Introduction to Text Mining

Pr T.RACHAD

QFM – 2024/2025

Objectifs

- Découvrir le processus d'extraction de connaissances à partir de données textuelles.
- Découvrir les techniques de prétraitement d'une collection de documents texte.
- Présentation des algorithmes courants de découverte de patterns à partir d'une collection de documents texte.
- Découvrir quelques cas d'utilisation courants en text mining.

Prérequis

- Les bases des statistiques
- Les bases d’analyse de données
- Les bases du machine learning
- Maitrise du langage python
  - Structures de données
  - Traitement des fichiers
  - Numpy
  - Scipy
  - plotlib
  - …

Bibliographie

- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

- Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press

- Zhai, C., & Massung, S. (2016). Text data management and analysis: a practical introduction to information retrieval and text mining. Association for Computing Machinery and Morgan & Claypool.

- KESELJ, Vlado. Speech and Language Processing Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009.

Plan

1. Introduction	
2. Prétraitement de texte	• **Atelier 1: Les tâches de base de prétraitement de texte** • **Détection de plagiat (Approche syntaxique vs approche sémantique)**
2. Modèles représentatifs de documents texte	• **Atelier 2: Vectorisation de documents** • **Détection de plagiat (Approche vectorielle)**
3. Classification de documents texte	• **Atelier 3: Classification de documents texte** • **Classification des news**
4. Clustering de documents texte	• **Atelier 4: Clustering de documents texte** • **Détection du plagiat (approche non supervisée)**
5. Topic Mining	• **Atelier 5: Analyse de topics** • **Détection des topics de news**
6. Découverte de patterns et de tendances dans les documents texte	• **Tendances et patterns dans les news**

7. Applications et topics avancés du text mining	• **Exposés/mini Projets**
	  • **Question Answering**
	  • **Summarization**
	  • **Intention mining**
	  • **Opinion mining**
	  • **Sentiment analysis**
	  • **LLMs**

Evaluation

- Travaux personnels 40%
- Examen final 60%

Introduction

C’est quoi le Text Mining?

- C’est un processus qui permet la découverte d’informations utiles et de patterns à partir d’une collection de documents texte **non structurées**.

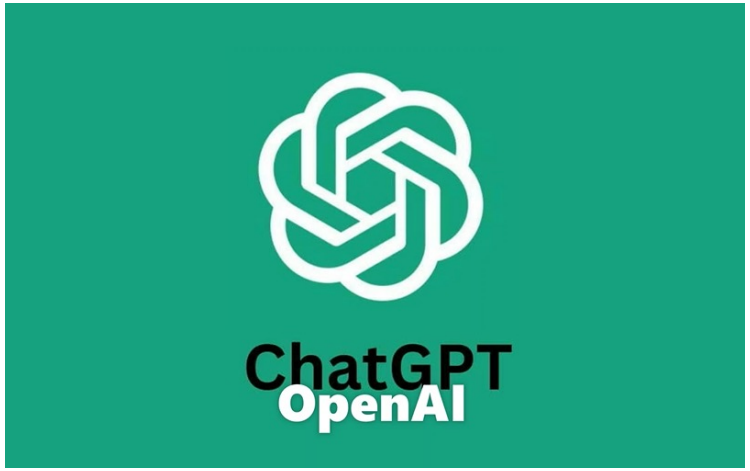
- Généralement le Text Mining est utilisé pour **catégoriser** les documents texte en fonctions des **sujets** traités, **sentiments** exprimés, **opinions** données ou en fonction des **intentions** annoncés.

Applications du Text Mining

- Automatic Question Answering
- Automatic Text Translation
- Text Summarization
- Document Indexing
- Sentiment Analysis
- Disease diagnosis
- Intention mining
- Fraud detection
- Opinion Mining
- Topic mining
- …

Applications du Text Mining

- Automatic Question Answering chatbot (2022)
- Apprentissage supervisé et apprentissage par renforcement.
- 100 millions utilisateurs en 2023
- 1 milliard de dollars au bout d’une année

