

Text mining

Statistiques textuelles - LZML041

Séance 1 : Introduction au Text mining

Chargé de cours : Loubna Serrar

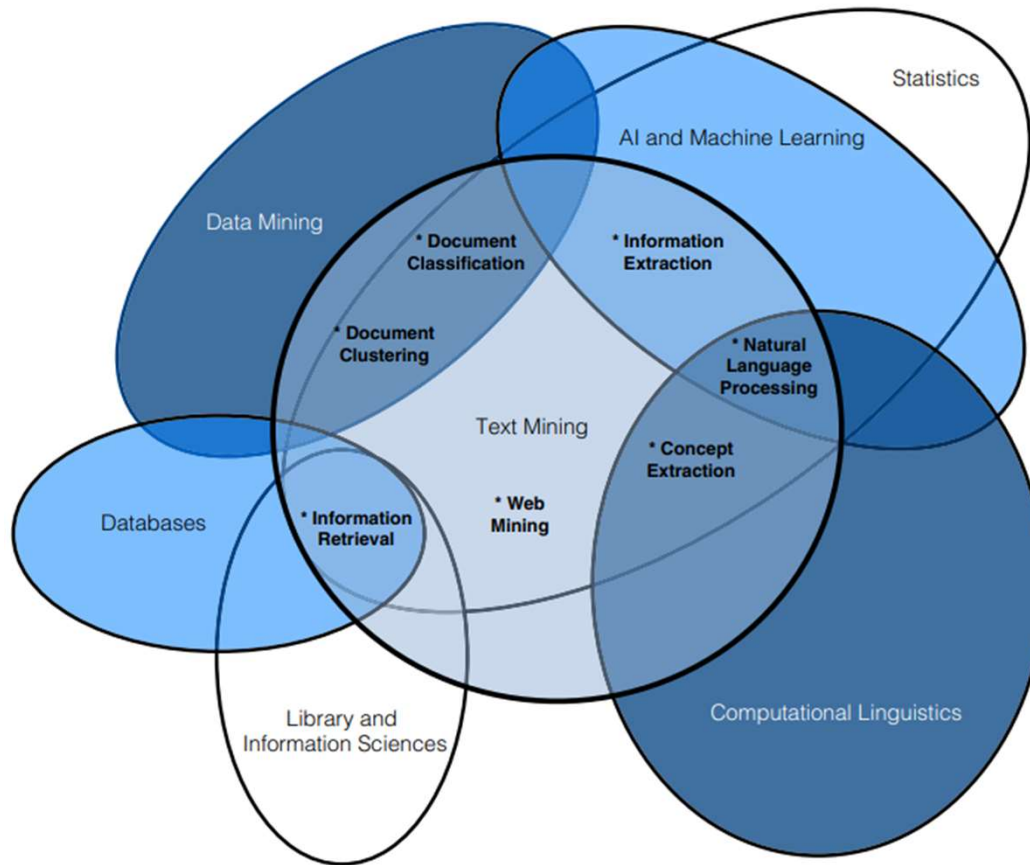
ANNÉE UNIVERSITAIRE 2022-2023

Description du cours

- ▶ 12 séances de 2h pour **s'initier au Text mining**
- ▶ Un mix de théorie et de pratique avec **l'outil Python/Jupyterlab**
- ▶ **Évaluation en controle continu** avec des tests courts toutes les 3 séances et un travail final à rendre en mai (50/50)
- ▶ **Plan des premières séances:**
 1. **Introduction générale**
 2. **pre-processing techniques** (Cleaning? Segmentation, tokenization, part of speech tagging, syntactic parser, data annotation...)
 3. **Text to numbers** (Bag of words, frequency, TF-IDF, n-gramm model...)

What is Text mining ?

- **Text mining**, also called Text Data Mining or Knowledge-Discovery in Text (KDT), is the process of deriving meaningful insights and patterns from a large collection of texts.



