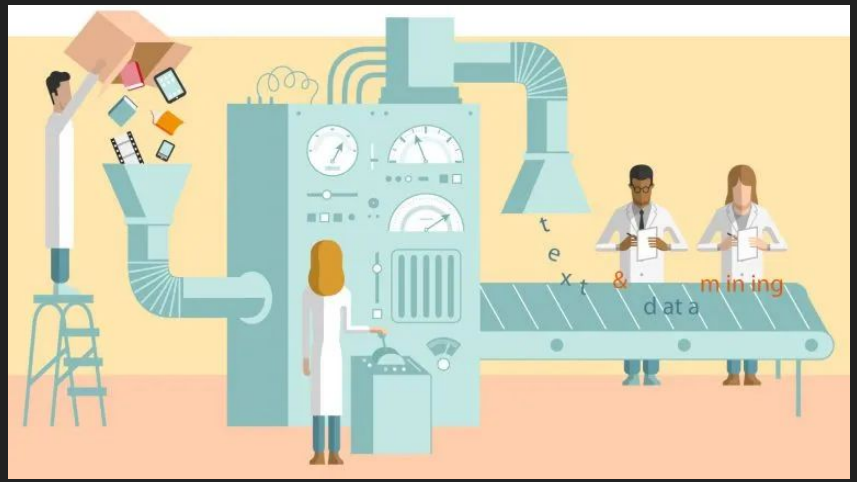
Natural Language processing

Sommaire

- Définition Data mining, Text mining et la différence entre les deux
- Processus du Text Mining (Etapes)
- Définition d'un document structuré et non structuré
- Définition du NLP
- Exemple d'applications du Text mining (Apprentissage supervisé, non-supervisé, recherche de l'information, extraction de l'information, etc)
- Représentation Bag Of Words (BOW)
- Explication de la réduction de dimensionnalité avec le retrait des stopwords, lemmatization et le stemming.

Data Mining

Le *Data Mining* englobant toute une famille d'outils facilitant l'exploration et l'analyse des données contenues au sein d'une base décisionnelle de base de donnée.



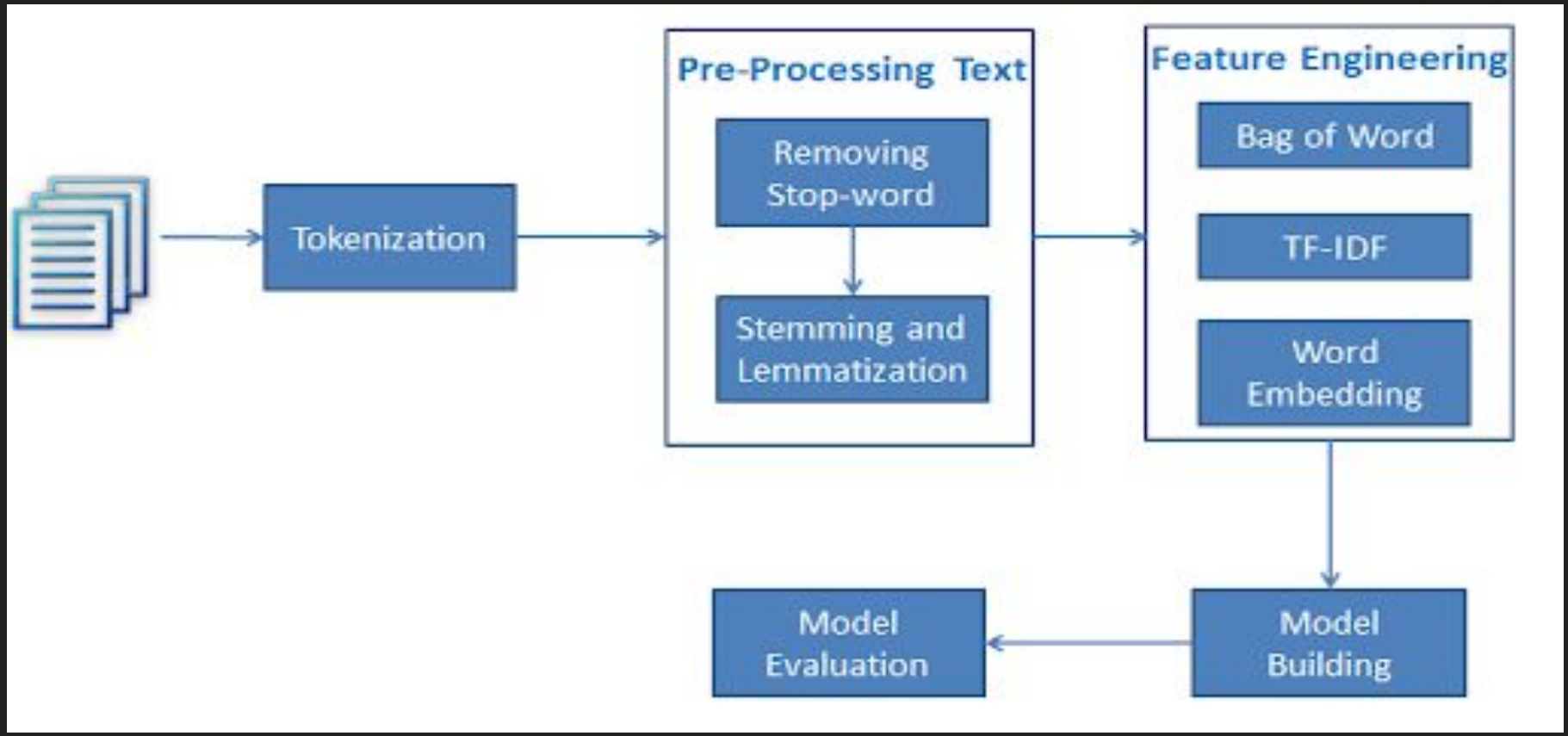
- **Association** – chercher des patterns au sein desquelles un événement est lié à un autre événement.
- **Analyse de séquence** – chercher des patterns au sein desquelles un événement mène à un autre événement plus tardif.
- **Classification** – chercher de nouvelles patterns, quitte à changer la façon dont les données sont organisées.
- **Clustering** – trouver et documenter visuellement des groupes de faits précédemment inconnus.
- **Prédiction** – découvrir des patterns de données pouvant mener à des prédictions raisonnables sur le futur. Ce type de data mining est aussi connu sous le nom d’analyse prédictive.

