

Natural Language Processing

Aude, Amaury & Thomas

Data Mining & Text Mining



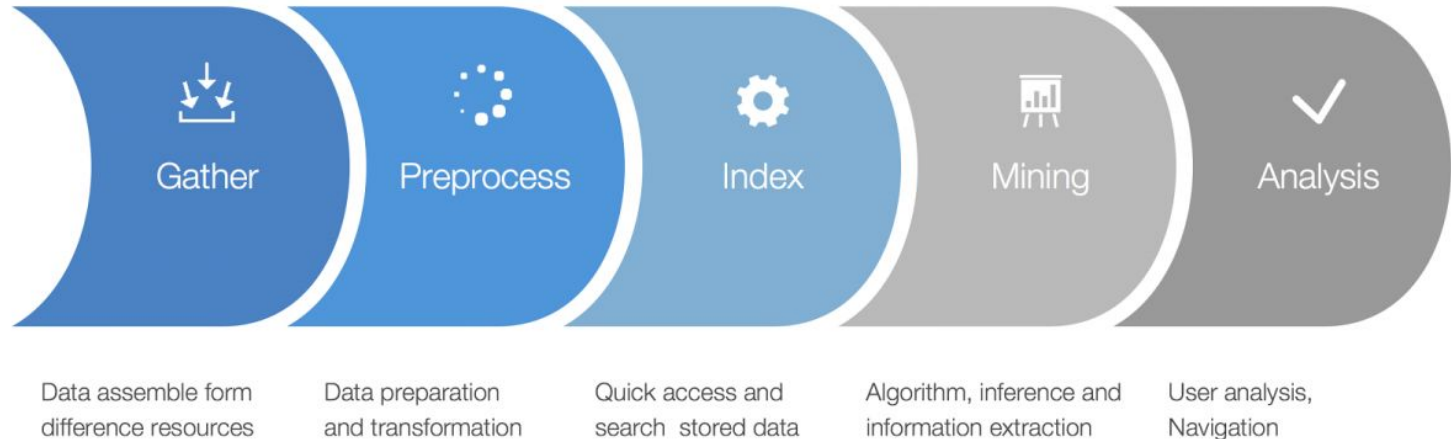
Les deux pratiques sont corrélées mais ne s'effectuent pas sur le même type de données.

- data mining : association, classification, clustering, régression sur des données déjà structurées
- text mining : même but que le data mining mais sur des données non structurées, qu'il faudra d'abord organiser et structurer.

Processus du Text Mining

Text Mining

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:



Documents (structurés et non structurés)

Documents structurés :

Jeu de données bien définies, design et positionnement des éléments constant d'un document à un autre.

Exemple : formulaires, questionnaires, données tabulaires...

Documents non-structurés :

Jeu de données non-structurées en aucune façon. Pas de champs de données explicites.

Exemple : XML, contrats, lettres, commandes, audio, video, chat logs...

Voir aussi : Données semi-structurées (ou “flexibles”)

Design, le nombre et le placement peuvent varier d'un document à l'autre.

Exemple : factures dont le nombre d'entrées dépend de l'entreprise émettrice.

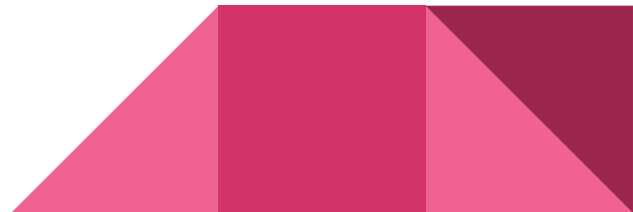
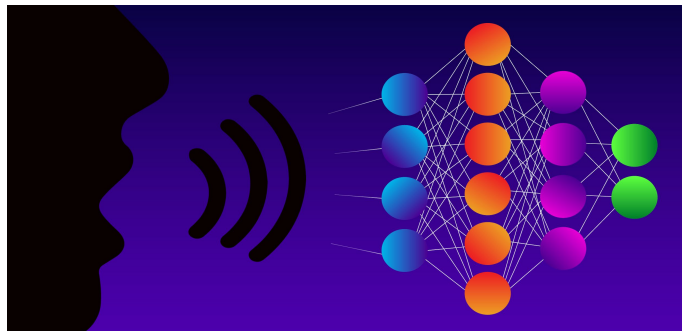


Natural Language Processing

L'objectif de cette technologie est de permettre aux machines de lire, de déchiffrer, de comprendre et de donner sens au langage humain.

