IEEE SB UPATRAS

Artificial Intelligence Scientific Group

Εισαγωγή στο Natural Language Processing

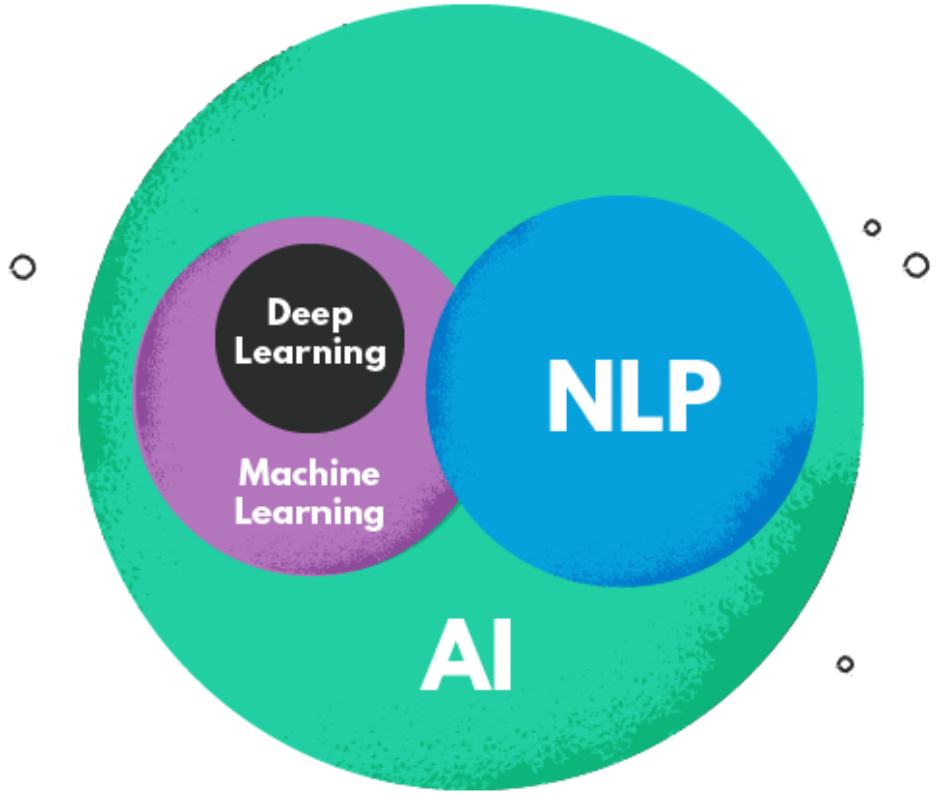
IEEE SB UPatras

Artificial Intelligence Group

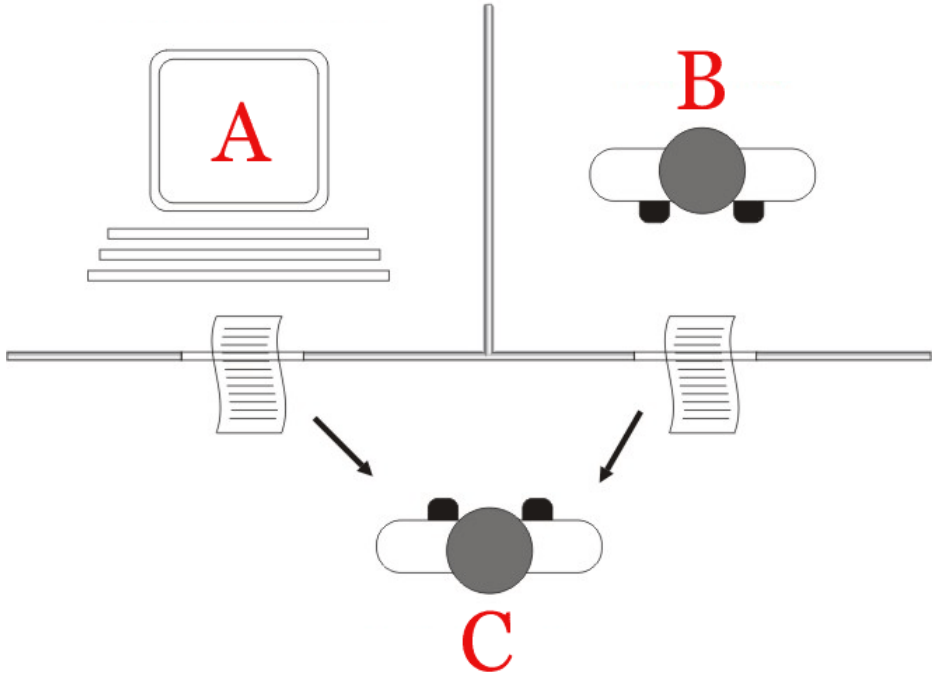
Μητακίδης Ανέστης

NLP: Μία έννοια πολλά ονόματα

- **Υπολογιστική Γλωσσολογία**: ασχολείται με τη στατιστική ή τη βασισμένη σε κανόνες μοντελοποίηση της φυσικής γλώσσας από υπολογιστική σκοπιά
