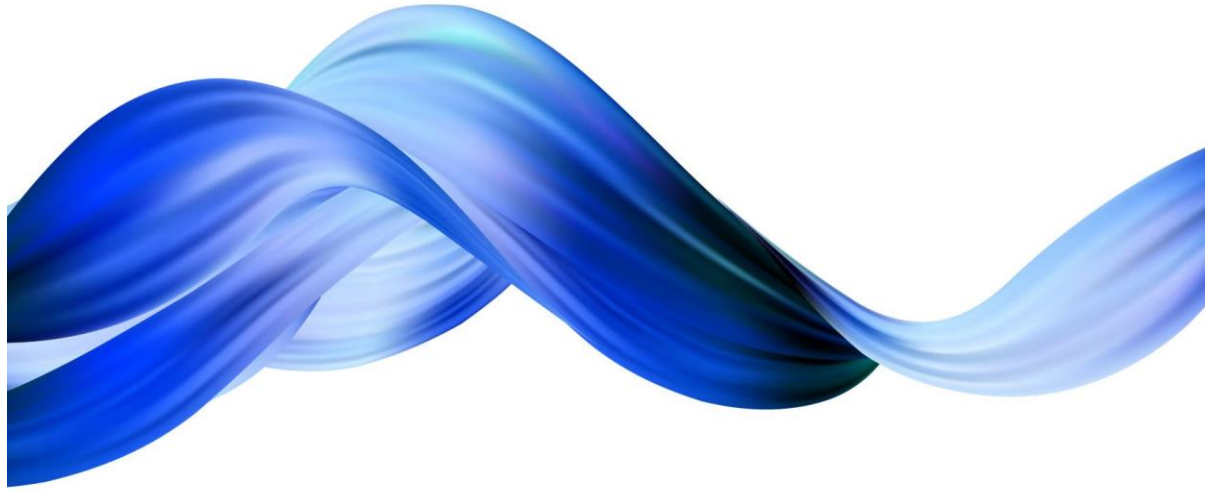
Traitement du langage naturel NLP

BUT SCIENCE DES DONNÉES

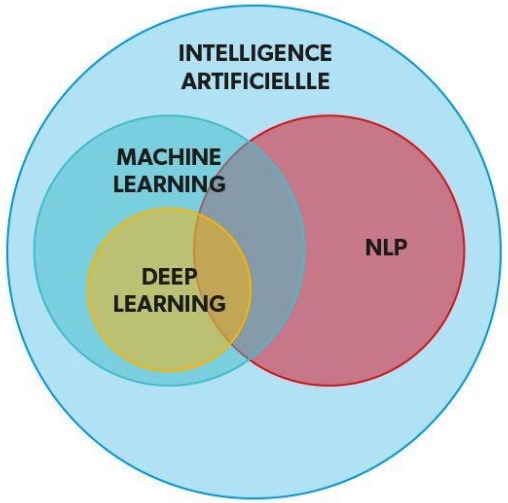
SÉBASTIEN QUINAULT –
JANVIER 2024



INTRODUCTION

Qu’est-ce que le NLP *(natural langage processing)* ?

Définition : Le traitement du langage naturel (NLP) est une branche de l’intelligence artificielle qui se concentre sur **l’interaction entre les ordinateurs et le langage humain**. Il vise à lire, décoder, comprendre et donner du sens au langage humain d’une manière précieuse et utilisable



Objectif du NLP

L’objectif du NLP est de permettre à un ordinateur de comprendre parfaitement le langage grâce à l'analyse, l'extraction d'informations, la classification et la génération de contenu écrit et parlé.

Les champs d'applications?

Recherche intelligente

Traduction automatique ex. deepl

Natural Language Processing

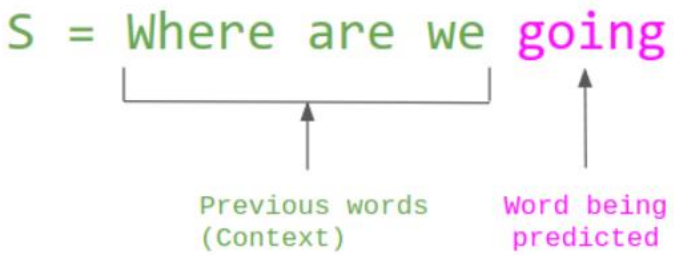
Extraction de caractéristiques
(ex. une adresse, un IBAN…)

