Natural Language Processing (NLP)

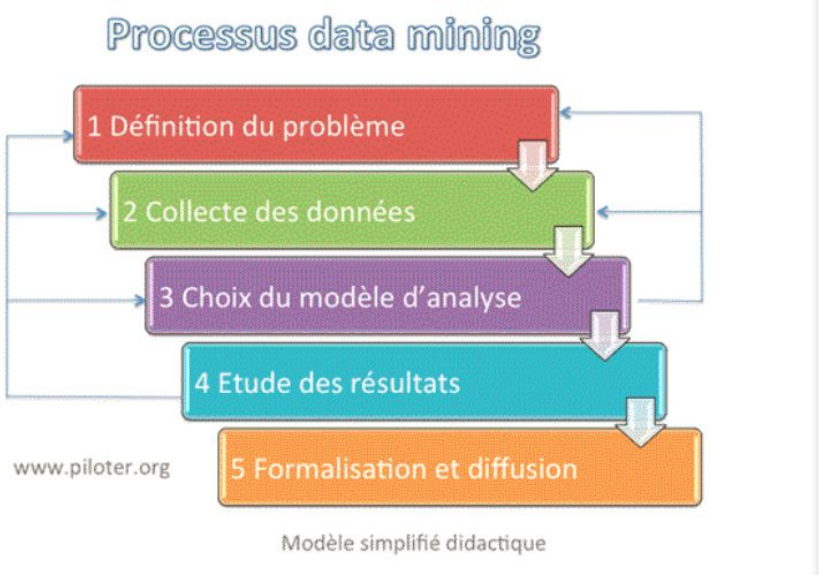
Baptiste - Erwan - Patricia - Paul

Sommaire

- Définition Data Mining
- Définition Text Mining et étapes
- Data Mining vs Text Mining
- Documents structuré et non-structuré
- Natural Language Processing (NLP)
  - Les différentes techniques de traitement NLP
- Applications du Text Mining
  - Apprentissage supervisé : exploiter une collection de document précédemment étiquetés
  - Apprentissage non supervisé : partitionner en groupes homogènes de documents
  - Recherche d'information
  - Autres possibilités
- Représentation Bag-of-Words (BOW)
- Stop words, lemmatisation, stemming
- Sources

Définition Data mining

Le data mining, est le fait d'analyser des données sous le prisme de différentes perspectives afin de transformer ces données en informations utiles, en établissant des relations entre ces dernières ou bien en repérant des patterns.



Modèle simplifié didactique

Définition Text mining et étapes

Le text mining, peut être défini comme étant un ensemble de techniques alliant plusieurs domaines : la linguistique, la sémantique, le langage, les statistiques et l'informatique. Combinées ensemble, ces techniques permettent d'extraire des données pour recréer de l'information à partir de corpus de textes en les classifiant et les analysant de manière à établir des tendances.

Le texte mining, respecte deux étapes principales :

- **L'analyse** : consiste à analyser les corpus de textes de manière à en reconnaître les mots, les phrases, les rôles grammaticaux ainsi que les relations et les sens de ces derniers entre eux.
- **L'interprétation de l'analyse.**

Le text mining a donc pour vocation d'automatiser la structuration de documents faiblement structurés, afin de générer de l'information sur le contenu d'un document texte, cette information n'étant alors pas présentée de manière explicite dans la forme initiale du document.

Data mining VS Text mining

Le text mining est une partie du data mining

S.No.	Data Mining	Text Mining
1.	Data mining is the statistical technique of processing raw data in a structured form.	Text mining is the part of data mining which involves processing of text from documents.
2.	Pre-existing databases and spreadsheets are used to gather information.	The text is used to gather high quality information.
3.	Processing of data is done directly.	Processing of data is done linguistically.
4.	Stastical techniques are used to evaluate data.	Computational linguistic principles are used to evaluate text.
5.	In data mining data is stored in structured format.	In text mining data is stored in unstructured format.
6.	Data is homogeneous and is easy to retrieve.	Data is heterogeneous and is not so easy to retrieve.
7.	It supports mining of mixed data.	In text mining, mining of text is only done.
8.	It combines artificial intelligence, machine learning and statistics and applies it on data.	It applies pattern recognizing and natural language processing to unstructured data.
9.	It is used in fields like marketing, medicine, healthcare.	It is used in fields like bioscience and customer profile analysis.

