DATA SCIENCE CONSULTING
Session 2

February 10th, 2020

Agenda

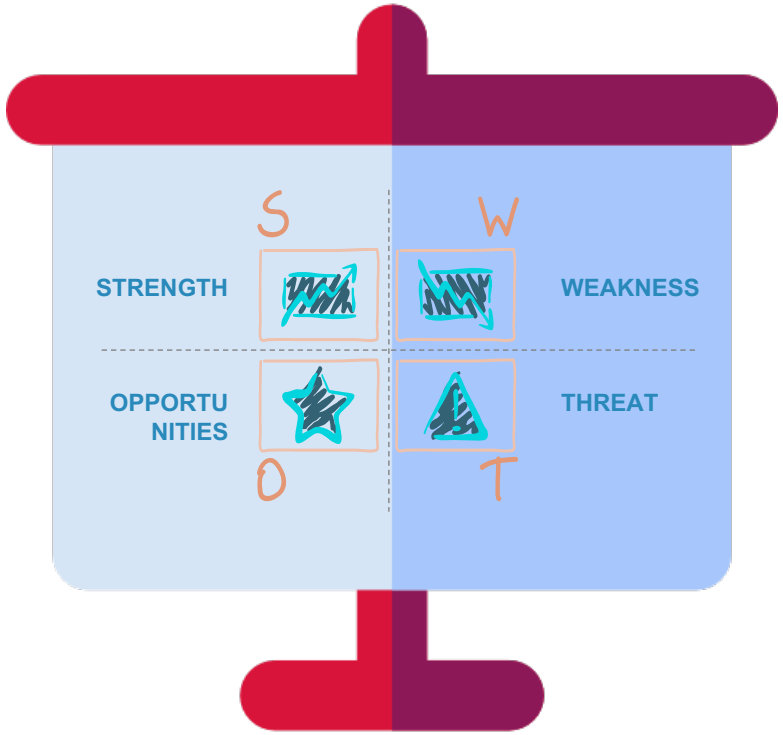
1. **SWOT analysis restitution**
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

Restitution

10’

SWOT Analysis ? What did you find ?



© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. **Hospitality industry & restaurants KPI definition**
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

Hospitality industry and Restaurants KPI definition

A world of services

Image by Laureate Hospitality

"Hospitality industry means businesses such as hotels, bars, and restaurants that offer people food, drink, or a place to sleep". In a nutshell, customer services…

While everything may feel like **a priority** and finding **the right areas to focus on** can feel overwhelming, it's important to identify the right goals to track periodically.

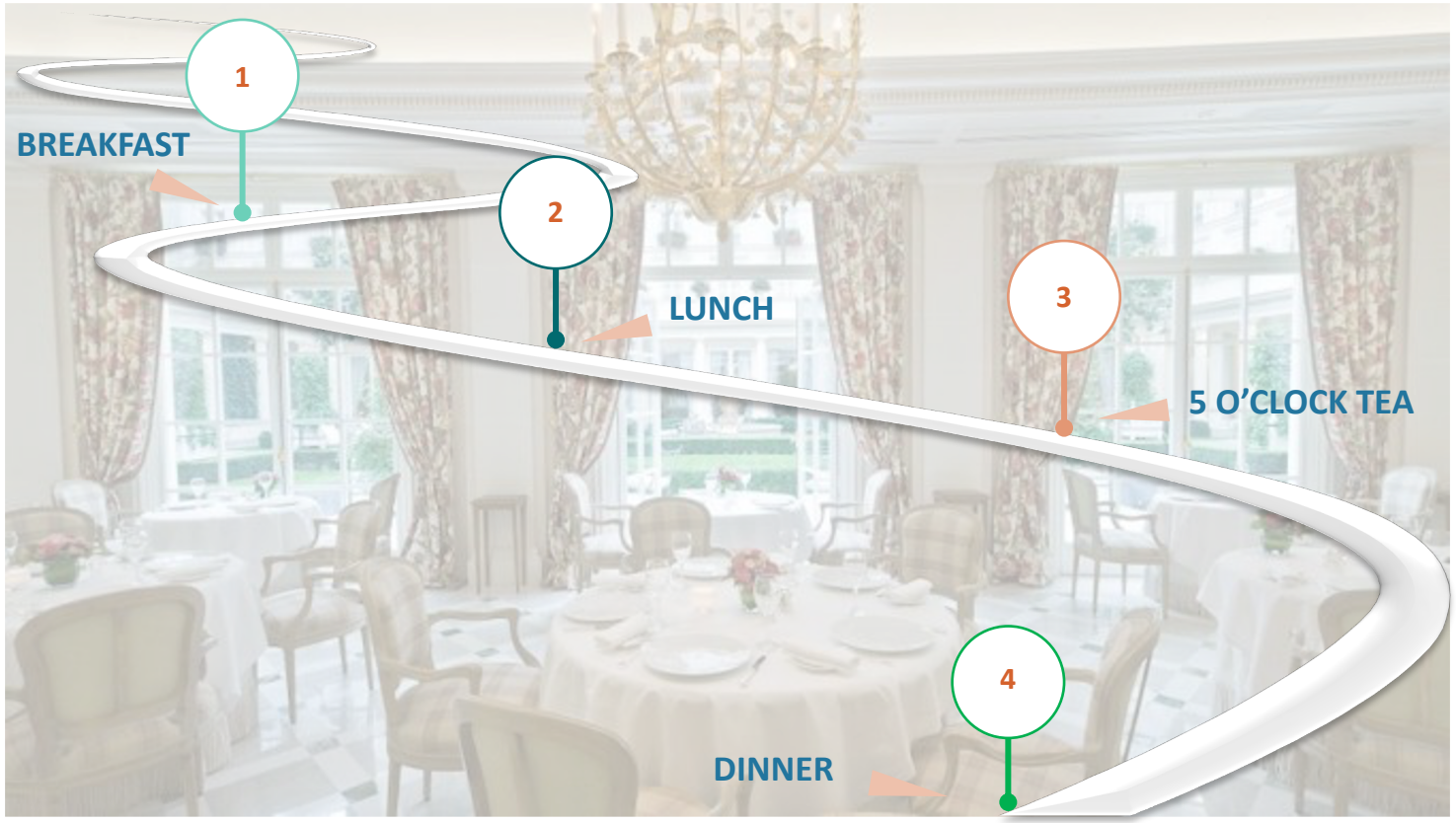
"KPI (Key Performance Indicator) is one of the most important indicators (= something that shows what a situation is like or how it is changing) that show how well an economy, company, project is doing, or how well an employee is working"