Q/R pour les chatbots par ex

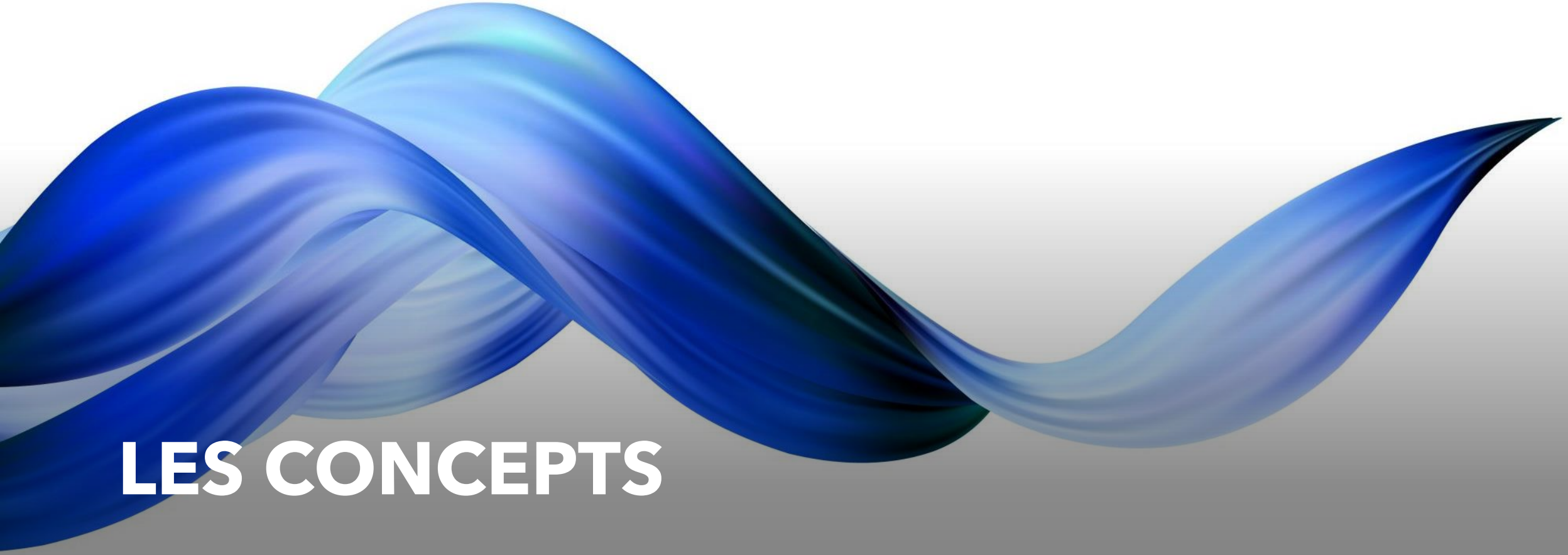
Les LLM (Large Langage Models)

Les LLM, comme GPT4, sont désormais capables de générer du texte de très haute qualité.
Il s'agit de réseaux de neurones profonds, entrainés sur des quantités colossales de données.

Il utilise ensuite une approche probabiliste pour prédire le prochain mot d'une séquence de mots.