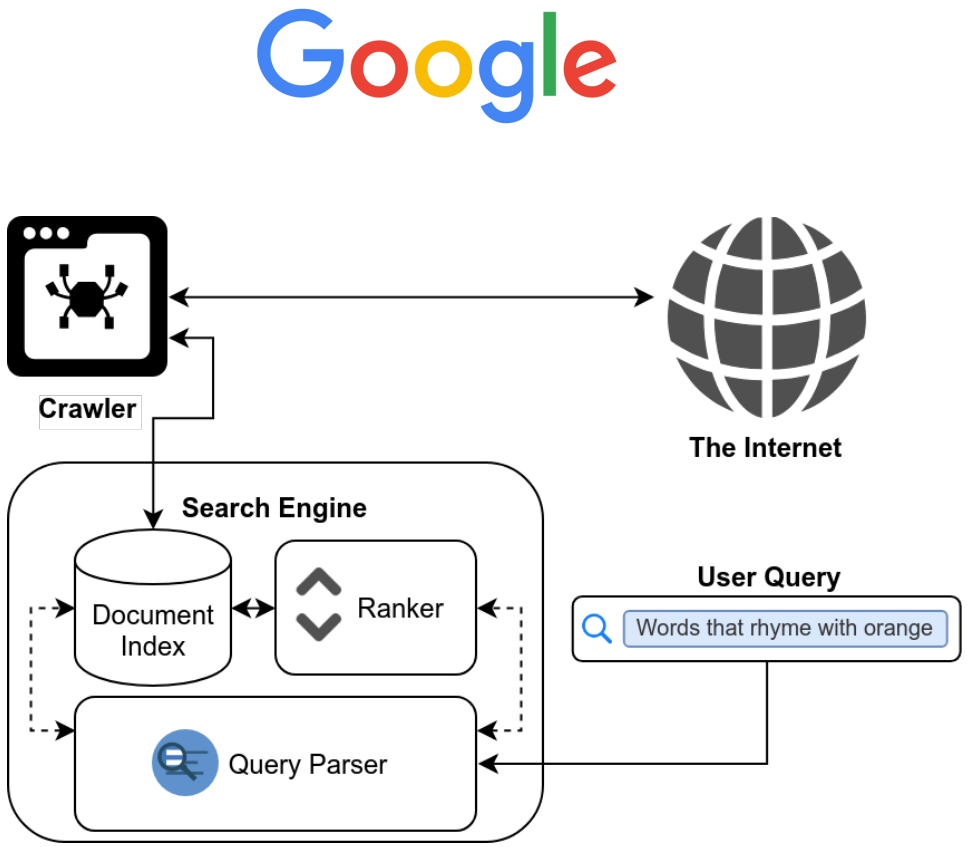


Applications du Text Mining

- Outil de traduction le plus Populaire.
- Disponible en +100 langues.
- 500 millions utilisateurs.
- Basé sur le modèle Seq2Seq développé par google.

Applications du Text Mining

- Moteur de recherche le plus populaire.
- Indexation de (x?) térabits de pages web.
- 3.5 billions de recherches par jour.
- Trois étapes: Crawling, Indexation et ranking.
- Ranking des pages web est basé sur l'algorithme PageRink

