NOVA IMS Information Management School

Text Mining

Introduction to Text Mining

Academic Year 2024-2025

Bruno Jardim
bjardim@novaims.unl.pt

Rita Oliveira
roliveira@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Lecture Plan

1. Introduction to Text Mining

2. Natural Language Processing (NLP) Pipeline

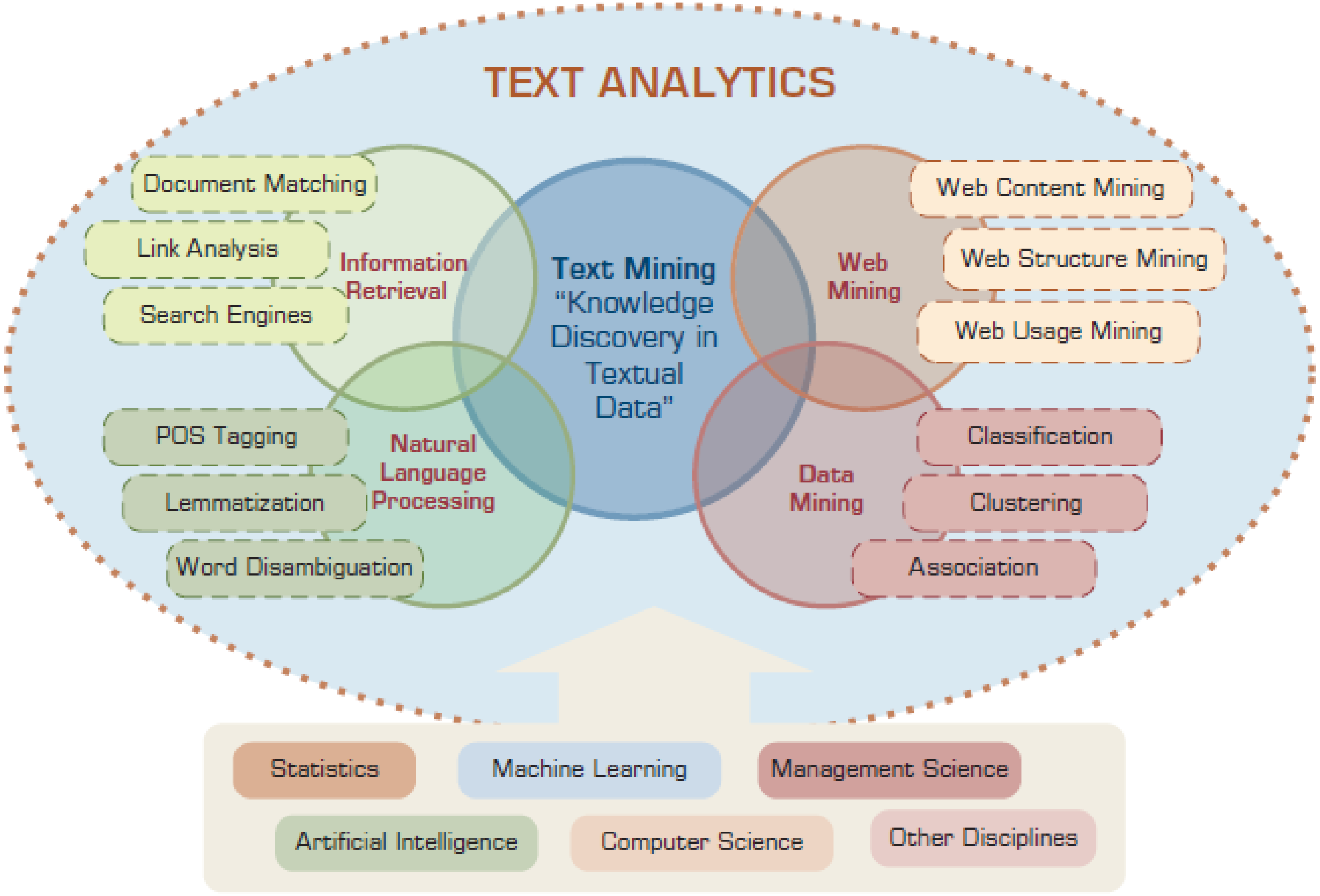
3. Text Preprocessing

NOVA IMS Information Management School

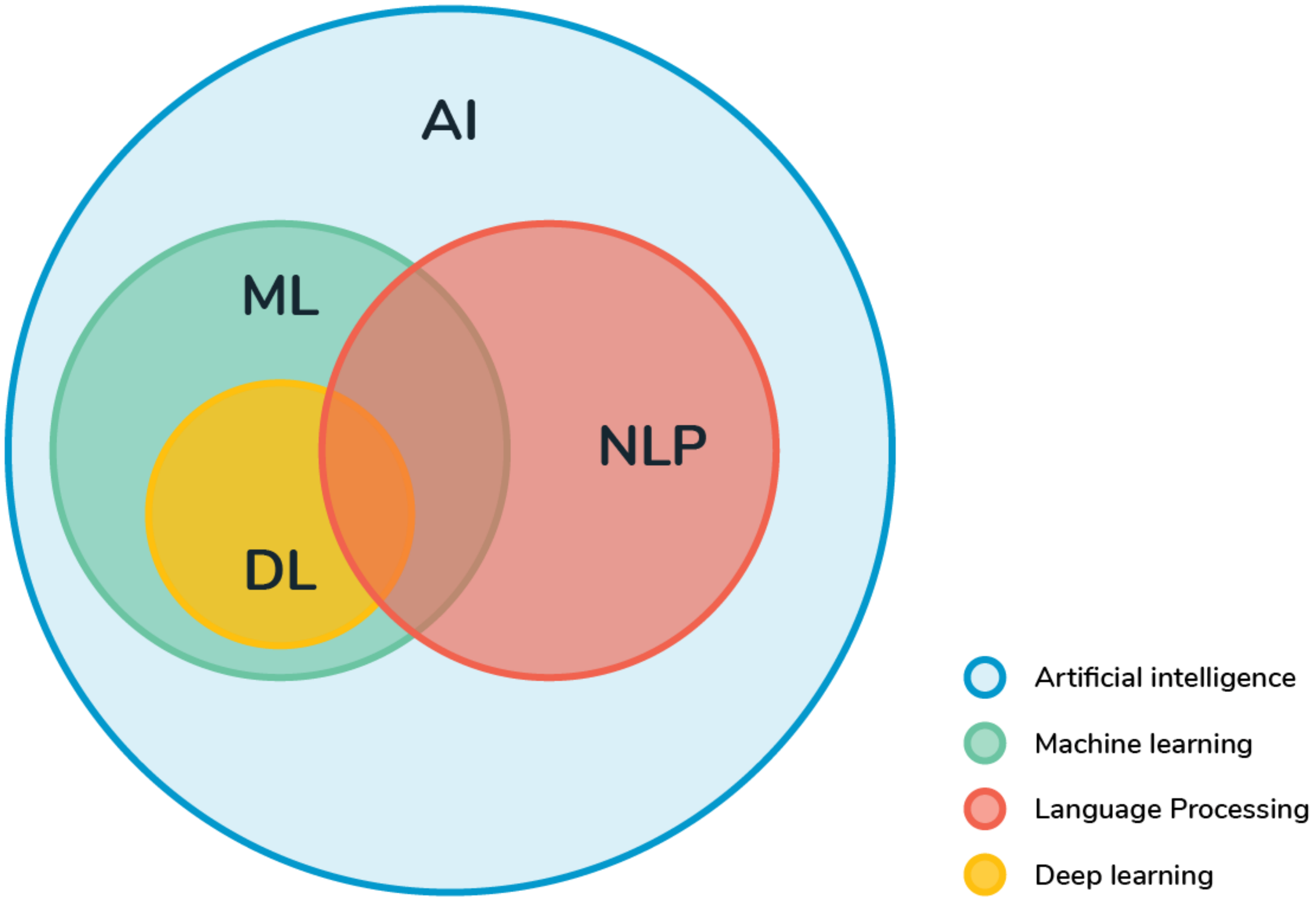
1. Introduction to Text Mining

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Introduction to Text Mining



Introduction to Text Mining

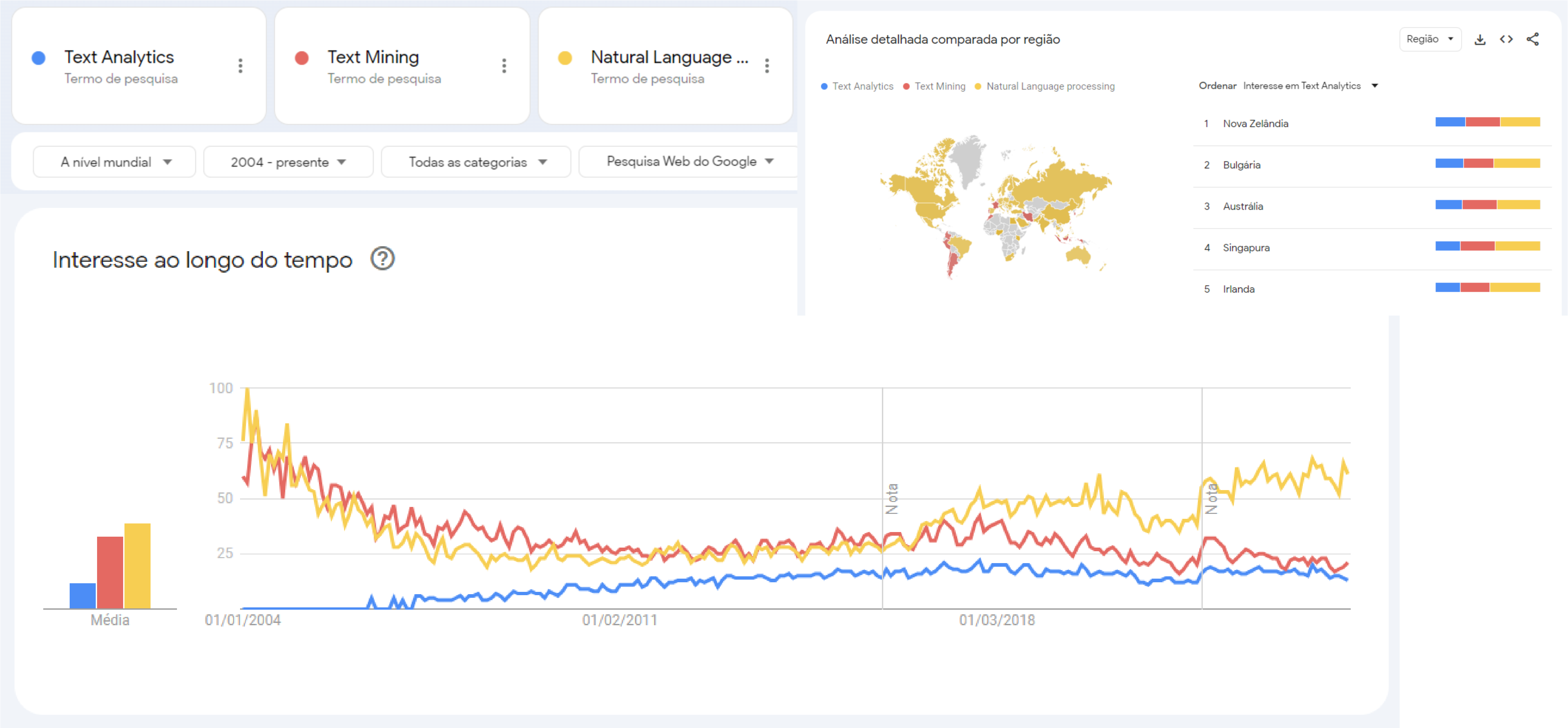


Introduction to Text Mining

Text mining (also referred to as ***text analytics***) is an **artificial intelligence (AI) technology that uses natural language processing (NLP)** to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

Natural language processing (NLP) is an area of computer science and artificial intelligence that is concerned with the interaction between computers and humans in natural language. The **ultimate goal of NLP is to enable computers to understand language as well as we do**.

Introduction to Text Mining



Applications

1. Machine translation
2. Dialog systems (chatbots)
3. Search Engines
4. Predictive Keyboards and auto correctors

…

Everything that deals with text!!!

Applications

Natural Language (NL):

1. Grammatical system, with its **own rules**, used by people to communicate
2. Natural **evolution** due to people communication
   - ✓ New words everyday: “bue”, “fixe”, “ups”, “lol”

Examples:

1. English

2. Portuguese

…

Python??