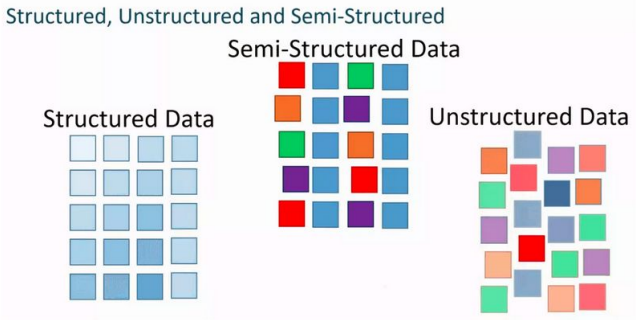
Documents structuré et non-structuré

Données structuré :

Les données structurées sont des informations qui ont été formatées et transformées en un modèle de données bien défini.

Données non-structuré :

Les données présentes sous forme brute absolue sont appelées non-structuré. Ces données sont difficiles à traiter en raison de leur disposition et de leur formatage complexes.
Les données non structurées peuvent être tout ce qui n'est pas dans un format spécifique (paragraphe d'un livre, articles, commentaires et publications sur les réseaux, etc.



Natural Language Processing (NLP)

Le traitement naturel du langage, aussi appelé Natural Language Processing ou NLP en anglais, est une technologie permettant aux machines de lire, de comprendre et de tirer du sens des langages humains grâce à l'intelligence artificielle.

Exemples d'applications :

- Identification des spams sur une boite mail, fausses informations
- ChatBot
- Interfaces vocales : Siri, Alexa, Cortana
- Recrutement (en identifiant les compétences des recrues potentielles)
- Trading (suivre les actualités, les rapports, les commentaires sur d'éventuelles fusions entre entreprises)

Les différentes techniques de traitement NLP

Les deux principales techniques utilisées pour le traitement naturel du langage sont **l'analyse syntaxique et l'analyse sémantique**.

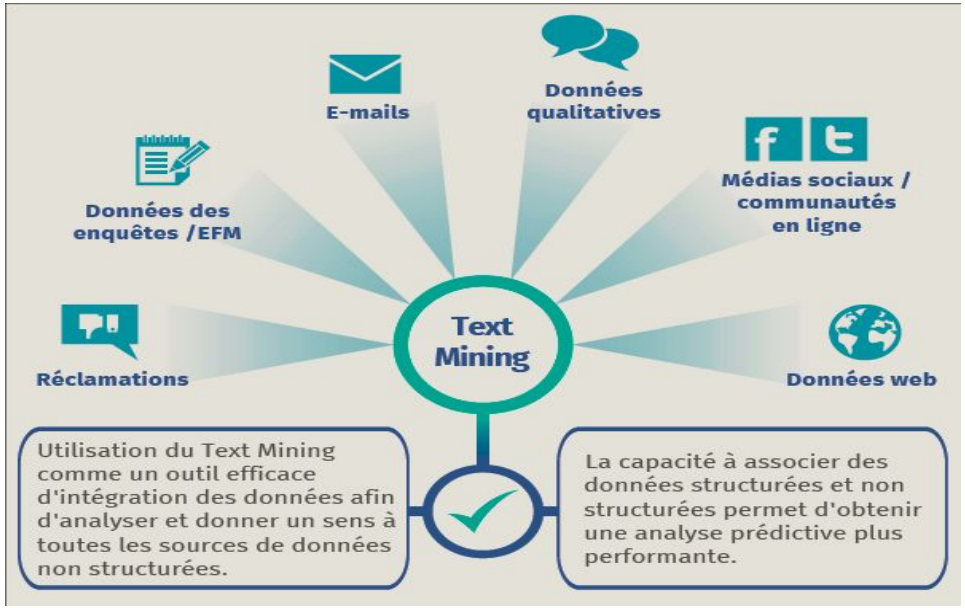
Analyse syntaxique :

L'analyse syntaxique consiste à identifier les règles grammaticales dans une phrase afin d'en déchiffrer le sens. Plusieurs techniques d'analyse sémantique existent. Le **parsing** consiste à analyser la grammaire d'une phrase. La **segmentation** par mot consiste à diviser un texte en unités, tandis que la **segmentation morphologique** divise les mots en groupes.

Analyse sémantique :

Consiste à **déchiffrer directement le sens d'un texte** en utilisant des algorithmes pour analyser les mots et la structure des phrases. Les algorithmes peuvent notamment se baser sur le contexte, ou comparer les textes avec des bases de données pour en comprendre le sens.

