DATA SCIENCE CONSULTING
Session 2

January 30th, 2023

The team with whom you will spend your session

Sami MHIRECH
Senior Data Scientist

sami.mhirech@ capgemini.com

Alice JAMET
Consultant

alice.jamet@ capgemini.com

Anne DUCOUT
Senior Data Scientist

anne.ducout@ capgemini.com

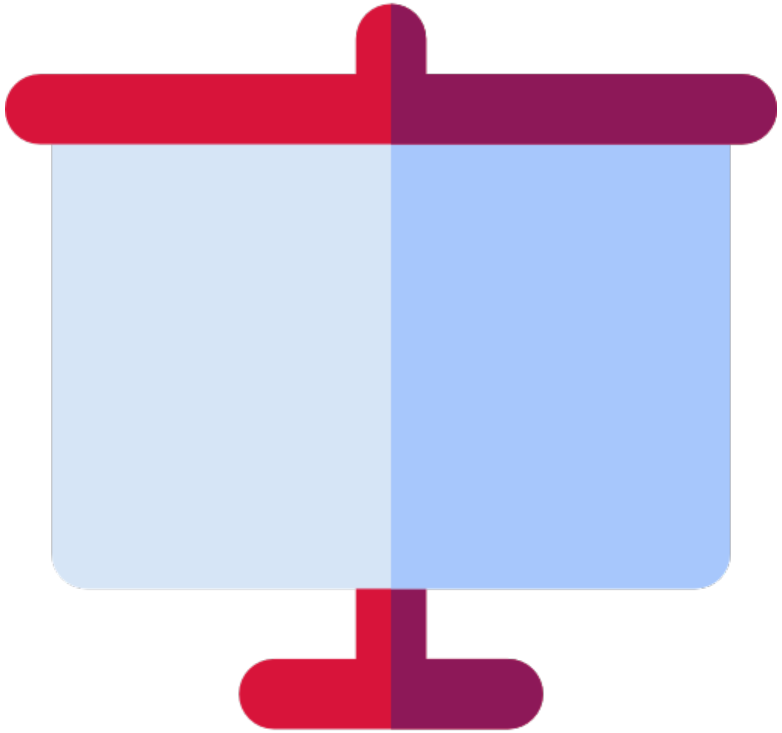
Naomi SERFATY
Senior Data Scientist

naomi.serfaty@ capgemini.com

Feedback

How did you find your homework for last week ?
What was challenging for you?

5’



Agenda

1. **Design of a customer journey**
2. Data pipeline : Data Preparation & Exploration
3. Text processing : Cleaning, Stemming & Lemmatization
4. Text representation
5. Summary of the session

When do people usually sign up for an energy contract?

When do people usually sign up for an energy contract?

In France, in 2016:

2,7 millions people when moving in

3 millions people at other occasions
A lot of online searches for energy suppliers in winter, as people wonder about heating options

Which touchpoint people use to subscribe to an energy contract?

Which touchpoint people use to subscribe to an energy contract?

In France, in 2016:

50% of people say they have subscribed online

In reality, only 2% of subscriptions are made online

From customer satisfaction to customer experience

Over the past two decades, with the growth of competition in retail and services, the B2C companies that have differentiated themselves and survived are those that have embraced the **paradigm shift** in customer relations: from implementing isolated retention and satisfaction measurement initiatives to building a **customer experience** vision.

Customer Experience is built step by step :

Our main points of focus for today

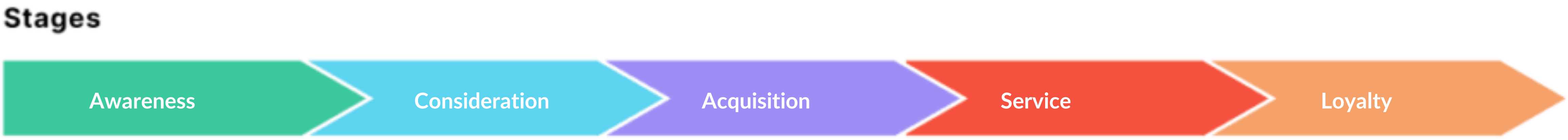
1. **Analyse :** the less obvious the insights are, the better positioned to define an innovative and differentiating strategy
2. **Design and test :** design the experience, after digesting the insights obtained in the previous analysis phase, using strategic design methodologies – like Design Thinking – and test it
3. **Control and correct :** to maintain a control over the contact points, to know and understand the reaction and impressions of real customers after the process and to be able to quickly react

"Customers don't just want to buy a product or service, but to feel the whole experience behind it." – Hartmut Esslinger (Founder of frog Design)

Which key elements matter when designing a customer journey?

1. The scope

When does the customer journey start and end?



