Introduction to natural language processing for health and biological questions

Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

université PARIS-SACLAY

Course overview

This class is an introduction to fundamental concepts in natural language processing for biomedical data, the subdiscipline of artificial intelligence that tries to make the computers "understand". This will survey a variety of interesting language problems and techniques.

Learning objectives

- Apply and evaluate statistical methods to real data and answer scientific questions.
- Write computer programs to analyse language data.
- Understand ethical issues relevant to using and working with real data.
- Identify and answer questions that involve applying statistical methods or machine learning algorithms to complex language data.
- Work in a team to solve NLP problems.
- Present the results and limitations of a data analysis at appropriate technical levels for the intended audience.

Evaluation

- Assignment (3 assignments; 20% each - TP notés)
- Project code and report (30%) - At least a two-page report in the form of a conference paper:
  - Motivation & Intro (what is the problem and the background and why are you trying to solve it)
  - Data (describe the data and the distribution of annotations, sizes of the data splits, i.e., train, dev, test)
  - Method and Evaluation setup (describe your method, what the frameworks and tools you are using, what baselines you are comparing with, what are the hyperparameters and you set them, what evaluation metrics you use)
  - Results and Discussion (what results you get in comparison with the baselines and why, error analysis)
  - Conclusion
- Project presentation (midterm and final) (10%)

AI assistance, Online resources, deadlines

- Do NOT use any online codes.
- Do NOT use any codes generated by AI.
- Assignments are due on Tuesdays at 5pm, no late submission allowed.

Lectures

1. Introduction

2. Linguistic essentials: tokenization, part-of-speech tagging, word sense disambiguation, sentence splitting

3. Data preparation: corpus annotation, evaluation measures

4. Biomedical resources and corpora

5. Language modeling

5. Named entity recognition

6. Classification

7. Ethical challenges (data and privacy, misrepresentation and bias, cost of prediction errors, dual use of technology)

8. Final project presentations

Course Calendar

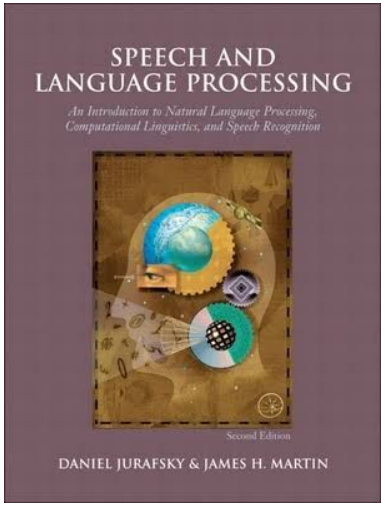
Jan 15	Intro, Linguistics essentials
January 22	Data preparation, Biomedical resources, Assignment 1
January 29	Language modeling,Classification, Assignment 2
February 5	Named entity recognition, Ethical challenges Assignment 3
February 12	Free session - project, project deadline 16 Feb on ecampus (latest 19)
February 19	Exam - project presentation - D104
February 26	Vacance

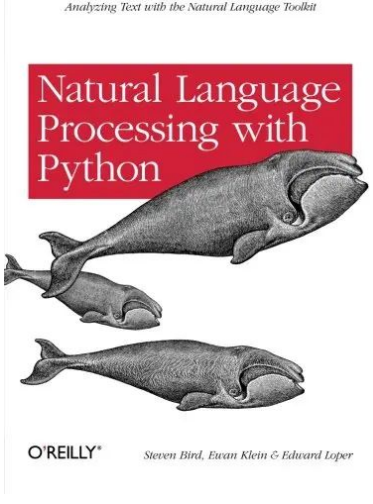
Reading list

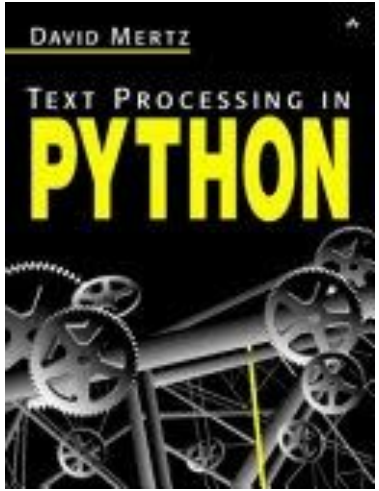
- D. Jurafsky & J. Martin, *Speech and Language Processing*, Prentice Hall, 2nd ed., 2009.