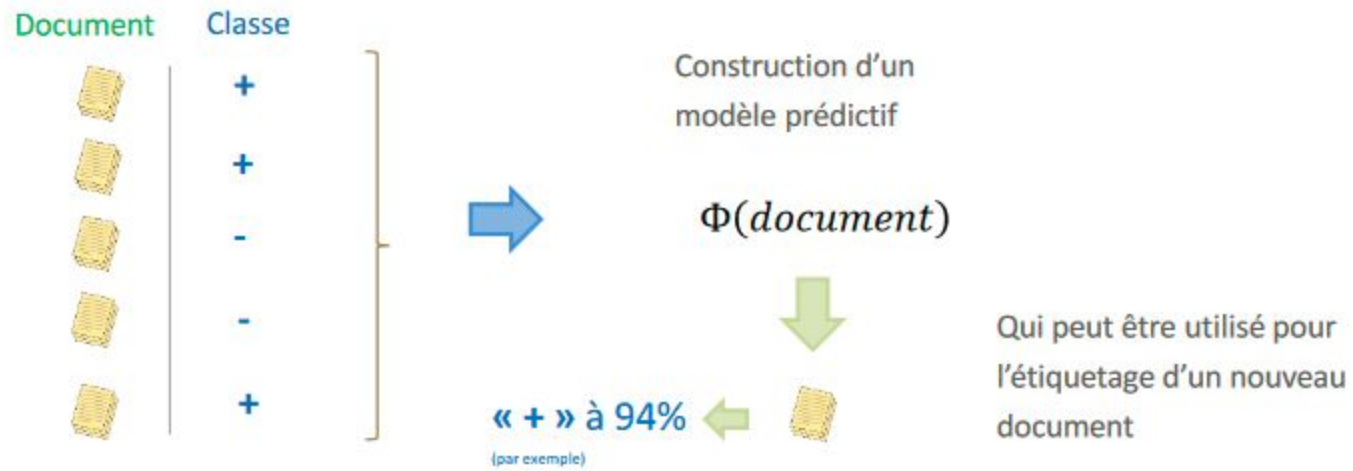
Applications du Text Mining



L'Analyse Textuel aide dans l'analyse de données multi-sources.

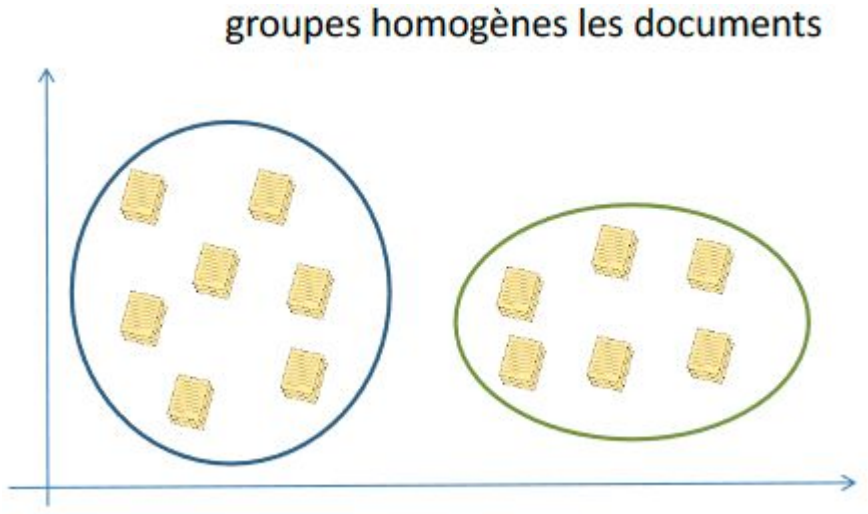
- Définir son objectif
- Gérer les attentes de son organisation
- Choisir le ou les outils d'analyse de données textuelles les mieux adaptés à notre objectif
- Se Familiariser avec les points forts et les points faibles de ses outils
- Ne pas abandonner !

Apprentissage supervisé : exploiter une collection de document précédemment étiquetés



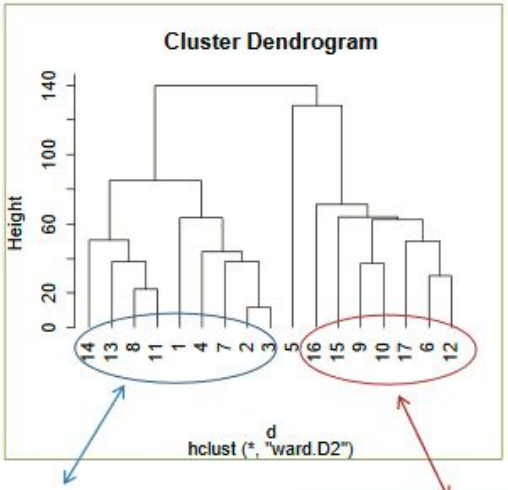
exemple emblématique : la détection de spam.
La base d'apprentissage est constituée de documents SPAM et non SPAM pour la construction du modèle prédictif