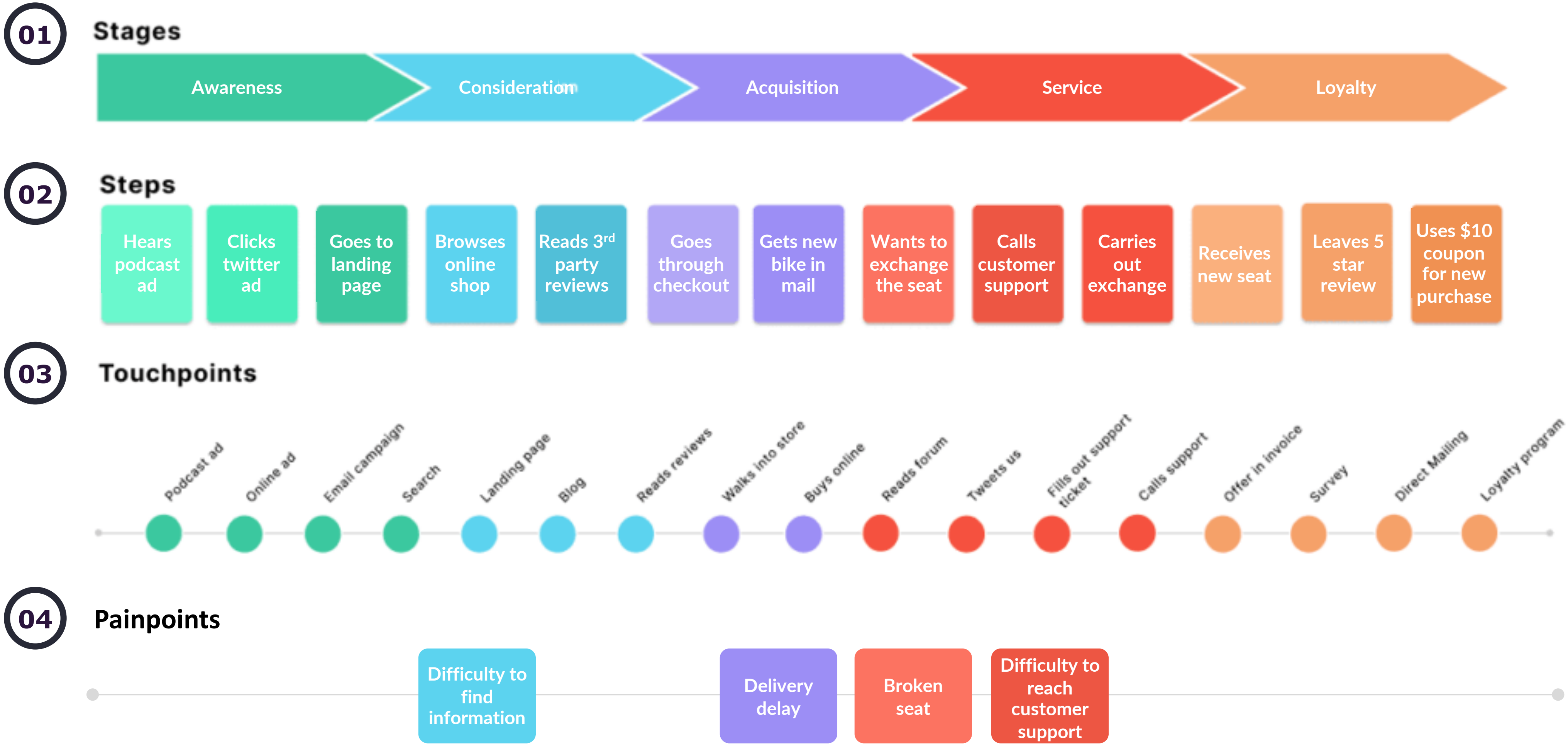
Stages

Awareness → Consideration → Acquisition → Service → Loyalty

Which key elements matter when designing a customer journey?

2. Analysis through big stages, smaller steps, touchpoints and pain points

Example of customer journey for purchasing a bike



01 Stages

Awareness

Consideration

Acquisition

Service

Loyalty

02 Steps

Hears podcast ad

Clicks twitter ad

Goes to landing page

Browses online shop

Reads 3rd party reviews

Goes through checkout

Gets new bike in mail

Wants to exchange the seat

Calls customer support

Carries out exchange

Receives new seat

Leaves 5 star review

Uses $10 coupon for new purchase

03 Touchpoints

Podcast ad

Online ad

Email campaign

Search

Landing page

Blog

Reads reviews

Walks into store

Buys online

Reads forum

Tweets us

Fills out support ticket

Calls support

Offer in invoice

Survey

Direct Mailing

Loyalty program

04 Painpoints

Difficulty to find information

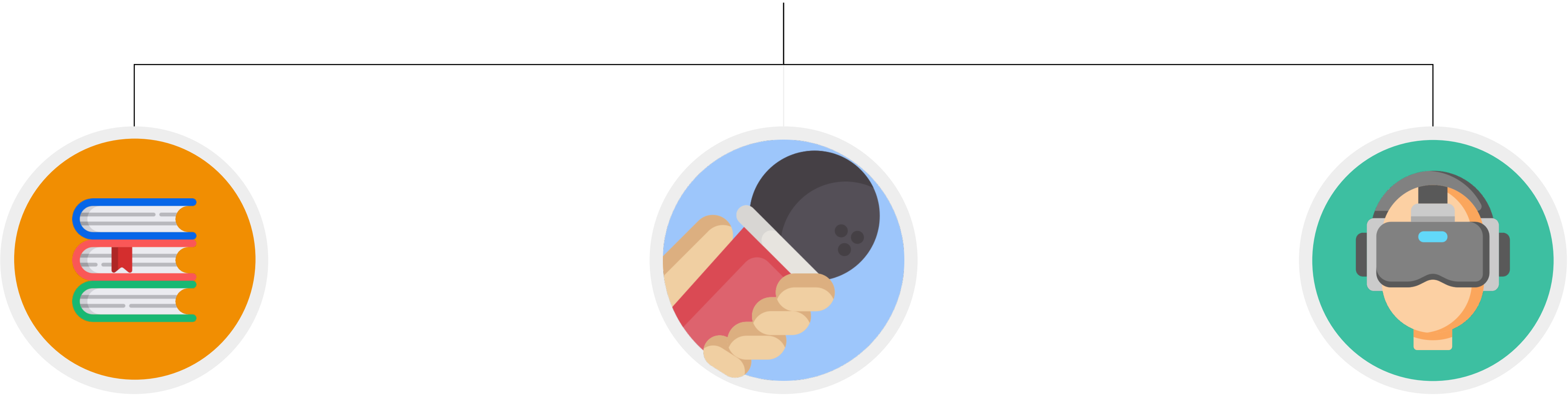
Delivery delay

Broken seat

Difficulty to reach customer support

A customer relations consulting project usually starts with the design of the legacy customer journey

Often, three areas of investigation are carried out in parallel :



Digging into the project **legacy** of the customer relations department

Conducting **interviews** within the company's **customer-facing departments** and outside the company directly with **customers**

Simulating various **customer journey scenarios**, trying to be as exhaustive as possible

Exercise (in groups)

Design what the customer journey would be when signing up an energy contract with TotalEnergies.

⏱ 25’

1. Draft first hypothesis about the customer journey

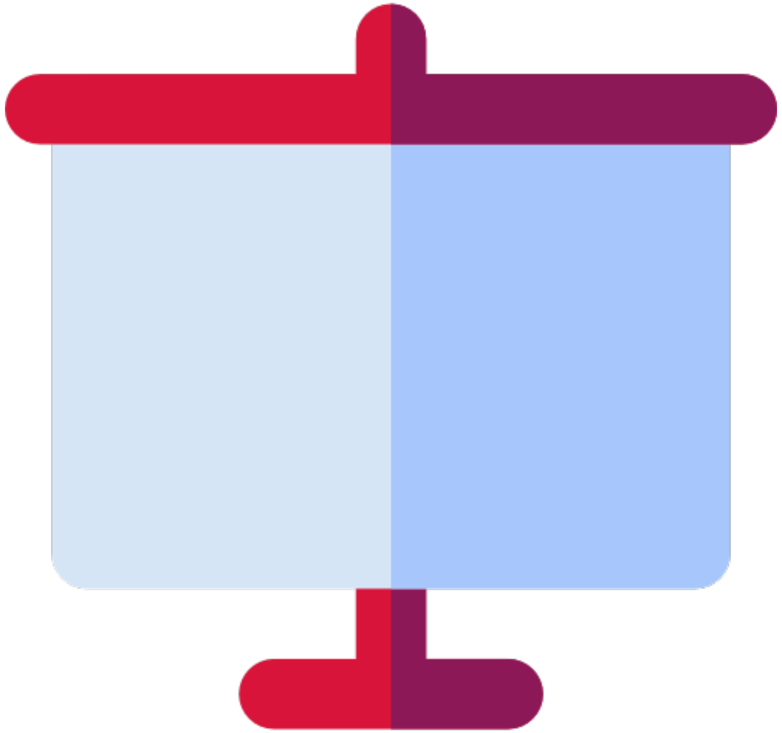
2. Simulate a customer journey

3. Quick interviews with Alice and Sami

Restitution

Could you identify the customer journey?
What was challenging for you?

