ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΜΙΑ ΓΕΝΙΚΗ (ΓΛΩΣΣΟΛΟΓΙΚΗ) ΠΡΟΣΕΓΓΙΣΗ

DR. ΑΘΑΝΑΣΙΟΣ Ν. ΚΑΡΑΣΙΜΟΣ

AKARASIMOS@GMAIL.COM || AKARASIMOS@ACADEMYOFATHENS.GR

ΕΙΣΑΓΩΓΗ

- Το πέρασμα στις ψηφιακές σπουδές των Ανθρωπιστικών Επιστημών
- Επεξεργασία Φυσικής Γλώσσας
- Ανάλυση Κειμένων: μια γενική προσέγγιση

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

A (SHORT) INTRODUCTION TO NATURAL LANGUAGE PROCESSING

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ (ΕΦΓ)

- **Υπολογιστική Γλωσσολογία**
  - Η χρήση υπολογιστικών τεχνικών για τη μελέτη και ανάλυση γλωσσολογικών φαινομένων
- **Γνωσιακή Επιστήμη**
  - Μελέτη της ανθρώπινης επεξεργασίας της πληροφορίας (αντίληψη, γλώσσα, λογική, κτλ.)
- **Επιστήμη Ηλεκτρικών Υπολογιστών**
  - Θεωρητική προσέγγιση των Η/Υ και πρακτικές τεχνικές για την υλοποίηση
- **Επιστήμη Πληροφορικής**
  - Ανάλυση, κατηγοριοποίηση, χειραγώγηση, εξαγωγή και διάδοση πληροφορίας

ΕΦΓ: ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΑΙ ΥΛΟΠΟΙΗΣΕΙΣ

- Η γλώσσα περιέχει *αμφισημία*.
  - Για σωστή δομή κειμένου, πρέπει να αφαιρέσουμε την *αμφισημία*.
- Λεξική ανάλυση και *τεμαχισμός* (tokenization)
  - Βρήκα λάθη παντού μεσ' το κείμενο.
  - >> Βρήκα/ λάθη/ παντού/ μεσ'το/ κείμενο (εναλλακτικά με/ σ'το | μεσ'/το)
- Απομάκρυνση των λειτουργικών λέξεων
  - Μια πεπερασμένη, αλλά μη-ορισμένη λίστα
    - Ο καλός, ο κακός και ο άσχημος.
    - To be or not to be.

ΕΦΓ ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΑΙ ΥΛΟΠΟΙΗΣΕΙΣ

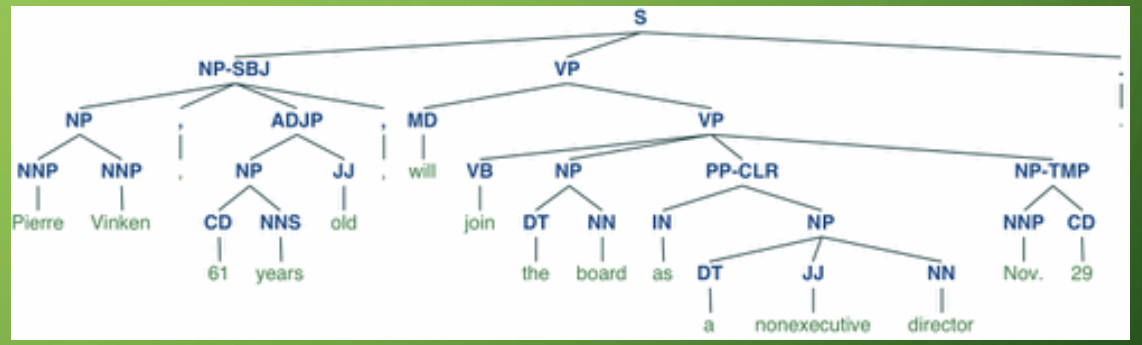
- Θεματοποίηση (stemming)
  - άνθρωπος, ανθρώπου, άνθρωπο, ανθρώπων, ανθρώπους > ανθρωπ-
  - χορός, χορεύω, χορεύουμε, χορευτής, χορευτικός > χορ-
- Λημματοποίηση (lemmatization)
  - Θεματοποίηση + μορφολογικές διαδικασίες + περικείμενο
    - λέγαμε > λέω +1P + Pl
    - είπα > λέω +past +1P + Sing
    - κυματάκι = κυ-ματ-άκι (?)
    - πλένω = πλ-έν-ω (?)
- Μορφολογία (ρίζες, προθήματα, επιθήματα, κτλ.)
  - σιδηροδρομικός > σιδηρ + ο + δρομ + ικ + ος