Text mining

Le text mining regroupe l'ensemble des techniques de data management et de data mining permettant le traitement des données particulières que sont les données textuelles.

- Expressions ou groupes de mots spécifiques
- Enlever tous les mots outils (« il y a », « un », « une », « dans », *etc*.)
- Retirer mots ayant la même racine
- Traiter les expressions ou groupes de mots spécifiques
- Classer des mails
- Détection des données sensibles

- Processus du Text Mining (Etapes)



Document structuré et non structuré

Document structuré

C’est un document dont la structure logique est décrite plutôt que la mise en forme physique, il n’est pas destiné à être directement utilisé pour la lecture humaine.

Document non structuré

C’est un document dont les données sont représentées ou stockées sans format prédéfini. Elles sont toujours destinées à des humains et peuvent contenir du texte, du multimédia, des dates, nombres, etc. Ici l'absence de format entraîne des irrégularités et des ambiguïtés pouvant rendre difficile la compréhension.

LE NLP

Définition: Le traitement naturel du langage, ou Natural Language Processing (NLP) en anglais, est une technologie d'intelligence artificielle visant à permettre aux ordinateurs de comprendre le langage humain.

Cas d'usage: Les applications de traduction(Google Translate), Apple Siri , Google,Microsoft Cortana, Amazon Alexa, les chatbot, traitement de texte (Microsoft word, Grammarly),les réseaux sociaux(étude de tendance, d'images,analyse de réactions).

Les techniques utilisées: -L'analyse syntaxique (consiste à identifier les règles grammaticales dans une phrase pour déchiffrer son sens).

-L' analyse sémantique (Le parsing qui consiste à analyser la grammaire d'une phrase, la segmentation par mot divise le texte en unités, la segmentation morphologique divise les mots en groupes).

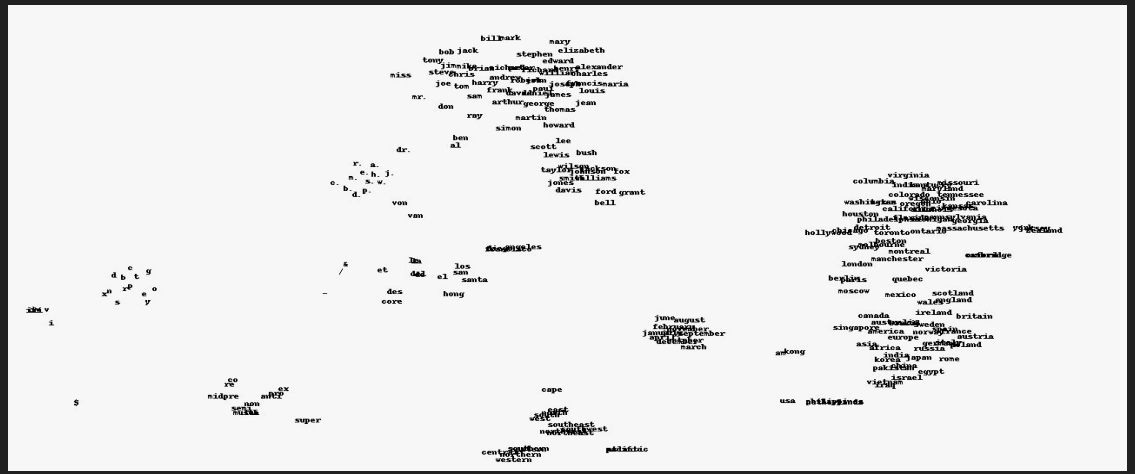
Exemples de Text Mining

- ➔ Analyse des médias sociaux
- ➔ Filtrage du spam
- ➔ Service clientèle
- ➔ Détection des fraudes (assurances)

Text Mining supervisé et non supervisé

La classification de texte avec des algorithmes tels que Naïves Bayes, Arbre de Décision , Machine à Vecteurs de Support ou K plus proches voisins , a besoin d’un apprentissage avant la phase de test.

