UNIVERSITÉ DE LORRAINE

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

Data Mining for NLP

1- Basics of Textual Data Exploration

These slides will be available on Arche

By Gaël Guibon, inspired by Benjamin Muller's course

Course Objective

Goal: Use NLP methods for Data Mining

- Get used to data handling (an often skipped part)
- Present methods to **explore textual data**
- Focus on **Machine Learning methods** to de
- Focus on **empirical approaches**

Course Logistics

- 3 sessions
- 2x 1h30 lectures
- 2x 1h30 and a 3h lab
- Q&A on Arche? Discord?

Material on Arche

Course Evaluation

- **Final Exam**
  - Mix between questions and code completion

Lectures Outline

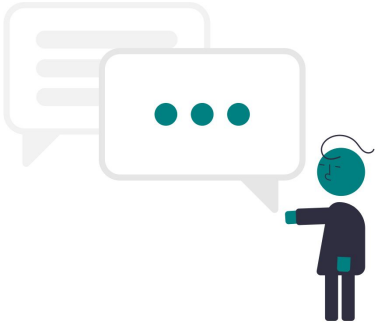
1. Basics of Textual Data Exploration
2. Data Exploration through NLP Tasks
3. Data Representation

Labs Outline

1. Describe Statistically Large Scale Corpora
2. Classifiers to Explore Data
3. Language Models and Clustering

Today Lecture Outline

- Why Natural Language Processing?
- Why Data Mining?
- Why is Language Hard to Model?
- Statistical Description of a Corpus
- How to Represent Data



Why Natural Language Processing?

Why Natural Language Processing?

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

Why NLP ?

- **Access Knowledge** (search engine, recommender system…)
- **Communicate** (e.g. Translation)
- **Linguistics** and **Cognitive Sciences** (Analyse Languages themselves)

Why Natural Language Processing?

Amount of online textual data…

- 70 billion web-pages online (1.9 billion websites)
- 55 million Wikipedia articles

…Growing at a fast pace

- 9000 tweets/second
- 3 million mail / second (60% spam)

Internet Stats

Why Natural Language Processing?

Potential Users of Natural Language Processing

- 7.9 billion people use some sort of language (January 2022)
- 4.7 billion internet users (January 2021) (~59%)
- 4.2 billion social media users (January 2021) (~54%)

Data Reportal

Why Natural Language Processing?

What Products ?

- Search: +2 billion Google users, 700 millions Baidu users
- Social Media: +3 billion users of Social media (Facebook, Instagram, WeChat, Twitter...)
- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)
- Machine Translation: 500M users for google translate

Data Reportal

That's a Lot of Data!

Why Data Mining?

Why Data Mining?

Data mining is used to handle:

- **Raw data**, not specifically filtered
- **Millions** of data
- Hundred of variables

Data Mining **focuses on "fast" calculus**.

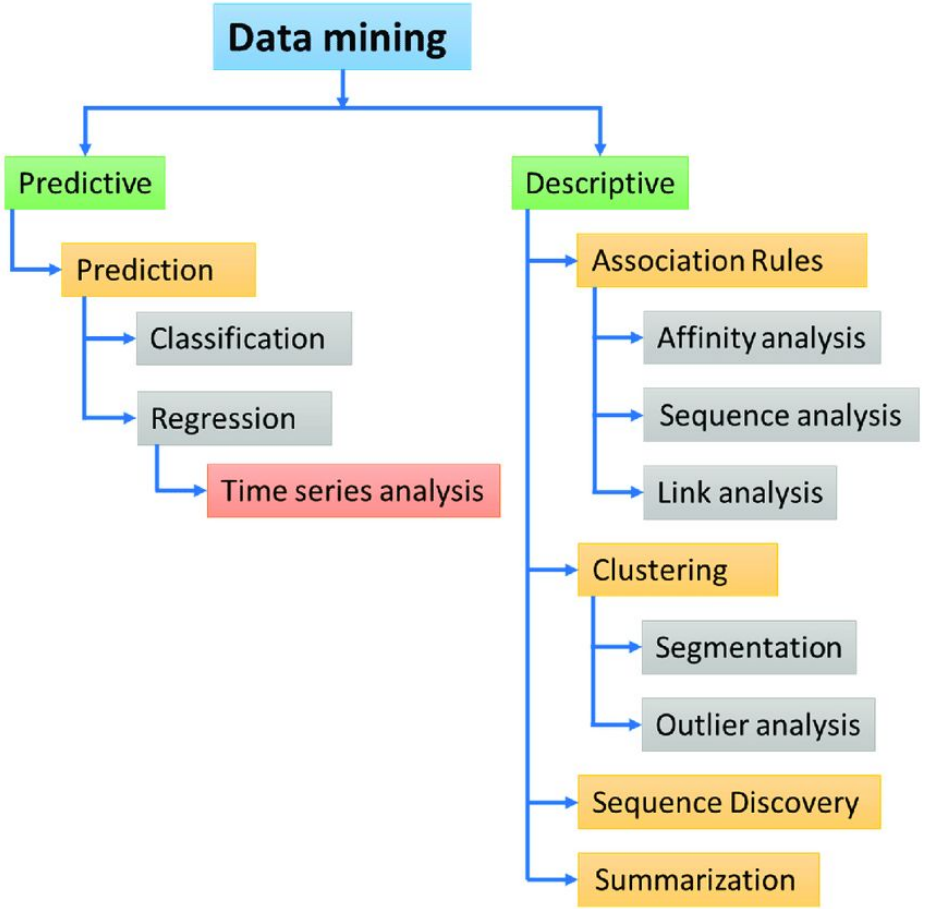
Why Data Mining?

It is a mix between:

- Databases
- Statistics
- Visualization
- Information Sciences

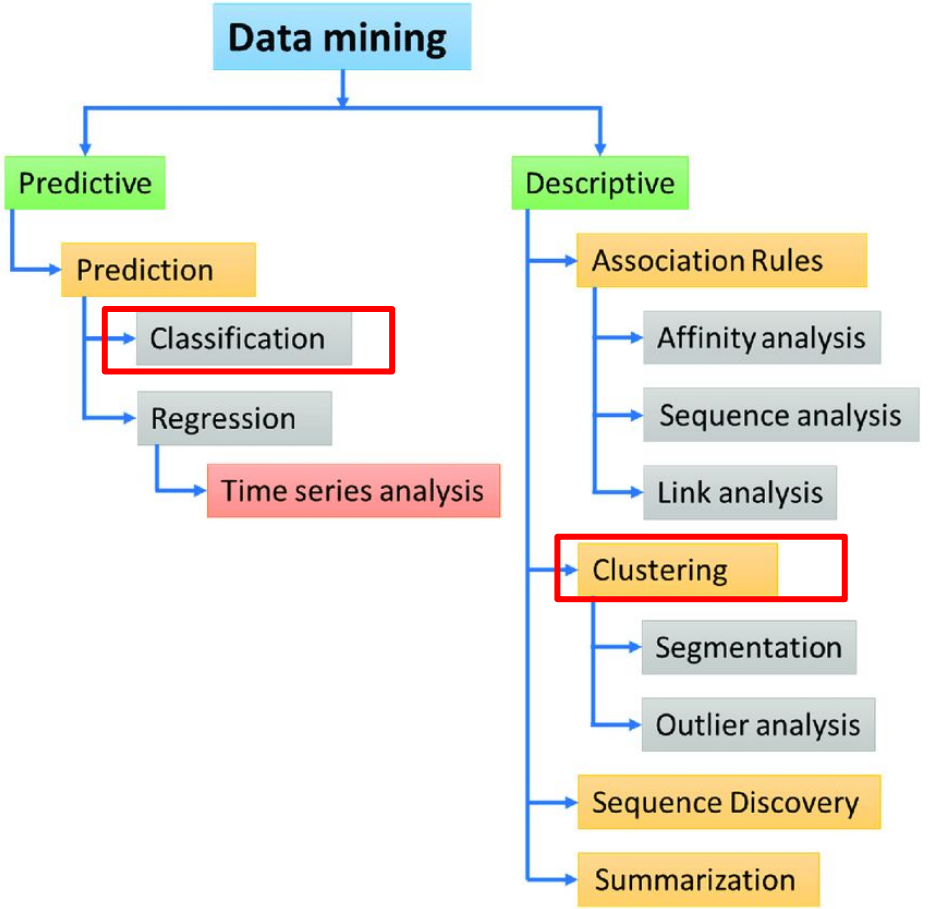
Why Data Mining?

