θετικές πτυχές της πιο χαλαρής ομοιότητας που κατασκευάσαμε είναι η εύκολη επέκταση από επίπεδο λέξεων σε επίπεδο φράσεων, η εύρεση ομοιότητας ακόμα και ανάμεσα σε λέξεις που δεν είναι άμεσα συνώνυμες και η εισαγωγή των συμφραζόμενων στην διαδικασία της εξαγωγής ομοιότητας. Επίσης, αποφεύγεται η εφαρμογή πολύπλοκων και αυστηρών γλωσσικών συντακτικών κανόνων για την εύρεση συντακτικών ομοιοτήτων, κανόνες οι οποίοι έχουν την τάση να μην ισχύουν πάντα.

Αρνητικές πτυχές της μεθόδου είναι η εξάρτησή της από το σύνολο δεδομένων που χρησιμοποιείται για την δημιουργία του γράφου και η πολυπλοκοτητα που εισάγει η διαχείριση και η επεξεργασία του γράφου και των μεθόδων διάσχισής του για την εξαγωγή της ομοιότητας.

Όπως ανέφερε σε μια παρατήρησή του ο *Alan Perlis* [Perlis, 1982],

> "Οι ανόητοι αγνοούν την πολυπλοκότητα.
> Οι ρεαλιστές υποφέρουν από αυτή.
> Μερικοί την αποφεύγουν.
> Οι ιδιοφυΐες την αφαιρούν."

Εγώ την προσέθεσα.

ΚΕΦΑΛΑΙΟ 2

Βιβλιογραφία

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δούμε την σχετική βιβλιογραφία χωρισμένη σε δύο σκέλη. Το πρώτο σκέλος αφορά εισαγωγικές έννοιες και το δεύτερο αφορά θέματα πιο εξειδικευμένα ως προς την ευθυγράμμιση προτάσεων.

Αναλυτικά:

- Στην ενότητα 2.2 θα δούμε κατηγορίες λέξειων και τις ομάδες σχέσεων ανάμεσα σε λέξεις. Επίσης θα περιγραφεί η διαδικασία της κατηγοριοποίησης δεδομένων και οι μετρικές για την αξιολόγησή της. Τέλος θα περιγραφούν μερικές βασικές διαδικασίες επεξεργασίας φυσικής γλώσσας.
- Στην ενότητα 2.3 θα δούμε τι είναι η πρόκληση RTE, καθώς και άλλες σχετικές εργασίες πάνω σε εύρεση αντιθέσεων και ευθυγράμμιση προτάσεων.

2.2 Γενική Βιβλιογραφία

Στην υποενότητα αυτή θα επεξηγήσουμε βασικές εισαγωγικές έννοιες οι οποίες θα χρειαστούν για την επεξήγηση των βημάτων που ακολουθήθηκαν κατά την υλοποίηση. Επίσης θα διευκρινήσουμε την ορολογία που χρησιμοποιείται καθώς και τι εννοούμε με την χρήση κάθε όρου.

2.2.1 Κατηγορίες λέξεων

Μέρη του λόγου

Οι γλωσσολόγοι έχουν δημιουργήσει πληθώρα κατηγοριοποιήσεων όσον αφορά τα σύνολα λέξεων. Αρχικά υπάρχει η κατηγοριοποίηση των λέξεων ανάλογα με το μέρος του λόγου στο οποίο ανήκουν.

Στοιχεία για χωρισμό των λέξεων σύμφωνα με κάποια γραμματική πρώτα έχουμε από την Αρχαία Ελλάδα από τους Αλεξανδρινούς Γραμματικούς τον $3_ο$ και $2_ο$ αιώνα π.Χ. Ο Διονύσιος ο Θράξ ήταν ο πρώτος, απ' όσο γνωρίζουμε, γραμματικός της αρχαιότητας που συνέταξε συστηματικό εγχειρίδιο γραμματικής της αρχαίας ελληνικής γλώσσας με τίτλο Τέχνη Γραμματική. Η Τέχνη Γραμματική αποτελούνταν από 3 βιβλία και η ανάλυση των λέξεων που περιείχε ήταν περισσότερο μορφολογική. Πλέον αμφισβητείται από μερικούς ερευνητές αν όντως έγραψε όλο το περιεχόμενο [De Jonge, 2008]. Σύμφωνα με τον Διονύσιο τα μέρη του λόγου ήταν 8:

1. ονόματα
2. ρήματα
3. αντωνυμίες
4. προθέσεις
5. επιρρήματα
6. σύνδεσμοι
7. άρθρα
8. μετοχές

Οι κατηγορίες που μαθαίνει κανείς δύο χιλίαδες χρόνια μετά στο σχολείο να χωρίζει τις λέξεις, για τα Αγγλικά είναι πάλι 8, ωστόσο παρουσιάζονται μερικές αλλαγές.

1. ουσιαστικά
2. ρήματα
3. αντωνυμίες
4. προθέσεις
5. επιρρήματα
6. συνδέσμοι
7. επιφωνήματα
8. επίθετα

Η εισαγωγή τους εδώ γίνεται καθαρά για λόγους ανάλυσης, καθώς θα πρέπει να δούμε παρακάτω τις επισημειώσεις που κάνει ένας συντακτικός αναλυτής στις λέξεις. Άλλωστε όλοι μας ξέρουμε πως μπορούμε να μάθουμε να χρησιμοποιήσουμε μια γλώσσα χωρίς να γνωρίζουμε τίποτα για την παραπάνω κατηγοριοποίηση. Αξίζει να σημειωθεί πως δεν υπήρχε πάντοτε απόλυτη συμφωνία ανάμεσα σε όσους δούλεψαν πάνω στα μέρη του λόγου, όπως επίσης πως παρουσιάζουν διαφορές από γλώσσα σε γλώσσα.

Σχέσεις λέξεων

Αν ψάξει κανείς μπορεί να βρεί πάρα πολλές ιδιότητες με τις οποίες μπορεί να φτιάξει ομάδες λέξεων και να τις ονοματίσει. Για παράδειγμα, αν χρησιμοποιήσουμε την ιδιότητα: οι λέξεις να έχουν ίδια κατάληξη, μπορεί να φτιάξει μια ομάδα με λέξεις που παρουσιάζουν ομοιοκαταληξία. Βέβαια, αν δεν γράφεις ποιήματα, κάτι τέτοιο μπορεί να είναι εντελώς ανούσιο!

Για αυτόν τον λόγο, θα περιοριστούμε σε μερικές ομάδες λέξεων που είχαν στα πλαίσια της πτυχιακής νόημα να αναλυθούν για την εύρεση ομοιότητας ανάμεσα σε φράσεις. Τέτοιες είναι:

- συνώνυμα
- υπερώνυμα
- ολώνυμα
- αντώνυμα
- υπώνυμα
- μερώνυμα

- τροπώνυμα
- συνεπαγωγές
- σχετικές φράσεις
- συμπληρωματικές λέξεις

Το wordnet [Miller, 1995] είναι μια λεξικολογική βάση δεδομένων που αναπτύχθηκε από το πανεπιστήμιο του Princeton και περιέχει πολλές από τις παραπάνω σχέσεις. Συγκεκριμένα, το wordnet απεικονίζει κάθε λέξη με την διαίρεση της σε νοηματικές κατηγορίες που ονομάζει synset [συνωνυμοσύνολο]. Περιέχει παραπάνω από 118.000 διαφορετικών μορφών λέξεις και 90.000 διαφορετικές νοηματικές κατηγορίες. Η σύνδεση των λέξεων με τις νοηματικές κατηγορίες δημιουργούν 166.000 διακριτά ζευγάρια λέξης - απόδοσης νοήματος. Πάνω από το 40% των λέξεων στο wordnet έχουν πάνω από ένα συνώνυμο. Ας δούμε αναλυτικότερα τι σημαίνει κάθε μια από αυτές.

Συνώνυμα

Συνώνυμες είναι μια λέξη με μια δεύτερη, όταν είναι παρόμοιες νοηματικά και μπορούν να χρησιμοποιηθούν ή πρώτη στην θέση της δεύτερης. Για παράδειγμα,
κοιτάω - βλέπω.

Απόλυτα συνώνυμα είναι λέξεις που έχουν το ίδιο ακριβώς νόημα. Είναι σπάνια, καθώς δεν έχουν κάποια χρησιμότητα και με τον καιρό εκλείπουν από την γλώσσα. Τα συνώνυμα συνεπώς συνήθως διαφέρουν σε ένα μικρό χαρακτηριστικό που έχει να κάνει είτε με το νόημα είτε με τον τρόπο χρήσης της λέξης. Επίσης υπάρχουν τα κοντινά συνώνυμα [near synonyms] τα οποία αποτελούνται από σχετικές λέξεις με πιο ελαστικούς όρους σύνδεσης [Edmonds and Hirst, 2002].

Όσον αφορά την σχέση των συνωνύμων, οι γλωσσολόγοι - με λίγες εξαιρέσεις - θεωρούν πως είναι συμμετρική [Murphy, 2003] , δηλαδή πως αν το *a* είναι συνώνυμο του *b* τότε συνεπάγεται και το αντίστροφο. Μάλιστα η άποψη αυτή είναι η επικρατέστερη. Όμως υπάρχουν πολλές προσεγγίσεις ειδικά στον χώρο της πληροφορικής που θεωρούν την σχέση μη συμμετρική κατά την ανάλυση και την απεικόνιση των σχέσεων [Sinha and Mihalcea, 2011] [Michelbacher et al., 2011] [Blondel, 2002] [Beeferman, 1998] .

Αντώνυμα

Αντώνυμες είναι μια λέξη με μια δεύτερη, όταν είναι αντίθετες νοηματικά και η αντικατάσταση της πρώτης με την δεύτερη, μπορεί να επιφέρει την άρνηση ολόκληρης της φράσης ή σημαντική αλλαγή του νοήματος. Και σε αυτήν την περίπτωση η επικρατούσα άποψη είναι πως η σχέση είναι συμμετρική. Για παράδειγμα,

καλός - κακός.

Υπερώνυμα

Υπερώνυμο μιας λέξης είναι μια δεύτερη λέξη, αν η δεύτερη έχει πιο ευρεία έννοια από την πρώτη - με την έννοια πως ενσωματώνει τα γενικά χαρακτηριστικά της δεύτερης, αλλά όχι τα ειδικά. Για παράδειγμα,
θηλαστικό - σκύλος.

Υπώνυμα

Υπώνυμο μιας λέξης είναι μια δεύτερη λέξη, αν η δεύτερη έχει πιο ειδική έννοια από την πρώτη - με την έννοια πως περιέχει τα γενικά χαρακτηριστικά της πρώτης, αλλά τα επεκτείνει με εξειδικεύσεις. Για παράδειγμα,
σκατζόχοιρος - ζώο.

Μερώνυμα

Μερώνυμο μιας λέξης είναι μια δεύτερη λέξη, αν η δεύτερη αποτελεί νοηματικό μέρος ή δομικό συστατικό της πρώτης. Για παράδειγμα,
μπουκάλι - καπάκι.

Ολώνυμα

Ολώνυμο μιας λέξης είναι μια δεύτερη λέξη, όταν η πρώτη αποτελεί νοηματικό μέρος ή δομικό συστατικό της δεύτερης. Για παραδειγμα,
καπάκι - μπουκάλι.

Τροπώνυμα

Τροπώνυμο ενός ρήματος είναι ένα δεύτερο ρήμα, αν το πρώτο δηλώνει πως γίνεται το δεύτερο με άλλον τρόπο. Για παράδειγμα,
τρέχω - περπατώ.

Συνεπαγωγές

Συνεπαγωγή ενός ρήματος είναι ένα δεύτερο ρήμα, αν σε περίπτωση που πραγματοποιηθεί το πρώτο, συνεπάγεται πως έχει γίνει και το δεύτερο. Για παράδειγμα,
περπατώ - κινούμαι.

Σχετικές φράσεις

Σχετικές φράσεις μιας λέξης, είναι φράσεις που περιέχουν το νόημα της λέξης. Για παράδειγμα, *τα φορτώνω στον κόκορα - τεμπελιάζω*.

Συμπληρωματικές λέξεις

Συμπληρωματικές λέξεις μια άλλης, είναι λέξεις που δεν ανήκουν σε καμία από τις παραπάνω κατηγορία, όμως έχουν δυνατή σχέση και εκφράζουν το νόημά της. Για παράδειγμα,
όμοιος - ομοιότητα.

2.2.2 Κατηγοριοποίηση δεδομένων

Σε αυτήν την υποενότητα θα περιγράψουμε την διαδικασία της κατηγοριοποίησης δεδομένων καθώς και τις μετρικές σύμφωνα με τις οποίες γίνεται η αξιολόγησή της.

Γενικά χαρακτηριστικά

Κατηγοριοποίηση δεδομένων κάνει κανείς όταν θέλει να χωρίσει τα δεδομένα του σε κατηγορίες σύμφωνα με ορισμένα χαρακτηριστικά. Συνήθως απώτερος σκοπός είναι η εξαγωγή συμπερασμάτων και γνώσης από και για τα δεδομένα. Για παράδειγμα, μπορεί κανείς να κατηγοριοποιήσει αυτοκινητιστικά ατυχήματα σύμφωνα με το φύλο και την ηλικία και να συμπεράνει ποιές κοινωνικές ομάδες πλήττονται περισσότερο.

Η κατηγοριοποίηση μπορεί να χωριστεί σε δυαδική κατηγοριοποίηση και κατηγοριοποίηση σε πολλαπλές κλάσεις. Στην πρώτη περίπτωση η απόφαση που καλείται να πάρει κανείς αφορά δύο εναλλακτικές. Μια βολική αναπαράσταση της δυαδικής κατηγοριοποίησης είναι η απόφαση ανάμεσα σε *Ναι* και *Όχι* για κάθε στιγμιότυπο δεδομένων που εκφράζεται με ένα σύνολο χαρακτηριστικών. Στην δεύτερη περίπτωση η απόφαση αφορά περισσότερες εναλλακτικές και μπορεί να αναπαρασταθεί ως διαμοιρασμός των δεδομένων στις ομάδες σύμφωνα με το ποια είναι η πιο αντιπροσωπευτική ομάδα για κάθε περίπτωση.

Παράδειγμα προβλήματος δυαδικής κατηγοριοποίησης είναι η αναγνώριση ονομάτων σε κείμενο - μια λέξη είτε είναι είτε δεν είναι όνομα. Παράδειγμα προβλήματος κατηγοριοποίησης πολλαπλών κλάσεων είναι για παράδειγμα η αναγνώριση μερών του λόγου. Κάθε λέξη εξετάζεται ως προς κάποια χαρακτηριστικά (επισημείωση προηγούμενης λέξης, στοιχεία προηγούμενων επισημειώσεων της ίδιας λέξης) και επισημειώνεται ως κάποιο μέρος του λόγου.

Μετρικές αξιολόγησης

Όπως είναι φυσικό, όταν κανείς κατηγοριοποιεί δεδομένα θα ήθελε να υπάρχει ένας τρόπος να κάνει αξιολόγηση των αποτελεσμάτων. Να ελέγξει δηλαδή ποιο ποσοστό των δεδομένων κατηγοριοποιήθηκε σωστά. Για τον λόγο αυτό έχουν δημιουργηθεί μετρικές οι οποίες ορίζουν ένα κοινό πλαίσιο σύμφωνα με το οποίο μπορούν ερευνητές να συγκρίνουν αποτελέσματα κατηγοριοποίησης. Οι μετρικές προϋποθέτουν πως έχει κανείς δεδομένα για τα οποία γνωρίζει ποια θα έπρεπε να είναι τα αποτελέσματα.

Οι μετρικές είναι αρκετά διαισθητικές και βασίζονται σε όλα τα πιθανά αποτελέσματα που μπορούν να βγούν από την διαδικασία. Κατά την δυαδική κατηγοριοποίηση ενός στιγμιότυπου *x* στις κατηγορίες *true* και *false* μπορούμε να διακρίνουμε τις εξής περιπτώσεις:

- Αληθές θετικό δείγμα - True positive

  Το *x* κατηγοριοποιήθηκε ως *true* και ήταν όντως *true*

- Αληθές αρνητικό δείγμα - True negative

  Το *x* κατηγοριοποιήθηκε ως *false* και ήταν όντως *false*

- Ψευδές θετικό δείγμα - False positive

  Το *x* κατηγοριοποιήθηκε ως *true* αλλά ήταν *false*

- Ψευδές αρνητικό δείγμα - False negative

  Το *x* κατηγοριοποιήθηκε ως *false* αλλά ήταν *true*

Όπως βλέπουμε υπάρχουν δύο τρόποι να πάρει κανείς λανθασμένη απόφαση και δύο τρόποι να πάρει κανείς σωστή απόφαση. Με βάση τα παραπάνω ορίζονται οι παρακάτω μετρικές αξιολόγησης.

Ακρίβεια [Precision] - εκφράζει το ποσοστό των αληθών στιγμιότυπων που κατηγοριοποιήθηκαν σωστά. Υψηλή βαθμολογία Ακρίβειας δείχνει πως ο κατηγοριοποιητής επιλέγει σωστά για όσα δείγματα επιλέγει να κατηγοριοποιήσει ως αληθή.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{2.1}$$

Ανάκληση [Recall] - εκφράζει το ποσοστό των αληθών στιγμιότυπων που εντοπίστηκαν σε σχέση με το σύνολο των αληθών στιγμιοτύπων που υπήρχαν στο σύνολο δεδομένων. Υψηλή βαθμολογία Ανάκλησης δείχνει πως από τα θετικά δείγματα που υπάρχουν, ο κατηγοριοποιητής

ανιχνεύει το μεγαλύτερο μέρος τους.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{2.2}$$

Ευστοχία [Accuracy] - εκφράζει το ποσοστό των σωστών αποφάσεων ως προς το σύνολο των αποφάσεων που πάρθηκαν.

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives}} \tag{2.3}$$

Γενικά η Ακρίβεια και η Ανάκληση είναι μετρικές που τείνουν να είναι αντιστρόφως ανάλογες. Αυτό γίνεται καθώς όταν κανείς προσπαθήσει να εντοπίσει περισσότερα θετικά στιγμιότυπα τείνει να είναι λιγότερο σίγουρος για το αν είναι θετικά ή όχι.

Ανάλογα με την εφαρμογή μπορεί να είναι πιο σημαντική σαν μετρική η Ακρίβεια ή η Ανάκληση. Η Ευστοχία δεν χρησιμοποιείται συνήθως σαν μετρική χωρίς τις άλλες δύο, καθώς μόνη της μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα ως προς τις αποδόσεις της κατηγοριοποίησης. Αυτό οφείλεται στο ότι συναθροίζουμε δύο μεταβλητές σε μια και συνεπώς χάνουμε την πληροφορία που αφορά το πόσο εύστοχη ήταν η ανίχνευση θετικών δειγμάτων σε σχέση με την ανίχνευση αρνητικών δειγμάτων.

Στην συνέχεια της βιβλιογραφίας όπου αναφερόμαστε σε Ακρίβεια, Ανάκληση και Ευστοχία θα εννοούμε τις παραπάνω μετρικές και θα τις υπολογίζουμε με βάση τους παραπάνω τύπους.

2.2.3 Διαδικασίες επεξεργασίας φυσικής γλώσσας

Σε αυτήν την υποενότητα θα δούμε ορισμένες διαδικασίες επεξεργασίας φυσικής γλώσσας που θα χρησιμοποιήσουμε παρακάτω.

Ο κατακερματισμός και ο λημματισμός αφορούν διαδικασίες επεξεργασίας που έχουν απώτερο σκοπό τον μετασχηματισμό κειμένου σε μορφές βολικές για επεξεργασία στον υπολογιστή.

Η συντακτική ανάλυση και ο χωρισμός προτάσεων σε συντακτικούς πλίνθους, αποτελούν διαδικασίες ανάλυσης, που επιτρέπουν στην συνέχεια να γίνει επεξεργασία της δομής και των σχέσεων που εκφράζει το κείμενο. προηγούμενες.

Κατακερματισμός

Ο κατακερματισμός [tokenization] αφορά τον χωρισμό του κειμένου σε λέξεις. Η διαδικασία αυτή είναι λίγο πολύ τετριμμένη. Συνήθως γίνεται με συγκεκριμένους κανόνες, ανάλογα με το τελικό αποτέλεσμα που επιθυμεί κανείς να πετύχει. Για παράδειγμα, μια απλή προσέγγιση είναι ο χωρισμός τον προτάσεων σε κάθε κενό. Ωστόσο, με αυτήν την προσέγγιση αρνήσεις όπως η

can't (δεν μπορώ), δεν θα είναι εύκολο να διαχωριστεί σε *can* και *not*. Έτσι μερικές προσεγγίσεις χρησιμοποιούν λίγο πιο σύνθετους κανόνες για τον κατακερματισμό για να διαχειρίζονται κατάλληλα και τα σημεία στίξης.

Λημματισμός

Ο λημματισμός [lemmatization], προσπαθεί να λύσει το πρόβλημα της ανάγκης για ενιαία αναπαράσταση των λέξεων που θέλουμε να συγκρίνουμε σε έναν υπολογιστή. Αν θεωρήσουμε, για παράδειγμα, πως θέλουμε να συγκρίνουμε το *τρέχω* με το *τρέχει*, παρατηρούμε πως για τον υπολογιστή οι δύο αυτές λέξεις είναι διαφορετικές. Αυτό ισχύει καθώς διαφέρουν στις καταλήξεις. Ωστόσο, για εμάς, είναι προφανές πως πρόκειται για το ίδιο ρήμα. Σε μερικές περιπτώσεις που δεν μας ενδιαφέρει η πτώση των ρημάτων, ή ο διαχωρισμός ανάμεσα σε πληθυντικό και ενικό, ο λημματισμός μπορεί να είναι μια καλή λύση. Λημματισμός μπορεί να γίνει με δύο προσεγγίσεις.

Η πρώτη είναι η δημιουργία κανόνων για την αποκοπή καταλήξεων [stemming]. Ένας από τους πιο γνωστούς αυτής της προσέγγισης είναι ο λημματιστής porter [M.F.Porter, 1980]. Με χρήση ενός τέτοιου λημματιστή, αποτέλεσμα για το προηγούμενο παράδειγμα θα ήταν να μεταμορφωθούν και οι δύο λέξεις στο κοινό λήμμα: *τρεχ*. Το πρόβλημα με αυτήν την προσέγγιση, είναι πως τα λήμματα στα οποία καταλήγουν οι λέξεις, δεν είναι πλέον έγκυρες λέξεις της γλώσσας. Επίσης, δεν μπορεί να ξέρει κανείς αν βρεθεί με ένα λήμμα, από ποια αρχική μορφή της λέξης προήλθε. Τέλος, παρουσιάζονται προβλήματα σε ορισμένες περιπτώσεις, όπως για παράδειγμα με τις λέξεις university και universe οι οποίες μετασχηματίζονται και οι 2 στο κοινό lemma univers ενώ εκφράζουν διαφορετική έννοια.

Η δεύτερη αφορά την χρήση ενός συνόλου δεδομένων που περιέχει όλες τις μορφές κάποιας λέξης [lemmatization]. Για παράδειγμα, μπορεί κανείς να κάνει λημματισμό με χρήση του wordnet, καθώς περιέχει αντιστοιχίσεις για κάθε μορφή μιας λέξης στη βασική της. Πλεονέκτημα της είναι πως το λήμμα πλέον είναι υπάρχουσα λέξη της γλώσσας. Ωστόσο, παρουσιάζονται προβλήματα όταν δύο διαφορετικές λέξεις έχουν το ίδιο λήμμα, όπως για παράδειγμα, η λέξη born (γεννήθηκε) και η λέξη bears (αρκούδες), καθώς και οι δύο μετασχηματίζονται στο λήμμα bear.

2.2.4 Διαδικασίες ανάλυσης φυσικής γλώσσας

Αναγνώριση μερών του λόγου

Η αναγνώριση μερών του λόγου [part of speach tagging], είναι η διαδικασία κατά την οποία οι λέξεις σε ένα κείμενο επισημειώνονται με μια ετικέτα που δηλώνει το μέρος του λόγου στο οποίο αντιστοιχούν στη συγκεκριμένη πρόταση. Η επισημείωση αυτή καθ' αυτή δεν έχει

πρακτική εφαρμογή. Όμως τα αποτελέσματα της είναι αναγκαία στην περίπτωση που πρέπει να γίνει χωρισμός της πρότασης σε συντακτικούς πλίνθους [chunking], ή συντακτική ανάλυση της πρότασης. Να κατανοηθεί, για παράδειγμα, ποιο είναι το υποκείμενο και ποιο το αντικείμενο ενός ρήματος.

Η δυσκολία στην αναγνώριση μερών του λόγου υφίσταται στην αμφισημία που παρουσιάζουν μερικές λέξεις ως προς το μέρος του λόγου. Για παράδειγμα η λέξη likes (του αρέσει) μπορεί να είναι ρήμα, μπορεί όμως να είναι και ουσιαστικό likes (γούστα). Αν και η πλειονότητα των λέξεων στα Αγγλικά δεν έχει αμφισημίες, η σωστή επισημείωση αποτελεί καθοριστικό παράγοντα στην επίδοση της συντακτικής ανάλυσης που πρόκειται να την χρησιμοποιήσει.

Η αναγνώριση μερών του λόγου γίνεται κυρίως με δύο τρόπους, στατιστικά με την εφαρμογή στατιστικών μοντέλων, και προγραμματιστικά με δυναμικό προγραμματισμό.

Η προσέγγιση με στατιστικά μοντέλα συνήθως περιλαμβάνει μοντέλα Markov [Charniak, 1997a]. Η λογική είναι πως η ετικέτα κάθε λέξης εξαρτάται από το περιεχόμενο στο οποίο θα βρεθεί κάθε λέξη. Για παράδειγμα, δεν υπάρχει περίπτωση ένα ρήμα να ακολουθεί άρθρο. Βλέπουμε κατά αυτόν τον τρόπο πως μπορούν να επιλυθούν ορισμένες αμφισημίες πιο έξυπνα. Με αυτόν τον σκοπό, εισάγωνται αλυσίδες δεσμευμένων πιθανοτήτων, οι οποίες υπολογίζονται από ένα μεγάλο σύνολο δεδομένων. Έπειτα χρησιμοποιούνται για να υπολογίσουν την επισημείωση πρότασης που συνολικά έχει την μεγαλύτερη πιθανότητα.

Η προσέγγιση με δυναμικό προγραμματισμό [DeRose, 1988] ανάγει το πρόβλημα σε εύρεση βέλτιστου μονοπατιού σε γράφο. Οι λέξεις που δεν παρουσιάζουν αμφισημία συμπληρώνονται ως κόμβοι σε ένα μονοπάτι. Στην συνέχεια για τις αμφισημίες εισάγονται όλοι οι πιθανοί κόμβοι με βάρος την πιθανότητα εμφάνισης της ετικέτας δοθέντος της προηγούμενης. Το συνολικό βέλτιστο μονοπάτι σχηματίζεται από την επέκταση κατά ένα κόμβο κάθε φορά, του βέλτιστου μονοπατιού του προηγούμενου βήματος.

Το ποσοστό ευστοχίας με το οποία επισημειώνονται σωστά λέξεις είναι υψηλό, γύρω στο 97%. Η εικόνα αυτή είναι σε μεγάλο βαθμό παραπλανητική όμως, καθώς το ποσοστό βγαίνει από την επισημείωση σε λέξεις και σημεία στίξης και όχι σε επίπεδο πρότασης. Όπως αναφέραμε πρίν, τα σημεία στίξης και η πλειοψηφία των λέξεων δεν έχουν αμφισημίες ως προς την επισημείωση. Για να καταλάβουμε τον βαθμό της παραπλάνησης του ποσοστού, αν για κάθε λέξη βρούμε την πιο πιθανή ετικέτα με βάση ένα μεγάλο σύνολο δεδομένων (300.000 λέξεων) και στην συνέχεια επισημειώσουμε με την ετικέτα αυτή κάθε λέξη για την οποία έχουμε στοιχεία, και με ετικέτα ουσιαστικού τις υπόλοιπες, μπορούμε να επιτύχουμε ποσοστό επιτυχίας 90%. Περισσότερη συζήτηση για τα ποσοστά ευστοχίας και τα περιθώρια βελτίωσης γίνονται στο άρθρο [Manning,

2011]

Οι επισημειώσεις λέξεων που γίνονται στο πλαίσιο της πτυχιακής είναι σύμφωνες με το πρότυπο που καθιερώθηκε στο σύνολο επισημειωμένων δεδομένων που δημιουργήθηκε από το Πανεπιστήμιο της Pennsylvania [Marcus et al., 1993] οι οποίες φαίνονται στο σχήμα 2.1. Τα δεδομένα αυτά αποτελούνταν από κείμενα τα οποία επισημειώθηκαν από ανθρώπινους επισημειωτές ως προς το μέρος του λόγου στο οποίο αντιστοιχεί κάθε λέξη. Η επισημείωση έγινε με βάση τις παρακάτω ειδικότερες κατηγορίες.

Ρηχή συντακτική ανάλυση

Στόχος της ρηχής συντακτικής ανάλυσης [shallow parsing] ή αλλιώς χωρισμός σε συντακτικούς πλίνθους [chunking], είναι η ανάλυση της φράσης σε μεγαλύτερες νοηματικές ενότητες. Προαπαιτούμενο της είναι οι λέξεις της πρότασεις να έχουν επισημειωθεί με τον μέρος του λόγου στο οποίο ανήκουν, όπως για παράδειγμα με επισημειώσεις του σχήματος 2.1.

Συνηθισμένες προσεγγίσεις είναι η στατιστική ανάθεση με βάση ένα μεγάλο σύνολο δεδομένων που είναι επισημειωμένο με το χέρι ή η σύνθεση τους με βάση κανόνες συνδυασμών λέξεων με βάση την συντακτική τους ανάλυση.

Η στατιστική ανάθεση γίνεται συνήθως μέσω μηχανικής μάθησης με κατηγοριοποιητές [classifiers] μέγιστης εντροπίας [maxent models] [Charniak, 1999] ή διανύσματα υποστήριξης μηχανής [support vector machines] [Kudo and Matsumoto, 2001]. Ως βάση χρησιμοποιείται ένα μεγάλο σύνολο δεδομένων όπου υπάρχει επισημειωμένη, πέρα από το μέρος του λόγου, ο πλίνθος στον οποίο ανήκει κάθε λέξη. Τα στατιστικά εργαλεία "μαθαίνουν" τα χαρακτηριστικά που τείνει να έχει κάθε πλίνθος και επισημειώνουν την πιο πιθανή κατηγορία.

Η δημιουργία με βάση κανόνες μπορεί να γίνει με χρήση κανονικών εκφράσεων [regular expressions]. Σε περίπτωση που οι κανόνες γίνεται να εκφραστούν με κανονικές εκφράσεις, γίνεται αναζήτηση στις συντακτικές ετικέτες των λέξεων για συγκεκριμένα μοτίβα. Στην συνέχεια επισημειώνονται οι κατηγορίες ανάλογα με τους κανόνες που αντιστοιχεί το κάθε μοτίβο.

Οι ετικέτες με τις οποίες επισημειώνονται οι συντακτικοί πλίνθοι στο σύνολο δεδομένων του πανεπιστημίου της Pennsylvania φαίνεται στον Πίνακα 2.1. Στον Πίνακα εξηγείται τι φανερώνει η ανάθεση κάθε είδους ετικέτας σε συντακτικό πλίνθο. Στην δεξιά τελευταία στήλη του πίνακα φαίνεται το ποσοστό εμφάνισης της κάθε ετικέτας στο σύνολο δεδομένων Penn treebank.

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition or subordinating conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PRP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non-3rd person singular present
32. VBZ Verb, 3rd person singular present
33. WDT Wh-determiner
34. WP Wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB Wh-adverb

Σχήμα 2.1: Τύποι επισημειώσεων για τα μέρη του λόγου σύμφωνα με το σύνολο δεδομένων Penn treebank

Ετικέτα	Chunk	Πλίνθος	% Εμφάνισης
NP	Noun Phrase	Ονοματική φράση	51
PP	Prepositional Phrase	Εμπρόθετη φράση	19
VP	Verb Phrase	Ρηματική φράση	9
ADVP	Adverb Phrase	Επιρρηματική φράση	6
ADJP	Adjective Phrase	Επιθετική φράση	3
SBAR	Subordinating Conjunction	Σύνδεσμος υποταγής	3
PRT	Particle	Μόριο	1
INTJ	Interjection	Επιφώνημα	0

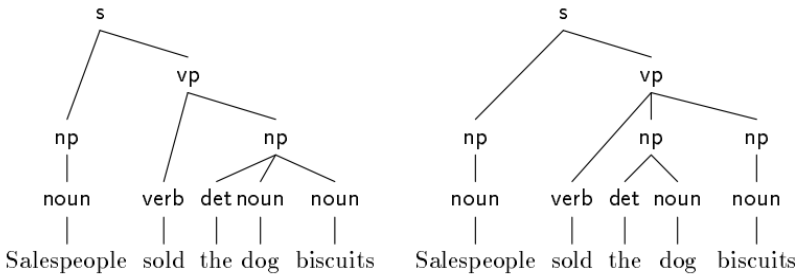
Πίνακας 2.1: Επεξήγηση ετικετών για τις επισημειώσεις συντακτικών πλίνθων [chunks].