Challenges of NL - Variability

Definition:

Different **sentences can have the same meaning**; thus, we can say the same thing in different ways. Two sentences with the same meaning are called **paraphrases**.

Examples:

"The president greets the press in Lisbon"

VS

"Marcelo speaks to the media in Campolide"

"He has tons of stuff to throw away"

VS

"He needs to get rid of a lot of junk."

Challenges of NL - Ambiguity

Definition:

A **single sentence can have different meanings**

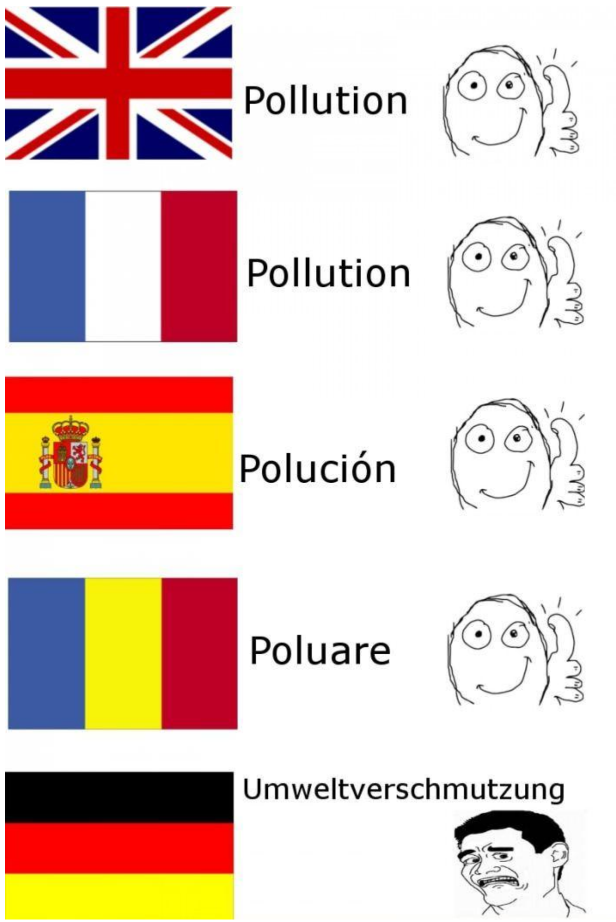
Dealing with it:

The only way to deal with ambiguity is through context!

Challenges of NL - Generalization

Often an **NLP system is trained with a corpus** in a specific domain, but it is **used in a different domain**.

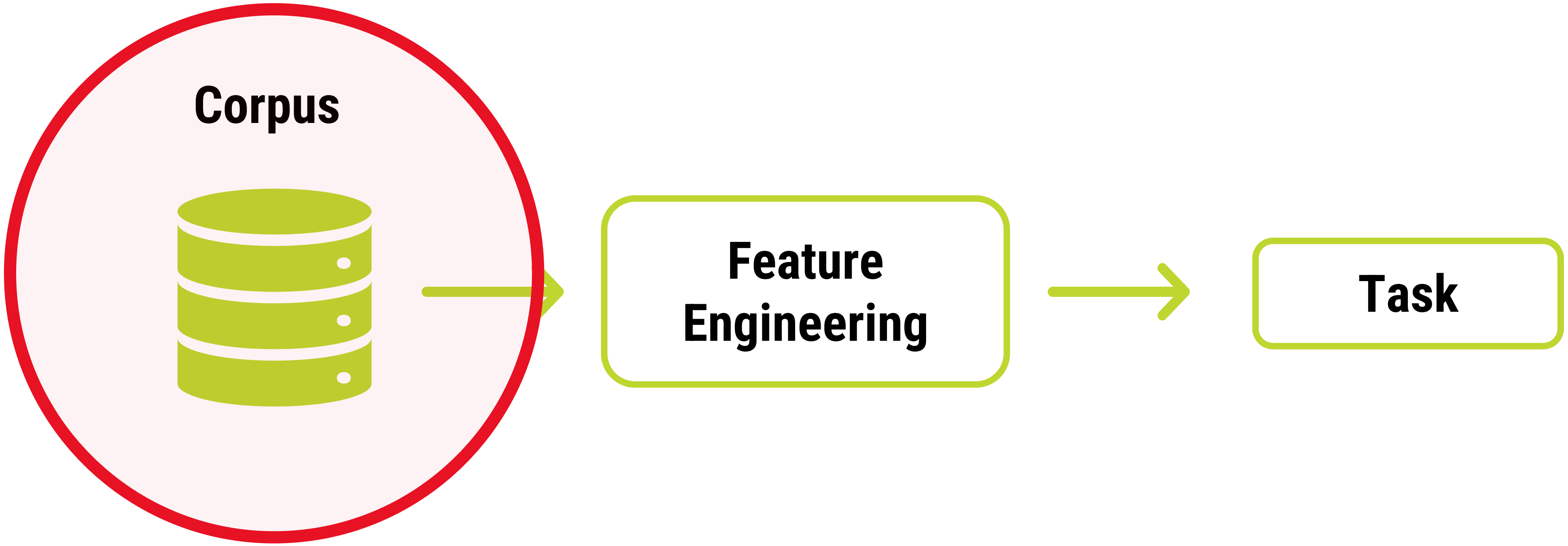
Thus, the system is confronted with input that he has never seen before, either because those inputs are Out-of-Domain (OOD) or because they contain words Out-of-Vocabulary (OOV).