Cambridge Dictionary

© Capgemini 2019. All rights reserved |

Today’s objectives

Identify the trends in London restaurant through KPI



© Capgemini 2019. All rights reserved |

Hotel restaurants processes & stages

Food presentation

People like to eat food which is well represented. So, a lot of focus should be made on food representation **to improve hospitality experience.**

Food management

It starts with the **production of food**. Many hotels and restaurants serve naturally organic food. Apart from production, **food transportation and storage** also make part of this division.

Beverage

Apart from general food items, **beverage storage, and representation** also make it into the list of food and catering services.

Restaurant management

Restaurant management is a science which is being taught in colleges these days. With proper skills to manage a restaurant, you can manage food, beverages, and maintain a quality representation of food, so that **customers will come again to eat at your place.**

© Capgemini 2019. All rights reserved |

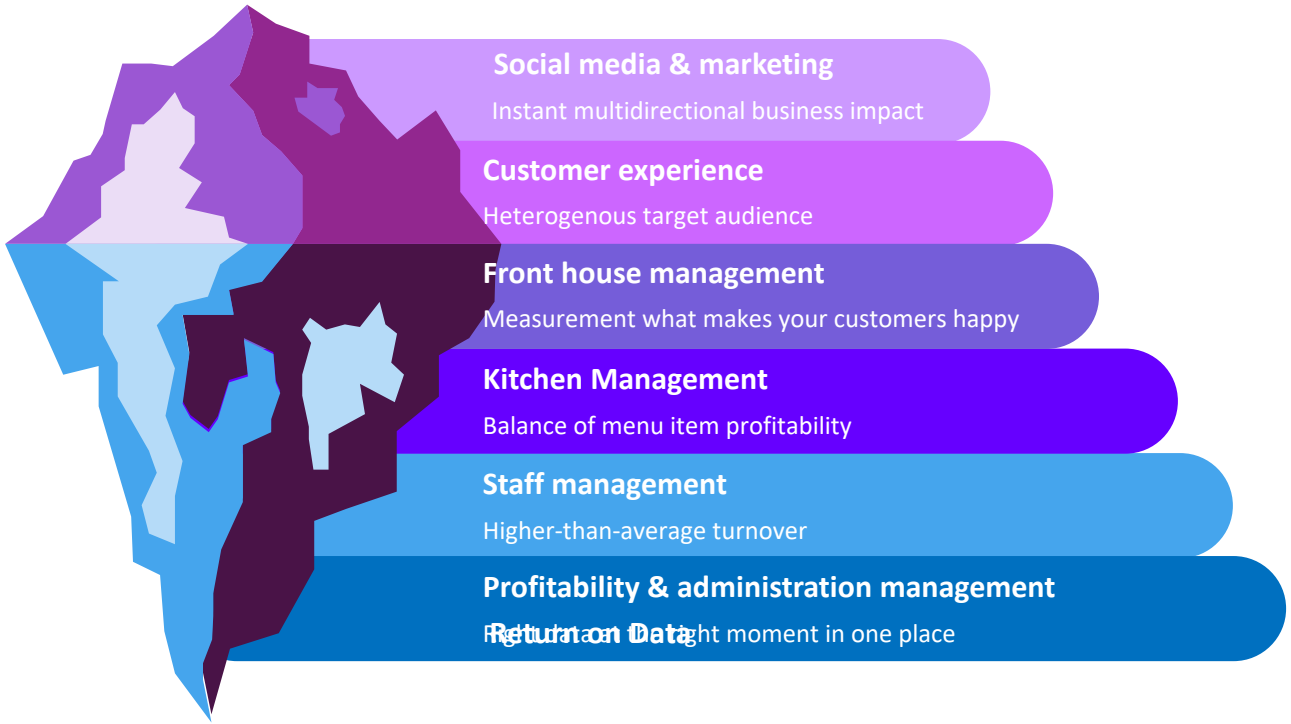
Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. **Deep dive into KPI**
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

KPI: objectives, features & main domains

Restaurant management



KPI: objectives, features

Essentially, a **KPI is a performance measurement** that is used to evaluate **how effectively** your company is **achieving its key strategic goals**.

KPIs must be:

- Applicable
- Generated and measured with actual data
- Properly defined
- Communicated clearly
- Monitored regularly

KPI monitoring **is not an end in itself**, but a mean to **achieve successful restaurant management**.
KPI monitoring should be coherente and be integrated into global strategy.

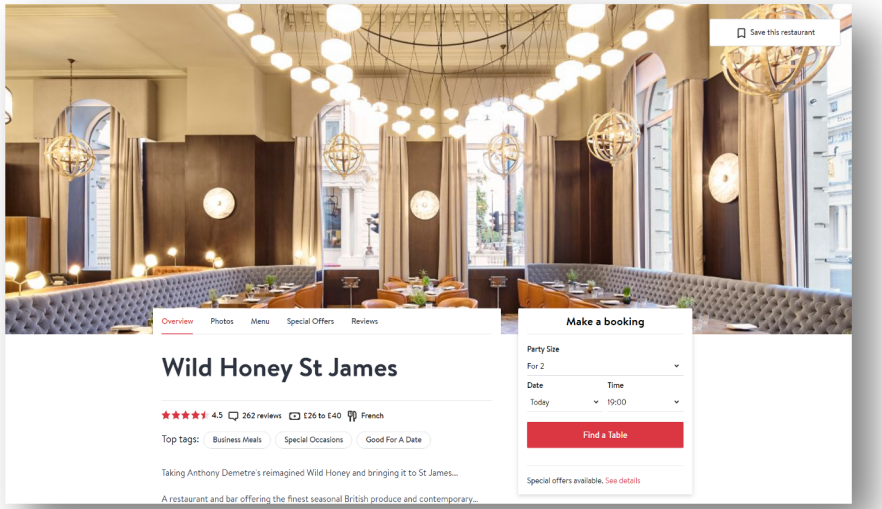
© Capgemini 2019. All rights reserved |

Key social media & Marketing KPI

SOCIAL ENGAGEMENT METRICS

Likes : The number of clicks on your post “Like” buttons.