Συντακτική ανάλυση

Σκοπός της συντακτικής ανάλυσης είναι η κατανόηση της δομής και των σχέσεων που υπάρχουν σε ένα κείμενο. Μια από τις δυσκολίες της δημιουργίας της συντακτικής ανάλυσης των φυσικών γλωσσών, σε αντίθεση με αυτή μιας γλώσσας προγραμματισμού, είναι η αμφισημία που εισάγουν ορισμένες λέξεις καθώς και οι ιδιότητες των γραμματικών της. Λέξεις οι οποίες εμφανίζονται άλλοτε ως ρήματα και άλλοτε ως ουσιαστικά, για παράδειγμα, δυσκολεύουν την ανάλυση. Αυτό γίνεται καθώς ο επισημειωτής πρέπει να αποφασίσει κάθε φορά τι είναι πιο λογικό να επισημειώσει. ο χρόνος συντακτικής ανάλυσης είναι εκθετικός.

Τα πιθανά συντακτικά δέντρα που μπορούν να αντιστοιχιστούν σε μια πρόταση αυξάνονται απότομα για κάθε αμφισημία που εισάγεται. Για έναν άνθρωπο τα πολλά πιθανά συνακτικά δέντρα μπορεί να σημαίνουν πως υπάρχουν πολλοί τρόποι να διαβάσεις την συγκεκριμένη πρόταση. Ωστόσο, η γενίκευση αυτής της εικόνας για τις αμφισημίες είναι κάπως λανθασμένη. Χάρη στην φύση των κανόνων που χρησιμοποιούνται για την κατασκευή συντακτικών δομών, πολλές πιθανές επιλογές δεν βγάζουν καν νόημα. Θεωρώ πως είναι άσχημο κανείς να μην φτιάχνει τα δικά του παραδείγματα για να εκφράσει τις ιδέες του, όμως το παράδειγμα του Charniak [Charniak, 1997a] ήταν τόσο εύστοχο και είχε τόσο μεστή μετάφραση και στα Ελληνικά, που επέλεξα να το παραθέσω εδώ. Στο σχήμα 2.2 φαίνονται δύο πιθανά συντακτικά δέντρα για την φράση : "Πωλητές πούλησαν τα μπισκότα του σκύλου". Παρατηρούμε εδώ πως μπορεί να πούλησαν τα μπισκότα που ανήκαν στον σκύλο, ή να πούλησαν τα μπισκότα στον σκύλο.

Οι δύο κύριοι τρόποι με τον οποίο γίνεται συντακτική ανάλυση είναι με κατασκευή συντακτικών δέντρων [parse trees] [Charniak, 1997b] και με κατασκευή εξαρτήσεων [dependencies]



Σχήμα 2.2: Παράδειγμα αμφισημίας στην κατασκευή συντακτικού δέντρου [Charniak, 1997a]

[de Marneffe et al., 2006].

2.3 Σχετική Βιβλιογραφία

Στην υποενότητα αυτή θα δούμε βιβλιογραφία που έχει άμεση σχέση με την ευθυγράμμιση προτάσεων και την χρήση συνωνύμων για την εύρεση ομοιοτήτων. Αρχικά θα δούμε την εύρεση συνεπαγωγής σε κείμενο και πως η εύρεση αντιθέσεων σε κείμενο προέκυψε ως υποπρόβλημά της.

2.3.1 Εύρεση λογικής συνεπαγωγής σε κείμενο

Η εύρεση λογικής συνεπαγωγής σε κείμενο είναι ένα ευρύ πρόβλημα του οποίου η λύση βρίσκει εφαρμογές στους τομείς της εξόρυξης πληροφορίας [information extraction] , της απάντησης ερωτήσεων [question answering], της περίληψης κειμένων [text summarization], της μηχανικής μετάφρασης [machine translation] και στην ανάπτυξη φυσικού κειμένου [natural language generation] [Androutsopoulos and Malakasiotis, 2010]. Μια από της πρώτες συλλογικές προσπάθειες να γίνει εκτεταμένη έρευνα για την εύρεση λογικής συνεπαγωγής σε κείμενο έγινε με την εισαγωγή της πρόκλησης RTE, η οποία δημιουργήθηκε με βάση την δημοσίευση [Dagan and Glickman, 2004].

Πρόκληση RTE

Η πρόκληση (challenge) RTE (recognizing textual entailment) [αναγνώριση συνεπαγωγής σε κείμενο] αφορούσε μια ετήσια ανάθεση προβλήματος στον χώρο της επεξεργασίας φυσικής γλώσσας με σκοπό την προώθηση της έρευνας στον τομέα αυτό. Στα πλαίσια αυτής της πρόκλησης έγινε και μια πρώτη έντονη προσπάθεια να αναλυθεί το πρόβλημα της εύρεσης αντιθέσεων σε κείμενο.

Αρχικά, το RTE διοργανώθηκε από την οργάνωση Pascal (Pattern Analysis, Statistical Modeling and Computational Learning) [Ανάλυση Μοτίβων, Στατιστική Μοντελοποίηση και Μάθηση με

Υπολογιστή]. Έπειτα συνεχίστηκε στα πλαίσια του TAC (Text Analysis Conference) [Συνέδριο Ανάλυσης Κειμένου].

Για κάθε πρόκληση (RTE 1 - RTE 7 , 2004 - 2011) αν και μερικές παράμετροι της πρόκλησης άλλαζαν, τα κύρια χαρακτηριστικά ήταν ίδια. Δημοσιευόταν σε κάθε πρόκληση ένα σύνολο δεδομένων από ζεύγη κειμένων. Κάθε ζεύγος αποτελούνταν από ένα μικρό απόσπασμα κειμένου 1-2 προτάσεων (text) και ένα συντομότερο απόσπασμα μιας πρότασης, την υπόθεση (hypothesis).

Το σύνολο δεδομένων του RTE 2 αποτελούνταν από 1.600 ζευγάρια και ήταν χωρισμένο μισό μισό σε σύνολο ανάπτυξης (dev) και σύνολο ελέγχου (test). Τα ζευγάρια του συνόλου ανάπτυξης ήταν επισημειωμένα με την πληροφορία ύπαρξης ή μη λογικής συνεπαγωγής από τα λεγόμενα του κειμένου στην υπόθεση. Σκοπός της πρόκλησης ήταν οι ενδιαφερόμενες ομάδες να προσπαθήσουν να διακρίνουν σε ποια ζεύγη του συνόλου ελέγχου υπάρχει λογική συνεπαγωγή από το κείμενο στην υπόθεση.

Λογική συνεπαγωγή από το κείμενο στην υπόθεση ορίζεται πως υπάρχει, όταν κάποιος άνθρωπος αφού διαβάσει το κείμενο θα συμπεράνει πως ισχύει η υπόθεση [Dagan and Glickman, 2005]. Δεδομένο θεωρείται πως το κείμενο περιέχει αληθή στοιχεία.

Ενώ η πρόκληση αυτή καθ' αυτή αφορούσε την κατηγοριοποίηση κάθε ζεύγους προτάσεων σε λογική συνεπαγωγή ή μη, στην τρίτη πρόκληση RTE υπήρχε πέρα από την κύρια πρόκληση, μια δευτερεύουσα.

Επεκτάθηκαν τα δεδομένα και έγινε περαιτέρω διαχωρισμός των περιπτώσεων για όσες ομάδες ήθελαν να ανιχνεύσουν αντιθέσεις [Voorhees, 2008]. Η περίπτωση της μη συνεπαγωγής αναλύθηκε σε ύπαρξη αντίφασης και σε έλλειψη γνώσης για λήψη απόφασης. Επίσης, δημιουργήθηκε ένα κοινό πλαίσιο για την κατηγοριοποίηση των αντιθέσεων. Ορίστηκε πως ένα κείμενο και μια υπόθεση παρουσιάζουν αντίθεση όταν δεν είναι πιθανό τα δεδομένα του κειμένου και της υπόθεσης να ισχύουν ταυτόχρονα. Έτσι πλέον οι ομάδες, εφόσον το ήθελαν, καλούνταν να ταξινομήσουν τα ζευγάρια σε 3 κατηγορίες.

- entailment - συνεπαγωγή
- contradiction - αντίθεση
- don't know - δεν ξέρω

Έτσι έγινε μια προώθηση της έρευνας προς την αναγνώριση αντιθέσεων σε κείμενο, εργασία της οποίας θετικά αποτελέσματα θα έδειχναν πως τα συστήματα καταφέρνουν να σχηματίσουν μια βασική κατανόηση - αναπαράσταση του νοήματος του κειμένου.

2.3.2 Εύρεση αντιθέσεων σε κείμενο

Τα παρακάτω συστήματα εξειδίκευσαν την ανάλυση τους στην ανίχνευση αντιθέσεων σε κείμενο. Συνεπώς τα αποτελέσματα που αναφέρονται αφορούν δυαδική κατηγοριοποίηση - ύπαρξη ή μη αντίθεσης στο κείμενο.

Κατόπιν ανάλυσης, η ομάδα De Marneffe et al προσδιόρισαν τις συνθήκες υπό τις οποίες εμφανίζονται συγκρούσεις σε κείμενα. Σε κείμενα που περιέχουν αντιθέσεις παρουσιάζεται:

1. Χρήση Αντωνύμων
2. Αρνήσεις
3. Αριθμητικές ασυνέπειες - διαφορές
4. Ασυνέπειες σχετικές με γεγονότα
5. Δομικές διαφορές στις προτάσεις
6. Λεξικές διαφορές
7. Ασυνέπειες που χρειάζονται γενική γνώση για τον κόσμο [world knowledge] για την ανίχνευσή τους

Αφού ανέλυσε τα παραπάνω χαρακτηριστικά, η ομάδα De Marneffe προχώρησε στο σχηματισμό ενός συστήματος ανίχνευσης αντιθέσεων. Η διαδικασία που ακολούθησε για την αντιμετώπιση του προβλήματος βασίστηκε σε μεγάλο βαθμό στην ευθυγράμμιση προτάσεων.

Αρχικά έγινε προεπεξεργασία κάθε ζεύγους κειμένου - υπόθεσης κατά την οποία έγινε μετατροπή τους σε γράφο εξαρτήσεων [typed dependency graph] [de Marneffe et al., 2006]. Κατά την διαδικασία αυτή έγινε σύμπτυξη κοινών φράσεων και ονομάτων σε μια λέξη και κάθε λέξη πλέον αποτελούσε κόμβο στον γράφο εξαρτήσεων.

Στην συνέχεια έγινε ευθυγράμμιση του γράφου της υπόθεσης με τον γράφο του κειμένου, κατά την οποία κάθε κόμβος της υπόθεσης αντιστοιχιζόταν σε έναν κόμβο του κειμένου ή στο κενό [null]. Η αντιστοίχιση έγινε με βάση την ομοιότητα των κόμβων καθώς και δομική πληροφορία που βασιζόταν στον γράφο εξαρτήσεων. Στην συνέχεια με χρήση μηχανικής μάθησης ανατέθηκαν βάρη στην συνεισφορά της ομοιότητας κόμβων και της δομής και έγινε εκπαίδευση στα δείγματα του RTE 3 ανάπτυξης [dev] [de Marneffe et al., 2007].

Το επόμενο βήμα αφορούσε την διήθιση ζευγαριών που δεν αναφέρονταν στο ίδιο γεγονός. Αυτό έγινε με απαίτηση ευθυγράμμισης της ρίζας του γράφου εξαρτήσεων της υπόθεσης στο κείμενο και την περαιτέρω αναζήτηση των κομματιών της υπόθεσης που είναι απαραίτητο να ευθυγραμιστούν με το κείμενο.

Στο τελευταίο βήμα έγινε εξαγωγή των χαρακτηριστικών για την ανίχνευση αντιθέσεων, τα οποία περιγράφηκαν παραπάνω, από τα ζεύγη κειμένου - υπόθεσης. Έδωσαν βάρη τα οποία θεώρησαν πως ήταν λογικά στα χαρακτηριστικά και εφάρμοσαν γραμμική παλινδρόμηση για

την κατηγοριοποίηση κάθε ζεύγους ως αντιφατικό ή μη.

Παρά την σε μεγαλύτερο βάθος ανάλυση τους όμως, τα αποτελέσματα τους δεν ήταν ιδιαίτερα καλά, καθώς πέρα από την άρνηση και την εύρεση αντωνύμων με την οποία ασχολήθηκαν άλλες ομάδες, οι υπόλοιπες κατηγορίες έχουν χαμηλά ποσοστά έγκυρης ανίχνευσης.

Στο σύνολο δεδομένων RTE 3 η ομάδα de Marneffe et al σημείωσε ποσοστό ανάκλησης [recall] 19.44% και ακρίβειας [precision] 22.95% στην ανίχνευση αντιθέσεων που ανήκαν σε όλες τις κατηγορίες που αναλύθηκαν παραπάνω [de Marneffe et al., 2008]. Όπως είδαμε στην υποενότητα 2.2.2 το παραπάνω σημαίνει πως έπιασαν κατά μέσο όρο 19.44% των αντιθέσεων που υπήρχαν συνολικά και σε ποσοστό 22.95% τα ζευγάρια που επισημείωναν πως εμφάνιζαν αντίθεση παρουσίαζαν όντως αντίθεση.

Στο ίδιο σύνολο δεδομένων αναφέρονται για την ανίχνευση αντιθέσεων μέσες βαθμολογίες ανάκλησης [recall] και ακρίβειας [precision] όλων των ομάδων που συμμετείχαν βαθμολογίες 11.69% και 10.72% αντίστοιχα.

Αξίζει να σημειωθεί πως στο RTE 3 σετ δεδομένων υπήρχε πολύ χαμηλό ποσοστό αντιθέσεων, περίπου 10% [Voorhees, 2008] και συνεπώς το σύνολο δεδομένων δεν ήταν ισορροπημένο.

Η ομάδα [Harabagiu et al., 2006] συμμετείχε στην πρόκληση RTE 2. Η δική τους προσέγγιση στο πρόβλημα ήταν να ψάξουν για συγκρούσεις μεταξύ κειμένου και υπόθεσης με το να μετέτρεψουν σε άρνηση μια από τις 2 προτάσεις και να ψάξουν για λογική συνεπαγωγή. Συνεπώς ανέλυσαν το πρόβλημα της εύρεσης αντιθέσεων για τις περιπτώσεις 1 και 2 που αναλύσαμε παραπάνω. Σε ένα σύνολο δεδομένων που δημιούργησαν οι ίδιοι από το RTE 2 σύνολο δεδομένων αφού μετέτρεψαν οι ίδιοι τις προτάσεις σε αρνήσεις αναφέρουν πως σημείωσαν 75.63% ευστοχία [accuracy]. Ωστόσο, το σύνολο δεδομένων αυτό ήταν ισορροπημένο (ίδιο ποσοστό αντιθέσεων και μη αντιθέσεων). Συγκριτικά, η ομάδα de Marneffe σημείωσε 63% ποσοστό ευστοχίας [accuracy] σε μια εξομοίωση του συνόλου δεδομένων της ομάδας Harabagiu et al.

Τέλος η ομάδα [Ritter et al., 2008] ασχολήθηκε με μια πολύ συγκεκριμένη κατηγορία φράσεων για να βρεί αντιθέσεις. Συγκεκριμένα βρήκε γραμματικές σχέσεις οι οποίες έχουν το χαρακτηριστικό να συμπεριφέρονται ως συναρτήσεις ένα προς ένα. Όταν έχουν ίδιο υποκείμενο θα έχουν αναγκαστικά και ίδιο αντικείμενο. Ως παράδειγμα αναφέρεται το ρήμα γεννιέμαι, καθώς αν γνωρίζουμε πως κάποιος γεννήθηκε κάπου, δεν γίνεται ο ίδιος άνθρωπος να έχει γεννηθεί και κάπου αλλού. Αυτές οι σχέσεις αναφέρονται ως συναρτησιακές σχέσεις (functional relations), ωστόσο δεν υπήρχαν αρκετές σε κάποιο από τα RTE datasets για να δοκιμάστει το σύστημα εκεί. Επίσης η δουλειά αυτή δίνει έμφαση στο πόσο σπάνιο είναι το φαινόμενο της αντίθεσης στα πραγματικά δεδομένα, καθώς παρόλο που το σύστημα αρχικά διαλέγει πολλές πιθανές περιπτώσεις, καταλήγει

πως οι περισσότερες δεν περιέχουν αντιθέσεις. Τα αποτελέσματα αυτής της προσέγγισης είναι ακρίβεια (precision) 62% με ανάκληση (recall) 19% στο σύνολο δεδομένων που δημιουργήθηκε από το διαδίκτυο και 92% ακρίβεια (precision) με 51% ανάκληση (recall) στο ισορροπημένο σύνολο δεδομένων.

Από τα παραπάνω είναι εμφανές πως η ανίχνευση αντιθέσεων σε κείμενο είναι δύσκολο πρόβλημα, καθώς παρόλο που περιορίστηκε σε ζεύγη προτάσεων με συγκεκριμένη μορφή, όταν αντιμετωπίστηκε στην γενική μορφή υπήρξαν χαμηλά ποσοστά ευστοχίας. Ωστόσο, η ανίχνευση αντιθέσεων που οφείλεται στην χρήση αντωνύμων μπορεί κανείς να πει πως είναι υλοποιήσιμη [de Marneffe et al., 2008] και θα πρέπει να δει κανείς αν είναι δυνατόν να βγεί κάποιο συμπέρασμα ως προς το που παρουσιάζεται η αντίθεση και τι σημαίνει αυτή.

ΚΕΦΑΛΑΙΟ 3

Υλοποίηση

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δούμε το σύστημα που κατασκευάστηκε, τις παραμέτρους από τις οποίες εξαρτάται, την προσέγγιση που ακολουθήθηκε κατά την υλοποίηση και τις ιδέες που οδήγησαν σε αυτή.

Αναλυτικά:

- Στην ενότητα 3.2 θα ρίξουμε μια αφαιρετική ματιά στην αρχιτεκτονική του συστήματος καθώς και τα κομμάτια που το απαρτίζουν από άποψη λειτουργικότητας και δομής.
- Στην ενότητα 3.3 θα περιγράψουμε τις πηγές πληροφορίας που χρησιμοποιήθηκαν, τα δεδομένα τους, την μορφή και το περιεχόμενό τους. Επίσης θα δούμε την μορφή που έχουν τα δεδομένα εξόδου του συστήματος.
- Στην 3.4 θα αναφερθούμε συνοπτικά στην προεπεξεργασία των δεδομένων και στην διαμόρφωση τους σε μια κοινή μορφή για το σύστημα, καθώς και στην επεξεργασία των αποτελεσμάτων.
- Στην ενότητα 3.5 θα εξηγηθεί η ιδέα πάνω στην οποία βασίζεται η λειτουργικότητα του πυρήνα του συστήματος και η ευελιξία της συγκεκριμένης προσέγγισης.

3.2 Σύστημα

Το σύστημα που υλοποιήθηκε στα πλαίσια της πτυχιακής έχει 2 βασικούς στόχους. Πρώτον, την χαλάρωση της έννοιας της λεξικής ομοιότητας από το αυστηρό ταίριασμα γραμμάτων σε μια ομοιότητα σχέσεων σε γράφο που απεικονίζει σχέσεις λέξεων (συνώνυμα και ορισμοί). Δεύτερον, την επέκταση της ομοιότητας από λεξικό επίπεδο σε επίπεδο φράσεων.

Το σύστημα μπορεί να χωριστεί σε 3 νοητά διακριτά επίπεδα.

1. *Δεδομένα* - Η εξαγωγή των αποτελεσμάτων και η εισαγωγή δεδομένων στο σύστημα
2. *Η προεπεξεργασία δεδομένων* εισόδου και εξόδου
3. *Ο πυρήνας* - η δημιουργία του γράφου, των διανυσμάτων και ο υπολογισμός ομοιότητας ανάμεσα σε λέξεις

Στο πρώτο επίπεδο ανήκουν τα δεδομένα εισόδου και τα δεδομένα εξόδου. Δεδομένα εισόδου αποτελούν οι πηγές συνωνύμων και ορισμών, ενώ δεδομένα εξόδου είναι οι βαθμολογίες ομοιότητας λέξεων καθώς και οι βαθμολογίες ευθυγράμμισης φράσεων. Σε αυτό το επίπεδο γίνεται η συλλογή των απαραίτητων δεδομένων συνωνύμων και ορισμών τα οποία είναι απαραίτητα για την λειτουργία του συστήματος. Επίσης αποθηκεύονται τα αποτελέσματα των αξιολογήσεων και των ευθυγραμμίσεων και επεξεργάζονται για να έρθουν στην τελική τους μορφή για να μπορούν να παρουσιαστούν. Περισσότερα στην ενότητα 3.3.

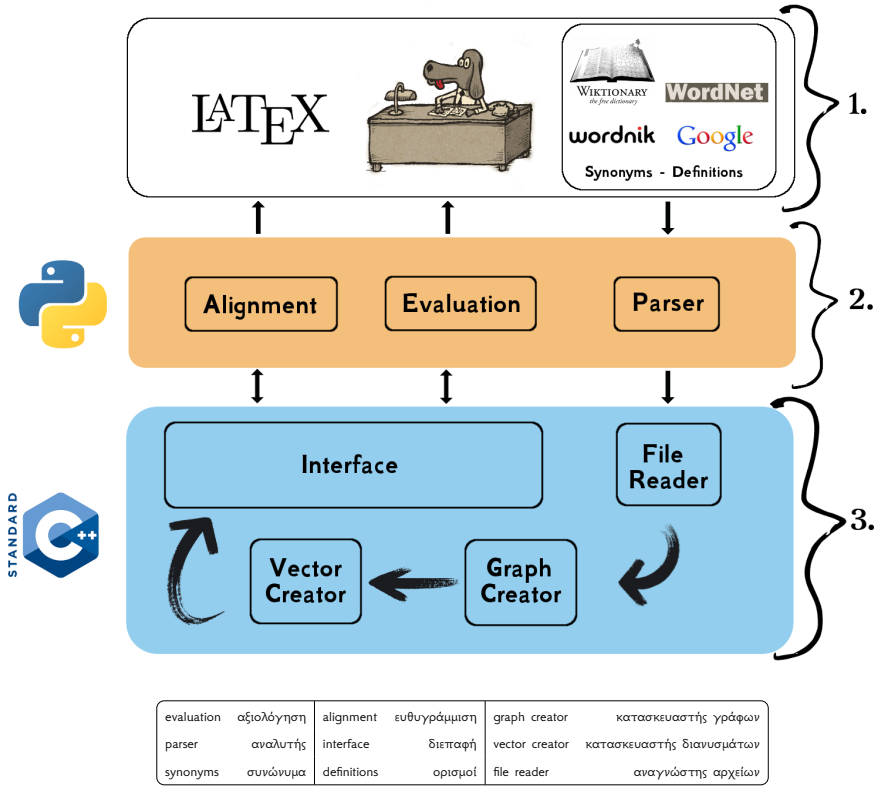
Στο δεύτερο γίνεται η μετατροπή των δεδομένων σε κοινή μορφή, συγκεκριμένα ένα αρχείο tsv (tab separated variables) για κάθε πηγή. Στην συνέχεια αυτά ενώνονται - πετώνται τα διπλότυπα και διατηρείται για κάθε λέξη μια συνεπής λίστα από συνώνυμα - σχετικές λέξεις. Επίσης σε αυτό το επίπεδο δημιουργείται η οπτικοποίηση της ευθυγράμμισης των προτάσεων καθώς και όποια προεπεξεργασία χρειάζονται τα δεδομένα των φράσεων που υπάρχουν στα RTE σύνολα δεδομένων. Αναλυτικότερα τα στάδια της προεπεξεργασίας στην ενότητα 3.4.

Το τρίτο επίπεδο δέχεται ως είσοδο αρχεία tsv τα οποία περιέχουν τις λέξεις και τις σχέσεις τους και χρησιμοποιούνται για την κατασκευή του γράφου. Μετά την κατασκευή του, ο γράφος μένει στην μνήμη και χρησιμοποιούνται τεχνικές για την κατασκευή διανυσμάτων (περισσότερα στην ενότητα 3.5.2). Αυτά στην συνέχεια συγκρίνονται και επιστρέφεται μια βαθμολογία ομοιότητας. Η επικοινωνία των αποτελεσμάτων με το 2ο επίπεδο γίνεται μέσω του λειτουργικού με σύζευξη [pipe]. Περισσότερα για τον πυρήνα του συστήματος στην ενότητα 3.5.

3.3 Δεδομένα

Όπως είδαμε στο σχήμα 3.1 τα δεδομένα που χρησιμοποιούνται ως βάση για να βρεθούν σχέσεις ανάμεσα σε λέξεις προέρχονται από πολλές πηγές. Στην υποενότητα αυτή θα δούμε λεπτομερώς τα δεδομένα, τις πηγές, τα χαρακτηριστικά τους, και την μορφή τους. Επίσης θα δούμε την μορφή των δεδομένων εξόδου και το περιεχόμενό τους.

Οι πηγές που δοκιμάστηκαν με χρονολογική σειρά είναι οι:



evaluation	αξιολόγηση	alignment	ευθυγράμμιση	graph creator	κατασκευαστής γράφων
parser	αναλυτής	interface	διεπαφή	vector creator	κατασκευαστής διανυσμάτων
synonyms	συνώνυμα	definitions	ορισμοί	file reader	αναγνώστης αρχείων

Σχήμα 3.1: Pipeline συστήματος

1. wiktionary
2. wordnet
3. google
4. wordnik

Οι λέξεις που αναζητήθηκαν ήταν στα Αγγλικά. Στις πρώτες δύο περιπτώσεις τα σύνολα των λέξεων εξήχθησαν από την πηγή, ενώ στις τελευταίες δύο χρησιμοποιήθηκαν οι ήδη υπάρχουσες λίστες από λέξεις για να βρεθούν οι σχέσεις που μας ενδιέφεραν. Μέχρι στιγμής έχουν δοκιμαστεί δύο μορφές δεδομένων στην υλοποίηση του συστήματος. Για λόγους απλότητας[1] δοκιμάστηκε η χρήση ορισμών και η χρήση συνωνύμων και αντωνύμων. Ωστόσο έχουν εξαχθεί και άλλες σχέσεις και το σύστημα είναι παραμετροποιήσιμο - δέχεται αρχεία σε μορφή tsv (tab separated variables).

Υπάρχουν αντίστοιχες δουλειές που αφορούν την συνάθροιση πολλαπλών συνόλων δεδομένων λέξεων σε ένα μεγαλύτερο σύνολο. Ένα τέτοιο παράδειγμα αποτελεί το wordnik το οποίο έχει συναθροίσει δεδομένα από άλλα λεξικά καθώς και από το βικιλεξικό (περισσότερα στην υποενότητα που αφορά το wordnik). Ωστόσο, δεν είναι ξεκάθαρο πως γίνεται η εσωτερική διαχείριση των λέξεων καθώς και σε ποιές σχέσεις λέξεων έχει κανείς πρόσβαση. Για παράδειγμα με χρήση

[1] Όπως θα δουμε παρακάτω, το σύστημα έχει πάρα πολλές παραμέτρους

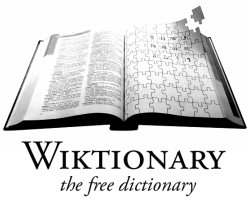
του wordnik είχαμε πρόσβαση σε συνώνυμα αντώνυμα και σχετικές λέξεις αλλά δεν ήταν ξεκάθαρο τι σχέσεις εξέφραζαν οι σχετικές λέξεις.

Το τελικό σύνολο δεδομένων συνωνύμων που χρησιμοποιήθηκε αποτελούνταν από τα συνώνυμα του wiktionary εμπλουτισμένα με συνώνυμα που βρέθηκαν με χρήση της υπηρεσίας της Google για λέξεις που υπήρχαν στο RTE σύνολο δεδομένων και ήταν πιθανή η ευθυγράμμιση τους. Το τελικό σύνολο δεδομένων ορισμών δημιουργήθηκε από το wordnet με αρχική εξαγωγή όλων των λέξεων. Στην συνέχεια για κάθε λέξη διατηρήθηκαν από τον ορισμό της ως σχετικές όσες λέξεις ήταν ρήματα, ουσιαστικά, επίθετα ή αριθμητικά (περισσότερα στην ενότητα 3.4.1).

Η συνάθροιση των πηγών όπως θα δούμε και παρακάτω δημιούργησε μια πλούσια πηγή συνωνύμων και ορισμών, καθώς περιείχε και τεχνικές λέξεις από το wordnet και λέξεις από την καθημερινότητα από το βικιλεξικό και το wordnik. Τα αποτελέσματα είναι ιδιαίτερα ικανοποιητικά αν σκεφτεί κανείς πως δεν χρησιμοποιήθηκε καμία εμπορική έκδοση λεξικού.

3.3.1 Πηγές

Βικιλεξικό



Το *Βικιλεξικό* είναι ένα ανοικτό, δωρεάν, συνεργατικό και πολυγλωσσικό διαδικτυακό λεξικό που πλέον έχει επεκταθεί και κάνει και χρέη θησαυρού. Είναι το αντίστοιχο της *Βικιπαίδειας* σε λεξικό - θησαυρό. Για κάθε λέξη που περιέχει παραθέτει ορισμούς, ετυμολογία, προφορά, παραδείγματα χρήσης της λέξης σε φράση, παρόμοιες εκφράσεις, συμπληρωματικές λέξεις, μεταφράσεις και σε μερικές περιπτώσεις συνώνυμα, αντώνυμα, υπόνυμα, υπέρνυμα και ολώνυμα. Βασικό χαρακτηριστικό της είναι πως συμπληρώνεται, συντηρείται και επεκτείνεται από ένα υποσύνολο της κοινωνίας των διαδικτυακών χρηστών που ενδιαφέρεται για αυτό και συμμετέχει. Ως εκ τούτου, οι εγγραφές δεν είναι εγγυημένο πως είναι έγκυρες ή πλήρεις. Ωστόσο υπάρχει ένα αξιόλογο ποσοστό των λέξεων που θα έβρισκε κανείς σε ένα λεξικό του εμπορίου καθώς και μερικές λέξεις που αποτελούν τοπικές ή διαδικτυακές εκφράσεις και πιθανώς να μην υπήρχαν σε άλλα λεξικά.

Τα δεδομένα του *Βικιλεξικού* στα Αγγλικά μπορεί να τα κατεβάσει κανείς από τον ιστό ως ένα συνονθύλευμα από ιστοσελίδες (html) από τον ιστότοπο της wiktionary. Ανα τακτά χρονικά διαστήματα αναρτάται ένα πιο πρόσφατο στιγμιότυπο στη συγκεκριμένη ιστοσελίδα. Ωστόσο, για να εξαχθούν οι λέξεις και οι σχέσεις που μας ενδιαφέρουν από τις σελίδες, απαιτείται χρήση

κανονικών εκφράσεων [*regular expressions*] καθώς οι σελίδες έχουν αυτοσχέδια μορφοποίηση κατα κύριο λόγο και δεν ακολουθούν κάποιο πρότυπο, δεν είναι για παράδειγμα μορφοποιημένα τα δεδομένα σε *xml* μορφή.

Θετικά:

- Ανοικτή - έχουμε πρόσβαση στα δεδομένα της
- Συνεχώς επεκτείνεται
- Περιέχει ορολογία που χρησιμοποιείται στο διαδίκτυο και στην καθημερινότητα
- Προσφέρει τις σελίδες συμπιεσμένες σε ένα αρχείο - δεν χρειάζεται έρπισμα [crawling] στις σελίδες

Αρνητικά:

- Δεν είναι πλήρης
- Τα συνώνυμα και τα αντώνυμα που περιέχει, καθώς και το πλήθος λέξεων για τα οποία τα έχει, είναι λίγα σε σχέση με ένα εμπορικό λεξικό
- Κακή μορφοποίηση δεδομένων - θέλουν αρκετό σκάλισμα
- Δεν εγγυάται κανείς την εγκυρότητα των δεδομένων, καθώς αυτή επαφίεται στην καλή θέληση της κοινότητας που το συντηρεί

Στατιστικά για τα Αγγλικά σύμφωνα με την προεπεξεργασία με κανονικές εκφράσεις (regular expressions).