⏱ 5’



An example of customer journey : a French distribution service for energy renovation grants

	1 I want to know more about the financial aid	2 I create my account	3 I submit my grant application	4 Ma demande de subvention est en cours d'instruction	*Optional* 5 My grant application is being processed	6 I carry out my renovation	7 I submit my request for payment	8 My payment request is being processed	*Optional* 9 I am subject to an on-site inspection	10 I receive the bonus payment	*Optional* 11 I make a claim	*Optional* 12 I make an appeal
	▨		🟩	🟩	▨		🟩	🟩		🟩	░	░
Phone	●	●	●	●	●	○	●	●	●	●	●	●
Forms	●	●	○	●	○	○	○	●	○	○	●	○
Website	●	●	●	●	●	○	●	●	○	●	○	○
Chat	○	○	○	○	○	○	○	○	○	○	○	○
E-mail	○	●	○	●	●	○	○	●	○	○	○	○
Mail									●		●	●
Social media												

Direct touchpoint : ● Available ○ Unavailable

Voice of the Customer Analysis Priority : 🟩 High ▨ Medium ░ Low

Homework (due February 3rd)

In group, finish to identify the customer journey for TotalEnergy users

- The main objective of this homework is to:
  - Identify the **different steps** of the customer journey and the **pain points** associated to each step
  - In a second time, think about the **different channels** used by consumers (online vs offline, type of devices used,…)

Please send your completed customer journey in a PowerPoint File by **Friday 3rd evening** to *alice.jamet@capgemini.com* and *sami.mhirech@capgemini.com*

If you have any questions, feel free to contact us by email

Agenda

1. Design of a customer journey
2. **Data pipeline : Data Preparation & Exploration**
3. Text processing : Cleaning, Stemming & Lemmatization
4. Text representation
5. Summary of the session

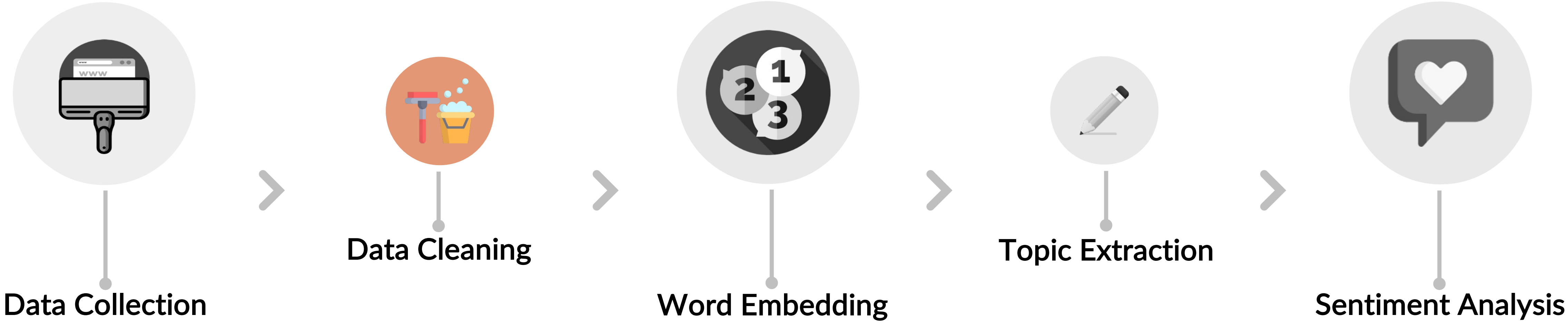
Review of last week's homework

5'

Data scraping using Selenium

- *What were your main takeaways ?*
- *Did you face any pain points during the scraping ?*
- *If you had to improve your code, what would you do ?*

Data pipeline



Data Collection

Data Cleaning

Word Embedding

Topic Extraction

Sentiment Analysis

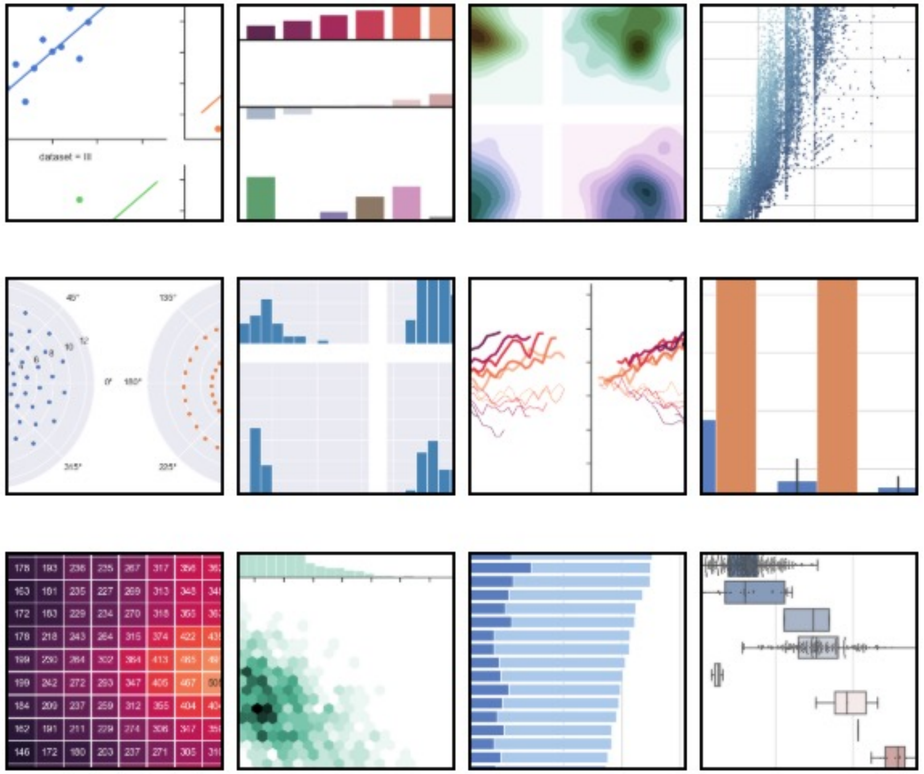
Data Preparation & Exploration

Data preparation is the step after data collection in the machine learning life cycle. It’s the process of cleaning and transforming the collected raw data before using it for statistical approaches.

