DATA SCIENCE CONSULTING Session 2

January 30th, 2023

Presentation Title | Author | Date Company Confidential © Capgemini Invent 2021. All rights reserved | 1

The team with whom you will spend your session

**Sami MHIRECH** Senior Data Scientist 

sami.mhirech@ capgemini.com

**Anne DUCOUT** Senior Data Scientist 

anne.ducout@ capgemini.com

**Alice JAMET** Consultant 

alice.jamet@

capgemini.com

**Naomi SERFATY** Senior Data 

Scientist

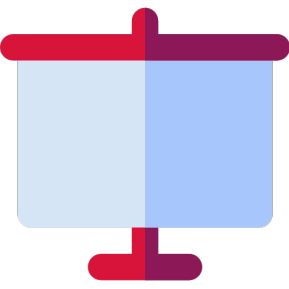
naomi.serfaty@ capgemini.com

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 2

Feedback

5’ **How did you find your homework for last week ?** 

**What was challenging for you?**

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 3

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 4

When do people usually sign up for an energy contract? 

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 5

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*

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 6

Which touchpoint people use to subscribe to an energy contract? 

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 7

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 8

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)*

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 9

Which key elements matter when designing a customer journey?

**1. The scope**

When does the customer journey start and end?

**Awareness Consideration Acquisition Service Loyalty**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 10

Which key elements matter when designing a customer journey? 

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

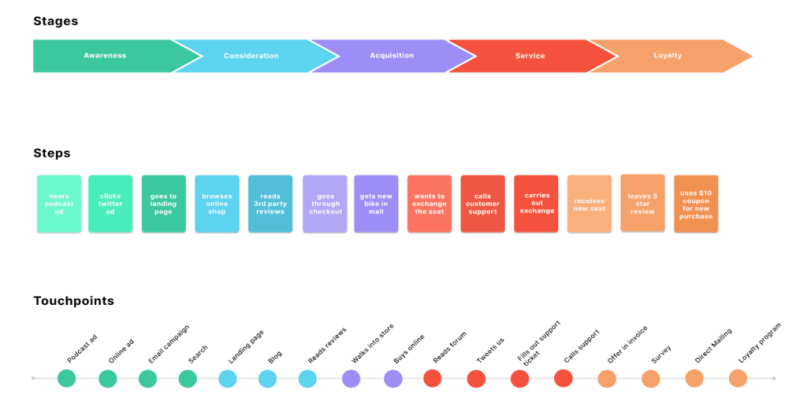
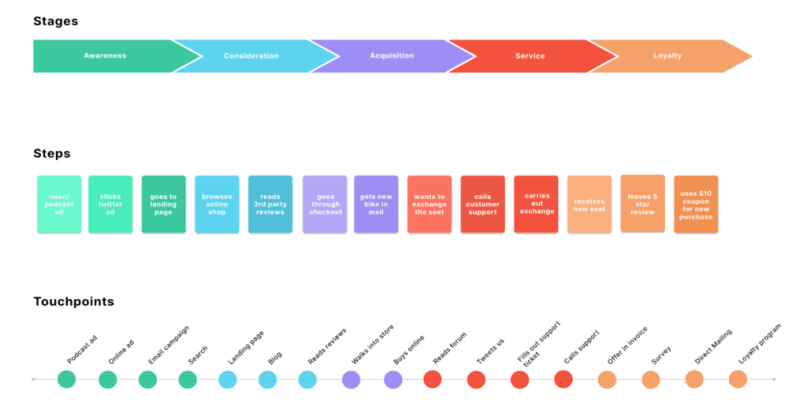
Example of customer journey for purchasing a bike

**01**

**Awareness Consideration Acquisition Service Loyalty**

**02**

**03 04**

**Hears** 

**podcast**

**ad**

**Painpoints**

**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**

**Difficulty to find**

**information**

**Delivery delay**

**Broken seat**

**Difficulty to reach**

**customer support**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 11

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 12

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 13

Restitution

5’ **Could you identify the customer journey?** 

**What was challenging for you?**

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 15

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

*Optional Optional Optional*

*Optional*

1 2 3 4 5 6 7 8 10 11 12 9

**I make an**

**appeal**

**Phone**

**Forms**

**Website**

**Chat**

**E-mail**

**Mail**

**Social media**

**I want to know more about the financial aid**

**I create my account**

**I submit my grant**

**application**

**Ma demande de subvention est en cours d’instruction**

**My grant application is being processed**

**I carry out my**

**renovation**

**I submit**

**my request for payment**

**My payment request is being processed**

**I am subject to an on-site inspection**

**I receive the bonus payment**

**I make a claim**

Direct touchpoint : Available Unavailable Voice of the Customer Analysis Priority : High Medium Low

