ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Extracting graph-structured information from simple text

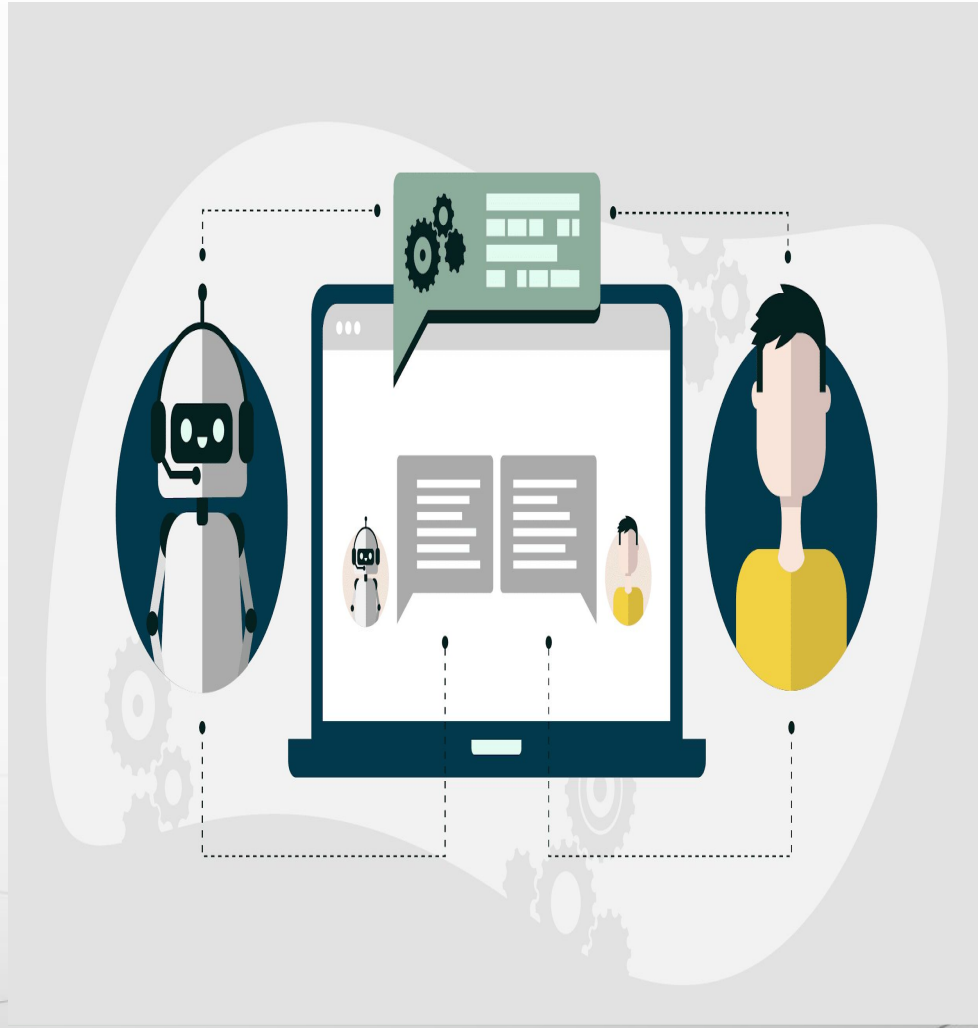
Ιωακειμίδου Δέσποινα

MSc Thesis

Επιβλέπουσα Καθηγήτρια: Κολωνιάρη Γεωργία

Περιεχόμενα Παρουσίασης

- Εισαγωγή
- Επεξεργασία Φυσικής Γλώσσας
- Βάσεις Δεδομένων
- Θεωρία Γράφων
- Προεπεξεργασία Κειμένου
- Μέθοδοι Προ-επεξεργασίας
- Μεθοδολογία
- Αξιολόγηση Εφαρμογής
- Συμπεράσματα -Προτάσεις για Περαιτέρω Έρευνα
- Βιβλιογραφία



Εισαγωγή

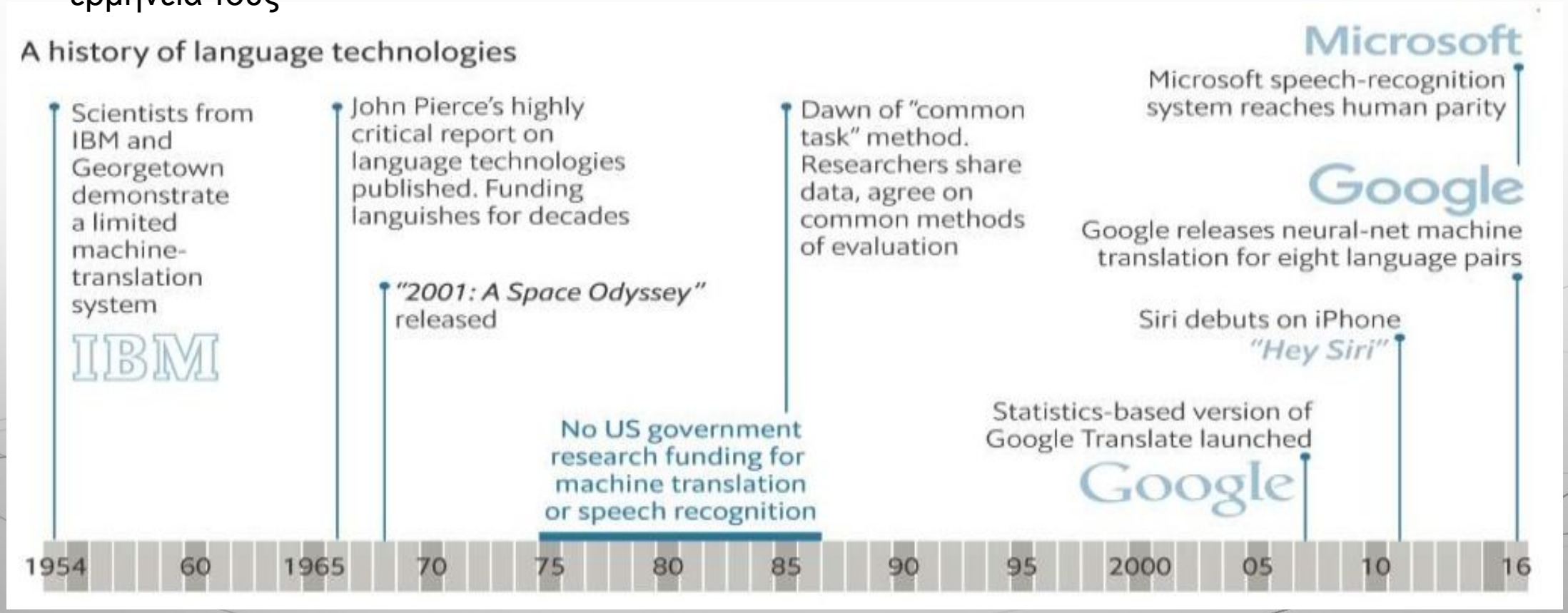
- **Μεγάλος Όγκος Δεδομένων.**
  Καθημερινά δημιουργούνται δισεκατομμύρια Megabytes δεδομένων.