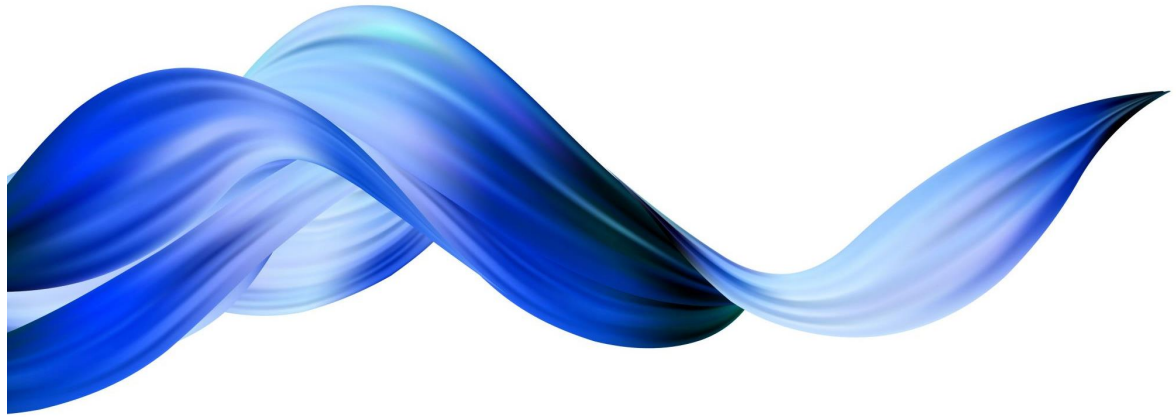


$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$



LES CONCEPTS

PRE-TRAITEMENT



Nettoyage

Cette étape consiste à supprimer les caractères indésirables tels que les signes de ponctuation, les chiffres, les symboles, etc. Elle permet également de convertir le texte en minuscules pour éviter les problèmes de casse. On peut également retirer les accents.

L'objectif est de réduire le bruit dans les données, simplifier le texte et faciliter son analyse

```
s = "Hello, World!"
s = s.replace(",", "")
s = s.replace("!", "")
print(s)
```

```
Hello World
```

StopWords

Des mots sans réelle signification mais un avec poids important dans un corpus.

L'objectif est de réduire la taille vocabulaire et d'éliminer le bruit dans les données. On se concentre sur les mots les plus pertinents.

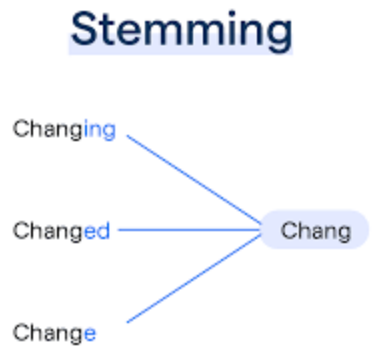
Attention à ne pas supprimer des mots qui changeraient le sens de la phrase (négation par ex.)

When was the first computer invented?
How do I install a hard disk drive?
How do I use Adobe Photoshop?
Where can I learn more about computers?
How to download a video from YouTube
What is a special character?
How do I clear my Internet browser history?
How do you split the screen in Windows?
How do I remove the keys on a keyboard?
How do I install a hard disk drive?

Stemming

Réduction d’un mot à sa forme racine, regroupement des variantes d’un mot sous une seule forme, mais ne tient pas compte du sens du mot généré.

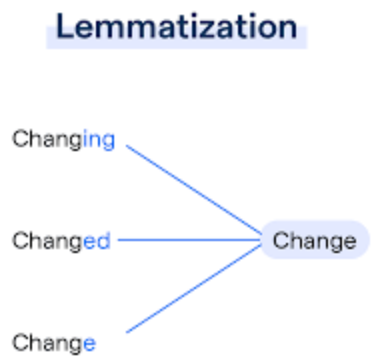
L’objectif du stemming est de réduire la taille du vocabulaire et de simplifier la représentation des mots



Lemmatization

Pocessus qui consiste à réduire un mot à sa forme de base (*lemme*), en tenant compte de son contexte grammatical et de son sens.

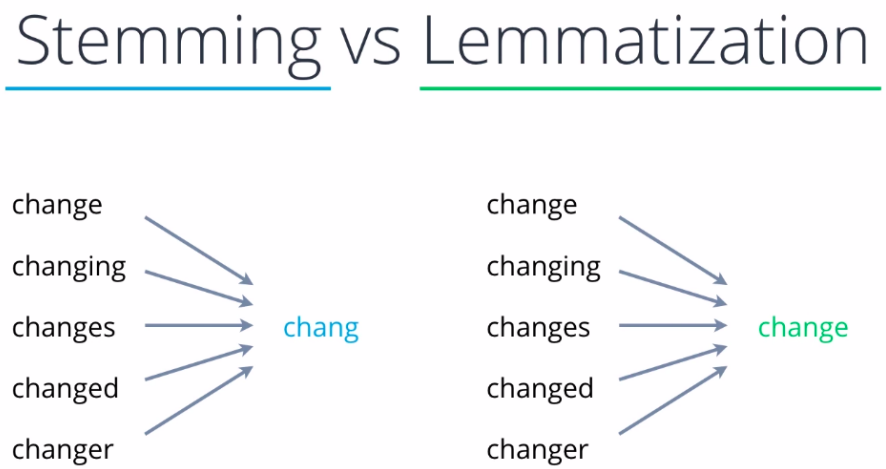
L'objectif de la lemmatisation est de représenter les mots de manière plus précise et plus cohérente, tout en conservant leur sens spécifique.