- 388165 λέξεις με ορισμούς
- 23641 λέξεις με συνώνυμα
- 5125 λέξεις αντώνυμα
- 108 λέξεις με ολώνυμα
- 114 λέξεις με μερώνυμα
- 954 λέξεις με υπέρνυμα
- 1134 λέξεις με υπώνυμα
- 17300 λέξεις με συμπληρωματικές λέξεις

Το βικιλεξικό επιλέχθηκε ως η αρχική πηγή για τις λέξεις, τους ορισμούς και τα συνώνυμα - αντώνυμα καθώς ήταν από τις λίγες λεξικές πηγές που είναι προσβάσιμες ανοικτά. Οι επιπλοκές της επιλογής αυτής θα αναλυθούν καλύτερα στο κεφάλαιο 4.

Wordnet

Το *wordnet* είναι μια λεξικολογική βάση δεδομένων αγγλικών λέξεων που αναπτύχθηκε από το πανεπιστήμιο του Princeton [Miller, 1995] (περισσότερα είδαμε στην ενότητα 2). Για πρόσβαση

στο wordnet χρησιμοποιήθηκε η βιβλιοθήκη για επεξεργασία φυσικής γλώσσας *NLTK* [Bird et al., 2009] η οποία παρέχει διεπαφή για πρόσβαση στο wordnet.

```
In [1]: from nltk.corpus import wordnet

In [2]: synsets = {synset for synset in wordnet.synsets('parrot')}

In [3]: definitions = [entry.definition() for entry in synsets]

In [4]: for each in enumerate(definitions):
   ...:     print str(each[0]) + '. ' + each[1] + '\n'
   ...:
0. usually brightly colored zygodactyl tropical birds with short hooked beaks and the ability to mimic sounds

1. repeat mindlessly

2. a copycat who does not understand the words or acts being imitated
```

Σχήμα 3.2: Παράδειγμα εξαγωγής ορισμών από το wordnet με χρήση του NLTK

Θετικά:

- Ανοικτό με εύχρηστη διεπαφή
- Περιέχει πολλά συνώνυμα, αντώνυμα, υπερώνυμα, υπώνυμα, μερώνυμα, ολώνυμα και τροπώνυμα
- Χρησιμοποιείται ευρέως και συνεπώς θεωρείται έγκυρη πηγή - είναι αναγνωρισμένο

Αρνητικά:

- Η λογική της δενδροειδής δομής με βάση της παραπάνω σχέσεις δεν εκφράζει πάντα τον τρόπο που αναλύει την ομοιότητα λέξεων ένας άνθρωπος

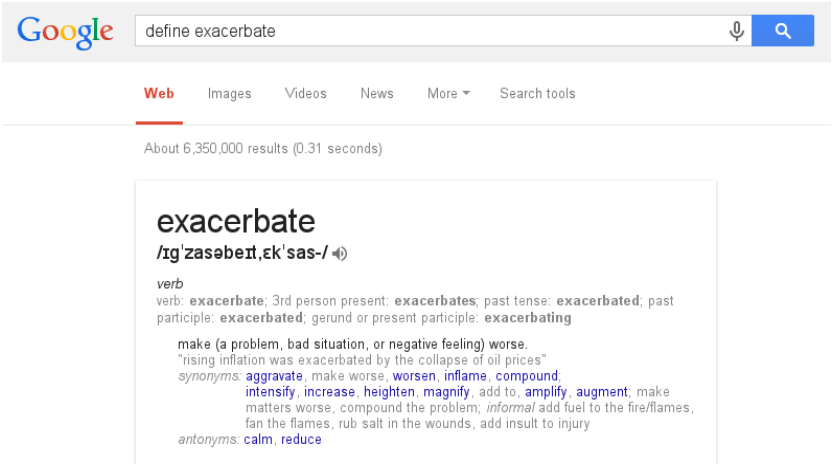
Στατιστικά:

- 155467 λέξεις με ορισμούς
- 124559 λέξεις με συνώνυμα

Google

Η Google παρέχει μια υπηρεσία στους χρήστες του διαδικτύου για αναζήτηση ορισμών λέξεων την οποία έχει ενσωματώσει στο παράθυρο αναζήτησης. Όταν πληκτρολογήσει κανείς define και μετά την λέξη στα Αγγλικά, επιστρέφονται πληροφορίες σχετικές με την λέξη. Πληροφορίες όπως ορισμοί, συνώνυμα και αντώνυμα, σχετικές εκφράσεις, καθώς και την ιστορία της λέξης, στατιστικά χρήσης της και την πιθανή προέλευσή της.

Αυτή η υπηρεσία χρησιμοποιήθηκε για τον εμπλουτισμό των συνωνύμων και αντωνύμων του βικιλεξικού, για όσες λέξεις θα μπορούσαν να βρεθούν συνώνυμες στο RTE2 σύνολο δεδομένων (περισσότερα για το RTE ειπώθηκαν στο κεφάλαιο 2).



Σχήμα 3.3: Παράδειγμα ορισμού και συνωνύμων - αντωνύμων σε αναζήτηση στο Google

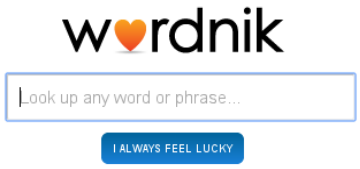
Θετικά:

- Πληρέστατα συνώνυμα για τις περισσότερες λέξεις
- Έγκυρότητα

Αρνητικά:

- Η διεπαφή που δίνει στους προγραμματιστές η Google επιτρέπει μόλις 100 επερωτήσεις την ημέρα
- Τα δεδομένα δεν είναι ανοικτά

Wordnik



Το *wordnik* είναι μια πλατφόρμα που προσφέρει στους χρήστες της πρόσβαση στην δημιουργία λιστών λέξεων, σε ορισμούς, παραδείγματα χρήσης λέξεων, συνώνυμα, αντώνυμα, λέξεις που κάνουν ομοιοκαταληξία, σχετικές λέξεις και φράσεις, ηχητικά αποσπάσματα ανάγνωσης της λέξης καθώς και εικόνες. Για την απόκτηση της πληροφορίας που συναθροίζει και πλαισιώνει, κάνει χρήση πηγών όπως το wordnet, το βικιλεξικό, καθώς και ορισμένων ακόμα λεξικών εμπορικών και μή. Στην συνέχεια δίνει προς τα έξω μια διεπαφή για προγραμματιστές με όριο 1500 κλήσεις την ώρα.

Θετικά:

- Συνάθροιση πηγών
- Λειτουργική και βολική διεπαφή

Αρνητικά:

- Μεγάλος ο αριθμός κλήσεων που χρειάζεται για τα συνώνυμα ακόμα και του RTE
- Δεν είναι ανοικτά τα δεδομένα

Στατιστικά:

- 78004 λέξεις με συνώνυμα
- 9191 λέξεις με αντώνυμα

3.3.2 Δεδομένα Εξόδου

Το σύστημα αποθηκεύει δεδομένα εξόδου με σκοπό να παρέχει δύο μορφές λειτουργικότητας. Η πρώτη αφορά την αξιολόγηση της απόδοσης των προσεγγίσεων κατασκευής των διανυσμάτων (περισσότερα στην ενότητα 3.5.2) με χρήση των υπαρχόντων δεδομένων. Η δεύτερη δίνει την δυνατότητα οπτικοποίησης της ευθυγράμμισης δύο προτάσεων σε επίπεδο φράσης ύστερα από μεταγλώττιση με χρήση του TEX.

```
1   doctor   nurse   0.0188
1   tiger   animal   0.0019
1   game   victory   0.0016
1   tiger   tiger   1.0000
0   king   cabbage   0.0005
0   professor   cucumber   0.0011
0   sugar   approach   0.0004
0   lad   wizard   0.0008
maxtotal   1.0000
mintotal   0.0000
posmean   0.0379
negmean   0.0012
posstd   0.0930
negstd   0.0009
overlap   -0.0573
overlap_norm   -0.0573
```

Σχήμα 3.4: Παράδειγμα δεδομένων αξιολόγησης

Στην πρώτη περίπτωση τα δεδομένα έξοδου έχουν μορφή κειμένου tsv (tab seperated variable) όπως φαίνεται στο Σχήμα 3.4.

Στις πρώτες 8 γραμμές βλέπουμε την βαθμολογία που έδωσε το σύστημα στα ζευγάρια λέξεων. Ο πρώτος αριθμός υποδεικνύει το αν είναι σχετικές οι λέξεις ή όχι σύμφωνα με την επισημείωση του συνόλου δεδομένων wordsimilarity-353 (0 αν δεν είναι σχετικές, 1 αν είναι σχετικές).

Στις υπόλοιπες γραμμές περιέχονται ορισμένα στατιστικά όπως μέσος όρος βαθμολογίας σχετικών και μη σχετικών ζευγαριών, μέγιστη βαθμολογία, ελάχιστη βαθμολογία και τυπική απόκλιση βαθμολογίας σχετικών και μη σχετικών ζευγαριών. Τις λεπτομέρειες για την δημιουργία των θετικών και των αρνητικών δειγμάτων θα τις δούμε στην επόμενη ενότητα.

Για κάθε επιλογή από δυνατές παραμέτρους δημιουργείται ένα αρχείο με τις παραπάνω μετρικές και προστίθεται μια εγγραφή στο αρχείο των αποδόσεων και μια στο αρχείο των επικαλύψεων η οποία προκύπτει από τα αποτελέσματα του αρχείου με τις μετρικές.

Το αρχείο των αποδόσεων περιέχει εγγραφές με τον συνδυασμό των χαρακτηριστικών, την βαθμολογία που είχε ο συνδυασμός, το κατώφλι ως βαθμολογία καθώς και το ζεύγος των λέξεων που λειτούργησαν ως κατώφλι.

Το αρχείο των επικαλύψεων περιέχει τον συνδυασμό των χαρακτηριστικών και έναν αριθμό, το μέγεθος της επικάλυψης. Περισσότερα για το πως ορίσαμε την απόδοση, τι είναι η επικάλυψη, τι είναι κατώφλι και πως κρίνουμε την απόδοση θα δούμε στην ενότητα 3.4.2.

Το αρχείο των αποδόσεων και το αρχείο των επικαλύψεων δημιουργούνται μια φορά στην αρχή της διαδικασίας και στην συνέχεια επισυνάπτεται μια εγγραφή για κάθε παραπάνω αρχείο αποτελεσμάτων.

Παράμετροι που επιλέγονται και δημιουργούν συνδυασμούς για να δημιουργηθούν τα παραπάνω αρχεία είναι:

- Προσέγγιση σχηματισμού διανυσμάτων- περισσότερα στην ενότητα 3.5.2
- Βάθος διάσχισης γράφου
- Χρήση συνωνύμων ή ορισμών

Στην δεύτερη περίπτωση επιλέγονται οι επιθυμητοί συνδυασμοί παραμέτρων, και για κάθε συνδυασμό, για κάθε ζεύγος φράσεων στο σύνολο δεδομένων RTE (δείτε το υποσύνολο του στο Παράρτημα Α'), παράγεται ένα αρχείο κειμένου με κατάληξη .sim που περιέχει έναν πίνακα με βαθμολογίες ευθυγράμμισης συντακτικών αποσπασμάτων των φράσεων σε μορφή κατανοητή από το TEX.

Η μορφή του αρχείου οπτικοποίησης της ευθυγράμμισης φαίνεται στο σχήμα 3.5. Περιέχει έναν πίνακα σε μορφή TEX. Κάθε κελί χωρίζεται με τον χαρακτήρα & και κάθε γραμμή με εξαίρεση τις 2 πρώτες και τις 2 τελευταίες εκφράζουν γραμμές του πίνακα. Η πρώτη γραμμή του πίνακα περιέχει την υπόθεση ενώ το πρώτο κελί κάθε επόμενης γραμμής περιέχει τις φράσεις του κειμένου μαζί με την ετικέτα που αντιστοιχεί στον συντακτικό πλίνθο που αντιπροσωπεύει

η κάθε μια. Οι αριθμοί εκφράζουν τις βαθμολογίες ευθυγράμμισης των φράσεων από την υπόθεση στο κείμενο μια προς μια.

Όπως και στην πρώτη περίπτωση οι παράμετροι που μπορούν να επιλεχθούν είναι ίδιες.

```
\begin{table}
\begin{tabular}{cccccc}
 & NP| Two people & VP| were wounded & PP| by & NP| a bomb & NT| . \\
NP| Police sources & 0.0002 & 0.0007 & 0.0001 & 0.0007 & 0.0000 \\
VP| stated & 0.0029 & 0.0010 & 0.0000 & 0.0002 & 0.0000 \\
NT| that & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
PP| during & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
NP| the bomb attack & 0.0005 & 0.0019 & 0.0000 & 0.4343 & 0.0000 \\
VP| involving & 0.0006 & 0.0014 & 0.0000 & 0.0004 & 0.0000 \\
NP| the Shining Path & 0.0001 & 0.0003 & 0.0000 & 0.0001 & 0.0000 \\
NT| , & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
NP| two people & 1.0000 & 0.0010 & 0.0000 & 0.0004 & 0.0000 \\
VP| were injured & 0.0010 & 0.5665 & 0.0000 & 0.0018 & 0.0000 \\
NT| . & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\
\end{tabular}
\end{table}
```

Σχήμα 3.5: Παράδειγμα δεδομένων ευθυγράμμισης

3.4 Προεπεξεργασία δεδομένων

Σε αυτήν την ενότητα θα αναλύσουμε το άτακτο κομμάτι του συστήματος. Έχει αυτό το χαρακτηριστικό καθώς η λειτουργικότητα αυτού του επιπέδου του συστήματος είναι γραμμένη ως μια πληθώρα εύκολα παραμετροποιήσιμων ανεξάρτητων προγραμμάτων που συνδέονται μέσω λειτουργικού συστήματος και χρήση της γραμμής εντολών. Απώτερος σκοπός αυτής της αρχιτεκτονικής δεν ήταν φυσικά η αταξία, αλλά η εύκολη παραμετροποίηση και η γρήγορη αλλαγή και προσαρμογή κομματιών, χαρακτηριστικό που κάνει αυτή την αρχιτεκτονική ιδανική για δοκιμές. Πρώτα θα δούμε με ποιο τρόπο έγινε η συνάθροιση και ο μετασχηματισμός σε ενιαία μορφή των δεδομένων από τις πηγές της ενότητας 3.3.1, στην συνέχεια την διαδικασία αξιολόγισης των αποτελεσμάτων και τέλος τα χαρακτηριστικά της ευθυγράμμισης.

3.4.1 Μετασχηματισμός

Το πρώτο πρόβλημα που συναντά κανείς όταν θέλει να συλλέξει ίδιου τύπου δεδομένα από διαφορετικές πηγές είναι πως συνήθως κάθε πηγή θα διατηρεί τα δεδομένα με λίγο διαφορετική μορφή, αν όχι εντελώς διαφορετική! Ως ένα παράδειγμα, για να πάρουμε τα δεδομένα συνωνύμων και ορισμών από το βικιλεξικό, έπρεπε να κάνουμε χρήση κανονικών εκφράσεων (regular expressions) για να εξάγουμε τα δεδομένα από τις σελίδες html. Στην υπηρεσία του Google, η μορφοποίηση ήταν διαφορετική, συνεπώς χρειαζόμασταν διαφορετικές κανονικές εκφράσεις. Για εξαγωγή των

συνωνύμων και των ορισμών από τις άλλες δύο πηγές είχαμε μια πολύ βολική διεπαφή, ωστόσο τίθεται το ερώτημα πως κανείς μπορεί να συμπτύξει όλα αυτά τα δεδομένα που έχει συλλέξει με διαφορετικούς τρόπους και να τα διατηρήσει σε ενιαία μορφή.

Η ενιαία μορφή που επιλέχθηκε για την αποθήκευση των συνωνύμων και των ορισμών έπρεπε παράλληλα να είναι και βολική για την δημιουργία γράφου. Συνεπώς η δομή που επιλέχθηκε να διατηρηθούν τα δεδομένα ήταν η tsv (tab separated variables) όπου η πρώτη λέξη είναι αυτή που ορίζεται ή αυτή της οποίας τα συνώνυμα αναφέρονται και όλες οι επόμενες λέξεις μέχρι την αλλαγή γραμμής είναι οι σχετικές. Συνεπώς η πρώτη λέξη - φράση μέχρι το διαχωριστικό \t εκφράζει έναν κόμβο, και κάθε επόμενη λέξη - φράση μέχρι την αλλαγή γραμμής εκφράζει τους κόμβους με τους οποίους υπάρχει σύνδεση.

```
widget  control sprocket        thingy  scraper gizmo
crouch  stoop   cower   squat   duck
jubilantly      delightedly     elatedly        joyfully
flocculent_spiral_galaxy        flocculent_galaxy       flocculent      flocculent_spiral
ching
existimation    opinion reputation      esteem
Venus'_comb     shepherd's_needle       lady's_comb
```

Σχήμα 3.6: Παράδειγμα μορφής συνωνύμων

Σύνολο δεδομένων συνωνύμων

Για την δημιουργία του συνόλου δεδομένων συνωνύμων χρησιμοποιήθηκαν ως πηγές αρχικά το βικιλεξικό και στην συνέχεια η υπηρεσία της Google. Απόσπασμα συνόλου δεδομένων συνωνύμων φαίνεται στο Σχήμα 3.6. Το σύνολο δεδομένων είναι σε μορφή tsv. Κάθε γραμμή αποτελεί μια εγγραφή του συνόλου δεδομένων. Η πρώτη λέξη αποτελεί την λέξη που ορίζεται ενώ οι επόμενες είναι τα συνώνυμά της.

Επειδή το σύνολο συνωνύμων από το βικιλεξικό ήταν αρκετά μικρό και για τις φράσεις του RTE2 - οι περισσότερες λέξεις δεν είχαν συνώνυμα - δοκιμάστηκε να εμπλουτιστεί το σύνολο δεδομένων για λέξεις που στις φράσεις του RTE2 θα είχε νόημα να ευθυγραμμιστούν. Συγκεκριμένα χρησιμοποιήθηκε υποσύνολο του RTE2 συνόλου δεδομένων το οποίο παραθέτεται στο παράρτημα Α'. Έτσι δημιουργήθηκε η παρακάτω λίστα και βρέθηκαν και προστέθηκαν τα συνώνυμά τους, τα συνώνυμα των συνωνύμων τους και τα συνώνυμα των συνωνύμων των συνωνύμων τους. Από την παραπάνω αναζήτηση συνωνύμων σε βάθος 3 βρέθηκαν 141369 συνώνυμα συνολικά για 7775 λέξεις. Στην συνέχεια αυτές προστέθηκαν στα 87141 συνώνυμα για 23641 λέξεις που είχαμε από το βικιλεξικό και αφαιρέθηκαν οι διπλοεγγραφές.

Στην συνέχεια προστέθηκαν συνώνυμα από το wordnet καθώς και συνώνυμα και αντώνυμα από το wordnik. Για την ακρίβεια, από το wordnet δεν προστέθηκαν μόνο συνώνυμα, καθώς

1. attack	12. politician	23. trip	34. computer
2. kill	13. mayor	24. sabotage	35. notebook
3. murder	14. rise	25. lifetime	
4. injure	15. baby	26. career	36. maker
5. wound	16. girl	27. ambush	37. delegation
6. produce	17. representative	28. assassination	
7. beat	18. spokesman	29. enhancers	38. member
8. assassinate	19. massacre	30. enrich	39. cause
9. defeat	20. shoot	31. creation	
10. clerk	21. victory	32. launch	40. reason
11. attorney	22. travel	33. computers	41. death

Σχήμα 3.7: Λέξεις πιθανών ευθυγραμμίσεων στο υποσύνολο του RTE2

τα synset [συνωνυμοσύνολα] αν χρησιμοποιηθούν χωρίς τα υπερώνυμα τους παραμένουν κατά κύριο λόγο ασύνδετα στον γράφο. Έτσι από το wordnet προστέθηκαν οι λέξεις που είχαν σχέση συνωνυμίας, αντωνυμίας, υπερώνυμου, ολώνυμου, συνεπαγωγής, και ομοιότητας.

Σε μερικές περιπτώσεις στα συνώνυμα υπήρχαν και φράσεις, συνεπώς αντικαταστάθηκαν τα κενά (space) που ήταν ανάμεσα σε λέξεις με τον χαρακτήρα _ για να διευκολυνθεί η απόσπαση των φράσεων αργότερα αν βρεθούν μέσα σε κείμενο. Το παραπάνω σύνολο δεδομένων αποτελεί το σύνολο δεδομένων συνωνύμων που χρησιμοποιήθηκε.

Σύνολο δεδομένων ορισμών

Το σύνολο δεδομένων ορισμών αποτελείται από ένα παρόμοιο αρχείο με αυτό που φαίνεται στο Σχήμα 3.6, με την διαφορά πως πλέον οι λέξεις που περιγράφουν αυτή που ορίζεται είναι λέξεις που βρέθηκαν στον ορισμό της.

Το σύνολο δεδομένων ορισμών δημιουργήθηκε αποκλειστικά από τις λέξεις του wordnet. Η διαδικασία που ακολουθήθηκε για να μετατραπεί ο ορισμός σε σχετικές λέξεις είναι να αφαιρεθούν από αυτόν λέξεις που έχουν δομικές ιδιότητες. Συνεπώς έγινε συντακτική ανάλυση και επισημείωση των λέξεων των ορισμών ως προς τη συντακτική τους ομάδα (ρήμα, ουσιαστικό, επίθετο κτλ.) και

διατηρήθηκαν μόνο τα λήμματα ρημάτων, ουσιαστικών, επιθέτων, και αριθμητικών. Οι φράσεις του wordnet ήταν ήδη χωρισμένες με τον χαρακτήρα _ αντί για κενό (space), οπότε δεν χρειάστηκε κάποια περαιτέρω επεξεργασία.

Σύνολο δεδομένων συνωνύμων και ορισμών

Το σύνολο δεδομένων συνωνύμων και ορισμών δημιουργήθηκε με την συνένωση των δύο παραπάνω συνόλων δεδομένων, αφού αφαιρέθηκαν τα διπλότυπα και έγινε συνάθροιση των σχέσεων.

3.4.2 Αξιολόγηση

Για την αξιολόγιση των αποτελεσμάτων ομοιότητας ζευγαριών λέξεων, χρησιμοποιήθηκε το σύνολο δεδομένων wordsimilarity-353 [Finkelstein et al., 2001] (δείτε το Παράρτημα Β΄) .

Το σύνολο δεδομένων wordsimilarity-353 είναι ένα σύνολο από ζευγάρια λέξεων βαθμολογημένα από ανθρώπους ως προς την ομοιότητα. Για την ακρίβεια, αποτελείται από δύο σύνολα δεδομένων. Το πρώτο περιέχει 153 ζευγάρια λέξεων βαθμολογημένα ως προς την ομοιότητα από 13 επισημειωτές. Το δεύτερο περιέχει 200 ζευγάρια λέξεων και η βαθμολόγηση τους έγινε από 16 επισημειωτές.

Οι βαθμολογίες κυμαίνονται από το 0 εώς το 10, με το 0 να εκφράζει λέξεις που είναι εντελώς άσχετες νοηματικά και το 10 να εκφράζει πολύ κοινές νοηματικά λέξεις ή ίδιες. Οι βαθμολογίες εξήχθησαν από τις επιμέρους βαθμολογίες των επισημειωτών ως ο μέσος όρος.

Ως αρχικό μέτρο σύγκρισης, για να παρθεί μια δυαδική απόφαση ως προς την σχετικότητα ή μη της λέξης, το σύνολο wordsim353 χωρίστηκε σε 2 κατηγορίες. Η πρώτη περιείχε τα ζευγάρια με βαθμολογίες στο διάστημα $[0, 3]$ και επισημειώθηκε ως αρνητικό δείγμα. Η δεύτερη περιείχε τα ζευγάρια με βαθμολογίες στο διάστημα $[7, 10]$ και επισημειώθηκε ως θετικό δείγμα.

Ο παραπάνω χωρισμός έγινε καθώς θεωρήθηκε πως οι λέξεις με βαθμολογίες στο $[3, 7]$ δεν είναι τόσο ξεκάθαρο αν είναι σχετικές ή όχι, και θα θέλαμε να δούμε πόσο καλά χωρίζει το σύστημά μας τα θετικά από τα αρνητικά δείγματα.

Μετά την κατασκευή των παραπάνω δύο ομάδων θετικών και αρνητικών δειγμάτων, ακολουθήθηκαν δύο διαφορετικοί τρόποι αξιολόγησης

Ο πρώτος προσπάθησε να αξιολογήσει την διαχωρισιμότητα των θετικών δειγμάτων από τα αρνητικά δείγματα με τις παρακάτω προσεγγίσεις

1. Αθροιστική

2. Κατωφλίωση

Ο δεύτερος εφαρμόστηκε σε όλο το σύνολο δεδομένων wordsimilarity-353 με την καθιερωμένη μέθοδο αξιολόγησης που συγκρίνει την κατάταξη των βαθμολογιών που αποδίδει το συστήμα στα ζευγάρια, με αυτή του συνόλου δεδομένων μέσω της στατιστικής συσχέτισης Spearman's ρ την οποία θα αναλύσουμε στην συνέχεια.

Αθροιστική προσέγγιση

Στην αθροιστική προσέγγιση συγκεντρώνονται ορισμένα στατιστικά όπως μέσος όρος βαθμολογίας θετικών και αρνητικών ζευγαριών, μέγιστη βαθμολογία, ελάχιστη βαθμολογία και τυπική απόκλιση βαθμολογίας θετικών και αρνητικών ζευγαριών. Τα στατιστικά αυτά βοηθούν στον σχηματισμό μιας συνοπτικής γενικής εικόνας για τα δεδομένα, καθώς δεν εστιάζει σε μεμονωμένες περιπτώσεις. Η ιδέα για την σύγκριση των στατιστικών είναι πως θα θέλαμε τα θετικά και τα αρνητικά δεδομένα να είναι όσο γίνεται περισσότερο διαχωρίσιμα. Αυτό θα σήμαινε πως η μέθοδος μας δίνει υψηλές βαθμολογίες σε ζευγάρια λέξεων που είναι σχετικά και χαμηλές σε όσα δεν σχετίζονται. Θα θέλαμε οι τιμές των θετικών δειγμάτων να είναι όσο γίνεται πιο μακριά από τις τιμές των αρνητικών καθώς όταν αργότερα προσθέσουμε τις βαθμολογίες λέξεων για να πάμε σε επίπεδο φράσεων, αν άσχετα ζευγάρια λέξεων έχουν υψηλές βαθμολογίες θα εισάγουν μεγάλο βαθμό θορύβου.

Μια μετρική που θα μπορούσαμε να σκεφτούμε είναι η επικάλυψη των διασπορών των θετικών (posstd) και αρνητικών (negstd) δειγμάτων γύρω από τους μέσους (posmean, negmean) όπως φαίνεται στο Σχήμα 3.8. Θα θέλαμε η τιμή αυτή να είναι κατά το δυνατό μεγαλύτερη, ει δυνατόν μη αρνητική.

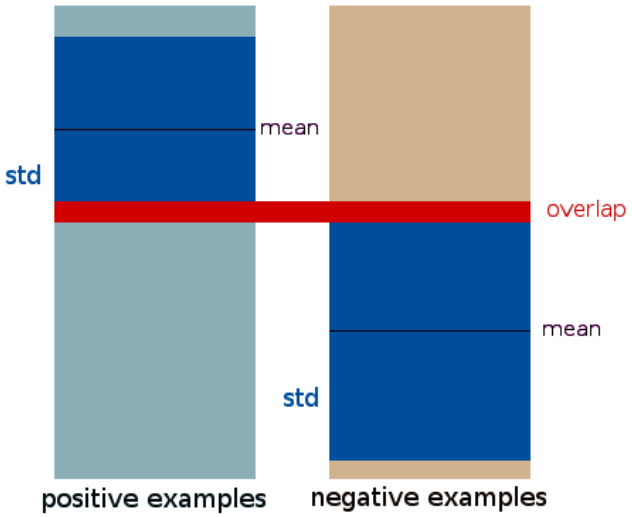
$$overlap = (posmean - posstd) - (negmean + negstd)$$

Επίσης η μετρική αυτή έχει νόημα να κανονικοποιηθεί, διότι κάθε μέθοδος παρουσιάζει άλλο μέγιστο. Με τον όρο κανονικοποίηση εννοούμε εφαρμογή της παρακάτω σχέσης, στην περίπτωση μας επειδή το ελάχιστο στις περισσότερες περιπτώσεις είναι 0, με τον όρο κανονικοποίηση πρακτικά εννοούμε διαίρεση με το μέγιστο. Σκοπός είναι να γίνει αντιστοίχιση των τιμών σε μια νέα τιμή ενός πιο αντιπροσωπευτικού κοινού εύρους τιμών για να μπορούμε να συγκρίνουμε διαφορετικές μεθόδους.

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)}$$

Για παράδειγμα αν μια μεταβλητή X έχει τιμή 6 και η μέγιστη τιμή που μπορεί να πάρει είναι 10 ενώ η ελάχιστη είναι 2, θα πάρει νέα κανονικοποιημένη τιμή $\frac{6-2}{10-2} = \frac{4}{8} = 0.5$.

Όπως αναφέραμε και στην ενότητα 3.3.2 η μετρική αυτή αποθηκεύεται για κάθε συνδυασμό παραμέτρων στο αρχείο επικαλύψεων με κανονικοποίηση και χωρίς. Στην περίπτωση χρήσης του



Σχήμα 3.8: Η επικάλυψη διασπορών θετικών και αρνητικών δειγμάτων με κόκκινο

353 ωστόσο, υπάρχει το ζεύγος tiger - tiger, και συνεπώς το μέγιστο από όλες τις βαθμολογίες είναι πάντα 1. Συνεπώς, η κανονικοποίηση εφόσον το ελάχιστον είναι 0 δεν δημιουργεί καμία διαφορά.

Μεμονωμένη προσέγγιση με κατωφλίωση

Στον αντίποδα της αθροιστικής προσέγγισης, στην προσέγγιση με κατωφλίωση κοιτάμε μια μια τις βαθμολογίες ευθυγράμμισης των θετικών και των αρνητικών ζευγαριών. Η ιδέα είναι πως αν η μέθοδος μας διαχωρίζει καλά τα θετικά από τα αρνητικά, μπορεί κανείς να βρεί το αρνητικό ζευγάρι με την μέγιστη βαθμολογία και να θεωρήσει ότι όλα τα ζευγάρια με χαμηλότερη βαθμολογία είναι αρνητικά, ενώ όσα έχουν μεγαλύτερη είναι θετικά. Έτσι μπορούμε να κατηγοριοποιήσουμε τα ζευγάρια μας με ανάθεση ετικέτας και στην συνέχεια να βρούμε πιο ποσοστό τους κατηγοριοποιήθηκε σωστά.

$$\forall i \in P \quad decision(i) = \begin{cases} 1 & \text{if} \quad i > maxneg(P) \\ 0 & \text{else} \end{cases}$$

$$\text{απόδοση} = \frac{\text{Πλήθος σωστών ετικετών}}{\text{Σύνολο ζευγαριών}}$$

Όπως αναφέρθηκε στην ενότητα 3.3.2, οι αποδόσεις αποθηκεύονται στο αρχείο αποδόσεων μαζί με το ζευγάρι που έδρασε ως κατώφλι και την βαθμολογία του.

Συσχέτιση

Στην συνέχεια στο σύνολο δεδομένων wordsimilarity-353, έγινε σύγκριση των βαθμολογιών που αποδίδει το σύστημα σε κάθε ζευγάρι σε σχέση με τις βαθμολογίες που υπάρχουν στο σύνολο δεδομένων.

Ο καθιερωμένος τρόπος σύγκρισης των αποτελεσμάτων στο σύνολο δεδομένων wordsim-353 είναι μέσω της στατιστικής συσχέτισης - Spearman's ρ. Η συσχέτιση αυτή βαθμολογεί την ομοιότητα της κατάταξης των ζευγαριών με βάση της βαθμολογία τους στις 2 περιπτώσεις.

