

Introduction to natural language processing for health and biological questions

Nona Naderi

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

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

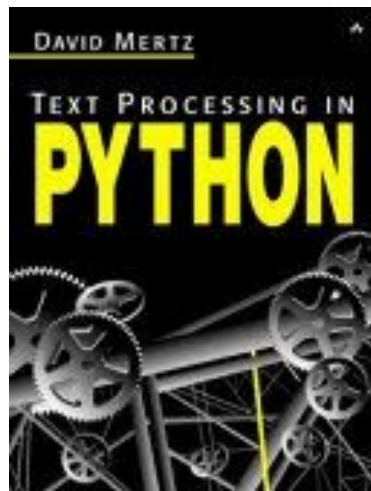
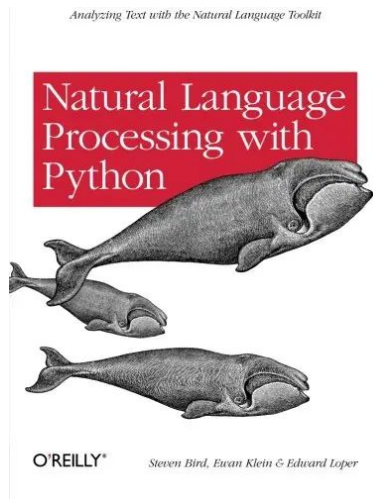
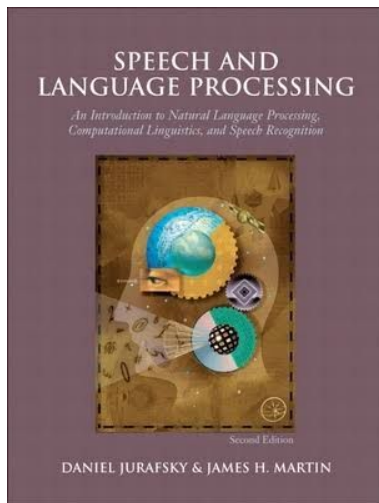
Jan 15	Intro, Linguistics essentials
January 22	Data preparation, Biomedical resources, Assignment 1
January 29	Language modeling, Classification, Assignment 2
February 5	Named entity recognition, Ethical challenges Assignment 3
February 12	Free session - project, project deadline 16 Feb on ecampus (latest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

