Ανάκτηση πληροφορίας και Επεξεργασία Φυσικής Γλώσσας

Ηρακλής Βαρλάμης

Οργανωτικά

Οργάνωση του μαθήματος

- Διδακτικά βοηθήματα
- Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2008. http://nlp.stanford.edu/IR-book/
- Search Engines: Information Retrieval in Practice. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009. http://ciir.cs.umass.edu/downloads/SEIRiP.pdf
- Ανάκτηση Πληροφορίας. Baeza-Yates Ricardo, Ribeiro-Neto Berthier, Έκδοση: 2η, 2014, ΕΚΔΟΣΕΙΣ Α. ΤΖΙΟΛΑ & ΥΙΟΙ Α.Ε.
- Papadopoulos, A., Manolopoulos, I., & Tsichlas, K. (2015). Information Retrieval [Undergraduate textbook]. Kallipos, Open Academic Editions. https://dx.doi.org/10.57713/kallipos-600

NLP textbooks (optional):
– Speech & Language Processing by D. Jurafsky and J.H. Martin, 2η edition, Pearson, 2009. (Free draft 3rd edition: http://web.stanford.edu/~jurafsky/slp3)
– Neural Network Models for Natural Language Processing, by Y. Goldberg, Morgan & Claypool, 2017.
– Introduction to Natural Language Processing by J. Eisenstein, MIT Press, 2019. (Free draft: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf)
– Foundations of Statistical Natural Language Processing by C.D. Manning and H. Schutze, MIT Press, 1999.
– Deep Learning in Python, by F. Chollet, Manning Publications, 2nd edition, 2021. (Free 1st edition: https://www.manning.com/books/deep-learning-with-python. But 2nd edition highly recommended.)
- Understanding Deep Learning, by S.J.D. Prince, MIT Press (in press). (Free pre-print: https://udlbook.github.io/udlbook/)
– Dive into Deep Learning, by Zhang et al. (Freely available: https://d2l.ai/)

Αξιολόγηση μαθήματος

- Εργασία
- Μέχρι 2 άτομα
- Επιλέγετε 1 από τα ακόλουθα 3 θέματα
  - Συσταδοποίηση ειδησεογραφικών άρθρων
  - Δημιουργία περιλήψεων (ή τίτλων) για ομάδες άρθρων
  - Δημιουργία μιας μηχανής αναζήτησης

  ή προτείνετε ένα δικό σας (σε συνεννόηση μαζί μου)

- Παρουσιάζετε τα αποτελέσματα σε 1 γραπτή αναφορά και 1 παρουσίαση στο τέλος του εξαμήνου
- Θα αξιολογηθεί ο βαθμός δυσκολίας όσων τεχνικών επιχειρήσετε, η δημιουργικότητα, η τεκμηρίωση των ενεργειών και αποτελεσμάτων σας, η παρουσίαση.

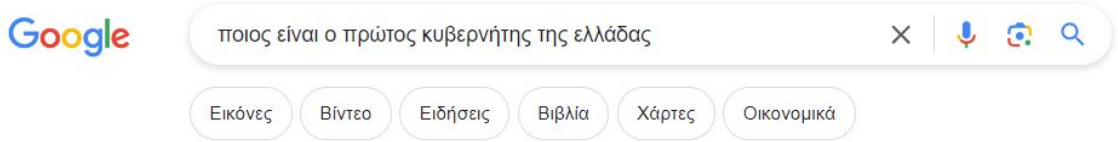
Ανάκτηση Πληροφορίας

Ανάκτηση πληροφορίας

- Πληροφοριακή ανάγκη = ερώτημα, τι ψάχνω να βρω
- Συλλογή = ένα δομημένο ή αδόμητο σύνολο από δεδομένα ή πληροφορίες

Μηχανή αναζήτησης

- Ένα πληροφοριακό σύστημα που συλλέγει και οργανώνει δεδομένα και πληροφορίες και μας επιτρέπει να καλύψουμε τις πληροφοριακές μας ανάγκες

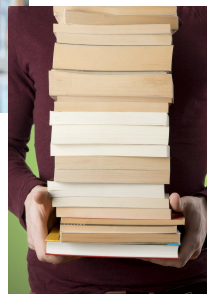


- Είναι μια βάση δεδομένων;

- Όχι ακριβώς. Όχι μόνο.

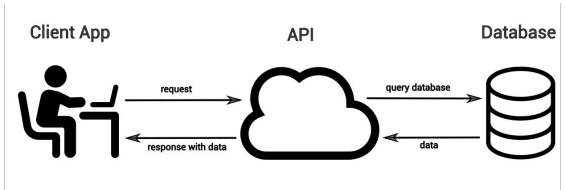
Τα βήματα

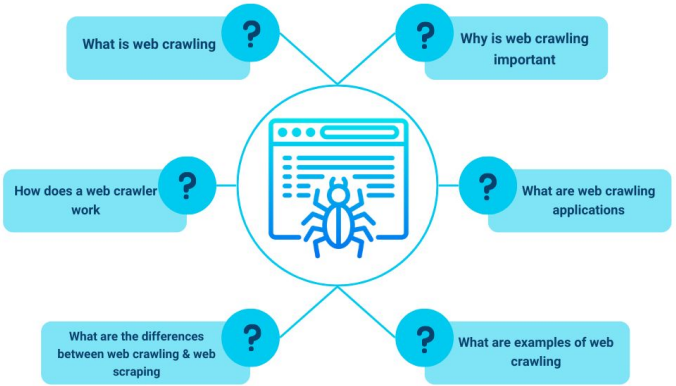
1. Συλλογή δεδομένων
2. Οργάνωση δεδομένων
3. Κατανόηση της πληροφοριακής ανάγκης
4. Αναζήτηση
5. Παρουσίαση αποτελεσμάτων