- **Raw data cannot be used without primary processing. Here are some usual points of attention necessiting some transformations**
  - **Data types** : convert string to integer/float, convert string to list
  - **Encoding** : text data need to be read with a proper encoding
  - **Duplicates** : check for duplicates, on a subset or all features. What strategy ? keep all, keep 1, keep none…
  - Feature Imputation for **missing values**
    - Value imputation (mean, median or mode of the feature)
    - Nearest neighboor : filling data with a value from another similar sample
    - Deleting the sample
  - Check for **outliers** and inspect : using standard deviations, inter quartile range (IQR).
  - **Feature encoding** : convert categorical feature to numerical value (label encoding, One-hot encoding)
  - Text, image and audio **representation** : embedding, pixels, signals…

- **Some needed transformations cannot be observed without deeper analysis : Exploratory Data Analysis**

Exploratory Data Analysis



EDA is the process of using **statistical** tools and ideas to **inspect** data in order to summarize their main features, discover patterns, spot anomalies, test hypothesis, or check assumptions.

- **Examining each variable** by itself. Python methods like `df.describe()` and `df.info()`
  - **Pandas-profiling**
  - Statistical tools (distribution, metrics, length of string, number of words…)

- Studying **relationships** among variables
  - Correlation
  - Feature selection

- Making graphs **to visualize** and have a better comprehension of the dataset. There are some powerful libraries like :
  - Matplotlib
  - Seaborn

Iterative process between data preparation & EDA

EDA on Jupyter Notebook

Agenda

1. Design of a customer journey
2. Data pipeline : Data Preparation & Exploration
3. **Text processing : Cleaning, Stemming & Lemmatization**
4. Text representation
5. Summary of the session

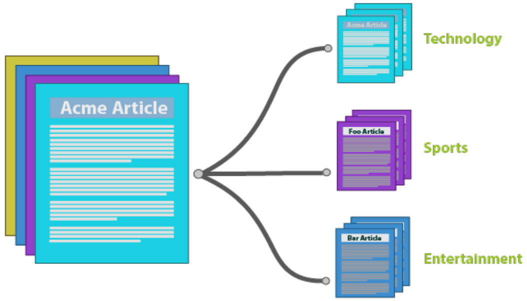
What is Natural Language Processing ?

> ***"Natural language processing (NLP) is a crucial part of artificial intelligence (AI), modeling how people share information"***

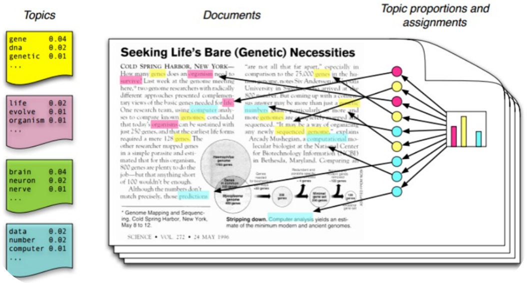
It is a field of study that focuses on making sense of human written text, blending both **linguistics** and **machine learning** in order to make machines learn how to process large amount of natural language data.

NLP applications include :

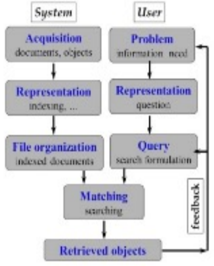
Classifying Documents
(supervised learning)



Understanding key topics & trends
(unsupervised learning)


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Finding relevant information
(information retrieval)



In this course, you will learn the basics of NLP along with some advanced techniques : text exploration, cleaning and processing, word embeddings, topic extraction and sentiment analysis.

Pre-processing: Introduction

According to you, what processes should we apply on our data to improve the quality of our analysis ?

```
['\n            Nous avons passé un excellent week-end, tout était bien entretenu\n         ',
 "\n            Parc très agréable, difficile de s'y retrouver au début. Mais on s'y fait à la longue. Le cottage un
peu vieilli mais cela peut aller. Au niveau cuisine, il manque un peu de choses (une vrai poele, des couverts en plus
quand les premiers sont au lave vaisselle). Cadre très agréable dans la foret surtout avec les fortes chaleurs. La lo
cation de la voiture électrique s'est bien passé et très vite (un peu cher sans doute pour la semaine !).",
 '\n            Pas grand chose ne marche, ni l'internet, ni wifi (ce n'en serais pas trop important s'il y avait du
réseau) Les cottage sont mal entretenus, j'ai trouvé au petit matin une puce et un tique dans les draps. ',
 "\n            Moi je vais parler aujourd'hui  du service commercial de center parc . J'ai fait une réservation pour
deux nuits sur le site de Center parcs au prix de 343 € quelques jours plus tard je vois sur vente-privee la même off
re avec une nuit de plus pour 269€. Lors ce que j'appelle le service pour obtenir un geste commercial sachant que je
suis cliente fidèle il me réponde: pas de chance pour vous et puis vous n'avez pas pris d'assurance annulation. Je le
urs explique que je veux pas annuler mais a titre commercial et car je suis cliente fidèle une compensation financièr
e. Je suis vraiment déçue\n        "]
```

```
"[ Alice ' s Adventures in Wonderland by Lewis Carroll 1865 ] CHAPTER I . Down the Rabbit - Hole Alice was beginning
to get very tired of sitting by her sister on the bank , and of having nothing to do : once or twice she had peeped i
nto the book her sister was reading , but it had no pictures or conversations in it , ' and what is the use of a book
,' thought Alice ' without pictures or conversation ?' So she was considering in her own mind ( as well as she could
, for the hot day made her feel very sleepy and stupid ), whether the pleasure of making a daisy - chain would be wor
th the trouble of getting up and picking the daisies , when suddenly a White Rabbit with pink eyes ran close by her .
There was nothing so VERY remarkable in that ; nor did Alice think it so VERY much out of the way to hear the Rabbit
say to itself , ' Oh dear ! Oh dear ! I shall be late !' ( when she thought it over afterwards , it occurred to her t
hat she ought to have wondered at this , but at the time it all"
```

Which pre-processing steps should we follow to have clean text data ?