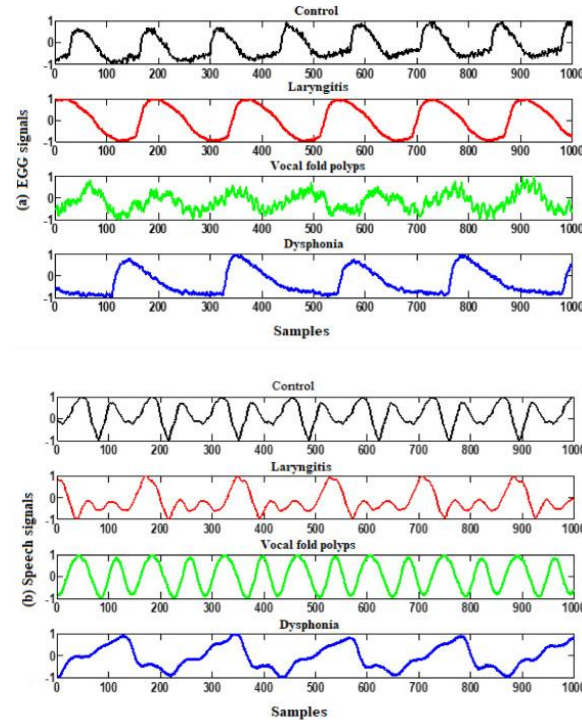
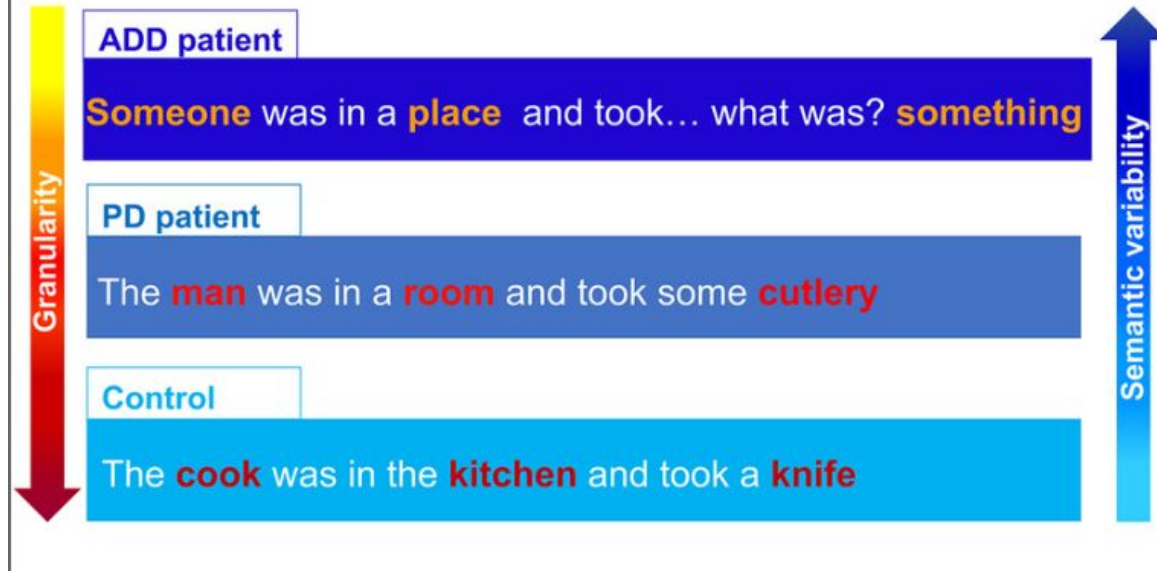


Figure source: <https://doi.org/10.1016/j.cmpbup.2022.100074>

Alzheimer disease detection

(A) Representative phrases of each group



ADD: Dementia, PD: Parkinson's

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

Machine translation

The screenshot displays a machine translation interface. At the top, the source language is set to 'anglais (langue détectée)' and the target language is 'français'. The translation mode is 'automatique'. The interface is split into two main panels. The left panel contains the original English text, and the right panel contains the translated French text. The English text discusses the 'Translation Industry Statistics' and mentions a 'Row chart of Global Language Services Market Distribution by region'. The French text is a direct translation of this content. At the bottom of the interface, there are navigation icons (back, forward), a character count '757 / 1500', and social media sharing icons (thumbs up, thumbs down, print, share).

anglais (langue détectée) ↕ français

automatique Glossaire

Translation Industry Statistics

Row chart of Global Language Services Market Distribution by region. In 2021, the global translation industry rose to \$56.18 billion USD (that's a growth of more than \$5 billion within 2 years). Europe is home to the largest language services market in the world, comprising almost half of the global market at 49%. North America follows this at 39.41%. In the U.S. alone, the market size of translation services is USD 6.6 billion, with TransPerfect, Lionbridge Technologies, and RWS Holdings taking the market lead. Growing Diversity in the United States has increased the need for professional interpreters and translators. As of 2020, the demand for healthcare translations, including telehealth services, increased by 49%.