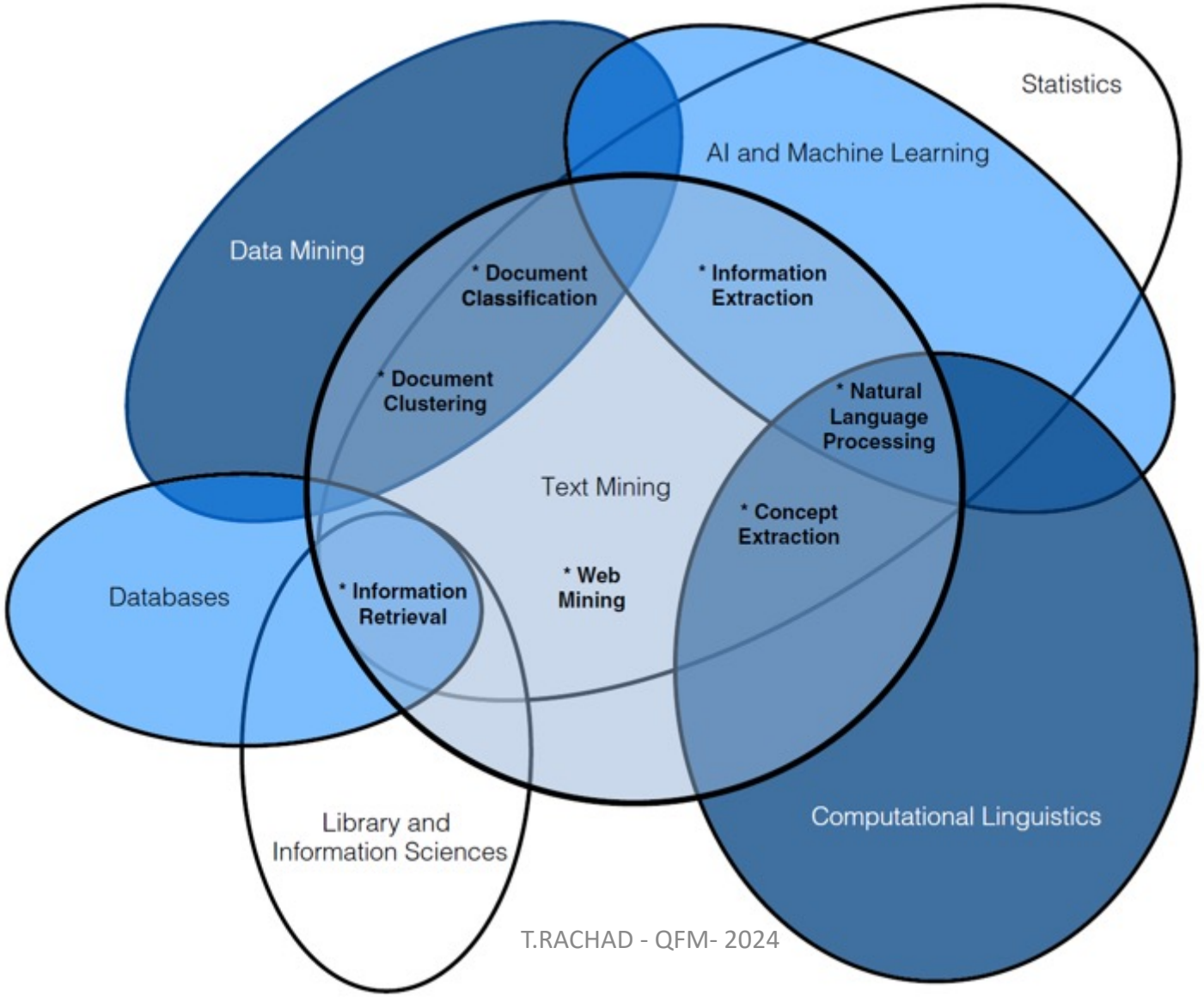


https://www.cs.toronto.edu/~muuo/img/Anatomy%20of%20Search.png

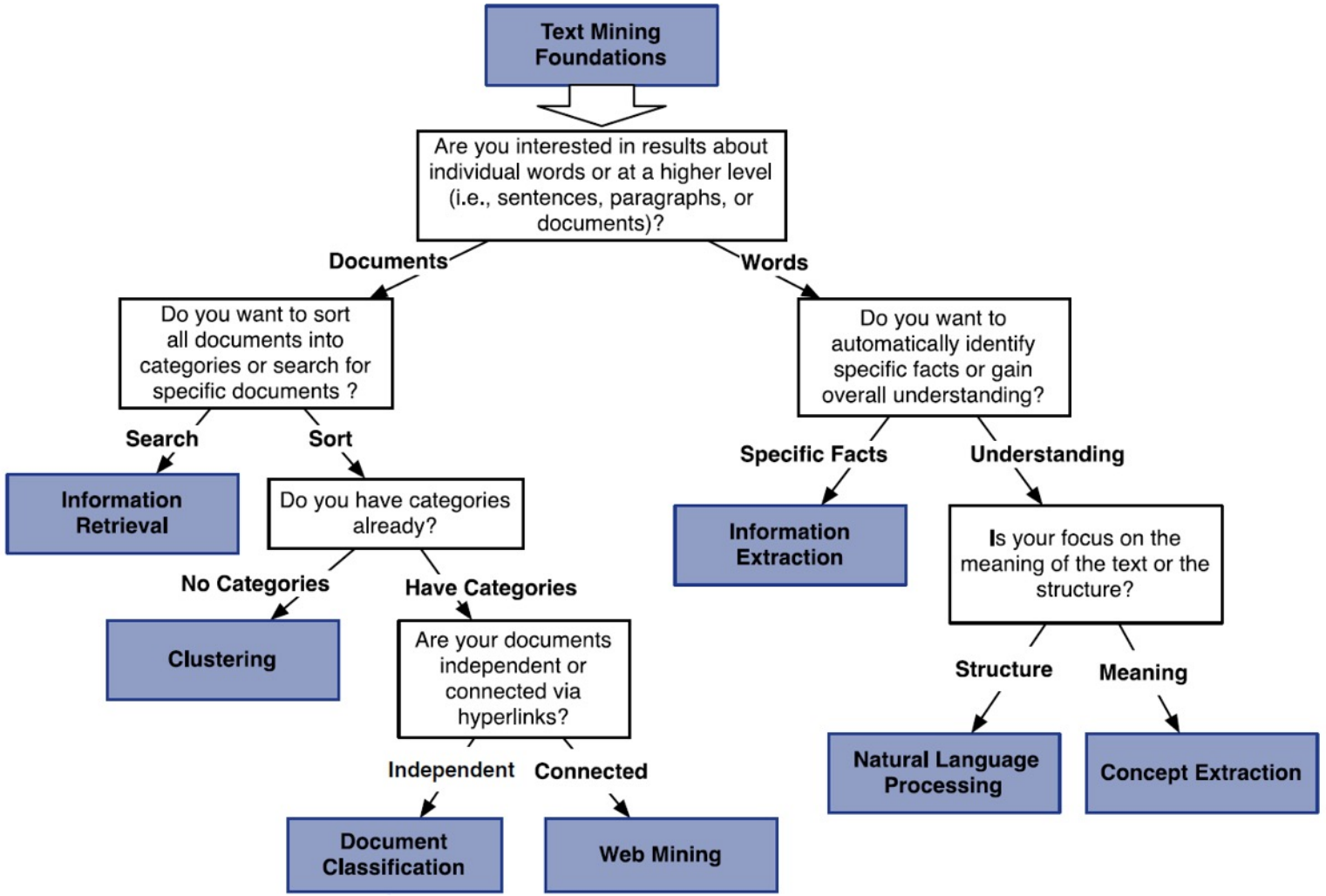
Applications du Text Mining

Translation	DeepL, ALEXA TRANSLATIONS, BING TRANSLATOR, OpenAI
Summarization	cohere, connexun, Azure, ONE AI, OpenAI