ΕΦΓ ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΑΙ ΥΛΟΠΟΙΗΣΕΙΣ

- Σύνταξη (POS tagger)
  - έγγραφο > ΟΥΣΙΑΤΙΚΟ
  - αυτής > ΑΝΤΩΝΥΜΙΑ
  - γράφουμε > ΡΗΜΑ
  - Εμείς ήρθαμε χτες > PN VPP ADV
- Αμφισημία
  - Ο αστυνομικός χτύπησε τη γριά με το μπαστούνι. (?)
  - Ο δάσκαλος με το μαντήλι έπιασε το χερούλι της πόρτας. (?)

ΕΦΓ ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΑΙ ΥΛΟΠΟΙΗΣΕΙΣ

- Συντακτική ανάλυση (syntactic parser)
  - Η Μαρία χαιρέτησε τον Κώστα.
  - Η Μαρία τον Κώστα χαιρέτησε.
  - Τον Κώστα χαιρέτησε η Μαρία.



- Οριοθέτηση προτάσεων (sentence boundaries)
  - Τα σημεία στίξης καθορίζουν τα όρια μιας πρότασης. Η τελεία δηλώνει το τέλος μιας πρότασης.
    - Όχι, όμως, πάντα!», είπε ο καθηγητής…

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΜΙΑ ΓΕΝΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

TEXT ANALYSIS VS. TAMING TEXT: AN OVERALL INTRODUCTION

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΓΙΑΤΙ;

- 80% της πληροφορίας και των κειμένων είναι αδόμητη
  - Πλήρεις αλυσίδες πληροφοριών σε κείμενα οργανισμών, φορέων, ιδρυμάτων
  - Προϊόντα και e-αγορές: παρουσιάσεις, διαφημίσεις, προωθήσεις
  - Υλικό από χρήστες: ιστολόγια, fora, wikis
  - Άποψη των πελατών: social media, προσωπική ανάλυση
- Ασύλληπτος όγκος δεδομένων
  - 161.000.000 Gigabytes σε ψηφιακό περιεχόμενο το 2006
  - ~1000 Exabytes σε ψηφιακό περιεχόμενο το 2010
  - Ήχος και εικόνα χρειάζονται περιλήψεις και ετικέτες
  - Πλήθος σωμάτων κειμένων χωρίς επισημείωση και μεταδεδομένα

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Σεισμική δόνηση, μεγέθους 3,7 βαθμών της κλίμακας Ρίχτερ (σύμφωνα με την αυτόματη λύση του EMSC και του Γεωδυναμικού Ινστιτούτου Αθηνών),σημειώθηκε χτες 14/3/2015. Το ακριβές επίκεντρο της δόνησης εντοπίζεται 93χλμ. νοτιοανατολικά του Αγίου Νικολάου Κρήτης και 128χλμ. νοτιοδυτικά της Καρπάθου. Το εστιακό βάθος του σεισμού υπολογίζεται στα 5χλμ. Ο σεισμός, όπως τον κατέγραψε ο σεισμογράφος του Σεισμολογικού Δικτύου του Γεωδυναμικού Ινστιτούτου του Εθνικού Αστεροσκοπείου Αθηνών, που είναι τοποθετημένος στην Ζάκρο του νομού Λασιθίου. Καταγράφηκαν μόνο μερικές καταστροφές σε παλιά σπίτια σε χωριά του νομού Λασιθίου.