- **Διαδικασία μετατροπής δεδομένων σε πληροφορία.**
  Η επεξεργασια, η κατηγοριοποίηση των δεδομένων οδηγεί σε χρήσιμη πληροφορία.

- **Επιλογή σωστού τύπου βάσης δεδομένων.**
  Τα δεδομένα πλέον είναι πολύτιμα, η σωστή αποθήκευση τους είναι αναγκαία.

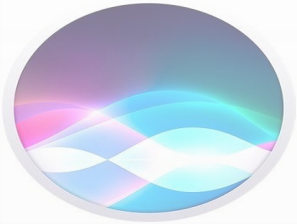
Επεξεργασία Φυσικής Γλώσσας – Natural Language Processing

- Διεπιστημονικός Τομέας της γλωσσολογίας και της πληροφορικής
- Διεπαφή ανθρώπου-μηχανής
- Πολύπλοκη διαδικασία, ξεκινά με την επεξεργασία των δεδομένων και έχει ως αποτέλεσμα την ερμηνεία τους



Εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας

- ✔ Machine Translation
- ✔ Speech Recognition
- ✔ Sentiment Analysis
- ✔ Question Answering
- ✔ Automatic Summarization
- ✔ Chatbots
- ✔ Market Intelligence
- ✔ Text Classification
- ✔ Character Recognition
- ✔ Spell Checking

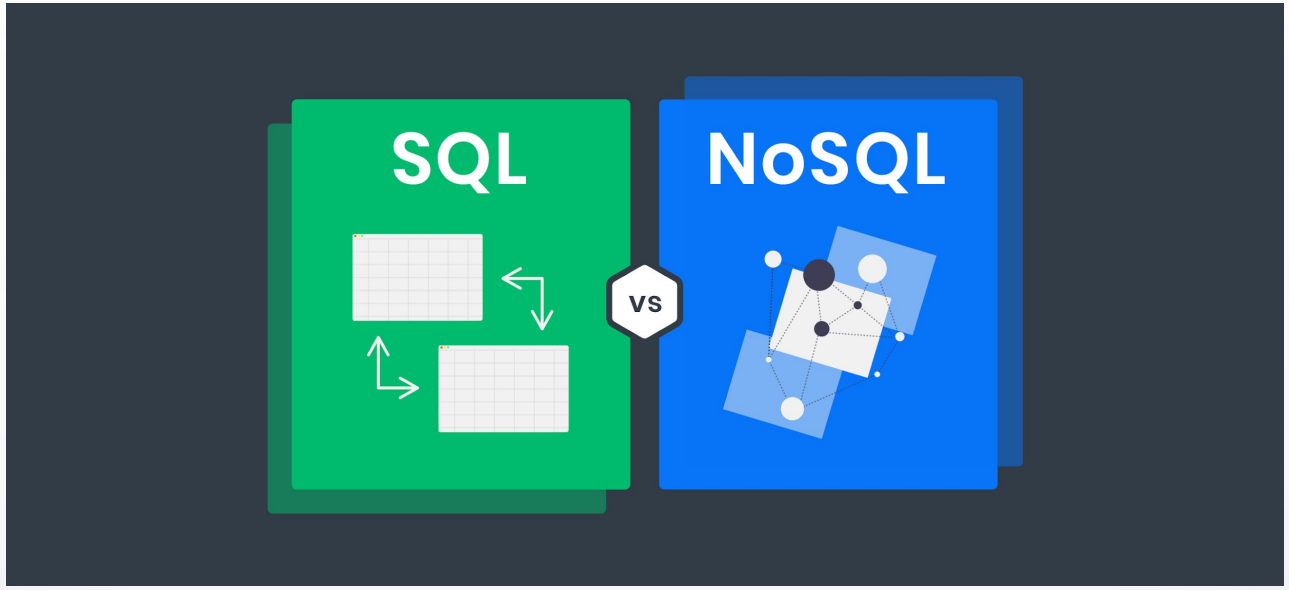


Hey Siri

Google Translate

Βάσεις Δεδομένων

Βάση Δεδομένων: συλλογή από σχετιζόμενα δεδομένα.



Οι Σχεσιακές Βάσεις Δεδομένων (ΣΒΔ) αποτελούνται από ένα σύνολο πινάκων στους οποίους αποθηκεύονται τα δεδομένα.

Vs

Οι Βάσεις Δεδομένων NoSQL αποτελούνται από αδόμητη ή ημιδομημένη αναπαράσταση δεδομένων, πράγμα το οποίο σημαίνει ότι η δομή καθορίζεται ανάλογα με τις ανάγκες που προκύπτουν.

Θεωρία Γράφων

- Γράφος είναι το διατεταγμένο ζεύγος G=(V,E), όπου V σύνολο σημείων και E διμελής σχέση πάνω στο V.
- Τα V καλούνται κορυφές και τα στοιχεία του E καλούνται ακμές που είναι ευθύγραμμα ή καμπύλα τμήματα με άκρα ένα ή δύο στοιχεία του συνόλου V, ενώ αποτελούν μια αναπαράσταση συσχετισμένων εννοιών.

