

Natural Language Processing

Introduction and Course Overview

Demetris Paschalides

Department of Computer Science

University of Cyprus

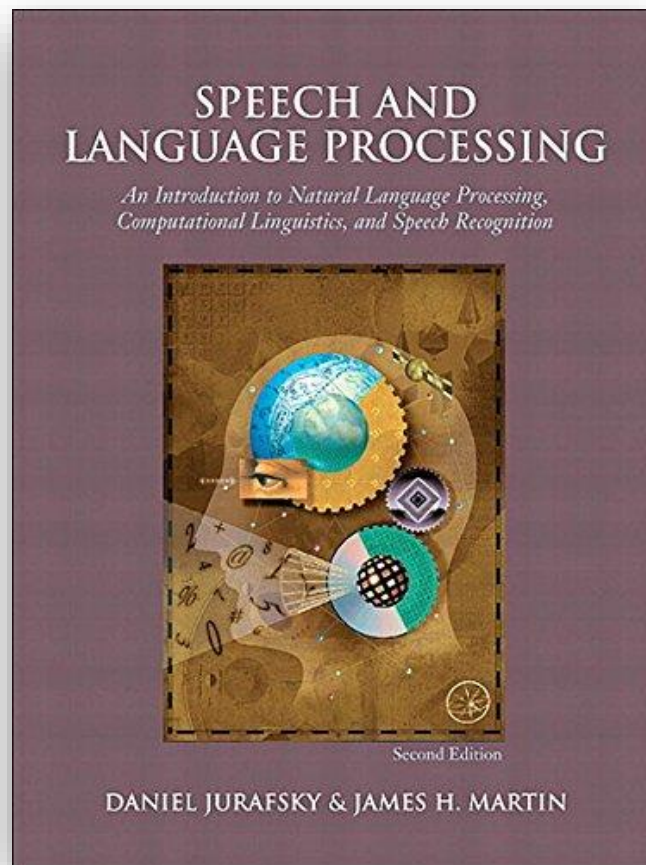


About the Instructor

- **Name:** Demetris Paschalides
- **Professional Experience:** 7.5 years of Academic Research
- **Research Interest:** Use of Natural Language Processing (NLP) and Machine Learning (ML) to address social and ethical challenges.
 - Examples: misinformation, hate-speech, and polarization (social or political).



Textbook



Introduction to NLP



From Language to Information

MAI4CAREU



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



From Language to Information

- Automatically extracting meaning and structure from:
 - Human language text and speech (news, social media, etc.)
 - Social networks
 - Genome sequences



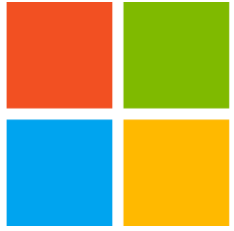
From Language to Information

- Automatically extracting meaning and structure from:
 - Human language text and speech (news, social media, etc.)
 - Social networks
 - Genome sequences

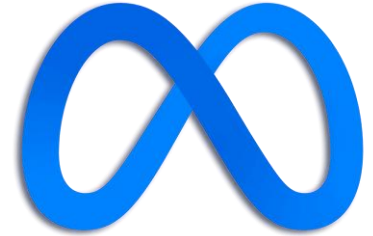
- Interacting with humans via language
 - Dialog systems/Chatbots
 - Question Answering
 - Recommendation Systems