Today, we will focus on the following next steps :

Modifying the structure

- Transforming the text into a “**corpus**”
- **Tokenizing**

Handling special characters

- Removing **punctuation**
- Removing or replacing (*highly recommended*) specific characters
  Replacing **accents** (*depending on language*)

Removing the noise

- Removing **stop words**
- **Lemmatization**
- **Stemming** (Optional)

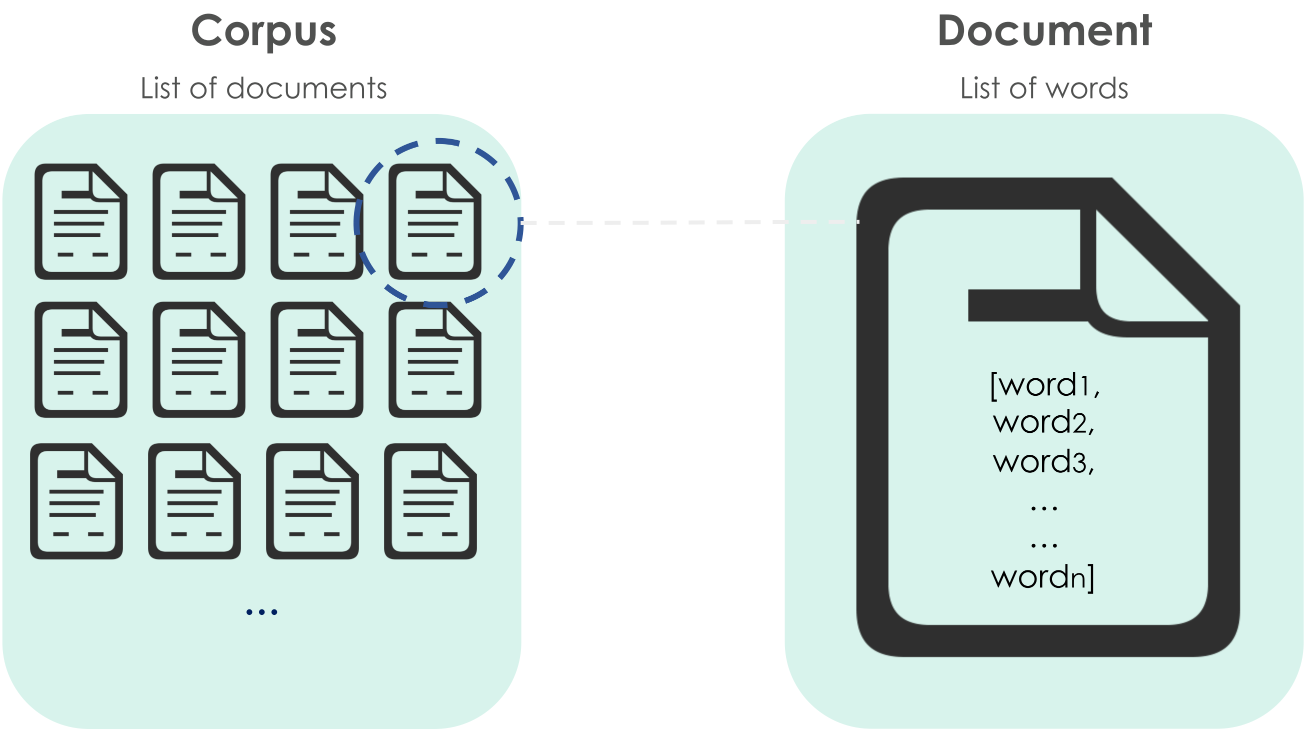
All of these is part of what we call **Natural Language Processing**, which leads to **Natural Language Understanding**, which we will focus on during the next session.

Corpus of texts

Before getting started, let’s introduce a basic concept: The Corpus

- A corpus or text corpus is a **large and structured set of texts** on which we will perform our analysis

- Each corpus contains separate documents, which might be articles, stories, or book volumes : each document is treated as a **separate entity or record**



Corpus

List of documents

…

Document

List of words

[word1,
word2,
word3,
…
…
wordn]

Tokenization

This method is used to **tokenize** the text, tokenization is the process of **breaking a stream of text** into words, phrases, symbols, or other meaningful elements called **“tokens”**

1st level: considering a sentence as a token

- Related to the structure of the text
- In a novel, dialogs should not be considered as the rest of the text

Input: [“How is it to be a Data Scientist? Olivier Auliard answered: It is super cool to be a Data Scientist”]

Output: [“How is it to be a Data Scientist”, “Olivier Auliard answered”, “ It is super cool to be a Data Scientist”]

2nd level: considering the word as a token

- The document is split word by word

Input: ” it is super cool to be a Data Scientist”

Output: [“It”, “is”, “super”, “cool”, “to”, “be”, “a”, “Data”, “Scientist”]

3rd level: N-grams

- We consider few words together
- Unigrams are tokens of one word, bi-grams are tokens of two, etc.

Input (Bi-grams): ” it is super cool to be a Data Scientist”

Output: [“It is”, “is super”, “super cool”, “cool to”, “to be”, “be a”, “a Data”, “Data Scientist”]

No tokenizer is better than the other : **the choice of your tokenizer should be made to fit your data and your NLP problematic**

Cleaning on Jupyter Notebook

Hands-on 1: Pre-processing your data

20’

Load, explore and clean your dataset

Break

10’

Let’s come back in ten minutes !

Vocabulary improvement: Orthographic correction

Detecting words with low frequency of apparition and removing them

Comparing with dictionaries (existing or non-existing words)

Using specific algorithms like Levenshtein Distance Spelling Correction :

- deletion, transposition, replacement or insertion of characters to create potential candidates
- Compare with dictionaries using edit distance

⚠ Those methods are often very time and memory consuming ⚠

Inflected language

In grammar, inflection is **the modification of a word to express different grammatical categories** such as tense, number, gender, case, voice, aspect and person.

An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change, to a common root.

- “person”
- “persons”
- “person’s”
- “personal”
- “unpersonal”

→ **person**

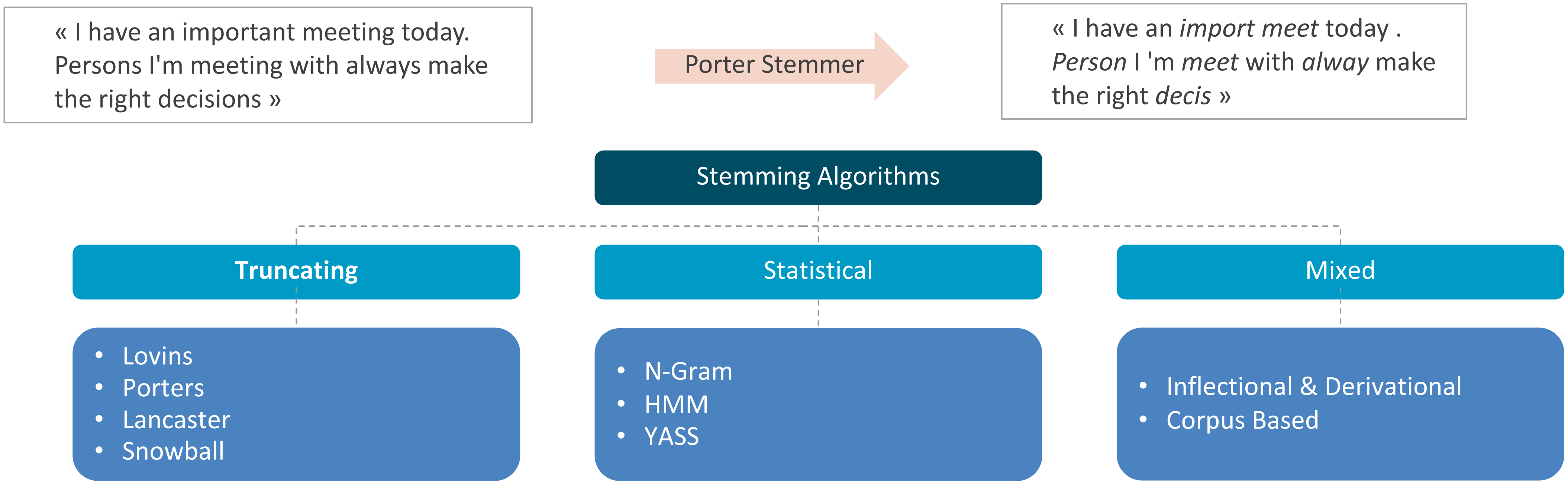
- “am”
- “are”
- “is”
- “was”
- “were”