- **Comments**: The number of comments on your posts.
- **Shares**: The number of times a post has been shared on Facebook, retweeted on Twitter or repinned on Pinterest.
- **Engagement rate**: This measures the level of interaction that happens between your social media account and audience.



By measuring these, you can **get an immediate sense of how your social media marketing efforts are paying off**. A social media reporting tool like Hootsuite can quickly gather this data for you.

WEBSITE TRAFFIC METRICS

Your website should be your number one promoter, tirelessly working for you 24/7 to attract new customers to your restaurant. These KPIs are surefire indicators of how well your website is doing and can easily be uncovered in your Google Analytics data.

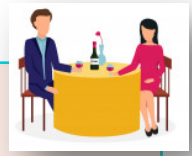
Conversions: With the help of call tracking software and Google Analytics, you can track key conversions for your restaurant, such as bookings over the phone.

© Capgemini 2019. All rights reserved |

Customer experience KPIs

- **Online reviews :** The number of online reviews, as well as the average review scores on Google My Business, Facebook, Tripadvisor and Yelp are also important metrics to monitor.
- **Bookings:** Track and measure your reservations throughout the year to identify patterns, and properly plan for restaurant usage during peak and low times.
- **Customer retention rate :** This KPI holds immense value when you consider the cost of acquiring new customers.

$$\text{CUSTOMER RETENTION RATE} = \frac{\text{THE NUMBER OF CUSTOMERS AT THE END OF A PERIOD} - \text{THE NUMBER OF NEW CUSTOMERS ACQUIRED DURING THAT PERIOD}}{\text{THE NUMBER OF CUSTOMERS AT THE START OF THAT PERIOD}} \times 100$$



TIPS & TRICKS

"Just take a few minutes to read posts. While preparing a trip, any advice is important. The references of www.booking.com and www.TripAdvisor.com are very helpful.

There is one interesting point to mention. I used to compare tourists' comments on any hotel in European countries posted in different languages (English, French and Russian). Results are absolutely amazing! Feedback provided in different languages prioritizes different aspects of customer experience.

You may see that **English-speaking** tourists mostly pay attention to good location, friendly staff. Ready to continue? Comments in **French** contribute to understand whether the breakfast is delicious, the room is noisy. Just one more tip in **Russian**: you may check whether the hotel's attractive description complies with the reality, everything works properly in the room."

Source: https://www.linkedin.com/pulse/preparing-your-multicultural-customer-journey-svetlana-ollivier/

© Capgemini 2019. All rights reserved |

Front-house management KPI

- **Food and beverage sales per guest**: get a handle on which menu items appeal most to guests, and whether time or day impact total spend. This metric can also help you determine which promotions to run – for example, happy hour to increase sales during slow periods.

- **RevPASH**: **Revenue per available seat per hour** is a daily/hourly calculation which takes time into account when calculating how effectively each seat in your restaurant is driving revenue.

$$\text{RevPASH} = \frac{\text{OVERALL REVENUE}}{\text{SEATS AVAILABLE}} \times \text{OPEN HOURS}$$

- **Table turnover rate:** you'll want to ensure your table turnover is optimally timed to allow your customers to fully enjoy their dining experience and new parties aren't waiting around to be seated.

$$\text{TABLE TURNOVER RATE} = \frac{\text{PERIOD OF TIME}}{\text{NUMBER OF TABLES SERVED DURING THAT TIME PERIOD}}$$

- **Average table occupancy :** this metric tells you how many customers on average visited your restaurant during a particular period of time.

$$\text{AVERAGE TABLE OCCUPANCY} = \frac{\text{NUMBER OF OCCUPIED TABLES}}{\text{TOTAL NUMBER OF AVAILABLE TABLES}}$$

© Capgemini 2019. All rights reserved |

Kitchen management KPI

- **Menu item profit and popularity:** pricing items on your menu impacts profitability. A high-profit menu item is great – but not if it produces a statistically insignificant percentage of your sales. **Items with a higher sales velocity but lower profit margin might be driving all your revenues**. Get a true handle on the profits of each menu item to help drive menu engineering and promotion decisions.

- **Production time per dish:** Knowing how long it takes to produce each dish is key, as it can help you determine the value of each single dish based on turnaround time and expectations. **Track popular (or star) items** as well dishes with low sales volume, and **consider ways to shorten production time and offer faster service**.

- **Food wasted per food purchased:** food waste is a huge concern for restaurateurs worldwide. **Monitoring food waste enables you to improve your demand forecasting**, forces you to reconsider how and where you procure your foods, and leads to better management and storage your food stock. It can also help you determine if there are better methods to prepare your food, and if you need to reconsider your portion sizes or serving methods.

TIPS & TRICKS

Is there any correlation between **napkins' choice (cost of linen) & sales increase per guest**?
In a high style restaurant, well known for its fantastic dishes, elegant design, great services, high prices, etc. the only stuff differs from everything: guest napkins *are not* of 100% linen, containing a few %tile of synthetic fabrics. It is not obvious while touching. However, napkins slide down sometimes. Why not to change these sliding napkins?
It is a matter of sales: every time a waiter assists the client to get the napkin back or to replace, it is **an opportunity** to replenish his … glass of wine, water etc.

© Capgemini 2019. All rights reserved |

Staff management KPI

- **Sales per employee per hour**: looks at how much revenue each employee generates.