2. Natural Language Processing

NLP Pipeline

NLP Pipeline



Corpus

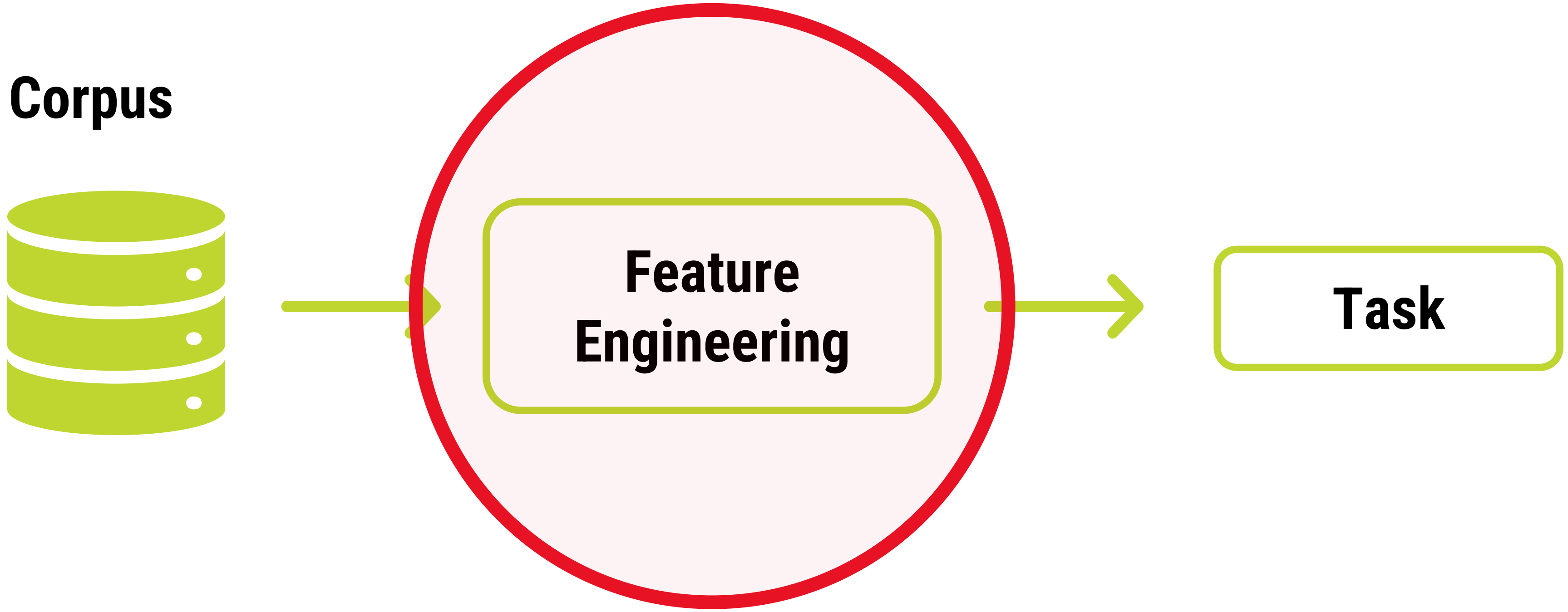
- A **Corpus** is a collection of **text** organized into datasets.
- A corpus can be made up of everything from **news**, **recipes**, **wikipedia pages**, **academic papers**, to **novels**, **movie** and **tv scripts**, and **social media posts** like **tweets**.
- A collection of Corpus is called **Corpora**.

Corpus

- The first step is to split a corpus into **Train/Validation/Test** or via **K-fold cross validation** (other methods can be considered).

- For **Train/Validation/Test**
  - 80%/10%/10% spill for small corpus (<10k samples)
  - More train percentage for bigger corpus.
  - Keep the original always!
  - Your split must be reproducible

Feature Engineering



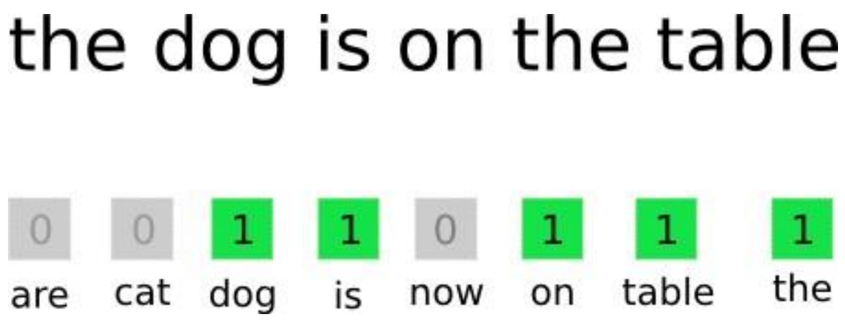
Feature Engineering

How can we represent text?

Feature Engineering

Bag-of-Words

- Each word is a feature.
- Our **feature space is defined by our vocabulary**
- Documents/pieces of text will be represented as **sparse vectors**.



Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

Text 2 – “ I live in France”

	France	I	in	live	love	Paris
Text 1	0	1	0	0	1	1
Text 2	1	1	1	1	0	0

Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

Text 2 – “I live in France”

Feature Engineering

Bag-of-Words Example

Text 1 – "I love Paris"

Text 2 – I live in France"

	France	I	in	live	love	Paris
Text 1						
Text 2						

Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

Text 2 – “I live in **France**”

	France	I	in	live	love	Paris
Text 1	0					
Text 2	1					

Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

Text 2 – “I live in France”

	France	I	in	live	love	Paris
Text 1	0	1				
Text 2	1	1				

Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

Text 2 – “I live **in** France”

	France	I	in	live	love	Paris
Text 1	0	1	0			
Text 2	1	1	1			

Feature Engineering

Bag-of-Words Example

Text 1 – “I love Paris”

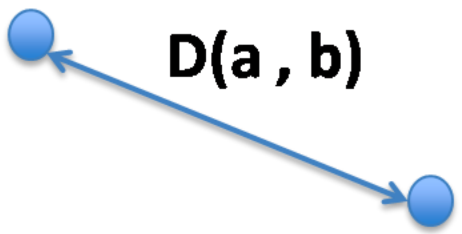
Text 2 – “I live in France”