→ **be**

Text normalization techniques have been developped to deal with inflected forms

Stemming

- **Stemming is the process of reducing the amount of inflected words** such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.
- By removing inflectional form of a word, we focus on the **meaning**
- “Rude” process as it could remove a part of a word even if it is not an inflectional form
- Used in Information Retrieval, basing the search process only on the word’s stem.



« I have an important meeting today. Persons I'm meeting with always make the right decisions »

Porter Stemmer

« I have an *import meet* today . *Person* I 'm *meet* with *alway* make the right *decis* »

Stemming Algorithms

Truncating
- Lovins
- Porters
- Lancaster
- Snowball

Statistical
- N-Gram
- HMM
- YASS

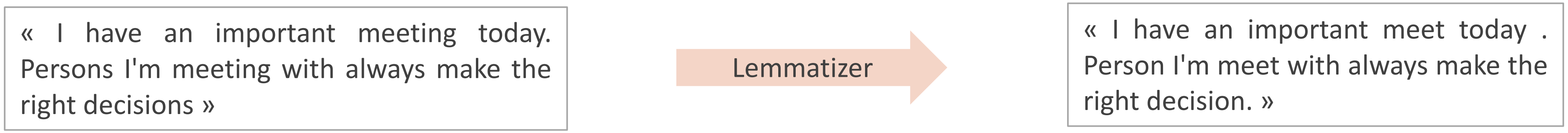
Mixed
- Inflectional & Derivational
- Corpus Based

https://www.researchgate.net/publication/284038938_A_Comparative_Study_of_Stemming_Algorithms

Lemmatization

- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the **root word belongs to the language**.
- In Lemmatization root word is called **Lemma**.
- A trivial way to do **lemmatization** is by simple **dictionary lookup**. It is also the most common way.

Example



« I have an important meeting today. Persons I'm meeting with always make the right decisions »

« I have an important meet today . Person I'm meet with always make the right decision. »

Trade-off between stemming & lemmatization

Stemming VS Lemmatization	
Stemming	**Lemmatization**
• Performed by “stemmers” • produces a word’s “stem”	• Performed by “lemmatizers” • produces a word’s “lemma”
• Better → Bet • am → am • having → hav	• Better → Good • am → be • having → have

- Stemmers are **faster**, and will better reduce vocabulary size
- **Lemmatizers** ensure you work with existing words, and deal with special cases
- These techniques are **language-specifics**
- More about this: http://stackoverflow.com/questions/17317418/stemmers-vs-lemmatizers

Let’s do it with Python

Hands-on 2: Stemming & Lemmatization

10’

Get ready for starting basic NLP : Stemming and lemmatization !

Agenda

1. Design of a customer journey
2. Data pipeline : Data Preparation & Exploration
3. Text processing : Cleaning, Stemming & Lemmatization
4. **Text representation**
5. Summary of the session

Bag-of-words (1/2)

Let’s consider :

- A corpus of documents where each **document** is a text
- Each document contains a certain number of **words** that we can tokenize
- Each word, or group of words will be a **token** (monogram, bigram, …)

A bag-of-word is the representation of the document by a numerical vector containing the **counts of each token**

In this part of the course, we will use **bag-of-words** techniques, which means each word will be considered independently, **whatever their position in the sentence**

Bag-of-words (2/2)

The Document Term Matrix

In order to get the most important topics in a corpus of documents, a first simple method would consist in looking at the words and **compare their occurrences**.

→ The **Document Term Matrix** synthetizes the words' occurrences in a corpus of documents : rows correspond to documents and columns correspond to tokens.

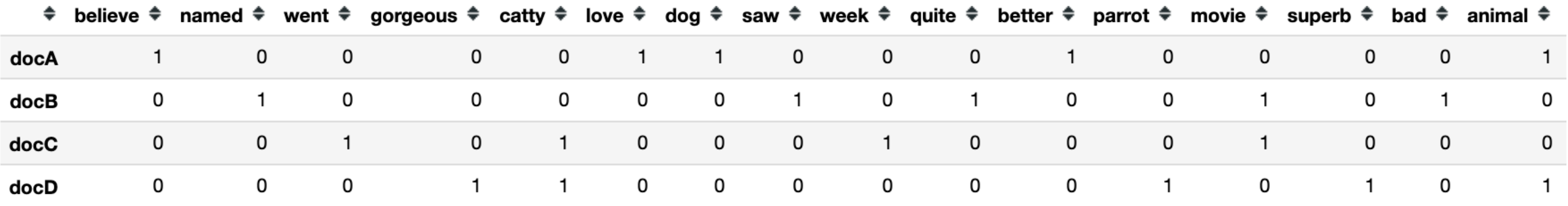
Example

docA = 'I believe cats are better animals than dogs, I love cats !'

docB = 'I saw this movie named cats, it was quite bad'

docC = 'I went to the movies with catty last week'

docD = 'Catty has a gorgeous animal : a superb green parrot !'



	believe	named	went	gorgeous	catty	love	dog	saw	week	quite	better	parrot	movie	superb	bad	animal
docA	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1
docB	0	1	0	0	0	0	0	1	0	1	0	0	1	0	1	0
docC	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0
docD	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1

…

…

TF-IDF

A score that better reflects how words are related to documents within a corpus :

$$\boldsymbol{tfidf(t,d) = tf(t,d)\times idf(t)}$$

tf: Term frequency

We have different possibilities to calculate the term frequency. The most frequent are raw count and term frequency.

idf: Inverse document frequency

A measure of how much information the word provides i.e. if the word is common or rare across all documents : it diminishes the weight of terms that occur very frequently in the corpus and increases the weight of terms that occur rarely.