- **Employee turnover:** expenses associated with finding, hiring and training new employees, a higher-than-average employee turnover rate can prove to be an expensive problem for your restaurant.

$$\text{EMPLOYEE TURNOVER RATE} = \frac{\text{NUMBER OF EMPLOYEES WHO LEFT DURING THE TIME PERIOD}}{\text{AVERAGE NUMBER OF EMPLOYEES}} \times 100$$

TIPS & TRICKS

When thinking on equip the team with tools/tablets/apps, think twice. In a high-turnover environment, you may not invest too much time to train new comers…

© Capgemini 2019. All rights reserved |

Administration et profitability management KPI

- **Cash flow** : It takes many things to start and run a successful restaurant – a passion for good food, determination and the right staff, but the best-laid plans will go nowhere without **one key ingredient.** And **that's cash**.

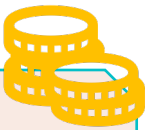
CASH FLOW = BEGINNING CASH – ENDING CASH

- **Cost of goods** sold (or COGS) : helps you measure the amount of money that goes into buying supplies and good ingredients for your menu items. Before you can even think about calculating your restaurant's profit, you need to know your cost of goods sold. If you track only one thing in your kitchen, make it this one. Cost of goods sold is **likely the biggest expense for any restaurant**; knowing where you can reduce the costs is a key factor in increasing profitability.

COGS = BEGINNING INVENTORY + PURCHASES DURING THE PERIOD – ENDING INVENTORY

- **Return on investment (ROI)** : As any business owner knows, this metric is essential to determining profit.

$$\text{ROI} = \frac{\text{NET VALUE OF CASH FLOWS}}{\text{COST OF INVESTMENT OR CAPITAL EMPLOYED}} \times 100$$



TIPS & TRICKS:
The proper way to calculate a return is using the "cash flow method", it should meet **at least 15% ROI minimum in your first year**, and you are in a good business if you could **reach 20 to 25% annual profit vs capital.** Good restaurant business require sustainability over 5 years. In fact **only 1 of 100 restaurants could reach over 25% profit vs capital.**

https://www.quora.com/What-is-a-good-ROI-on-a-new-restaurant

© Capgemini 2019. All rights reserved

Top tips for 2020

BE SENSIBLE

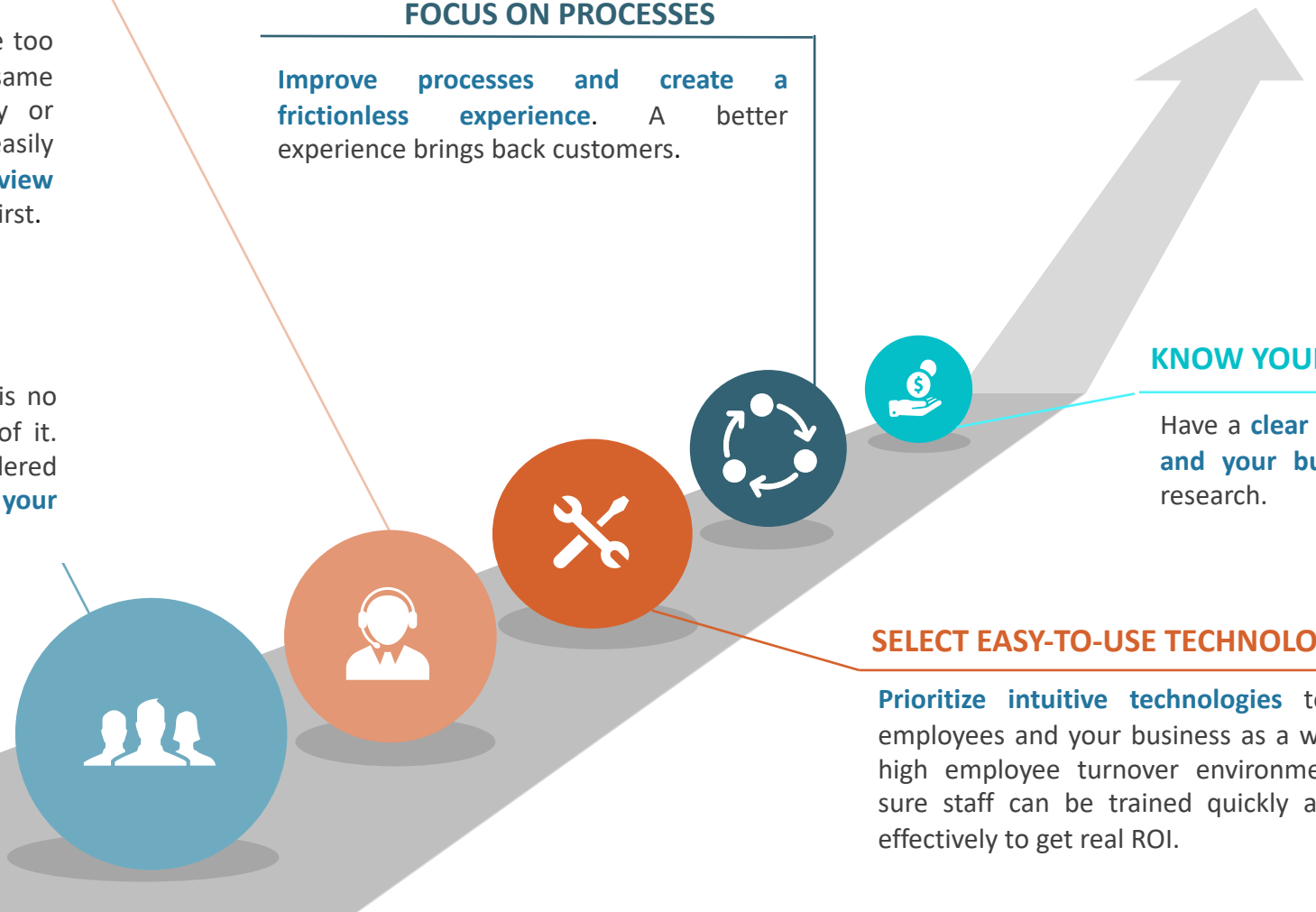
In a challenging climate it's easy to be too reserved and stagnate, but at the same time don't get caught up in shiny or distracting technology that can't be easily implemented into your business. **Review business, customers' and your needs** first.

KNOW YOUR CUSTOMER

Know who are you targeting. There is no point using technology for the sake of it. Your decision should be well considered and depend entirely on the needs of **your customer base**.