1. Συλλογή δεδομένων

- Μια λύση είναι να έχουμε κάποιον να μας δώσει τα δεδομένα σε ψηφιακή μορφή ή να τα ψηφιοποιήσουμε (π.χ. Μια συλλογή κειμένων, ιστοσελίδων, αλληλουχιών, εικόνων, κλπ)
- Μια άλλη λύση είναι να συλλέξουμε τα δεδομένα από τρίτες πηγές (π.χ. κάνοντας κλήσεις σε ένα API ή κάνοντας crawling ή scrapping)





2. Οργάνωση

- Ταξινόμηση αντικειμένων

Dewey Decimal Classification

000 GENERAL WORKS
020 Library and Information Science
030 General Encyclopedias
050 General Periodicals
060 General Organizations

100 PHILOSOPHY
110 Metaphysics
130 Psychology, occultism
140 Philosophy
150 Psychology
160 Logic

200 RELIGION
220 The Bible
230 Christian Doctrine
290 Comparative and Other Religions

300 SOCIAL SCIENCES
310 Statistics
320 Political Science
330 Economics
340 Law
350 Public Administration
360 Social Welfare
370 Education
380 Public Service
390 Customs and Folklore

400 LANGUAGE
410 Comparative Linguistics
420 English and Anglo Saxon
430 German Language
440 French
450 Italian, Rumanian
460 Spanish, Portuguese
470 Latin and Other Italic Languages
480 Classical and Modern Greek
490 Other Languages

500 SCIENCE
510 Mathematics
520 Astronomy
530 Physics
540 Chemistry
550 Earth Sciences
560 Paleontology
570 Life Sciences
580 Botanical Sciences
590 Zoological Sciences

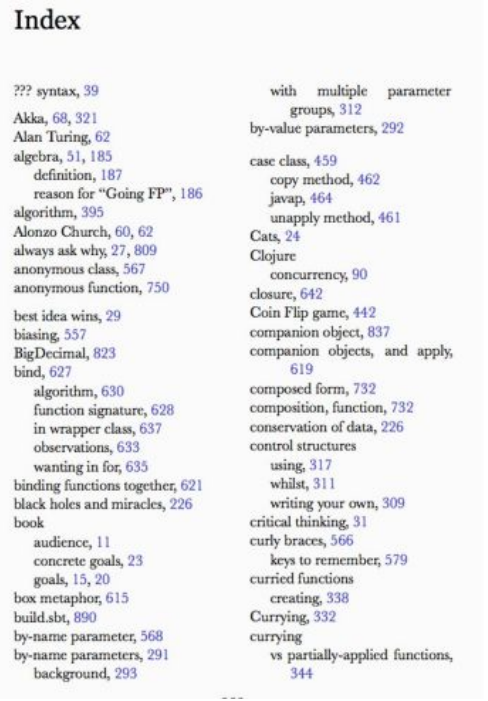
600 TECHNOLOGY
610 Medical Sciences
620 Engineering
630 Agriculture
640 Domestic Sciences
650 Business and Management
660 Chemical Technology
670 Manufacturers
690 Building Construction

700 THE ARTS
710 Landscape and Civic Art
720 Architecture
730 Sculpture, Plastics
740 Drawing, Decorative Arts
750 Painting
760 Prints and Print Making
770 Photography
780 Music
790 Recreation, Performing Arts

800 LITERATURE
810 American Literature
820 English Literature
830 German Literature
840 French Literature
850 Italian, Rumanian Literature
860 Spanish, Portuguese Literature
870 Latin and Other Italic Literatures
880 Classical and Modern Greek Literature
890 Other Literature

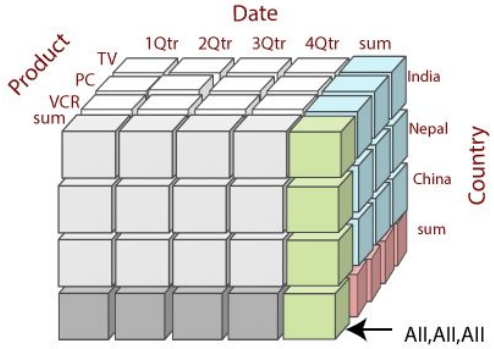
900 HISTORY, GEOGRAPHY
910 Geography, Travel
920 Biography, Genealogy
930 Ancient History
940 Europe
950 Asia
960 Africa
970 North America
980 South America
990 Pacific Ocean Islands
991 Indonesia
993 New Zealand and Melanesia
994 Australia
995 New Guinea (Papua)
996 Polynesia
997 Atlantic Ocean Islands
998 Arctic Region
999 Antarctic Regions