See also the draft 3rd edition

- S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, O'Reilly, 2009.
- D. Mertz, Text Processing in Python, Addison Wesley, 2003.







Why natural language processing?

Computers have to talk and understand like humans.

Computers as personal assistants.

Computers as researchers: answering questions, classifying information.

Computers as language experts, perform translations.

Pathology speech recognition

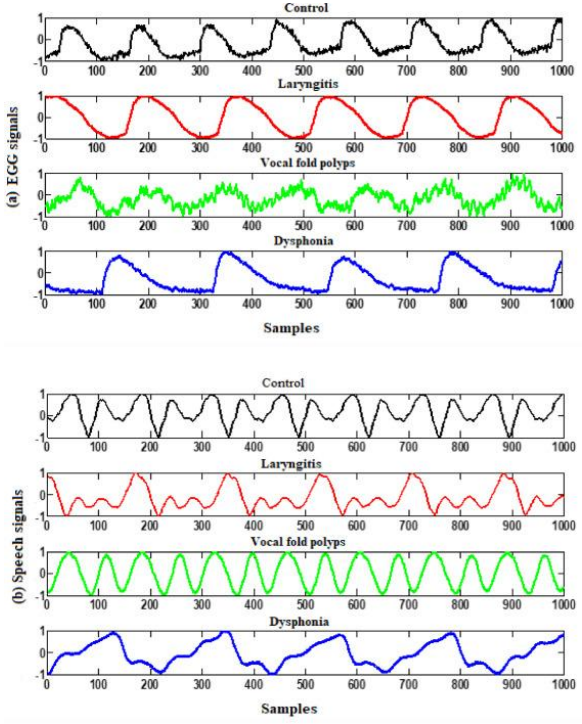


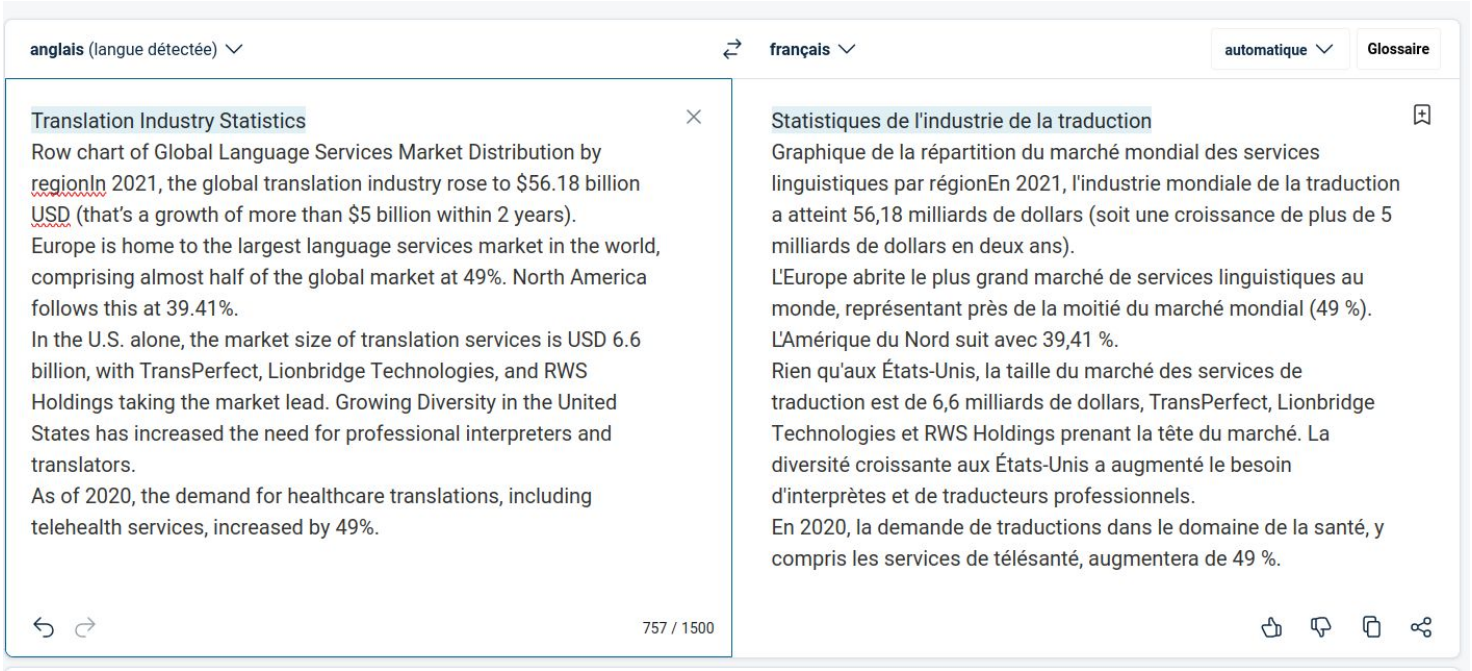
Figure source: https://doi.org/10.1016/j.cmpbup.2022.100074