- Τύπος καταστροφής: σεισμός
  - τοποθεσία: *Κρήτη*
  - ημερομηνία: *14/3/2015*
  - μέγεθος: *3,7*
  - επίκεντρο: 93χλμ. νοτιοανατολικά του Αγίου Νικολάου Κρήτης και 128χλμ. νοτιοδυτικά της Καρπάθου.
  - Πηγή: EMSC και Γεωδυναμικό Ινστιτούτο Αθηνών
  - ζημιές:
    - ανθρώπινες:
      - victim: -
      - number: -
      - outcome: -
    - υλικές:
      - object: *χωριά του νομού Λασιθίου*
      - outcome: ζημιές

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ

- υγεία
  - διατροφή
  - ασθένειες
- επιστήμες
  - Ανθρωπιστικές επιστήμες
  - Κοινωνικές επιστήμες
- αθλητισμός
  - ποδόσφαιρο
  - μπάσκετ

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ

- Ροή νέων και ειδήσεων
  - Κατηγοριοποίηση των εισερχόμενων ειδήσεων, νέων και ιστοριών
- Ερωτήματα στις μηχανές αναζήτησης
  - Google: αναζήτηση «συγγραφέας των Μεταφυσικών»
- Εντοπισμός των spam emails
  - http://www.paulgraham.com/spam.html
- Κατεύθυνση των emails στα κατάλληλα άτομα και ομάδες
- Δοκιμή:
  - http://cogcomp.cs.illinois.edu/page/run_demo/Dataless
  - http://text-processing.com/demo/sentiment/

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

Συλλογή κειμένων

Σύστημα Εξαγωγής Πληροφορίας

Ποιος: _____
Γιατί: _____
Που:_____
Πότε: _____
Πως: _____

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

- Αναγνώριση (recognition), ετικετοποίηση (tagging), και εξαγωγή (extraction) από μια δομημένη αναπαράσταση σωμάτων κειμένων: ορισμένα βασικά στοιχεία των πληροφοριών, π.χ. πρόσωπα, εταιρείες, τοποθεσίες, οργανώσεις, ημερομηνίες.
- Αυτές οι εξαγωγές μπορούν στη συνέχεια να χρησιμοποιηθούν για ένα ευρύ φάσμα εφαρμογών, συμπεριλαμβανομένων συστημάτων ερωταποκρίσεων, οπτικοποίησης πληροφορίας και την εξόρυξη δεδομένων.
- Δοκιμή:
  - http://services.gate.ac.uk/annie/

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΕΞΑΓΩΓΗ ΟΡΟΛΟΓΙΑΣ

- Διαφοροποίηση ανάμεσα στις χρήσιμες πληροφορίες και τον «θόρυβο».
- Βοηθάει τους λεξικογράφους να εντοπίσουν νέους όρους:
  - Η κυβέρνηση ενέκρινε την διανομή γενόσημων.
  - Τα νέα αυτοκίνητα έχουν αυξημένη προστασία λόγων των ανθρακονημάτων.
  - Η Apple προσπαθεί να κυριαρχήσει στην νέα αγορά των phablets.
- Εξαγωγή ορολογίας σαρώνει επιστημονικά άρθρα για τον εντοπισμό όρων, πιθανώς συγκρίνοντάς τα με μία λίστα δεδομένων όρων.
- Δοκιμή:
  - http://fivefilters.org/term-extraction/

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΕΡΩΤΑΠΟΚΡΙΣΕΙΣ

- Σε αντίθεση με την ανάκτηση πληροφορίας, η οποία παρέχει μια λίστα σχετικών εγγράφων ως απάντηση σε ένα ερώτημα του χρήστη
- Ένα σύστημα ερωταποκρίσεων παρέχει στο χρήστη είτε μόνο το κείμενο της ίδιας της απάντησης ή οι διαφορετικές εναλλακτικές απαντήσεις.
- Δοκιμή:
  - http://start.csail.mit.edu/index.php
  - http://www.wikiqa.de/index.php

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΕΡΩΤΑΠΟΚΡΙΣΕΙΣ

- Ζητάμε να μας απαντήσει, όχι να μας επιστρέψει έγγραφα.
- Αναζήτηση γεγονότων vs. επεξηγηματική αναζήτηση
  - Απαντήσεις *Ναι/Όχι*: «Είναι ο Παπούλιας πρόεδρος της Δημοκρατίας;»
  - Ερωτήσεις τύπου *Ποιος*: «Ποιος κέρδισε το Νόμπελ Ειρήνης το 2004;»
  - Ερωτήσεις λίστας: «Ποιοι κέρδισαν στο τελικό των 100 μ. στους Ολυμπιακούς αγώνες;»
  - Ερωτήσεις οδηγίες: «Πώς να πάω στην Ακρόπολη από το Μαρούσι;»
  - Επεξηγήσεις: «Γιατί διασπάστηκε η ΕΣΣΔ;»
  - Εντολές: «Δώσε μου το ύψος του Ολύμπου»
- ανάλυση ερώτησης – εξαγωγή εγγράφων – παροχή απάντησης