Apprentissage non supervisé : partitionner en groupes homogènes de documents

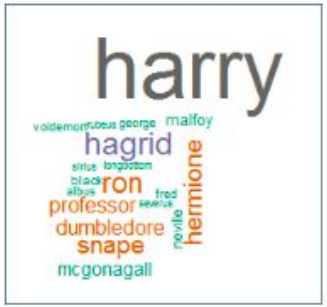


Pour l'a**lgorithme d'apprentissage non supervisé** le but de classifier de manière automatique un ensemble de textes sans apprentissage préalable. Dans ce scénario, le modèle prend en entrée des textes et il tente de les séparer en ensembles cohérents afin de comprendre la nature des groupes, en les associant à des thèmes par exemple

Clustering de textes – Exemple « Harry Potter »

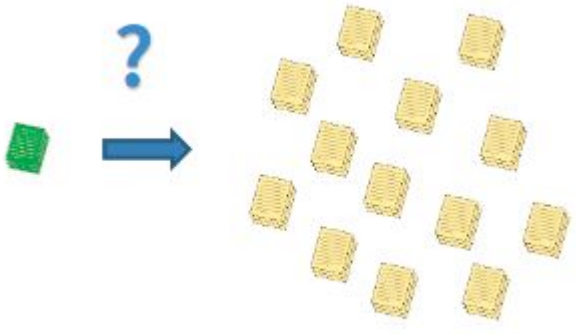


Analyse par chapitre des occurrences des personnages dans le 1er tome de Harry Potter



Harry est présent partout. Normal. Mais, selon les chapitres, Hagrid, Ron ou Hermione se distinguent.

Recherche d’information



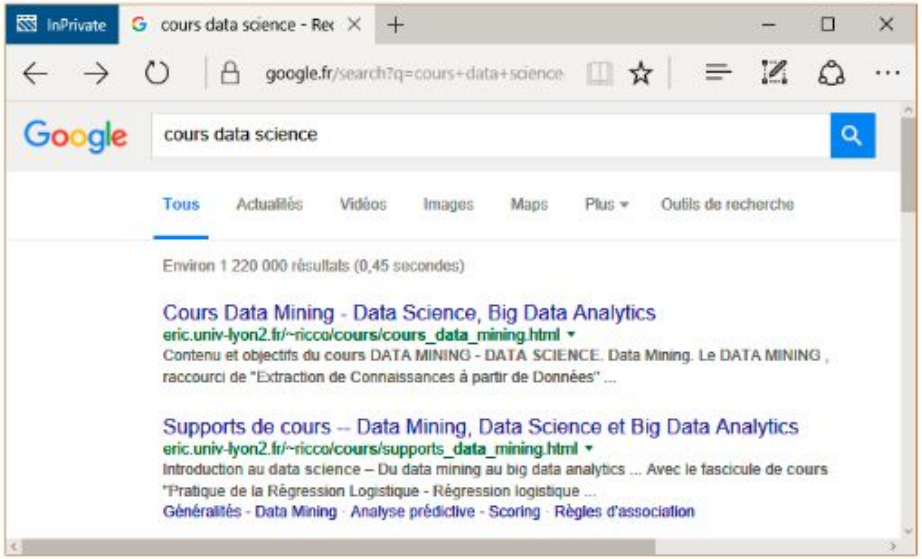
Objectif
mettre en place les stratégies permettant d’identifier, dans un corpus, les documents pertinents relatifs à un document requête.

Il s’agit d’une recherche par le contenu, le texte est concerné, mais elle peut s’étendre à l’image, la vidéo, le son...

Recherche d'information – Le cas particulier des moteurs de recherche web

Les moteurs de recherche web constituent l'exemple le plus emblématique de la recherche d'information, … mais avec des spécificités : le document requête est très court, constitué de quelques mots clés ; des stratégies de spamming viennent polluer la recherche (Voir Historique des moteurs de recherche)

La popularité de Google avec le concept PageRank repose en très grande partie sur sa capacité à parer les assauts des spammeurs, en tenant compte des liens, et de la crédibilité (critère d'autorité) de l'auteur.



Extraction d’information

L’extraction d’informations consiste à rechercher des champs prédéfinis dans un texte plus ou moins rédigé en langage naturel. On s’appuie sur de l’analyse lexicale et morphosyntaxique pour identifier les zones d’intérêts.