Stemming & lemmatization

stemming : Plus rapide et plus simple à implémenter, **mais** il peut produire des racines qui ne sont pas des mots valides ou qui changent le sens du mot. Utile pour des applications qui ne nécessitent pas une grande précision, comme la recherche d'information ou l'analyse de sentiment.

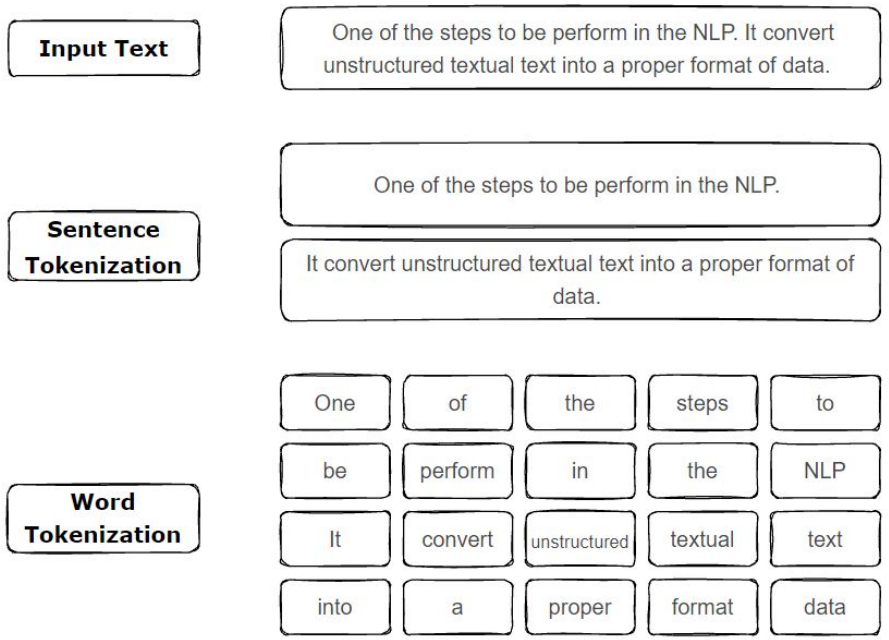
lemmatisation : plus lente et plus complexe à implémenter, mais elle produit des lemmes qui sont des mots valides et qui conservent le sens du mot. La lemmatisation peut être utile pour des applications qui nécessitent une représentation plus précise et plus cohérente des mots, comme les chatbots, les systèmes de questions-réponses, ou l'analyse sémantique.



Tokenization

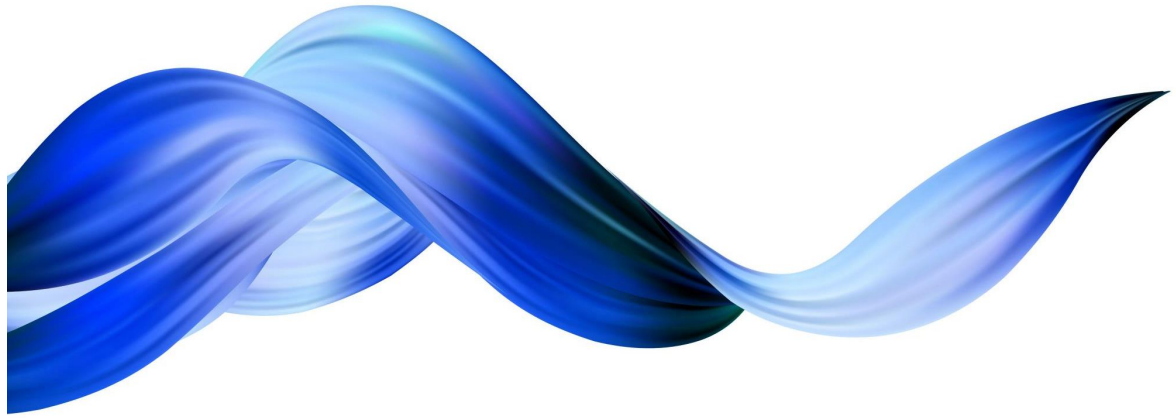
La tokenisation : processus qui consiste à découper un texte en unités plus petites appelées tokens. Un token peut être un mot, une partie de mot ou un caractère comme la ponctuation.