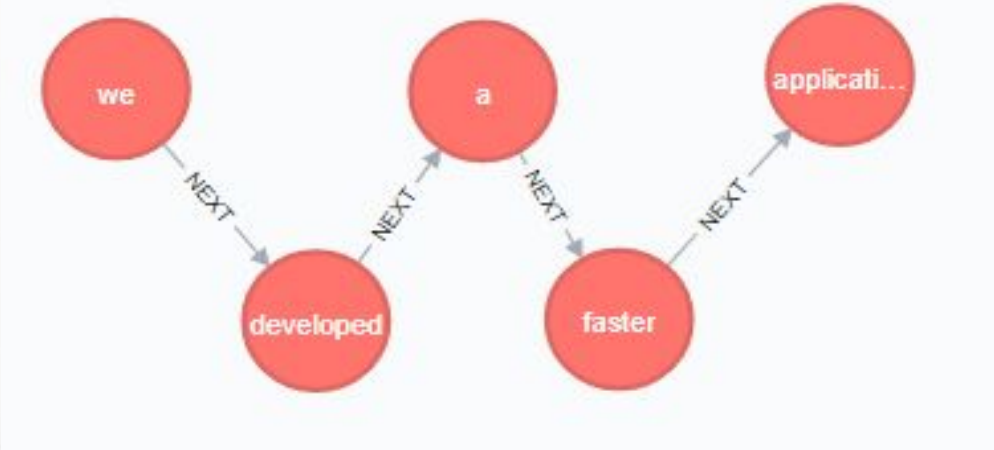
Βάσεις Δεδομένων με Γράφους

- Ένας γράφος αποτελείται από κόμβους (nodes) και ακμές (edges), όπου οι κόμβοι αντιπροσωπεύουν συνήθως τις οντότητες (entities) ενός πεδίου ορισμού (domain) και οι ακμές τις σχέσεις (relationships) που τους συνδέουν.
- Τόσο οι κόμβοι, όσο και οι ακμές μπορούν να αντιπροσωπεύουν και να αποθηκεύουν δεδομένα.
- Εμείς θα κάνουμε χρήση του μοντέλου γράφων με ετικέτες και ιδιότητες (labeled property graph model) με την Neo4j.

Στόχος

Λειτουργικότητα στην Neo4j

- Οι λέξεις της προτασης συνδέονται μεταξύ τους με έναν τύπο σχέσης ‘NEXT’.
- Όλες οι λέξεις χρησιμοποιούνται ακριβώς με τον ίδιο τρόπο.



- Βελτίωση παραγόμενου γράφου

- Απαλοιφή λέξεων που προκαλούν θόρυβο.
- Διαφοροποίηση στην χρήση των λέξεων ανάλογα με το μέρος του λόγου στο οποίο ανήκουν.
- Δημιουργία διαφορετικού τύπου σχέσεων με την χρήση των ρημάτων.
- Δημιουργία κόμβων με ιδιότητες.

Προ-επεξεργασία Κειμένου

Διαδικασία με την οποία οι οντότητες ενός κειμένου διαχωρίζονται με στόχο την αύξηση της αποτελεσματικότητας της εφαρμογής.

Α) Λεξιλογική ανάλυση
- Part-of-speech tagging (Αναγνώριση μέρους του λόγου της κάθε λέξης)
- Tokenization (Αυτόματος χωρισμός σε λέξεις και προτάσεις)

Β) Αποκλεισμός λέξεων
- Stopwords removal (Απαλοιφή λέξεων με πολύ μικρή διακριτική ικανότητα)

Γ) Στελέχωση των εναπομεινουσών λέξεων
- Stemming ή lemmatization

Δ) Επιλογή των λέξεων που θα χρησιμοποιηθούν
- Tag-patterns (αποτελούν μια αλληλουχία από χαρακτήρες ή μοτίβο και χρησιμοποιούνται για την αντιστοίχιση σε δεδομένα κειμένου)

Μέθοδοι Προεπεξεργασίας

- Tokenization

Διαδικασία χωρίσματος του κειμένου σε τμήματα. Το μικρότερο τμήμα με νόημα είναι η λέξη. Στην εφαρμογή ο διαχωρισμός γίνεται σε κάθε λέξη της κάθε πρότασης χωρίς όμως να αποσπάται από την πρόταση.

- Part-of-speech tagging

Η διαδικασία αναγνώρισης του μέρους του λόγου που ανήκει η κάθε λέξη. Στην εφαρμογή χρησιμοποιούνται κάποια από τα μέρη του λόγου όπως αναφέρονται στο Penn tree bank .

- Stopwords removal

Διαδικασία με την οποία αφαιρείται η μη χρήσιμη πληροφορία για την συνέχεια της επεξεργασίας του κειμένου. Μη χρήσιμη πληροφορία αποτελούν οι πιο κοινές λέξεις όπως τα άρθρα. Χρησιμοποιείται το σύνολο των stopwords μέσα από την εργαλείο nltk.

Μέθοδοι Προεπεξεργασίας

- Κανονικές Εκφράσεις - Tag-Patterns

Κανονικές εκφράσεις ορίζουν αλληλουχίες μεταξύ των μερών του λόγου.

π.χ.

NP: <NN><VB><NN>

Αποτελεί μια αλληλουχία ενός ουσιαστικού, ενός ρήματος και ενός ουσιαστικού.
Στην εφαρμογή τα tag-patterns αποτελούν τους κανόνες που δημιουργήθηκαν.