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΔΗΜΙΟΥΡΓΙΑ ΠΕΡΙΛΗΨΗΣ

- Η δημιουργία περίληψης (summarization) μειώνει ένα μεγαλύτερο κείμενο μετατρέποντας σε ένα μικρότερο, αλλά διατηρεί τα κεντρικό νόημα έχοντας πλούσια συγκροτημένη δομική αναπαράσταση του αρχικού εγγράφου με καίριες πληροφορίες.
- Δοκιμή:
  - http://apidemo.pingar.com/Summarize.aspx#wrapper
  - http://textsummarization.net/text-summarizer

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΜΗΧΑΝΙΚΗ ΜΕΤΑΦΡΑΣΗ

- Η Μηχανική (ή Αυτόματη) Μετάφραση (Machine Translation) είναι ίσως η παλαιότερη όλων των εφαρμογών ΕΦΓ, ενώ στη σύγχρονη εποχή διάφορες υλοποιήσεις της ΕΦΓ έχουν χρησιμοποιηθεί σε συστήματα μηχανικής μετάφρασης για τη βελτίωση απόδοσης και ακρίβειας. Τα συστήματα μηχανικής μετάφρασης κυμαίνονται από την λεξοκεντρική προσέγγιση (word-based approach) σε εφαρμογές που περιλαμβάνουν υψηλότερα επίπεδα ανάλυσης (higher-level analysis).
- Δοκιμή:
  - http://www.babelfish.com
  - http://translate.google.com

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΔΙΑΛΟΓΙΚΑ ΣΥΣΤΗΜΑΤΑ

- Ίσως η πανταχού παρούσα εφαρμογή του μέλλοντος, ως το σύστημα που οραματίζονται όλοι οι μεγάλοι πάροχοι εφαρμογών προς τον τελικό χρήση.
- Τα διαλογικά συστήματα (dialogue systems) συνήθως επικεντρώνονται σε μια περιορισμένη εφαρμογή (π.χ. από ψυγεία, έξυπνες τηλεοράσεις, κινητά νέας γενιάς, έως online ψηφιακούς βοηθούς σε εταιρίες ή τηλεφωνικά κέντρα),
- Χρησιμοποιούν σήμερα το φωνητικό ή λεξικό επίπεδο της γλώσσας. Πιστεύεται ότι η χρησιμοποίηση όλων των βαθμίδων επεξεργασίας γλώσσας σε επίπεδα που αναφέρθηκε προηγουμένως θα δώσει τη δυνατότητα για πραγματικά διαλογικά συστήματα.

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΔΙΚΤΥΑ

- Δεδομένα επεξεργάσιμα από Η/Υ για τη σύνδεση και δημιουργία υπερκειμένου.
- Απαραίτητη η παρουσία μεταδεδομένων σε έγγραφα
- Ρητά: τίτλος, συγγραφέας, ημερομηνία δημιουργίας
- Έμμεσα: συναφείς πληροφορίες όπως τα ονόματα οντοτήτων και η σχέση τους

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΡΟΗ ΕΡΓΑΣΙΑΣ

- Ταξινόμηση - Κατηγοριοποίηση
- Τεμαχισμός
- Διαχωρισμός προτάσεων
- Ετικετοποίηση Μερών του Λόγου
- Ετικετοποίηση Ονομάτων και Οντοτήτων
- Εξαγωγή πληροφορίας

ΑΝΑΖΗΤΗΣΗ ΚΕΙΜΕΝΩΝ: ΑΝΑΖΗΤΗΣΗ

- Απλή αναζήτηση
- Σύνθετη αναζήτηση
- Πολύπλοκη αναζήτηση
- Παραμετροποιήσιμη αναζήτηση

ΑΝΑΖΗΤΗΣΗ ΚΕΙΜΕΝΩΝ: ΑΝΑΖΗΤΗΣΗ

- **Κανονιστικές εκφράσεις** (regular expressions)
  - Μια έκφραση που περιγράφει μια σειρά από ακολουθίες (=κανονιστική γλώσσα) ή ένα σύνολο από διατεταγμένα ζεύγη ακολουθιών (= μια κανονιστική σχέση). Κάθε γλώσσα ή σχέση που περιγράφεται από μια κανονιστική έκφραση μπορεί να αναπαρίσταται από ένα αυτόματο πεπερασμένων καταστάσεων. Υπάρχουν πολλοί διαφορετικοί φορμαλισμοί για κανονιστικές εκφράσεις. Οι πιο κοινοί τελεστές είναι η αλυσιδωτή σύνδεση, η ένωση, η τομή, το συμπλήρωμα (= άρνηση), η επανάληψη και η σύνθεση. Επίσης καλείται και ως ρητή έκφραση.