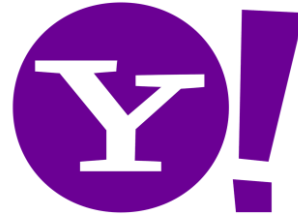
Industry and commercial applications



Microsoft



Meta



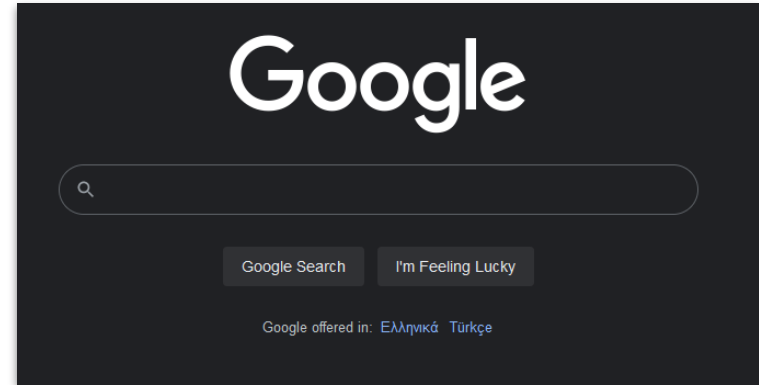
YouTube



Extracting Information from Language

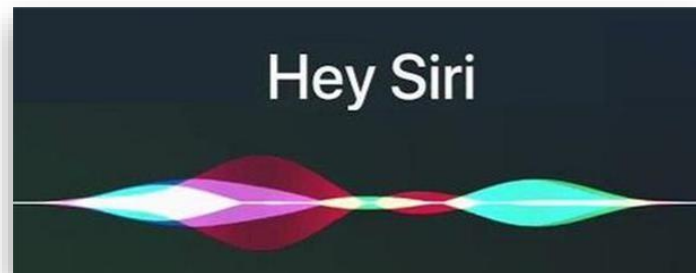
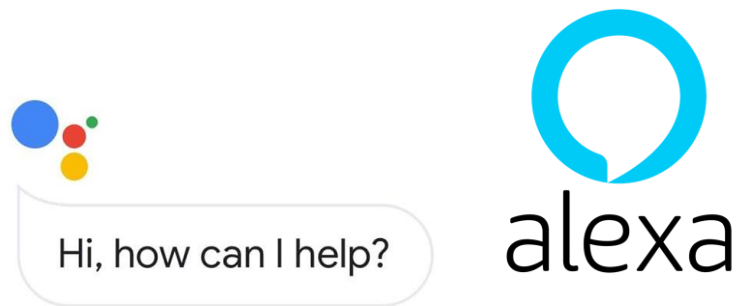
□ Information retrieval

- Year 2020: **6.9 billion** daily Google searches (estimate).
- Text-based information retrieval → the most frequently used software in the world.



Conversational Agents

- Speech Recognition
- Language Analysis
- Dialogue Processing
- Information Retrieval
- Text-to-speech



ChatGPT Revolution

DP

The students of MAI NLP course are here. Please introduce yourself and welcome them.



Hello students of the MAI NLP course, it is my pleasure to introduce myself as ChatGPT, a large language model developed by OpenAI. I am here to assist you with any questions or tasks related to natural language processing. I am excited to work with you and help you learn more about this fascinating field. Welcome to the course!



[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.



ChatGPT **LLMs** Revolution



Stanford
Alpaca



MISTRAL
AI_



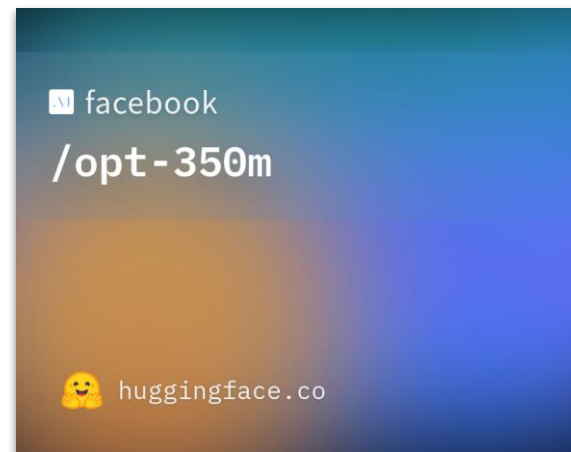
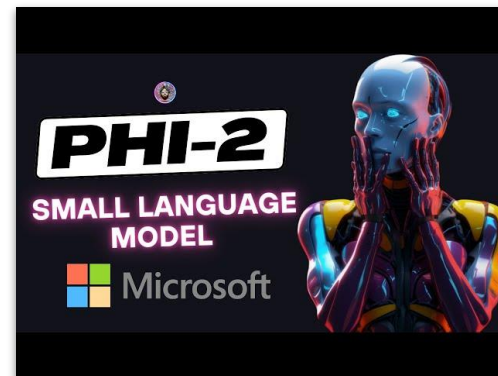
ORCA