- **Επεξεργασία Φυσικής Γλώσσας**: ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρωπίνων γλωσσών
- **Γλωσσική Τεχνολογία**: ασχολείται με την ανάπτυξη συστημάτων επεξεργασίας φυσικής γλώσσας

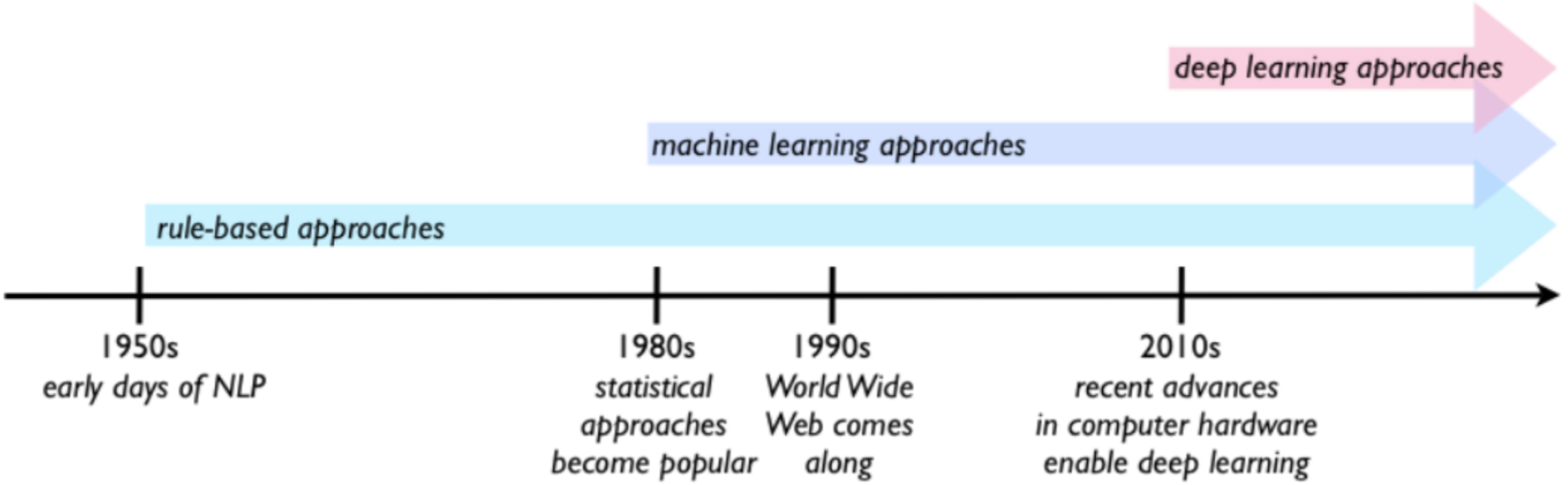


The Turing Test



Υπάρχει «τέλεια» Τεχνητή Νοημοσύνη που χρησιμοποιεί επεξεργασία φυσικής γλώσσας?