It possesses **multiple tasks** (Zia et al., 2022)



Why Data Mining?

It possesses **multiple tasks** (Zia et al., 2022)



Why Data Mining?

In this course, we will focus on **textual data**.

Data Mining on textual data ⇒ **Text Mining**

Workflow and logic behind it can apply to other modalities!

Why Data Mining?

In this course, we will focus on **textual data**.

Data Mining on textual data ⇒ **Text Mining**

Workflow and logic behind it can apply to other modalities!

But how to model the language?!

Why is Language Hard to Model?

A Definition of Language

Definition 1: *Language is a means to communicate, it is a semiotic system. By that we simply mean that it is a set of signs. A sign is a pair consisting in [...] a signifier and a signified.*

Definition 2: *A sign consists in a phonological structure, a morphological structure, a syntactic structure and a semantic structure*

Krach 2007

The Six Levels of Linguistics Analysis

Analysis in context

Extra-linguistic context

*Found **him** in the street inside a bag. I think **he** is happy with his new life*

Linguistic context

— *You know what? John gave Peter a Christmas present yesterday*
— *Wow, was he surprised? What was it like?*
— *Surprisingly good. He spent quite a bit on it.*

Semantic level

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$.*DNI*$_{\text{TRIGGER}}$

Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

Sentence-level analysis

Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

Phonological level

International Phonetic Alphabet

[aɪ pʰiː eɪ]

Graphemic level

enough, cough, draught, although, brought, through, thorough, hiccough

cf. Sagot, Algorithms for speech and natural language processing, MVA course Material

The 5 Challenges of NLP

1. Productivity
2. Ambiguous
3. Variability
4. Diversity
5. Sparsity

Productivity

Definition

"property of the language-system which enables native speakers to construct and understand an indefinitely large number of utterances, including utterances that they have never previously encountered." (Lyons, 1977)

➔ **New words, senses, structure** are **introduced in languages all the time**

Examples: *staycation* and *social distance* were added to the Oxford Dictionary in 2021

Ambiguous

Most linguistic observations (speech, text) are open to **several interpretations**

We (Humans) disambiguate - i.e. **find the correct interpretation** - using all kind of signals (linguistic and extra linguistic)

Ambiguity can appear at all levels (phonology, graphemics, morphology, syntax, semantics)

Ambiguous

Syntactic Ambiguity

cf. Sagot

Ambiguous

Semantic Ambiguity

- Polysemy: *e.g.* ***set*** *,* ***arm, head***
  Head of New-Zealand is a woman
- Name Entity: *e.g.* ***Michael Jordan***
  Michael Jordan is a professor at Berkeley
- Object/Color: *e.g.* ***cherry***
  Your cherry coat

Ambiguous

Pragmatic Ambiguity (i.e. needs context)

*Two Soviet ships collide, **one dies***

*Dealers will hear **car talk** at noon*

Ambiguous

Disambiguating can requires Discourse Knowledge

Where can I find **a vegetarian restaurant** in Paris

Here is a list of restaurant in Paris: ….

Give me the top ranked ones, in the 14th arrondissement

Here are the top ranked restaurant in the 14th arrondissement in Paris

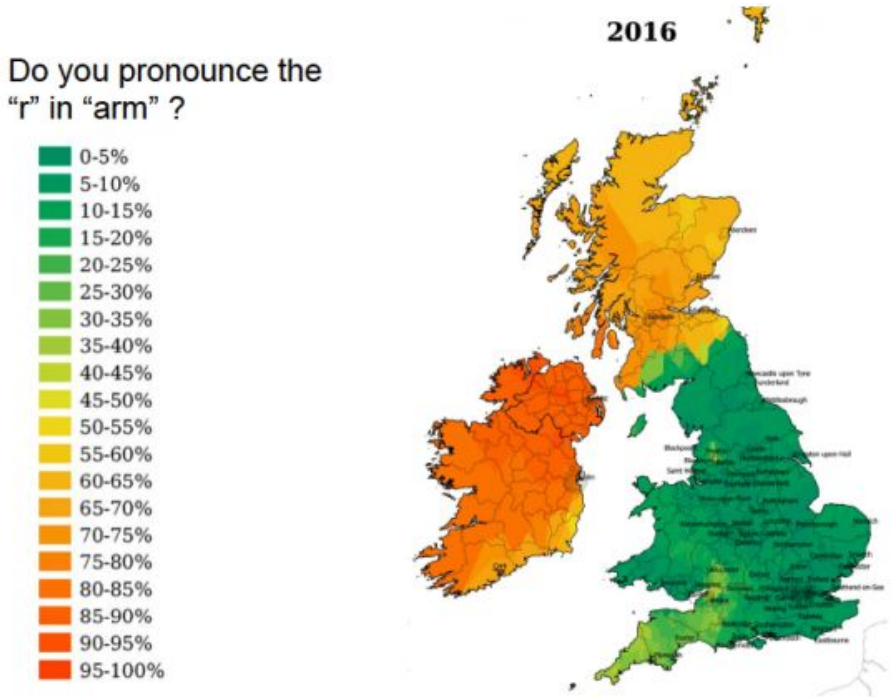
How far is the closest **one** from my current location?