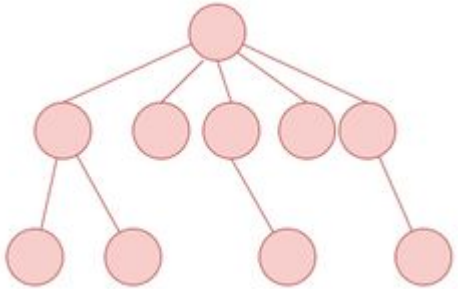
- Ευρετηρίαση δεδομένων



Index

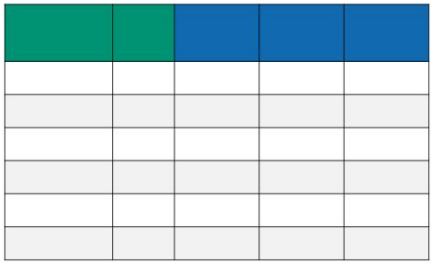
Σχήματα δεδομένων

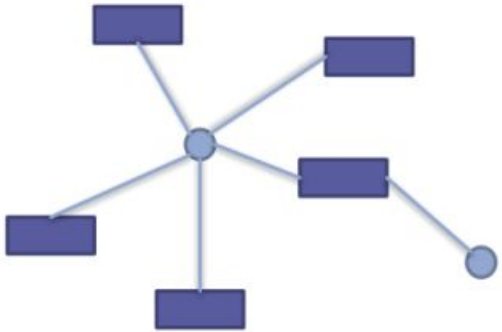




Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vel erat nec dui aliquet vulputate sed quis nulla. Donec eget ultricies magna, eu dignissim elit. Nullam sed urna nec nisl rhoncus ullamcorper placerat et enim. Integer varius ornare libero quis consequat. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean eu efficitur orci. Aenean ac posuere tellus. Ut id commodo turpis.

Praesent nec libero metus. Praesent at turpis placerat, congue ipsum eget, scelerisque justo. Ut volutpat, massa ac lacinia cursus, nisl dui volutpat arcu, quis interdum sapien turpis in tellus. Suspendisse potenti. Vestibulum pharetra justo massa, ac venenatis mi condimentum nec. Proin viverra tortor non orci suscipit rutrum. Phasellus sit amet euismod diam. Nullam convallis nunc sit amet diam suscipit dapibus. Integer porta hendrerit nunc. Quisque pharetra congue porta. Suspendisse vestibulum sed mi in euismod. Etiam a purus suscipit, accumsan nibh vel, posuere ipsum. Nulla nec tempor nibh, id venenatis lectus. Duis lobortis id urna eget tincidunt.





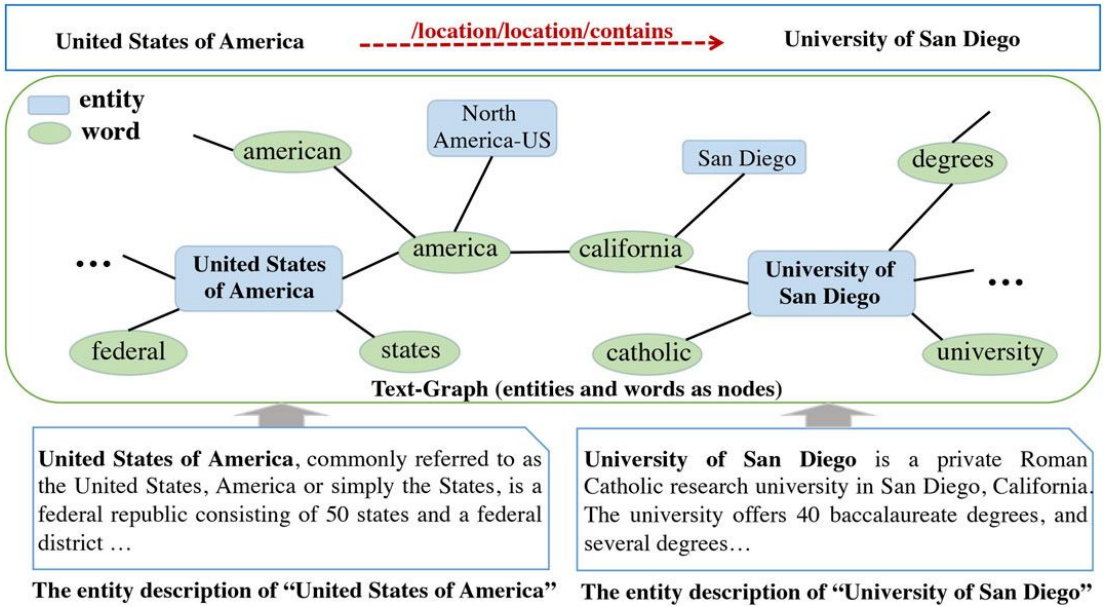
Δομημένα Δεδομένα

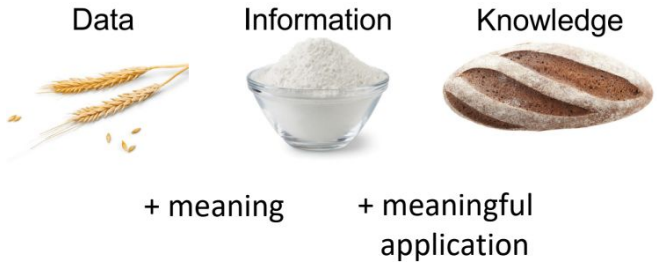
Ημι-Δομημένα Δεδομένα

Αδόμητα Δεδομένα

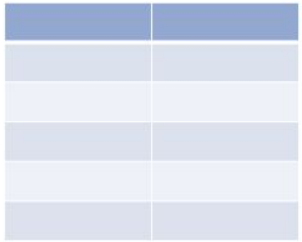
Εξαγωγή πληροφορίας - Information Extraction