Formula of Inverse Document Frequency:

$$idf(t) = \log\left|\frac{n}{\{d \in D : t \in d\}}\right|$$

?

Could you comment this formula ?
Why do we need a log ?

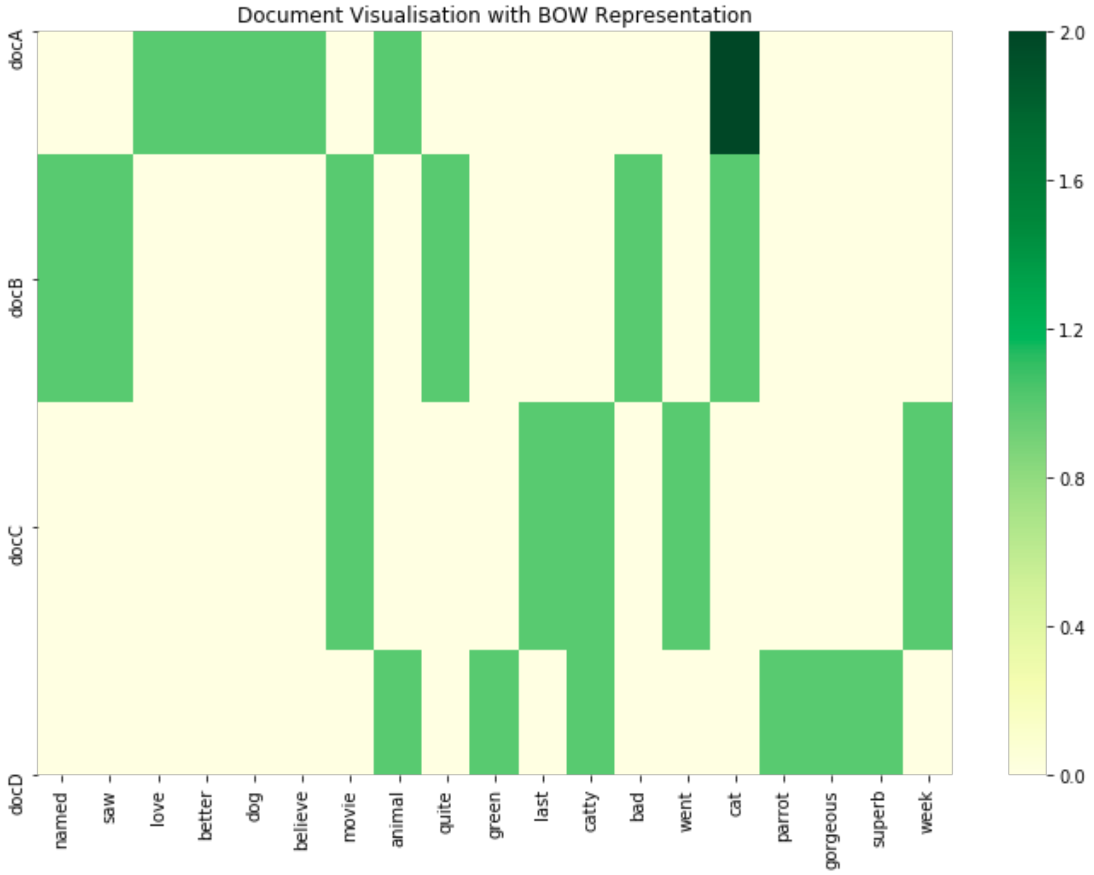
With n the total number of documents, D the set of all documents and d a given document in D, t is a given term.

Thus, $\{d \in D: t \in d\}$ is the number of documents in which the term t appears.

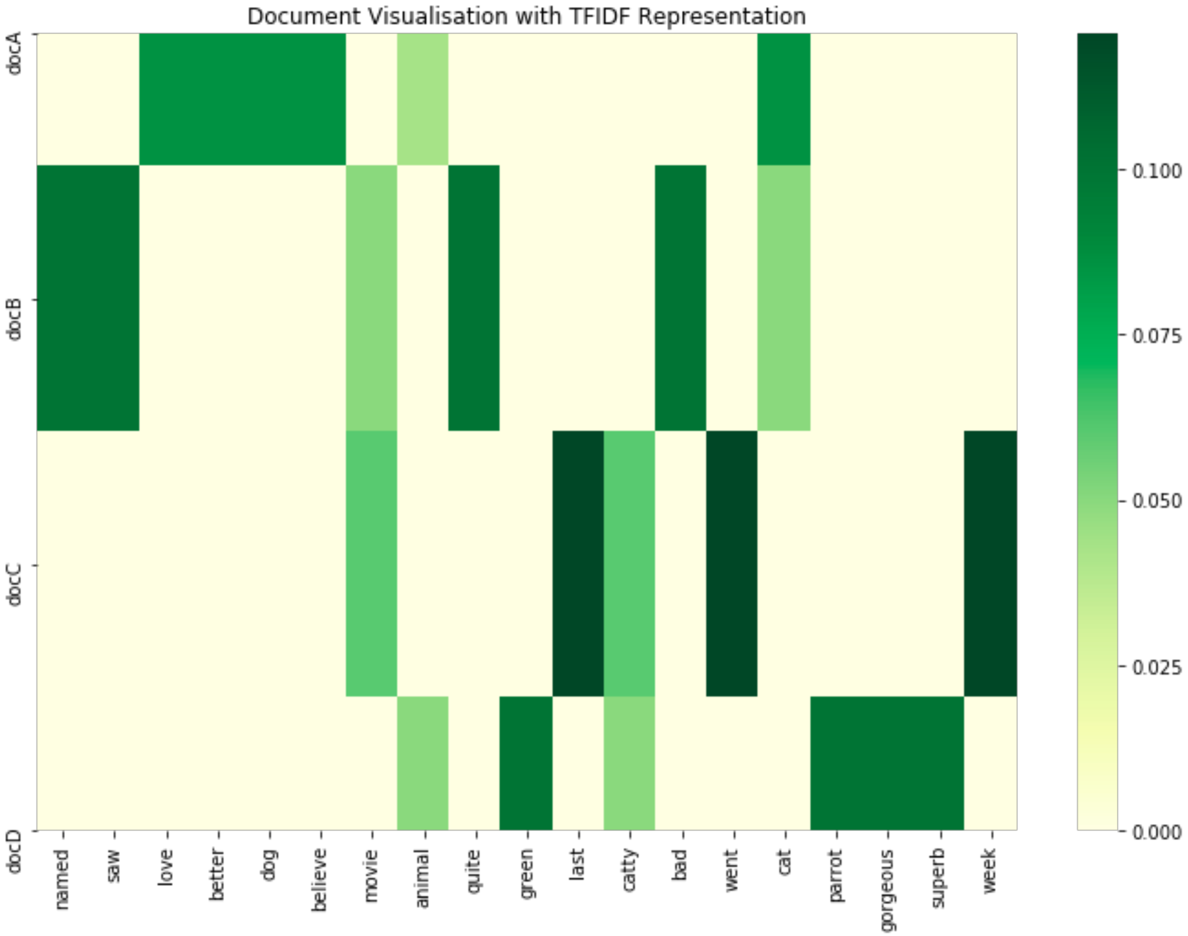
N.B. They are various formulas for tf and idf, see on Wikipedia