Variation

Language Varies at all levels

- Phonetic (accent)
- Morphological, Lexical (spelling)
- Syntactic
- Semantic

Phonetic Variation



cf. Sagot

Spelling and Syntactic Variation

T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> tabuz, lavé rien fé
> ki ca ? le mec ou son chien ?
> Wtf is wrong with him ? #PETA4EVER
> ki ca ? le chien ?
> looool

BING translation:
You saw coming it #AperoChezRicard wsh
> +10000, shah!
> tabuz, washed anything fe
> Ki ca? the guy or his dog?
> WTF is wrong with him? #PETA4EVER
> Ki ca? the dog?
> looool

Variation Determiners

- Who is talking?
- To Whom?
- Where? *Work, Home, Restaurant*
- When? *19th century, 2008, 2022…*
- About what? *Specialised domain, the Weather,…*

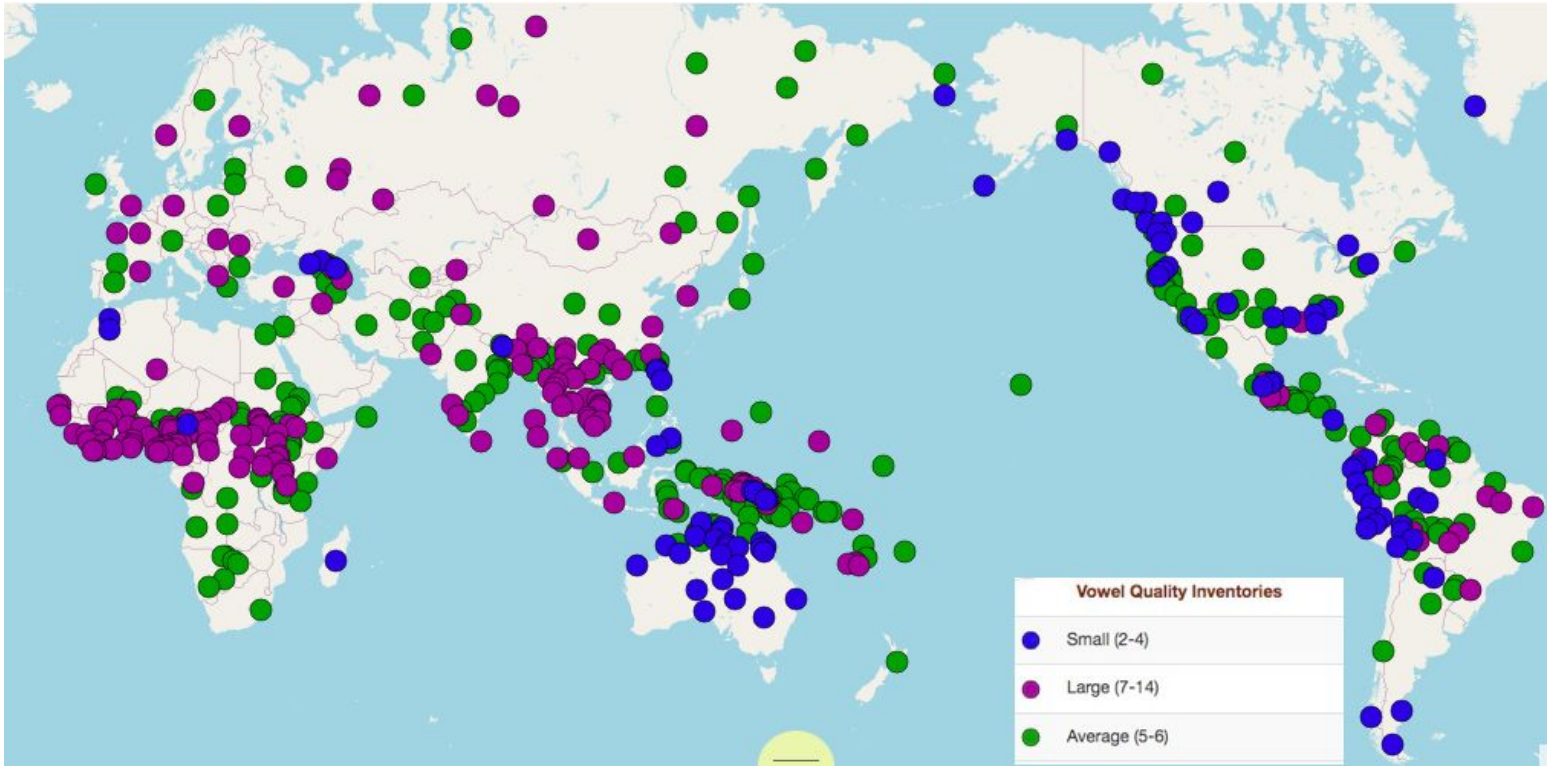
Essentially, the Variability of a language depends on:

- Social Context
- Geography
- Sociology
- Date
- Topic

Diversity

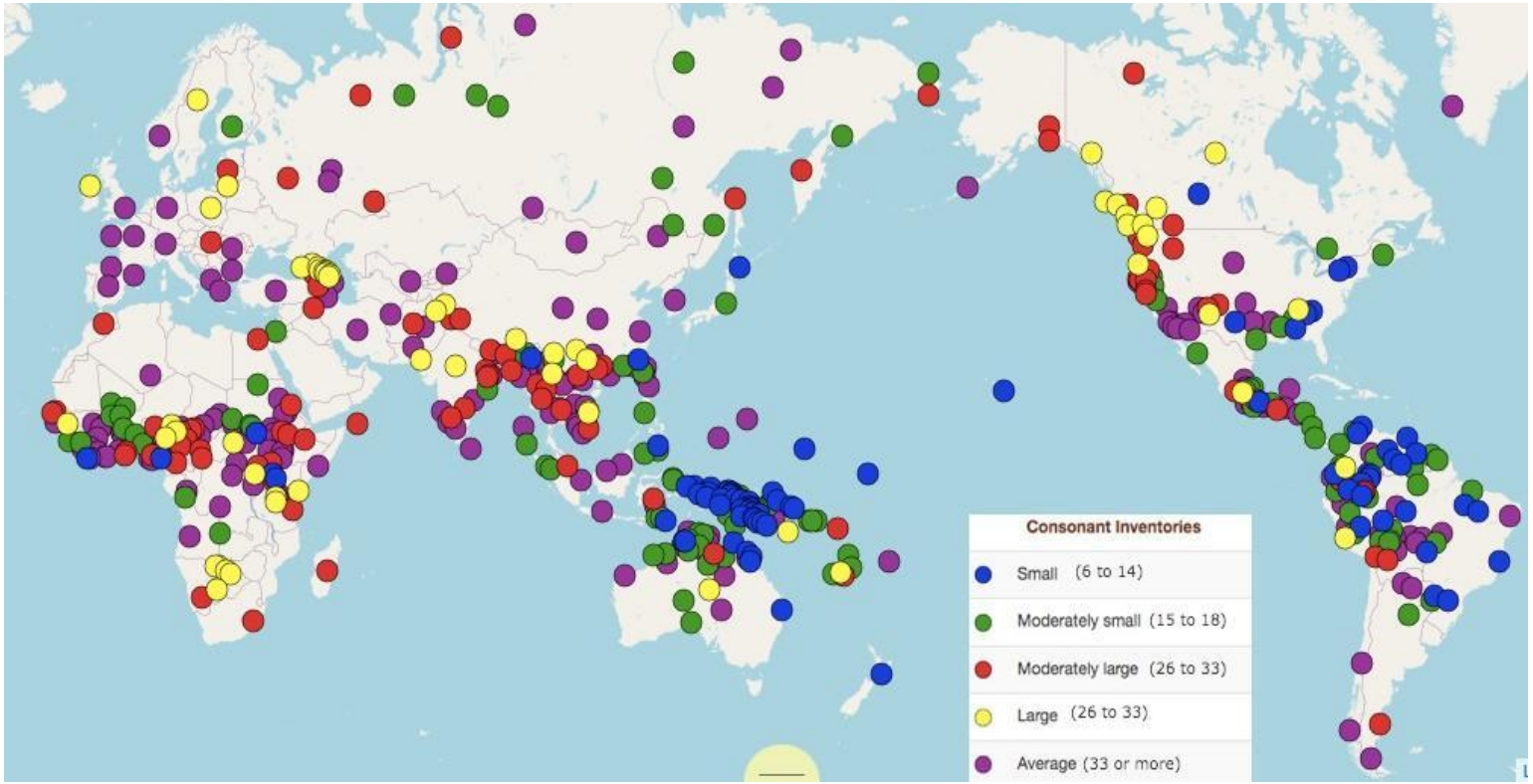
- About **7000 languages** spoken in the world
- About **60%** are found in the **written form** (cf. Omniglot)