FOCUS ON PROCESSES

Improve processes and create a frictionless experience. A better experience brings back customers.



KNOW YOUR NEEDS AND BUDGET

Have a **clear understanding of your needs and your budget** before you start your research.

SELECT EASY-TO-USE TECHNOLOGIES

Prioritize intuitive technologies to benefit employees and your business as a whole. In a high employee turnover environment, make sure staff can be trained quickly and use it effectively to get real ROI.

© Capgemini 2019. All rights reserved |

Brainstorm

20’

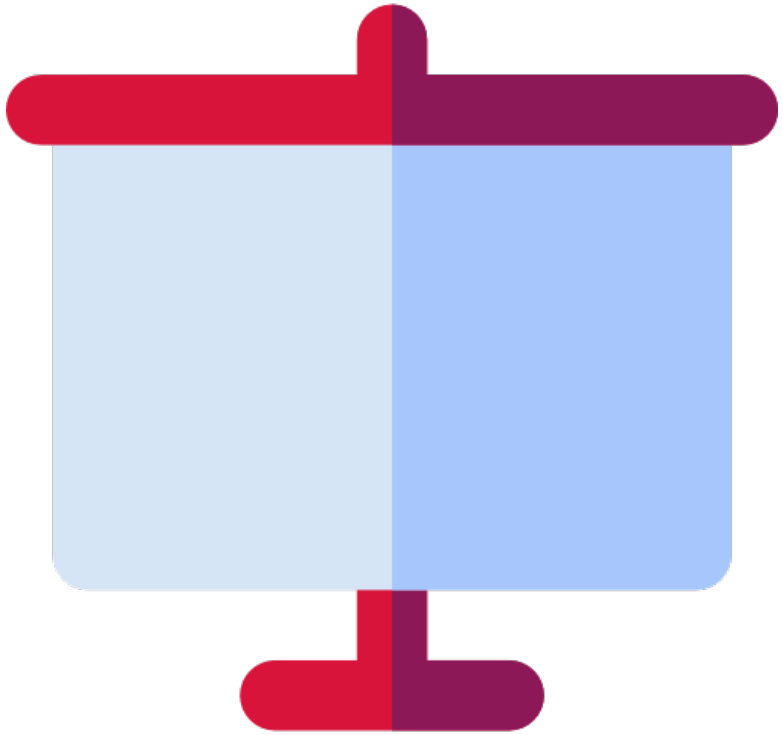
Select and identify your KPIs to identify trends

© Capgemini 2019. All rights reserved |

Restitution

10’

What KPIs did you find ? Which one seems relevant ?



© Capgemini 2019. All rights reserved |

Example of a project presentation

OBJECTIVES

- Define KPI to identify trends and develop a new business model for London restaurant
- Identify key stakeholders and define RACI & deliverables

PRE-REQUISITES

- Define the **scope** of the project
- Determine **objectives & stakes** (new services, cost killing, new target audience…)
- Check **strategic constraints to integrate** into the project (costs, target operational business models)
- Validate **key client expectations** (deliverables, etc)?

CONDITIONS & DURATION

- **Key success factors:** sponsor et stakeholders availability, access to internal information
- **Duration:** 2 weeks / per sprint

APPROACH

1. Interviews
2. Final brainstorming
3. Gap analysis
4. Prototyping
5. Go-no Go
6. Development
7. Change management
8. Go live

METHODS

- AGILE
- Interviews, 2-5 persons maximum
- Design Thinking
- Brainstorm workshops, 10 persons maximum
- Framework PESTEL
- Porter 5 Forces
- SWOT
- Canvas business model

DELIVERABLES

- Reference framework
- Diagnosis (internal & external)
- Planning

© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. **Data pipeline - Scrapping restitution**
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

Presentation of the achievements so far

5’

Restitution of data scraped using Scrapy

- Pain Points
- Main takeaways
- How to improve ?

© Capgemini 2019. All rights reserved |

Data pipeline

Data Collection

Data Cleaning

Word Embedding

Topic Extraction

Sentiment Analysis

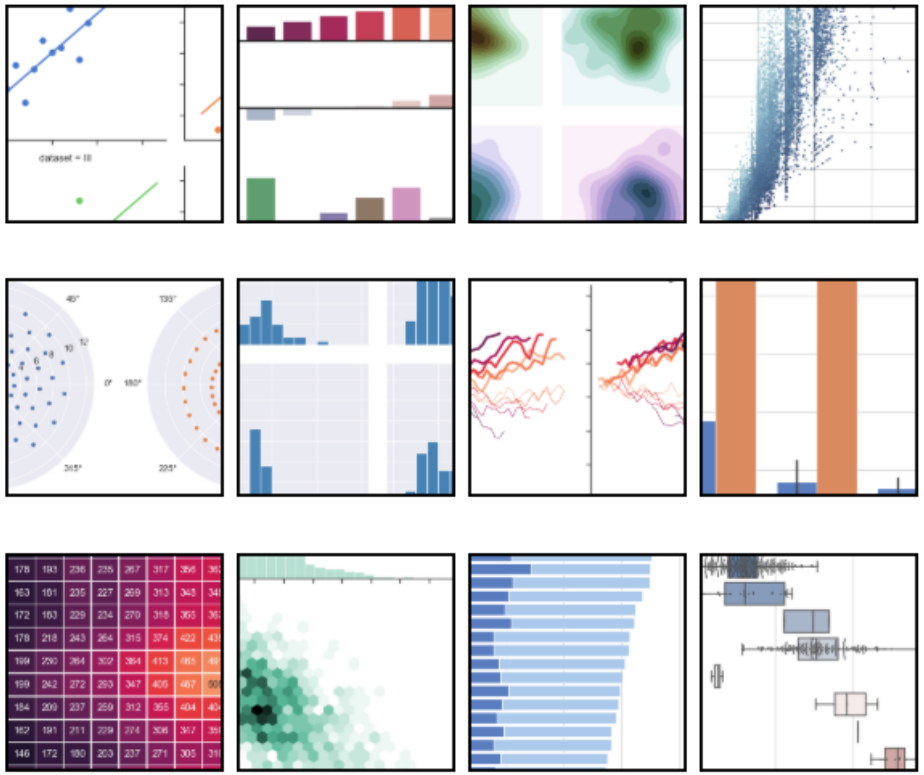
© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. **Data Cleaning : introduction & basic processing**
6. Stemming & Lemmatization
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

Exploratory Data Analysis



EDA is the process of using **statistical** tools and ideas to **examine** data in order to describe their main features

Exploring data begins with :

- **Examining each variable** by itself. Then move on to study the **relationships** among the variables.
  - Python methods like `df.describe()` and `df.info()`
  - Statistical tools
  - **Pandas-profiling**

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

© Capgemini 2019. All rights reserved |

EDA on Jupyter Notebook

© Capgemini 2019. All rights reserved |

What is Natural Language Processing ?