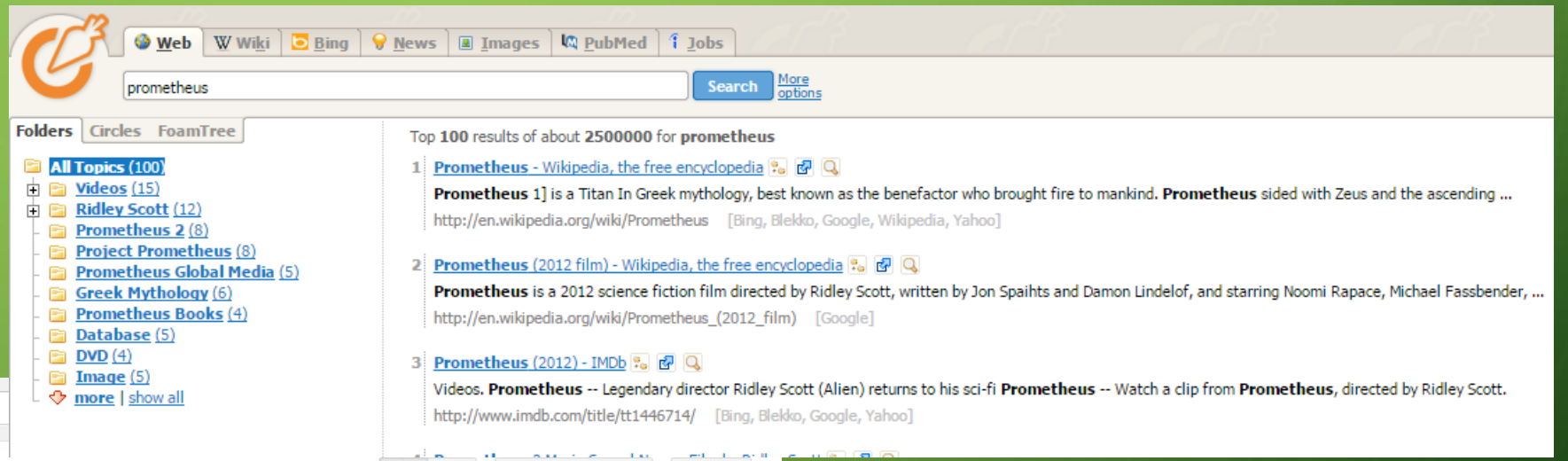
ΑΝΑΖΗΤΗΣΗ ΚΕΙΜΕΝΩΝ: ΑΝΑΖΗΤΗΣΗ

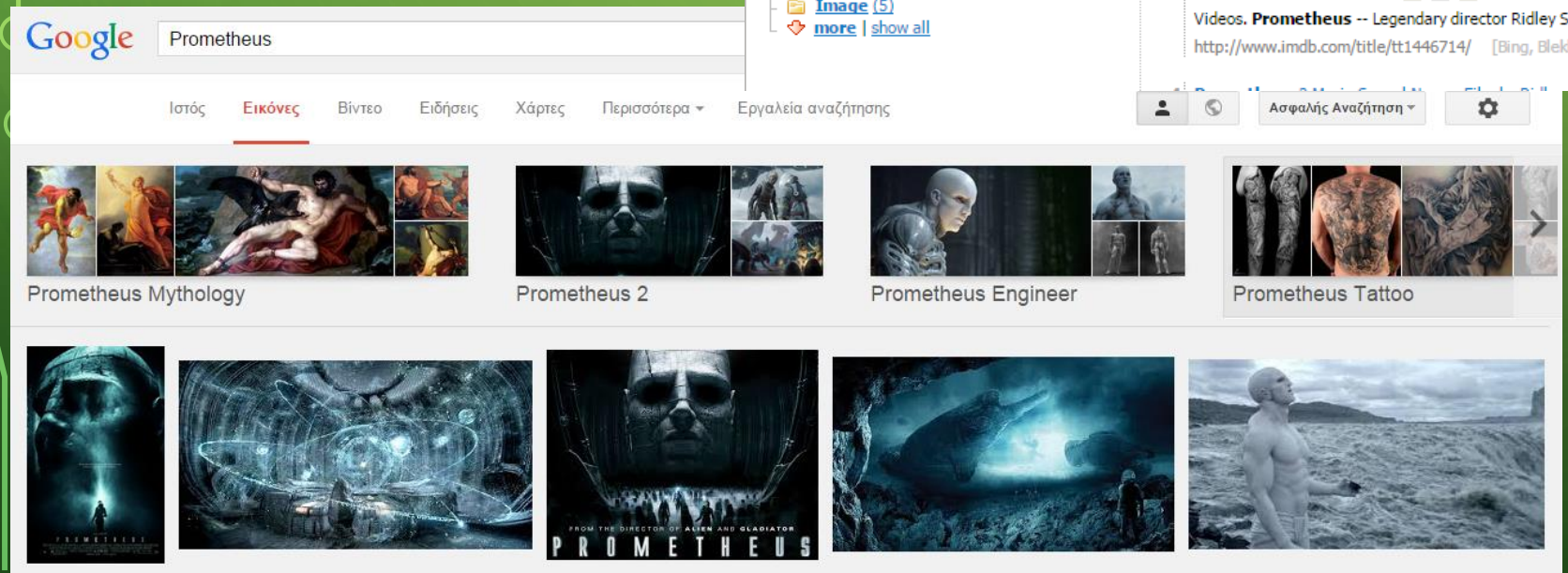
Expression	Syntax	Description
Any character	.	Matches any single character except a line break.
Zero or more	*	Matches zero or more occurrences of the preceding expression, making all possible matches.
One or more	+	Matches at least one occurrence of the preceding expression.
Beginning of line	^	Anchors the match string to the beginning of a line.
End of line	$	Anchors the match string to the end of a line.
Beginning of word	<	Matches only when a word begins at this point in the text.
End of word	>	Matches only when a word ends at this point in the text.
Line break	\n	Matches a platform-independent line break. In a Replace expression, inserts a line break.
Any one character in the set	[]	Matches any one of the characters within the []. To specify a range of characters, list the starting and ending character separated by a dash (-), as in [a-z].
Any one character not in the set	[^...]	Matches any character not in the set of characters following the ^.