Phonologic Diversity



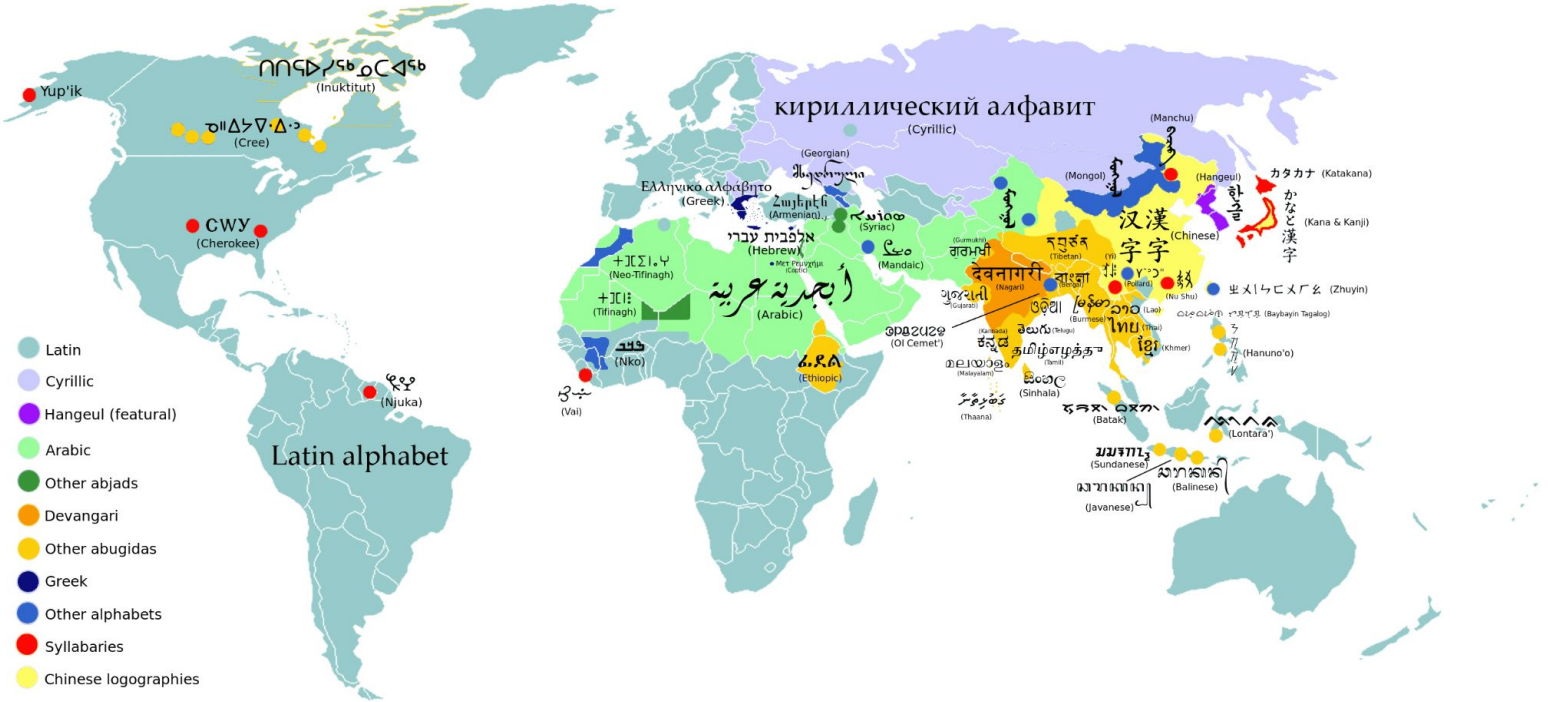
(Dyer et. al 2013)

Phonologic Diversity



(Dyer et. al 2013)