A brief History

- ▶ One of the earliest examples (Frost 1976) was the library catalog attributed to Thomas Hyde (1674) for the Bodelian Library at the University of Oxford. In the mid-1800s, a library catalog card was introduced by Melvil Dewey (2012) containing bibliographic information, including author's name, book title, abstract, and index.
- ▶ Almost a century later, Hans Peter Luhn Luhn (1958) demonstrated an early IBM 701 computer indexing texts using its KWIC (Key Word in Context) algorithm to generate document abstracts.

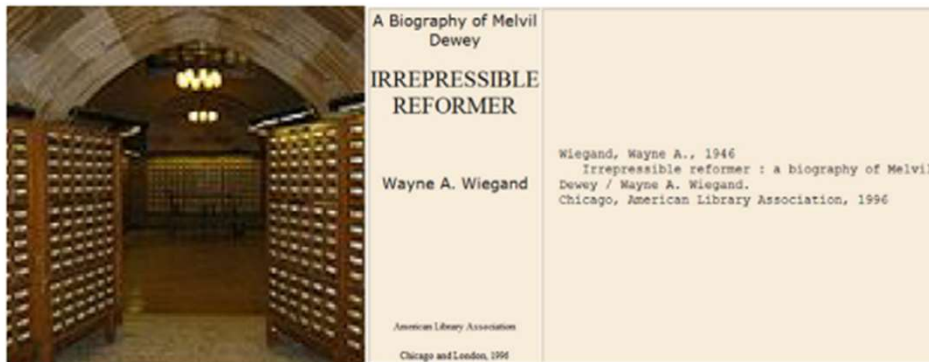


FIGURE 1.1

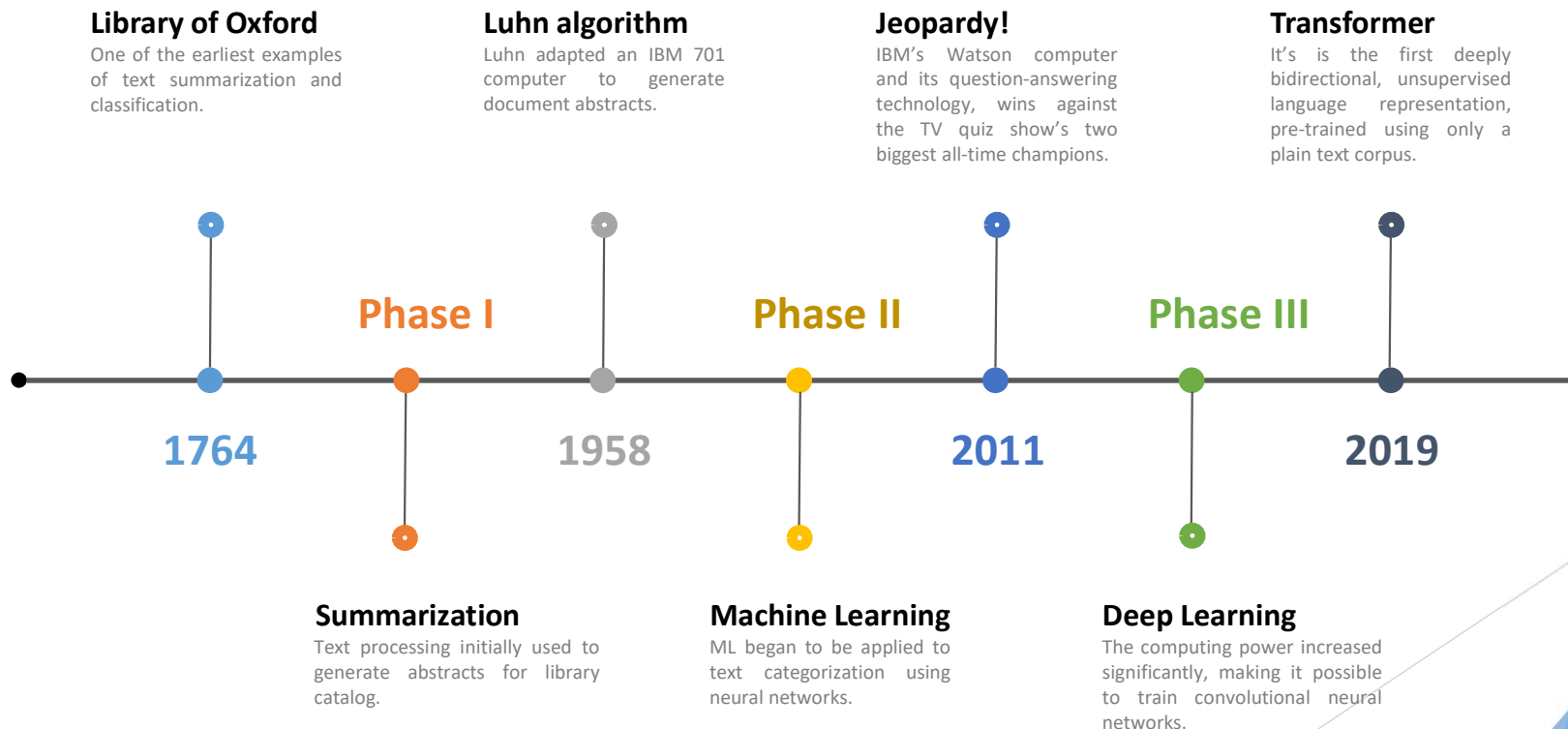
The library card catalog at Yale University and an index card. Source: http://commons.wikimedia.org/wiki/File:Yale_card_