QA	Gemini, Ask2End (Ask anything, get the ultimate answer!), NLPCloud, Hugging Face
Disease diagnosis	twill™, IBM Watson, iz.ai, ENLITIC, regard

Techniques du Text Mining



Tâches du Text Mining



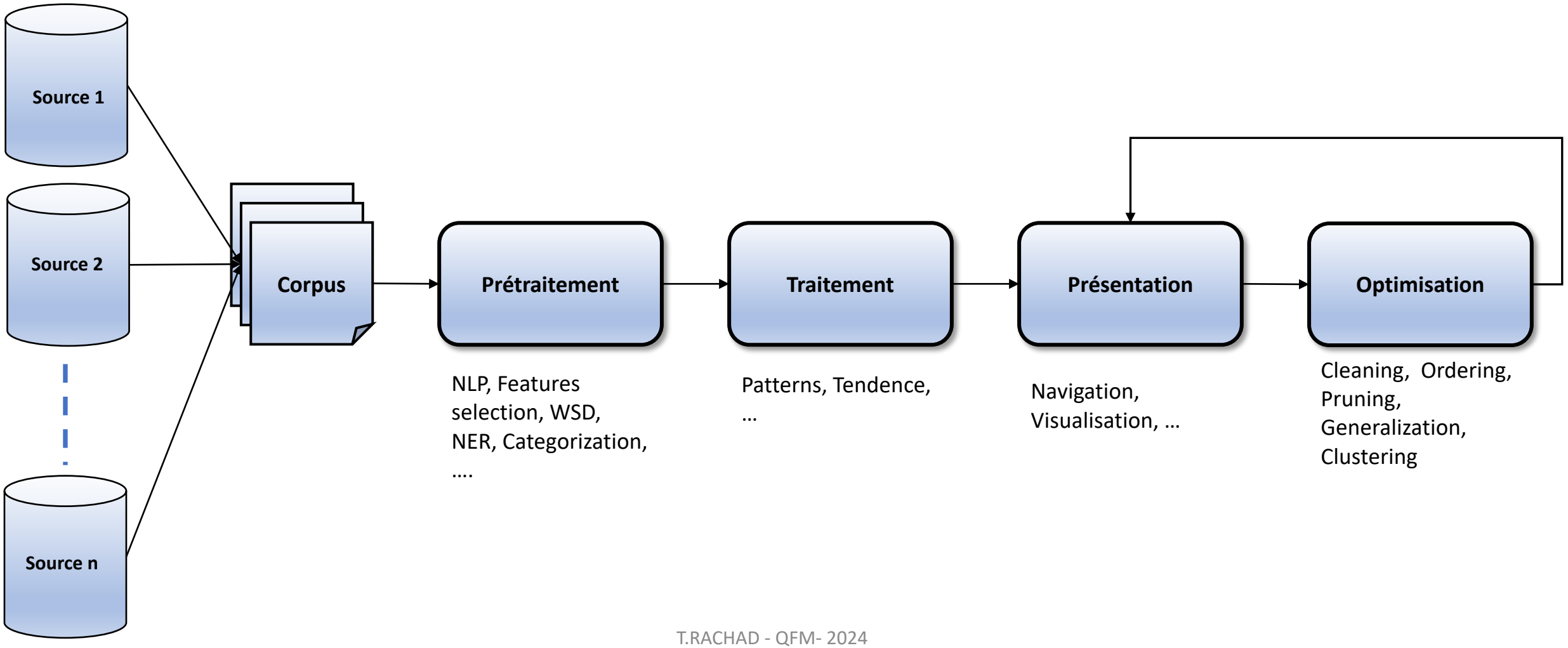
Extracted from https://towardsdatascience.com/a-text-analytics-primer-key-factors-in-a-text-analytics-strategy-d24dc84a5576