	France	I	in	live	love	Paris
Text 1	0	1	0	0	1	1
Text 2	1	1	1	1	0	0

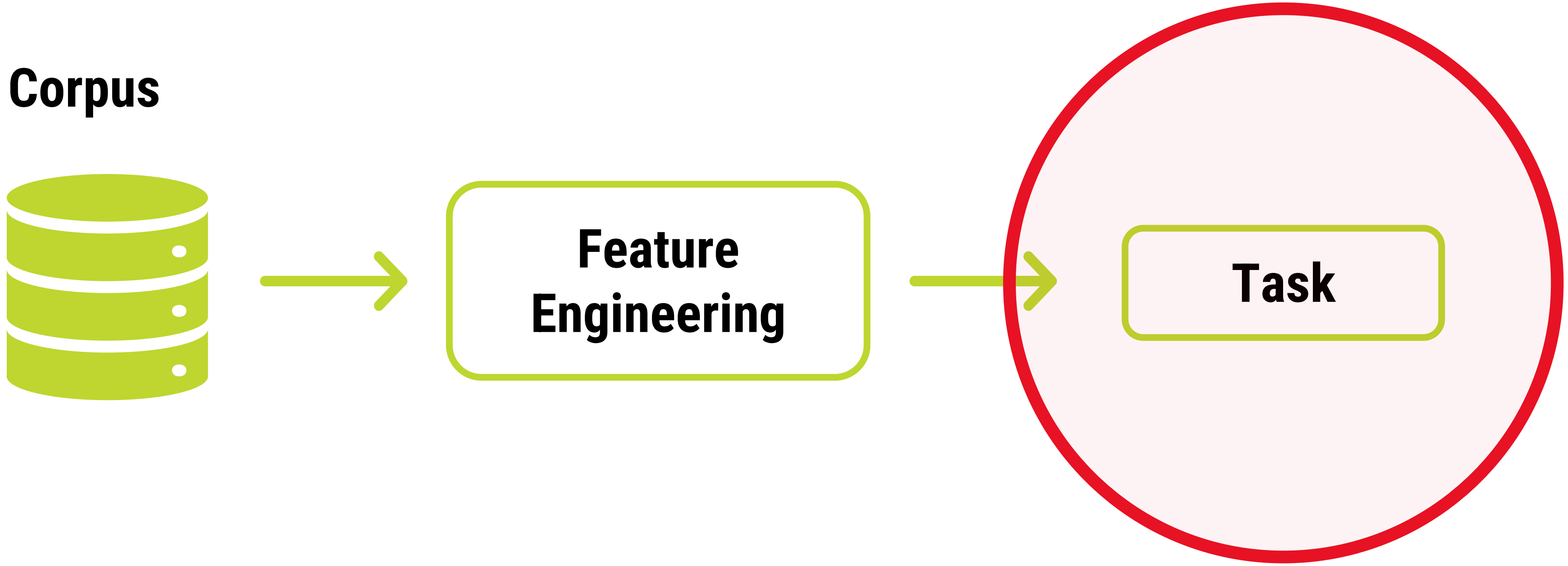
Feature Engineering

- Given 2 documents **we can transform them into sparse vectors** and **compare them with simple distance metrics**.
- If they share a lot of words, they will be close to each other.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$



NLP Tasks



NLP Tasks

- Keyword Extraction
- Text Similarity
- Document Classification
- Sentiment Analysis
- Topic Modelling
- Information Retrieval
- Q&A
- Text Generation

(…)

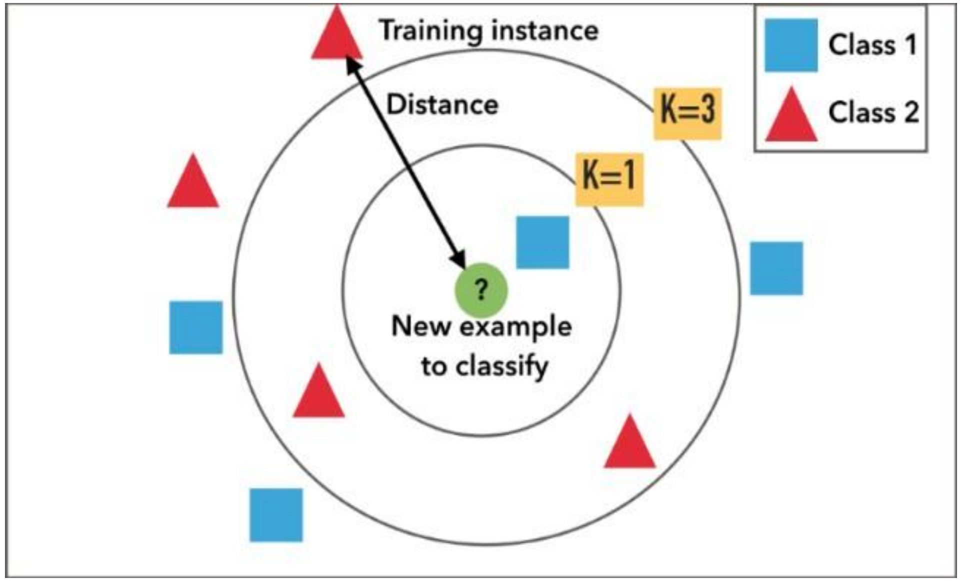
NLP Tasks

K-Nearest-Neighbour: (**our baseline**)

The simplest classification algorithm!

We take our new document **x**, and we represent **x** in the **same feature space of our training documents**.

Then we **compare x with all training documents** and we **give x the label of the closest** document.



Example of an applied case (cont.)

{you have the original academic article in https://doi.org/10.1016/j.cstp.2022.07.011}

"Predicting illegal parking classes from tickets' description":

- Covers 3-year period (2017-2020)
- 89,126 tickets included in analysis
- 7 classes were considered (crosswalk, sidewalk, conditions access, disabled, reserved, others and unknown).

Example of an applied case (cont.)

Descriptive features of illegal parking occurrences in Lisbon

Feature	Description
Datetime	Date and time of illegal parking occurrence
Description	Details and explanation of the illegality as written by the responsible officer
Latitude	Latitude of illegal parking occurrence
Longitude	Longitude illegal parking occurrence
Address	Address of illegal parking occurrence

Example of an applied case (cont.)

Descriptive features of illegal parking occurrences in Lisbon

Feature	Description
Datetime	Date and time of illegal parking occurrence
Description	Details and explanation of the illegality as written by the responsible officer
Latitude	Latitude of illegal parking occurrence
Longitude	Longitude illegal parking occurrence
Address	Address of illegal parking occurrence

Unstructured and not systematized!