Description :
FIAT COUPE 2L 20V 155 CV
1999
136000 km
Contrôle technique ok
Prix: 2990 €
Equipement :
Cuir. Climatisation. Vitres électriques. Rétroviseurs électriques. Direction assistée. Fermeture centralisée. ABS. Airbag. Jantes aluminium

Fiat coupe T 16 (moteur lancia delta dorigine) an 95 tuning leger rabaisse av et ar blistein b8 ressourt court ct ok en cours le vehicule roule tres peu (suite perte emploi) il es entretenu marche tres fort environ 200 ch il peut parcourir tous les trajet. je ne suis pas presse alors les marchand de tapis pas la peine de venir la peinture a ete refait par lancien proprietaire ainsi que lembrayage mais elle as des inperfections demande de renseignement par tel uniquement

Ex. extraire automatiquement les informations additionnelles sur les annonces automobiles.

Annonce	2L	5 cylindres	Turbo	Bolidée	1ère main	Réparations à prévoir
1	oui	oui	non	non	?	non
2	oui	non	oui	oui	?	oui
3	...	...	...	...	...	...

Objectif : Récupérer automatiquement les informations qui ne sont pas standardisées sur leboncoin.fr (ex. pour les voitures : année modèle, kilométrage, carburant, boîte)

Autres possibilités

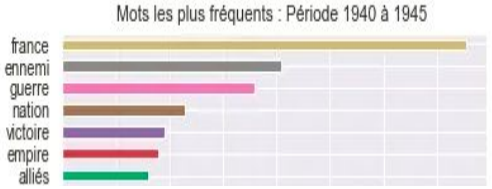
- résumé automatique:
  - phrase les plus représentatives dans son document
- identification des tendances ;
  - Analyser l'évolution de la popularité des thèmes, les sujets émergents (topic détection)
- analyse des liens
  - Analyser les relations entre les termes (ex dans le cas de l'analyse des réseaux sociaux)

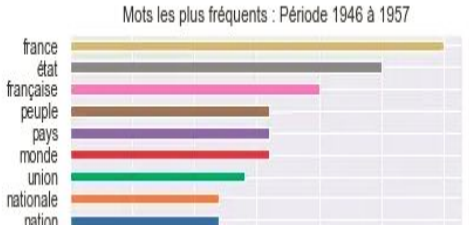
exemple : un corpus de textes constitué de **25 discours du Général de Gaulle**

Discours 1940-1945

Discours 1946-1957

Discours 1958-1969





Représentation Bag-of-Words (BOW)

La manière la plus simple de formater un texte pour l'analyse consiste à ignorer l'ordre pour se concentrer sur la fréquence de chaque mot et les retourner sous forme d'un dictionnaire clés-valeurs, le **Bag-of-Words** :

"Un verre en verre vert." -> {"Un":1, "verre":2, "en":1, "vert":1}

Il est aussi possible de former des clés de plusieurs mots qui se suivent (modèle **N-gramme**). Un modèle bigramme donnera ceci :

"Un verre en verre vert." ->

{"Un verre":1, "verre en": 1, "en verre":1, "verre vert": 1}.

Stop words, lemmatisation, stemming

- L'analyse de fréquence de mots se heurte au fait que certains mots banals reviennent très souvent ("il", "un", "à" en français par exemple). On peut les identifier comme une liste de ***stop words*/mots vides** à ignorer.
- On peut trouver des mots aux sens proches par la **lemmatisation**, où une famille de mots est renvoyée à sa racine lexicale, son **lemme**. Par exemple, "meilleur" a pour lemme "bien" même s'ils n'ont pas la même étymologie.
- Une technique plus automatique de lemmatisation est le ***stemming*/racinisation**, où l'algorithme identifie des mots de même lemme en isolant la racine du mot (ex : "agrandir" -> "grand"). Une **approche algorithmique** est plus rapide mais fait des erreurs, une **approche par dictionnaire** peut donner de meilleurs résultats (ex : algorithme de Porter, avec des règles pour retirer les préfixes, etc.).

Sources

Définition NLP: https://www.lebigdata.fr/traitement-naturel-du-langage-nlp-definition

Documents structuré/ non-structuré :
https://www.astera.com/fr/type/blog/structured-semi-structured-and-unstructured-data/

https://eric.univ-lyon2.fr/~ricco/cours/slides/TM.A%20-%20introduction%20text%20mining.pdf

https://www.geeksforgeeks.org/difference-between-data-mining-and-text-mining/

https://ia-data-analytics.fr/logiciel-data-mining/text-mining/definition/