- Η προσπάθεια εξαγωγής δομημένης ή ημι-δομημένης πληροφορίας από αδόμητα δεδομένα

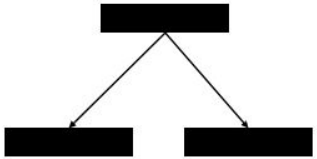




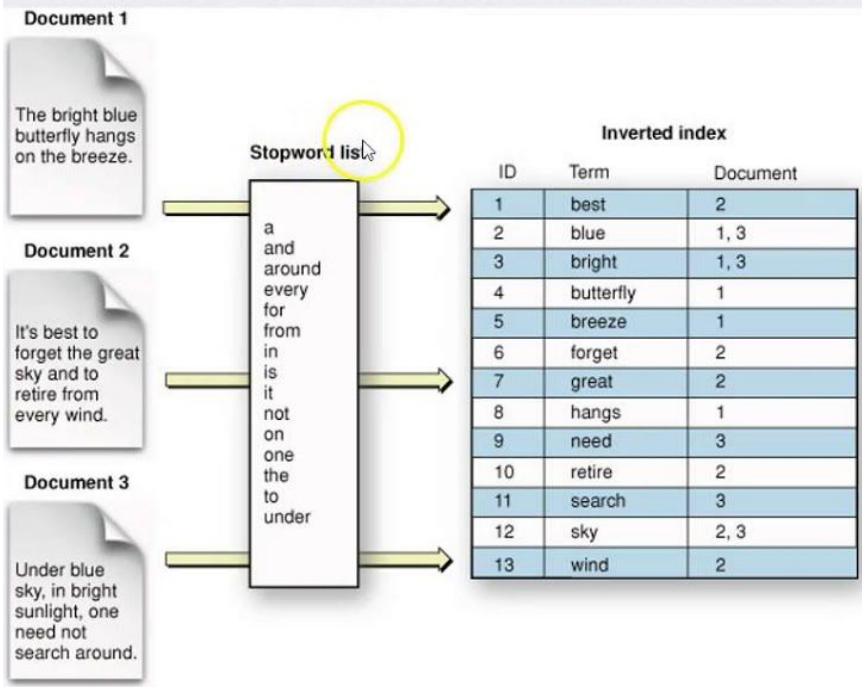
Ευρετηρίαση πληροφορίας



Πίνακας κατακερματισμού

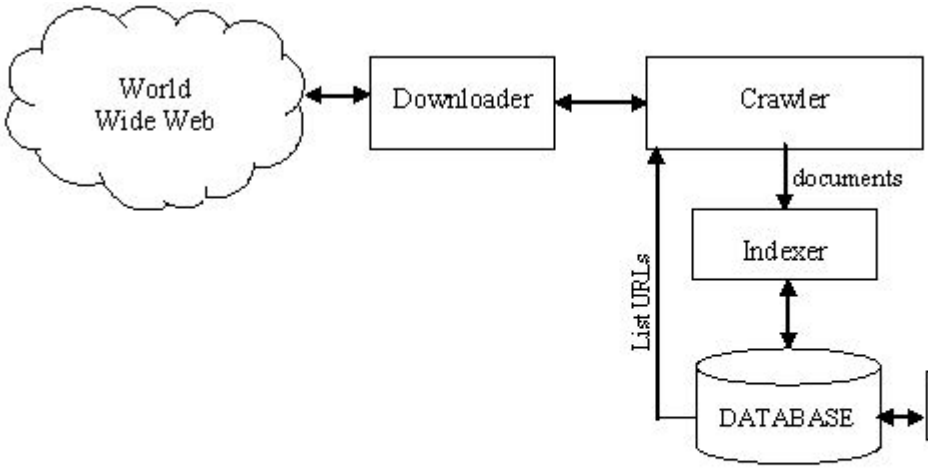


Δένδρα (B, B+)