IBM 701 Electronic analytical control unit



What is Text mining today?

- ▶ Today, Text mining is a broad umbrella term describing a range of technologies for analysing and processing text data. We might define modern text mining as *the process of extracting meaningful insights from unstructured text with the application of advanced analytical techniques stemming from statistics, machine learning and linguistics.*



Why is Text mining useful today ?

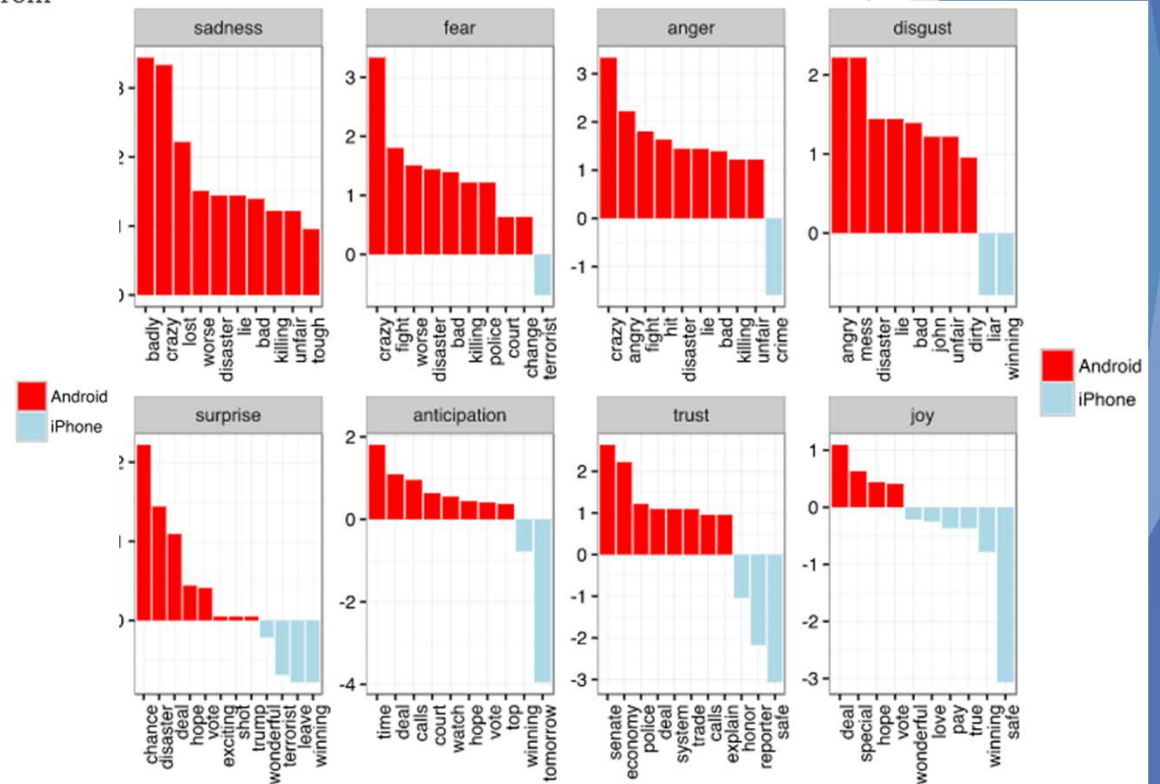
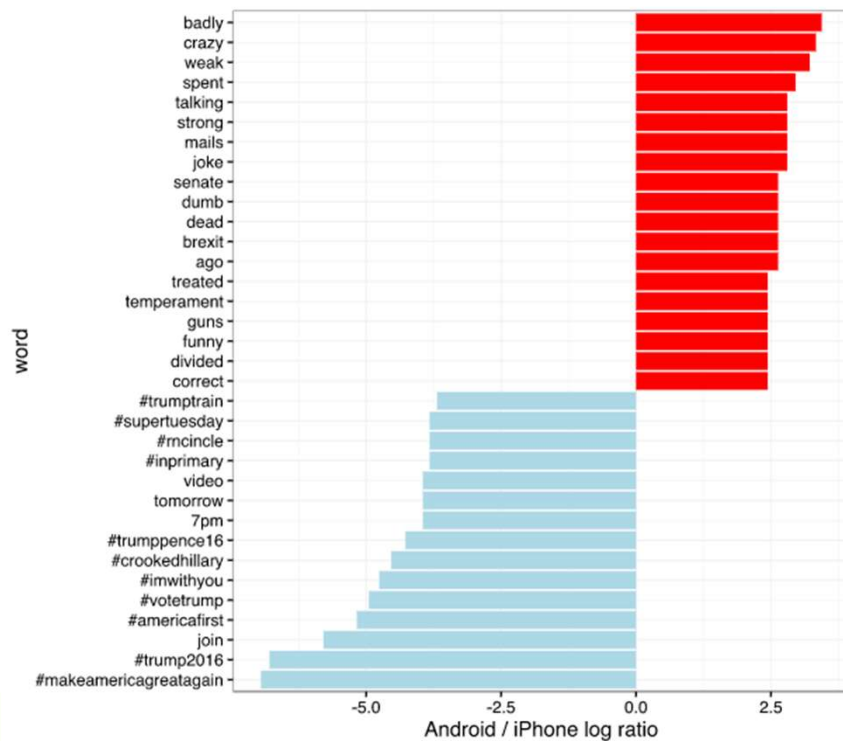
- ▶ In general, most information available in the “real world” exists as written or recorded words.
- ▶ **The Explosion of Text Stored in Electronic Format !**



Case Study - Trump's Tweets

- ▶ One of many studies : <http://varianceexplained.org/r/trump-tweets/>
- ▶ lots of words annotated as negative sentiments (with a few exceptions like “crime” and “terrorist”) are more common in Trump’s Android tweets than the campaign’s iPhone tweets.

Which are the words most likely to be from Android and most likely from iPhone?