Συγκεκριμένα, η στατιστική συσχέτιση Spearman's ρ, ταξινομεί κάθε ζευγάρι $pair_i$ ως προς την βαθμολογία του συνόλου δεδομένων. Το i είναι δείκτης που αντιστοιχεί σε συγκεκριμένο ζευγάρι πρίν την ταξινόμηση. Στην συνέχεια, αναθέτει έναν αυξανόμενο κατά 1 ακέραιο αριθμό α_i , ξεκινώντας από το 1. Η ανάθεση στο κάθε ζευγάρι $pair_i$ γίνεται σύμφωνα με την κατάταξη τους μετά την ταξινόμηση. Έπειτα, ταξινομεί ως προς την βαθμολογία του συστήματος και αναθέτει έναν δεύτερο αυξανόμενο κατά 1 ακέραιο αριθμό b_i, ξεκινώντας από το 1. Η ανάθεση στο ζευγάρι $pair_i$ γίνεται αντίστοιχα σύμφωνα με την νέα κατάταξη. Τέλος συγκρίνει τις παραπάνω κατατάξεις μέσω των διαφορών των α_i, b_i μέσω του παρακάτω τύπου.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad , \quad d_i = \alpha_i - b_i \tag{3.1}$$

3.4.3 Ευθυγράμμιση

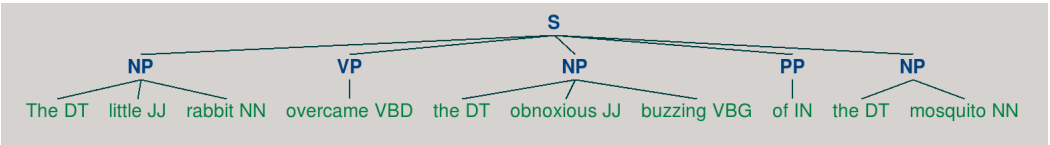
Στην ευθυγράμμιση των ζευγαριών προτάσεων του υποσυνόλου του RTE2 (παράρτημα A') υλοποιήθηκαν δύο ειδών ευθυγραμμίσεις. Ευθυγράμμιση λέξη προς λέξη, και ευθυγράμμιση φράσεων.

Λέξεων

Στην ευθυγράμμιση λέξη προς λέξη αρχικά χωρίστικαν οι προτάσεις στις επιμέρους λέξεις με χρήση του κατακερματιστή [tokenizer] του NLTK [Bird et al., 2009] . Στην συνέχεια έγινε λημματισμός των λέξεων με χρήση του wordnet και της διεπαφής του μέσω NLTK για να μετασχηματιστούν τα ρήματα στην βασική τους μορφή και τα ουσιαστικά από τον πληθυντικό στον ενικό. Για παράδειγμα bought (αγόρασα) → buy (αγοράζω) και rugs (χαλιά) → rug (χαλί). Τέλος, για κάθε πιθανό ζευγάρι έγινε κλήση του πυρήνα του συστήματος και αποθηκεύθηκαν οι βαθμολογίες σε μορφή πίνακα TEX.

Φράσεων

Πρώτο βήμα για την ευθυγράμμιση φράσεων ήταν να σπάσουμε τις προτάσεις σε συντακτικούς πλίνθους [chunks]. Για να γίνει ωστόσο αυτό, προαπαιτούμενα είναι να έχει κατακερματιστεί η πρόταση σε λέξεις και να έχουν ανατεθεί ετικέτες στην κάθε μια με τον συντακτικό της ρόλο [part of speech tag]. Για τον κατακερματισμό χρησιμοποιήθηκε ο κατακερματιστής [tokenizer] του NLTK, ενώ για την ανάθεση ετικετών συντακτικού ρόλου χρησιμοποιήθηκε ο αναλυτής [parser] του Stanford [Socher et al., 2013]. Μετά την επισημείωση των λέξεων με τον συντακτικό τους ρόλο εκπαιδεύτηκε ένα μοντέλο μέσω του NLTK στο σύνολο δεδομένων conll2000 με χρήση του MegaM [Daumé III, 2004] κατηγοριοποιητή, ένα μοντέλο μέγιστης εντροπίας. Στην συνέχεια χρησιμοποιήθηκε το μοντέλο για τον χωρισμό των προτάσεων σε συντακτικούς πλίνθους [chunks], αποτέλεσμα του οποίου είναι ένα δέντρο με τα μέρη που χωρίστηκε η πρόταση.



Σχήμα 3.9: Παράδειγμα Συντακτικού δέντρου

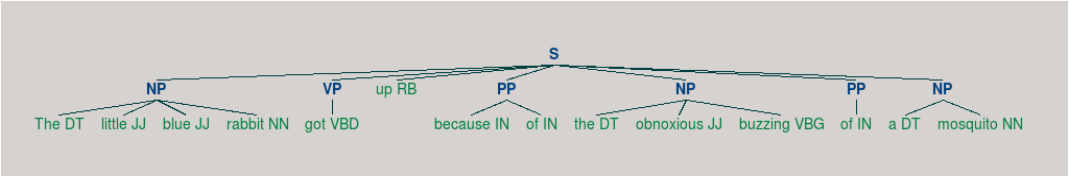
Ωστόσο, υπήρχαν μεμονωμένες περιπτώσεις που δεν υπήρξε ανάθεση σε κόμβο του δέντρου κάποια ετικέτα, όπως στην περίπτωση του σχήματος 3.10. Επίσης, στην ευθυγράμμιση φράσεων θα θέλαμε ορισμένες φράσεις που υπάρχουν στο λεξικό και έχουν συνώνυμα να αντικαθίστανται ολόκληρες και να μην αναλύονται σε λέξεις. Για αυτόν τον λόγο αναζητήθηκαν στο δέντρο οι παραπάνω φράσεις, αποσπάστηκαν από τους κόμβους του δέντρου στους οποίους ανήκαν και ενώθηκαν σε έναν νέο κόμβο με ετικέτα PR. Περισσότερα για τις ετικέτες μπορεί να δει κανείς στον πίνακα 2.1 στο κεφάλαιο 2. Οι ετικέτες του πίνακα 3.1 δεν αποτελούν μέρος της συντακτικής ανάλυσης, έγινε εισαγωγή τους για να φαίνεται στην ευθυγράμμιση που είναι οι φράσεις που υπήρχαν στο λεξικό.

Ετικέτα	Αγγλικά	Ελληνικά
NT	No Tag	Έλλειψη ετικέτας
PR	Phrase from dictionary	Φράση που υπήρχε στο λεξικό

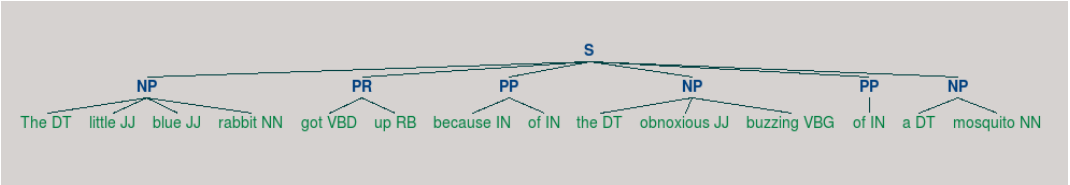
Πίνακας 3.1: Λίστα επισημειώσεων συντακτικών πλίνθων [chunks]

Πρίν την κλήση του πυρήνα του συστήματος, έγινε λημματισμός των φράσεων αντίστοιχος με την διαδικασία που ακολουθήθηκε στην ευθυγράμμιση λέξεων και μετατράπηκαν όσοι κόμβοι

(α′) Πρίν



(β′) Μετά



Σχήμα 3.10: Παράδειγμα Συντακτικού δέντρου πρίν και μετά την επεξεργασία

του δέντρου είχαν ετικέτα PR σε μια φράση με τις λέξεις ενωμένες με τον χαρακτήρα _. Τέλος, και εδώ αποθηκεύθηκαν οι βαθμολογίες σε μορφή πίνακα TEX. Για την μορφή των αποτελεσμάτων μπορείτε να ανατρέξετε στην ενότητα 3.3.2 και στον πίνακα 3.5.

3.5 Πυρήνας

3.5.1 Δημιουργία Γράφου

Η ιδέα για την δημιουργία γράφου συνωνύμων ή σχετικών λέξεων όπως αναφέρθηκε και στην βιβλιογραφία έχει πολλές εφαρμογές. Ωστόσο, παρατηρούμε πως το τι πληροφορία θα κρατήσει κανείς, πως θα την εξάγει από τον γράφο και το πως θα ερμηνεύσει τα αποτελέσματα, δεν αποτελούν κάτι το τετριμμένο καθώς το πλήθος των δυνατών συνδυασμών μεθόδων είναι εκθετικό στον αριθμό των μεθόδων και κατά έναν τρόπο η εφαρμογή κάθε μεθόδου έχει ευριστικό χαρακτήρα.

Πέρα από τις υπάρχουσες υλοποιήσεις, το *wordnet* [Miller, 1995] για παράδειγμα, θα μπορούσαμε να ισχυριστούμε πως η απεικόνιση σχέσεων ανάμεσα σε λέξεις σε γράφο είναι ενστικτώδης. Μέσω της ένωσης κόμβων με ακμές μπορούμε να απεικονίσουμε την πολυπλοκότητα των σχέσεων που εμφανίζονται ανάμεσα στις λέξεις [2], τον τρόπο χρήσης τους μέσα σε μια πρόταση και τις έννοιες τους. Η απεικόνιση των δεδομένων με αυτόν τον τρόπο δίνει ένα γενικό μοντέλο που είναι επεκτάσιμο, και μιας που το πρόβλημα της κατανόησης της φυσικής γλώσσας είναι περίπλοκο θα ήταν αφελές να ξεκινήσει κανείς με ένα πιο απλό μοντέλο. Τρόποι επέκτασης θα μπορούσαν

[2] συνώνυμο, αντώνυμο, υπέρνυμο, υπώνυμο, ετερώνυμο κτλ.

να είναι η εισαγωγή βαρών στις ακμές καθώς και η δημιουργία πιο σύνθετων κόμβων. Επίσης με τη χρήση γράφων, η πληροφορία που διατηρούμε είναι αχανής και η εξόρυξη των σχέσεων που μπορεί να βρεί κανείς περιορίζεται πλεόν μόνο από το υπολογιστικό του δυναμικό, δοθέντος πως ο γράφος που επεξεργάζεται χωράει στην κύρια μνήμη ενός ή περισσοτέρων μηχανημάτων.

Μια αρχική αναπαράσταση που θα μπορούσε να σκεφτεί κανείς, που εκμεταλλεύεται την δενδρική δομή 3.11 των δεδομένων, θα ήταν η κατασκευή ενός κατευθυνόμενου γράφου χωρίς βάρη.

Παραδείγματα πηγών πληροφορίας πάνω σε λέξεις που μπορούν να αναπαρασταθούν με δενδρική δομή είναι:

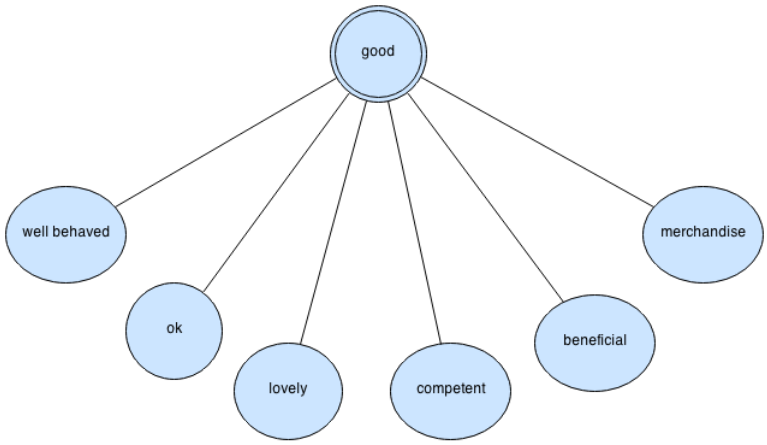
- λεξικά

  συνώνυμα

  ταυτόσημες έννοιες

- θησαυροί

  ορισμοί εννοιών



Σχήμα 3.11: Παράδειγμα δενδρικής αναπαράστασης συνωνύμων - τα συνώνυμα της λέξης good (καλός)

Η προσέγγιση μας είναι να συσσωρεύσουμε τις παραπάνω δομές και να δημιουργήσουμε ένα συνολικό ευρύτερο σώμα πληροφορίας. Για να χρησιμοποιήσουμε συνεπώς τις παραπάνω πηγές πληροφορίας και να φτιάξουμε έναν γράφο:

1. Προσθέτουμε όλες τις λέξεις που υπάρχουν στο σώμα πληροφορίας ως κόμβους
2. Για κάθε δενδροειδή μορφή, βρίσκουμε τους αντίστοιχους κόμβους στον γράφο και τραβάμε ακμή από τη ρίζα στα φύλλα

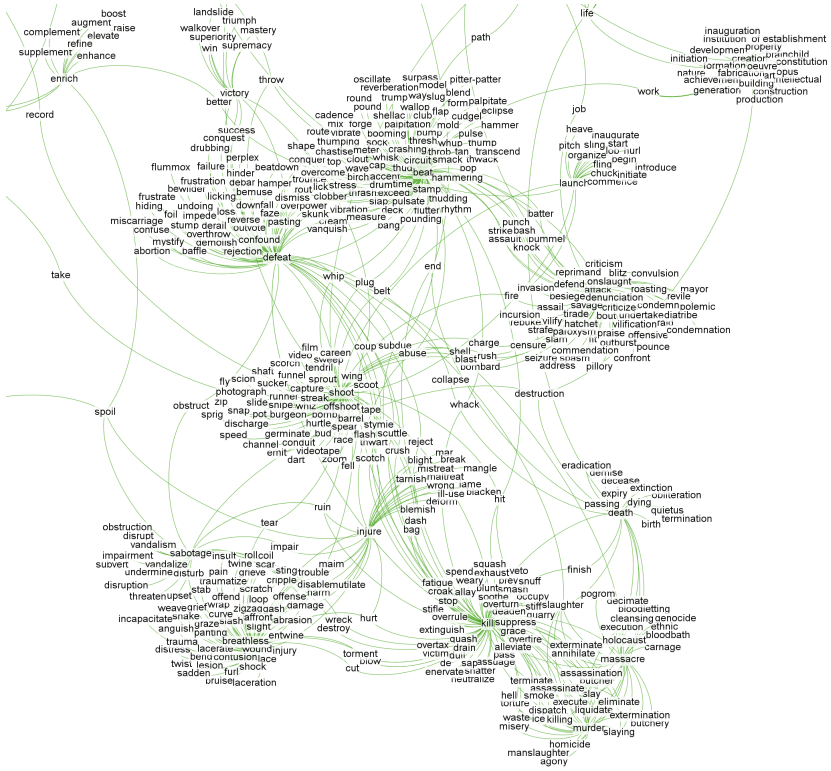
Με την παραπάνω προσέγγιση το σώμα πληροφορίας που δημιουργείται διατηρεί τις αρχικές σχέσεις και τις επεκτείνει καθώς δημιουργεί σχέσεις ανάμεσα σε δομές που ήταν αρχικά

ασυσχέτιστες. Για παράδειγμα, αν έχουμε τα συνώνυμα της λέξης καλός και τα συνώνυμα της λέξης συνεπής, και το καλός είναι συνώνυμο του συνεπής, πλέον έχουμε πως τα συνώνυμα του καλός είναι συνώνυμα του συνεπής σε βάθος 2. Το παραπάνω σκεπτικό επεκτείνεται χωρίς τελικά να υπάρχει περιορισμός στο βάθος, αν και λογικά οι σχέσεις των λέξεων από ένα σημείο και πέρα πιθανώς να μην είναι αντιπροσωπευτικές. Για παράδειγμα:

$$\text{τρέχω} \rightarrow \text{κινούμαι} \rightarrow \text{ρέω} \rightarrow \text{κελαρύζω}$$

Η λέξη τρέχω σίγουρα δεν έχει την ίδια σχέση με το κελαρύζω όπως έχει με το κινούμαι, ωστόσο η μετάβαση έχει λογική. Σε κάθε μετάβαση συνδέονται λέξεις που μπορούν να αντικαταστήσουν η μια την άλλη σε μεγάλο ποσοστό εκφράσεων. Εδώ φαίνεται άλλωστε πως η έννοια του συνωνύμου είναι εως ένα σημείο ασαφής, καθώς σε αντίθετη περίπτωση είτε καμία λέξη δεν θα ήταν συνώνυμη της άλλης είτε θα είχαμε δύο διαφορετικές λέξεις με ολόιδια χρηστικότητα - πράγμα λίγο χαζό.

Ωστόσο η μικρή αυτή αλλαγή στο νόημα που επιτρέπει την μετάβαση μεταφράζεται ως χαλάρωση της έννοιας και συνεπώς επιτρέπει την σύνδεση της έννοιας και με μια τρίτη, πράγμα που αρχικά θα ήταν αδύνατο. Παρατηρούμε τελικά πως ενώ στο μυαλό μας έχουμε πιθανώς δύο διαφορετικές ομάδες λέξεων, "θετικές" και "αρνητικές", άλλοτε σε μικρό και άλλοτε σε μεγάλο βάθος, σε ένα γράφο συνωνύμων, οι λέξεις αυτές συνδέονται.



Σχήμα 3.12: Παράδειγμα γράφου συνωνύμων και αντωνύμων

Αυτό μπορεί να δημιουργήσει προβλήματα αν προσπαθήσουμε να βρούμε ομοιότητες ανάμεσα σε κόμβους και κοιτάξουμε πέρα από την άμεση γειτονιά - πιο βαθιά στον γράφο. Περισσότερα για τις επιπλοκές που μπορεί να δημιουργήσει η αναζήτηση συνωνύμων βαθιά στον γράφο θα δούμε στο κεφάλαιο 4.

Επίσης παρατηρούμε πως αν μια λέξη είναι συνώνυμη της άλλης δεν σημαίνει ξεκάθαρα πως και η άλλη είναι συνώνυμη της πρώτης. Η σχέση δεν είναι πάντα αμφίδρομη, ή μάλλον, για να τεθεί πιο σωστά, δεν έχει πάντα την ίδια βαρύτητα και στις δύο κατευθύνσεις. Για παράδειγμα η λέξη "τρέχω" αν χρησιμοποιηθεί στην πρόταση "Πήγα να τρέξω στο δάσος" έχει έμφυτη την έννοια της κίνησης. Αντίθετα αν κάποιος κινείται, δεν είναι ανάγκη να τρέχει και αυτό οφείλεται στην γενικότητα και στο επίπεδο αφαίρεσης που έχει η λέξη "κίνηση" σαν έννοια.

shoot	πυροβολώ
fire projectile	πυροβολώ
film	γυρίζω ταινία
dash	κινούμαι γρήγορα
sprout	φύτρα
sprout	φυτρώνω
damn	να πάρει

Σχήμα 3.13: Διαφορετικοί τρόποι χρήσης της λέξης shoot στα Αγγλικά

Όπως φαίνεται και στο σχήμα 3.12 το οποίο κατασκευάστηκε στο σύνολο δεδομένων των συνωνύμων που αναλύσαμε στη ενότητα 3.4.1 με χρήση του εργαλείου Gephi [Bastian et al., 2009], υπάρχουν λέξεις που δεν είναι άμεσα συνδεδεμένες όμως σε περίπτωση που βρεθούν με συγκεκριμένα συμφραζόμενα γίνεται πιο έντονη η σύνδεση τους. Βλέπουμε για παράδειγμα πως το shoot (πυροβολώ) δεν είναι συνώνυμο με την λέξη murder (δολοφονώ). Ωστόσο, σε μια πρόταση όπως η "Mark David Chapman shot John Lennon dead" είναι ξεκάθαρο πως η λέξη shot είναι συνώνυμη της murder.

Επιχειρηματολογώ εδώ πως τελικά κάθε λέξη περιέχει ταυτόχρονα πολλά νοήματα που κρύβονται πιο βαθιά σε έναν γράφο συνωνύμων και έρχονται στο φως ανάλογα με τον τρόπο χρήσης της λέξης και τα συμφραζόμενα. Άλλωστε στα Ελληνικά η λέξη πυροβολώ και η λέξη δολοφονώ φαίνεται να είναι πιο σχετικές, και αυτό μπορεί να εξηγηθεί αν σκεφτεί κανείς πως η λέξη πυροβολώ δεν έχει όσες διαφορετικές έννοιες έχει η λέξη shoot. Για την επιβεβαίωση της συνωνυμίας αρκεί κανείς να μπορεί να διακρίνει από τα συμφραζόμενα αν έγινε επίθεση ή αν το θύμα είναι νεκρό [3].

Συνεπώς, βασικό χαρακτηριστικό που θα πρέπει να έχει η μέθοδος που θα αναπτυχθεί για εύρεση ομοιότητας λέξεων, θα είναι να μπορεί να επεκταθεί σε επίπεδο φράσεων για να μπορούν να επιλυθούν τυχόν διενέξεις με βάση τα συμφραζόμενα. Όπως θα δούμε στην επόμενη ενότητα,

[3]Ζητώ συγνώμη για το άσχημο παράδειγμα. το RTE2 σύνολο δεδομένων ήταν γεμάτο τέτοια γεγονότα!

η εξαγωγή διανυσμάτων για κάθε λέξη από τον γράφο και η μετέπειτα χρήση τους για εξαγωγή ομοιότητας έχει αυτό το χαρακτηριστικό και θα δούμε πως επηρρεάζεται από την επιλογή προσέγγισης της κατασκευής των διανυσμάτων και από το βάθος διάσχισης του γράφου.

3.5.2 Δημιουργια διανυσμάτων

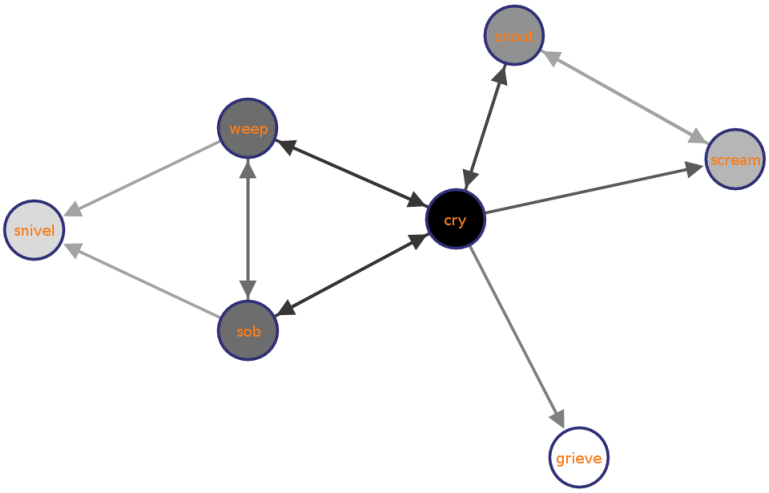
Όπως προαναφέραμε, ένας τρόπος να μπορέσουμε να επεκτείνουμε την ομοιότητα λέξεων σε ομοιότητα φράσεων είναι να κατασκευάσουμε ένα διάνυσμα για κάθε λέξη με μια διάσχιση του γράφου σε ορισμένο βάθος. Η δημιουργία των διανυσμάτων γίνεται ως προσπάθεια να εκφραστεί κάθε λέξη σε συνιστώσες άλλων σχετικών λέξεων. Έτσι θα είναι δυνατή η συσχέτιση λέξεων που έχουν μη μηδενική τιμή στις αντίστοιχες θέσεις τους.

Στην συνέχεια, αν θεωρήσουμε πως οι μη μηδενικές τιμές σε ένα διάνυσμα εκφράζουν την υπέρθεση των πιθανών νοημάτων που μπορεί να έχει μια λέξη εκφρασμένη ως άλλες λέξεις, μπορούμε να φανταστούμε μια φράση ως το άθροισμα αυτών των εκφράσεων. Μπορούμε να παρατηρήσουμε εδώ πως θέσεις στις οποίες υπάρχουν επικαλύψεις θα έχουν την τάση να ενισχύουν την επιρρόη των λέξεων στις οποίες αντιστοιχούν αυτές οι επικαλύψεις στο τελικό "νόημα" ενώ η έλλειψη επικαλύψεων θα την μειώνει.

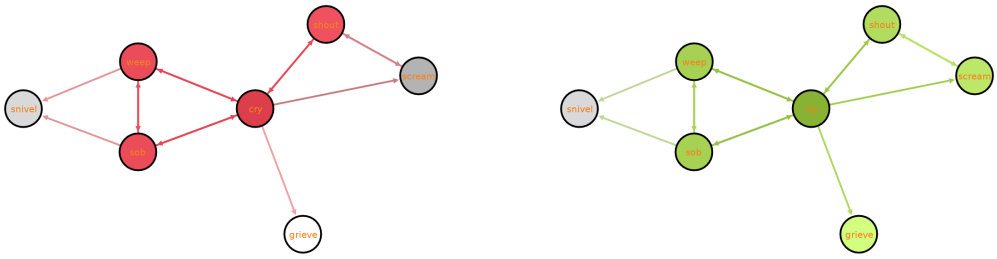
Για να δούμε λίγο οπτικά την κατάσταση ενός γράφου συνωνύμων ας πάρουμε το παράδειγμα του σχήματος 3.14. Οι κόμβοι είναι σκιασμένοι με βάση το πλήθος τον εισερχόμενων και εξερχόμενων ακμών. Στο σχήμα 3.15 βλέπουμε με κόκκινο όσους κόμβους έχουν εξερχόμενες ακμές προς τον κόμβο cry (κλαίω) και πράσινο όσους κόμβους έχουν εισερχόμενες ακμές από τον κόμβο cry. Αυτό σημαίνει πως οι κόκκινοι κόμβοι έχουν ως συνώνυμο την λέξη cry ενώ οι πράσινοι είναι λέξεις που έχει ως συνώνυμα η λέξη cry. Όπως είδαμε και στην προηγούμενη ενότητα, η σχέση συνωνύμων δεν είναι ανάγκη να είναι αμφίδρομη και για αυτό οι πράσινοι κόμβοι δεν είναι ίδιοι με τους κόκκινους.

$$V_{cry} = \begin{bmatrix} 0 & cry \\ 0 & sob \\ 0 & weep \\ 0 & snivel \\ 0 & shout \\ 0 & scream \\ 0 & grieve \end{bmatrix}$$

Με τον παραπάνω γράφο συνωνύμων πριν ξεκινήσουμε να κάνουμε οποιαδήποτε διάσχιση ξεκινώντας από οποιονδήποτε κόμβο, μπορούμε να σχηματίσουμε το μηδενικό διάνυσμα για οποιαδήποτε λέξη επιθυμούμε. Στο παράδειγμά μας, ας είναι για το cry (κλαίω). Το γεγονός πως σχηματίσαμε το διάνυσμα για το cry υπαινίσσεται πως θα ξεκινήσουμε την διάσχιση του γράφου από εκεί. Άλλωστε στην γειτονιά του κόμβου cry βρίσκονται και τα πιο "ισχυρά" συνώνυμα του. Παρομοίως με το cry μπορούμε να φτιάξουμε ένα ίδιο διάνυσμα για κάθε λέξη. Τα παραγόμενα διανύσματα έχουν μια θέση για κάθε λέξη στο σύνολο



Σχήμα 3.14: Παράδειγμα γράφου συνωνύμων



Σχήμα 3.15: Αριστερά με κόκκινο οι κόμβοι με ακμές προς τον κόμβο cry, δεξιά με πράσινο οι κόμβοι με εισερχόμενες ακμές από τον κόμβο cry

λέξεων, έστω $\mathbf{W}$. Μπορούμε να αναφερθούμε σε κάθε θέση του διανύσματος για ευκολία με χρήση της λέξης στην οποία αντιστοιχεί.

Πέρα από τη λέξη για την οποία κατασκευάζεται το διάνυσμα υπάρχει ως παράμετρος το βάθος διάσχισης του γράφου. Για παράδειγμα το $V_{cry,3}$ εκφράζει το διάνυσμα που δημιουργείται ξεκινώντας από τον κόμβο cry με διάσχιση του υπογράφου που ορίζεται από τα μονοπάτια που ξεκινούν από τον κόμβο cry και έχουν μήκος μικρότερο ή ίσο του 3.

Τέλος, αυτό που έμεινε να προσδιορίσουμε είναι η μέθοδος απόδοσης βαρών στα διανύσματα κατά τη διάσχιση. Ακολουθούν οι προσεγγίσεις που υλοποιήθηκαν.

- boolean
- count
- countdepth
- dispersion
- dispersion+
- concentration+
- robinhood

Θα ορίσουμε σε αυτό το σημείο την ορολογία του πίνακα 3.2 για να μπορέσουμε να εκφράσουμε στην συνέχεια πιο αφαιρετικά τις παραπάνω προσεγγίσεις.

Με τα παραπάνω υπ' όψην, παρατηρούμε πως για όλες τις παρακάτω προσεγγίσεις τα διανύσματα παράγονται με τον παρακάτω τρόπο 3.2.

$$V_{word,depth} = \begin{bmatrix} V_{word,depth}[w_1] \\ V_{word,depth}[w_2] \\ .. \\ .. \\ V_{word,depth}[w_n] \end{bmatrix} \quad \forall w_i \in W \quad | \quad V_{word,depth}[w_i] = f(w_i) \tag{3.2}$$

Η παράμετρος που διαφέρει για κάθε προσέγγιση είναι η συνάρτηση $f(w_i)$.

$\mathbf{depth}$	Βάθος αναζήτησης, $depth \in \mathbb{N}^*$.
$\mathbf{W}$	Το σύνολο όλων των λέξεων.
$\lvert\mathbf{X}\rvert$	Πληθικότητα - αριθμός στοιχείων του σύνολου ή της λίστας $\mathbf{X}$.
$\mathbf{G}(\mathbf{W}, \mathbf{E})$	Ο κατευθυνόμενος γράφος που περιέχει το σύνολο των λέξεων ως κόμβους και ως ακμές το σύνολο των κατευθυνόμενων διμερών σχέσεων που μας ενδιαφέρουν, όπως για παράδειγμα τις σχέσεις συνωνύμων $\mathbf{synonym}(\mathbf{word}, \mathbf{synonym})$.
$\mathbf{W}(\mathbf{G})$	Το σύνολο των κόμβων του γράφου $\mathbf{G}$
$\mathbf{E}(\mathbf{G})$	Το σύνολο των ακμών του γράφου $\mathbf{G}$.
$\mathbf{I}(\mathbf{w})$	Το σύνολο των εισερχόμενων ακμών του κόμβου $\mathbf{w}$ στον γράφο $\mathbf{G}$.
$\mathbf{O}(\mathbf{w})$	Το σύνολο των εξερχόμενων ακμών του κόμβου $\mathbf{w}$ στον γράφο $\mathbf{G}$.
$\mathbf{P}_{\alpha,\beta}$	Μονοπάτι στον γράφο $\mathbf{G}$ από τον κόμβο α στον κόμβο β το οποίο θα εκφράσουμε ως μια διατεταγμένη λίστα με βάση τη σειρά διάσχισης από $\mathbf{n}$ στο πλήθος ακμές εκφρασμένες ως διατεταγμένα ζεύγη κόμβων $(\mathbf{x}, \mathbf{y})$, όπου $\mathbf{n} \in \mathbb{N}^*$. Το μονοπάτι θεωρείται πως μπορεί να περιέχει και κυκλώματα, δηλαδή η λίστα $\mathbf{P}_{\alpha,\beta}$ είναι δυνατό να περιέχει παραπάνω από μια φορές την ίδια ακμή.
$\mathbf{S}_{node,depth}$	Σύνολο από μονοπάτια $\mathbf{P}_{node,w}$ που πληρούν για το βάθος την σχέση $\lvert\mathbf{P}_{node,w}\rvert \leq depth$. $\mathbf{S}_{node,depth} = \{\mathbf{P}_{node,w} \quad \lvert \quad \forall w \in W(G), \quad \lvert\mathbf{P}_{node,w}\rvert \leq depth\}$
$\mathbf{V}_{node,depth}$	Διάνυσμα μεγέθους $\mathbf{n} \times 1$ όπου $\mathbf{n} = \lvert\mathbf{W}\rvert$. Το διάνυσμα σχηματίζεται κατά την διάσχιση του συνόλου μονοπατιών $\mathbf{S}_{node,depth}$.
$\mathbf{V}_{node,depth}[\mathbf{word}]$	Η θέση στο διάνυσμα $\mathbf{V}_{node,depth}$ που αντιστοιχεί στην λέξη $\mathbf{word}$.