Μία γρήγορη ιστορική αναδρομή



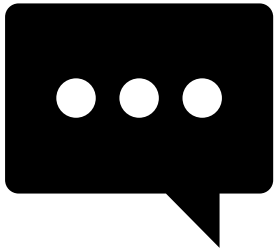
Virtual Assistants

This is so sad, Alexa play Despacito…

καλό 2018…

Εφαρμογές και Projects…

- **Language Identifier**
- **Language Translation**
- **Text Generation**
- **Autocomplete**
- **Movie Recommendation**
- **Sentiment Analysis**
- **Text Summarization**
- **Fake News Detection**
- **Entity Recognition**
- **Question Answering**



Ποιες κοινές διαδικασίες μοιράζεται η εκπόνηση των παραπάνω projects?

Προ-επεξεργασία κειμενικών δεδομένων

- Regular Expressions
- Tokenization, Lower Casing and Stopwords
- Stemming
- Lemmatization
- Vectorization

Regular Expressions

- **Τι είναι;**

Κανόνες που ορίζουν μία γλώσσα (πρότυπο) L

- **Γιατί τις χρησιμοποιούμε;**

Εύκολη και γρήγορη εύρεση κομματιών ενός κειμένου που μας ενδιαφέρουν

- **Που αλλού χρησιμοποιούνται;**

Μπορούν να χρησιμοποιηθούν στο web scraping.

https://cheatography.com/davechild/cheat-sheets/regular-expressions/
https://docs.python.org/3/library/re.html

ΣΥΜΒΟΛΙΣΜΟΣ	ΕΠΕΞΗΓΗΣΗ	ΠΑΡΑΔΕΙΓΜΑ	ΠΡΟΤΥΠΟ
/…/	**Ακολουθία**: συμβολοσειρά μεταξύ των “/”.	/abcdef/	{abc}
[…]	**Επιλογή**: Επιλογή ενός όρου από όσους είναι μέσα στις αγκύλες	/a[bcd]e/	{abe, ace, ade}
(…)	**Ομαδοποίηση**: Οι όροι μέσα στις παρενθέσεις θεωρούνται ένας χαρακτήρας	/a[b(cd)e]/	{ab, acd, ae}
?	**Προαιρετική εμφάνιση**: Ο χαρακτήρας πριν το ? Εμφανίζεται 0 ή 1 φορά	/ab?c/	{ac, abc}
*	**Kleene star**: Ο χαρακτήρας πριν το * Εμφανίζεται 0 ή περισσότερες φορές	/ab*c/	{ac, abc, abbc, abbbc, ..., abb...bc, ...}
+	**Kleene cross**: Ο χαρακτήρας πριν το + Εμφανίζεται 1 ή περισσότερες φορές	/ab+c/	{abc, abbc, abbbc, ..., abb...bc, ...}
.	**Τελεία** : Επιλογή ενός συμβόλου από όλο το αλφάβητο	/a.b/	{aab, abb, acb, adb, aeb}
-	**Διάστημα**: ακολουθία συνεχόμενων συμβόλων	/[1-4][f-i]/	{1f, 1g, 1h, 1i, 2f, 2g, 2h, 2i, 3f, 3g, 3h, 3i, 4f, 4g, 4h, 4i,}
\	**Ειδικός Χαρακτήρας**: Ο χαρακτήρας που ακολουθεί ανήκει στο πρότυπο	/ab\?c/	{ab?c}

https://cheatography.com/davechild/cheat-sheets/regular-expressions/
https://docs.python.org/3/library/re.html

Tokenization, Lower Casing and Stopwords

- **Τι είναι το Tokenization;**

Μία απλή διαδικασίας μετατροπής ενός string σε λίστα λέξεων

- **Γιατί το χρησιμοποιούμε;**

Διευκόλυνση στην κωδικοποίηση των δεδομένων και καλύτερη κατανόηση από το μοντέλο μας

Stemming

- **Τι είναι;**

Διαδικασία απλοποίησης των λέξεων χωρίς να χαθεί η σημαντική πληροφορία

- **Γιατί το χρησιμοποιούμε;**

Αφαίρεση περιττής πληροφορίας από τα δεδομένα για καλύτερη κατανόηση από το μοντέλο

- **Ποιοι είναι οι πιο γνωστοί αλγόριθμοι stemming;**
  - Porter’s Stemmer
  - Lovins Stemmer
  - Paice Stemmer

Lemmatization

- **Τι είναι;**

Διαδικασία σωστής μείωσης στην μορφολογία των λέξεων

- **Γιατί το χρησιμοποιούμε;**

Αφαίρεση περιττής πληροφορίας από τα δεδομένα για καλύτερη κατανόηση από το μοντέλο

Vectorization

- **Τι είναι;**

Διαδικασία μετατροπής των κειμενικών δεδομένων σε δομές αριθμών

- **Γιατί το χρησιμοποιούμε;**

Γιατί τα μοντέλα μας καταλαβαίνουν αριθμούς και όχι λέξεις

Tools - Libraries

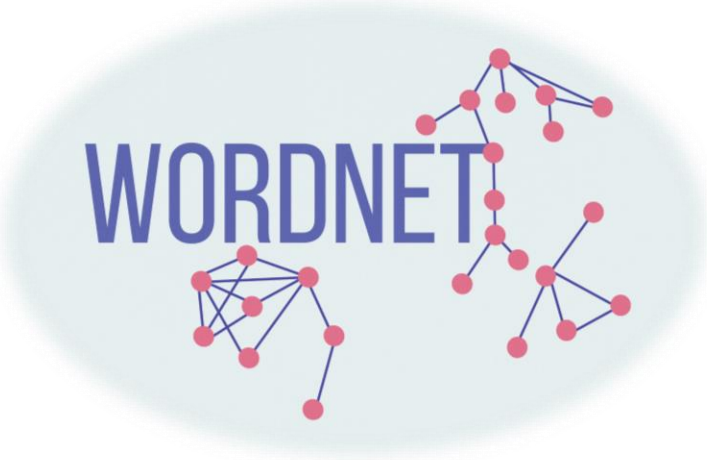
https://textblob.readthedocs.io/en/dev/

https://pypi.org/project/gensim/

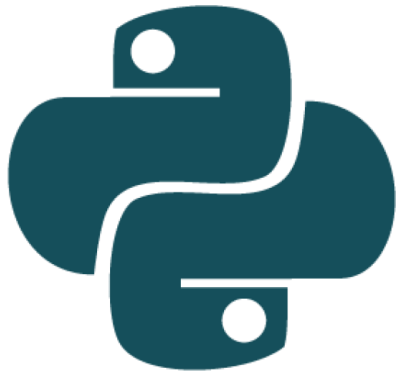
https://wordnet.princeton.edu/

https://www.nltk.org/

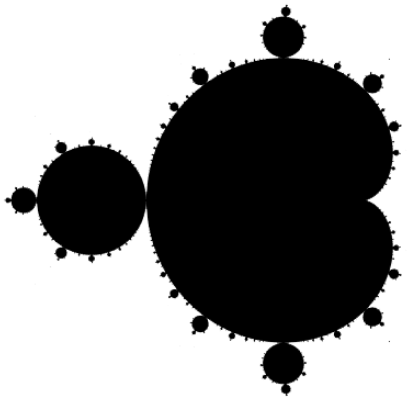
https://spacy.io/



spaCy



NLTK



TextBlob

GENSIM

topic modelling for humans

Time to play