Graphemic Diversity



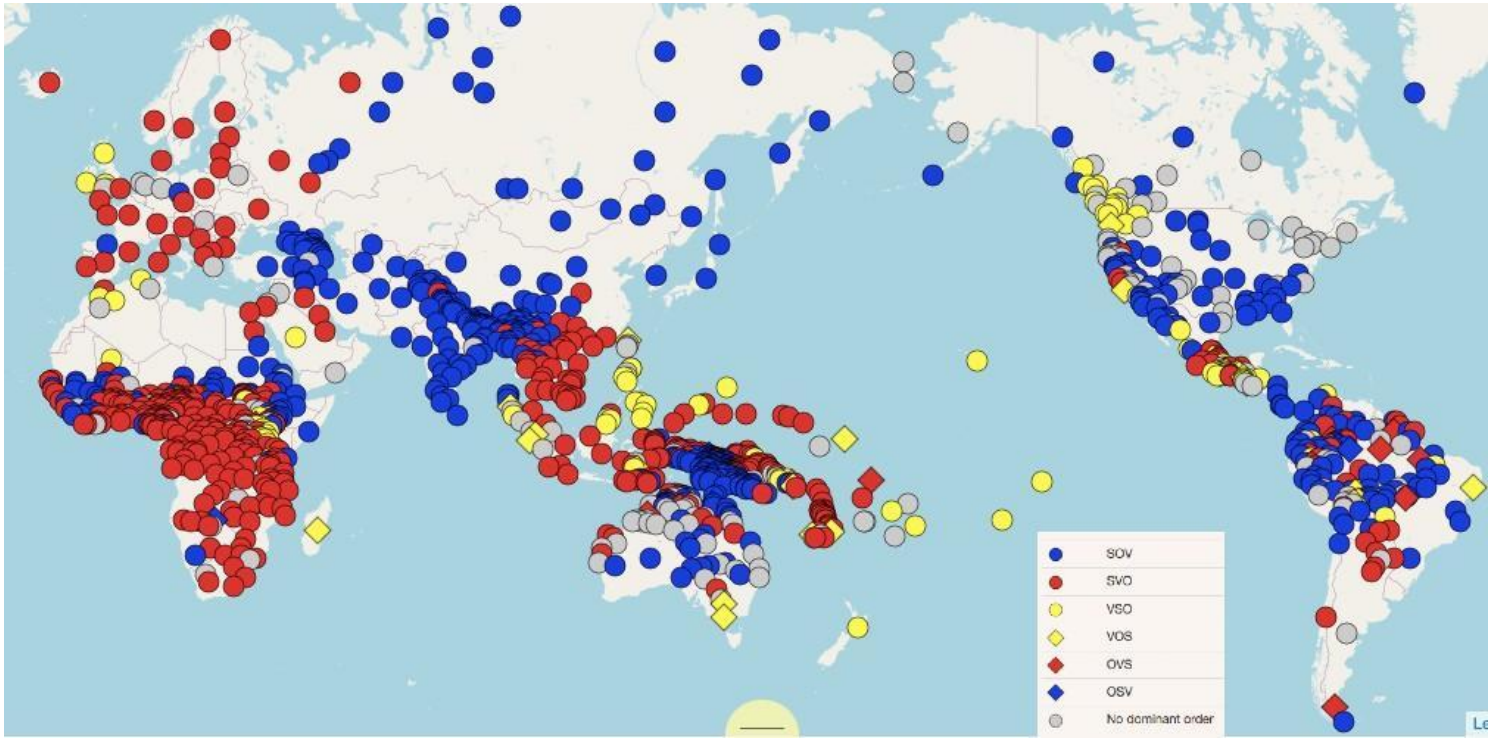
wikipedia

Syntactic Diversity

A key characteristics of the syntax of a given language is **the word order**

- **Word order differs** across languages
- Word order degree of freedom also differs across languages
- We characterize word orders with: **Subject (S) Verb (V) Object (O) order**

Syntactic Diversity



(Dyer et. al 2013)

Word Order Freedom And Morphology

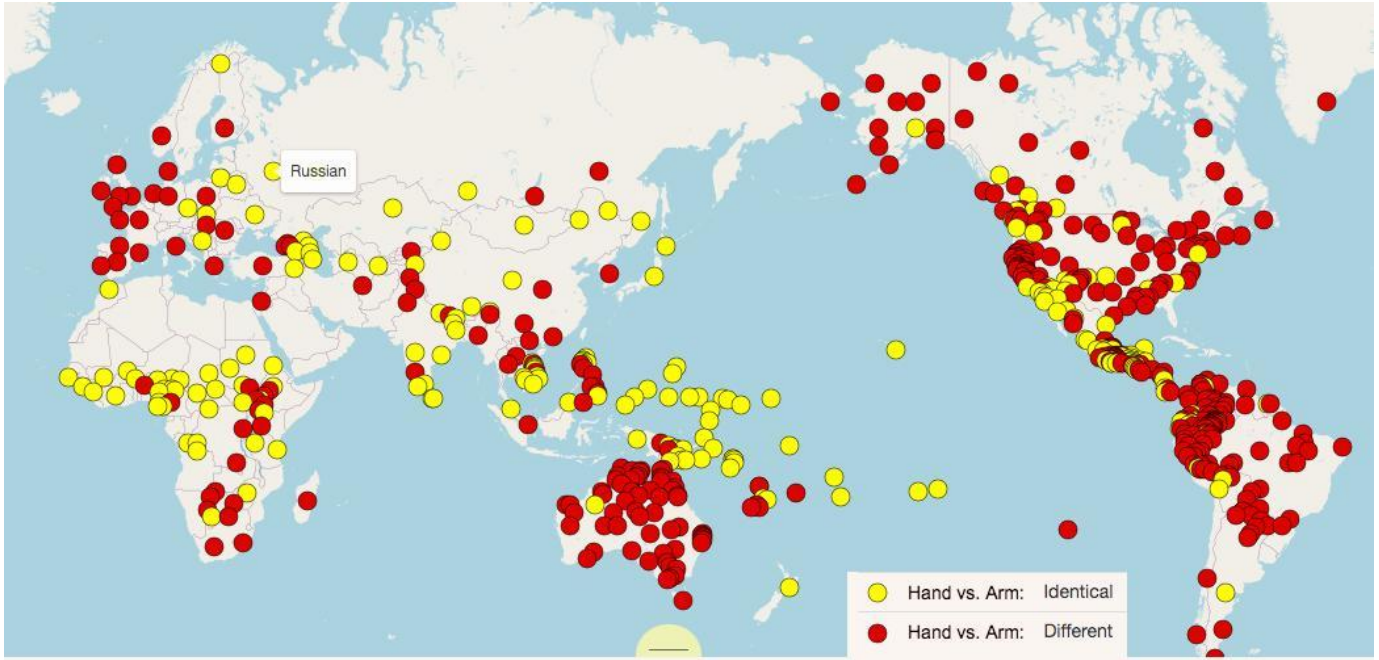
- Word orders freedom and morphology are usually related
- **The more freedom in word orders**
  → the less information is conveyed by word positions
  → the more information is carried by each word
  → **the richer the morphology**

English *cats eat mice*

Russian (O: -ей) *Кошки едят мышей*
Мышей едят кошки
Едят кошки мышей.
Едят мышей кошки.

Semantic Diversity

- Words partition the semantic space
- This partition is very diverse across language



(Dyer et. al 2013)

Statistical Description of a Corpus

Statistical Description of a Corpus

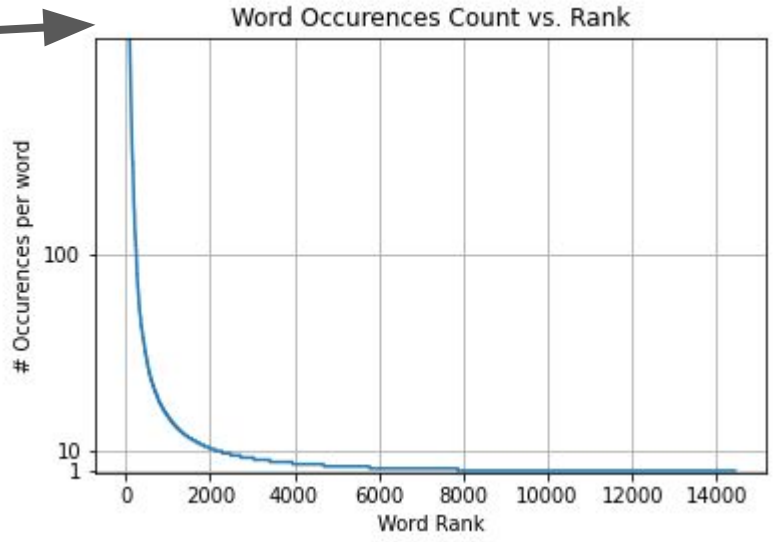
We describe statistically a corpus of 800 scientific articles

Question: If we plot the number of occurrences of each word vs. the rank, what will we observe?