Processus du Text Mining

- Les processus de Text Mining et data Mining sont similaires dans le sens où ils partagent les mêmes étapes de:
  - Prétraitement,
  - Application d'algorithmes pour la découverte de patterns,
  - Exploration et visualisation des résultats
  - Application des techniques de raffinement

- Ça n'implique pas forcément que ces étapes se font de la même façon ou qu'elles ont les mêmes objectifs.

Processus du Text Mining



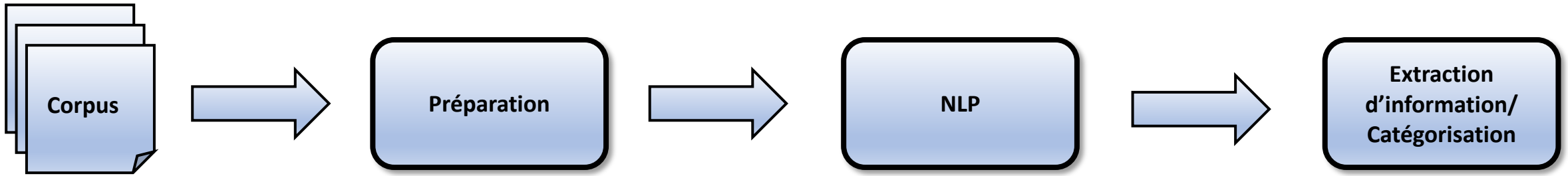
Prétraitements

- Les prétraitements en text mining possèdent une grande importance et affectent considérablement les performances des traitements qui vont venir par la suite.

- Ils visent principalement à produire à partir des données non structurées de nouvelles représentations de données qui sont plus structurées. Cela se fait via l'extraction des concepts clés (**Features extraction**) qui représentent mieux le contenu des documents texte.

- Finalement, il faut assurer l'enrichissement de la structure d'un document en affinant les concepts clés actuelles (**Features selection, réduction de la dimensionnalité**) et en en ajoutant de nouveaux concepts (**Extraction d'information, catégorisation**).

Prétraitements



OCR (Optical Character Recognition)
Speech recognition
….

Tokeniszation
Lemmatization
POS tagging
syntactic parsing
….