Κανόνες που αναπτύχθηκαν

Οι κανόνες που αναπτύχθηκαν είναι 9:

- "NP1:{(<JJ.?><,>)*<JJ.?>+<NN.?>}"
- "NP2:{<NN.?><NN>}"
- "NP3:{<NN.?><VB.?><NN.?><.>}"
- "NP4:{<NN.?><VB.?><JJ.?><NN.?>}"
- "NP5:{<PRP><VB.?><RB.?><NN.?><.>}"
- "NP6:{<PRP><VB.?><NN.?><.>}"
- "NP7:{<PRP><VB.?><JJ.?><NN.?><.>}"
- "NP7:{<PRP><VB.?><JJ.?><NN.?><.>}"
- "NP9:{<PRP><VB.?><NN.?><RB.?><.>}"

Χωρίζονται σε δύο κατηγορίες: στους κανόνες που δημιουργουν σχέσεις και σε αυτους που δημιουργούν κόμβους.
Το ερωτηματικό (?) μετα την τελεία αναφέρεται σε όλα τα είδη του συγκεκριμένου μέρους του λόγου.
Δημιουργήθηκαν μετα απο μελέτη απλών προτάσεων και της συντακτικής μορφής τους.
Ο κανόνας 1 είναι ο πιο γενικός κανόνας που κατασκευάστηκε.
Μία πρόταση μπορεί να περιέχει παραπάνω από έναν κανόνα.

Αντιστοίχιση Κειμένου σε Κανόνες

Η διαδικασία με την οποία οι κανόνες αντιστοιχούν στο κείμενο και το κείμενο αντιστοιχεί στα μέρη του γράφου.

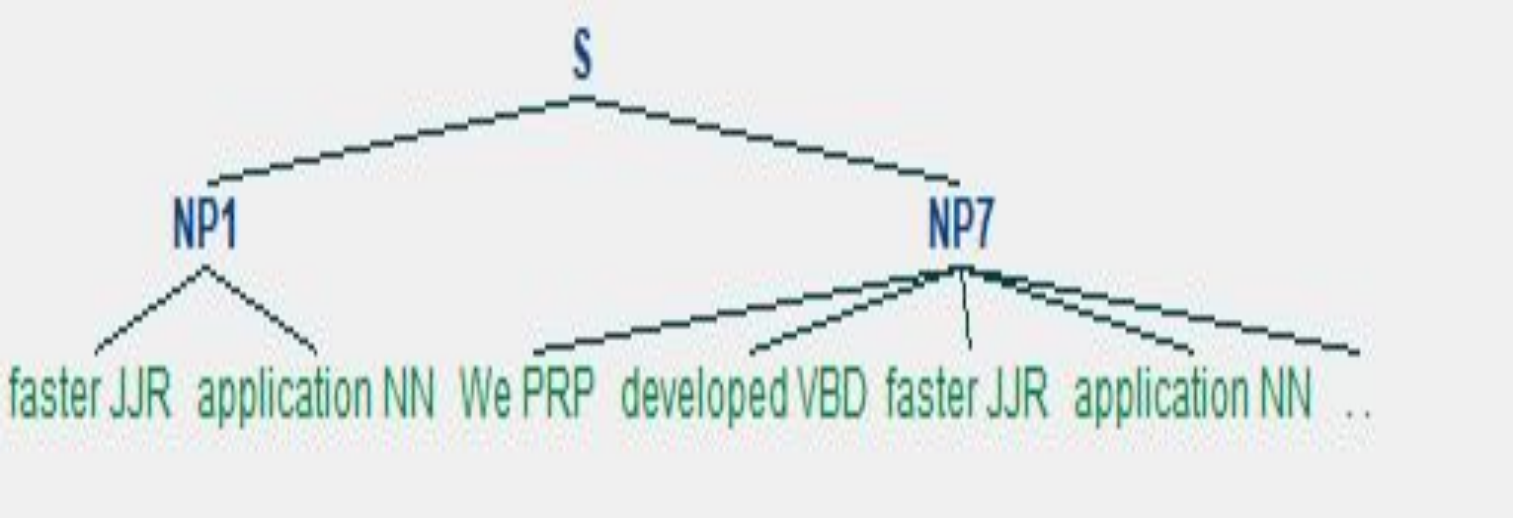
- NP:{<PRP><VB.?><JJ.?><NN.?>}
  Αυτό το tag pattern αποτελεί μία αλληλουχία προσωπικής αντωνυμίας, ενός ρήματος οποιουδήποτε είδους, ενος επιθέτου οποιουδήποτε είδους και ενός ουσιαστικού οποιουδήποτε είδους.

Κείμενο εισόδου : “We developed a faster application.”

- Αποτελεί μια κύρια πρόταση με ρήμα (VB) το “developed”, υποκείμενο το “We” (PRP) και αντικείμενο “application” (NN) με επίθετο το “faster”(JJ).