L'objectif de la tokenisation est de préparer le texte pour des tâches plus avancées. En convertissant le texte en tokens, on obtient une représentation structurée qui facilite l'analyse automatique



Pipeline de preprocessing

VISUALISATION

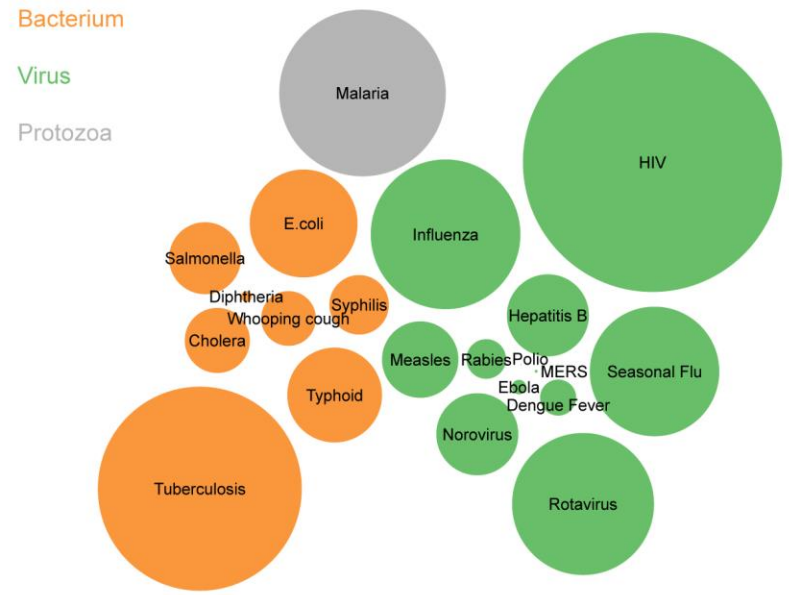


Wordcloud

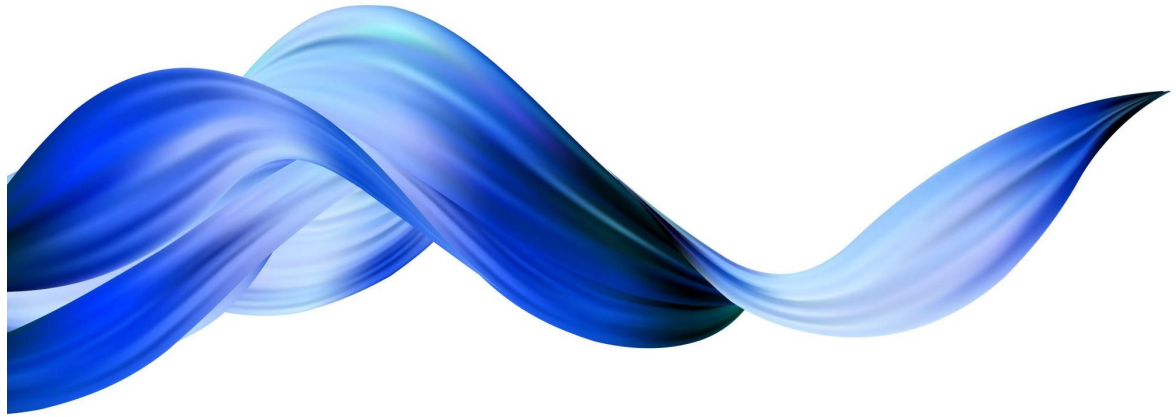
Visuel permettant de visualiser l’importance des mots dans un corpus.

S’utilise avec des mots seuls ou des n-grams (tuples de mots)

Bubble chart



EXTRACTION DE FEATURES



Feature extraction

Tâche qui consiste à transformer un texte brut en un format qui peut être facilement traité par des algorithmes d’apprentissage automatique.

Bag of Words

Un bag of words est une méthode de représentation d'un texte par un ensemble de mots (ou de n-grammes) sans tenir compte de leur ordre ou de leur contexte. Chaque mot est associé à une fréquence d'apparition dans le texte, ce qui permet de mesurer son importance.