OUPOCO - La Boîte à poésie

- Ce travail est le fruit d'un travail de recherche mené au LATTICE (CNRS – École Normale Supérieure/PSL – Sorbonne Nouvelle), et inspiré du **livre Cent mille milliards de poèmes** publié en 1961 par Raymond Queneau.
- [Oupoco](https://oupoco.org/fr/le-projet/index.html) intègre plusieurs milliers de poèmes d'auteurs du 19e siècle, et permet ensuite de produire de nouveaux poèmes à la façon de Queneau, en respectant les contraintes essentielles de la métrique française.

Site : <https://oupoco.org/fr/le-projet/index.html>

vidéo : <https://savoirs.ens.fr/expose.php?id=3929>

OUPOCO

Le projet Générateur de poème Sonnet féminin Poètes contributeurs La Boîte à poésie

VISUALISATION DE L'ENSEMBLE DES FEMMES POÈTES CONTRIBUTEURICES

La base de sonnets contient 439 sonnets écrits par des femmes (soit 107 auteures).



La Lune, lentement, s'élève à l'horizon...

Retentissent, en vain, sur les monts, dans les bois,

Jours divins, émus d'un vague frisson,

Mais dans mon coeur reste la foi.

Pour entendre la voix que leurs cris couvriront

N'obéissent jamais un instant à sa loi;

À l'heure de l'aurore aux roses pavillons,

Couronne d'empereur, diadèmes de rois.

Mon sang jaillit, semant des rubis, et le flux

Puis, à Clotaire deux sa liberté vendue,

Déclin, rides, pâleur, fausses dents, faux cheveux!!

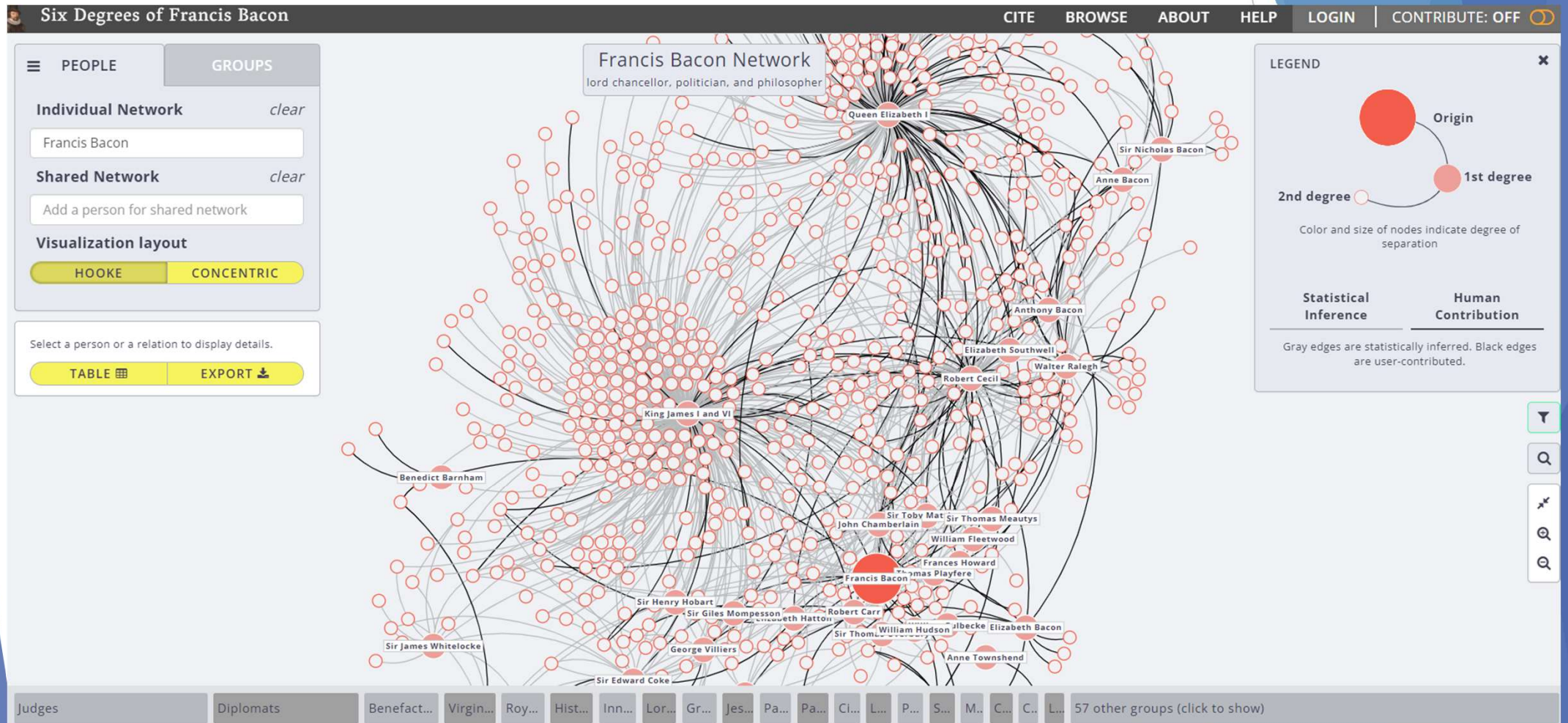
Le prestige divin, l'extase qui l'enlève,