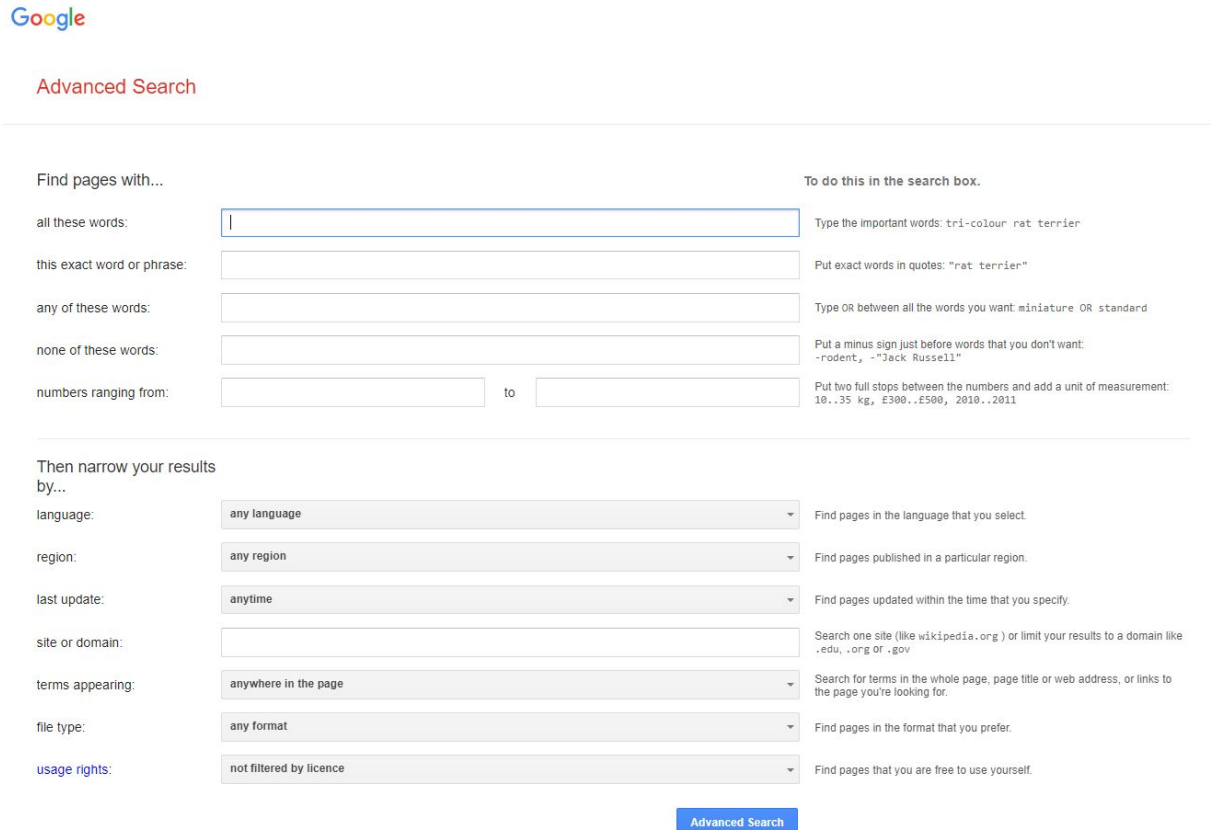
Και τι γίνεται όταν έρθουν νέα δεδομένα;

- Ο κύκλος αρχίζει από την αρχή
- Τα νέα δεδομένα προστίθενται στα παλιά



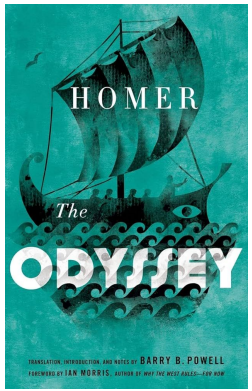
3. Κατανόηση της πληροφοριακής ανάγκης

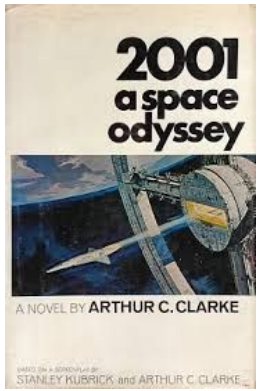
- Το ερώτημα του χρήστη είναι συχνά μια λέξη ή μια ομάδα λέξεων
- Μπορεί όμως να είναι και μια ολόκληρη φράση, πρόταση, κείμενο ή εικόνα
- Μπορεί και να διατυπώνεται με ένα πιο σύνθετο τρόπο

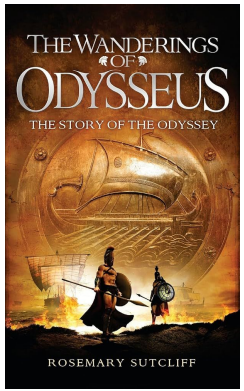


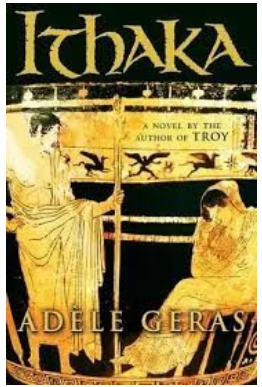
4. Αναζήτηση

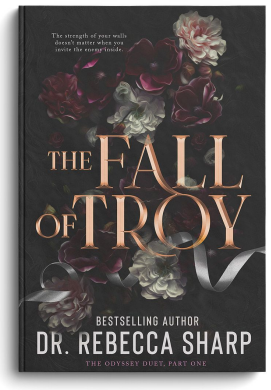
- Τα κείμενα που έχουν όλες τις λέξεις του χρήστη είναι σχετικά
- Τα κείμενα που έχουν έστω και μια από τις λέξεις του χρήστη είναι σχετικά
- Τα κείμενα που έχουν συνώνυμες λέξεις με τις λέξεις του χρήστη είναι σχετικά
- Κείμενα που σχετίζονται με άλλο τρόπο με τα σχετικά κείμενα (π.χ. δείχνουν ή δείχνονται με κάποιο σύνδεσμο ή αναφορά, έχουν κοινά χαρακτηριστικά) είναι σχετικά