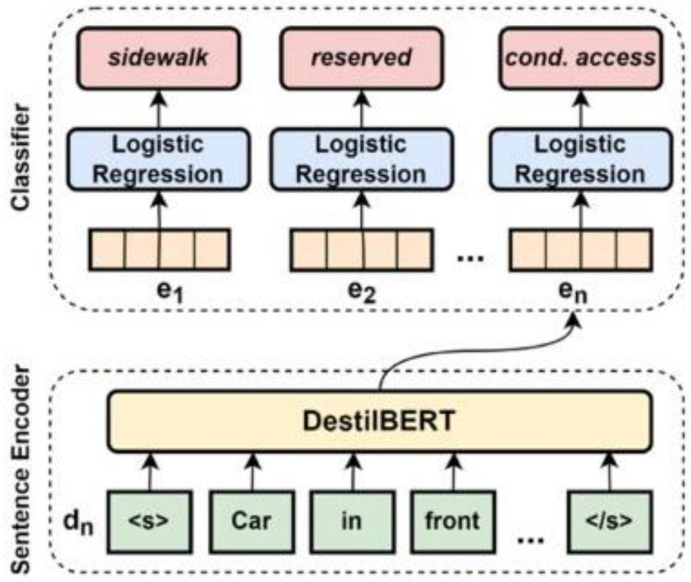
Examples:

- "Parking ticket issued for 'conditioning access' - obstructing driveway entrance."
- "Violation for car parked on sidewalk - hindering pedestrian passage."
- "Ticket given for impeding safe crossing for pedestrians."

Example of an applied case (cont.)

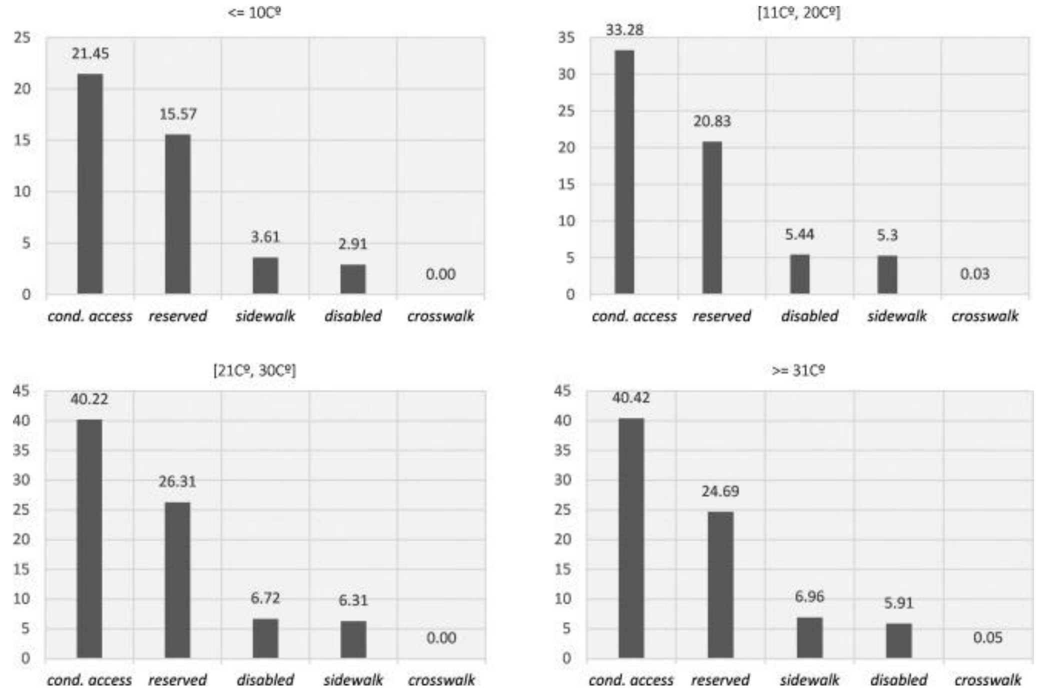
Text Classification model using NLP state-of-the art models

- NLP was used to classify the descriptions into pre-defined classes.
- First, to turn the description text into machine-readable input, it is required to transform it into a vectorial representation (embeddings).
- So, a sentence encoder (DestilBERT - we will learn how to use it later 😉) was used to transform each sentence into a vector.
- A logistic regression receives the input vectors and classifies them accordingly.



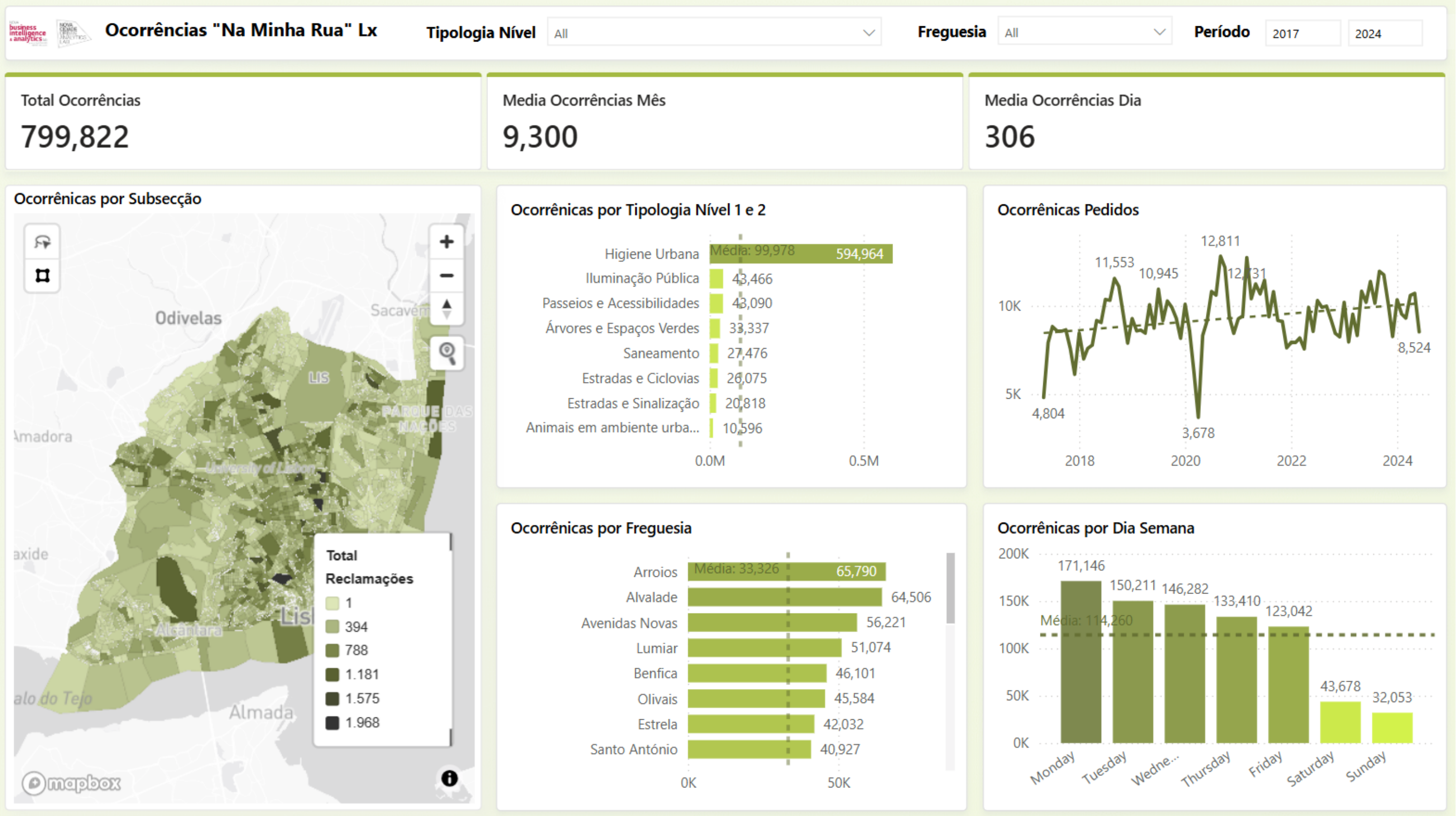
Example of an applied case (cont.)