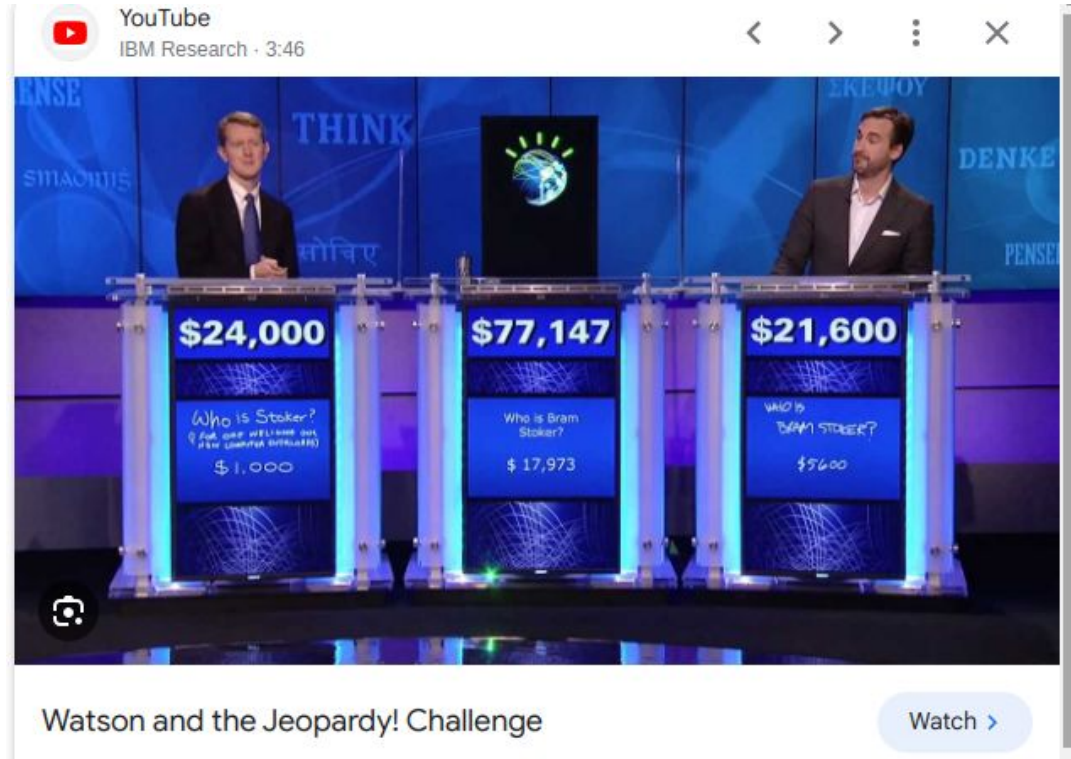
Statistiques de l'industrie de la traduction

Graphique de la répartition du marché mondial des services linguistiques par région. En 2021, l'industrie mondiale de la traduction a atteint 56,18 milliards de dollars (soit une croissance de plus de 5 milliards de dollars en deux ans). L'Europe abrite le plus grand marché de services linguistiques au monde, représentant près de la moitié du marché mondial (49 %). L'Amérique du Nord suit avec 39,41 %. Rien qu'aux États-Unis, la taille du marché des services de traduction est de 6,6 milliards de dollars, TransPerfect, Lionbridge Technologies et RWS Holdings prenant la tête du marché. La diversité croissante aux États-Unis a augmenté le besoin d'interprètes et de traducteurs professionnels. En 2020, la demande de traductions dans le domaine de la santé, y compris les services de télésanté, augmentera de 49 %.

↶ ↷ 757 / 1500

👍 👎 🖨️ 🔗

Question and answering



Information retrieval



who won the Turing award 2022



Actualités

Images

Vidéos

Livres

Maps

Flights

Finance

Tous les filtres ▾

Outils

Environ 1440000 résultats (0,51 secondes)

Prix Turing / Gagnants (2022)

Robert Metcalfe



ACM has named **Bob Metcalfe** as recipient of the 2022 ACM A.M. Turing Award for the invention, standardization, and commercialization of Ethernet.

 Association for Computing Machinery
<https://awards.acm.org/about/2022-turing>

2022 Turing Award - ACM Awards

Recherches associées



David
Boggs



Howard
Charney



Jack
Dongarra



Jeffrey
P. Buzen



Peter J.
Denning



David
Vaskevili



Grace
Hopper

Commentaires

Autres questions :

Robert Metcalfe

Ingénieur et entrepreneur américain :

Robert Melancton Metcalfe est un ingénieur et entrepreneur américain qui a contribué au développement d'Internet dans les années 1970. Il a co-inventé Ethernet, co-fondé 3Com et formulé la loi de Metcalfe, qui décrit l'effet d'un réseau de télécommunications. [Wikipédia](#)