5. Παρουσίαση (βαθμονόμηση) αποτελεσμάτων

- Τα αποτελέσματα που έχουν όλες τις λέξεις αναζήτησης είναι πιο σχετικά
- Τα αποτελέσματα που έχουν πολλές φορές όλες τις λέξεις αναζήτησης είναι πιο σχετικά
- Τα αποτελέσματα που έχουν μόνο τις λέξεις αναζήτησης είναι πιο σχετικά
- Τα αποτελέσματα που τα έχουν επιλέξει πολλοί χρήστες είναι πιο σημαντικά και πρέπει να προηγηθούν
- κλπ κλπ

Επεξεργασία Φυσικής Γλώσσας

Επεξεργασία Φυσικής Γλώσσας

- Natural Language Processing
- Βασικοί στόχοι:
  - Η κατανόηση της φυσικής γλώσσας - Natural Language Understanding
  - Η παραγωγή φυσικού λόγου - Natural Language Generation
- Εφαρμογές:
  - Απάντηση ερωτήσεων - question answering
  - Ανίχνευση γνώμης - opinion mining
  - Ανάλυση συναισθήματος - sentiment analysis
  - Εντοπισμός ψευδών ειδήσεων
  - Εντοπισμός γλώσσας μίσους
  - Εξαγωγή ή και απόκρυψη ευαίσθητων πληροφοριών από κείμενα
  - Διαλογικά συστήματα
  - Αυτόματη μετάφραση
  - Συμπλήρωση γραπτού λόγου
  - Αυτόματη δημιουργία περιλήψεων
  - Αυτόματη δημιουργία περιγραφών
  - Αυτόματος υποτιτλισμός

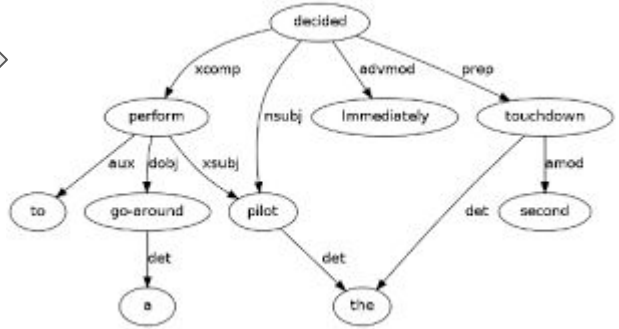
Τι αλλάζει στην περίπτωση αυτή;

Η βασική αρχή

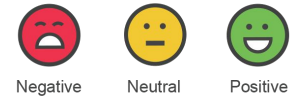
Er hörte leise Schritte hinter sich. Das bedeutete nichts Gutes. Wer würde ihm schon folgen, spät in der Nacht und dazu noch in dieser engen Gasse mitten im übel beleumundeten Hafenviertel? Gerade jetzt, wo er das Ding seines Lebens gedreht hatte und mit der Beute verschwinden wollte! Hatte einer seiner zahllosen Kollegen dieselbe Idee gehabt, ihn beobachtet und abgewartet, um ihn nun um die Früchte seiner Arbeit zu erleichtern? Oder gehörten die Schritte hinter ihm zu einem der unzähligen Gesetzeshüter dieser Stadt, und die stählerne Acht um seine Handgelenke würde gleich zuschnappen? Er konnte die Aufforderung stehen zu bleiben schon hören. Gehetzt sah er sich um. Plötzlich erblickte er den schmalen Durchgang. Blitzartig drehte er sich nach rechts und verschwand zwischen den beiden Gebäuden.

κείμενο

0.1	–0.3	0.8
0.4	0.2	0.6
0.7	–0.1	0.3
0.2	0.4	0.5
0.5	0.5	0.6

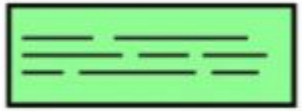


αναπαράσταση



ετικέτα

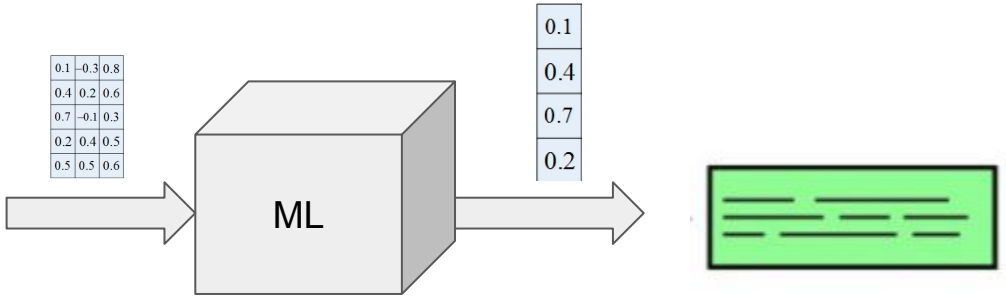
He had quiet footsteps behind him. That didn't mean anything good. Who would follow him, late at night and in this narrow alley in the middle of the Libel-famous harbor district? Just now, when he had done the thing of his life and wanted to disappear with the loot, had one of his countless colleagues had the same idea, watched him and waited, in order to relieve him of the burden of his work? Or did the steps behind him belong to one of the countless law enforcement officers in this city, and the steel figure eight around his wrists was about to snap shut? He could already hear the request to stop. He looked around in a rush. Suddenly he saw the narrow passage. In a flash he turned to the right and disappeared between the two buildings.