Le Clustering est pour l’instant l'algorithme le plus utilisé en classification non supervisée.



Recherche et extraction d’information

NER ou Named Entity Recognition est une des techniques du Text Mining qui consiste à attribuer une étiquette à un ensemble de texte.
Ce système est utilisé dans l’analyse de sentiment , en service clientèle ou marketing ou dans l’administration.

Bag of words

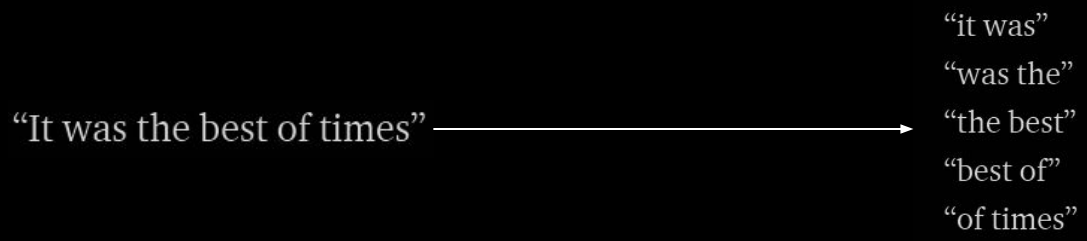
Solution pour extraire des features d’un texte :

- Tokenisation de chaque mot, puis recherche de fréquence dans une phrase afin de les transformer en vecteur.

“It was the best of times”	*“It was the best of times” = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]*
“It was the worst of times”	*“It was the worst of times” = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]*
“It was the age of wisdom”	*“It was the age of wisdom” = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]*
“It was the age of foolishness”	*“It was the age of foolishness” = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]*

Bag of words

L’approche présentée précédemment, chaque mot (ou token) est appelé un gram. On peut également procéder en regroupant les mots par deux (bigram).



Convertir le texte en nombre est appelé vectorisation. Il y a deux solutions pour effectuer la conversion :

- Comptage du nombre d’occurence
- Calcul des fréquences d’apparition

Bag of words

Méthodes de vectorisation :

- Scikit learn CountVectorizer :
  - Compte les occurrences de chaque token dans le document et construit une matrice (document x token)

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
vect = CountVectorizer()
from nltk.tokenize import TreebankWordTokenizer
tokenizer = TreebankWordTokenizer()
vect.set_params(tokenizer=tokenizer.tokenize)
vect.set_params(stop_words='english')
vect.set_params(ngram_range=(1, 2))
vect.set_params(max_df=0.5)
vect.set_params(min_df=2)
```

Bag of words

- Scikit learn TF-IDF Vectorizer :
  - Mesure statistique du poid d'un mot dans un document. Son importance augmente en fonction du nombre d'occurence.

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF(t) = log_e(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

$$TF - IDF score = TF * IDF$$

Réduction dimensionnalité

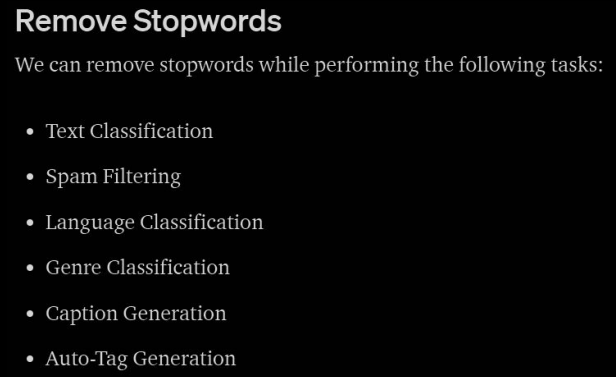
Pour une augmentation des performances, on doit appliquer une réduction de dimensionnalité en réduisant le nombre de features :

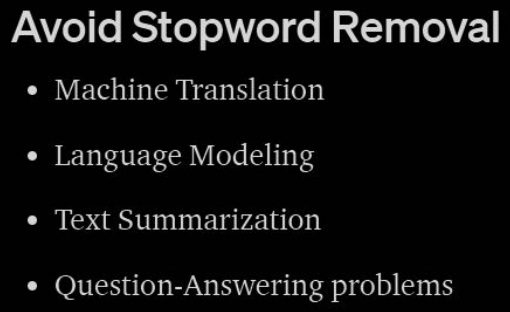
- Meilleures performance de calcul
- Plus de représentativité des données

Réduction dimensionnalité

Principe suppression des stopwords :

- Retirer les mots “vides” qui ne rajoutent pas de sens (ex: des, de,le …)
- En fonction de l’application, il ne faut pas nécessairement les retirer :





Normalisation du texte

Un même mot peut prendre plusieurs forme. Par exemple, les verbes peuvent être conjugués sans affecté le sens de celui ci. Pour gérer cela, on utilise 2 méthodes :

- Stemming : Suppression des préfixes et suffixes
- Lemmatisation : Analyse le mot et retourne sa racine

La lemmatisation est plus souvent utilisée et aide à créer de meilleure features.