Alzheimer disease detection

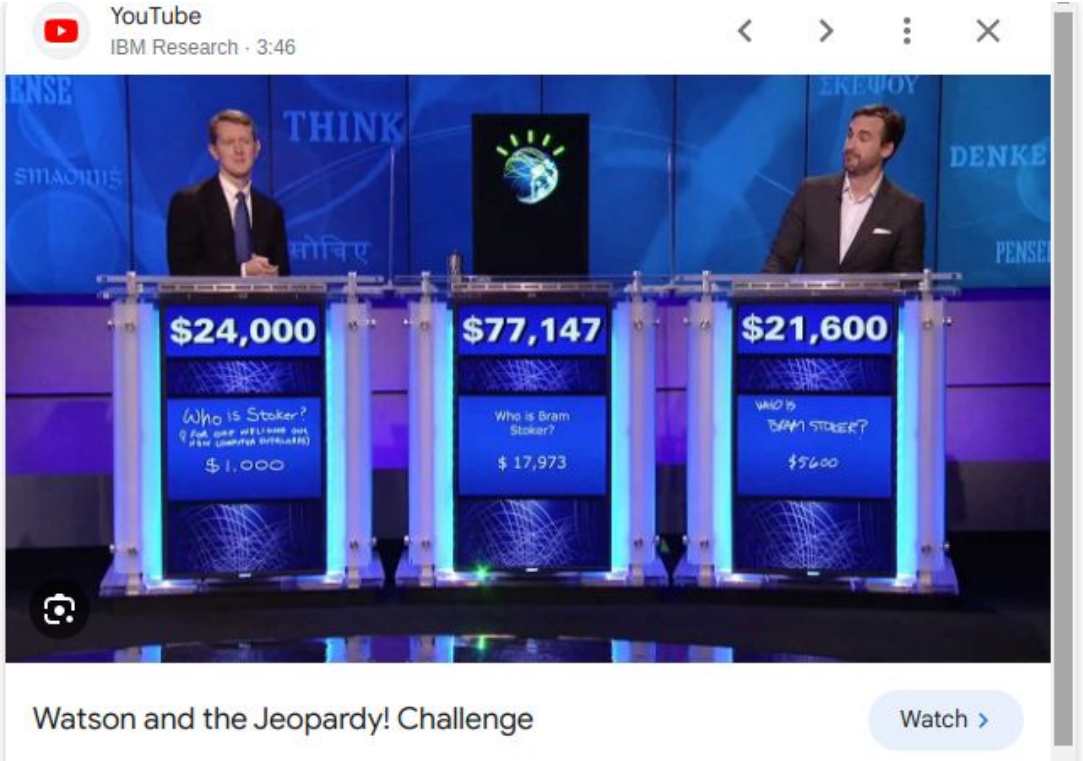
ADD: Dementia, PD: Parkinson's

Figure source: https://doi.org/10.1002/dad2.12276

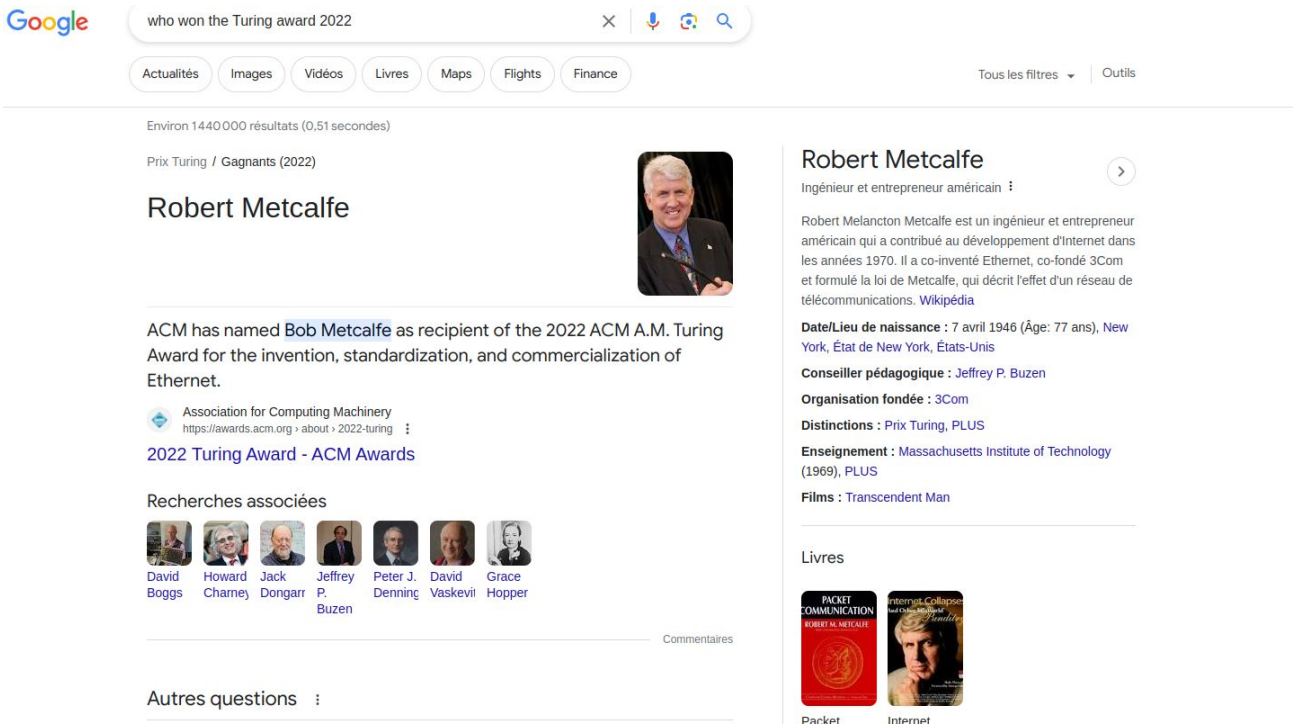
Machine translation



Question and answering



Information retrieval



Chatbots

Hi, how can I help?

Other example NLP tasks

- Document summarization
- Information extraction
- Natural language inference
- Author identification
- Image caption generation
- Document classification

Why is NLP difficult?

- Highly ambiguous at different levels,
- Is fuzzy and probabilistics,
- Involves world knowledge reasoning,
- Is complex and subtle.

Tokenization

A token can be:

- A word
- A sub-word
- A character
- A sequence of characters

Tokenization is the task of segmenting raw textual data into tokens

Exercise

Tokenize the following abstract https://pubmed.ncbi.nlm.nih.gov/10090885/ by white space.

Exercise

Use Scispacy (https://github.com/allenai/scispacy) to tokenize the following abstract https://pubmed.ncbi.nlm.nih.gov/10090885/.

Challenges with biomedical text tokenization

- Biomedical terms contain digits, capitalized letters, Latin and Greek letters, Roman digits, measurement units, hyphens and other special symbols…
- Includes abbreviations, acronyms, …

Sentence splitting

Dividing the text into sentences.

Example: SARS-CoV-2 has rapidly spread across the globe and infected hundreds of millions of people worldwide. As our experience with this virus continues to grow, our understanding of both short-term and long-term complications of infection with SARS-CoV-2 continues to grow as well. Just as there is heterogeneity in the acute infectious phase, there is heterogeneity in the long-term complications seen following COVID-19 illness. The purpose of this review article is to present the current literature with regards to the epidemiology, pathophysiology, and proposed management algorithms for the various long-term sequelae that have been observed in each organ system following infection with SARS-CoV-2. We will also consider future directions, with regards to newer variants of the virus and their potential impact on the long-term complications observed.

Stemming

The process of removing word suffixes.

Programmer, programming, programs → program

Exercise

Use Porter Stemmer to perform the stemming of words in the abstract.

Lemmatization

Reducing the word into its base form. Requires morphological analysis to identify the lemma of each word.