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

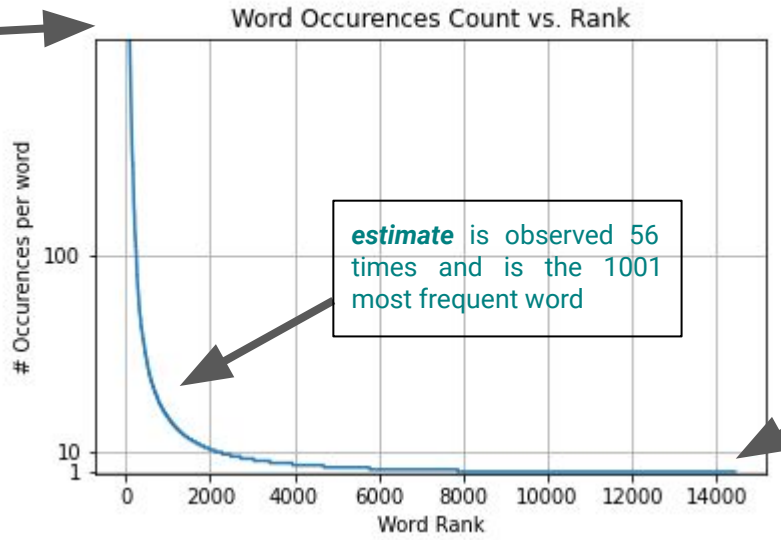
the is the most observed (rank 1) word with 8119 occurrences



Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

the is the most observed (rank 1) word with 8119 occurrences



estimate is observed 56 times and is the 1001 most frequent word

About 6000 Words are observed only 1 time in the dataset (e.g. ***stakeholders, pending, score…)***

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, word distributions follows ***a Zipf Law ie:***

f_w frequency of entity w
k frequency rank of entity w

$$f_w(k) \ \alpha \ \frac{1}{k^\theta}$$

Statistical Description of a Corpus

We describe statistically a corpus of 800 scientific articles

➔ In a large enough corpus, word distributions follows ***a Zipf Law ie:***

f_w frequency of entity w
k frequency rank of entity w

$$f_w(k) \ \alpha \ \frac{1}{k^\theta}$$

- Zipf law is a Power relation between the rank and frequency
  *The most frequent entities are **much more frequent** than the less frequent ones*

- Under a Zipf law, log(fw) and log(k) are linearly related

Statistical Description of Language

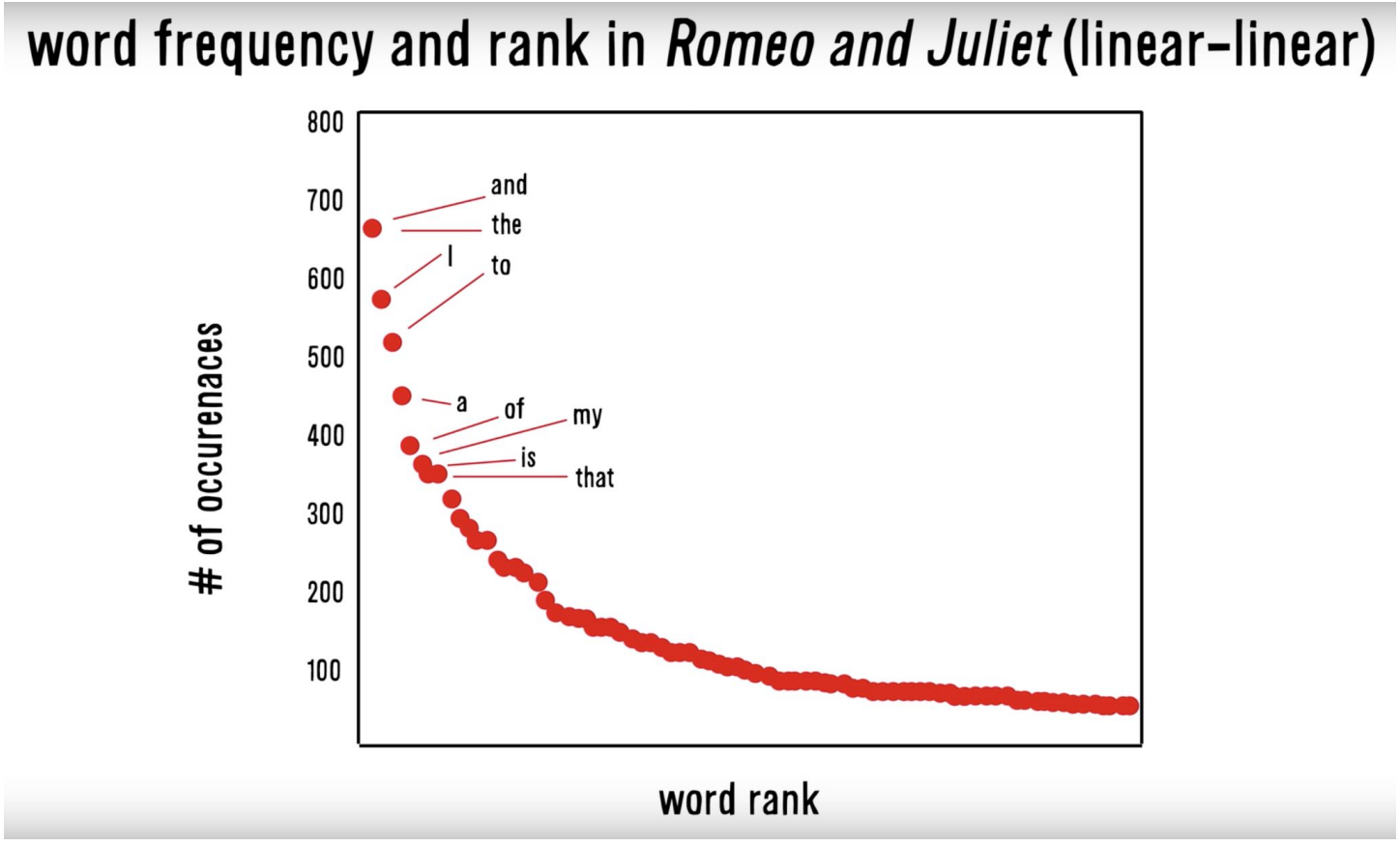
Zipf Distributions are observed not only for words but with many other units of language (sounds, syntactic structure, name entities…)

Consequence

➡ A large number of units are observed in language with very low frequency i.e. **Sparsity**