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 16

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 17

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 18

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 ?*

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 19

Data pipeline



Data Collection

Data Cleaning

Word Embedding

Topic Extraction

Sentiment Analysis

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 20

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**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 21

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**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 22

EDA on Jupyter Notebook



Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 23

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 24

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)***

**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.

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 25

Pre-processing: Introduction

**According to you, what processes should we apply on our data**

**to improve the quality of our analysis ?**

****Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 26

Which pre-processing steps should we follow to have clean text data ? **Today, we will focus on the following next steps :**

**Modifying the structure**

**Handling special characters**

**Removing the noise**

▪ Transforming the text into a “**corpus**” 

▪ **Tokenizing**

▪ Removing **punctuation**

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

▪ 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.

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 27

Corpus of texts

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

**Corpus**

List of documents

▪ 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**

**…**

**Document** List of words

[word1, 

word2,

word3,

…

…

wordn]

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 28

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

**2nd level: considering the word as a token**

▪ The document is split word by word

**3rd level: N-grams**

▪ We consider few words together

▪ Unigrams are tokens of one word, bi-grams are tokens of two, etc.

**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”]

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

**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**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 29

Cleaning on Jupyter Notebook



Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 30

Hands-on 1: Pre-processing your data

20’

**Load, explore and clean your dataset**

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 31

Break

10’

**Let’s come back in ten minutes !**

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 32

Vocabulary improvement: Orthographic correction

Detecting words with low frequency 

of apparition and removing them

Comparing with dictionaries 

(existing or non-existing words)

Those methods are often very time and memory consuming

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





Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 33

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”

• “am”

• “are”

**person be**

• “is”

• “was”

• “were”

**Text normalization techniques have been developped to deal with inflected forms**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 34

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

Stemming Algorithms

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

• Lovins

**Truncating** Statistical Mixed • N-Gram

• Porters • Lancaster • Snowball

• HMM • YASS

• Inflectional & Derivational • Corpus Based

https://www.researchgate.net/publication/284038938\_A\_Comparative\_Study\_of\_Stemming\_Algorithms

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 35

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 »

Lemmatizer

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 36

Trade-off between stemming & lemmatization

**Stemming VS Lemmatization**

**Stemming Lemmatization**

• Performed by “stemmers” • produces a word’s “stem”

• Better 🡪 Bet

• am 🡪 am

• having 🡪 hav

• Performed by “lemmatizers” • produces a word’s “lemma”

• 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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 37

Let’s do it with Python



Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 38

Hands-on 2: Stemming & Lemmatization

10’

**Get ready for starting basic NLP : Stemming and lemmatization ! **

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 39

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 40

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**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 41

**The Document Term Matrix**

Bag-of-words (2/2)

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 !'*

… 

…

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 42

TF-IDF

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

����� �, � = �� �, � ×���(�)

***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:

��� � = log� � ∈ �:� ∈ �

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

With � the total number of documents, � the set of all documents and � a given document in �, � is a given term. Thus, {� ∈ �:� ∈ �} is the number of documents in which the term � appears.

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 43

BOW vs TF-IDF

**BOW TFIDF**

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

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 44

Interpret the TF-IDF matrix

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 45

Let’s do it with Python !



Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 46

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?**

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 47

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**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 48

Today we learnt

**Cleaning steps Text representation**

• **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**

• Represent the corpus as a **Document-Term Matrix**

• Identify and represent terms that are important

• Tools : bag-of-words, TF-IDF, wordcloud

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 49

Work for next week

**Instructions for February 3rd !**

Send your homework by February 3rd to :

• 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.

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

• Feel free to contact us by email !

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 50

Course evaluation

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

****

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 51

**Thank you for your attention See you next week !**

**GOODBYE !**

**January 30th, 2023**

Master X-HEC – 30012023 © Capgemini 2023. All rights reserved | 52