ChatGPT LLMs Revolution



Stanford
Alpaca



Text classification: Disaster Response

- Haiti earthquake 2010
- Classifying SMS messages

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.”*



Meier, P., & Munro, R. (2010). The unprecedented role of SMS in disaster response: Learning from Haiti. *SAIS Rev. Int'l Aff.*, 30, 91.

Caragea, C., McNeese, N. J., Jaiswal, A. R., Traylor, G., Kim, H. W., Mitra, P., & Yen, J. (2011, May). Classifying text messages for the Haiti earthquake. In ISCRAM.



Recommendation engines

The good:

- Products: Amazon, ebay
- Content: Netflix, Spotify



NETFLIX



Recommendation engines

The good:

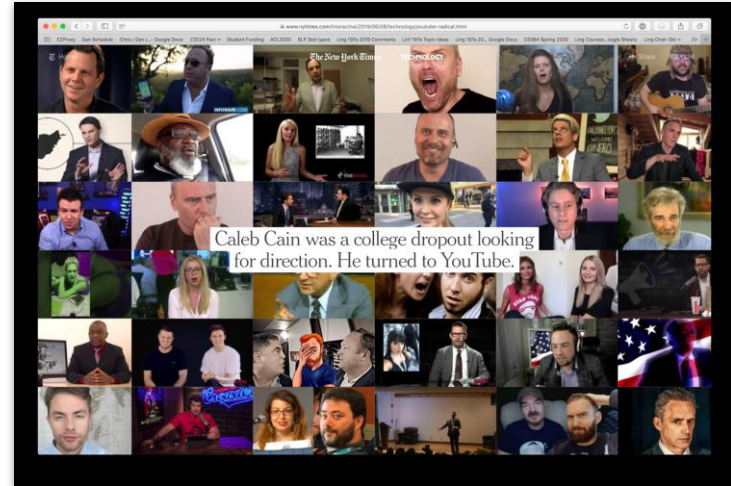
- Products: Amazon, ebay
- Content: Netflix, Spotify



NETFLIX

The bad

- Youtube radicalization



Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). "How over is it?" Understanding the Incel Community on YouTube. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25.



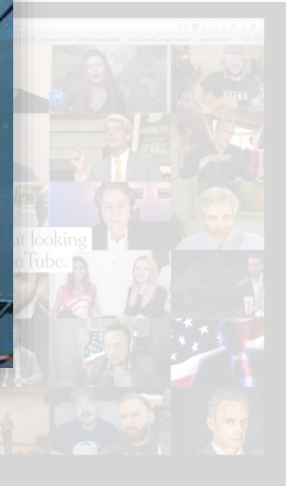
Recommendation engines

The good:

- Products: Amazon
- Content: Netflix



ization

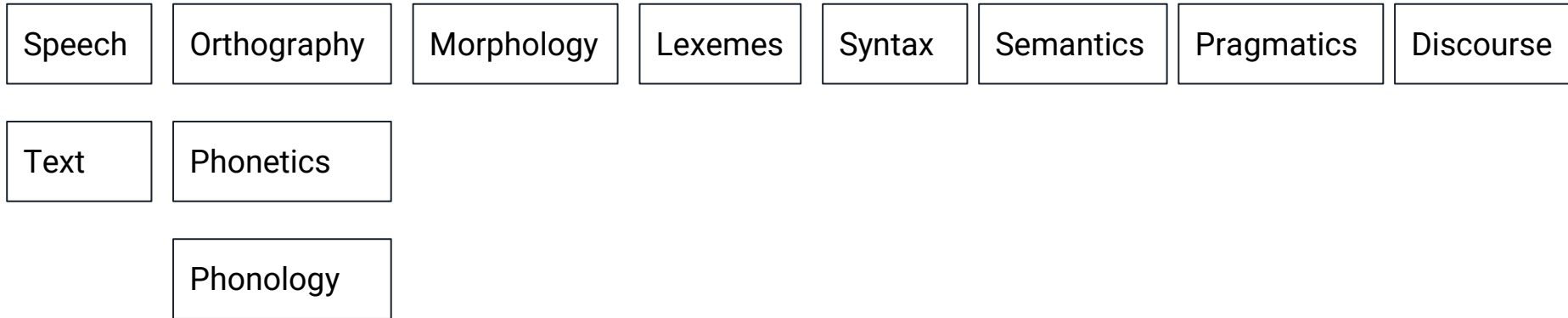


NETFLIX

Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). "How over is it?" Understanding the Incel Community on YouTube. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25.



Linguistic Knowledge Levels



Phonetics and Phonology

□ Pronunciation Modeling

Sounds:

n

T h i a s i e



Words

- Language Modeling
- Tokenization
- Spelling Correction

Words:
n c e

T h i s i s a s i m p l e s e n t e



Morphology

- Morphology Analysis
- Tokenization
- Lemmatization

Words:

n c e

This is a simple sente

Morphology:

be
present



Part-of-Speech

□ Part-of-Speech (PoS) Tagging

PoS: DT VBZ DT JJ NN

Words: This is a simple sentence

Morphology: *be*
present



Syntax

- Syntactic Parsing

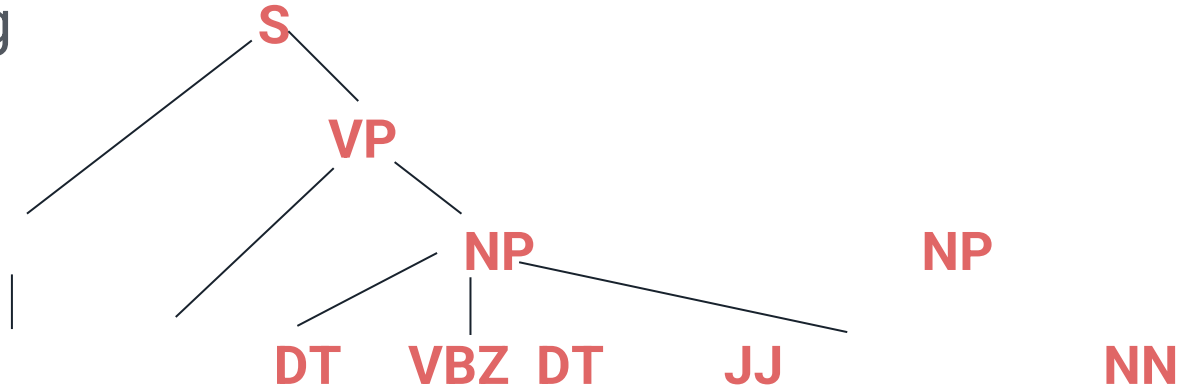
Syntax:

PoS:

Words:

n c e

Morphology:



This is a simple sentence

be

present



Semantics

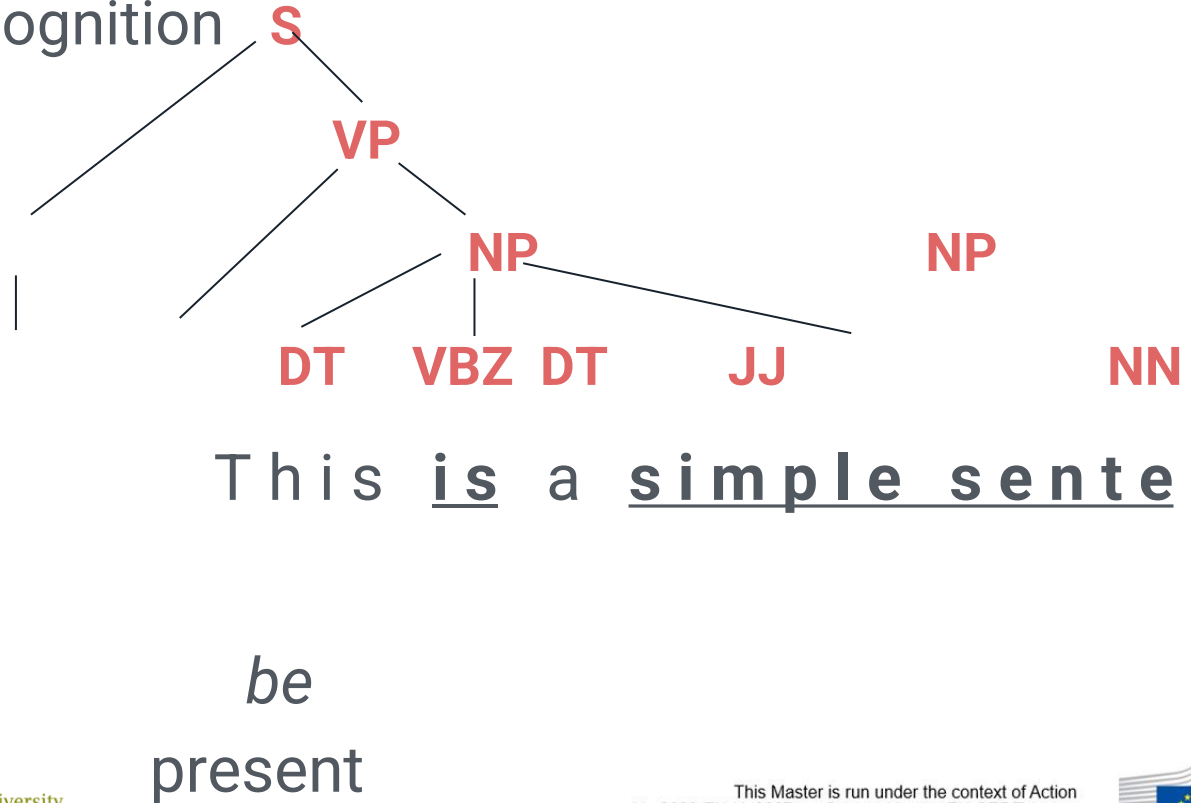
☐ Named Entity Recognition

Syntax:

PoS:

Words:
n c e

Morphology:



Language Interpretation is Hard

MAI4CAREU



Co-financed by the European Union
Connecting Europe Facility



University
of Cyprus

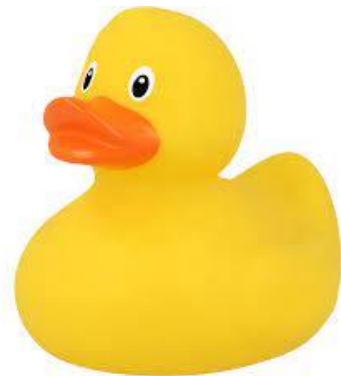
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



Language Interpretation is Hard

Ambiguity

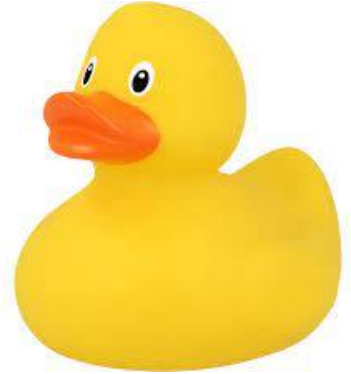
- Sentence: “*I made her duck*”



Language interpretation is hard

Ambiguity

- Sentence: *“I made her duck”*
- At least 6 different meanings:
 - *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*



Language interpretation is hard

Ambiguity

- Sentence: “I made her **duck**”
- At least 6 different meanings:
 - *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