Or	|	Matches either the expression before or the one after the OR symbol (|). Mostly used within a group. For example, (sponge|mud) bath matches "sponge bath" and "mud bath."
Escape	\	Matches the character that follows the backslash (\) as a literal. This allows you to find the characters used in regular expression notation, such as { and ^. For example, \^ Searches for the ^ character.
Tagged expression	{}	Matches text tagged with the enclosed expression.
C/C++ Identifier	:i	Matches the expression ([a-zA-Z_$][a-zA-Z0-9_$]*).
Quoted string	:q	Matches the expression (("[^"]*")|('[^']*')).

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΣΥΣΤΑΔΕΣ ΑΝΑΖΗΤΗΣΗΣ

- Ανάλυση συστάδας ή ομαδοποίησης είναι η λειτουργία της ομαδοποίησης μια σειρά από αντικείμενα, θέμάτα, έννοιες με τέτοιο τρόπο που τα μέλη μία ομάδας ταιριάζουν περισσότερο (κατά κάποιο τρόπο ή τον άλλο) μεταξύ τους παρά με τα μέλη άλλων ομάδων.
- Δεν χρησιμοποιούνται κείμενα και δεδομένα με ανθρώπινη επισημείωση.
- Ομαδοποίηση στις μηχανές αναζήτησης (βλ. Carrotsearch), αλλά και σε ηλεκτρονικά καταστήματα.

ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ: ΣΥΣΤΑΔΕΣ ΑΝΑΖΗΤΗΣΗΣ





ΕΥΧΑΡΙΣΤΩ ΓΙΑ ΤΗΝ ΠΡΟΣΟΧΗ ΣΑΣ!