Field of study focused on making sense of human written text

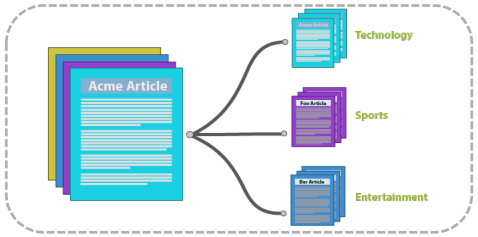
It includes both linguistics and machine learning in order to make machines learn how to process large amount of natural language data

You will learn the basics of NLP

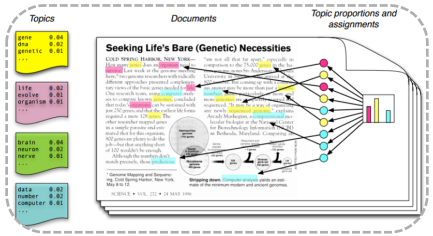
- Word Embedding
- Topic Extraction
- Sentiment Analysis

NLP applications include

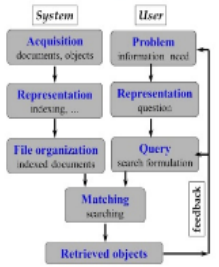
Classify Documents
(supervised learning)



Understanding key topics & trends
(unsupervised learning)



Find relevant information
(information retrieval)



© Capgemini 2019. All rights reserved |

Pre-processing: Introduction

What processes should we apply on our data before anything 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 location 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 offre 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 leurs explique que je veux pas annuler mais a titre commercial et car je suis cliente fidèle une compensation financière. 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 into 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 worth 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 that she ought to have wondered at this , but at the time it all"

© Capgemini 2019. All rights reserved |

What are the preprocessing steps to get clean text data ?

Today's next steps

Structure

- **Transforming the text into a "corpus"**
- **Tokenizing**

Special Characters

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

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.

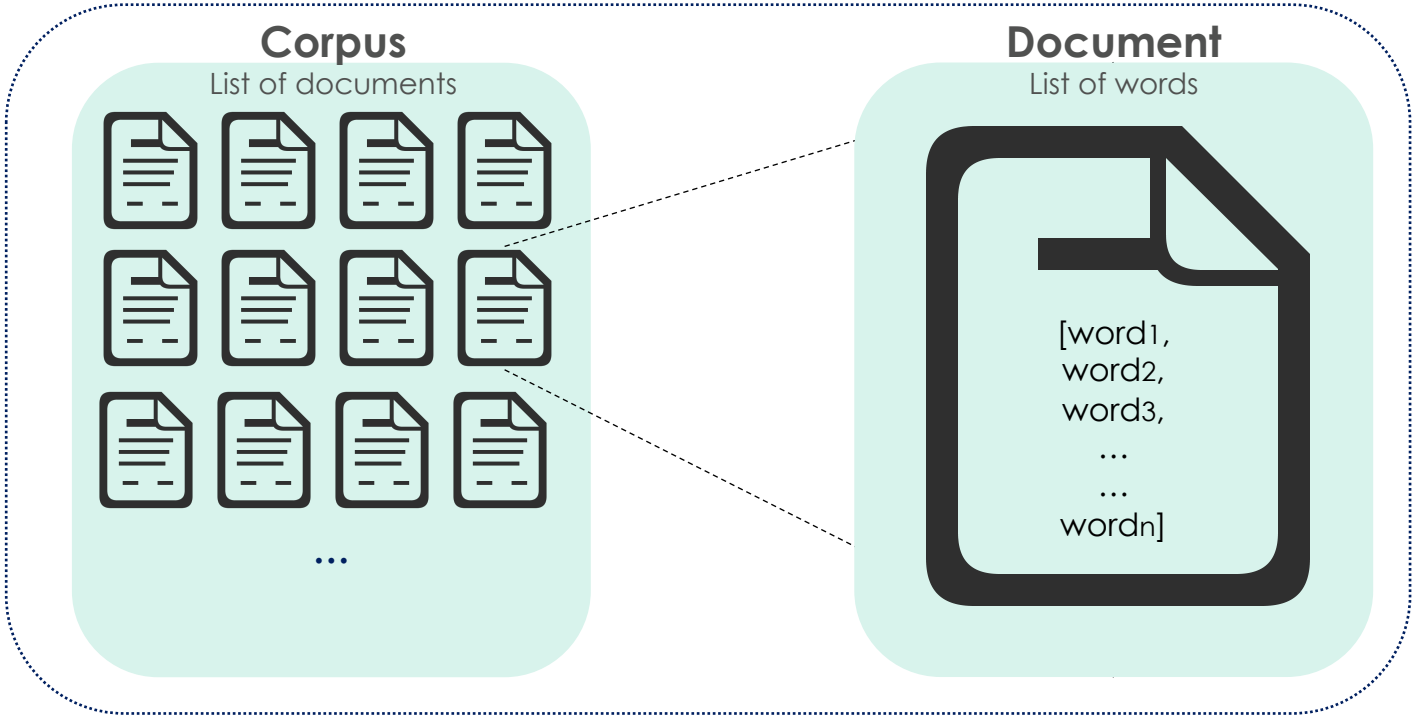
© Capgemini 2019. All rights reserved |

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** from which we perform our analysis

- Within each corpus we will have separate documents, which might be articles, stories, or book volumes, each document is treated as a separate entity or record



© Capgemini 2019. All rights reserved |

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

- We split the document 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"]

© Capgemini 2019. All rights reserved |

Cleaning on Jupyter Notebook

© Capgemini 2019. All rights reserved |

Hands-on 1: Pre-processing your data

30’

Load, explore and clean your dataset

© Capgemini 2019. All rights reserved |

Break

10’

Feel free to help yourself !

© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. **Stemming & Lemmatization**
7. Text representation
8. Summary of the session

© Capgemini 2019. All rights reserved |

One possibility to clean text : Orthographic correction

Detection of words with low frequence of apparition and removing them

Making a comparaison with a dictionnnary

Using specific algorithms like Levenshtein Distance Spelling Correction

Those methods are often very time and memory consuming

© Capgemini 2019. All rights reserved |

Inflected language

In grammar, inflection is **the modification of a word to express different grammatical categories** such as tense, case, voice, aspect, person, number, and gender. An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change.

- “person”
- “persons”
- “person’s”
- “persons’”

→ **person**

- “processing”
- “processes”
- “processed”

→ **process**

- “university”
- “universe”

→ **univers**

- “am”
- “are”
- “is”

→ **be**

© Capgemini 2019. All rights reserved |

Stemming

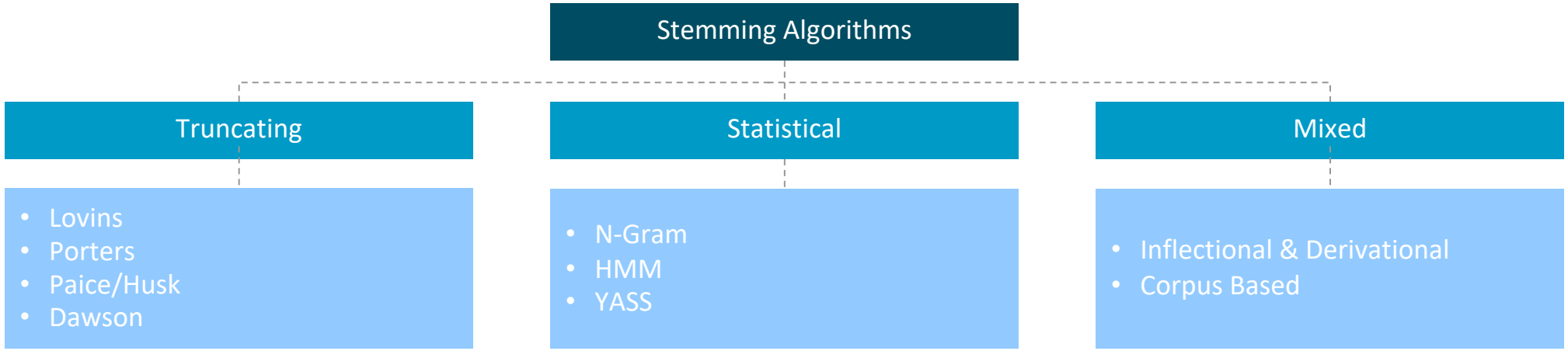
- **Stemming is the process of reducing inflection in words to their root** forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.
- By removing inflectional form of a word we focus on the **meaning**
- “Rude” process in the sense that it could remove the end of a word even if it is not an inflectional form

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

Porter Stemmer

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

- Stemming algorithms can be classified into three groups: **truncating methods**, statistical methods, and mixed methods:



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

© Capgemini 2019. All rights reserved |

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. »

Can you see some limits of lemmatization from this example?

© Capgemini 2019. All rights reserved |

Trade-off between stemming & lemmatization

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

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

© Capgemini 2019. All rights reserved |

Let’s do it with Python

© Capgemini 2019. All rights reserved |

Hands-on 2: Stemming & Lemmatization

10’

Get ready for starting basic NLP : Stemming and lemmatization !

© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. **Text representation**
8. Summary of the session

© Capgemini 2019. All rights reserved |

Bag-of-words (1/2)

Principle

- Consider a corpus of documents, each **document** is a text
- Each document contains a certain number of **words** for which we can do **tokenization**
- Each word, or group of words will be a **token** (monogram, bigram, …)
- **Bag-of-word**: 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 individually, **whatever its place in the sentence.**

© Capgemini 2019. All rights reserved |

Bag-of-words (2/2)

Description

- To get the most important topics in a collection of documents, a natural thing would be to look at the words and **compare their occurrences**.
- The **Document Term Matrix** synthetizes the words' occurrences in a collection 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 !'

→ [1,0,0,0,0,1,1,0,0 ...]

Document Term Matrix

	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	

© Capgemini 2019. All rights reserved |

TF-IDF

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

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

Term frequency

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

Inverse document frequency

A measure of how much information the word provides, i.e., if it's common or rare across all documents. It diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Formula of Inverse Document Frequency:

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

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

With n the total number of documents, D the set of all documents, 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.