Language interpretation is hard

Ambiguity

- Sentence: “I made **her** duck”
- At least 6 different meanings:
 - *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”



Language Interpretation is Hard

Ambiguity

- Sentence: “I **made** her duck”
- At least 6 different meanings:
 - 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”



Challenges: Slang, Emojis and Hashtags

- “OMG” =
- “w8” =
- “brb” =



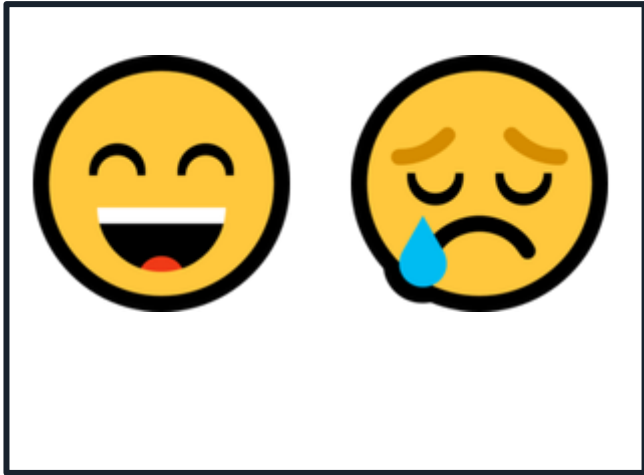
Challenges: Slang, Emojis and Hashtags

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



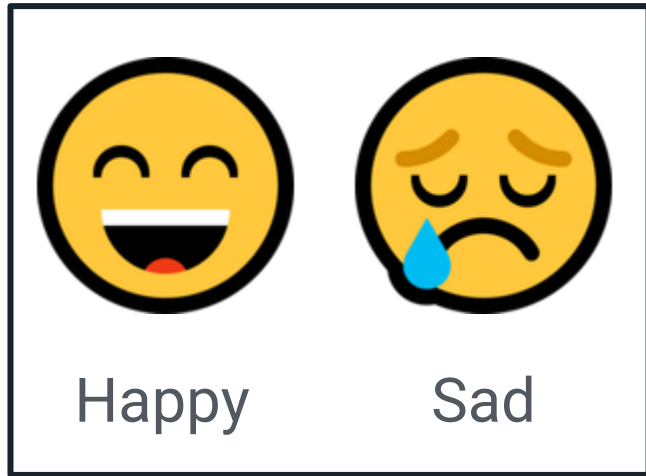
Challenges: Slang, Emojis and Hashtags

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



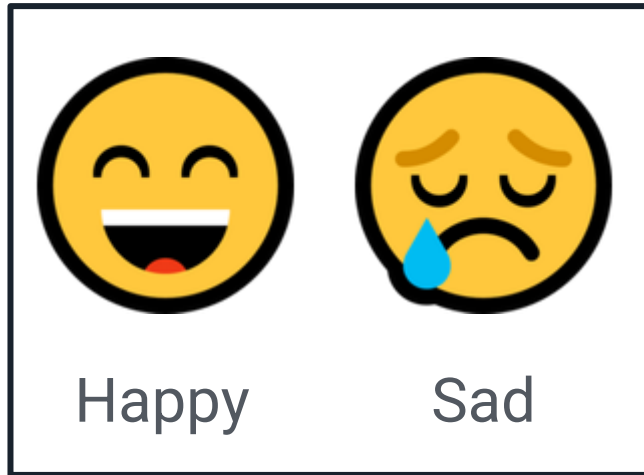
Challenges: Slang, Emojis and Hashtags

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



Challenges: Slang, Emojis and Hashtags

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



Challenges on PoS Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

27



Challenges on PoS Tagging

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

so he can add you on Facebook lololol

Challenges on PoS Tagging

| | | | | | | | |
|---------------|---------------|---------|-------|-------|------|------|------|
| I know, right | shake my head | | | for | your | | |
| ikr | smh | he | asked | fir | yo | last | name |
| ! | G | O | V | P | D | A | N |
| interjection | acronym | pronoun | verb | prep. | det. | adj. | noun |

| | | | | | | | |
|-------------|----|-----|-----|-----|-------------|----------------|---------|
| | | | | you | Facebook | laugh out loud | |
| so | he | can | add | u | on | fb | lololol |
| P | O | V | V | O | P | ^ | ! |
| preposition | | | | | proper noun | | |

Challenging Morphology and Syntax

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



Tackling the problem

What tools do we need?