Avant: Les algorithmes de Machine Learning de l'époque cherchaient des mots et des phrases dans un texte et donnaient des réponses spécifiques en fonction.

Maintenant: le Deep Learning permet une approche plus flexible, plus proche du langage naturel.



Use Cases of NLP



Translation Application



Fake News Detection



Classifying Emails



Predicting Disease



Error Detection



IVR Application



Sentiment Analysis



Personal Voice Assistant

Applications

Supervisé :


Tokenization : Casser un texte en bouts que la machine peut comprendre. Pour les langues logographiques sans espace comme le Chinois, permet d'entraîner la machine à comprendre les motifs.

Sentiment Analysis : Déterminer le contenu émotionnel d'un texte (positif, négatif, neutre). Score de sentiment (poids) associé à chaque entité du document. Tâche extrêmement complexe (à cause des figures de style, par ex).

Non supervisé :

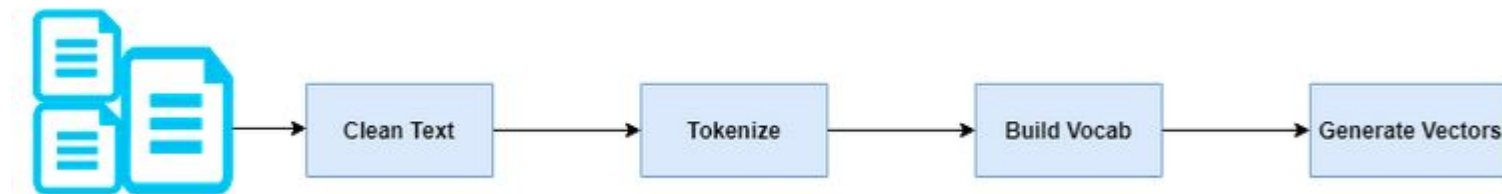
Clustering des documents similaires (non pré-tagués) en groupes. Puis, classement hiérarchique de ces clusters.

Latent Semantic Indexing (LSI) : identifier des mots et phrases qui surviennent souvent ensemble. Utilisé pour retourner des résultats de recherche proches mais pas avec le mot exact.



Représentation Bag Of Words

C'est une collection de mots pour représenter une phrase via le comptage des mots et sans accorder d'importance à leur ordre d'apparition.



Processus de vectorisation

1. "John likes to watch movies. Mary likes movies too."

2. "John also likes to watch football games."

1. {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1}

2. {"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1}

{"John":2,"likes":3,"to":2,"watch":2,"movies":2,"Mary":1,"too":1,"also":1,"football":1,"games":1}

Résultat

```
John likes to watch movies. Mary likes movies too.[1, 2, 1, 1, 2, 1, 1, 0, 0, 0]
```

```
John also likes to watch football games.[1, 1, 1, 1, 0, 0, 0, 1, 1, 1]
```

possibilité d'utiliser des bibliothèques comme CountVectorizer in sci-kit learn.



Réduction de dimensionnalité

Stop words (mot vide)

Mots filtrés avant/après NLP sur du texte. Souvent les mots les plus communs de la langue. Parfois conservés.

Exemple : “le, la, les” qui ne sont pas pris en compte par un moteur de recherche en français pour améliorer la performance.

Lemmatisation

Traitement lexical apporté au texte pour analyse. Remplacement de toute occurrence d'un lexème donné par sa forme canonique (son lemme).

Exemple : lexèmes “petit, petite, petits, petites” -> lemme “petit”

Stemming (racinisation)

Similaire à la lemmatisation mais plus simple. On tronque le mot pour avoir sa racine.

Exemple : mots “continu, continuait, continuation” -> racine “continu”

Une lemmatisation aurait donné “continu, continuer, continuation”.