Remarks

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

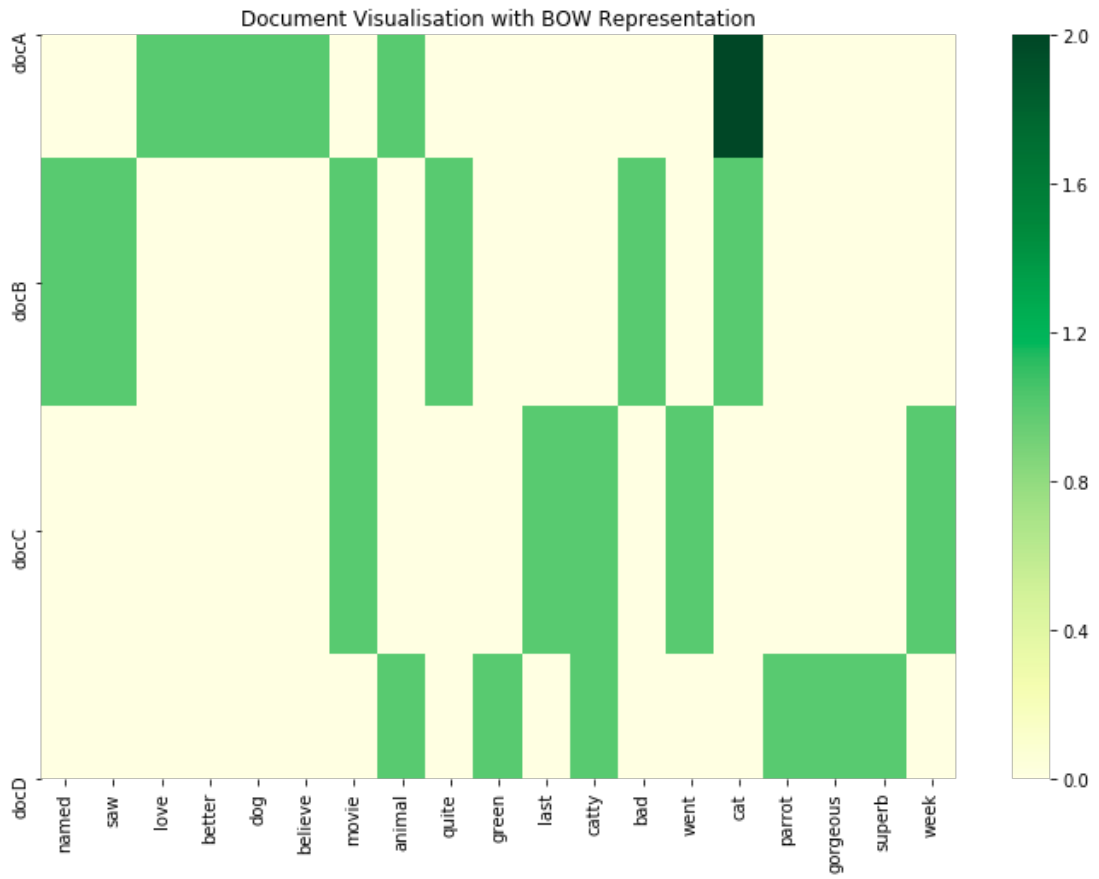
Choices for the value of tf (Wikipédia)

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

© Capgemini 2019. All rights reserved |

BOW vs TF-IDF

BOW

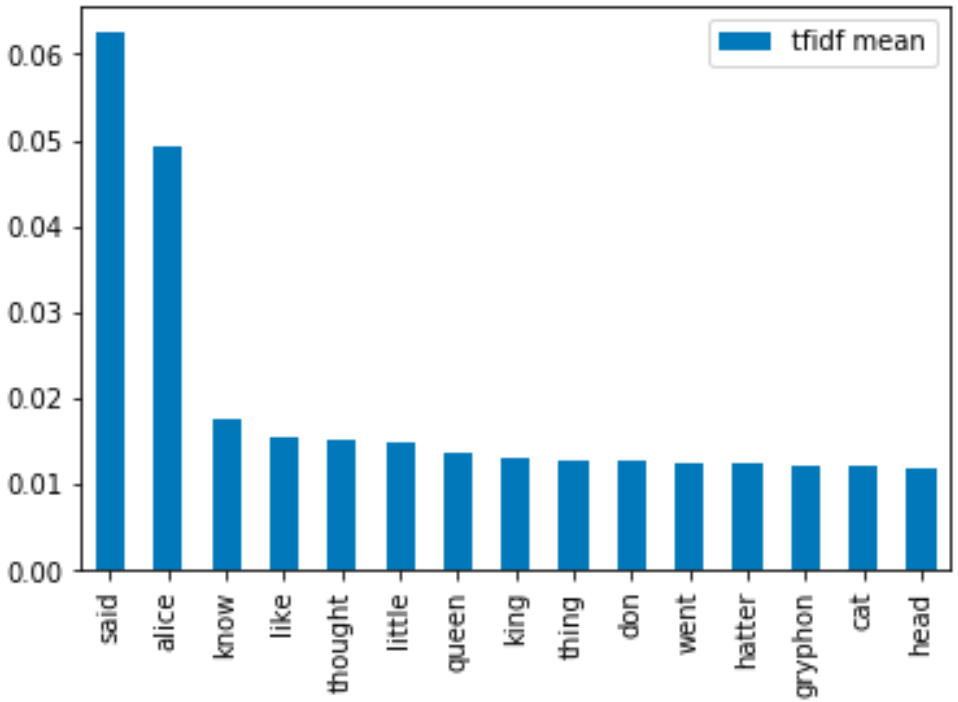


TFIDF

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

© Capgemini 2019. All rights reserved |

Interpret the TF-IDF matrix



How can we improve our analysis ?

© Capgemini 2019. All rights reserved |

Let’s do it with Python !

© Capgemini 2019. All rights reserved |

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 the comments of your scrapped reviews ?

© Capgemini 2019. All rights reserved |

Agenda

1. SWOT analysis restitution
2. Hospitality industry & restaurants KPI definition
3. Deep dive into KPI
4. Data pipeline - Scrapping restitution
5. Data Cleaning : introduction & basic processing
6. Stemming & Lemmatization
7. Text representation
8. **Summary of the session**

© Capgemini 2019. All rights reserved |

Today we learnt

Cleaning steps

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

Approach & organization of a data science consulting project

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

© Capgemini 2019. All rights reserved |

Work for next week

Instructions for Sunday 23rd !

- Be sure that you have applied every steps of text processing to your reviews
- Create a TF-IDF Matrix with all the reviews 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 lemmatize text.
  - One that takes a corpus (or a dataframe text column) and creates a Wordcloud from it.

- Create sets of KPI (both evident and shadow) you may collect from available open information
- The idea is to anticipate what you'll present on the fourth session for the client meeting simulation

Our email addresses :

- francois.lemeille@capgemini.com
- ismail.mebsout@capgemini.com
- Thibaud.Lamothe@Capgemini.com

- Feel free to contact us by email or Slack !

© Capgemini 2019. All rights reserved |

Course evaluation

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

© Capgemini 2019. All rights reserved |

Thank you for your attention

See you next time @147

GOODBYE !

February 10th, 2020