Πίνακας 3.2: Πίνακας μαθηματικής σημειογραφίας

Ακολουθούν οι προσεγγίσεις στην εξής μορφή:

- Λεξική περιγραφή
- Μαθηματική περιγραφή - ορισμός του $f(w_i)$ (όπου είναι δυνατόν)
- Ψευδοκώδικας
- Παραδείγματα (όπου είναι δυνατόν)

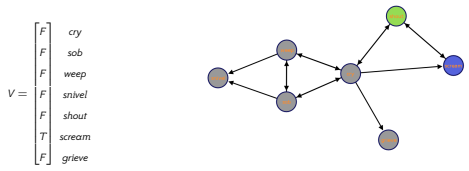
boolean

Τα διανύσματα είναι τύπου boolean. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος depth για τον σχηματισμό του διανύσματος για την λέξη word, προσπελαύνονται όσα μονοπάτια ξεκινούν από τον κόμβο word και έχουν ακμές λιγότερες ή ίσες με depth. Σημειώνονται ως αληθείς οι θέσεις στο διάνυσμα που αντιστοιχούν στις λέξεις που συναντώνται.

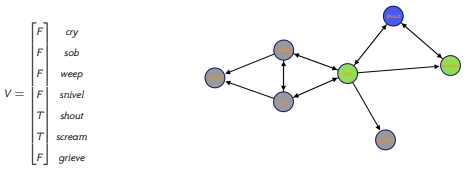
$$f(w_i) = \begin{cases} true & \text{if} \quad P_{word,w_i} \in S_{word,depth} \\ false & \text{else} \end{cases} \tag{3.3}$$

Αλγόριθμος 3.1 Ψευδοκώδικας μεθόδου boolean

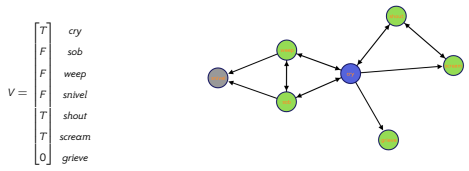
```
Συνάρτηση BOOLEANSCORE(vector,node,depth)
    if depth > 0 then
        depth ← depth − 1
        neighbours ← neighbours(node)
        for neighbour in neighbours do
            booleanscore(vector,neighbour,depth)
        τέλος for
    τέλος if
    vector[node] ← True
τέλος Συνάρτηση
```



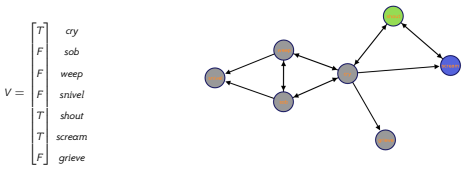
(α′) Βάθος 0



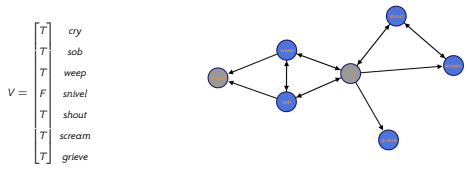
(β′) Βάθος 1



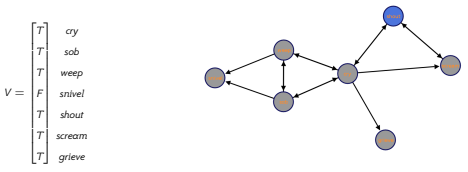
(γ′) Βάθος 2, φάση 1



(δ′) Βάθος 2, φάση 2



(ε′) Βάθος 3, φάσεις 1-5 μαζί



(στ′) Βάθος 3, φάση 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 boolean -debug
scream
sob : 1 weep : 1 cry : 1 grieve : 1 shout : 1 scream : 1
```

(ζ′) Διάνυσμα συστήματος

Σχήμα 3.16: Παράδειγμα boolean ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις.

count

Τα διανύσματα είναι τύπου int. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος depth για τον σχηματισμό του διανύσματος για την λέξη word, προσπελαύνονται όσα μονοπάτια ξεκινούν από τον κόμβο word και έχουν ακμές λιγότερες ή ίσες με depth. Μετρούνται οι εμφανίσεις της κάθε λέξης και σημειώνονται στην αντίστοιχη θέση του διανύσματος.

$$f(w_i) = \sum_{path}^{S_{word,depth}} score, \quad score = \begin{cases} 1 & \text{if} \quad path = P_{word,w_i} \\ 0 & \text{else} \end{cases} \tag{3.4}$$

Αλγόριθμος 3.2 Ψευδοκώδικας μεθόδου count

Συνάρτηση COUNTSCORE(*vector*,*node*,*depth*)

if $depth > 0$ **then**

$depth \leftarrow depth - 1$

$neighbours \leftarrow neighbours(node)$

for *neighbour* *in* *neighbours* **do**

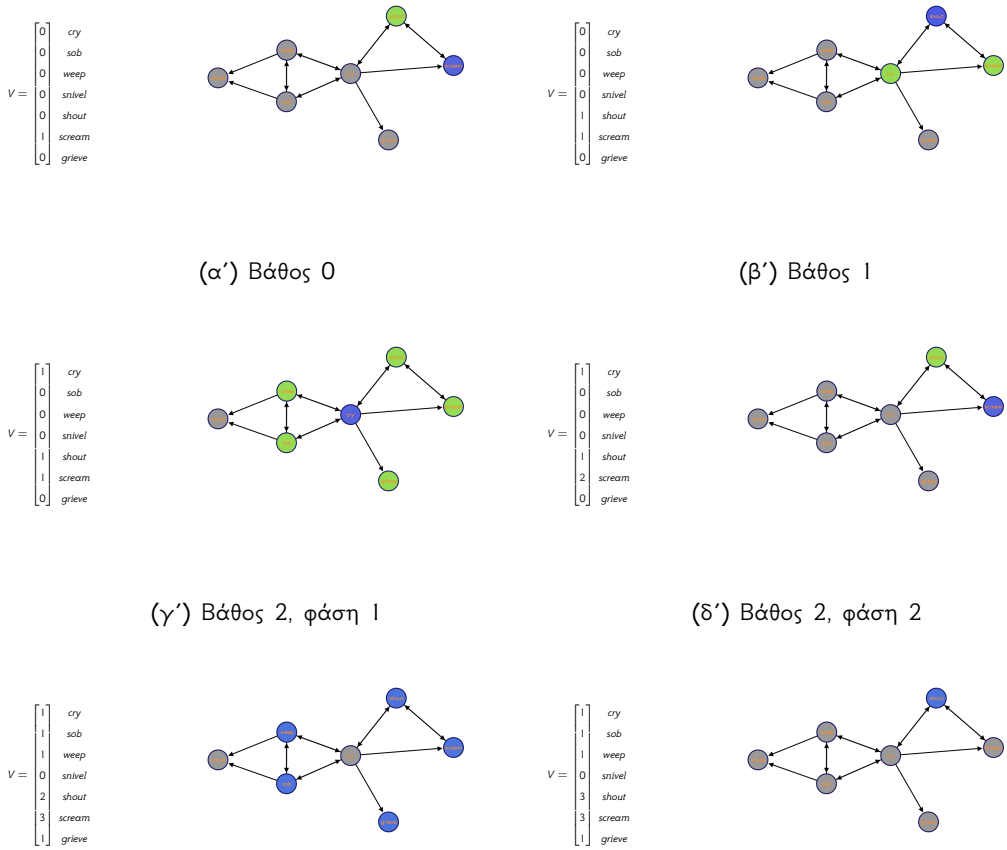
countscore(*vector*,*neighbour*,*depth*)

τέλος for

τέλος if

$vector[node] \leftarrow vector[node] + 1$

τέλος Συνάρτηση



(α′) Βάθος 0

(β′) Βάθος 1

(γ′) Βάθος 2, φάση 1

(δ′) Βάθος 2, φάση 2

(ε′) Βάθος 3, φάσεις 1-5 μαζί

(στ′) Βάθος 3, φάση 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 count -debug
scream
sob : 1 weep : 1 cry : 1 grieve : 1 shout : 3 scream : 3
```

(ζ′) Διάνυσμα συστήματος

Σχήμα 3.17: Παράδειγμα count ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις.

countdepth

Τα διανύσματα είναι τύπου double. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος depth για τον σχηματισμό του διανύσματος για την λέξη word, προσπελαύνονται όσα μονοπάτια ξεκινούν από τον κόμβο word και έχουν ακμές λιγότερες ή ίσες με depth. Προστίθεται στην θέση του διανύσματος που αντιστοιχεί στη λέξη word $\frac{1}{depth+1}$ για κάθε συνάντηση της κατά την προσπέλαση.

$$f(w_i) = \sum_{path}^{S_{word,depth}} score, \quad score = \begin{cases} \frac{1}{depth+1} & \text{if} \quad path = P_{word,w_i} \\ 0 & \text{else} \end{cases} \tag{3.5}$$

Αλγόριθμος 3.3 Ψευδοκώδικας μεθόδου countdepth

Συνάρτηση COUNTDEPTHSCORE(*vector*,*node*,*depth*,*totaldepth*)

if $depth > 0$ **then**

$realdepth \leftarrow totaldepth - depth$

$depth \leftarrow depth - 1$

$neighbours \leftarrow neighbours(node)$

for *neighbour* *in* *neighbours* **do**

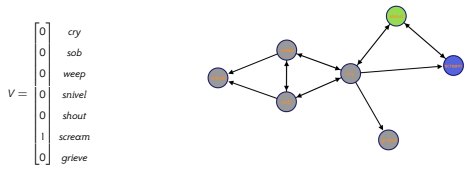
countdepthscore(*vector*,*neighbour*,*depth*,*totaldepth*)

τέλος for

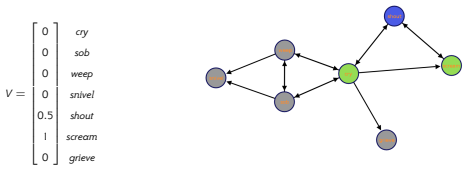
τέλος if

$vector[node] \leftarrow vector[node] + \frac{1}{realdepth}$

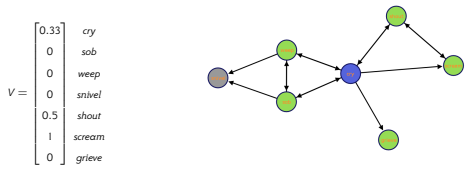
τέλος Συνάρτηση



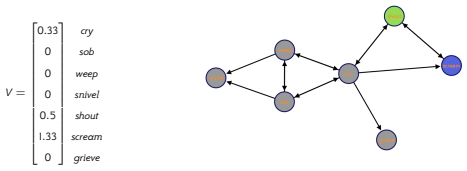
(α′) Βάθος 0



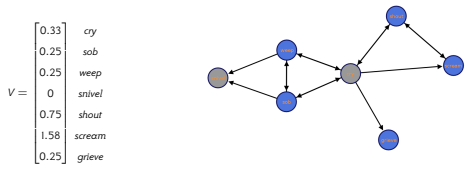
(β′) Βάθος 1



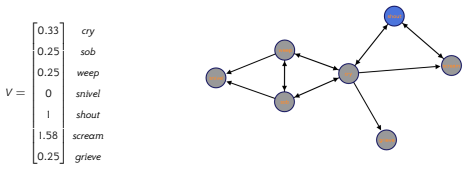
(γ′) Βάθος 2, φάση 1



(δ′) Βάθος 2, φάση 2



(ε′) Βάθος 3, φάσεις 1-5 μαζί



(στ′) Βάθος 3, φάση 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 countdepth -debug
scream
sob : 0.25 weep : 0.25 cry : 0.333333 grieve : 0.25 shout : 1 scream : 1.58333
```

(ζ′) Διάνυσμα συστήματος

Σχήμα 3.18: Παράδειγμα countdepth ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις.

recursive

Τα διανύσματα είναι τύπου double. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος υπολογίζεται η ομοιότητα της κάθε λέξης με την προηγούμενη λέξη στο μονοπάτι σε βάθος : βάθος -1. Χρησιμοποιείται η countdepth για την κατασκευή των διανυσμάτων και επιστρέφεται η ομοιότητα συνωνύμων τους. Στην συνέχεια αυτή η ομοιότητα που προκύπτει χρησιμοποιείται ως βάρος στην countdepth για τον συγκεκριμένο κόμβο. Σκοπός της μεθόδου είναι να περιορίσει το βάρος κόμβων που έχουν μικρή σχέση με την λέξη για την οποία δημιουργείται το διάνυσμα.

Αλγόριθμος 3.4 Ψευδοκώδικας μεθόδου recursive

Συνάρτηση getSimilarity(*node*1,*node*2,*depth*,*weight*)
  *vector*1 ← *initialize*(0)
  *vector*2 ← *initialize*(0)
  countdepthscore(*vector*1,*node*1,*depth*,*depth* + 1)
  countdepthscore(*vector*2,*node*2,*depth*,*depth* + 1)
  similarity ← cosineSimilarity(*vector*1,*vector*2)
  return similarity
τέλος Συνάρτηση

Συνάρτηση recursivescore(*vector*,*node*,*depth*,*totaldepth*,*weight*)
  realdepth ← *totaldepth* − *depth*
  if *depth* > 0 **then**
    depth ← *depth* − 1
    neighbours ← *neighbours*(*node*)
    for *neighbour* *in* *neighbours* **do**
      similarity ←getSimilarity(*node*,*neighbour*,*depth*,*weight*)
      recursivescore(*vector*,*neighbour*,*depth*,*similarity*)
    τέλος for
  τέλος if
  $vector[node] \leftarrow vector[node] + \frac{weight}{realdepth}$
τέλος Συνάρτηση

dispersion

Τα διανύσματα είναι τύπου double. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος σχηματίζεται βάρος $\frac{1}{\text{πληθικότητα εξερχόμενων ακμών}}$ σε κάθε κόμβο. Το βάρος αυτό προστίθεται στο υπάρχων βάρος καθενός γείτονα με τους οποίους ενώνεται με εξερχόμενες ακμές ο συγκεκριμένος κόμβος. Ο κόμβος για τον οποίο γίνεται η διάσχιση, δηλαδή ο πρώτος κόμβος που διασχίζεται ανατίθεται τιμή 1.

$$f(w_i) = \sum_{path}^{S_{word,depth}} score, \quad score = \begin{cases} 1 & \text{if} \quad path = P_{word,w_i} \quad \text{and} \quad O(w_i) = 0 \\ \frac{1}{O(w_i)} & \text{else if} \quad path = P_{word,w_i} \\ 0 & \text{else} \end{cases} \tag{3.6}$$

Αλγόριθμος 3.5 Ψευδοκώδικας μεθόδου dispersion

Συνάρτηση DISPERSIONSCORE(*vector*,*node*,*depth*,*weight*)

if $depth > 0$ **then**

$depth \leftarrow depth - 1$

$neighbours \leftarrow neighbours(node)$

$numneighbours \leftarrow size(neighbours)$

$dispersion \leftarrow 1$

if $numneighbours \neq 0$ **then**

$dispersion \leftarrow \frac{1}{numneighbours}$

τέλος if

for *neighbour* *in* *neighbours* **do**

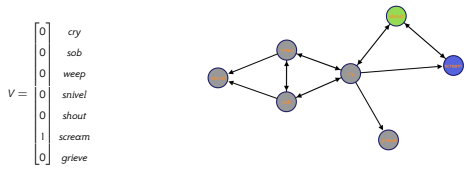
dispersionscore(*vector*,*neighbour*,*depth*,*dispersion*)

τέλος for

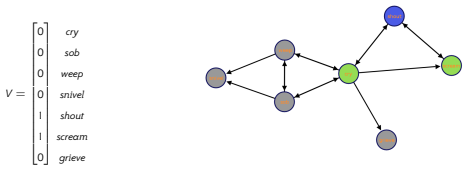
τέλος if

$vector[node] \leftarrow vector[node] + weight$

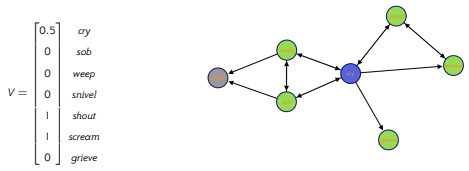
τέλος Συνάρτηση



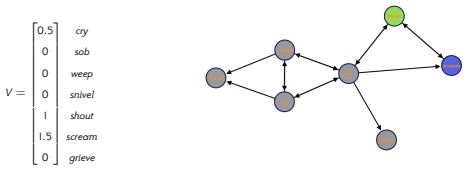
(α′) Βάθος 0



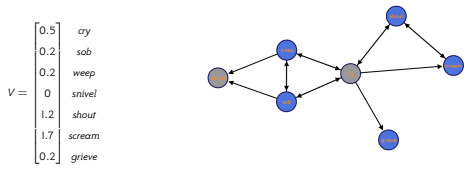
(β′) Βάθος 1



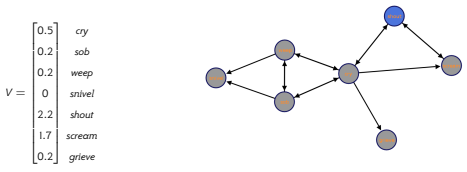
(γ′) Βάθος 2, φάση 1



(δ′) Βάθος 2, φάση 2



(ε′) Βάθος 3, φάσεις 1-5 μαζί



(στ′) Βάθος 3, φάση 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 dispersion -debug
scream
sob : 0.2 weep : 0.2 cry : 0.5 grieve : 0.2 shout : 2.2 scream : 1.7
```

(ζ′) Διάνυσμα συστήματος

Σχήμα 3.19: Παράδειγμα dispersion ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις.

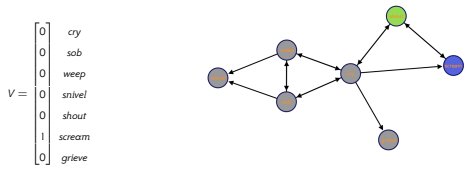
dispersion+

Τα διανύσματα είναι τύπου double. Κατά την διάσχιση του γράφου στο επιθυμητό βάθος σχηματίζεται βάρος $\frac{\text{εισερχόμενο βάρος}}{\text{πληθικότητα εξερχόμενων ακμών}}$ σε κάθε κόμβο, όπου το εισερχόμενο βάρος είναι αυτό που υπολογίστηκε στον κόμβο που διασχίστηκε στο προηγούμενο βήμα. Το βάρος αυτό προστίθεται στο υπάρχων βάρος καθενός γείτονα με τους οποίους ενώνεται με εξερχόμενες ακμές ο συγκεκριμένος κόμβος. Ο κόμβος για τον οποίο γίνεται η διάσχιση, δηλαδή ο πρώτος κόμβος που διασχίζεται ανατίθεται τιμή 1.

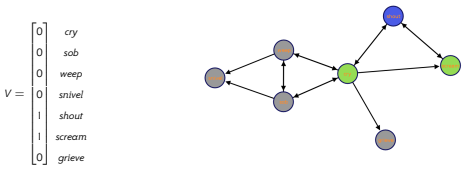
$$f(w_i) = \sum_{path}^{S_{word,depth}} score, \quad score = \begin{cases} 1 & \text{if} \quad path = P_{word,w_i} \quad \text{and} \quad O(w_i) = 0 \\ \prod_{(\alpha,b)}^{path} \frac{1}{O(\alpha)} & \text{else if} \quad path = P_{word,w_i} \\ 0 & \text{else} \end{cases} \tag{3.7}$$

Αλγόριθμος 3.6 Ψευδοκώδικας μεθόδου dispersion+

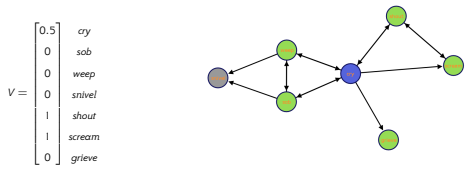
```
Συνάρτηση DISPERSION+SCORE(vector,node,depth,weight)
    if depth > 0 then
        depth ← depth − 1
        neighbours ← neighbours(node)
        numneighbours ← size(neighbours)
        dispersion ← weight
        if numneighbours ≠ 0 then
            dispersion ← weight / numneighbours
        τέλος if
        for neighbour in neighbours do
            dispersion+score(vector,neighbour,depth,dispersion)
        τέλος for
    τέλος if
    vector[node] ← vector[node] + weight
τέλος Συνάρτηση
```



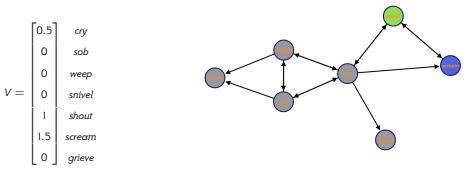
(α′) Βάθος 0



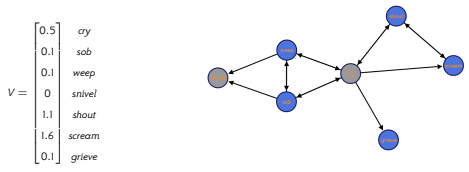
(β′) Βάθος 1



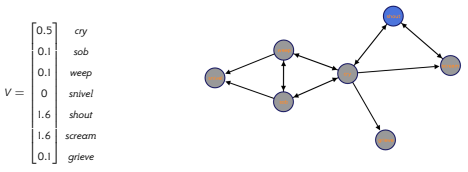
(γ′) Βάθος 2, φάση 1



(δ′) Βάθος 2, φάση 2



(ε′) Βάθος 3, φάσεις 1-5 μαζί



(στ′) Βάθος 3, φάση 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 dispersion+ -debug
scream
sob : 0.1 weep : 0.1 cry : 0.5 grieve : 0.1 shout : 1.6 scream : 1.6
```

(ζ′) Διάνυσμα συστήματος

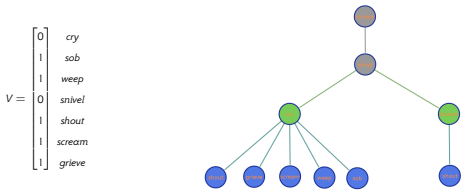
Σχήμα 3.20: Παράδειγμα dispersion+ ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις.

concentration+

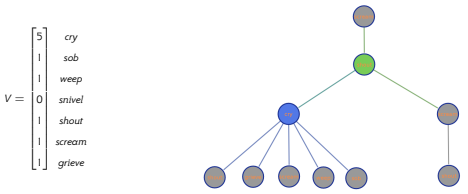
Αντίστροφη λογική της dispersion. Κάθε κόμβος στο τελευταίο επίπεδο επιθυμητού βάθους ανατίθεται βάρος 1. Στην συνέχεια για κάθε υψηλότερο επίπεδο προστίθονται οι τιμές του προηγούμενου επιπέδου μέχρι τη ρίζα και συγκεντρώνεται το βάρος κάθε κόμβου. Στην θέση του διανύσματος που αντιστοιχεί στην λέξη του κάθε κόμβου ανατίθεται το βάρος κανονικοποιημένο.

Αλγόριθμος 3.7 Ψευδοκώδικας μεθόδου concentration+

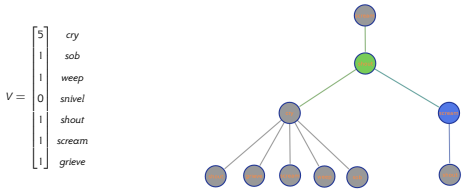
```
Συνάρτηση CONCENTRATION+SCORE(vector,node,depth,weight)
    if depth > 0 then
        depth ← depth − 1
        neighbours ← neighbours(node)
        numneighbours ← size(neighbours)
        if numneighbours = 0 then
            vector[node] ← weight
            return weight
        τέλος if
        sumdeeper ← 0
        for neighbour in neighbours do
            sumdeeper ← sumdeeper + concentration+score(vector,neighbour,depth,weight)
        τέλος for
        vector[node] ← sumdeeper
        return sumdeeper
    τέλος if
    vector[node] ← weight
    return weight
τέλος Συνάρτηση
```



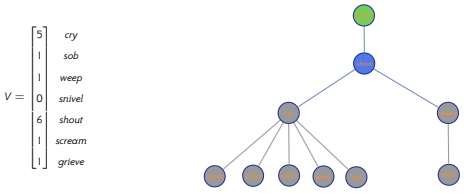
(α′) Βάθος 3, φάσεις 1-6 μαζί



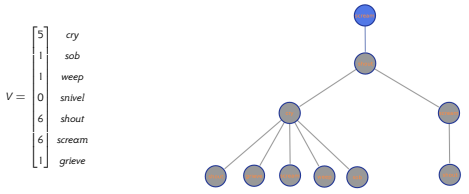
(β′) Βάθος 2, φάση 1



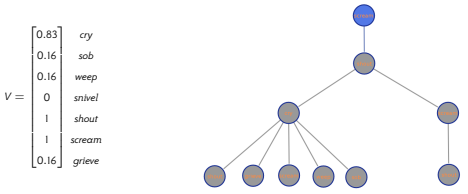
(γ′) Βάθος 2, φάση 2



(δ′) Βάθος 1



(ε′) Βάθος 0



(στ′) Κανονικοποίηση, διαιρώντας δια 6

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 concentration+ -debug
scream
sob : 0.166667 weep : 0.166667 cry : 0.833333 grieve : 0.166667 shout : 1 scream : 1
```

(ζ′) Διάνυσμα συστήματος

Σχήμα 3.21: Παράδειγμα concentration+ ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις. Ο γράφος μετασχηματίστηκε σε δενδροειδή μορφή με επανάληψη κόμβων για να φαίνεται πως ο αλγόριθμος πρακτικά ξεκινάει από τα "φύλλα".

robinhood

Κλέβει από τους πολλούς (δυνατούς) και μοιράζει στους φτωχούς (αδύναμους). Με την συγκεκριμένη μέθοδο δίνεται περισσότερο βάρος σε όσους κόμβους έχουν λίγους εξερχόμενους κόμβους, και λιγότερο σε όσους έχουν πολλούς.

Αλγόριθμος 3.8 Ψευδοκώδικας μεθόδου robinhood

Συνάρτηση ROBINHOODSCORE(*vector,node,depth,weight*)

if $depth > 0$ **then**

$depth \leftarrow depth - 1$

$neighbours \leftarrow neighbours(node)$

$numneighbours \leftarrow size(neighbours)$

$total \leftarrow 0$

for *neighbour* *in* *neighbours* **do**

$total \leftarrow total + numSubgraphNodes(neighbour, depth)$

τέλος for

$max \leftarrow max(1, \frac{total}{maxSubgraphNodes(neighbours,depth)})$

for *neighbour* *in* *neighbours* **do**

$outweight \leftarrow \frac{1}{numneighbours}$

$currentSubNodes \leftarrow numSubgraphNodes(neighbour, depth)$

if $currentSubNodes \neq 0$ **then**

$outweight \leftarrow \frac{total}{max \times currentSubNodes}$

τέλος if

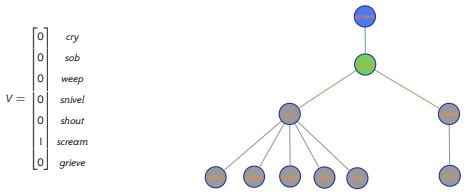
robinhoodscore(*vector,neighbour,depth,weight* $\times$ *outweight*)

τέλος for

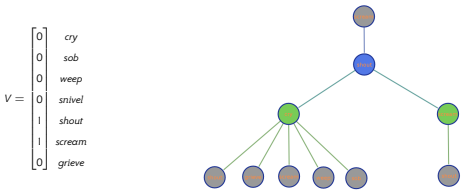
τέλος if

$vector[node] \leftarrow vector[node] + weight$

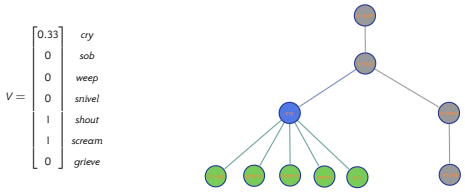
τέλος Συνάρτηση



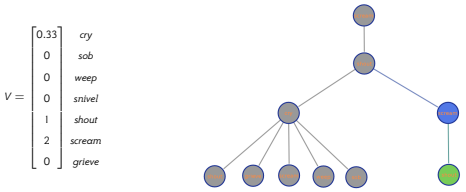
(α′) Βάθος 0



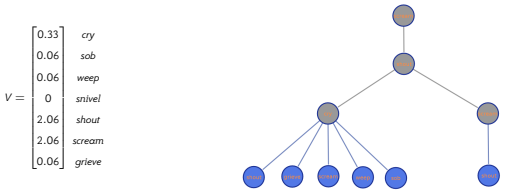
(β′) Βάθος 1



(γ′) Βάθος 2, φάση 1



(δ′) Βάθος 2, φάση 2



(ε′) Βάθος 3, φάσεις 1-6 μαζί

```
grv@Porcupine:~/wikinet$ ./wikinetword 3 robinhood -debug
scream
sob : 0.0666667 weep : 0.0666667 cry : 0.333333 grieve : 0.0666667 shout : 2.06667 scream : 2.06667
```

(στ′) Διάνυσμα συστήματος

Σχήμα 3.22: Παράδειγμα robinhood ξεκινώντας από τον κόμβο scream. Με μπλέ χρώμα απεικονίζεται η τωρινή κατάσταση. Με πράσινο οι επόμενες μεταβάσεις. Ο γράφος μετασχηματίστηκε σε δενδροειδή μορφή με επανάληψη κόμβων για να είναι εύκολο να μετρηθεί το πλήθος πιο βαθιών κόμβων.

3.5.3 Εύρεση ομοιότητας

Αφου κατασκευαστούν τα διανύσματα για τις λέξεις που μας ενδιαφέρουν με την μέθοδο και το βάθος που επιλέξαμε έρχεται η ώρα για τον υπολογισμό ομοιότητας ανάμεσα σε αυτά. Άλλωστε για αυτό μπήκαμε από την αρχή σε ολόκληρη αυτή τη διαδικασία.

Η προσέγγιση που χρησιμοποιήθηκε είναι η μέθοδος των συνημιτόνων (cosine similarity) η οποία ορίζεται ως:

$$cosine(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{3.8}$$

Όπου A,B διανύσματα ίδιου τύπου και μεγέθους.

Το αποτέλεσμα της μεθόδου των συνημιτόνων είναι στο διάστημα $[-1, 1]$ και φανερώνει αν τα διανύσματα δείχνουν προς την ίδια κατεύθυνση. Η ομοιότητα είναι 1 μεταξύ ίδιων διανυσμάτων η διανυσμάτων που ο λόγος των στοιχείων τους είναι ίδιος και σταθερός, 0 η ομοιότητα δύο καθέτων διανυσμάτων και -1 η ομοιότητα δύο αντίρροπων διανυσμάτων.

Στην περίπτωση μας ωστόσο δεν θα μας απασχολήσουν οι αρνητικές τιμές καθώς τα διανύσματά μας έχουν θετικά βάρη και συνεπώς οι τιμές ομοιότητας θα είναι στο $[0, 1]$.

Γενικά υπάρχουν αρκετές προσεγγίσεις για την εύρεση ομοιότητας ανάμεσα σε διανύσματα. Ωστόσο στην περίπτωση μας που ο γράφος είναι μεγάλος, η μέθοδος των συνημιτόνων (cosine similarity) έχει ένα θετικό. Αυτό είναι πως στον υπολογισμό της δεν λαμβάνονται καθόλου υπόψην οι μηδενικές τιμές που υπάρχουν στο διάνυσμα. Συνεπώς μπορούμε να διατηρούμε στην μνήμη αραιά διανύσματα [sparse vectors] με τις μη μηδενικές τιμές [4].

Σε περίπτωση που θέλουμε να βρούμε ομοιότητα φράσεων, η διαδικασία που ακολουθείται είναι η προσθήκη των διανυσμάτων των λέξεων που αποτελούν κάθε φράση και στην συνέχεια η εφαρμογή ομοιότητας συνημιτόνων σε αυτά τα διανύσματα.

[4] Η υλοποίηση των σποραδικών διανυσμάτων που έγινε στο σύστημα δεν ελέγχει την ανάθεση μηδενικής τιμής σε θέση διανύσματος, συνεπώς τα διανύσματα δεν είναι όσο αραιά γίνεται να είναι.

ΚΕΦΑΛΑΙΟ 4

Αποτελέσματα

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δούμε τα αποτελέσματα της αξιολόγησης των προσεγγίσεων και των ευθυγράμμισεων στο υποσύνολο του RTE2 A′. Επίσης, θα συζητήσουμε τα συμπεράσματα που προκύπτουν και θα αναφέρουμε τα μελλοντικά βήματα και τις διορθώσεις που πρέπει να γίνουν για να μπούν πιο γερές βάσεις στο σύστημα. που αναφέραμε στην ενότητα 3.4.1.