Date/Lieu de naissance : 7 avril 1946 (Âge: 77 ans), New York, État de New York, États-Unis

Conseiller pédagogique : Jeffrey P. Buzen

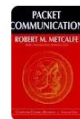
Organisation fondée : 3Com

Distinctions : Prix Turing, PLUS

Enseignement : Massachusetts Institute of Technology (1969), PLUS

Films : Transcendent Man

Livres



Packet



Internet

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](https://pubmed.gov)

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

PubMed

covid 19

[Advanced](#) [Create alert](#) [Create RSS](#) [User Guide](#)

Sort by: Best match

MY NCBI FILTERS

408,700 results

Page 1 of 40,870

RESULTS BY YEAR

TEXT AVAILABILITY

☐ Abstract
☐ Free Full text
☐ Full text

ARTICLE ATTRIBUTE

☐ Associated data

ARTICLE TYPE

☐ Books and Documents
☐ Clinical Trial
☐ Meta-Analysis

☐ **COVID-19 diagnostic methods in developing countries.**
1 Maniruzzaman M, Islam MM, Ali MH, Mukerjee N, Maitra S, Kamal MA, Ghosh A, Castrosanto MA, Alexiou A, Ashraf GM, Tagde P, Rahman MH.
Environ Sci Pollut Res Int. 2022 Jul;29(34):51384-51397. doi: 10.1007/s11356-022-21041-z. Epub 2022 May 27.
PMID: 35619009 **Free PMC article.** Review.
To bring an end to this pandemic, scientists had put their all effort into discovering the vaccine for SARS-CoV-2 infection. For their dedication, now, we have a handful of **COVID-19 vaccines**. Worldwide, millions of people are at risk due ...

☐ **Postvaccination COVID-19 among Healthcare Workers, Israel.**
2 Amit S, Beni SA, Bilber A, Grinberg A, Leshem E, Regev-Yochay G.
Emerg Infect Dis. 2021 Apr;27(4):1220-1222. doi: 10.3201/eid2704.210016. Epub 2021 Feb 1.
PMID: 33522478 **Free PMC article.**
Coronavirus disease (COVID-19) symptoms can be mistaken for vaccine-related side effects during initial days after immunization. Among 4,081 vaccinated healthcare workers in Israel, 22 (0.54%) developed **COVID-19** from 1-10 days (median 3.5 days) ...

☐ **Resolution of coronavirus disease 2019 (COVID-19).**
3 Habas K, Nganwucho C, Shahzad F, Gopalan R, Haque M, Rahman S, Majumder AA, Nasim T.
Expert Rev Anti Infect Ther. 2020 Dec;18(12):1201-1211. doi: 10.1080/14787210.2020.1797487. Epub 2020 Aug 4.
PMID: 32749914 **Review.**

PubMed Central (PMC)

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

The screenshot displays the PubMed Central (PMC) website interface. At the top, there is a navigation bar with the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". Below this, a search bar contains the text "COVID 19". To the left of the search bar, there is a dropdown menu with "PMC" selected. Below the search bar, there are links for "Create alert", "Journal List", and "Advanced".

On the left side of the page, there is a sidebar with various filters and options, including "Article attributes", "Text availability", "Publication date", and "Research Funder".