κείμενο

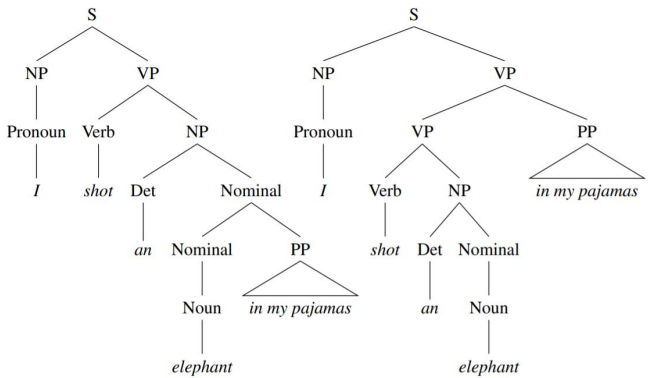
Δύο προσεγγίσεις

Er hörte leise Schritte hinter sich. Das bedeutete nichts Gutes. Wer würde ihm schon folgen, spät in der Nacht und dazu noch in dieser engen Gasse mitten im übel beleumundeten Hafenviertel? Gerade jetzt, wo er das Ding seines Lebens gedreht hatte und mit der Beute verschwinden wollte! Hatte einer seiner zahllosen Kollegen dieselbe Idee gehabt, ihn beobachtet und abgewartet, um ihn nun um die Früchte seiner Arbeit zu erleichtern? Oder gehörten die Schritte hinter ihm zu einem der unzähligen Gesetzeshüter dieser Stadt, und die stählerne Acht um seine Handgelenke würde gleich zuschnappen? Er konnte die Aufforderung stehen zu bleiben schon hören. Gehetzt sah er sich um. Plötzlich erblickte er den schmalen Durchgang. Blitzartig drehte er sich nach rechts und verschwand zwischen den beiden Gebäuden.



end-to-end-training

Er hörte leise Schritte hinter sich. Das bedeutete nichts Gutes. Wer würde ihm schon folgen, spät in der Nacht und dazu noch in dieser engen Gasse mitten im übel beleumundeten Hafenviertel? Gerade jetzt, wo er das Ding seines Lebens gedreht hatte und mit der Beute verschwinden wollte! Hatte einer seiner zahllosen Kollegen dieselbe Idee gehabt, ihn beobachtet und abgewartet, um ihn nun um die Früchte seiner Arbeit zu erleichtern? Oder gehörten die Schritte hinter ihm zu einem der unzähligen Gesetzeshüter dieser Stadt, und die stählerne Acht um seine Handgelenke würde gleich zuschnappen? Er konnte die Aufforderung stehen zu bleiben schon hören. Gehetzt sah er sich um. Plötzlich erblickte er den schmalen Durchgang. Blitzartig drehte er sich nach rechts und verschwand zwischen den beiden Gebäuden.



Syntactic analysis

Οι εργασίες και δυσκολίες της συντακτικής ανάλυσης

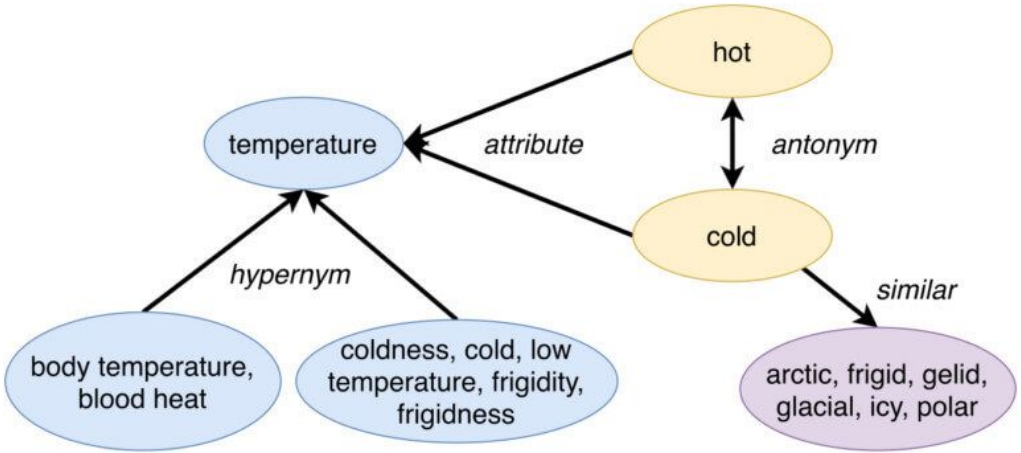
- Part of Speech (POS) tagging: χαρακτηρισμός των λέξεων της πρότασης ως noun, verb, adjective, adverb, κλπ.
- Dependency Parsing: Παραγωγή ενός parse/syntactic tree με τις λέξεις της πρότασης και τις μεταξύ τους σχέσεις και εξαρτήσεις.
- Phrase Structure Analysis: αναγνώριση και κατηγοριοποίηση φράσεων μέσα σε μια πρόταση, όπως τα noun phrases (NP), τα verb phrases (VP) κλπ.
- Ο ρόλος μιας λέξης αλλάζει ανάλογα με τις περιβάλλουσες λέξεις στην πρόταση:

π.χ. *The trophy doesn't fit into the brown suitcase because **it** is too [small/large].*