BOW vs TF-IDF

BOW

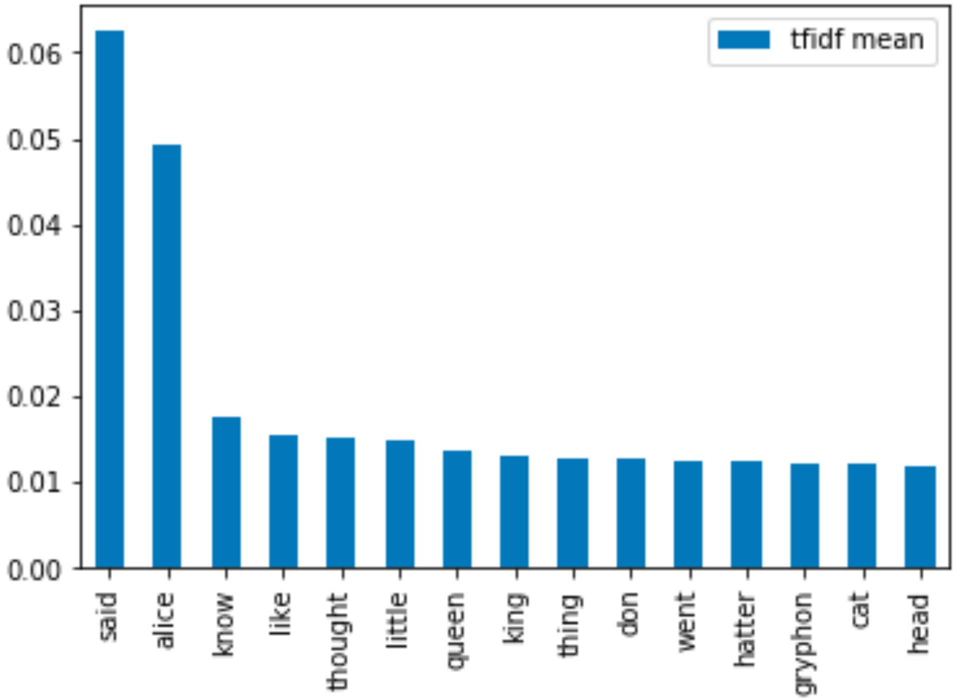


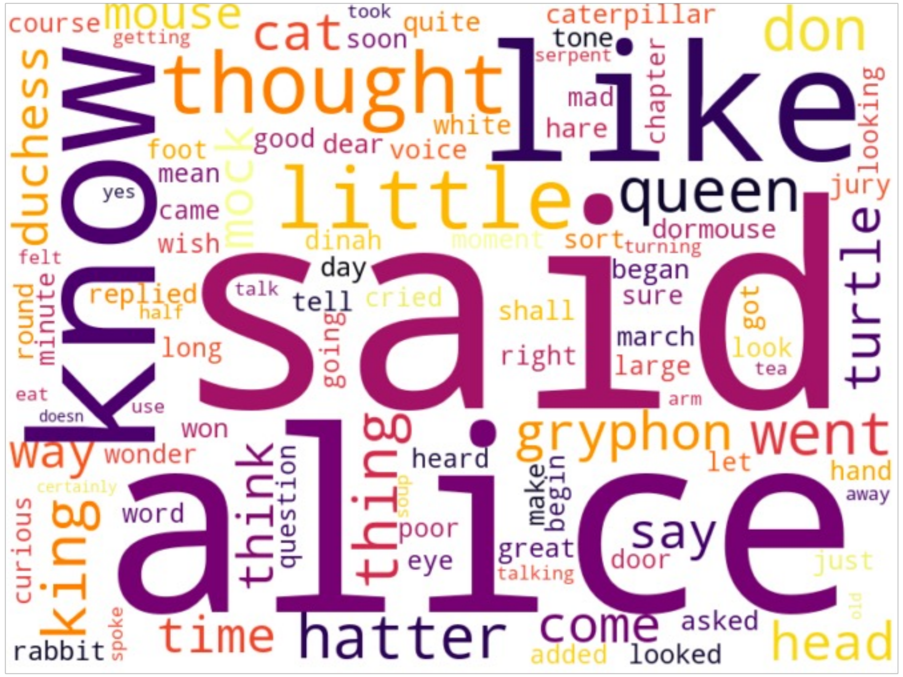
TFIDF



? What are the pro’s and con’s of each representation ?

Interpret the TF-IDF matrix





Let’s do it with Python !

Hands-on 3: BOW - TFIDF

10’

Use the notebook provided to build your own TF-IDF matrix from scratch, then use a Python library (*TfidfVectorizer*).

What are the most relevant terms in of your scrapped data?

Agenda

1. Design of a customer journey
2. Data pipeline : Data Preparation & Exploration
3. Text processing : Cleaning, Stemming & Lemmatization
4. Text representation
5. **Summary of the session**

Today we learnt

Cleaning steps

- **Convert documents into a corpus**
- **Adjust characters:**
  - Convert to lowercase
  - Remove special characters, punctuation, accents, double spaces, numbers…
- **Careful :** Some punctuation sets can be useful to keep (ex. : “:)”)
- **Process some words:** spell-checking, **stemming** / **lemmatizing**, remove stop-words & **tokenization**

Text representation

- Represent the corpus as a **Document-Term Matrix**
- Identify and represent terms that are important
- Tools : bag-of-words, TF-IDF, wordcloud

Work for next week

Instructions for February 3rd !

- Be sure that you have applied every steps of text processing to your reviews
- Create a TF-IDF Matrix with all the posts you scrapped on the web, and find the best way to represent it (WordCloud ?)
- Bonus reward : build 2 functions :
  - One that takes a corpus of raw text and creates a new corpus with cleaned and option for lemmatizion or stemming.
  - One that takes a corpus (or a dataframe text column) and creates a Wordcloud from it.

Send your homework by February 3rd to :

- naomi.serfaty@capgemini.com
- anne.ducout@capgemini.com

- Feel free to contact us by email !

Course evaluation

Did you like that first course? It’s time to share your feedbacks!

Chapter 2 - X-HEC NLP Bootcamp 2023

Capgemini

Thank you for your attention

See you next week !

GOODBYE !

January 30th, 2023