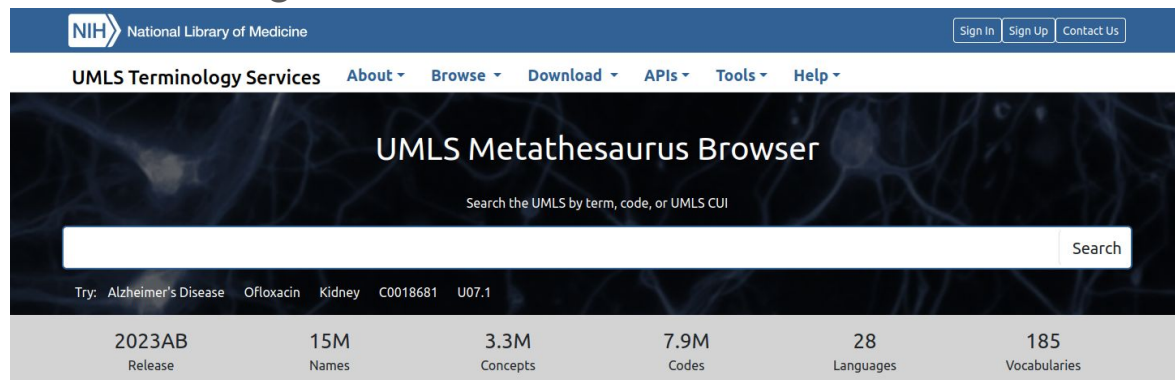
The main content area shows the search results for "COVID 19". It includes a summary of the search results, stating "You searched 8+ million full text articles" and "Try this search in 34+ million citations and abstracts". Below this, there is a section titled "PMC Full-Text Search Results" with "Items: 1 to 20 of 654449". The results are displayed in a list format, with each item showing the title, authors, publication date, and links to the full text and PDF.

The first three results are:

- Rapid point-of-care antigen tests for diagnosis of SARS-CoV-2 infection**
Cochrane Database Syst Rev. 2022; 2022(7): CD013705. Published online 2022 Jul 22.
doi: 10.1002/14651858.CD013705.pub3
PMCID: PMC9305720
[Article](#) [PDF-9.9M](#)
- COVID-19 vaccines: The status and perspectives in delivery points of view**
Jee Young Chung, Melissa N. Thone, Young Jik Kwon
Adv Drug Deliv Rev. 2021 Mar; 170: 1–25. Published online 2020 Dec 24. doi: 10.1016/j.addr.2020.12.011
PMCID: PMC7759095
[Article](#) [PDF-3.3M](#)
- COVID-19 vaccine development: milestones, lessons and prospects**
Maochen Li, Han Wang, Lili Tian, Zehan Pang, Qingkun Yang, Tianqi Huang, Junfen Fan, Lihua Song, Yigang Tong, Huahao Fan
Signal Transduct Target Ther. 2022; 7: 146. Published online 2022 May 3. doi: 10.1038/s41392-022-00996-y
PMCID: PMC9062866
[Article](#) [PDF-4.7M](#)

Unified Medical Language System (UMLS)

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



The screenshot shows the UMLS Metathesaurus Browser interface. At the top is the NIH National Library of Medicine logo and navigation links (Sign In, Sign Up, Contact Us). Below is a menu bar with UMLS Terminology Services, About, Browse, Download, APIs, Tools, and Help. The main header reads 'UMLS Metathesaurus Browser' with a search prompt 'Search the UMLS by term, code, or UMLS CUI'. A search bar with a 'Search' button is present. Below the search bar, a row of suggestions includes 'Try: Alzheimer's Disease', 'Ofloxacin', 'Kidney', 'C0018681', and 'U07.1'. At the bottom, a statistics bar displays: 2023AB Release, 15M Names, 3.3M Concepts, 7.9M Codes, 28 Languages, and 185 Vocabularies.

What is the UMLS Metathesaurus Browser?

This is an interface for searching and browsing the UMLS Metathesaurus data. Our goal here is to present the UMLS Metathesaurus data in a useful way. We welcome your feedback. Please submit your comments to the NLM Help Desk.

[Submit Feedback](#)

What is the UMLS Metathesaurus?

The UMLS Metathesaurus is a large biomedical thesaurus that is organized by concept, or meaning. It links synonymous names from over 200 different source vocabularies. The Metathesaurus also identifies useful relationships between concepts and preserves the meanings, concept names, and relationships from each vocabulary. [More information...](#)

Annotated, freely available, biomedical corpora examples

Corpora	Text Genre	Annotations
<u>NCBI disease</u>	Scientific articles	Disease entities
<u>MedNLI</u>	Patient records	Natural language inference, textual entailment
<u>BioASQ</u>	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.