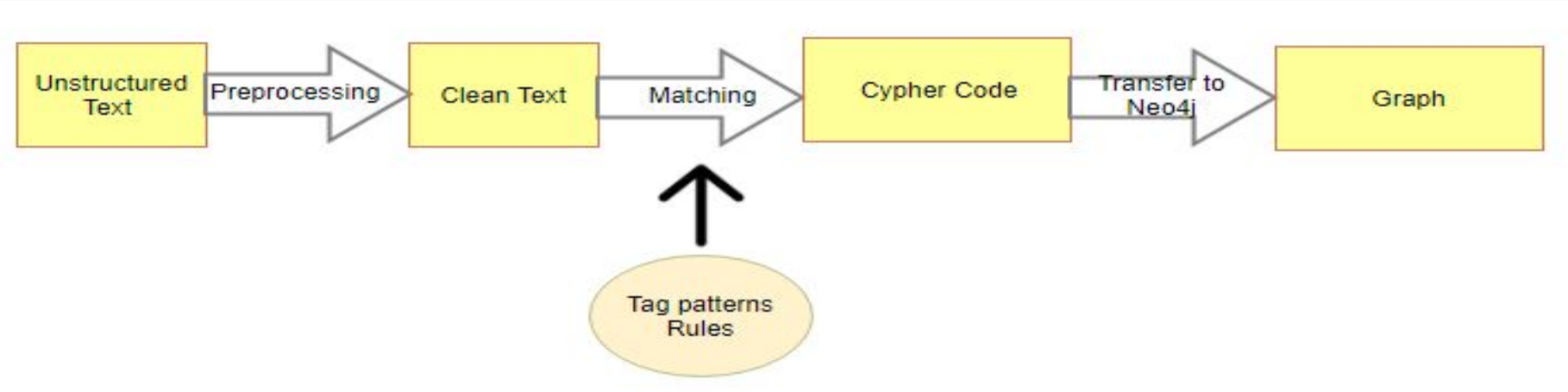
Συντακτικό Δένδρο - Parse tree

Αποτελεί μια ιεραρχική δομή που αντιπροσωπεύει την διάσπαση της πρότασης ως προς τους κανόνες.
Ρίζα του δέντρου στην περίπτωση μας θα είναι η πρόταση (S) και σαν κόμβος - γονέας είναι το όνομα του κάθε κανόνα που θα αντιστοιχεί στο μέρος της πρότασης ενώ φύλλο είναι η λέξη και το μέρος του λόγου που ανήκει. Κατα την υλοποίηση το δένδρο είναι πολύ σημαντικό καθώς αναφερόμαστε στις λέξεις που χρησιμοποιούνται από το κείμενο με βαση την θέση τους στο δένδρο.



Μεθοδολογία

Διάγραμμα της Μεθοδολογίας που αναπτύχθηκε για την εφαρμογή.



Διαδικασία Υλοποίησης και Αξιολόγησης

- Για την υλοποίηση της εφαρμογής έγινε χρήση των παρακάτω εργαλείων:
  - Neo4j (1.1.7)
  - NLTK (3.2.5)
  - Python
  - Pycharm (Edu 2017.3)
  - Py2neo (2.0.8)

- Για την αξιολόγηση της εφαρμογής χρησιμοποιήθηκε μία συλλογή δεδομένων από τρία κείμενα που είχαν ως θέμα την κλιματική αλλαγή.

- Για την μέτρηση της επίδοσης του μοντέλου μας, αξιολογήθηκε η ποσοστιαία μεταβολή της χρήσης των λέξεων, των ρημάτων και των κανόνων που δημιουργήθηκαν στον τελικό γράφο.

Αξιολόγηση – Παράδειγμα

Αδόμητο κείμενο: *We developed a faster application .*
Κείμενο μετά το stop word removal: *We developed faster application.*
Part-of-speech-tagging: [('We', 'PRP'), ('developed', 'VBD'), ('faster', 'JJR'), ('application', 'NN'), ('.', '.')]
Cypher:
(a:Person),(b:Person) WHERE a.name = 'We'AND b.name = 'application' CREATE (a)-[r:developed]->(b) RETURN type(r)

Τελικός Γράφος

Αξιολόγηση – 1η Περίπτωση

Our generation is facing the greatest threat. Human activity has increased the average temperature of the Earth. Climate change is affecting wildlife, many animals face risk. Planet Earth needs our help. We have to act efficiently. We have to change quickly our lifestyle. Planet Earth is a unique and beautiful home. Some easy changes can make a big difference. We should all of us use reusable cups. Single-use plastic are made from fossil-fuel. We should think our transportations very carefully. The unnecessary use of cars is a major cause of global warming. We have to work for the environment now. The environment improves the quality of life.

Αξιολόγηση – 1η Περίπτωση

- Σύγκριση του τελικού γράφου προερχόμενου από το ίδιο κείμενο που δημιουργείται απο την Neo4j και την εφαρμογή. Το κείμενο αποτελείται από 107 λέξεις.

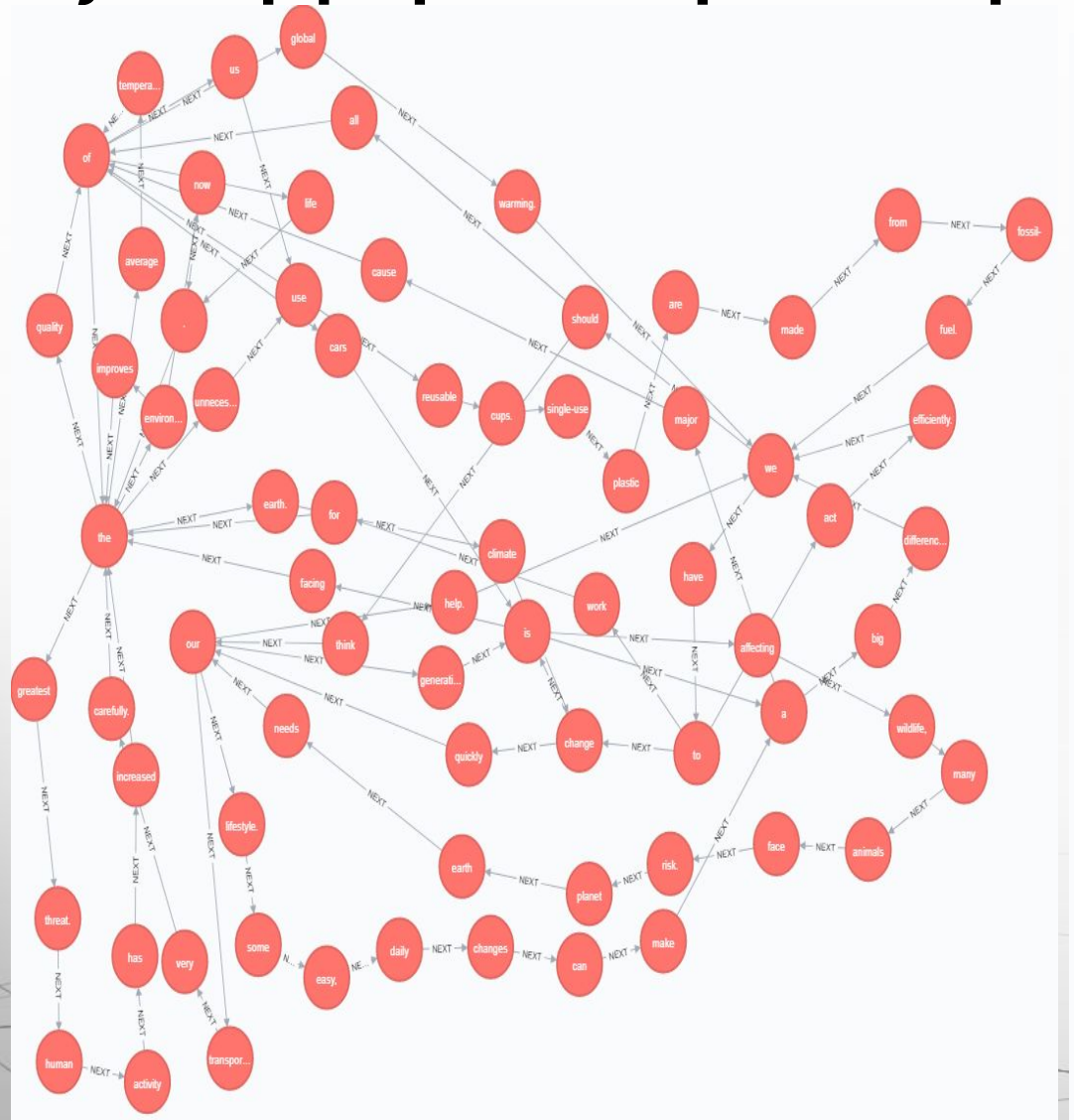
- Ο γράφος που δημιουργείται από την βάση περιέχει 76 κόμβους και 98 σχέσεις.
- Είναι πολύπλοκος
- Αποτελεί μια απεικόνιση της αλληλουχίας των λέξεων που υπάρχουν στο κείμενο και όχι των σχέσεων που υπάρχουν μεταξύ τους.
- Είναι δυσανάγνωστος.

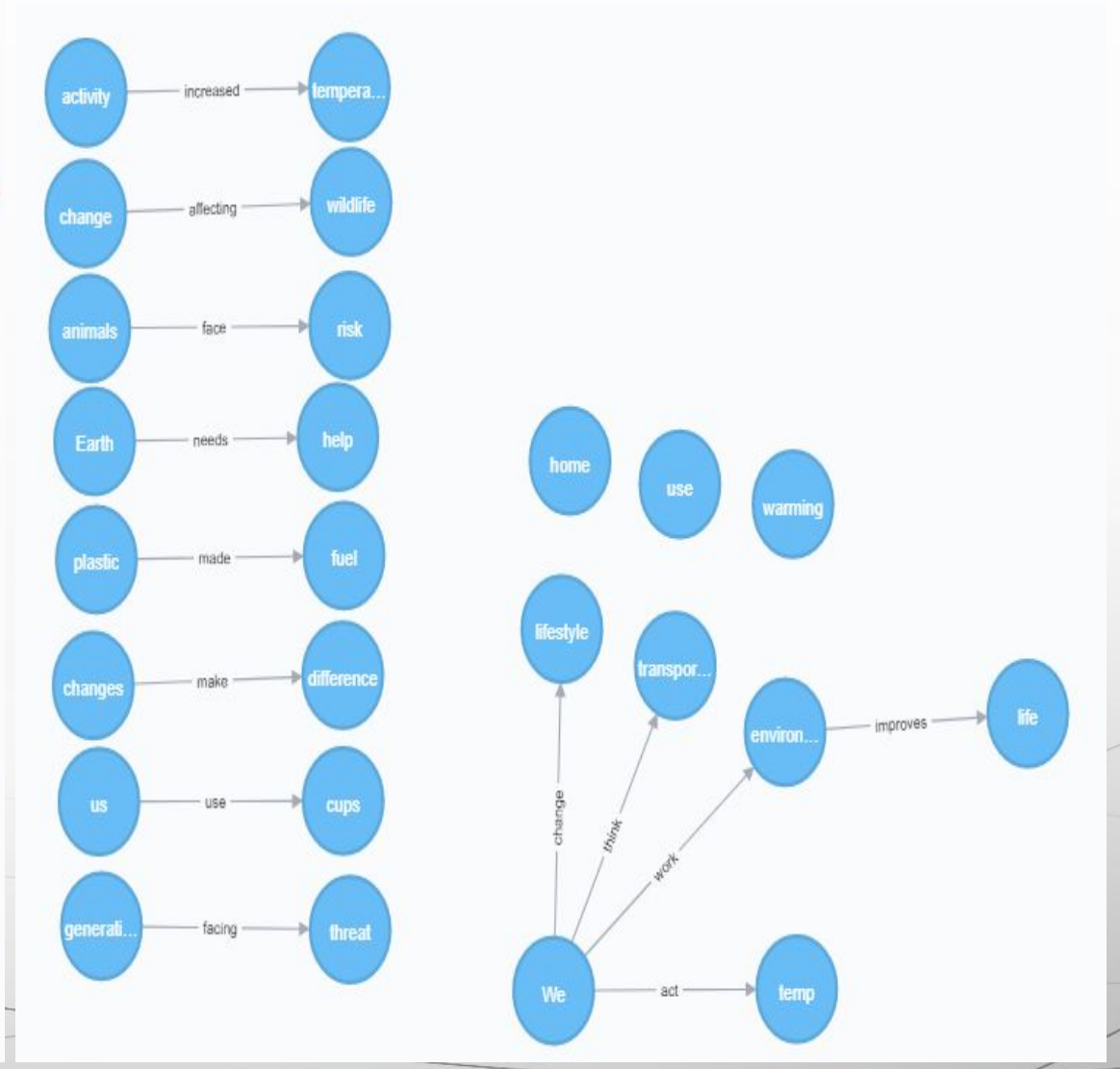
Αξιολόγηση – 1η Περίπτωση

- Το μέγεθος του κειμένου μέσω της προεπεξεργασίας και συγκεκριμένα μέσω του stopwords removal μειώνεται κατά 37,3% .