Allows for downstream analysis of illegal parking classes in different contexts



Example of an applied case

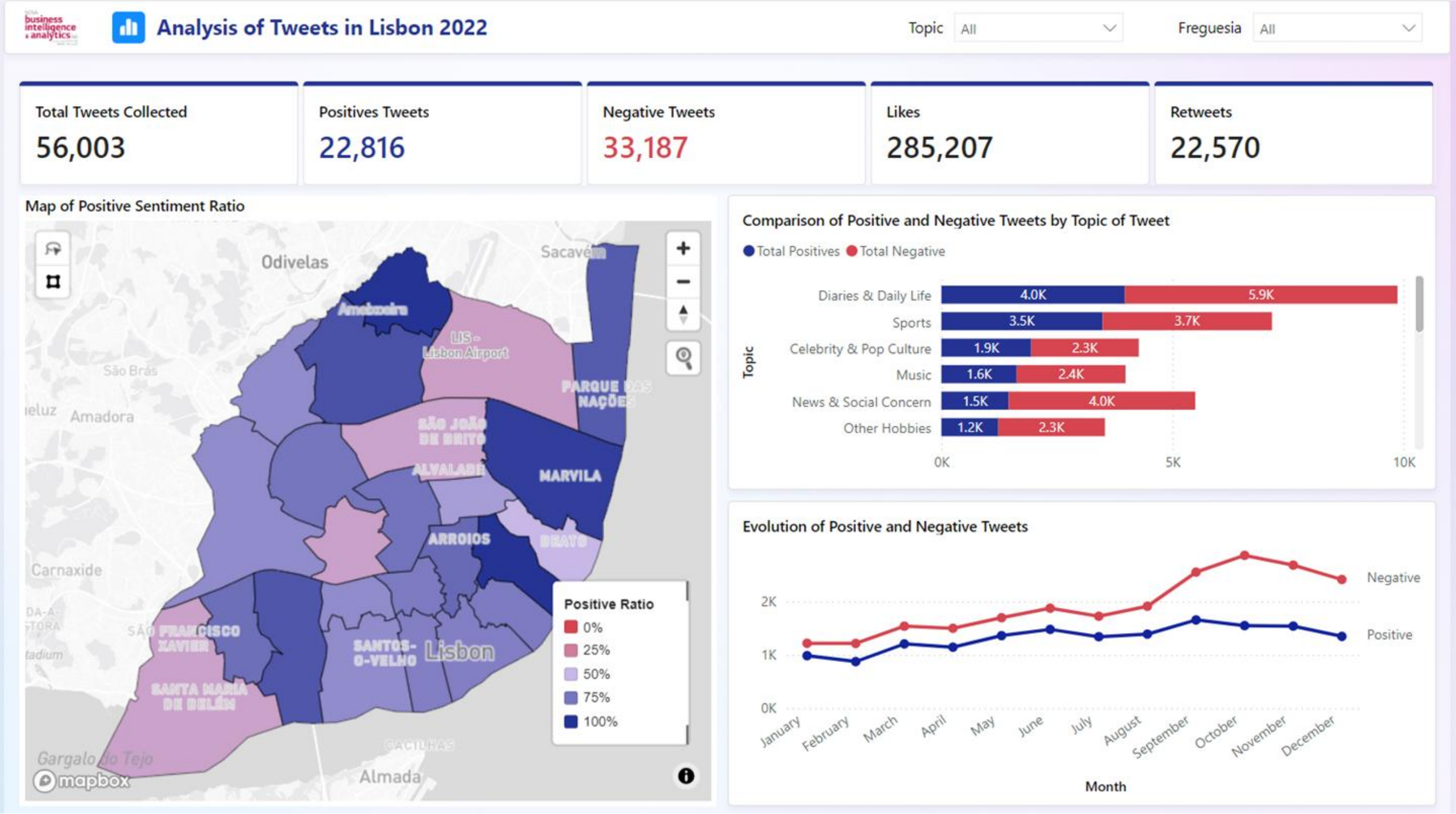
Public Space Occorrences and Claims



https://lnkd.in/dnzeNEQC

Example of an applied case

Analysis of Tweets (X) in Lisbon 2022



https://lnkd.in/dRrfDybH

NOVA IMS Information Management School

3. Text Preprocessing

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Text Preprocessing

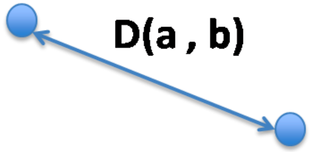
Bag-of-Words

Each word is a feature!

Now is the time for all good man to come to the aid of their party.

The quick brown fox jumped over the lazy dog's back.

$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$



	aid	all	back	brown	come	dog	fox	good	jump	lazy	men	now	over	party	quick	their	Time
Doc1	0	0	1	1	0	1	1	0	1	1	0	0	1	0	1	0	0
Doc2	1	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	1

Tokenization

Tokenization (or word segmentation) is the process of **identifying tokens/features in text**.

This process is **not trivial**:

- United Kingdom ? Is it two words or one?
- Sexta-feira?
- I am vs I'm?