	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

	king	was	the	not	But	him	one	north	kings	is	in	he	Eddard	everywhere	different	could	change	but	are	Stark	North	Hodor	without
0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
1	2	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0
2	1	2	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) est une mesure statistique qui reflète l'importance d'un mot dans un document ou un corpus. Elle est calculée comme le produit de la fréquence des termes (nombre de fois qu'un mot apparaît dans un document) et de la fréquence inverse des documents (logarithme du nombre total de documents divisé par le nombre de documents contenant le mot).

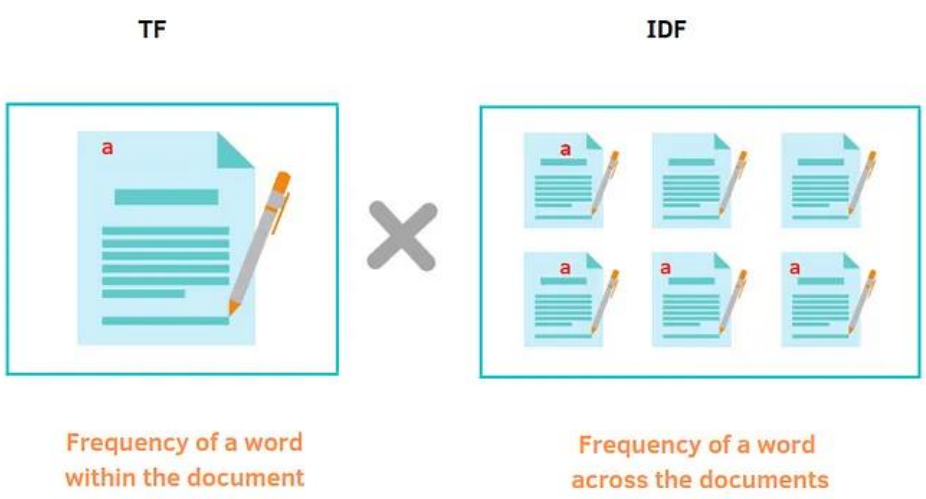
Les vecteurs TF-IDF qui en résultent représentent chaque document comme un vecteur dans un espace à haute dimension où les mots qui sont plus importants dans le document ont un poids plus élevé.

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents



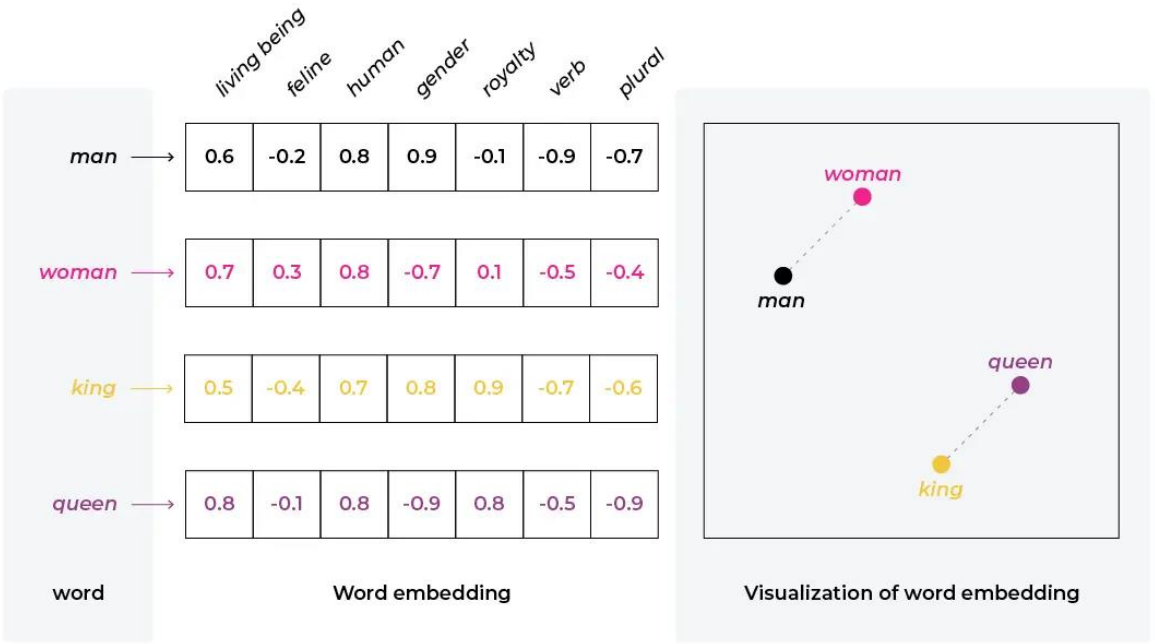
TF-IDF

	text	tf	idf
0	Eddard Stark is a king in the north.	1	3
1	A king but one king : kings are everywhere.	2	3
2	Hodor was different : he was not a king .	1	3
3	But the North could not change without him.	0	3