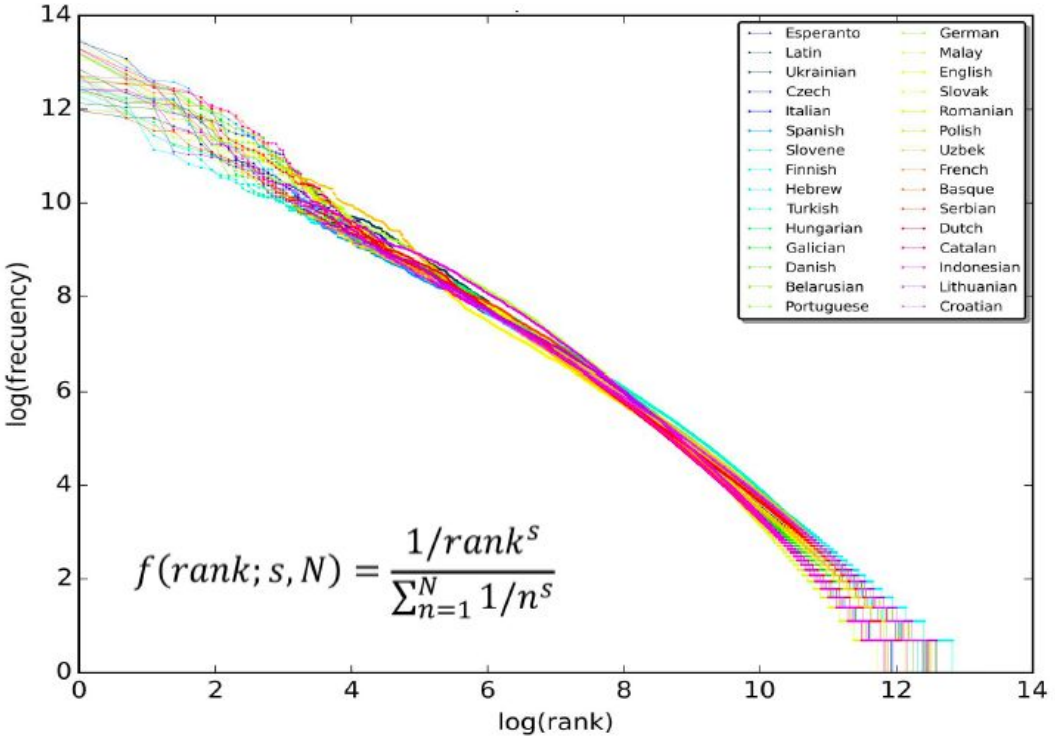
➡ Very challenging for NLP

Statistical Description: Word Frequency



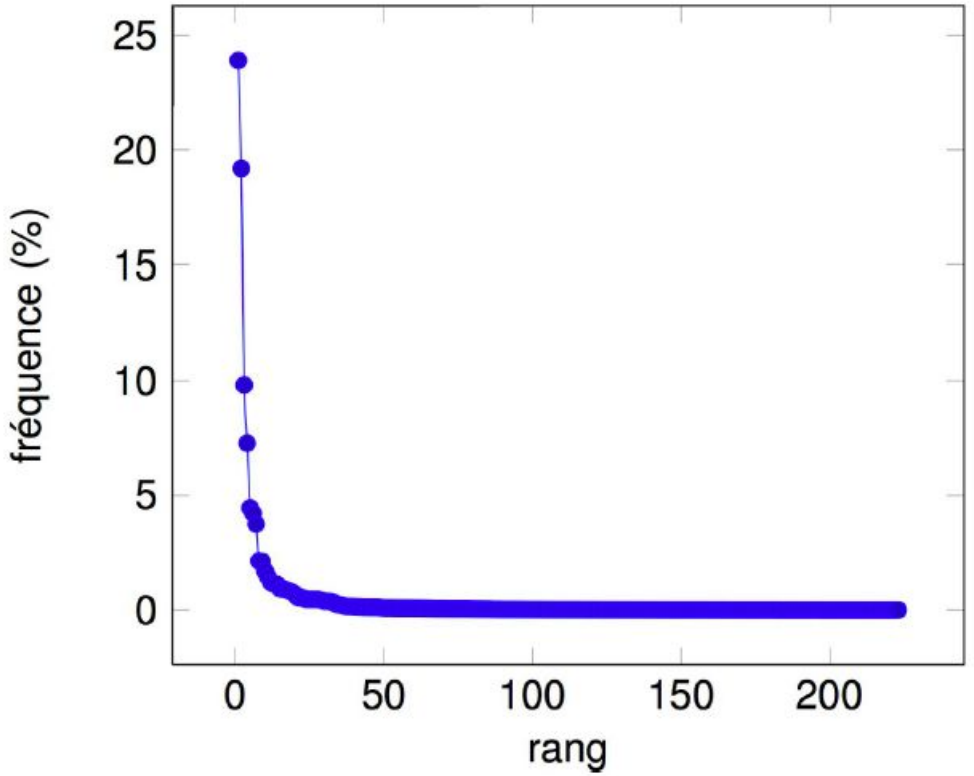
Statistical Description: Lexicon

Top 10M words in wikipedia



Statistical Description: Syntax

Automatically parsed corpus



How to Represent Data?

What is Natural Language Processing?

In a nutshell, NLP consists in handling the complexities of natural languages "to do something"

- Raw Text / Speech → Structured Information
- Raw Text / Speech → (Controlled) Text/Speech

In this course we will focus **on textual data**

Framework

We assume:

- A **token** is the basic unit of discrete data, defined to be an item from a vocabulary indexed by 1, ..., V.
- A **document** is a sequence of N words denoted by **d = (w1,w2, ...,wN)**, where wn is the N-th word in the sequence.
- A **corpus** is a collection of M documents denoted by D = (d1, d2, ..., dM)

Example: *Wikipedia, All the articles of the NYT in 2021…*

Token

With regard to our end task, a token can be:

- A **word**
- A **sub-word**: *e.g. a sequence of 3 characters*
- A **character**
- An **sequence of characters** (sometimes a word, sometimes several words, sometimes a sub-word…)

Document

A Document can be:

- A **Sentence**
- A **Paragraph**
- A **sequence of character**s

Text Segmentation

Definition: Text Segmentation is the process of splitting raw text (i.e. list of characters) into **units of interest**.

Two level of segmentation (usually) required :

- Split raw text into **modeling units** (ex: sentence, paragraph, 1000 characters, web-page...)
- Split modeling units into sequence of **basic units** (referred as tokens) (e.g: words, word-pieces, characters, ...)

Two distinct approaches:

- **Linguistically informed** e.g. word, sentence segmentation...
- **Statistically informed** e.g. frequent sub-words (word pieces, sentence pieces...)

Tokenization

Definition: Tokenization consists in *segmenting* raw textual data into tokens:

Tokenization

Definition: Tokenization consists in *segmenting* raw textual data into tokens:

Can be framed as a character level task

input: *une industrie métallurgique existait.*
output: IIIEIIIIIIIIIIIIIEIIIIIIIIIIIIIIIIIEIIIIIIIIIIEE

- **Easy task** for most languages and domains
- Can be **very complex in some cases** (Chinese, Social Media...)

Part-of-Speech Tags

POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

Part-of-Speech Tags

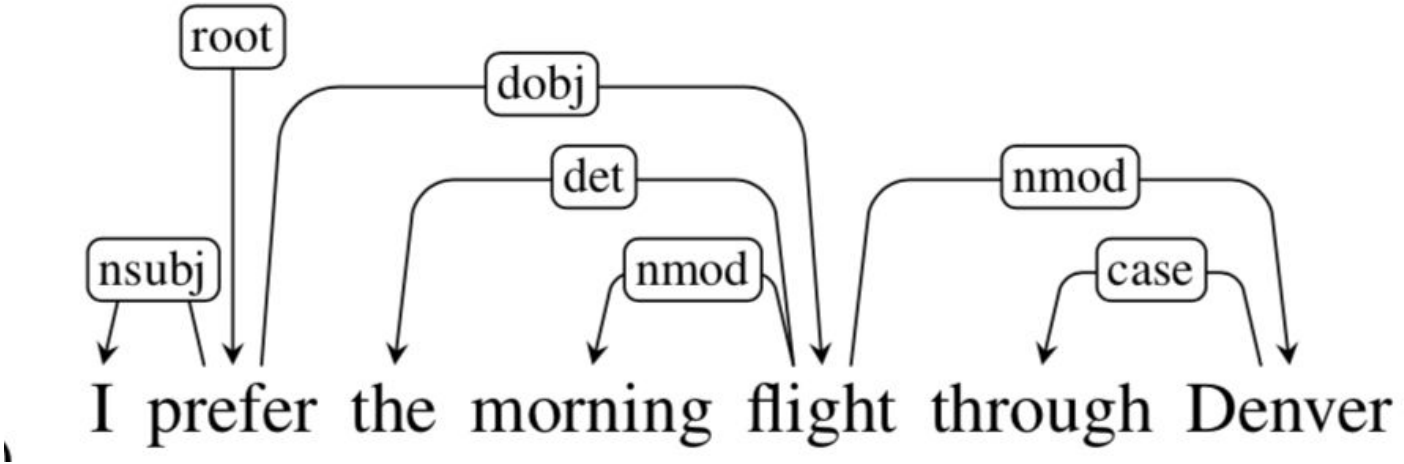
POS Tagging: Find the **grammatical category** of each word

[My , name, is, Bob, and, I, live, in, NY, !]