Better → good

Parts of Speech tagging

Finding the grammatical category of each word

Nouns: denote an object, a concept, a place, …

• Count nouns: dog, spleen, Band-Aid, …

• Mass nouns: water, wheat, …

• Proper nouns: Fred, New York City, …

• Pronouns: he, she, you, I, they, …

• Adjectives: denote an attribute of the denotation of a noun. • Intersective: pink, furry, … • Measure: big, … • Intensional: former, alleged,

Exercise

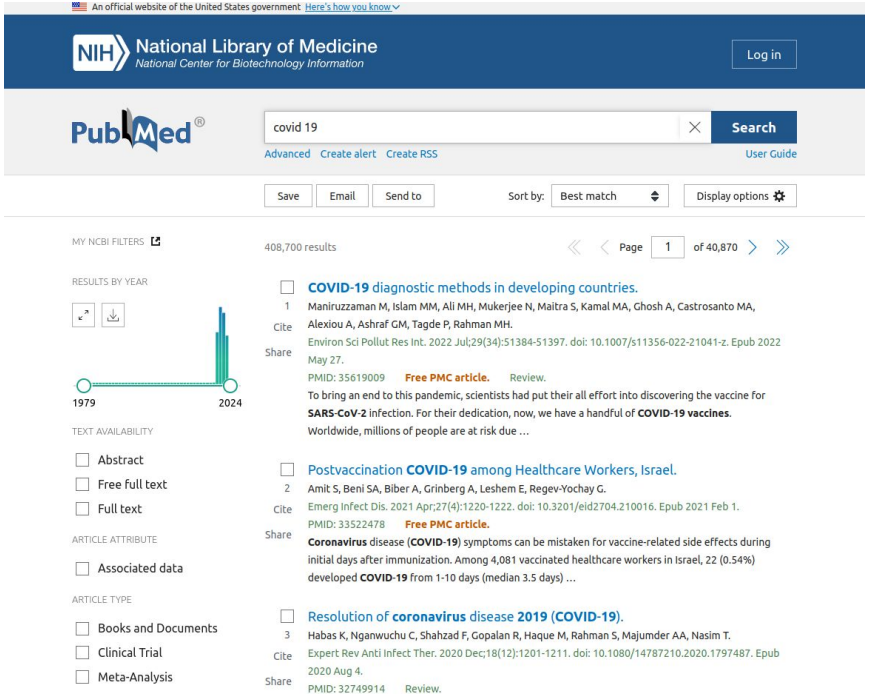
Use NLTK (https://www.nltk.org/book/ch05.html) to POS tag the previous abstract.

Exercise

Find the vocabulary list of the abstract.

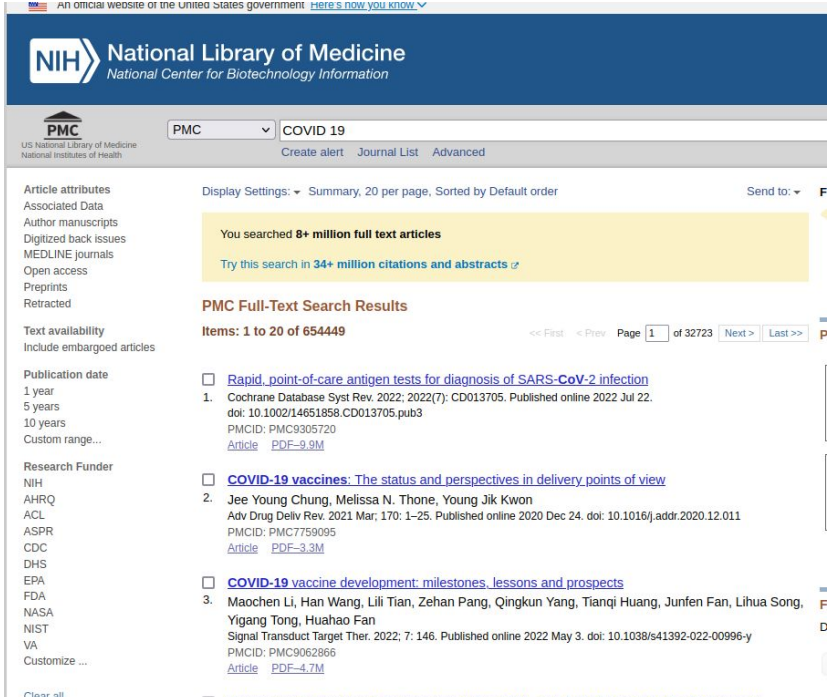
Biomedical resources

Biomedical and life sciences articles dating from the 1950s to the present (MEDLINE): //pubmed.gov