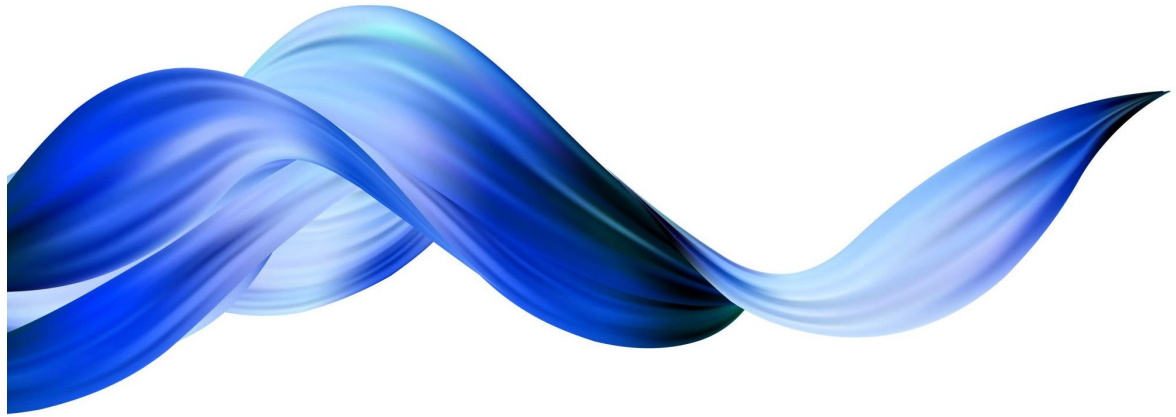
	king	was	the	not	a		he	one	north	kings	is	in	him	everywhere	A	different	could	change	but	are	Stark	North	Hodor	Eddard
0	0.333333	0.0	0.5	0.0	0.5	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.666667	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
2	0.333333	2.0	0.0	0.5	0.5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	0.000000	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0

Word Embedding

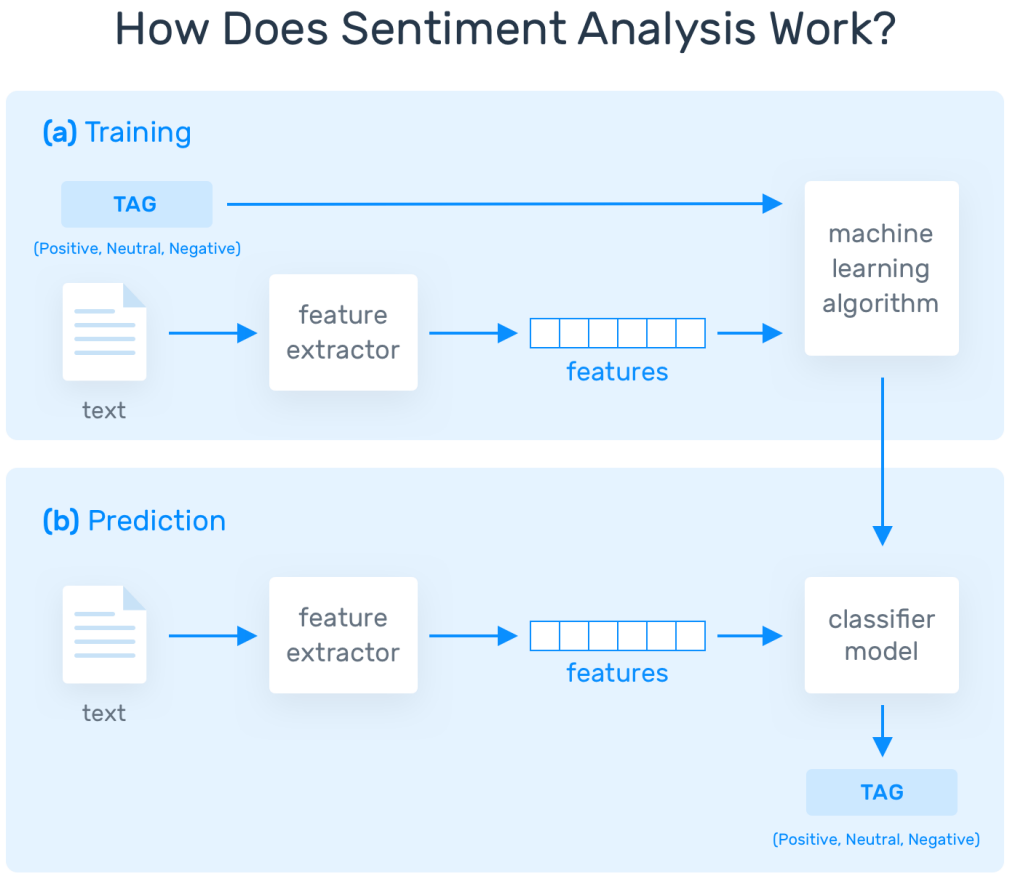
Les word embeddings représentent chaque mot comme un vecteur dans un espace à haute dimension où les mots similaires sont proches les uns des autres et les mots dissemblables sont éloignés les uns des autres.
Les word embeddings sont généralement appris à partir de grandes quantités de données textuelles à l'aide de méthodes d'apprentissage non supervisées telles que Word2Vec et GloVe.