Αναλυτικά:

- Στις ενότητες 4.2.1, 4.2.2 και 4.2.3 θα δούμε τα αποτελέσματα της αξιολόγησης των προσεγγίσεων που αναλύσαμε στην ενότητα 3.5.2 με χρήση των συνόλων δεδομένων που αναλύσαμε στην ενότητα 3.4.1. Στην ενότητα 4.2.4 θα αναλύσουμε τα συμπεράσματα της αξιολόγησης.
- Στις ενότητες 4.3.1 και 4.4 θα εξετάσουμε τις ευθυγραμμίσεις που παρήγαγε το σύστημα σε επίπεδο λέξης και επίπεδο φράσεων αντίστοιχα.
- Στην ενότητα 4.5, θα κλείσουμε την πτυχιακή με μια σύνοψη συμπερασμάτων και προτάσεις για μελλοντικές βελτιώσεις.

4.2 Αξιολόγηση

4.2.1 Συνώνυμα

Με βάση το σύνολο δεδομένων συνωνύμων που κατασκευάσαμε στην ενότητα 3.4.1 από τις πηγές μας, θα παραθέσουμε εδώ τα στατιστικά της αξιολόγησης και θα προσπαθήσουμε να βγάλουμε ορισμένα συμπεράσματα για τις προσεγγίσεις έχοντας ως δεδομένο πως τα δεδομένα που έχουμε είναι καλά - σωστά. Θα δούμε πρώτα την αθροιστική προσέγγιση, στην συνέχεια τα

αποτελέσματα για την μεμονωμένη προσέγγιση με χρήση κατωφλίωσης και τέλος την στατιστική προσέγγιση με συσχέτιση Spearson's ρ.

Αθροιστική προσέγγιση

Σχηματίσαμε τον πίνακα 4.1 με τις επικαλύψεις των θετικών και αρνητικών δειγμάτων με σκοπό να διακρίνουμε τις μεθόδους που διαχωρίζουν περισσότερο τα θετικά από τα αρνητικά ζευγάρια όπως είδαμε στην ενότητα 3.4.2. Στον παρακάτω πίνακα η τιμή της αλληλοκάλυψης θέλουμε να είναι όσο το δυνατόν πιο μεγάλη γίνεται, δηλαδή στην περίπτωση μας που τα δείγματα είναι αρνητικά, όσο πιο κοντά στο 0 γίνεται.

μέθοδος	επικάλυψη
dispersion+ depth 2	-0.0633
concentration+ depth 1	-0.0660
dispersion depth 1	-0.0660
dispersion+ depth 1	-0.0660
robinhood depth 1	-0.0660
recursive depth 2	-0.0694
robinhood depth 2	-0.0764
concentration+ depth 2	-0.0773
recursive depth 1	-0.0802
countdepth depth 1	-0.0827
dispersion depth 2	-0.0839
boolean depth 1	-0.0891
count depth 1	-0.0891
countdepth depth 2	-0.1539
count depth 2	-0.1560
boolean depth 2	-0.1719

Πίνακας 4.1: Δεδομένα συνωνύμων. Επικάλυψη σε φθίνουσα σειρά

Μπορούμε καταρχάς να διακρίνουμε πως γενικά με την αύξηση του βάθους η επικάλυψη τείνει να αυξάνεται. Αυτό το χαρακτηριστικό είναι λογικό αν σκεφτεί κανείς πως από βάθος 2 και μετά οι λέξεις αρχίζουν και ξεφεύγουν στο νόημα και συνεπώς βρίσκεται ομοιότητα ανάμεσα σε μη όμοιες λέξεις. Επίσης, σε μεγάλο βάθος για τις προσεγγίσεις που δεν χρησιμοποιούν καθόλου στο βάρος την πληθικότητα των εξερχόμενων ακμών, αλλοιώνονται πολύ τα αποτελέσματα σε λέξεις που είναι γενικές ή έχουν πολλές διαφορετικές χρήσεις. Αυτό γίνεται καθώς έχουν πολλές εξερχόμενες ακμές και συνεπώς βρίσκεται επικάλυψη με άσχετες λέξεις πολύ πιο εύκολα.

Οι μέθοδοι boolean, count και countdepth παρατηρούμε γενικά ότι έχουν την ίδια μοίρα παρόλο που τα διανύσματα χτίζονται με αρκετά διαφορετικό τρόπο. Από όσο φαίνεται η διαίρεση με το βάθος δεν κάνει μεγάλη διαφορά και γενικά ο πειραματισμός με σταθερές που εφαρμόζωνται σε πλήθος κόμβων δεν φαίνεται να έχουν μεγάλη επιρροή σε αντίθεση με μεταβλητές που αφορούν κάθε κόμβο όπως η πληθικότητα των εξερχόμενων ακμών. Αυτη η παρατήρηση πιθανώς να έχει να κάνει και με την φύση της ομοιότητας συνημιτόνων (cosine similarity) καθώς εγγενώς δίνει έμφαση στον λόγο

- σχέση που έχουν οι τιμές του διανύσματος μεταξύ τους και όχι στο μέγεθος. Για παράδειγμα δύο διανύσματα που το ένα είναι ισούται με το άλλο πολλαπλασιασμένο με μια σταθερά έχουν ομοιότητα 1 με χρήση ομοιότητας συνημιτόνων.

Γενικά οι προσεγγίσεις dispersion+ και recursive φαίνονται σαν καλές επιλογές αν θέλουμε να κατασκευάσουμε το διάνυσμα πιο βαθιά στον γράφο. Το θετικό με την αύξηση του βάθους είναι πως μπορούμε να πιάσουμε μερικές φορές έννοιες οι οποίες, αν και δεν είναι άμεσα συνώνυμα, είναι σχετικές.

Τέλος σε βάθος 1 παρατηρούμε πως όλες οι μέθοδοι έχουν σχετικά χαμηλή επικάλυψη.

Μη Αθροιστική προσέγγιση με κατωφλίωση

Όπως αναφέραμε στην ενότητα 3.4.2 θα προσπαθήσουμε να βρούμε μια μέθοδο που διαλέγει με τον τρόπο που βαθμολογεί, ένα καλό υποψήφιο αρνητικό ζευγάρι που να χωρίζει καλά τα θετικά από τα αρνητικά δείγματα.

Στο σημείο αυτό επιβεβαιώνεται η παρατήρηση της προηγούμενης ενότητας πως οι μέθοδοι boolean, count και countdepth παρόλο που είναι αρκετά διαφορετικος ο τρόπος που αναθέτουν τα βάρη, έχουν παρόμοια συμπεριφορά και στην μεμονωμένη μελέτη. Παρουσιάζουν παρόμοια βαθμολογία σε ίδια βάθη και έχουν υψηλά κατώφλια που μεγαλώνουν καθώς μεγαλώνει το βάθος - άλλη μια ένδειξη πως σε μεγάλο βάθος τα διανύσματα στα οποία ορίζονται βάρη χωρίς να ληφθούν υπόψη τα χαρακτηριστικά του κόμβου όπως ο αριθμός εξερχόμενων ακμών, πνίγονται στον θόρυβο. Επίσης βρίσκουν κοινό κατώφλι - το sign - recess.

Η dispersion+ και η recursive φαίνονται να είναι οι πιο βάσιμες επιλογές μέχρι στιγμής, καθώς και στις δύο περιπτώσεις είναι καλές οι επιδόσεις της.

Αξιοσημείωτο είναι πως το ζευγάρι που λειτουργεί τις περισσότερες φορές ως κατώφλι πέρα από το sign - recess είναι το peace insurance, όμως λαμβάνει ένα σχετικά μεγάλο εύρος από τιμές ανάλογα με την προσέγγιση. Επίσης με την εξαίρεση της concentration+ τα κατώφλια έχουν την τάση να ανεβαίνουν όσο αυξάνεται το βάθος, πράγμα που σημαίνει πως λέξεις που δεν έχουν σχέση αρχίζουν και βρίσκουν κοινούς κόμβους στους οποίους δημιουργούνται επικαλύψεις.

μέθοδος	βαθμολογία	κατώφλι	ζευγάρι
dispersion depth 2	0.6836	0.0590	medium gain
dispersion+ depth 2	0.6328	0.0054	peace insurance
recursive depth 2	0.6328	0.0002	peace insurance
robinhood depth 2	0.6102	0.0153	peace insurance
robinhood depth 1	0.5763	0.0035	peace insurance
dispersion+ depth 1	0.5763	0.0035	peace insurance
dispersion depth 1	0.5763	0.0035	peace insurance
concentration+ depth 1	0.5763	0.0035	peace insurance
countdepth depth 1	0.5593	0.0799	sign recess
concentration+ depth 2	0.5424	0.0016	line insurance
count depth 1	0.5367	0.0823	sign recess
boolean depth 1	0.5367	0.0823	sign recess
countdepth depth 2	0.4746	0.3701	sign recess
count depth 2	0.4689	0.3797	sign recess
boolean depth 2	0.3785	0.3987	sign recess
recursive depth 1	0.2825	0.0000	king cabbage

Πίνακας 4.2: Δεδομένα συνωνύμων. Μέθοδοι σε σειρά φθίνουσας βαθμολογίας

Στατιστική συσχέτιση

Η στατιστική προσέγγιση με την συσχέτιση Pearson's ρ, όπως αναφέραμε και στην ενότητα 3.4.2 είναι ο καθιερωμένος τρόπος σύγκρισης αποτελεσμάτων στο wordsimilarity-353 σύνολο δεδομένων.

Από τις βαθμολογίες συσχέτισης που πέτυχαν οι προσεγγίσεις με την χρήση των συνωνύμων, βλέπουμε πως γενικά αποτυπώνεται κάποια ομοιότητα στην κατάταξη των ζευγαριών. Ωστόσο, οι μέγιστες βαθμολογίες σίγουρα χρειάζονται βελτίωση, και μέρος της βελτίωσης μπορεί να έρθει με την αύξηση του βάθους διάσχισης και την προσθήκη ορισμών.

Αυτό μπορεί να το δεί κανείς, καθώς στα δεδομένα του wordsimilarity-353 υπάρχουν λέξεις που έχουν υψηλή βαθμολογία χωρίς να είναι συνώνυμες. Έχουν απλά κοινά χαρακτηριστικά, που είναι πιθανό να μπορούν να εντοπιστούν αν αναλυθεί ο ορισμός τους.

Παρατηρούμε και εδώ πως οι προσεγγίσεις boolean, count και countdepth σε βάθη μεγαλύτερα από 1 εισάγουν πολύ θόρυβο, καθώς παρουσιάζεται σε βάθος 2 αισθητά μειωμένος βαθμός συσχέτισης.

μέθοδος	συσχέτιση
recursive depth 2	0.4487
robinhood depth 1	0.4425
dispersion+ depth 1	0.4425
dispersion depth 1	0.4425
concentration+ depth 1	0.4425
countdepth depth 1	0.4424
dispersion+ depth 2	0.4415
count depth 1	0.4388
boolean depth 1	0.4388
robinhood depth 2	0.4051
concentration+ depth 2	0.3881
dispersion depth 2	0.3739
count depth 2	0.2848
countdepth depth 2	0.2846
boolean depth 2	0.2415
recursive depth 1	0.0077

Πίνακας 4.3: Δεδομένα συνωνύμων. Συσχέτιση Spearman's ρ

Οι μέθοδοι recursive και dispersion+ φαίνεται να υπερισχύουν και εδώ σε μεγάλα βάθη, γεγονός που συμφωνεί με τις παρατηρήσεις μας με χρήση των προσεγγίσεων επικάλυψης και κατωφλίωσης.

4.2.2 Ορισμοί

Αντίστοιχα με τα συνώνυμα, σχηματίσαμε τον πίνακα 4.4 με τις αλληλοκαλύψεις από δημιουργία του γράφου με χρήση του συνόλου δεδομένων του οποίου περιγράψαμε την κατασκευή στην ενότητα 3.4.1.

Αθροιστική προσέγγιση

Μπορούμε να διακρίνουμε πως εδώ η γενική εικόνα είναι αρκετά διαφορετική όσον αφορά την επικάλυψη σε υψηλό βάθος. Γενικά οι μέθοδοι φαίνεται να πηγαίνουν χειρότερα στα μεγάλα βάθη και καλά στα μικρότερα. Αυτό έχει μια λογική, υπό την έννοια πως οι λέξεις στους ορισμούς σε μεγάλο βάθος θα είναι πιο πιθανό να χάνουν το νόημα τους. Γενικά, οι λέξεις στους ορισμούς, έχουν την τάση να σχολιάζουν χαρακτηριστικά του αντικειμένου, και συνεπώς η σχέση των λέξεων μπορεί να χαθεί πολύ πιο γρήγορα σε μεγάλα βάθη.

μέθοδος	επικάλυψη
recursive depth 2	-0.0610
dispersion+ depth 2	-0.0642
concentration+ depth 1	-0.0686
dispersion depth 1	-0.0688
dispersion+ depth 1	-0.0688
robinhood depth 1	-0.0688
concentration+ depth 2	-0.0745
countdepth depth 1	-0.0745
recursive depth 1	-0.0802
count depth 1	-0.1007
boolean depth 1	-0.1083
robinhood depth 2	-0.1955
dispersion depth 2	-0.2049
boolean depth 2	-0.2195
countdepth depth 2	-0.3034
count depth 2	-0.3082

Πίνακας 4.4: Δεδομένα ορισμών. Επικάλυψη σε φθίνουσα σειρά

Βλέπουμε πως η μέθοδος recursive αποδίδει εδώ, πράγμα λογικό καθώς ο λόγος κατασκευής της ήταν να μειώσει τον θόρυβο από λέξεις που είναι ασυσχέτιστες ήδη από μικρό βάθος. Όπως είναι λογικό με τον τρόπο που επιλέχθηκε να δημιουργηθεί το σύνολο δεδομένων των ορισμών, θα υπάρχουν αρκετές λέξεις που δεν θα σχετίζονται με την λέξη στης οποίας τον ορισμό ανήκει. Συνεπώς, η recursive μειώνει λίγο τον θόρυβο που προέρχεται από αυτήν την πηγή.

Και εδώ η dispersion+ είναι ανάμεσα σε αυτές που αποδίδουν καλά, παρόλο που τα δεδομένα είναι άλλου είδους. Παρατηρούμε πως όσες μέθοδοι αποδίδουν καλά με χρήση ορισμών επιτυγχάνουν χαμηλότερες συσχετίσεις από ότι με την χρήση συνωνύμων.

Τέλος, αξίζει να παρατηρηθεί η διαφορά στην επικάλυψη των μεθόδων που δεν πήγαν καλά. Σε σχέση με τα συνώνυμα, βλέπουμε πως η επικάλυψη για τις boolean, count και countdepth σε βάθος 2 είναι σχεδόν διπλάσια από αυτή που είχαν στα συνώνυμα. Το γεγονός αυτό μας δείχνει πως για τις προσεγγίσεις που έχουν χαμηλή ανοχή στον θόρυβο, οι ορισμοί είναι χειρότερο σύνολο δεδομένων από ότι τα συνώνυμα σε περίπτωση διάσχισης σε μεγάλα βάθη.

Μη Αθροιστική προσέγγιση με κατωφλίωση

Το πρώτο πράγμα που παρατηρούμε για την περίπτωση της κατωφλίωσης με χρήση των δεδομένων ορισμών είναι η μεγαλύτερη ποικιλία στα κατώφλια που επιλέγονται από τις διαφορετικές μεθόδους σε σχέση με την περίπτωση των συνωνύμων.

Παρατηρούμε πως σε βάθος 2 πολλά κατώφλια είναι πάνω από 0.40. Μάλιστα, η μέθοδος count σε βάθος 2 έχει κατώφλι στο stock life με βαθμολογία 0.70! Επιβεβαιώνονται συνεπώς οι παρατηρήσεις που έγιναν στην εξέταση των επικαλύψεων. Οι μέθοδοι boolean, count και countdepth είναι ακατάλληλες για μεγάλα βάθη.

Όπως και στην αθροιστική προσέγγιση, η recursive και η dispersion+ τα πάνε καλά σε μέτριο βάθος, ενώ γενικά οι βαθμολογίες είναι πιο χαμηλές από την περίπτωση χρήσης των συνωνύμων.

μέθοδος	βαθμολογία	κατώφλι	ζευγάρι
recursive depth 2	0.5819	0.0130	opera industry
countdepth depth 1	0.5650	0.1250	cemetery woodland
concentration+ depth 2	0.4972	0.0094	holy sex
dispersion+ depth 2	0.4746	0.0553	cemetery woodland
robinhood depth 1	0.4633	0.0500	cemetery woodland
dispersion+ depth 1	0.4633	0.0500	cemetery woodland
dispersion depth 1	0.4633	0.0500	cemetery woodland
concentration+ depth 1	0.4633	0.0500	cemetery woodland
count depth 1	0.4068	0.2000	cemetery woodland
boolean depth 2	0.4011	0.4739	direction combination
boolean depth 1	0.3955	0.2000	cemetery woodland
count depth 2	0.3503	0.7068	stock life
countdepth depth 2	0.3503	0.6941	stock life
robinhood depth 2	0.3390	0.4455	medium gain
dispersion depth 2	0.3277	0.4882	medium gain
recursive depth 1	0.2825	0.0000	king cabbage

Πίνακας 4.5: Δεδομένα ορισμών. Μέθοδοι σε σειρά φθίνουσας βαθμολογίας

Επίσης οι βαθμολογίες των μεθόδων που τα πήγαν καλά παρατηρούμε πως είναι ελαφρώς χαμηλότερες από τις αντίστοιχες με χρήση συνωνύμων. Συνεπώς, η χρήση ορισμών ως σύνολο δεδομένων, είναι χειρότερη επιλογή σε περίπτωση που θέλει κανείς να πάρει δυαδική απόφαση

για την ομοιότητα ή μη δύο λέξεων.

Στατιστική συσχέτιση

Την παραπάνω θέση ενισχύει και η παρατήρηση πως τα κατώφλια είναι γενικά μεγαλύτερα σε σχέση με τα αντίστοιχα με την χρήση συνωνύμων. Αυτό σημαίνει πως ίδιες μέθοδοι όταν χρησιμοποιούν ορισμούς αντί για συνώνυμα, βγάζουν μεγαλύτερες βαθμολογίες σε αρνητικά ζευγάρια.

Οι βαθμολογίες συσχετίσεων ωστόσο, σε αντίθεση με τις βαθμολογίες από την κατωφλίωση δείχνουν πως οι ορισμοί είναι πιο κατάλληλη από τα συνώνυμα ως προς την κατάταξη των ζευγαριών. Αυτό ισχύει φυσικά για τις μεθόδους που δείχνουν καλά αποτελέσματα, καθώς οι υπόλοιπες δεν παρουσίαζουν μεγάλες διαφόρες.

Σημειώνουμε επίσης πως κάνει την εμφάνιση της ψηλά στις βαθμολογίες συσχέτισης και η μέθοδος concentration+ σε βάθος 2. Παρατηρούμε πως ήταν ψηλά και στις βαθμολογίες κατωφλίωσης, ωστόσο είχε σχετικά υψηλή βαθμολογία επικάλυψης.

Συνεπώς, με βάση τα παραπάνω μπορούμε να βγάλουμε ένα συμπέρασμα. Με χρήση ορισμών σε σχέση με τη χρήση συνωνύμων, έχουμε καλύτερες επιδόσεις όσον αφορά την επικάλυψη και την συσχέτιση, ενώ έχουμε χειρότερη όσον αφορά την κατωφλίωση.

μέθοδος	συσχέτιση
dispersion+ depth 2	0.4801
concentration+ depth 2	0.4537
robinhood depth 1	0.4399
dispersion+ depth 1	0.4399
dispersion depth 1	0.4399
concentration+ depth 1	0.4396
recursive depth 2	0.4390
countdepth depth 1	0.4292
count depth 1	0.4140
boolean depth 1	0.3904
robinhood depth 2	0.2605
dispersion depth 2	0.1875
boolean depth 2	0.0983
countdepth depth 2	0.0856
count depth 2	0.0721
recursive depth 1	0.0077

Πίνακας 4.6: Δεδομένα ορισμών. Συσχέτιση Spearman's ρ

4.2.3 Συνώνυμα + ορισμοί

Παρατηρήσαμε με την χρήση συνωνύμων και την χρήση ορισμών πως τα αποτελέσματα έχουν διαφορετικά χαρακτηριστικά.

Φαίνεται πως οι προσεγγίσεις που δίνουν καλά αποτελέσματα, έχουν υψηλότερες βαθμολογίες συσχέτισης και επικάλυψης με τη χρήση ορισμών, ενώ καλύτερες βαθμολογίες σημειώνονται στην περίπτωση της κατωφλίωσης με τη χρήση συνωνύμων.

Λογικό ήταν συνεπώς να φτιάξουμε ένα σύνολο δεδομένων όπως περιγράψαμε στην ενότητα 3.4.1 που να περιέχει και συνώνυμα και ορισμούς, για να δούμε τι αποτελέσματα θα έχουμε.

Το λογικό που θα περίμενε κανείς να δει, είναι μια μέση εικόνα στα αποτελέσματα. Θα περίμενε να δει επιδόσεις στην κατωφλίωση λίγο πιο ψηλές από αυτές των ορισμών, και επιδόσεις στην επικάλυψη και την συσχέτιση λίγο καλύτερες από αυτές των συνωνύμων.

Επίσης, λογικό θα ήταν να περιμένει κανείς και την προσθήκη θορύβου, καθώς το σύνολο δεδομένων με την συνάθροιση ορισμών και συνωνύμων βγαίνει περίπου διπλάσιο. Η λογική είναι πως ειδικά σε μεθόδους που είναι ευαίσθητες στον θόρυβο όπως είναι η boolean, η count και η countdepth, η προσθήκη τόσης παραπάνω πληροφορίας στον γράφο θα επεκτείνει περαιτέρω το πρόβλημα αυτό.

Αθροιστική προσέγγιση

Όσον αφορά τις επικαλύψεις για την χρήση του συνόλου δεδομένων συνωνύμων + ορισμών, με έκπληξη διαπιστώνουμε πως είναι χαμηλότερες! Για μεθόδους που σημείωσαν χαμηλές επικαλύψεις, η βαθμολογία επικάλυψης είναι χαμηλότερη και από την χρήση συνωνύμων αλλά και από την χρήση ορισμών. Αυτό είναι πολύ θετικό, αλλά περίεργο, καθώς κανείς θα ανησυχούσε πως όταν αυξηθεί πολύ το μέγεθος του γράφου θα αυξανόταν και ο θόρυβος.

Βλέπουμε ως προς την επικάλυψη, πως προσεγγίσεις που πήγαιναν καλά και στα επιμέρους σύνολα δεδομένων, τα πάνε καλά και στο ενιαίο. Ωστόσο, πλέον φαίνεται να ξεχωρίζει η dispersion+.

Μέθοδοι όπως η boolean, η count και η countdepth αποδίδουν πάλι άσχημα σε βάθος 2, και έχουν τιμές επικάλυψης ανάμεσα σε αυτές που είχαν στα επιμέρους σύνολα δεδομένων.

Μη Αθροιστική προσέγγιση με κατωφλίωση

Και στην περίπτωση της κατωφλίωσης εκτελίσσεται το ίδιο φαινόμενο. Και εδώ οι βαθμολογίες όσων μεθόδων τα πήγαιναν καλά και στα επιμέρους σύνολα δεδομένων, τα πάνε καλύτερα. Συνεπώς μπορούμε να συμπεράνουμε πως η προσθήκη των ορισμών στα συνώνυμα τελικά βοηθάει στον καλύτερο διαχωρισμό των δεδομένων. Πιθανά αυτό σημαίνει πως αρκετές σχέσεις για να βρεθούν, δεν ήταν αρκετό να έχει κανείς συνώνυμα.

Όσων αφορά τα κατώφλια, βλέπουμε πως και εδώ όπως και με την περίπτωση των ορισμών, πως μερικές προσεγγίσεις έχουν πολύ υψηλά. Αυτό γίνεται κυρίως σε βάθος 2 για τις μεθόδους boolean, count και countdepth, πράγμα που επιβεβαιώνει πως δεν είναι κατάλληλες για χρήση σε

μέθοδος	επικάλυψη
dispersion+ depth 2	-0.0573
concentration+ depth 1	-0.0625
dispersion depth 1	-0.0626
dispersion+ depth 1	-0.0626
robinhood depth 1	-0.0626
countdepth depth 1	-0.0645
recursive depth 2	-0.0676
count depth 1	-0.0718
boolean depth 1	-0.0760
concentration+ depth 2	-0.0790
recursive depth 1	-0.0802
robinhood depth 2	-0.0963
dispersion depth 2	-0.1148
boolean depth 2	-0.1827
countdepth depth 2	-0.2169
count depth 2	-0.2199

(α′) Δεδομένα συνωνύμων + ορισμών. Επικάλυψη σε φθίνουσα σειρά

μέθοδος	συσχέτιση
dispersion+ depth 2	0.5676
recursive depth 2	0.5362
robinhood depth 1	0.5260
dispersion+ depth 1	0.5260
dispersion depth 1	0.5260
concentration+ depth 1	0.5257
concentration+ depth 2	0.5115
countdepth depth 1	0.5062
count depth 1	0.4905
boolean depth 1	0.4764
robinhood depth 2	0.3996
dispersion depth 2	0.3828
countdepth depth 2	0.2475
count depth 2	0.2403
boolean depth 2	0.1664
recursive depth 1	0.0077

(β′) Δεδομένα συνωνύμων + ορισμών. Συσχέτιση Spearman's ρ

μεγάλο βάθος.

Στατιστική συσχέτιση

Όσον αφορά την συσχέτιση, παρατηρούμε και εδώ την ίδια τάση. Και εδώ τα αποτελέσματα για τις αποτελεσματικές μεθόδους έχουν σαφέστατη βελτίωση.

Μπορούμε να συμπεράνουμε συνεπώς πως η χρήση του συνδυαστικού συνόλου δεδομένων καταλήγει σε πιο πλήρη αποτελέσματα. Αυτό σημαίνει πως υπάρχουν ζευγάρια τα οποία δεν μπορεί να βαθμολογήσει καλά ως προς την ομοιότητα αν αρκεστεί μόνο σε συνώνυμα ή σε ορισμούς.

Τέλος, και εδώ για ακόμα μια φορά φαίνεται πως οι μέθοδοι boolean, count και countdepth έχουν παρουσιάζουν χαμηλή συσχέτιση σε βάθος 2.

μέθοδος	βαθμολογία	κατώφλι	ζευγάρι
dispersion+ depth 2	0.7627	0.0048	cemetery woodland
recursive depth 2	0.7345	0.0004	opera industry
robinhood depth 1	0.7232	0.0022	cemetery woodland
dispersion+ depth 1	0.7232	0.0022	cemetery woodland
dispersion depth 1	0.7232	0.0022	cemetery woodland
concentration+ depth 1	0.7232	0.0022	cemetery woodland
concentration+ depth 2	0.6667	0.0005	lad wizard
countdepth depth 1	0.6610	0.0739	sign recess
count depth 1	0.6328	0.0768	sign recess
boolean depth 1	0.6102	0.0786	sign recess
dispersion depth 2	0.4802	0.3194	medium gain
count depth 2	0.4237	0.6085	sign recess
countdepth depth 2	0.4237	0.6004	sign recess
robinhood depth 2	0.3955	0.1558	stock egg
boolean depth 2	0.3729	0.4535	sign recess
recursive depth 1	0.2825	0.0000	king cabbage

Πίνακας 4.7: Δεδομένα συνωνύμων + ορισμών. Μέθοδοι σε σειρά φθίνουσας βαθμολογίας

4.2.4 Συμπεράσματα

Από την παραπάνω ανάλυση ως προς τις μεθόδους και την χρήση των συνωνύμων, των ορισμών και του συνδυασμού τους ως σύνολα δεδομένων, μπορεί κανείς να προβεί στα παρακάτω συμπεράσματα.

Καταρχάς, η χρήση του συνδυασμού των συνωνύμων και ορισμών ως σύνολο δεδομένων δεν φαίνεται να χειροτερεύει τα αποτελέσματα με κανέναν από τους τρείς τρόπους αξιολόγησης. Το μόνο αρνητικό στοιχείο χρήσης του που μπορεί να επισημειωθεί σε σχέση με τα επιμέρους σύνολα δεδομένων είναι ο χρόνος επεξεργασίας του γράφου λόγω του μεγέθους του.

Επίσης, φαίνεται πως και οι 3 τρόποι αξιολόγησης συμφωνούν σε αρκετά μεγάλο βαθμό. Ειδικά η σχέση της επικάλυψης με τον βαθμό συσχέτισης φαίνεται να έχει αρκετά κοινά, καθώς και στις 3 περιπτώσεις προσεγγίσεις που είχαν μικρή επικάλυψη είχαν μεγάλη συσχέτιση και το αντίστροφο. Η περίπτωση της κατωφλίωσης όπως είδαμε από την περίπτωση των ορισμών μερικές φορές δεν συμβαδίζει με τις άλλες 2. Φαίνεται να είναι πιο άπληστη και μη γενική.

Γενικά οι προσεγγίσεις που φαίνεται να αποδίδουν καλύτερα είναι η dispersion+ και h recursive, ενώ θετικά στοιχεία δείχνει σε μερικές περιπτώσεις η concentration+.

Οι προσεγγίσεις boolean, count και countdepth δεν ενδείκνυνται για χρήση σε βάθος μεγαλύτερο από 1, καθώς εισάγουν θόρυβο - βρίσκουν υψηλές ομοιότητες ακόμα και ανάμεσα σε μη σχετικές λέξεις.

Καλύτερη βαθμολογία επικάλυψης πέτυχε η dispersion+ σε βάθος 2 με 0.0573. Καλύτερη βαθμολογία κατωφλίωσης η dispersion+ σε βάθος 2 με 0.7627. Τέλος καλύτερη βαθμολογία συσχέτισης πέτυχε πάλι η dispersion+ σε βάθος 2 με 0.5676

4.3 Ευθυγράμμιση

Σε αυτήν την ενότητα θα δούμε τα αποτελέσματα εφαρμογής του συστήματος στην ευθυγράμμιση φράσεων από το RTE2. Αρχικά θα δούμε την ευθυγράμμιση σε επίπεδο λέξης και στην συνέχεια σε επίπεδο φράσεων.

Η διαδικασία της ευθυγράμμισης όπως αναλύσαμε στην σχετική βιβλιογραφία στην ενότητα 2 αφορά την αντιστοίχιση σχετικών λέξεων ή φράσεων ανάμεσα σε δύο κείμενα. Στην υλοποίηση μας, το σύστημα βαθμολογεί την ομοιότητα ανάμεσα σε λέξεις ή συντακτικούς πλίνθους [chunks] των δύο κειμένων μια προς μια. Στην συνέχεια θεωρεί πως υπάρχει ευθυγράμμιση αν η τιμή της ομοιότητας είναι μεγαλύτερη από αυτή που βρέθηκε ως κατώφλι στην περασμένη ενότητα.

Για διευκόλυνση στην οπτικοποίηση των αποτελεσμάτων, εμφανίζεται με πορτοκαλί οποια-

δήποτε ευθυγράμμιση έχει τιμή από την τιμή κατωφλίωσης μέχρι και τρείς φορές την τιμή κατωφλίωσης. Με κίτρινο εμφανίζονται αντίστοιχα τιμές ομοιότητας στο εύρος από τρείς μέχρι 10 φορές την τιμή κατωφλίωσης. Με πράσινο εμφανίζονται όποιες τιμές είναι μεγαλύτερες από 10 φορές την τιμή κατωφλίωσης. Η οπτικοποίηση αυτή εκφράζει κατά έναν τρόπο την βεβαιότητα για την ευθυγράμμιση.

Η προσέγγιση που χρησιμοποιείται σε κάθε περίπτωση είναι η dispersion+ σε βάθος 2.

4.3.1 Ευθυγράμμιση λέξεων

Στην προηγούμενη ενότητα είδαμε πως σύμφωνα με τα δεδομένα του similarity-353, καλύτερες αναθέσεις ομοιότητας γίνονται με την μέθοδο dispersion+ σε βάθος 2 με χρήση του συνδιαστικού συνόλου δεδομένων. Λογικά, θα πρέπει και οι ευθγραμμίσεις που πετυχαίνει να είναι καλές.

Συνώνυμα + ορισμοί

Σε αυτό το σημείο ας δούμε μερικά παραδείγματα για το πως τα πάει το σύστημα μας στην ευθυγράμμιση λέξη προς λέξη.