Et l'impie, atterré, restant silencieux,

L'éclair d'amour jailli d'une heure brève.

Six Degrees of Francis Bacon

Site : <http://www.sixdegreesoffrancisbacon.com>



ChatGPT - <https://chat.openai.com/chat>



https://www.youtube.com/watch?v=8klxMUrdLQw&ab_channel=Numerama

Business applications examples

- In the business world, text mining techniques are used to reveal insights, patterns and trends from large volumes of unstructured data:

1. Social Media Data Analysis
2. Contextual Advertising
3. Content Enrichment
4. Spam Filtering
5. Cybercrime Prevention
6. Enhanced Customer Service
7. Streamlined Claims Investigation
8. Risk Management
9. Knowledge Management
10. Business Intelligence



Manufacturers

- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products



Government

- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy



Financial Institutions

- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation



Retail

- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media



Legal

- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications



Healthcare

- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data



Telecommunications

- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments



Life Sciences

- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials



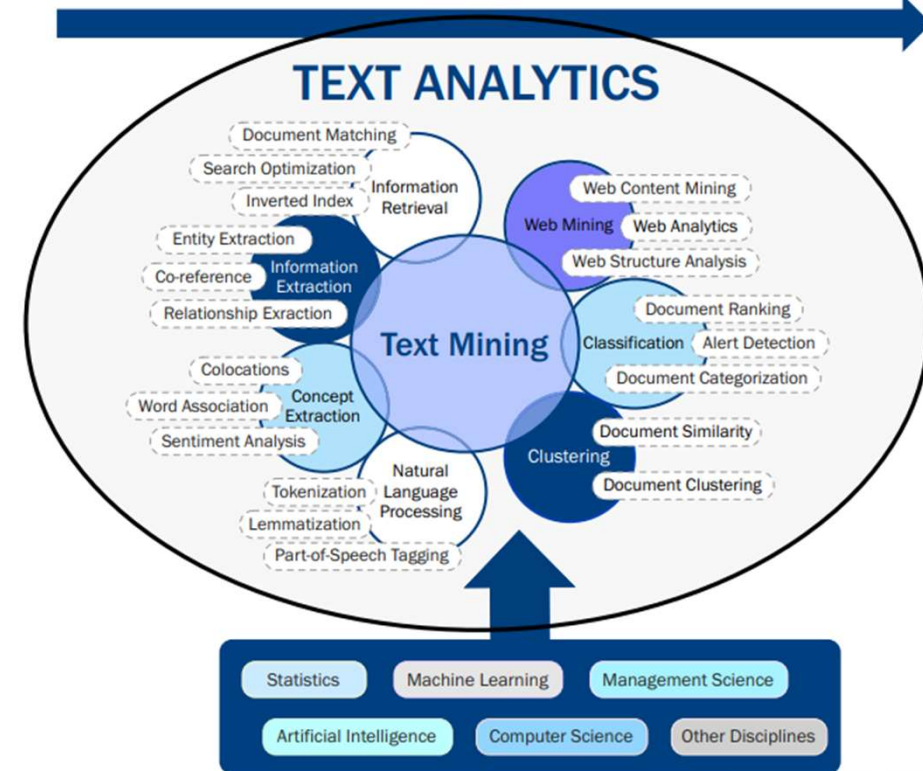
Insurance

- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos

Text Mining practices

1. **Information Retrieval (IR):** use algorithms based on a predefined set of queries to retrieve documents that match a keyword search. IR is commonly used in library catalog systems and popular search engines, like Google;
2. **Document Clustering:** aim at grouping similar (*lexically or semantically*) paragraphs, or documents, together. Clustering techniques are broadly used in search results classification, marketing research, taxonomy definition, and in identifying fake news or spam emails;
3. **Document Classification:** is labelling - or tagging - documents using pre-defined categories, depending on their content. Text classification finds wide application in improving browsing websites (SEO), filtering emails, negative content or hate speech on social media...;