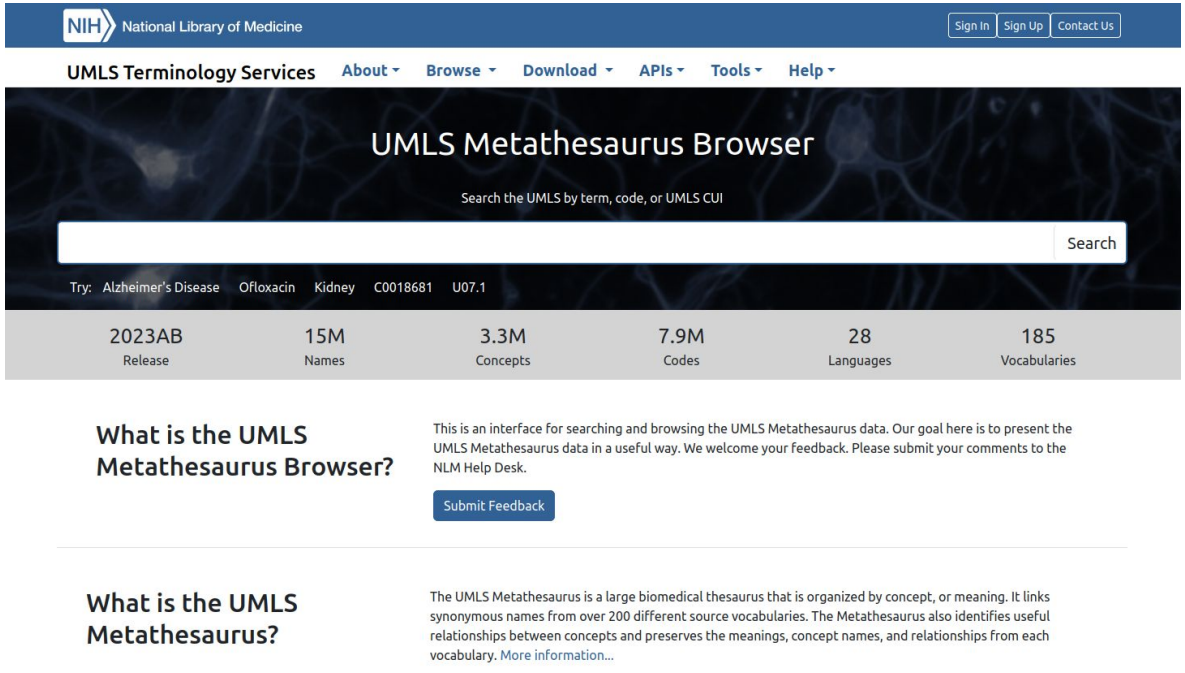
PubMed Central (PMC)

Digital archive of free biomedical and life sciences journal literature developed by NLM: https://www.ncbi.nlm.nih.gov/pmc/



Unified Medical Language System (UMLS)

The Metathesaurus of UMLS [Bodenreider, 2004] combines concepts from over 200 source ontologies.



Annotated, freely available, biomedical corpora examples

Corpora	Text Genre	Annotations
NCBI disease	Scientific articles	Disease entities
MedNLI	Patient records	Natural language inference, textual entailment
BioASQ	Medical articles	Question and answering

Document representation

1- hot encoding: most basic representation

“I walked my dog”

I =[1,0,0,0]

walked=[0,1,0,0]

my=[0,0,1,0]

dog=[0,0,0,1]

Challenges with 1-hot encoding

- The meaning of words are ignored.
- Sparse.

Distributional word representation

- A word is defined by its context.
- Approaches:
  - Count-based
  - Prediction-based

Count-based

Measure the frequency of words in the context of each word in the vocabulary.

Vector representations are defined based on those frequency.

Co-occurrence matrix

	walk	bark	pet
dog	2	5	3
car	1	0	0

Limits of count-based representation

High sensitivity to frequent words or to very infrequent words.

Exercise

Use Scikitlearn to represent a document.