Όπως βλέπουμε στο πρώτο από τα παραδείγματα που ακολουθούν, υπάρχουν αρκετά θέματα με τις ευθυγραμμίσεις. Καταρχάς έχει ευθυγραμμίσει ορισμένες λέξεις που είναι εντελώς άσχετες - το two με το be για παράδειγμα. Επίσης έχει ευθυγραμμίσει το be με το source, γεγονός που φαίνεται και αυτό προβληματικό.

Και στην περίπτωση του πίνακα 4.9, το σύστημα φαίνεται να πιάνει αρκετές ομοιότητες οι οποίες δεν είναι σωστές. Για παράδειγμα βρίσκει μεγάλη ομοιότητα ανάμεσα στο 30, το four και το cost με το be.

Παρατηρούμε σε αυτό το σημείο πως το πρόβλημα φαίνεται να δημιουργείται σε γενικές λέξεις. Το σύνολο δεδομένων wordsim-353 γενικά δεν είχε πολύ κοινές και γενικές λέξεις όπως το be, και συνεπώς αυτός ο παράγοντας δεν εξετάστηκε ως προς την αξιολόγηση, κακώς. Ωστόσο, παρατηρούμε πως έπιασε ομοιότητα ανάμεσα στο sheet και το paper, γεγονός ενθαρρυντικό.

Στο ζεύγος 4.10 πάλι βλέπουμε το ίδιο μοτίβο να συνεχίζεται. Το σύστημα βρίσκει σωστά ομοιότητες όπως του civilian με το people, όμως παράλληλα βρίσκει υψηλή και με το be.

Πλέον το γεγονός πως το σύστημα έχει την τάση να βρίσκει πάντα ομοιότητες ανάμεσα σε γενικές και συχνά χρησιμοποιούμενες λέξεις είναι ολοφάνερο αν ρίξει κανείς μια ματιά και στον πίνακα 4.11. Μήπως να δοκιμάσουμε να δούμε τι γίνεται στην περίπτωση χρήσης συνωνύμων;

	0 two	1 people	2 be	3 wound	4 by	5 a	6 bomb	7 .
police	0.0003	0.0018	0.0017	0.0006	0.0004	0.0009	0.0007	0.0000
source	0.0013	0.0030	0.0189	0.0017	0.0015	0.0009	0.0013	0.0000
state	0.0028	0.0214	0.0139	0.0011	0.0074	0.0023	0.0015	0.0000
that	0.0002	0.0002	0.0032	0.0001	0.0026	0.0003	0.0001	0.0000
during	0.0009	0.0001	0.0036	0.0002	0.0106	0.0048	0.0001	0.0000
the	0.0000	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bomb	0.0006	0.0006	0.0041	0.0006	0.0004	0.0003	1.0000	0.0000
attack	0.0016	0.0011	0.0029	0.0018	0.0009	0.0006	0.0200	0.0000
involve	0.0016	0.0023	0.0063	0.0062	0.0011	0.0007	0.0004	0.0000
the	0.0000	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
shining	0.0017	0.0008	0.0141	0.0009	0.0006	0.0005	0.0008	0.0000
path	0.0006	0.0010	0.0016	0.0016	0.0021	0.0006	0.0004	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
two	1.0000	0.0019	0.0457	0.0012	0.0014	0.0037	0.0006	0.0000
people	0.0019	1.0000	0.0146	0.0017	0.0005	0.0015	0.0006	0.0000
be	0.0457	0.0146	1.0000	0.0091	0.0027	0.0040	0.0041	0.0000
injure	0.0005	0.0006	0.0019	0.0342	0.0004	0.0002	0.0007	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.8: Ευθυγράμμιση λέξεων Ζεύγους 137, χρήση συνωνύμων και ορισμών

	0 the	1 cost	2 of	3 paper	4 be	5 rise	6 .
the	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
cost	0.0000	1.0000	0.0003	0.0012	0.0254	0.0065	0.0000
for	0.0000	0.0003	0.0101	0.0002	0.0000	0.0000	0.0000
some	0.0000	0.0013	0.0056	0.0008	0.0077	0.0009	0.0000
30	0.0000	0.0024	0.0004	0.0016	0.0891	0.0014	0.0000
million	0.0000	0.0005	0.0004	0.0010	0.0037	0.0010	0.0000
sheet	0.0000	0.0009	0.0004	0.0290	0.0045	0.0012	0.0000
of	0.0000	0.0003	1.0000	0.0002	0.0013	0.0004	0.0000
paper	0.0000	0.0012	0.0002	1.0000	0.0034	0.0007	0.0000
use	0.0000	0.0048	0.0112	0.0048	0.0198	0.0020	0.0000
each	0.0000	0.0009	0.0012	0.0018	0.0064	0.0015	0.0000
year	0.0000	0.0017	0.0004	0.0019	0.0035	0.0010	0.0000
by	0.0000	0.0018	0.1581	0.0007	0.0027	0.0009	0.0000
cal	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
state	0.0000	0.0023	0.0022	0.0109	0.0139	0.0022	0.0000
long	0.0000	0.0017	0.0002	0.0005	0.0115	0.0052	0.0000
beach	0.0000	0.0005	0.0002	0.0005	0.0027	0.0010	0.0000
college	0.0000	0.0006	0.0002	0.0008	0.0047	0.0009	0.0000
and	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
department	0.0000	0.0012	0.0009	0.0012	0.0030	0.0005	0.0000
go	0.0000	0.0031	0.0004	0.0012	0.0180	0.0080	0.0000
up	0.0000	0.0012	0.0007	0.0004	0.0147	0.0043	0.0000
wednesday	0.0000	0.0003	0.0001	0.0020	0.0027	0.0004	0.0000
for	0.0000	0.0003	0.0101	0.0002	0.0000	0.0000	0.0000
the	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
first	0.0000	0.0013	0.0006	0.0007	0.0049	0.0017	0.0000
time	0.0000	0.0158	0.0007	0.0012	0.0132	0.0019	0.0000
in	0.0000	0.0007	0.0333	0.0005	0.0066	0.0008	0.0000
four	0.0000	0.0014	0.0012	0.0010	0.0287	0.0014	0.0000
year	0.0000	0.0017	0.0004	0.0019	0.0035	0.0010	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.9: Ευθυγράμμιση λέξεων Ζεύγους 43, χρήση συνωνύμων και ορισμών

	0 four	1 people	2 be	3 assassinate	4 by	5 the	6 pilot	7 .
in	0.0011	0.0013	0.0066	0.0004	0.0028	0.0000	0.0006	0.0000
that	0.0002	0.0002	0.0032	0.0001	0.0026	0.0000	0.0001	0.0000
aircraft	0.0005	0.0100	0.0034	0.0004	0.0005	0.0000	0.0131	0.0000
accident	0.0011	0.0006	0.0055	0.0011	0.0004	0.0000	0.0004	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
four	1.0000	0.0022	0.0287	0.0013	0.0008	0.0000	0.0005	0.0000
people	0.0022	1.0000	0.0146	0.0023	0.0005	0.0005	0.0013	0.0000
be	0.0287	0.0146	1.0000	0.0032	0.0027	0.0000	0.0029	0.0000
kill	0.0010	0.0014	0.0072	0.0274	0.0005	0.0000	0.0006	0.0000
:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
the	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
pilot	0.0005	0.0013	0.0029	0.0007	0.0010	0.0000	1.0000	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
who	0.0034	0.0005	0.0024	0.0004	0.0005	0.0000	0.0004	0.0000
be	0.0287	0.0146	1.0000	0.0032	0.0027	0.0000	0.0029	0.0000
wear	0.0015	0.0034	0.0031	0.0015	0.0015	0.0000	0.0008	0.0000
civilian	0.0018	0.0614	0.0493	0.0017	0.0007	0.0000	0.0013	0.0000
clothe	0.0008	0.0013	0.0016	0.0012	0.0004	0.0000	0.0008	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
and	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
three	0.0742	0.0030	0.0296	0.0006	0.0009	0.0000	0.0003	0.0000
other	0.0023	0.0017	0.0093	0.0004	0.0041	0.0000	0.0014	0.0000
people	0.0022	1.0000	0.0146	0.0023	0.0005	0.0005	0.0013	0.0000
who	0.0034	0.0005	0.0024	0.0004	0.0005	0.0000	0.0004	0.0000
be	0.0287	0.0146	1.0000	0.0032	0.0027	0.0000	0.0029	0.0000
wear	0.0015	0.0034	0.0031	0.0015	0.0015	0.0000	0.0008	0.0000
military	0.0003	0.0064	0.0026	0.0007	0.0013	0.0000	0.0016	0.0000
uniform	0.0009	0.0016	0.0049	0.0003	0.0005	0.0000	0.0007	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.10: Ευθυγράμμιση λέξεων Ζεύγους 404, χρήση συνωνύμων και ορισμών

	0 organic	1 fertilizer	2 be	3 use	4 a	5 soil	6 enhancer	7 .
organic	1.0000	0.0266	0.0180	0.0075	0.0027	0.0053	0.0012	0.0000
fertilizer	0.0266	1.0000	0.0027	0.0857	0.0017	0.0778	0.0009	0.0000
slowly	0.0008	0.0081	0.0016	0.0921	0.0014	0.0014	0.0002	0.0000
enrich	0.0018	0.0082	0.0050	0.0025	0.0003	0.0022	0.0083	0.0000
and	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
feed	0.0019	0.0149	0.0039	0.0207	0.0007	0.0017	0.0008	0.0000
the	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
soil	0.0053	0.0778	0.0170	0.0138	0.0005	1.0000	0.0032	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
fast	0.0005	0.0007	0.0044	0.0060	0.0003	0.0005	0.0010	0.0000
act	0.0022	0.0021	0.0127	0.0125	0.0011	0.0048	0.0036	0.0000
synthetic	0.0125	0.0051	0.0069	0.0035	0.0005	0.0014	0.0010	0.0000
fertilizer	0.0266	1.0000	0.0027	0.0857	0.0017	0.0778	0.0009	0.0000
harm	0.0027	0.0012	0.0105	0.0024	0.0003	0.0033	0.0022	0.0000
soil	0.0053	0.0778	0.0170	0.0138	0.0005	1.0000	0.0032	0.0000
life	0.0092	0.0012	0.0273	0.0026	0.0007	0.0019	0.0019	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.11: Ευθυγράμμιση λέξεων Ζεύγους 192, χρήση συνωνύμων και ορισμών

Συνώνυμα

Στην περίπτωση χρήσης συνωνύμων βλέπουμε πως τα πράγματα είναι πολύ πιο ομαλά. Δεν φαίνεται να υπάρχει τόσο θόρυβος σε ευθυγραμμίσεις όπου η μια από τις δύο λέξεις είναι γενική.

Στον πίνακα 4.12 φαίνεται μια γενικά έγκυρη ευθυγράμμιση.

Στην περίπτωση του 4.13 με την εξαίρεση των ευθυγραμμίσεων by - of και in - of, πάλι τα πράγματα φαίνεται να είναι πιο ομαλά.

Και στους πίνακες 4.14 και 4.15, οι ευθυγραμμίσεις του civilian με το people και του feed με το fertilizer νοηματικά δεν είναι λανθασμένες. Σε αυτό το σημείο μάλλον θα είχε νόημα να δούμε πιο πολύ σε βάθος άλλα κριτήρια για τις ευθυγραμμίσεις πέρα από την νοηματική ομοιότητα.

Όπως και να έχει, η χρήση συνωνύμων φάνηκε να δημιουργεί καλύτερες ευθυγραμμίσεις από το σύνολο συνωνύμων + ορισμών.

	0 two	1 people	2 be	3 wound	4 by	5 a	6 bomb	7 .
police	0.0000	0.0002	0.0005	0.0001	0.0001	0.0008	0.0001	0.0000
source	0.0001	0.0019	0.0012	0.0004	0.0001	0.0001	0.0002	0.0000
state	0.0001	0.0027	0.0014	0.0003	0.0010	0.0000	0.0002	0.0000
that	0.0000	0.0000	0.0000	0.0000	0.0052	0.0001	0.0000	0.0000
during	0.0000	0.0000	0.0001	0.0002	0.0162	0.0093	0.0001	0.0000
the	0.0000	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
bomb	0.0002	0.0007	0.0003	0.0003	0.0000	0.0000	1.0000	0.0000
attack	0.0001	0.0003	0.0009	0.0014	0.0001	0.0002	0.0268	0.0000
involve	0.0002	0.0017	0.0028	0.0005	0.0002	0.0002	0.0002	0.0000
the	0.0000	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
shining	0.0003	0.0001	0.0136	0.0006	0.0001	0.0001	0.0005	0.0000
path	0.0000	0.0003	0.0001	0.0003	0.0002	0.0002	0.0001	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
two	1.0000	0.0001	0.0016	0.0002	0.0001	0.0000	0.0002	0.0000
people	0.0001	1.0000	0.0036	0.0011	0.0000	0.0005	0.0007	0.0000
be	0.0016	0.0036	1.0000	0.0073	0.0006	0.0001	0.0003	0.0000
injure	0.0001	0.0005	0.0006	0.0387	0.0001	0.0000	0.0005	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.12: Ευθυγράμμιση λέξεων Ζεύγους 137, χρήση συνωνύμων

	0 the	1 cost	2 of	3 paper	4 be	5 rise	6 .
the	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
cost	0.0000	1.0000	0.0001	0.0006	0.0317	0.0005	0.0000
for	0.0000	0.0004	0.0116	0.0004	0.0000	0.0000	0.0000
some	0.0000	0.0001	0.0076	0.0001	0.0001	0.0001	0.0000
30	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
million	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
sheet	0.0000	0.0001	0.0002	0.0298	0.0007	0.0003	0.0000
of	0.0000	0.0001	1.0000	0.0000	0.0001	0.0001	0.0000
paper	0.0000	0.0006	0.0000	1.0000	0.0004	0.0002	0.0000
use	0.0000	0.0034	0.0002	0.0003	0.0045	0.0004	0.0000
each	0.0000	0.0000	0.0007	0.0000	0.0000	0.0000	0.0000
year	0.0000	0.0000	0.0000	0.0004	0.0001	0.0002	0.0000
by	0.0000	0.0002	0.1648	0.0001	0.0006	0.0001	0.0000
cal	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
state	0.0000	0.0011	0.0001	0.0004	0.0014	0.0003	0.0000
long	0.0000	0.0004	0.0000	0.0000	0.0018	0.0002	0.0000
beach	0.0000	0.0002	0.0000	0.0001	0.0003	0.0008	0.0000
college	0.0000	0.0001	0.0000	0.0001	0.0005	0.0001	0.0000
and	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
department	0.0000	0.0012	0.0009	0.0003	0.0006	0.0002	0.0000
go	0.0000	0.0029	0.0001	0.0007	0.0222	0.0103	0.0000
up	0.0000	0.0001	0.0006	0.0000	0.0008	0.0050	0.0000
wednesday	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000
for	0.0000	0.0004	0.0116	0.0004	0.0000	0.0000	0.0000
the	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
first	0.0000	0.0003	0.0003	0.0001	0.0010	0.0009	0.0000
time	0.0000	0.0003	0.0003	0.0002	0.0013	0.0006	0.0000
in	0.0000	0.0003	0.0345	0.0001	0.0016	0.0004	0.0000
four	0.0000	0.0003	0.0000	0.0001	0.0005	0.0005	0.0000
year	0.0000	0.0000	0.0000	0.0004	0.0001	0.0002	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.13: Ευθυγράμμιση λέξεων Ζεύγους 43, χρήση συνωνύμων

	0 four	1 people	2 be	3 assassinate	4 by	5 the	6 pilot	7 .
in	0.0001	0.0009	0.0016	0.0004	0.0043	0.0000	0.0001	0.0000
that	0.0000	0.0000	0.0000	0.0001	0.0052	0.0000	0.0000	0.0000
aircraft	0.0001	0.0003	0.0003	0.0004	0.0000	0.0000	0.0006	0.0000
accident	0.0000	0.0000	0.0013	0.0005	0.0001	0.0000	0.0001	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
four	1.0000	0.0001	0.0005	0.0014	0.0001	0.0000	0.0000	0.0000
people	0.0001	1.0000	0.0036	0.0011	0.0000	0.0006	0.0002	0.0000
be	0.0005	0.0036	1.0000	0.0014	0.0006	0.0000	0.0008	0.0000
kill	0.0007	0.0008	0.0070	0.0307	0.0001	0.0000	0.0001	0.0000
:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
the	0.0000	0.0006	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
pilot	0.0000	0.0002	0.0008	0.0001	0.0001	0.0000	1.0000	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
who	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
be	0.0005	0.0036	1.0000	0.0014	0.0006	0.0000	0.0008	0.0000
wear	0.0001	0.0015	0.0022	0.0014	0.0002	0.0000	0.0003	0.0000
civilian	0.0000	0.1061	0.0002	0.0001	0.0000	0.0000	0.0005	0.0000
clothe	0.0002	0.0008	0.0012	0.0014	0.0000	0.0000	0.0001	0.0000
,	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
and	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
three	0.0184	0.0001	0.0008	0.0001	0.0001	0.0000	0.0000	0.0000
other	0.0000	0.0002	0.0009	0.0001	0.0060	0.0000	0.0003	0.0000
people	0.0001	1.0000	0.0036	0.0011	0.0000	0.0006	0.0002	0.0000
who	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
be	0.0005	0.0036	1.0000	0.0014	0.0006	0.0000	0.0008	0.0000
wear	0.0001	0.0015	0.0022	0.0014	0.0002	0.0000	0.0003	0.0000
military	0.0002	0.0042	0.0002	0.0001	0.0000	0.0000	0.0001	0.0000
uniform	0.0002	0.0002	0.0009	0.0001	0.0001	0.0000	0.0000	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.14: Ευθυγράμμιση λέξεων Ζεύγους 404, χρήση συνωνύμων

	0 organic	1 fertilizer	2 be	3 use	4 a	5 soil	6 enhancer	7 .
organic	1.0000	0.0450	0.0035	0.0016	0.0000	0.0014	0.0000	0.0000
fertilizer	0.0450	1.0000	0.0003	0.0000	0.0000	0.0027	0.0000	0.0000
slowly	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
enrich	0.0018	0.0149	0.0018	0.0009	0.0000	0.0016	0.0160	0.0000
and	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
feed	0.0017	0.0211	0.0023	0.0095	0.0001	0.0006	0.0006	0.0000
the	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
soil	0.0014	0.0027	0.0013	0.0002	0.0000	1.0000	0.0031	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
fast	0.0002	0.0001	0.0006	0.0005	0.0000	0.0002	0.0000	0.0000
act	0.0001	0.0000	0.0087	0.0097	0.0000	0.0003	0.0003	0.0000
synthetic	0.0009	0.0051	0.0002	0.0001	0.0001	0.0001	0.0002	0.0000
fertilizer	0.0450	1.0000	0.0003	0.0000	0.0000	0.0027	0.0000	0.0000
harm	0.0013	0.0012	0.0012	0.0019	0.0001	0.0049	0.0003	0.0000
soil	0.0014	0.0027	0.0013	0.0002	0.0000	1.0000	0.0031	0.0000
life	0.0021	0.0001	0.0142	0.0007	0.0000	0.0003	0.0004	0.0000
.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.15: Ευθυγράμμιση λέξεων Ζεύγους 192, χρήση συνωνύμων

4.4 Ευθυγράμμιση φράσεων

Συνώνυμα + ορισμοί

Ας δούμε και μερικά παραδείγματα για την περίπτωση ευθυγράμμισης φράσεων. Και εδώ, γρήγορα παρατηρεί κανείς λανθασμένες ευθυγραμμίσεις, και μάλιστα πιο πολλές για την περίπτωση χρήσης συνωνύμων και ορισμών. Οι περισσότερες οφείλονται στο ίδιο πρόβλημα που συναντήσαμε και στην περίπτωση της ευθυγράμμισης λέξεων που αφορά την ευθυγράμμιση γενικών λέξεων όπως το be.

Στην περίπτωση των φράσεων ωστόσο, συναντούμε και ένα επιπλέον πρόβλημα. Όταν μια φράση περιέχει ίδια λέξη με μια άλλη, η βαρύτητα με την οποία συνεισφέρει στην ομοιότητα των φράσεων είναι πολύ μεγαλύτερη από λέξεις που παρουσιάζουν σχέσεις αλλά δεν είναι ίδιες. Για παράδειγμα, η ευθυγράμμιση του "the first time" με το "The cost" στον πίνακα 4.18 οφείλεται στην ύπαρξη του "the" και στις δύο φράσεις.

Από την παραπάνω παρατήρηση φαίνεται πως μάλλον έχει νόημα να μην λαμβάνονται υπ' όψιν λέξεις όπως το "the" κατά την απόδοση ομοιοτήτων . Είτε αυτό, είτε να εξομαλυνθεί με κάποιον τρόπο η επιρροή που έχουν ίδιες λέξεις σε σχέση με τις λέξεις που είναι σχετικές στην

εξαγωγή ομοιότητας σε επίπεδο φράσεων.

	NP| Two people	VP| were wounded	PP| by	NP| a bomb	NT| .
NP| Police sources	0.0032	0.0113	0.0014	0.0019	0.0000
VP| stated	0.0169	0.0106	0.0074	0.0027	0.0000
NT| that	0.0003	0.0023	0.0026	0.0002	0.0000
PP| during	0.0007	0.0027	0.0106	0.0035	0.0000
NP| the bomb attack	0.0017	0.0039	0.0007	0.4241	0.0000
VP| involving	0.0027	0.0087	0.0011	0.0008	0.0000
NP| the Shining Path	0.0019	0.0074	0.0016	0.0010	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| two people	1.0000	0.0318	0.0013	0.0032	0.0000
VP| were injured	0.0309	0.5232	0.0021	0.0045	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.16: Ευθυγράμμιση λέξεων Ζεύγους 137, χρήση συνωνύμων και ορισμών

	NP| Organic fertilizers	VP| are used	PP| as	NP| soil enhancers	NT| .
NP| Organic fertilizer	1.0000	0.0564	0.0030	0.0423	0.0000
NT| slowly	0.0063	0.0657	0.0014	0.0011	0.0000
VP| enriches	0.0071	0.0052	0.0003	0.0075	0.0000
NT| and	0.0000	0.0000	0.0000	0.0000	0.0000
VP| feeds	0.0119	0.0173	0.0007	0.0017	0.0000
NP| the soil	0.0423	0.0155	0.0004	0.5018	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000
NT| Fast	0.0009	0.0073	0.0003	0.0011	0.0000
VP| acting	0.0030	0.0177	0.0011	0.0059	0.0000
NP| synthetic fertilizers	0.5331	0.0495	0.0016	0.0406	0.0000
VP| harm	0.0027	0.0090	0.0003	0.0039	0.0000
NP| soil life	0.0469	0.0300	0.0009	0.4982	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.17: Ευθυγράμμιση λέξεων Ζεύγους 192, χρήση συνωνύμων και ορισμών

	NP| The cost	PP| of	NP| paper	VP| is rising	NT| .
NP| The cost	1.0000	0.0002	0.0009	0.0160	0.0000
PP| for	0.0002	0.0101	0.0002	0.0000	0.0000
NP| some 30 million sheets	0.0018	0.0033	0.0159	0.0391	0.0000
PP| of	0.0002	1.0000	0.0002	0.0012	0.0000
NP| paper	0.0009	0.0002	1.0000	0.0028	0.0000
VP| used	0.0034	0.0112	0.0048	0.0154	0.0000
NP| each year	0.0013	0.0012	0.0026	0.0062	0.0000
PP| by	0.0013	0.1581	0.0007	0.0025	0.0000
NP| Cal State	0.0017	0.0014	0.0051	0.0119	0.0000
PR| went up	0.0008	0.0008	0.0004	0.0227	0.0000
NP| Wednesday	0.0002	0.0001	0.0020	0.0022	0.0000
PP| for	0.0002	0.0101	0.0002	0.0000	0.0000
NP| the first time	0.4045	0.0008	0.0011	0.0089	0.0000
PP| in	0.0005	0.0333	0.0005	0.0053	0.0000
NP| four years	0.0016	0.0012	0.0021	0.0173	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.18: Ευθυγράμμιση λέξεων Ζεύγους 43, χρήση συνωνύμων και ορισμών

	NP| Four people	VP| were assassinated	PP| by	NP| the pilot	NT| .
PP| In	0.0016	0.0050	0.0028	0.0004	0.0000
NP| that aircraft accident	0.0052	0.0055	0.0021	0.0058	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| four people	1.0000	0.0235	0.0009	0.0011	0.0000
VP| were killed	0.0230	0.5207	0.0022	0.0018	0.0000
NT| :	0.0000	0.0000	0.0000	0.0000	0.0000
NP| the pilot	0.0011	0.0018	0.0007	1.0000	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| who	0.0028	0.0020	0.0005	0.0003	0.0000
VP| was wearing	0.0241	0.5035	0.0030	0.0019	0.0000
NP| civilian clothes	0.0329	0.0274	0.0008	0.0011	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NT| and	0.0001	0.0000	0.0000	0.0000	0.0000
NP| three other people	0.4372	0.0233	0.0031	0.0014	0.0000
NP| who	0.0028	0.0020	0.0005	0.0003	0.0000
VP| were wearing	0.0241	0.5035	0.0030	0.0019	0.0000
NP| military uniforms	0.0046	0.0043	0.0013	0.0011	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.19: Ευθυγράμμιση λέξεων Ζεύγους 404, χρήση συνωνύμων και ορισμών

Συνώνυμα

Και στην περίπτωση της ευθυγράμμισης φράσεων, η χρήση συνωνύμων αντί για την χρήση του συνδυαστικού συνόλο δεδομένων φαίνεται να είναι πιο καλή επιλογή.

Γενικά οι ευθυγραμμίσεις είναι καλύτερης ποιότητας, ωστόσο συναντούμε και εδώ το πρόβλημα που αναφέραμε και νωρίτερα για την επιρροή συνδετικών λέξεων όπως το "the" και τον άνισο καταμερισμό της βαρύτητας με την οποία συνεισφέρουν στην ομοιότητα ίδιες λέξεις σε σχέση με σχετικές.

	NP| Two people	VP| were wounded	PP| by	NP| a bomb	NT| .
NP| Police sources	0.0011	0.0011	0.0001	0.0006	0.0000
VP| stated	0.0019	0.0012	0.0010	0.0001	0.0000
NT| that	0.0000	0.0000	0.0052	0.0001	0.0000
PP| during	0.0000	0.0002	0.0162	0.0066	0.0000
NP| the bomb attack	0.0007	0.0012	0.0001	0.4259	0.0000
VP| involving	0.0013	0.0023	0.0002	0.0003	0.0000
NP| the Shining Path	0.0006	0.0059	0.0002	0.0004	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| two people	1.0000	0.0032	0.0001	0.0007	0.0000
VP| were injured	0.0029	0.5212	0.0005	0.0005	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.20: Ευθυγράμμιση φράσεων Ζεύγους 137, χρήση συνωνύμων

	NP| The cost	PP| of	NP| paper	VP| is rising	NT| .
NP| The cost	1.0000	0.0000	0.0005	0.0162	0.0000
PP| for	0.0003	0.0116	0.0004	0.0000	0.0000
NP| some 30 million sheets	0.0001	0.0036	0.0138	0.0004	0.0000
PP| of	0.0000	1.0000	0.0000	0.0001	0.0000
NP| paper	0.0005	0.0000	1.0000	0.0004	0.0000
VP| used	0.0024	0.0002	0.0003	0.0035	0.0000
NP| each year	0.0000	0.0005	0.0003	0.0001	0.0000
PP| by	0.0002	0.1648	0.0001	0.0005	0.0000
NP| Cal State	0.0008	0.0004	0.0003	0.0017	0.0000
PR| went up	0.0002	0.0003	0.0001	0.0267	0.0000
NP| Wednesday	0.0000	0.0000	0.0004	0.0000	0.0000
PP| for	0.0003	0.0116	0.0004	0.0000	0.0000
NP| the first time	0.3951	0.0004	0.0002	0.0016	0.0000
PP| in	0.0002	0.0345	0.0001	0.0014	0.0000
NP| four years	0.0001	0.0000	0.0004	0.0006	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.21: Ευθυγράμμιση φράσεων Ζεύγους 43, χρήση συνωνύμων

	NP| Four people	VP| were assassinated	PP| by	NP| the pilot	NT| .
PP| In	0.0007	0.0014	0.0043	0.0001	0.0000
NP| that aircraft accident	0.0002	0.0010	0.0031	0.0003	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| four people	1.0000	0.0033	0.0001	0.0004	0.0000
VP| were killed	0.0027	0.5180	0.0005	0.0005	0.0000
NT| :	0.0000	0.0000	0.0000	0.0000	0.0000
NP| the pilot	0.0004	0.0005	0.0001	1.0000	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NP| who	0.0004	0.0000	0.0000	0.0000	0.0000
VP| was wearing	0.0028	0.4997	0.0006	0.0005	0.0000
NP| civilian clothes	0.0557	0.0014	0.0000	0.0003	0.0000
NT| ,	0.0000	0.0000	0.0000	0.0000	0.0000
NT| and	0.0002	0.0000	0.0000	0.0000	0.0000
NP| three other people	0.4071	0.0027	0.0035	0.0005	0.0000
NP| who	0.0004	0.0000	0.0000	0.0000	0.0000
VP| were wearing	0.0028	0.4997	0.0006	0.0005	0.0000
NP| military uniforms	0.0025	0.0006	0.0001	0.0001	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.22: Ευθυγράμμιση φράσεων Ζεύγους 404, χρήση συνωνύμων

	NP| Organic fertilizers	VP| are used	PP| as	NP| soil enhancers	NT| .
NP| Organic fertilizer	1.0000	0.0024	0.0000	0.0020	0.0000
NT| slowly	0.0000	0.0001	0.0000	0.0000	0.0000
VP| enriches	0.0124	0.0019	0.0000	0.0129	0.0000
NT| and	0.0000	0.0000	0.0000	0.0001	0.0000
VP| feeds	0.0171	0.0084	0.0001	0.0009	0.0000
NP| the soil	0.0021	0.0008	0.0000	0.4833	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000
NT| Fast	0.0002	0.0008	0.0000	0.0002	0.0000
VP| acting	0.0001	0.0130	0.0000	0.0005	0.0000
NP| synthetic fertilizers	0.6051	0.0003	0.0001	0.0016	0.0000
VP| harm	0.0017	0.0022	0.0001	0.0035	0.0000
NP| soil life	0.0030	0.0082	0.0000	0.4795	0.0000
NT| .	0.0000	0.0000	0.0000	0.0000	1.0000

Πίνακας 4.23: Ευθυγράμμιση φράσεων Ζεύγους 192, χρήση συνωνύμων

4.5 Συμπεράσματα

Στα πλαίσια της πτυχιακής εργασίας είδαμε πως έχει νόημα η δημιουργία γράφου από δεδομένα συνωνύμων και ορισμών. Οι σχέσεις αυτές περικλείουν αρκετή χρήσιμη πληροφορία για την εύρεση ομοιότητας ανάμεσα σε λέξεις. Η διατήρηση τους σε γράφο μας δίνει την δυνατότητα να κοιτάξουμε για ομοιότητες σε βάθος μεγαλύτερο του ενός, και όπως είδαμε στην αξιολόγηση των προσεγγίσεων, κάτι τέτοιο μας δίνει καλύτερα αποτελέσματα.

Δείξαμε πως είναι δυνατόν κανείς να εξάγει πληροφορία από γράφους με χρήση διανυσμάτων. Πως τελικά, είναι δυνατή η επέκταση σε φράσεις, αν και θα πρέπει να αναζητηθεί τρόπος να εξομαλυνθούν οι διαφορές που υπάρχει στην βαθμολογία ομοιότητας ίδιων λέξεων και παρόμοιων. Αυτό είναι αναγκαίο καθώς στην περίπτωση ευθυγράμμισης φράσεων η βαθμολογία ίδιων λέξεων έχει πολύ μεγαλύτερη βαρύτητα στην ευθυγράμμιση της φράσης από ότι έχουν παρόμοιες λέξεις.

Ως προς την βαθμολογία συσχέτισης που βρήκαμε στην ενότητα 4.2.3 μπορούμε να πούμε πως είναι μέτρια σε σχέση με τα αποτελέσματα που φαίνονται στο σχήμα 4.1 [tau Yih and Qazvinian, 2012]. Ωστόσο, υπάρχουν μεγάλα περιθώρια βελτίωσης αν θέλει κανείς να επεκτείνει το σύστημα προς αυτήν την κατεύθυνση. Άλλωστε, οι άλλες ομάδες χρησιμοποίησαν πολλά δεδομένα από το διαδίκτυο, ενώ το δικό μας σύστημα λειτουργεί με ένα αρχείο πληροφορίας ορισμών και συνωνύμων 30 Megabyte.