4. **Web Mining:** adapt text mining techniques to the new format of information on the Web (blogs, emails, tweets, ...), which offers opportunities and challenges different from standard text;
5. **Information Extraction (IE):** aim at discovering structured data from unstructured data, which often requires specialized algorithms, considerable tuning, and may require subject matter expertise. A typical application of IE is the recognition of entity names, for example: people, product, or location names, sentiment, opinion, or terminology extraction;
6. **Natural Language Processing:** use machine learning algorithms to perform different tasks such as **summarization**, **text categorization**, **sentiment analysis** or **Part-of-Speech Tagging** (tag each word as noun, verb, adjective, adverb, etc.);
7. **Concept Extraction:** is the trickiest component as a concept could represent multiple underlying terms depending on the text collection, and the set of linguistic resources available. It requires usually a combination of human expertise and machine intelligence to try to determine text meaning within context.



Supervised vs. unsupervised

- ▶ **Supervised Learning:** The algorithms learn a classifier or infer a function from the labeled training dataset in order to perform predictions on unseen data.
- ▶ **Unsupervised Learning:** also called learning without a teacher. So unlike supervised learning, only input data is provided to the model, to find the hidden patterns and useful insights from this unknown dataset. Unsupervised Learning can be further classified in:
 - Clustering: looks for text data that are similar to each other and find natural clusters (groups) if they exist.
 - Associations: is a rule-based method for finding relationships between variables in a given dataset.

Unsupervised Learning

- No label data provided
- Finds hidden structure in unlabeled data
- Uses techniques such as clustering and dimensionality reduction
- Number of classes is not known
- Higher risk of inaccurate results as no prior knowledge is provided