- Χρησιμοποιούνται όλοι οι κανόνες τουλάχιστον μια φορά.
- Όλες οι προτάσεις του κειμένου ανταποκρίνονται τουλάχιστον σε έναν κανόνα.
- Οι πιο συχνοί κανόνες ως προς την χρήση τους είναι αυτοί που προσθέτουν ιδιότητες σε έναν κόμβο.
- Ο τελικός γράφος αποτελείται από 25 κόμβους και 13 σχέσεις.

Αξιολόγηση – 1η Περίπτωση





Αξιολόγηση – 2η Περίπτωση

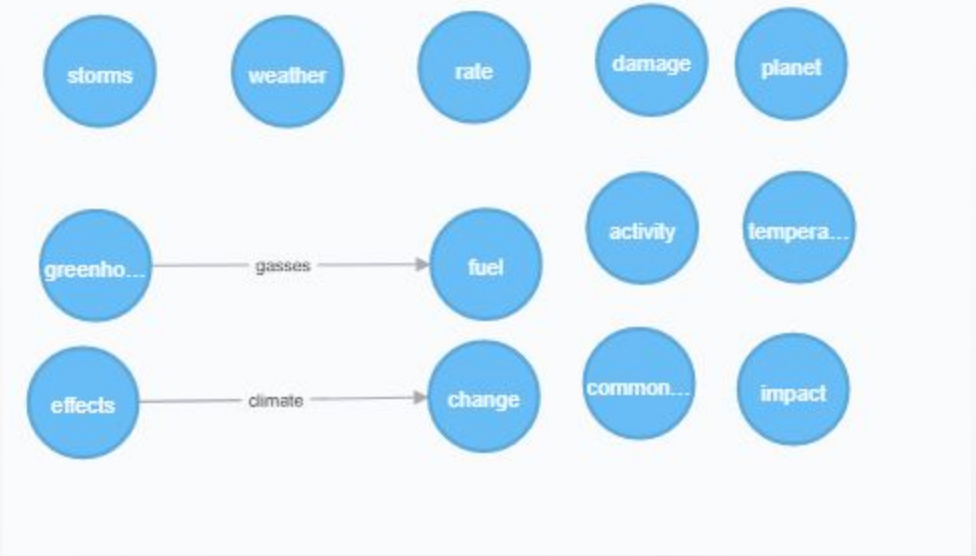
It's real. It's happening. It's accelerating. And it's our fault. Human activity — particularly the production of greenhouse gasses fossil fuel emissions — is reshaping our planet, effecting rapid environmental change at a rate never seen before. Global temperature averages are creeping upward, seas are warming, rising and becoming more acidic, and extreme weather events such as droughts, wildfires, floods and powerful storms are more commonplace. Here's where you'll find the latest on the effects of climate change, and the measures that scientists, world leaders and innovators are taking to reduce our harmful impact on the planet and mitigate the damage already done.

Αξιολόγηση – 2η Περίπτωση

- Το κείμενο που χρησιμοποιείται είναι διαφορετικό από την πρώτη περίπτωση.
- Κείμενο που δεν ανταποκρίνεται στην συντακτική μορφή με την οποία έχουν δημιουργηθεί οι γράφοι. Περιέχει δευτερεύουσες και μεγάλες προτάσεις. Το μέγεθος του κειμένου είναι 102 λέξεις.
- Μετα την προεπεξεργασία το κείμενο μειώνεται κατα 31%.

- Από το σύνολο των κανόνων χρησιμοποιούνται οι 4 και συνολικα πραγματοποιούνται μόνο 15 κανόνες
- Δημιουργούνται 13 κόμβοι και 2 σχέσεις
- Υπάρχουν ολόκληρες προτάσεις που δεν ανταποκρίνονται σε κανένα κανόνα.
- Ο γράφος που παράγεται είναι ελλιπής ως προς την πληροφορία του κειμένου.

Αξιολόγηση – 2η Περίπτωση



Αξιολόγηση – 3η Περίπτωση

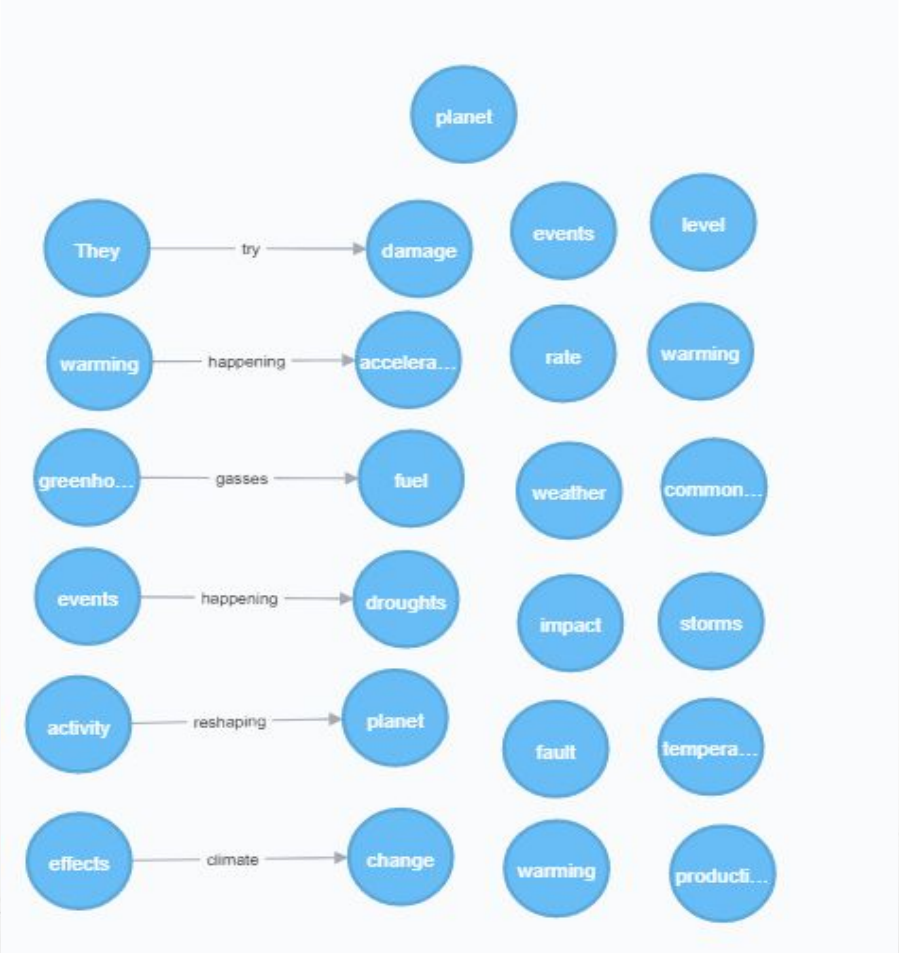
Climate change is real. Climate change is happening. Climate change is accelerating and is our fault. Human activity is reshaping our planet. Particularly the production of greenhouse gasses from fossil fuel emissions . Human activity is affecting rapid environmental change at a rate never seen before. Global temperature averages are creeping upward. Seas are warming. Seas are rising and becoming more acidic. Extreme weather events are happening such as droughts, wildfires, floods and powerful storms . Those weather events are more commonplace. Here is where you will find the latest on the effects of climate change. Here are the measures. The scientists, world leaders and innovators are taking to reduce our harmful impact on the planet. They try to mitigate the damage.