- ❑ Knowledge about language and the world.
- ❑ Ways to combine knowledge sources.

How we do this?

- ❑ Neural and other machine learning models build from language data



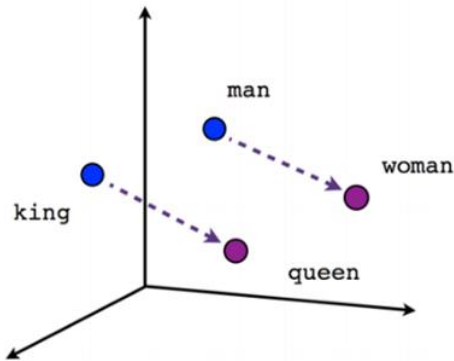
Models and tools

- Regular Expressions
- Edit Distance
- Language Models
- Neural Word Embeddings
- Machine Learning Classifiers
- Sentiment Lexicons
- Emotion Lexicons
- Network Algorithms
- Recommendation Algorithms

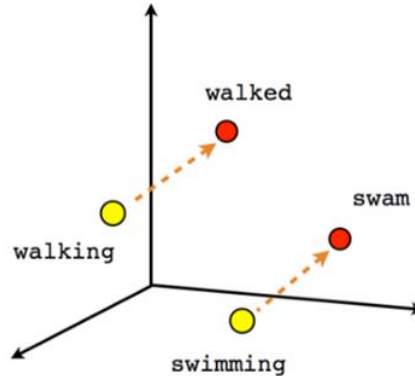


Word embeddings

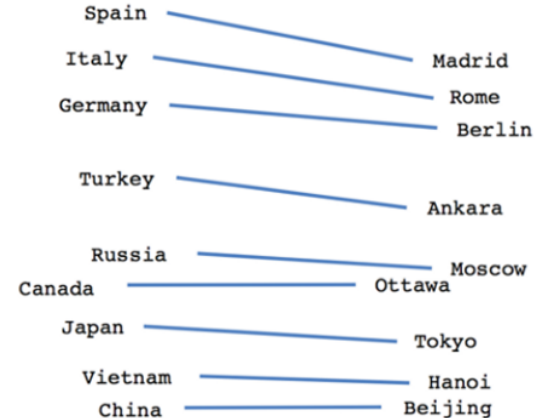
A word's semantic meaning as a 300-dimensional vector



Male-Female



Verb tense



Country-Capital

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

Embeddings are the core of NLP

Word embeddings are the core technology for any NLP task:

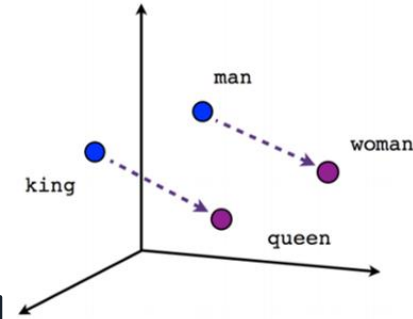
- Finding synonyms of words.
- Deciding the similarity of two sentences.
- Capturing the context of a text.



How to learn the embeddings?

Push co-occurring words together in space:

- ☐ Read millions of words → Study their co-occurrence.



“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.”

Course Outline

1. Text Pre-processing
2. Language Modeling
3. Text Classification
4. Word Vector Representation
5. Distributed Contextual Embeddings
6. Application of NLP in:
 - a. Hate-speech Identification
 - b. Fake News Detection
 - c. Political Polarization
7. Introduction to Large Language Models



Thank You