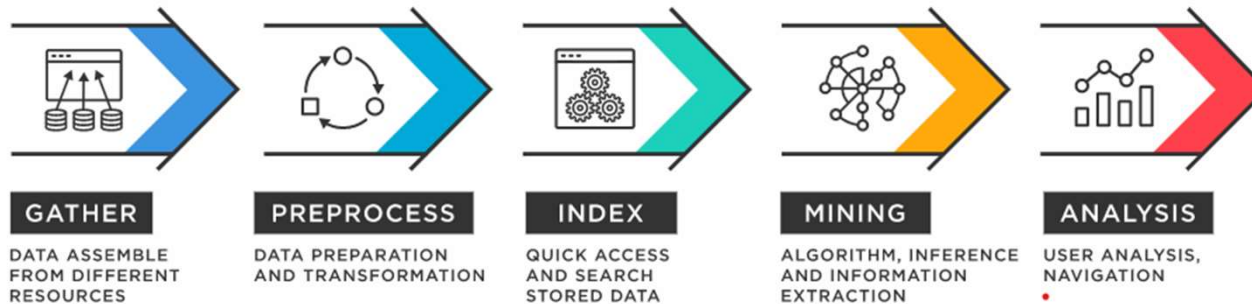
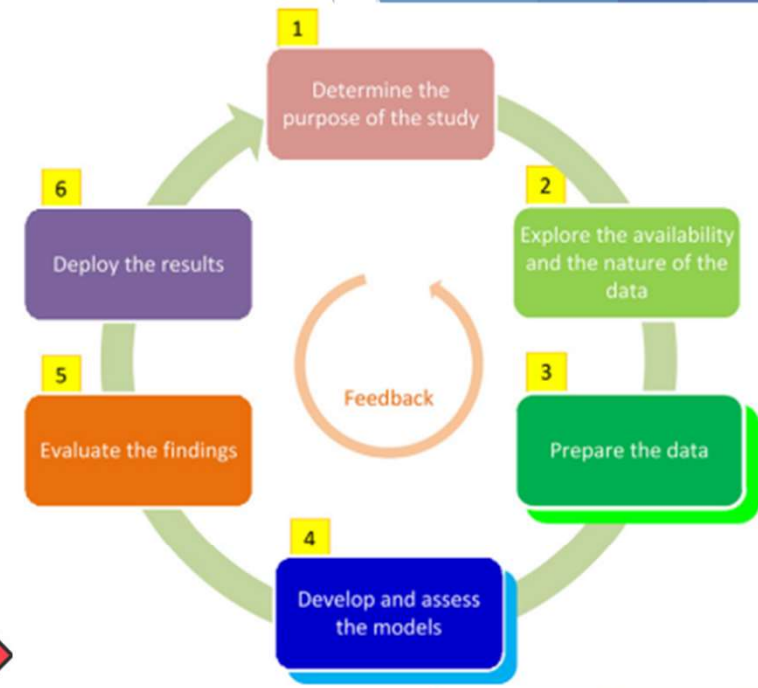
Supervised Learning

- Labels provided
- Finds patterns in existing structure to make predictions
- Uses techniques such as regression and classification
- Number of classes is known
- Classifying big data can be a real challenge in Supervised Learning

How does Text Mining work?

► The six phases that naturally describes the data science life cycle of a data mining project:

1. Determining the purpose of the study: what is the question we want to answer?
2. Data understanding: what data do we have / need?
3. Data preparation: how do we organize the data for modeling?
4. Modeling: what modeling techniques should we apply?
5. Evaluation: which model best meets the study objectives?
6. Deployment: how do users access the results?








Text Mining project in practice ?

- ▶ The text mining process contains the following steps :
 1. Build the corpus : In the first step, the text documents are collected, which can have several formats.
 2. Pre-Processing: Text Cleanup, stop words, Tokenization, Stemming....
 3. Transformation: from text to numbers, bag of words
 4. Applying a technique: Word frequency, TF-IDF, clustering, sentiment analysis
 5. Evaluate: did we completed the task ? Keyword Extraction or main trends...
 6. Applications/visualizations

For next week :

- ▶ Repository in GitHub : <https://github.com/lserrar/LZML041>
- ▶ Créer un compte personnel Github
- ▶ Installer python3 et Jupyter notebook sur votre poste (utilisez Anaconda !)
- ▶ M'envoyer un email sur loubna.serrar@gmail.com pour me confirmer l'installation de Python, et le compte Github avec une capture d'écran des résultats du notebook Homework1_Test : 5pts si avant le prochain cours...

 Iserrar Add files via upload		b2ba95b now	🕒 14 commits
 Homework1_JupyterLab	Update Homework1_JupyterLab		4 days ago
 Homework1_Test.ipynb	Add files via upload		now
 LICENSE	Initial commit		5 days ago
 README.md	Update README.md		5 days ago