Αξιολόγηση – 3η Περίπτωση

- Στο κείμενο που χρησιμοποιείται στην δεύτερη περίπτωση πραγματοποιήθηκε η κατάλληλη προεπεξεργασία, βελτίωση της συντακτικής του δομής ετσι ώστε να ανταποκριθεί στους κανόνες.
- Οι δευτερεύουσες προτάσεις εξαλείφθησαν και στις περιπτώσεις που το υποκείμενο εννοούνταν, προστέθηκε στην πρόταση κανονικά.
- Μετα την προεπεξεργασία το κείμενο μειώνεται κατα 36,2%.

- Χρησιμοποιούνται 5 απο τους 9 κανόνες και συνολικά πραγματοποιούνται 29 κανόνες
- Οι πιο συχνοί κανόνες ως προς την χρήση τους είναι αυτοι που προσθέτουν ιδιότητες σε έναν κόμβο
- Ο τελικός γράφος αποτελείται από 25 κόμβους και 6 σχέσεις.

Αξιολόγηση – 3η Περίπτωση



Σύγκριση γράφων

- Το μέγεθος των κειμένων είναι σχεδόν ίσο
- Το ποσοστό μείωσης των λέξεων κυμαίνεται στα ίδια επίπεδα
- Ο αριθμός των nodes που δημιουργούνται στις περιπτώσεις των text_1 και text_3 είναι ο ίδιος. Πολύ μικρότερος στο κείμενο της δεύτερης περίπτωσης.
- Η βασική διαφορά είναι στην δημιουργία σχέσεων
- Ο αριθμό των σχέσεων που παράγονται απο το text_2 επισημαίνει την κακή απόδοση της εφαρμογής σε αυτό όπως και το ποσοστό των ρημάτων που χρησιμοποιούνται για την δημιουργία σχέσεων.
- Οι κανονες σε σύνολο που χρησιμοποιούνται στο text_1 και στο text_3 είναι σχεδον ίσοι αλλα διαφέρει το ποιοί χρησιμοποιήθηκαν σε κάθε περίπτωση.

	$text_1$	$text_2$	$text_3$
number of words	107	102	120
number of words after stop words removal	67	71	76
% reduction of words	37,3%	31%	36,6%
nodes	25	15	25
relationships	13	2	6
% of verbs that creates relationship	100%	16,6%	50%
nodes with properties	14	13	17
number of rules used	29	15	27

Συμπεράσματα

- Η σωστή προ-επεξεργασία του κειμένου είναι αρκετά σημαντική για την επιτυχία της εφαρμογής.
- Η εφαρμογή που έχει δημιουργηθεί αποδίδει αρκετά καλά σε περιπτώσεις όπου η συντακτική μορφή του κειμένου συμφωνεί με αυτην των κανόνων που υπάρχουν.
- Ακόμη και σε κείμενα που δεν έχουν την κατάλληλη μορφή η εφαρμογή δίνει κάποια αποτελέσματα.
- Η εφαρμογή αποτελεί μια πρώτη προσπάθεια και είναι αρκετά ελπιδοφόρα καθώς αξιοποίησε αρκετή πληροφορία και ο τελικός γράφος που δημιουργεί βοηθά στην περαιτέρω ανάλυση των δεδομένων κειμένου.

Προτάσεις για Περαιτέρω Έρευνα

- Αναπτυξη κανόνων για περισσότερα μέρη του λόγου
- Διαφορετική προσέγγιση στην ανάπτυξη των κανόνων
- Συνεργασία με γλωσσολόγο για την καλύτερη κατανόηση της σύνταξης της πρότασης.

Σας ευχαριστώ για τον χρόνο σας!

Βιβλιογραφία

1. Σακελλαρίου Ηλίας, Βασιλειάδης Νικόλαος, Κεφαλάς Πέτρος, Σταμάτης Δημοσθένης “Επεξεργασία Φυσικής Γλώσσας και Γραμματικές Οριστικών Προτάσεων.”Κεφάλαιο Συγγράμματος ,2015.
2. Γιάννης Τζίτζικας , Υλικό μαθήματος “Συστήματα Ανάκτησης Πληροφοριών Information Retrieval (IR) Systems” , Πανεπιστήμιο Κρήτης Τμήμα Επιστήμης Υπολογιστών, 2009.
3. Ισίδωρος Περίκος, Διαδακτορική Διατριβή “Τεχνικές Επεξεργασίας Φυσικής Γλώσσας, Εξαγωγής Γνώσης και Αυτόματης Δημιουργίας Προσαρμοσμένης στον Χρήστη Ανάδρασης για Ευφυή Συστήματα Διδασκαλίας” Πανεπιστήμιο Πατρών Πολυτεχνική Σχολή Τμήμα Μηχανικών Η/Υ & Πληροφορικής, 2016.