Ένα σημαντικό συμπέρασμα είναι πως η επιλογή του συνόλου δεδομένων για αξιολόγηση

	Spearman's ρ					
Method	**WS-353**	**WS-sim**	**WS-rel**	**MC-30**	**RG-65**	**MTurk-287**
(Radinsky et al., 2011)	0.80	-	-	-	-	0.63
(Reisinger and Mooney, 2010)	0.77	-	-	-	-	-
(Agirre et al., 2009)	0.78	0.83	0.72	**0.92**	**0.96**	-
(Gabrilovich and Markovitch, 2007)	0.75	-	-	-	-	0.59
(Hughes and Ramage, 2007)	0.55	-	-	0.90	0.84	-
Web Search	0.56	0.56	0.54	0.48	0.44	0.44
Wikipedia	0.73	0.80	0.73	0.87	0.83	0.62
Bloomsbury	0.45	0.60	0.60	0.71	0.78	0.29
WordNet	0.37	0.49	0.49	0.79	0.78	0.25
Combining VSMs	**0.81**	**0.87**	**0.77**	0.89	0.89	**0.68**

Σχήμα 4.1: Άλλες επιδόσεις στο σύνολο δεδομένων wordsimiliarity-353

πρέπει να γίνεται πολύ προσεχτικά, καθώς τελικά καταλήγει να επηρεάζει τα αποτελέσματα και τις αποφάσεις που παίρνει κανείς στα πλαίσια της πτυχιακής.

Η χρήση του συνόλου δεδομένων wordsimilarity-353 για αξιολόγηση δεν ήταν η καλύτερη δυνατή επιλογή. Αυτό φαίνεται από τις αστοχίες που συναντήθηκαν στην ευθυγράμμιση. Η εξήγηση είναι πως το είδος ομοιότητας λέξεων το οποίο εξέφραζε το σύνολο δεδομένων δεν ήταν ίδιο με αυτό που θα θέλαμε σε μια ευθυγράμμιση προτάσεων. Στο μέλλον θα ήταν καλό να βρεθεί ένα σύνολο δεδομένων που να αντιπροσωπεύει καλύτερα τις ομοιότητες που αναζητώνται κατά την διαδικασία ευθυγράμμισης.

Παρατηρήσαμε επίσης από τις ευθυγραμμίσεις και την αξιολόγηση πως, ανάλογα με το τι ομοιότητα αναζητεί κανείς, αλλάζουν και οι σχέσεις που θα πρέπει να περιέχει ο γράφος. Για παράδειγμα, σε περίπτωση επέκτασης του συστήματος για καλύτερα αποτελέσματα στο σύνολο wordsimilarity-353, θα μπορούσε να προστεθεί στον γράφο πληροφορία από ορισμούς σε εγκυκλοπαίδεια για να μπορέσει να βρεί καλύτερες συσχετίσεις ανάμεσα στα φυσικά πρόσωπα που περιέχει το wordsimilarity-353.

Εξετάσαμε τις ευθυγραμμίσεις που επιλέγει το σύστημα και είδαμε πως στην περίπτωση τουλάχιστον της χρήσης συνωνύμων είναι εως έναν βαθμό ικανοποιητικές. Για την περίπτωση χρήσης συνωνύμων και ορισμών είδαμε πως υπάρχουν αρκετές αστοχίες, και συνεπώς θα πρέπει πέρα από την αναζήτηση πιο αντιπροσωπευτικού συνόλου δεδομένων για αντιπαράθεση στην αξιολόγηση, να ελεγχθεί αν υπάρχει κάποιος καλύτερος τρόπος σχηματισμού του συνόλου δεδομένων.

Σημαντικά μελλοντικά βήματα είναι η βελτίωση του τρόπου εξαγωγής των σχετικών λέξεων από τους ορισμούς, η εισαγωγή βαρών στις ακμές του γράφου ανάλογα με το χαρακτηριστικό της σχέσης που απεικονίζεται, καθώς και η αναζήτηση προσεγγίσεων που συγκλίνουν δίχως

την ανάγκη περιορισμού σε κάποιο βάθος. Τέλος, κρίσιμη μελλοντική προσθήκη είναι η συστηματική αξιολόγηση των ευθυγραμμίσεων με αντιπαράθεση κάποιου επισημειωμένου ως προς την ευθυγράμμιση συνόλου του RTE.

ΠΑΡΑΡΤΗΜΑ Α'

Υποσύνολο RTE2 συνόλου δεδομένων

```
<pair id="476" entailment="YES" task="SUM">
<t>Clonaid scientist, Brigitte Boisselier, said the first human clone - a girl nicknamed
    Eve - was born on Thursday to an American mother.</t>
<h>Brigitte Boisselier announced that a cloned baby had been born.</h>
<pair id="137" entailment="YES" task="IE">
<t>Police sources stated that during the bomb attack involving the Shining Path, two
    people were injured.</t>
<h>Two people were wounded by a bomb.</h>
<pair id="404" entailment="NO" task="IE">
<t>In that aircraft accident, four people were killed: the pilot, who was wearing civilian
     clothes, and three other people who were wearing military uniforms.</t>
<h>Four people were assassinated by the pilot.</h>
<pair id="430" entailment="NO" task="IE">
<t>A.S. Roma suffered their second home defeat of the campaign, with a 3-2 loss to 10-man
    Siena.</t>
<h>Siena was beaten by A.S. Roma.</h>
<pair id="431" entailment="NO" task="IE">
<t>Helena Brighton, an attorney for Eliza May who served as executive director of the
    Texas Funeral Services Commission -the state agency that regulates the funeral
    business- claimed that she was fired from her state job because she raised questions
    about SCI.</t>
<h>Helena Brighton is a clerk of the Texas Funeral Services Commission.</h>
<pair id="434" entailment="YES" task="IE">
<t>Self-sufficiency has been turned into a formal public awareness campaign in San
    Francisco, by Mayor Gavin Newsom.</t>
<h>Gavin Newsom is a politician of San Francisco.</h>
<pair id="481" entailment="YES" task="IE">
<t>Lima, Jan. 10, '90 - The national police reported that over 15,000 people have been
    arrested in Lima in a dragnet aimed at uncovering the assassins of former Defense
    Minister Enrique Lopez Albujar Trint, who was murdered in a terrorist attack,
    yesterday.</t>
```

```
<h>Enrique Lopez Albujar Trint was killed on Jan. 9 '90.</h>
<pair id="508" entailment="YES" task="IE">
<t>Lao government spokesman, Yong Chantalangsy, said Rangoon has promised to announce its
    decision in Vientiane.</t>
<h>Yong Chantalangsy is a representative of Laos.</h>
<pair id="512" entailment="NO" task="IE">
<t>The murders of Galan, a high-ranking police officer, and a judge last week, were
    probably carried out by an Israeli mercenary paid by the Medellin cartel - the world's
     most powerful drug-trafficking mafia - which has set up paramilitary squads accused
    of hundreds of individual murders and massacres against political and labor leaders,
    as well as peasants and patriotic union leftist militants in different parts of the
    country.</t>
<h>Galan was killed by patriotic union leftist militants. </h>
<pair id="535" entailment="YES" task="IE">
<t>Toshiba, the world's third-largest notebook computer maker behind Dell Inc. and Hewlett
    -Packard Co., said the PC would be introduced in Japan in early 2006.</t>
<h>Toshiba produces notebook computers.</h>
<pair id="562" entailment="YES" task="IE">
<t>Salvadoran reporter Mauricio Pineda, a sound technician for the local Canal Doce
    television station, was shot to death today in Morazan Department in the eastern part
    of the country.</t>
<h>Mauricio Pineda was killed by gunfire.</h>
<pair id="570" entailment="YES" task="IE">
<t>Alleged terrorists today, killed Dolores Hinostroza, the mayor of Mulqui district,
    shooting her five times.</t>
<h>The mayor of Mulqui district was murdered with a firearm.</h>
<pair id="581" entailment="UNKNOWN" task="IE">
<t>A mercenary group, faithful to the warmongering policy of former Somozist colonel
    Enrique Bermudez, attacked an IFA truck belonging to the interior ministry at 0900 on
    26 March in El Jicote, wounded and killed an interior ministry worker and wounded five
     others.</t>
<h>A mercenary group was injured in El Jicote.</h>
<pair id="620" entailment="YES" task="IE">
<t>Ordonez Reyes accused Jose Jesus Pena, alleged chief of security for the Nicaraguan
    embassy in Tegucigalpa, of masterminding the January 7th assassination of contra-
    commander Manuel Antonio Rugama. </t>
<h>Jose Jesus Pena is accused of the assassination of Manuel Antonio Rugama.</h>
<pair id="625" entailment="UNKNOWN" task="IE">
<t>The loss offered a minor moral victory for Liverpool, as they scored only the second
    goal this season against Chelsea in league play.</t>
<h>Liverpool beat Chelsea.</h>
```

<pair id="627" entailment="UNKNOWN" task="IE">
<t>La Paz, 30 May 89 - La Paz Department Police authorities have disclosed that investigations into the murder of two young U.S. citizens are being conducted by a specialized group summoned specially to clarify this crime.</t>
<h>Two young U.S. citizens were killed on 30 May 89. </h>
<pair id="653" entailment="NO" task="IE">
<t>Mr. Olsen paid $20 million for the space trip, but says that in the future space travel will likely become as routine as air travel is today.</t>
<h>Air travel costs $20 million.</h>
<pair id="663" entailment="YES" task="IE">
<t>FMLN militias sabotaged power lines along the road linking Nueva Concepcion to the northern trunk highway at 1800 on 12 August.</t>
<h>Power lines were attacked by FMLN.</h>
<pair id="668" entailment="UNKNOWN" task="IE">
<t>Their work has stimulated research into microbes as possible reasons for other chronic inflammatory conditions, such as Crohn's disease, ulcerative colitis, rheumatoid arthritis and atherosclerosis, the Nobel assembly said.</t>
<h>Rheumatoid arthritis is caused by microbes.</h>
<pair id="670" entailment="YES" task="IE">
<t>A member of the Burmese delegation, Thaung Tun, told reporters on the eve of the meeting that his government did not wish to place ASEAN in a difficult position by insisting on assuming the chairmanship.</t>
<h>Thaung Tun is a representative of Burma.</h>
<pair id="679" entailment="YES" task="IE">
<t>Referring to the April 2 car bomb explosion perpetrated in Santa Tecla, the FMLN-FDR claimed responsibility for the attack in a communique, and said that it was an action aimed directly at the police and that they regret the death of a civilian, but they did not mention compensation for the relatives.</t>
<h>A civilian was killed by a car bomb.</h>
<pair id="701" entailment="UNKNOWN" task="IE">
<t>FMLN guerrilla units ambushed the 1st company of military detachment no. 2 Jr. Battalion at la Pena Canton, Villa Victoria Jurisdiction.</t>
<h>FMLN guerrilla units attacked a commercial company.</h>
<pair id="702" entailment="UNKNOWN" task="IE">
<t>The chaotic situation unleashed in Bogota last night, with the assasination of Justice Carlos Valencia, began on 28 July in Medellin, when motorized paid assasins murdered third public order Judge Maria Elena Diaz.</t>
<h>Justice Carlos Valencia was killed in Medellin.</h>
<pair id="776" entailment="UNKNOWN" task="IE">
<t>According to an Israeli official, the murders of Galan, a high-ranking police officer, and a judge last week were probably carried out by an Israeli mercenary paid by the

```
    Medellin cartel - the world's most powerful drug-trafficking mafia - which has set up
    paramilitary squads accused of hundreds of individual murders and massacres against
    political and labor leaders, as well as peasants and patriotic union leftist militants
     in different parts of the country.</t>
<h>An Israeli official was killed by patriotic union leftist militants. </h>
<pair id="43" entailment="YES" task="IR">
<t>The cost for some 30 million sheets of paper used each year by Cal State Long Beach
    colleges and departments went up Wednesday for the first time in four years.</t>
<h>The cost of paper is rising.</h>
<pair id="192" entailment="YES" task="IR">
<t>Organic fertilizer slowly enriches and feeds the soil. Fast acting synthetic
    fertilizers harm soil life.</t>
<h>Organic fertilizers are used as soil enhancers.</h>
<pair id="78" entailment="YES" task="QA">
<t>German automaker, Volkswagen AG, launched a special collector's edition of its original
     Beetle, on Thursday, to mark the end of the line for the most popular car in history.
    </t>
<h>Volkswagen AG produces the 'Beetle'.</h>
<pair id="393" entailment="UNKNOWN" task="QA">
<t>The Beetle has achieved an unbelievable cult status across the world, not least because
     its creation marked the launch of what has since become the largest car company in
    Europe.</t>
<h>Europe produces the 'Beetle'.</h>
<pair id="676" entailment="YES" task="QA">
<t>Joe Friday wore badge No. 714-in honor of the number of home runs Ruth hit in his
    career. </t>
<h>Ruth hit 714 home runs in his lifetime. </h>
<pair id="437" entailment="UNKNOWN" task="SUM">
<t>While the beet is roasting, make the vinaigrette: In a bowl, whisk together the orange
    juice, balsamic vinegar, and extra-virgin olive oil.</t>
<h>In a bowl, toss the olive oil with salt and pepper.</h>
```

ΠΑΡΑΡΤΗΜΑ Β΄

Σύνολο δεδομένων wordsimilarity-353

admission,ticket,7.69
alcohol,chemistry,5.54
aluminum,metal,7.83
announcement,effort,2.75
announcement,news,7.56
announcement,production,3.38
announcement,warning,6.00
Arafat,Jackson,2.50
Arafat,peace,6.73
Arafat,terror,7.65
architecture,century,3.78
arrangement,accommodation,5.41
arrival,hotel,6.00
asylum,madhouse,8.87
atmosphere,landscape,3.69
attempt,peace,4.25
baby,mother,7.85
bank,money,8.12
baseball,season,5.97
bed,closet,6.72
benchmark,index,4.25
bird,cock,7.10
bird,crane,7.38
bishop,rabbi,6.69
board,recommendation,4.47
book,library,7.46
book,paper,7.46
boxing,round,7.61
boy,lad,8.83
bread,butter,6.19
brother,monk,6.27
calculation,computation,8.44
canyon,landscape,7.53
car,automobile,8.94
car,flight,4.94
cell,phone,7.81
cemetery,woodland,2.08
century,nation,3.16
century,year,7.59
championship,tournament,8.36
chance,credibility,3.88
change,attitude,5.44
chord,smile,0.54
closet,clothes,8.00
coast,forest,3.15
coast,hill,4.38
coast,shore,9.10
company,stock,7.08
competition,price,6.44
computer,internet,7.58
computer,keyboard,7.62
computer,laboratory,6.78
computer,news,4.47
computer,software,8.50
concert,virtuoso,6.81
consumer,confidence,4.13
consumer,energy,4.75
country,citizen,7.31
crane,implement,2.69
credit,card,8.06
credit,information,5.31
cucumber,potato,5.92
cup,article,2.40
cup,artifact,2.92
cup,coffee,6.58
cup,drink,7.25
cup,entity,2.15
cup,food,5.00
cup,liquid,5.90
cup,object,3.69
cup,substance,1.92
cup,tableware,6.85
currency,market,7.50
day,dawn,7.53
day,summer,3.94

death,inmate,5.03
death,row,5.25
decoration,valor,5.63
delay,news,3.31
delay,racism,1.19
deployment,departure,4.25
deployment,withdrawal,5.88
development,issue,3.97
direction,combination,2.25
disability,death,5.47
disaster,area,6.25
discovery,space,6.34
dividend,calculation,6.48
dividend,payment,7.63
doctor,liability,5.19
doctor,nurse,7.00
doctor,personnel,5.00
dollar,buck,9.22
dollar,loss,6.09
dollar,profit,7.38
dollar,yen,7.78
drink,car,3.04
drink,ear,1.31
drink,eat,6.87
drink,mother,2.65
drink,mouth,5.96
drug,abuse,6.85
energy,crisis,5.94
energy,laboratory,5.09
energy,secretary,1.81
environment,ecology,8.81
equipment,maker,5.91
exhibit,memorabilia,5.31
experience,music,3.47
family,planning,6.25
FBI,fingerprint,6.94
FBI,investigation,8.31
fertility,egg,6.69
fighting,defeating,7.41
five,month,3.38
focus,life,4.06
food,fruit,7.52
food,preparation,6.22
food,rooster,4.42
football,basketball,6.81
football,soccer,9.03
football,tennis,6.63
forest,graveyard,1.85
fuck,sex,9.44
furnace,stove,8.79
game,defeat,6.97
game,round,5.97
game,series,6.19
game,team,7.69
game,victory,7.03
gem,jewel,8.96
gender,equality,6.41
glass,magician,2.08
glass,metal,5.56
government,crisis,6.56
governor,interview,3.25
governor,office,6.34
grocery,money,5.94
Harvard,Yale,8.13
holy,sex,1.62
hospital,infrastructure,4.63
hotel,reservation,8.03
hundred,percent,7.38
image,surface,4.56
impartiality,interest,5.16
investigation,effort,4.59
investor,earning,7.13
jaguar,car,7.27
jaguar,cat,7.42
Japanese,American,6.50
Jerusalem,Israel,8.46
Jerusalem,Palestinian,7.65
journal,association,4.97
journey,car,5.85
journey,voyage,9.29
king,cabbage,0.23
king,queen,8.58
king,rook,5.92
lad,brother,4.46
lad,wizard,0.92
law,lawyer,8.38
lawyer,evidence,6.69
liability,insurance,7.03
life,death,7.88
life,lesson,5.94
life,term,4.50
line,insurance,2.69
liquid,water,7.89
listing,category,6.38
listing,proximity,2.56
lobster,food,7.81
lobster,wine,5.70
lover,quarrel,6.19
love,sex,6.77
luxury,car,6.47
magician,wizard,9.02
man,governor,5.25
man,woman,8.30
Maradona,football,8.62

marathon,sprint,7.47
Mars,scientist,5.63
Mars,water,2.94
media,gain,2.88
media,radio,7.42
media,trading,3.88
Mexico,Brazil,7.44
midday,noon,9.29
mile,kilometer,8.66
minister,party,6.63
ministry,culture,4.69
minority,peace,3.69
money,bank,8.50
money,cash,9.08
money,cash,9.15
money,currency,9.04
money,deposit,7.73
money,dollar,8.42
money,laundering,5.65
money,operation,3.31
money,possession,7.29
money,property,7.57
money,wealth,8.27
money,withdrawal,6.88
monk,oracle,5.00
monk,slave,0.92
month,hotel,1.81
morality,importance,3.31
morality,marriage,3.69
movie,critic,6.73
movie,popcorn,6.19
movie,star,7.38
movie,theater,7.92
murder,manslaughter,8.53
museum,theater,7.19
music,project,3.63
nature,environment,8.31
nature,man,6.25
network,hardware,8.31
news,report,8.16
noon,string,0.54
observation,architecture,4.38
oil,stock,6.34
OPEC,country,5.63
OPEC,oil,8.59
opera,industry,2.63
opera,performance,6.88
peace,atmosphere,3.69
peace,insurance,2.94
peace,plan,4.75
phone,equipment,7.13
physics,chemistry,7.35
physics,proton,8.12
plane,car,5.77
planet,astronomer,7.94
planet,constellation,8.06
planet,galaxy,8.11
planet,moon,8.08
planet,people,5.75
planet,space,7.92
planet,star,8.45
planet,sun,8.02
population,development,3.75
possibility,girl,1.94
practice,institution,3.19
precedent,antecedent,6.04
precedent,cognition,2.81
precedent,collection,2.50
precedent,example,5.85
precedent,group,1.77
precedent,information,3.85
precedent,law,6.65
prejudice,recognition,3.00
preservation,world,6.19
president,medal,3.00
problem,airport,2.38
problem,challenge,6.75
production,crew,6.25
production,hike,1.75
professor,cucumber,0.31
professor,doctor,6.62
profit,loss,7.63
profit,warning,3.88
psychology,anxiety,7.00
psychology,clinic,6.58
psychology,cognition,7.48
psychology,depression,7.42
psychology,discipline,5.58
psychology,doctor,6.42
psychology,fear,6.85
psychology,Freud,8.21
psychology,health,7.23
psychology,mind,7.69
psychology,psychiatry,8.08
psychology,science,6.71
reason,criterion,5.91
reason,hypertension,2.31
record,number,6.31
registration,arrangement,6.00
report,gain,3.63
rock,jazz,7.59
rooster,voyage,0.62
school,center,3.44
seafood,food,8.34

seafood,lobster,8.70
seafood,sea,7.47
secretary,senate,5.06
seven,series,3.56
shore,woodland,3.08
shower,flood,6.03
shower,thunderstorm,6.31
sign,recess,2.38
situation,conclusion,4.81
situation,isolation,3.88
size,prominence,5.31
skin,eye,6.22
smart,student,4.62
smart,stupid,5.81
soap,opera,7.94
space,chemistry,4.88
space,world,6.53
start,match,4.47
start,year,4.06
stock,CD,1.31
stock,egg,1.81
stock,jaguar,0.92
stock,life,0.92
stock,live,3.73
stock,market,8.08
stock,phone,1.62
street,avenue,8.88
street,block,6.88
street,children,4.94
street,place,6.44
stroke,hospital,7.03
student,professor,6.81
sugar,approach,0.88
summer,drought,7.16
summer,nature,5.63
telephone,communication,7.50
television,film,7.72
television,radio,6.77
tennis,racket,7.56
territory,kilometer,5.28
territory,surface,5.34
theater,history,3.91
tiger,animal,7.00
tiger,carnivore,7.08
tiger,cat,7.35
tiger,fauna,5.62
tiger,feline,8.00
tiger,jaguar,8.00
tiger,mammal,6.85
tiger,organism,4.77
tiger,tiger,10.00
tiger,zoo,5.87
tool,implement,6.46
train,car,6.31
travel,activity,5.00
treatment,recovery,7.91
type,kind,8.97
victim,emergency,6.47
video,archive,6.34
viewer,serial,2.97
vodka,brandy,8.13
vodka,gin,8.46
volunteer,motto,2.56
war,troops,8.13
water,seepage,6.56
weapon,secret,6.06
weather,forecast,8.34
Wednesday,news,2.22
wood,forest,7.73
word,similarity,4.75

ΠΑΡΑΡΤΗΜΑ Γ΄

Στατιστικά Συνωνύμων

boolean depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1210
negmean	0.0055
posstd	0.1876
negstd	0.0170
overlap	-0.0891
overlap_norm	-0.0891

boolean depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.2240
negmean	0.1138
posstd	0.1917
negstd	0.0903
overlap	-0.1719
overlap_norm	-0.1719

concentration+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0304
negmean	0.0001
posstd	0.0957
negstd	0.0005
overlap	-0.0660
overlap_norm	-0.0660

concentration+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0145
negmean	0.0002
posstd	0.0913
negstd	0.0003
overlap	-0.0773
overlap_norm	-0.0773

count depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1210
negmean	0.0055
posstd	0.1876
negstd	0.0170
overlap	-0.0891
overlap_norm	-0.0891

count depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.2568
negmean	0.0860
posstd	0.2454
negstd	0.0814
overlap	-0.1560
overlap_norm	-0.1560

countdepth depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1208
negmean	0.0052
posstd	0.1825
negstd	0.0158
overlap	-0.0827
overlap_norm	-0.0827

countdepth depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.2529
negmean	0.0829
posstd	0.2444
negstd	0.0794
overlap	-0.1539
overlap_norm	-0.1539

dispersion depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0304
negmean	0.0001
posstd	0.0957
negstd	0.0005
overlap	-0.0660
overlap_norm	-0.0660

dispersion+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0304
negmean	0.0001
posstd	0.0957
negstd	0.0005
overlap	-0.0660
overlap_norm	-0.0660

dispersion depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.1757
negmean	0.0190
posstd	0.2206
negstd	0.0200
overlap	-0.0839
overlap_norm	-0.0839

dispersion+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0413
negmean	0.0005
posstd	0.1033
negstd	0.0009
overlap	-0.0633
overlap_norm	-0.0633

recursive depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0078
negmean	0.0000
posstd	0.0880
negstd	0.0000
overlap	-0.0802
overlap_norm	-0.0802

recursive depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0318
negmean	0.0000
posstd	0.1011
negstd	0.0000
overlap	-0.0694
overlap_norm	-0.0694

robinhood depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0304
negmean	0.0001
posstd	0.0957
negstd	0.0005
overlap	-0.0660
overlap_norm	-0.0660

robinhood depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0848
negmean	0.0012
posstd	0.1571
negstd	0.0029
overlap	-0.0764
overlap_norm	-0.0764

ΠΑΡΑΡΤΗΜΑ Δ΄

Στατιστικά Ορισμών

boolean depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1273
negmean	0.0295
posstd	0.1608
negstd	0.0452
overlap	-0.1083
overlap_norm	-0.1083

boolean depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.3322
negmean	0.2829
posstd	0.1684
negstd	0.1004
overlap	-0.2195
overlap_norm	-0.2195

concentration+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0412
negmean	0.0023
posstd	0.1000
negstd	0.0075
overlap	-0.0686
overlap_norm	-0.0686

concentration+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0268
negmean	0.0012
posstd	0.0987
negstd	0.0014
overlap	-0.0745
overlap_norm	-0.0745

count depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1327
negmean	0.0280
posstd	0.1619
negstd	0.0435
overlap	-0.1007
overlap_norm	-0.1007

count depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.4631
negmean	0.4194
posstd	0.1984
negstd	0.1536
overlap	-0.3082
overlap_norm	-0.3082

countdepth depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1264
negmean	0.0230
posstd	0.1445
negstd	0.0335
overlap	-0.0745
overlap_norm	-0.0745

countdepth depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.4530
negmean	0.4044
posstd	0.1981
negstd	0.1539
overlap	-0.3034
overlap_norm	-0.3034

dispersion depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0407
negmean	0.0022
posstd	0.0999
negstd	0.0074
overlap	-0.0688
overlap_norm	-0.0688

dispersion depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.2674
negmean	0.1966
posstd	0.1539
negstd	0.1219
overlap	-0.2049
overlap_norm	-0.2049

dispersion+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0407
negmean	0.0022
posstd	0.0999
negstd	0.0074
overlap	-0.0688
overlap_norm	-0.0688

dispersion+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0517
negmean	0.0063
posstd	0.1015
negstd	0.0081
overlap	-0.0642
overlap_norm	-0.0642

recursive depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0078
negmean	0.0000
posstd	0.0880
negstd	0.0000
overlap	-0.0802
overlap_norm	-0.0802

recursive depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0394
negmean	0.0010
posstd	0.0972
negstd	0.0022
overlap	-0.0610
overlap_norm	-0.0610

robinhood depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0407
negmean	0.0022
posstd	0.0999
negstd	0.0074
overlap	-0.0688
overlap_norm	-0.0688

robinhood depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.1122
negmean	0.0466
posstd	0.1671
negstd	0.0940
overlap	-0.1955
overlap_norm	-0.1955

ΠΑΡΑΡΤΗΜΑ Ε΄

Στατιστικά Συνωνύμων + Ορισμών

boolean depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1294
negmean	0.0169
posstd	0.1689
negstd	0.0196
overlap	-0.0760
overlap_norm	-0.0760

boolean depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.3241
negmean	0.2594
posstd	0.1538
negstd	0.0936
overlap	-0.1827
overlap_norm	-0.1827

concentration+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0288
negmean	0.0003
posstd	0.0906
negstd	0.0005
overlap	-0.0625
overlap_norm	-0.0625

concentration+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0093
negmean	0.0002
posstd	0.0880
negstd	0.0001
overlap	-0.0790
overlap_norm	-0.0790

count depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1336
negmean	0.0166
posstd	0.1697
negstd	0.0192
overlap	-0.0718
overlap_norm	-0.0718

count depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.4345
negmean	0.3199
posstd	0.1976
negstd	0.1369
overlap	-0.2199
overlap_norm	-0.2199

countdepth depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.1366
negmean	0.0155
posstd	0.1676
negstd	0.0180
overlap	-0.0645
overlap_norm	-0.0645

countdepth depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.4304
negmean	0.3127
posstd	0.1984
negstd	0.1362
overlap	-0.2169
overlap_norm	-0.2169

dispersion depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0286
negmean	0.0003
posstd	0.0905
negstd	0.0005
overlap	-0.0626
overlap_norm	-0.0626

dispersion depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.2588
negmean	0.1062
posstd	0.1942
negstd	0.0732
overlap	-0.1148
overlap_norm	-0.1148

dispersion+ depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0286
negmean	0.0003
posstd	0.0905
negstd	0.0005
overlap	-0.0626
overlap_norm	-0.0626

dispersion+ depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0379
negmean	0.0012
posstd	0.0930
negstd	0.0009
overlap	-0.0573
overlap_norm	-0.0573

recursive depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0078
negmean	0.0000
posstd	0.0880
negstd	0.0000
overlap	-0.0802
overlap_norm	-0.0802

recursive depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0292
negmean	0.0000
posstd	0.0967
negstd	0.0001
overlap	-0.0676
overlap_norm	-0.0676

robinhood depth 1

maxtotal	1.0000
mintotal	0.0000
posmean	0.0286
negmean	0.0003
posstd	0.0905
negstd	0.0005
overlap	-0.0626
overlap_norm	-0.0626

robinhood depth 2

maxtotal	1.0000
mintotal	0.0000
posmean	0.0771
negmean	0.0082
posstd	0.1419
negstd	0.0232
overlap	-0.0963
overlap_norm	-0.0963

Βιβλιογραφία

Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods, 2010.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Doug Beeferman. Lexical discovery with an enriched semantic network. In *In Proceedings of the ACL/COLING Workshop on Applications of WordNet in Natural Language Processing Systems*, pages 358–364, 1998.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

Vincent D. Blondel. Automatic extraction of synonyms in a dictionary. In *in Proceedings of the SIAM Text Mining Workshop*, 2002.

Eugene Charniak. Statistical techniques for natural language parsing. *AI Magazine*, 18:33–44, 1997a.

Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. pages 598–603. AAAI Press/MIT Press, 1997b.

Eugene Charniak. A maximum-entropy-inspired parser. pages 132–139, 1999.

Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *In PASCAL Workshop on Learning Methods for Text Understanding and Mining*, 2004.

Ido Dagan and Oren Glickman. The pascal recognising textual entailment challenge. In *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/, 2004.

Casper Constantijn De Jonge. *Between grammar and rhetoric: Dionysius of Halicarnassus on language, linguistics, and literature*, volume 301. Brill, 2008.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *IN PROC. INT'L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC*, pages 449–454, 2006.

Marie-Catherine de Marneffe, Trond Grenager, Bill Maccartney, Daniel Cer, Daniel Ramage, Chloé Kiddon, and Christopher D. Manning. Aligning semantic graphs for textual inference and machine reading. In *In Proc. of the AAAI Spring Symposium at*, 2007.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *In ACL 2008*, 2008.

Steven J. DeRose. Grammatical category disambiguation by statistical optimization. *Comput. Linguist.*, 14(1):31–39, jan 1988. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=49084.49087.

Philip Edmonds and Graeme Hirst. Near-synonymy and lexical choice. *Computational Linguistics*, 28: 105–144, 2002.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited, 2001.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing, 2006.

Taku Kudo and Yuji Matsumoto. Chunking with support vector machines, 2001.

Christopher D. Manning. Part-of-speech tagging from 97 linguistics?, 2011.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993.

M.F.Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, 7(2):245—276, 2011.

George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38: 39–41, 1995.

M. Lynne Murphy. *Semantic Relations and the Lexicon*. Cambridge University Press, 2003. ISBN 9780511486494. URL http://dx.doi.org/10.1017/CBO9780511486494. Cambridge Books Online.

Alan J. Perlis. Special feature: Epigrams on programming. *SIGPLAN Not.*, 17(9):7–13, September 1982. ISSN 0362-1340. doi: 10.1145/947955.1083808. URL http://doi.acm.org/10.1145/947955.1083808.

Alan Ritter, Doug Downey, Stephen Soderland, and Oren Etzioni. It's a contradiction—no, it's not: A case study using functional relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 11–20, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613718.

Ravi Sinha and Rada Mihalcea. Using centrality algorithms on directed graphs for synonym expansion, 2011.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*, 2013.

Wen tau Yih and Vahed Qazvinian. Measuring word relatedness using heterogeneous vector space models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 616–620, 2012.

Ellen M. Voorhees. Contradictions and justifications: Extensions to the textual entailment task, 2008.