- Word Sense Disambiguation: η ίδια λέξη μπορεί να έχει πολλαπλές σημασίες

Σημασιολογικές Οντολογίες

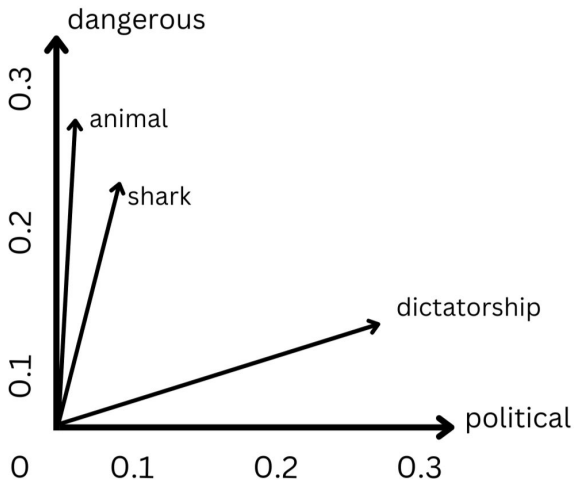
WORDNET: https://wordnet.princeton.edu/



Δεν είναι διαθέσιμα σε όλες τις γλώσσες.
Το ίδιο και πολλά εργαλεία συντακτικής ανάλυσης.

Οι λέξεις ως κατανομές

- Βασιζόμαστε στη συνεμφάνιση των λέξεων σε πολλά κείμενα
- Συνεμφάνιση = Σχέση
- Αναπαριστούν τις λέξεις με διανύσματα αριθμών (σημεία) σε έναν πολυδιάστατο χώρο.



NLP και machine learning

- Πολλές από τις εφαρμογές ΕΦΓ τις αντιμετωπίζουμε ως προβλήματα κατηγοριοποίησης
  - Κατηγοριοποίηση γνώμης σε θετική/αρνητική/ουδέτερη,
  - Κατηγοριοποίηση συναισθήματος σε ένα από τα βασικά συναισθήματα
  - Κατηγοριοποίηση μιας είδησης ως ψευδούς, ενός σχολίου ως κακόβουλο
  - Κατηγοριοποίηση πρόθεσης του χρήστη από ένα διαλογικό σύστημα (τι τύπου πληροφορίες ψάχνει ο χρήστης)
- Σε άλλες περιπτώσεις ως προβλήματα πρόβλεψης της επόμενης λέξης σε μια ακολουθία λέξεων
  - Συμπλήρωση κειμένου
  - Μετάφραση
  - Περίληψη κλπ

Πλεονέκτημα σε ειδικές καταστάσεις

- Επεξεργασία κώδικα

```python
def quicksort(arr):
    if len(arr) <= 1:
        return arr
    else:
        pivot = arr[0]
        less_than_pivot = [x for x in arr[1:] if x <= pivot]
        greater_than_pivot = [x for x in arr[1:] if x > pivot]
        return quicksort(less_than_pivot) + [pivot] + quicksort(greater_than_pivot)
```

- Επεξεργασία αλληλουχιών

```
TGYHDNTWPIHIHPMRVPSKRHQRYGSVLNFAHMAGHMCHSVYGTSWLTMKKIWLWQAEF
WAECRHVWAICKYCIRLNNWNTFDYTMVTQEEYGQEQYYYLTEGHYSHMEENERYVNVHW
AFSGGFYQAGRVAYNIFTTGHVVDACYKCDCQRVMANDWMFSFRGFSWNPFHKEEHSDPQ
RPFADPMWPYPRQCIEVYDWRHITARWNMAPNSKQTPGMHWIYYLSDPIDHTTSHWVYEV
CGHEWQIHHCAQTENSACRHEPNTMTMNDCFMIFSCDQILQVIHTQDLPRGMQDINWMTE
FIPTEMIIWKLDSNLGCMAYRLQDVNHDDYMRFLLWVHDIYWARSCFSNRYSHSKFLSEV
YFLAFTAQMKQDALYFYQAGHMFKHQRASWDDGLYHADLRLMINSKFCWNPPEHLPSAEP
LNDRKCYLFSNFIVQELMVDCVGMWHMYVNMDVFCGWYGGWICKVCGYKKGFDAASKCFV
KRLLSTWIVCDFISFDVGIACRFRAKESIDQGYCKGAEPGPIFNHLWRGDQNFNWRPPVT
NNNNITMQSPSERPYNFQTGVRVRVWKYHWYNYPNTRYANEGAEVEYAQNMDADDHGSNF
NLMQHFSRTNYILHKVHVYKMMRYYYRGEVCNYKPCNTQPQESSLFRCSMWGQRYRWITQ
IPSQSQMMPHLCCKNASKGKHMWCRYDWGIHFRTICWYPLANGNMLEPSTAKGCPLHVMC
YKFWCTNDIWPTWYWSSCEPTCQKKKYGHRGCCGVSYLAAAKLNTSQSAPMYLFQMYENA
SPKTCEDGKSNAQECQAYMDQRIIPFESDAAQFIFWKLHTYQAMKWNKKHNCTGGQGKYE
NNVQGMADILMWNNRYNKANNIYMLCVELWIQTYYRPCAFMVKMTFATNAIGIWVMWWFH
CKYNAWNASKIVGHMGNNRDYKENLTMYQYVKFNTFGWSTTKFTLPTIMAPDGADYKCNQ
TLGYCNTMTHCRYDLNNCAAAVWTAPWKSGCQVEGFRHCQ
```