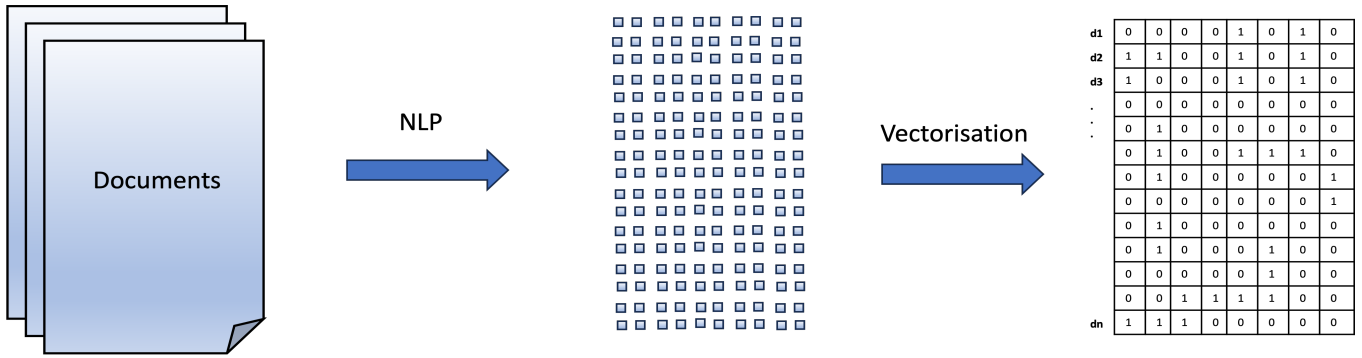
Prétraitements (Préparation)

- Les tâches de préparation font la conversion des données à partir de leur format brut vers une structure plus adaptée aux traitements linguistiques.

- Plusieurs techniques sont utilisées:
  - Les OCR (Optical character recognition)
  - Reconnaissance de la parole (speech recognition)
  - conversion de fichiers électroniques à partir des formats propriétaires
  - …

Prétraitements (NLP)

- Les tâches de la NLP traitent des documents texte en utilisant les connaissances générales sur un langage naturel:
  - Segmentation (Tokenization),
  - Analyse morphologique
  - Étiquetage morpho-syntaxique (POS tagging),
  - Analyse syntaxique (syntactic parsing)
  - Analyse sémantique
  - Vectorisation: BOW, TFIDF, WORD2VEC…



Prétraitements (Extraction d'informations)

- L'extraction d'information vise à identifier les informations pertinentes (mots-clés, noms d'entités, adresses, e-mails, etc.) dans les documents (en considérant les dimensions syntaxique et sémantique du langage) et à les présenter dans un format structuré (généralement sous format tabulaire) pour les exploiter par la suite dans des opérations de traitement.

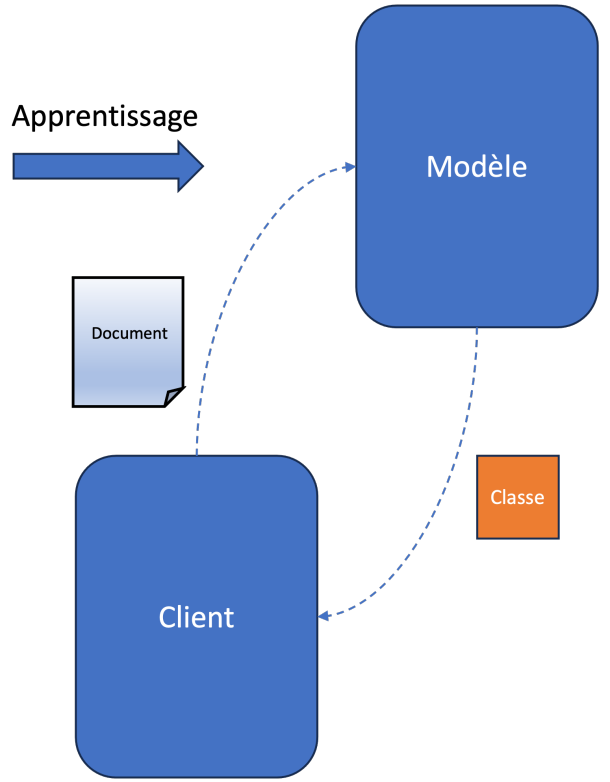
- Quelques exemples d'extraction d'information:
  - NER (named-entity recognition)
  - Détection de relation
  - Extraction d'évènements
  - Analyse temporelle
  - Template filling
  - …

Prétraitements (Catégorisation)

- La tâche de catégorisation (classification) marque chaque document avec un petit nombre de concepts ou de mots-clés qui sont reconnus à l'avance.