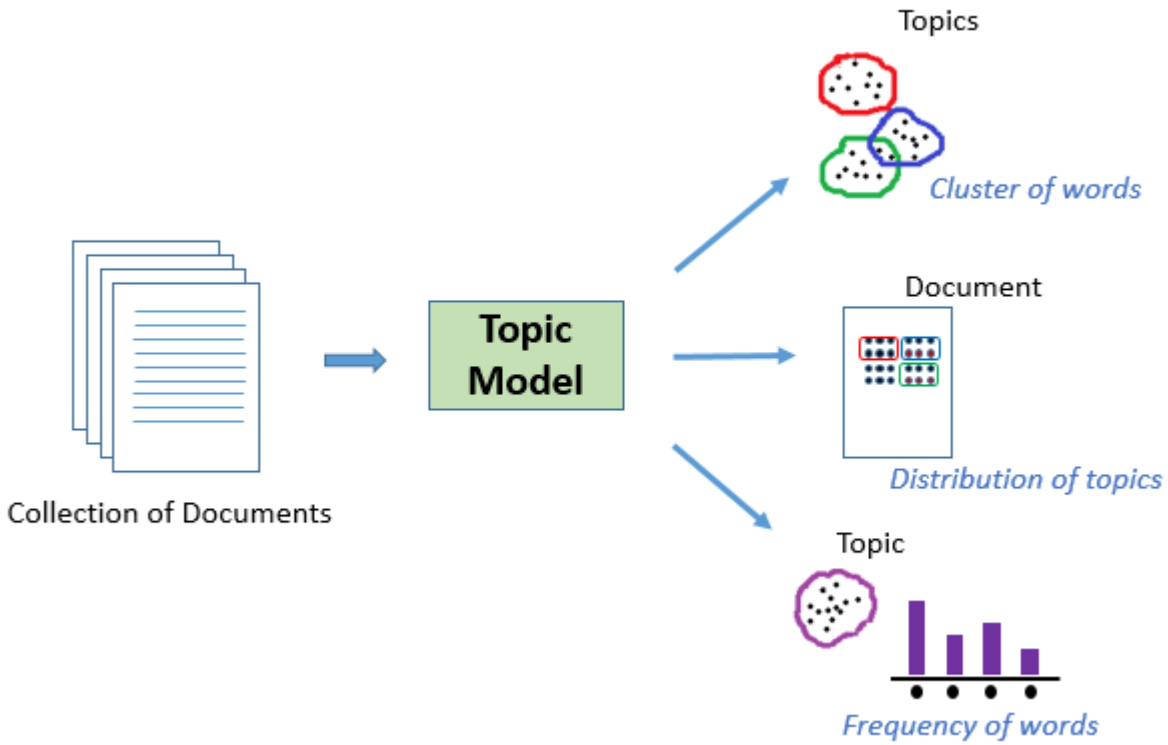
MODELISATION



Sentiment analysis



Topic modeling



Topic classification