



ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου

kerman@ionio.gr

Τα Θεμέλια στην
Επεξεργασία
Κειμένου

Μερικές εικόνες είναι δανεισμένες από τους Sun & Chen στο University of Cambridge

Σώματα Κειμένων (Text Corpora)

- ❖ Συλλογή κειμένων που έχει δημιουργηθεί με διαδικασίες δειγματοληψίας προκειμένου να χρησιμοποιηθεί για έναν συγκεκριμένο σκοπό, σε μια εργασία Γλωσσικής Τεχνολογίας.
- ❖ "Σώμα κειμένων θεωρείται κάθε συλλογή τμημάτων μιας συγκεκριμένης γλώσσας, τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα της γλώσσας αυτής" (Sinclair, 1996)
- ❖ Wall Street Journal Corpus (αμερικάνικο)
- ❖ Brown Corpus (αμερικάνικο)
- ❖ British National Corpus (αγγλικό)
 - ❖ <https://www.english-corpora.org/>
- ❖ Wikipedia Corpus
- ❖ Edinburgh Twitter Corpus
- ❖ Εθνικός Θησαυρός Ελληνικής Γλώσσας
 - ❖ <http://hnc.ilsp.gr/>
- ❖ Σώμα ελληνικών κειμένων
 - ❖ <http://www.sek.edu.gr/>

Ένα Σώμα Κειμένων

- Είναι αντιπροσωπευτικό μιας γλώσσας
- Φανερώνει τι είναι σύνηθες σε μια γλώσσα
- Μπορεί να δώσει ακριβείς στατιστικές μετρήσεις των φαινομένων της γλώσσας
- Αποθηκεύεται και ανακαλείται οποιαδήποτε στιγμή η πληροφορία σε αυτό
- Παρέχει φυσικά/πραγματικά παραδείγματα της γλώσσας
- Αποτελεί αντικειμενικό δείγμα της χρήσης της γλώσσας
- Είναι διαθέσιμο σε όλους
- Μπορεί να επικαιροποιείται συνεχώς και αν αντανακλά τις πρόσφατες αλλαγές στην γλώσσα
- Δεν μπορεί να παρέχει αρνητική ένδειξη για μη δυνατά φαινόμενα
- Δεν μπορεί από μόνο του να εξηγήσει αυτά που παρατηρούνται σε αυτό
- Αποτελεί περιορισμό για την όποια έρευνα στηρίζεται σε αυτό
- Είναι πολωμένο/στρεβλωμένο (skewed) – N. Chomsky

Είδη Σωμάτων Κειμένων

- Γραπτού λόγου/προφορικού λόγου
- Σύγχρονα/διαχρονικά
- Γενικού σκοπού (ισορροπημένα)/εξειδικευμένα
 - Γενικού σκοπού: για χρήση σε εφαρμογές ανεξάρτητες θεματικής περιοχής, π.χ. WSJ corpus
 - Εξειδικευμένα: για εφαρμογές σε συγκεκριμένες θεματικές περιοχές, π.χ. chat text για ανάλυση τέτοιου είδους επικοινωνίας
- Μονόγλωσσα/Παράλληλα
- Επίπεδο επισημείωσης
 - Γυμνό σώμα κειμένων (raw corpus)
 - Λημματοποιημένο
 - Επισημειωμένες λέξεις ως προς το μέρος του λόγου
 - Συντακτικά αναλυμένο
 - Σημασιολογικά αναλυμένο

Επισημείωση - Annotation

- Το κείμενο φέρει φωνολογικές, γραμματικές, συντακτικές ή σημασιολογικές πληροφορίες.
- Μέθοδοι επισημείωσης: Χειρωνακτική, Αυτόματη, Ημι-αυτόματη

```
Perdita&NN1-NPO; ,&PUN; covering&VVG; the&ATO; bottom&NN1; of&PRF;  
the&ATO; lorries&NN2; with&PRP; straw&NN1; to&TOO; protect&VVI;  
the&ATO; ponies&NN2; '&POS; feet&NN2; ,&PUN; suddenly&AVO; heard&VVD-  
VVN; Alejandro&NN1-NPO; shouting&VVG; that&CJT; she&PNP; better&AVO;  
dig&VVB; out&AVP; a&ATO; pair&NNO; of&PRF; clean&AJO; breeches&NN2;  
and&CJC; polish&VVB; her&DPS; boots&NN2; ,&PUN; as&CJS; she&PNP;  
'd&VMO; be&VBI; playing&VVG; in&PRP; the&ATO; match&NN1; that&DTO;  
afternoon&NN1; .&PUN;
```

Επίπεδα Επισημείωσης

Η επισημείωση μπορεί να πραγματοποιηθεί σε διάφορα επίπεδα γλωσσολογικής γνώσης

- Φωνολογικό επίπεδο
 - Όρια φωνητικών συλλαβών
 - Στοιχεία προσωδίας
- Μορφολογικό επίπεδο
 - Προθέματα
 - Επιθέματα
 - Λήμματα/Θέματα
 - Επισημείωση μορφολογικής πληροφορίας
 - Μέρη του Λόγου
- Συντακτικό επίπεδο
 - Treebanks
 - Ομαδοποίηση λέξεων σε φράσεις
- Σημασιολογικό επίπεδο
 - Έννοιες λέξεων
- Πραγματολογικό επίπεδο
 - Αντωνυμικές αναφορές
 - Επισημείωση υφολογικών στοιχείων

Μετρήσεις σε ένα σώμα κειμένων

...they lay back on the San Francisco grass and looked at the stars and their...

Type: μια λέξη του λεξικού μου (13)

Token: μια εμφάνιση της λέξης αυτής στο κείμενό μου (15)

Ωμή συχνότητα (Raw Frequency) $\text{count}(\text{the})=2$

Κανονικοποιημένη συχνότητα ανα 1 εκατ λέξεις (Normalized Frequency)

$$\text{NF}(\text{the}) = 2 * 1,000,000 / 15$$

Πιθανότητα λέξης $P(\text{the}) = 2/15$

Πιθανότητα διγράμμου $P(\text{stars} \mid \text{the}) = \text{count}(\text{the stars}) / \text{count}(\text{the}) = 1/2$

N-γραμμα

...they lay back on the San Francisco grass and looked at the stars and their...

- ❖ Δίγραμμα = ακολουθίες δυο λέξεων/χαρακτήρων
they lay, lay back, back on, on the, the San, San Francisco,
- ❖ Τρίγραμμα = ακολουθίες τριων λέξεων/χαρακτήρων
they lay back, lay back on, back on the, on the San, the San Francisco, San Francisco grass
- ❖ N-γραμμα = ακολουθίες N λέξεων/χαρακτήρων

Πιθανότητα διγράμμου $P(\text{stars} \mid \text{the}) = \text{count}(\text{the stars}) / \text{count}(\text{the}) = \frac{1}{2}$

Το Μοντέλο Bag of Words (BoW)

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

15



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...