[PRON , NOUN, VERB, NOUN, CC, PRON, VERB, PREP, NOUN, PUNCT]

Dependency Trees

Syntactic Parsing consists in **extracting the syntactic structure** of a sentence. For instance, **Dependency Parsing** (here) predicts an acyclic directed graph (a **tree**)



Slot Filling / Intent Detection

Intent Detection is a sequence classification task that consists in **classifying the intent of a user** in a pre-defined category.

Slot-Filling is a sequence labelling task that consists in identifying **specific parameters in a user request**.

Can you please play Hello from Adele ?

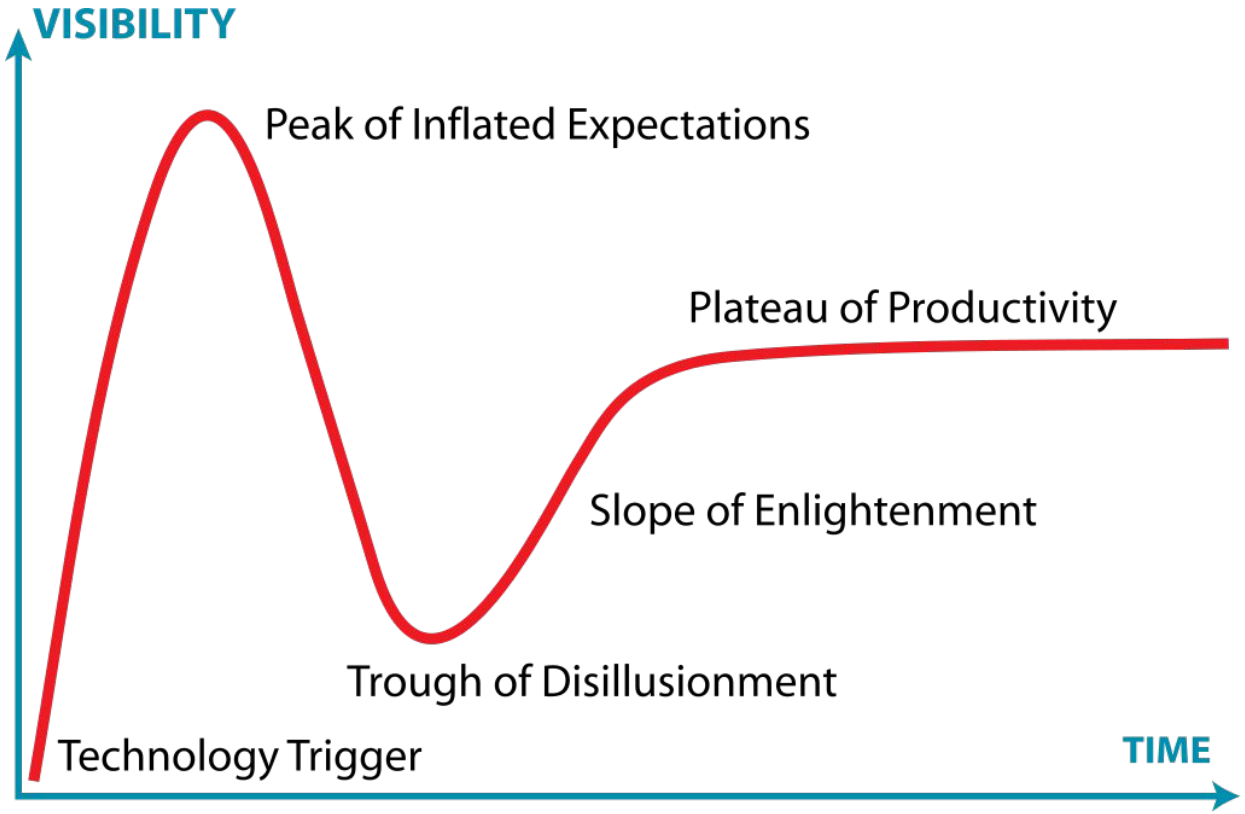
Intent: ***play_music***
Slots: [*Can, you, please, play, Hello, from, Adele, ?]*
*[O , O , O , O , **SONG**, O , ARTIST, O]*

NLP Data Representation Frameworks

(Zia et al., 2022)

	Natural Language Toolkit	SpaCy	Scikit-Learn NLP Toolkit	Gensim
What is it?	open-source python platform for handling human language data	open-source python library for advanced natural language processing	machine learning software library for the Python programming language	fastest python library for the training of vector embedding
Features			• Based on NumPy, SciPy, and Matplotlib • An easy and efficient way to analyze predictive data • Easily accessible and reusable in different contexts	
Advantage	• Most well-known and comprehensive NLP libraries with many extensions • offers support in the largest number of languages	• easy to use • fully integrated with Python • compatible with other deep learning frameworks • many already trained statistical models available • applicable to many different languages • high speed and performance • freely available • able to process long texts • platform-independent usable	• simple and efficient tools for machine learning, data mining, and data analysis • freely available for everyone • applicable to different application areas, like natural language processing	• Provides ready-to-use models and corpora • Models pre-trained for specific areas such as health care • Processes large amounts of data using streaming data
NLP Tasks	• Classification • Tokenization • Stemming • Tagging • Parsing	• Classification • Tokenization • Stemming • Tagging • Parsing • Named Entity recognition • Sentiment Analysis	• Classification • Topic Modeling • Sentiment Analysis	• Text similarity • Text summarization • Topic Modeling
GitHub stars	10.4 k	22.4 k	49 k	12.9 k
Website	nltk.org (accessed on 16 March 2022)	spacy.io (accessed on 16 March 2022)	scikit-learn.org (accessed on 16 March 2022)	radimrehurek.com/gensim/ (accessed on 16 March 2022)
Reference	Bird et al. [59]	Honnibal [60]	Pedregosa et al. [61], Pinto et al. [62]	Rehurek and Sojka [63]

NLP in a Nutshell



To Be Seen in Lab

- Basic Description of Corpus
- Zipf Law Check
- Simple Topic Modelling

Bibliography and Acknowledgment

Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial Intelligence-Based Medical Data Mining. *Journal of Personalized Medicine*, *12*(9), 1359.
[Benoit Sagot 2022], Algorithms for speech and natural language processing, MVA course Material
[Warren Weaver, 1949] Memorandum on Translation
[Weizenbaum, 1966] Eliza
[Dryer, Matthew S. & Haspelmath, Martin (eds.) The WALS]