- Quelques exemples de catégorisation de texte:
  - Topic analysis
  - Sentiment analysis
  - Langage détection
  - Intent detection
  - Opinion mining

d1	0	0	0	0	1	0	1	0	1
d2	1	1	0	0	1	0	1	0	0
d3	1	0	0	0	1	0	1	0	1
.	0	0	0	0	0	0	0	0	0
.	0	1	0	0	0	0	0	0	0
.	0	1	0	0	1	1	1	0	0
	0	1	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	1	1
	0	1	0	0	0	0	0	0	0
	0	1	0	0	0	1	0	0	0
	0	0	0	0	0	1	0	0	1
	0	0	1	1	1	1	0	0	0
dn	1	1	1	0	0	0	0	0	1



Traitements

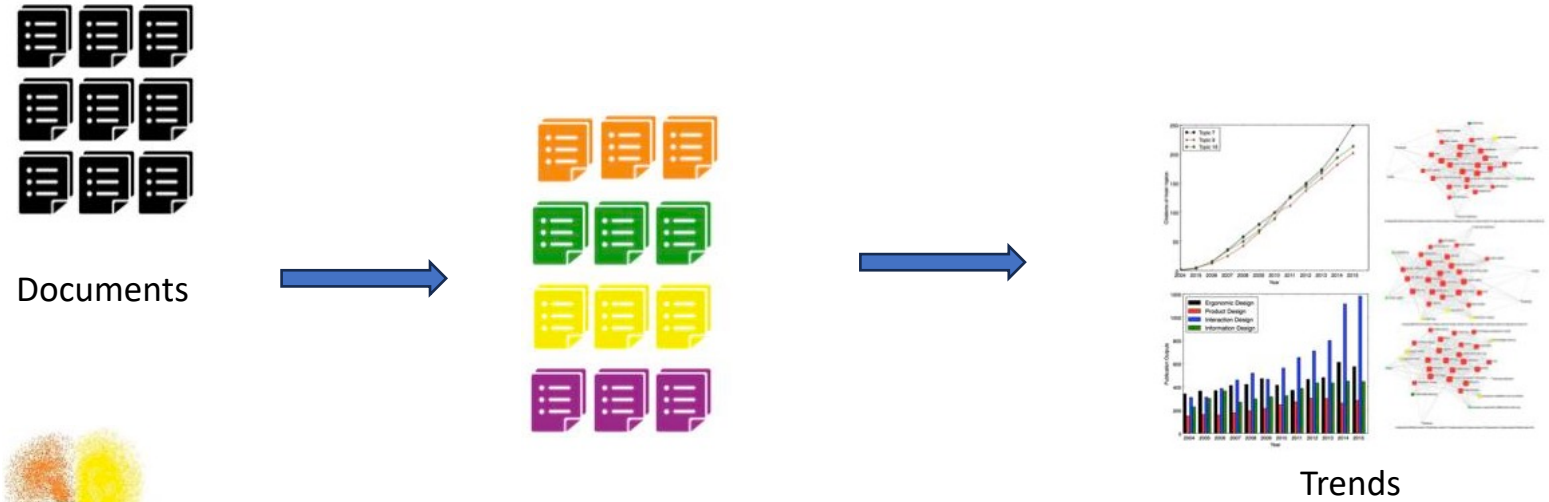
- Les traitements en Text Mining se réfèrent à:
  - la découverte de patterns de cooccurrence de concepts dans une collection de documents texte.
  - la détection de tendances dans une collections de documents texte.
  - Le regroupent de documents (Clustering)

Traitements (Découverte de patterns)

- Trois patterns sont souvent rencontrés en Text Mining :
  - Les distributions et les proportions de concepts
  - Ensemble des concepts fréquents,
  - Les associations de concepts.

Traitements (Analyse de tendances)

- L’analyse des tendances dans des documents texte repose souvent sur l’aspect temporel pour comparer des collections de documents qui proviennent de la même source et qui possèdent la même structure mais qui sont collectées dans des périodes différentes.



Documents

Clusters

Trends

Présentation

- Les utilisateurs auront besoin d'outils de navigation et de visualisation qui leur permettront l'exploration des patterns découverts lors des étapes de traitement:
  - Outils graphiques de visualisations
  - Outils graphiques d'exploration
  - Editeurs
  - …

Optimisation

- Des techniques permettant le filtrage des redondances et le regroupement des données similaires:
  - Cleaning,
  - Ordering,
  - Pruning,
  - Generalization
  - Clustering