Tokenization

Compounds:

It might be important to treat all compound terms, such as “United Kingdom” as unique features (they share the same semantics)

This does not occur only in names: “soap opera”, “environmentally-friendly”, “fat free”

Assuming a Bag-of-words model:

“This cake is free fat”

vs

“This cake is fat free”

Punctuation

Punctuation can bring several problems to the tokenization.

Typically, we **treat punctuation as individual tokens** but sometimes this can bring problems:

"…" – [. , . , .]
"Dr." – [Dr, .]
"www.google.com" – [www, . , google, . , com]

vs

Hey! – [Hey, !]
I love NLP. – [I, love, NLP, .]

Problems with BoW

1. Curse of Dimensionality

2. No Semantic Relationships

Curse of Dimensionality

It has the **curse of dimensionality** issue as the **total dimension of the BoW is the vocabulary size**. It can easily over-fit your model.

The remedy is to **use some well-known dimensionality reduction** technique to your input data.

Curse of Dimensionality

Bag-of-Words size equals the size of the vocabulary

Text 1 – “I love Paris”

Text 2 – “ I live in France”

	France	I	in	live	love	Paris	(…)	word n-1	word n
Text 1	0	1	0	0	1	1	…	0	0
Text 2	1	1	1	1	0	0	…	0	0

Curse of Dimensionality

Bag of words representation **doesn't consider the semantic relation** between words.

- **Word order is discarded**
- **Similar words are treated as different features** and do not share any information…

E.g.: Paris and France should share some information since, for example, they are both locations

Text Preprocessing

Eliminating Stop Words

Some words are not very informative, and **their presence only represents more rules and/or more processing**. Thus, the first thing we can do is to get rid of them.

Examples: on, the, a, of, from, and, that, it, in (…)

"Who **is the** main actor **of** Pulp Fiction"

- [who, main, actor, Pulp Fiction]

Less tokens same meaning!

"What actor played **the** main role **in** Pulp Fiction"

- [what, main, actor, Pulp Fiction, played, role]

Lowercasing

Converting everything to lowercase also helps reducing the vocabulary size.
Examples:

“Comi sopa e comi muito bem” – [Comi, sopa, comi, muito, bem]

“Comi sopa e comi muito bem” – [sopa, comi, muito, bem]

Sometimes it can go wrong…. Bush (name) != bush (plant)

Regular Expressions

In some situations, you may want to normalize dates, numbers, names, etc.

Examples:
“03-04-18" vs. "03/04/18”
“John Fitzgerald Kennedy" vs. "John F. Kennedy" vs. "John Kennedy ”

Using **regular expressions** we can **convert dates, names and number from one format to another**.

Or, we can **replace them with a default token**. In industry this is an important step for anonymization!

Regular Expressions

Text 1 – “I love Paris”

Text 2 – “ I live in France”

	France	I	in	live	love	Paris
Text 1	0	1	0	0	1	1
Text 2	1	1	1	1	0	0

Regular Expressions

Text 1 – “I love Paris”

Text 2 – “ I live in France”

	France	I	in	live	love	Paris
Text 1	0	1	0	0	1	1
Text 2	1	1	1	1	0	0

$$\mathbf{D(Text\ 1,\ Text\ 2)} = \sqrt{(1-0)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2} = \mathbf{2.24}$$

Regular Expressions

Text 1 – “I love ~~Paris~~” → “I love **LOCATION**”

Text 2 – “ I live in ~~France~~” → “I live in **LOCATION**”

	~~France~~	**I**	**in**	**live**	**love**	~~Paris~~
Text 1	~~0~~	1	0	0	1	~~1~~
Text 2	~~1~~	1	1	1	0	~~0~~

Regular Expressions

Text 1 – “I love **LOCATION**”

Text 2 – “ I live in **LOCATION**”

	I	in	live	love	LOCATION
Text 1	1	0	0	1	1
Text 2	1	1	1	0	1

Regular Expressions

Text 1 – “I love **LOCATION**”

Text 2 – “ I live in **LOCATION**”

	I	in	live	love	LOCATION
Text 1	1	0	0	1	1
Text 2	1	1	1	0	1

D(Text 1, Text 2) = $\sqrt{(1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (1-1)^2}$ = **1.73** (<2.24)

Stemming & Lemmatization

Transform words in a way that if they represent the same meaning they are captured by the same token. Reduce inflection in language (Inflected Language)

- **Stemming** is the process of removing the last few characters of a given word, to obtain a shorter form, even if that form doesn't have any meaning.
  Ex: **Change, changing, changes, changer** becomes **chang**

- **Lemmatization** is the process of turning words into their root word.
  Ex: **Change, changing, changes, changer** becomes **change**

Part-of-speech filtering:

Consider only certain type of words (e.g: consider only verbs, nouns, adjectives and adverbs).

~~The~~	Dog	Is	Sitting	~~On~~	~~The~~	Table
DET	NOUN	VERB	VERB	ADP	DET	NOUN

Preprocessing

1. **Normalization**
   - Replace links with a special token
   - Normalize Dates
2. **Lowercasing**
3. **Tokenization:**
   - Compounds
   - Punctuation
4. **Remove Stop-Words**
5. **Stemming and lemmatisation**
6. **POS filtering**

Typically, in this order, but everything is optional!!

Preprocessing

Corpus

Bag-of-Words

- Tokenization
- Lowercasing
- Stop words
- Regular Expression

- Keyword Extraction
- Text Similarity
- Document Classification
- Sentiment Analysis
- Topic Modelling
- Information Retrieval
- Q&A
- Text Generation

Instituto Superior de Estatística e Gestão de Informação

Questions and doubts

Introduction to Text Mining

Complete these slides by reading the following material:

https://web.stanford.edu/~jurafsky/slp3/6.pdf
pp.1 to pp.11

Bibliography

- Dan Jurafsky and James H. Martin. **Speech and Language Processing** 3rd ed. https://web.stanford.edu/~jurafsky/slp3/
- Alammar, J., & Grootendorst, M. (2024). **Hands-On large language models: Language Understanding and Generation**. Sebastopol : O'Reilly Media.
  NOVA IMS Access: https://search.library.novaims.unl.pt/cgi-bin/koha/opac-detail.pl?biblionumber=97234

NOVA IMS

Information Management School

Thank you!

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa