MAI4CAREU

Επεξεργασία Φυσικής Γλώσσας

Εισαγωγική Παρουσίαση

Δημήτρης Πασχαλίδης

Τμήμα Πληροφορικής

Πανεπιστήμιο Κύπρου

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Απο τις γλώσσες στην πληροφορία

- Αυτόματη εξαγωγή εννοιών και δομής απο:
  - Κείμενο και λόγο (ειδήσεις, αρθρα κτλ.)
  - Κοινωνικά δίκτυα
  - Ακολουθίες Genome
- Αλληλεπίδραση με ανθρώπους μέσω της γλώσσας
  - Συστήματα διαλόγων/Chatbots
    Απάντηση ερωτήσεων
  - Συστήματα Συστάσεων

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Βιομηχανικές και εμπορικές εφαρμογές

Co-financed by the European Union
Connecting Europe Facility

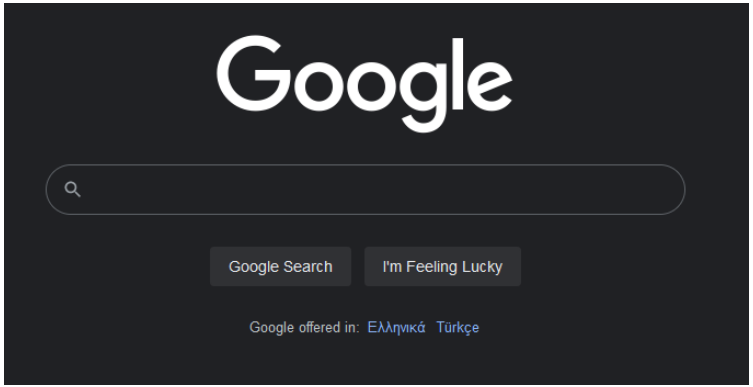
University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Εξαγωγή πληροφοριών από τη γλώσσα

❑Information retrieval

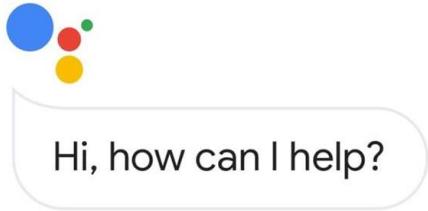
- Χρονολογια 2020: **6.9 δισεκατομμυρια** Google searches καθημερινά.
- Information retrieval κειμένου→ το πιο χρησιμοποιησιμο software στον κόσμο

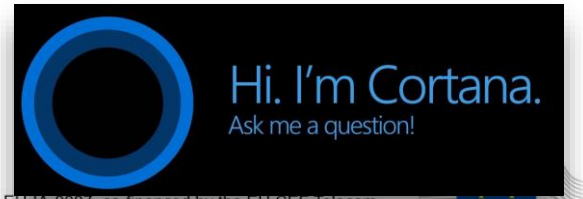


Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Conversational Agents

- Αναγνώριση ομιλίας
  - Γλωσσική Ανάλυση
  - Επεξεργασία Διαλόγου
  - Ανάκτηση πληροφοριών
  - Μετατροπή κειμένου σε ομιλία





Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Text classification: Αντιμετώπιση καταστροφών

- Σεισμός στην Αϊτή 2010
- Classifying SMS Μηνύματα

Haitian Creole: *"Mwen thomassin 32 nan pyron mwen ta renmen jwen yon ti dlo gras a dieu bo lakay mwen anfom se sel dlo nou bezwen"*

English: *"I am in Thomassin number 32, in the area named Pyron. I would like to have some water. Thank God we are fine, but we desperately need water."*

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

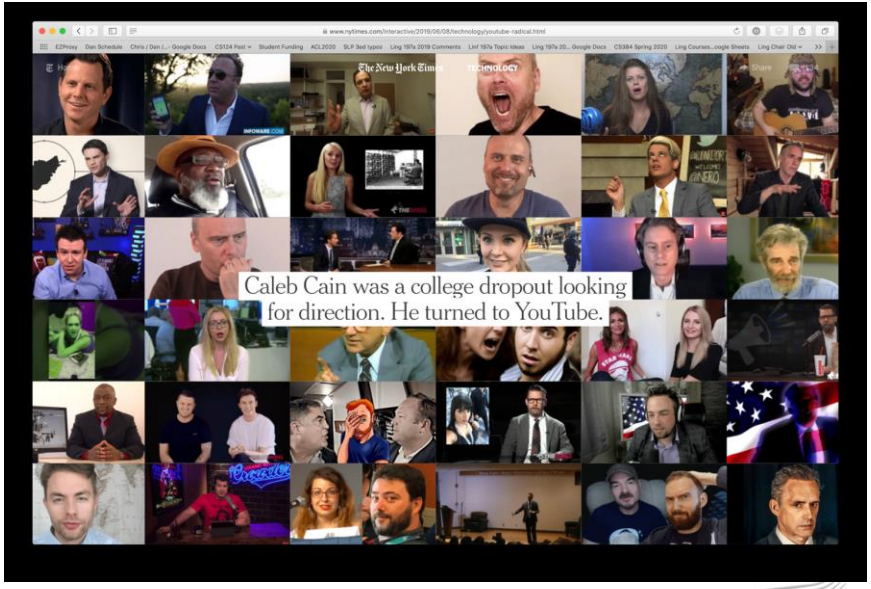
Μηχανές συστάσεων

Τα καλά:

- Προϊόντα: Amazon, ebay
- Περιεχόμενο: Netflix, Spotify

Τα κακά:

- Ριζοσπαστικοποίηση Youtube



Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επίπεδα Γλωσσικών Γνώσεων

Speech	Orthography	Morphology	Lexemes	Syntax	Semantics	Pragmatics	Discourse
Text	Phonetics						
	Phonology						

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Φωνητική και Φωνολογία

- Μοντελοποίηση προφοράς

Ήχοι: Th i a si e n

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Words

- Μοντελοποίηση Γλώσσας
- Tokenization
- Διόρθωση ορθογραφίας

Λέξεις: This is a simple sentence

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Morphology

- Ανάλυση Μορφολογίας
  Tokenization
- Lemmatization

Λέξεις: This **is** a simple sentence

Μορφολογία: *be*
present

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Part-of-Speech

- Σήμανση Part-of-Speech (PoS)

PoS: DT VBZ DT JJ NN

Λέξεις: This **is** a simple sentence

Μορφολογία: *be*
present

Co-financed by the European Union
Connecting Europe Facility

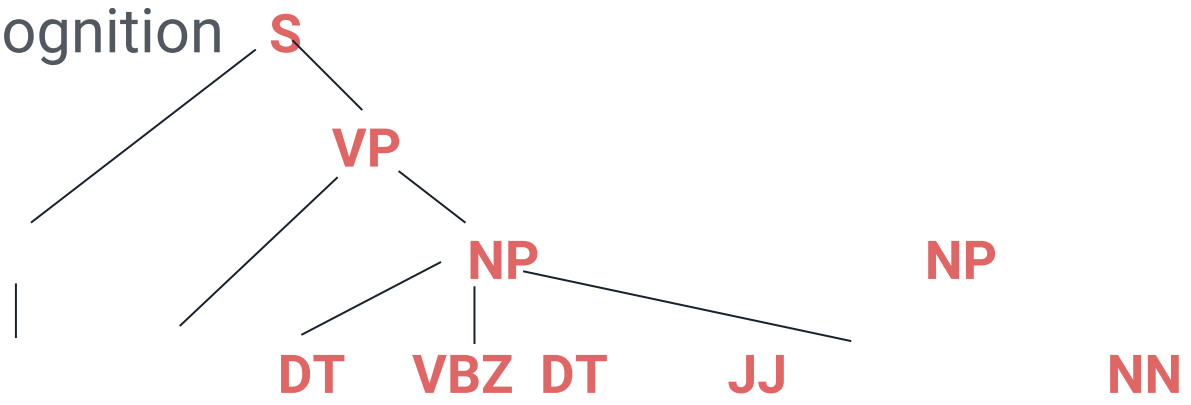
University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Semantics

- Named Entity Recognition

Syntax:



PoS:

Words: This **is** a **simple sente**

n c e

Morphology: *be*

present

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Η γλωσσική διερμηνεία είναι δύσκολη

Ασάφεια

- Πρόταση: “*I made her duck*”
- Τουλάχιστον 6 διαφορετικές έννοιες:
  - *I cooked waterfowl for her (to eat)*
  - *I cooked waterfowl of her*
  - *I created the plastic waterfowl she owns*
  - *I caused her to quickly lower her head or body*

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Η γλωσσική διερμηνεία είναι δύσκολη

Ασάφεια

- Πρόταση: “*I made her **duck***”
- Τουλάχιστον 6 διαφορετικές έννοιες:
  - *I cooked waterfowl for her (to eat)*
  - *I cooked waterfowl of her*
  - *I created the plastic waterfowl she owns*
  - *I caused her to quickly lower her head or body*

“*Duck*” can be a Noun or Verb

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Η γλωσσική διερμηνεία είναι δύσκολη

Ασάφεια

- Πρόταση: “*I made **her** duck*”
- Τουλάχιστον 6 διαφορετικές έννοιες:
  - *I cooked waterfowl **for her** (to eat)*
  - *I cooked waterfowl **of her***
  - *I created the plastic waterfowl she owns*
  - *I caused her to quickly lower her head or body*

“*her*” can be:
- a **possessive** pronoun “*of her*”
- a **dative** pronoun “*for her*”

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Η γλωσσική διερμηνεία είναι δύσκολη

Ασάφεια

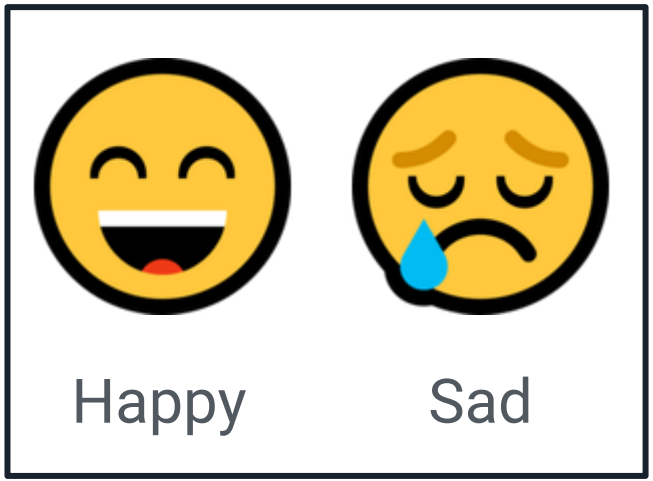
- Πρόταση: “*I **made** her duck*”
- Τουλάχιστον 6 διαφορετικές έννοιες:
  - *I **cooked** waterfowl for her (to eat)*
  - *I **cooked** waterfowl of her*
  - *I **created** the plastic waterfowl she owns*
  - *I **caused** her to quickly lower her head or body*

“*make*” can mean “*cooked*”, “*created*”, or “*caused*”

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Επιπλέον δυσκολίες: Slang, emojis και hashtags

- *“OMG”* = Oh my god
- *“w8”* = wait
- *“brb”* = be right bac



Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Προκλήσεις επί PoS Tagging

ikr smh he asked fir yo last name

so he can add u on fb lolololol

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Προκλήσεις επί PoS Tagging

I know, right	shake my head			for	your		
ikr	smh	he	asked	fir	yo	last	name

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Προκλήσεις επί PoS Tagging

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Προκλητική μορφολογία και σύνταξη

- *“A ship-shipping ship, shipping shipping-ships”.*

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Αντιμετώπιση του προβλήματος

Τι εργαλεία χρειαζόμαστε?

- Γνώση για τη γλώσσα και τον κόσμο.
  Τρόποι συνδυασμού πηγών γνώσης.

Πώς το κάνουμε αυτό?

- Νευρωνικά και άλλα μοντέλα μηχανικής μάθησης που βασίζονται σε γλωσσικά δεδομένα.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Μοντέλα και εργαλεία

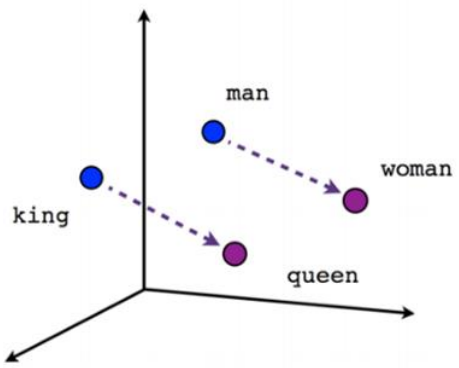
- Regular expressions
- Edit distance
- Γλωσσικά μοντέλα
  Neural word embeddings
- Machine learning classifiers

- Λεξικά Sentiment
- Λεξικά Emotion
- Αλγόριθμοι δικτύου
  Αλγόριθμοι συστάσεων

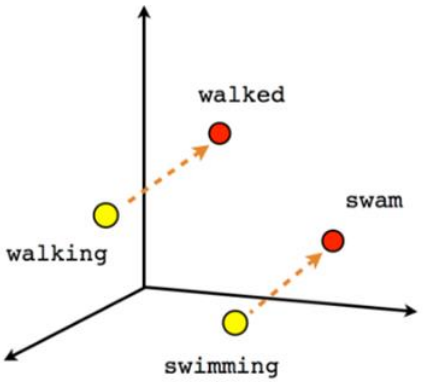
Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Word embeddings

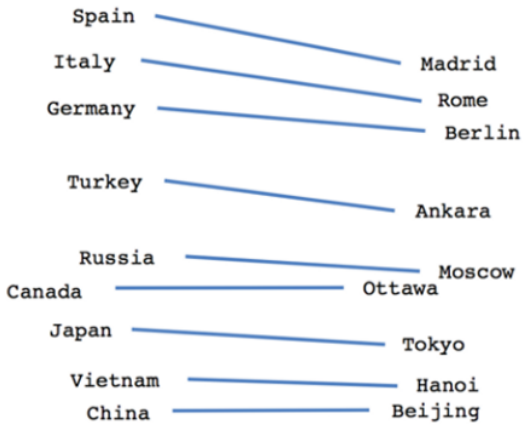
Η σημασιολογική έννοια μιας λέξης ως διάνυσμα 300 διαστάσεων



Male-Female



Verb tense



Country-Capital

Image taken from: https://towardsdatascience.com
Plots are a product of dimensionality reduction to 3D and 2D.

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Embeddings είναι ο πυρήνας του NLP

Word embeddings είναι η βασική τεχνολογία για οποιαδήποτε εργασία NLP:

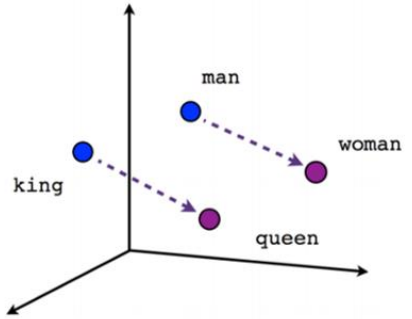
- ❏ Εύρεση συνωνύμων λέξεων.
  Αποφασίζοντας την ομοιότητα δύο προτάσεων.
- ❏ Αποτύπωση του context ενός κειμένου.

Co-financed by the European Union
Connecting Europe Facility

University of Cyprus

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Πώς να μάθετε τα embeddings?



Σπρώξτε τις συν-εμφανιζόμενες λέξεις μαζί στο διάστημα:

Διαβάστε εκατομμύρια λέξεις → Μελετήστε τη συν-εμφάνισή τους.

"Elizabeth II is **Queen** of the United Kingdom ... **Her father** ascended the throne in 1936 upon the abdication of **his** brother, **King** Edward VIII ... **She** was educated privately at home ... In November 1947, **she** married Philip Mountbatten, a former prince of Greece and Denmark ... When **her** father died in February 1952, Elizabeth—then 25 years old—became **queen** regnant ..."

Co-financed by the European Union
Connecting Europe Facility
University of Cyprus
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423