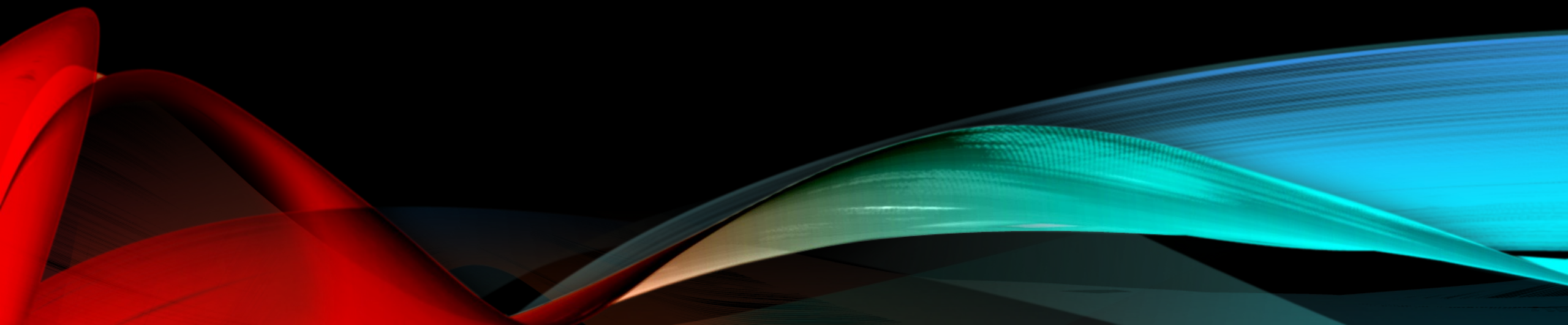
NATURAL LANGUAGE PROCESSING (NLP)

Définition et principes

Abdoul Kader KABORE
abdoulkader.kabore@protonmail.com

A PROPOS DU COURS



A PROPOS DU COURS

- Vous êtes-vous déjà demandé comment des assistants personnels IA tels que Siri ou Cortana fonctionnent ? Comment votre correcteur d'orthographe a été capable de détecter des erreurs de syntaxe que vous-même n'auriez pas repérées ? Comment votre moteur de recherche réussit à deviner les mots que vous étiez sur le point d'écrire dès les premières lettres ?
- Si ces outils sont utilisés à des fins radicalement différentes, ils reposent tous sur des méthodes communes : celles du Natural Language Processing (NLP) ou Traitement Automatique du Langage Naturel (TALN) en français.

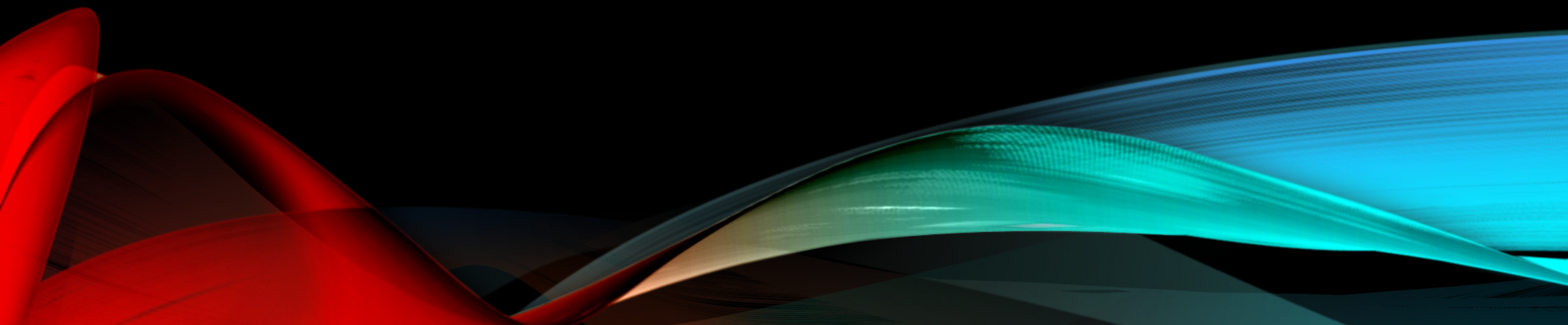
A PROPOS DU COURS

- L'objectif de ce cours est de donner un aperçu global du NLP. Plus précisément, à la fin de votre lecture, vous saurez :
  - Qu'est-ce que c'est le NLP ?
  - Quels sont les principaux domaines d'application du NLP ?
  - Quelles sont les méthodes les plus répandues en NLP ?

PLAN

1. Introduction au NLP
   - Définition NLP
   - Principaux domaines d'application du NLP
   - Méthodes les plus répandues en NLP
   - Perspectives et enjeux
2. NLP Concepts avancés
3. Travaux Pratique

QU'EST-CE QUE LE NLP ?

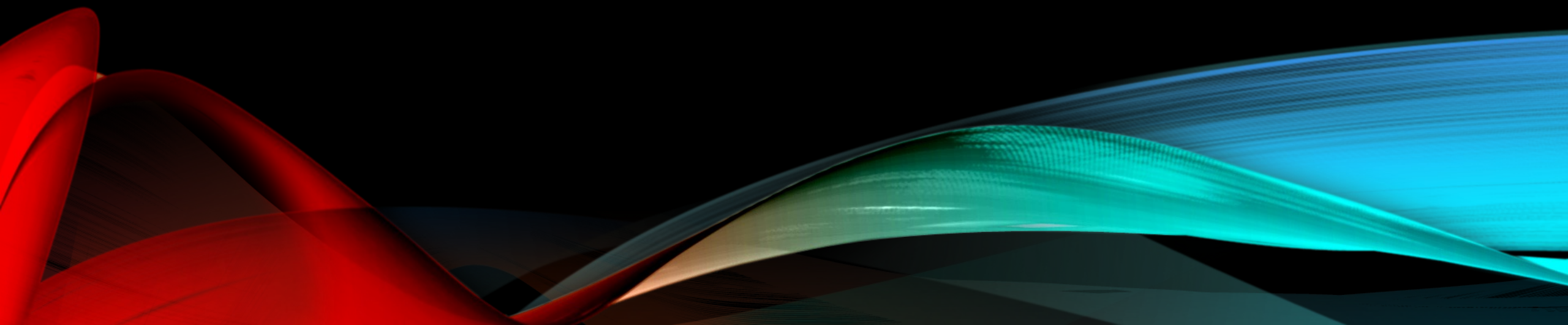


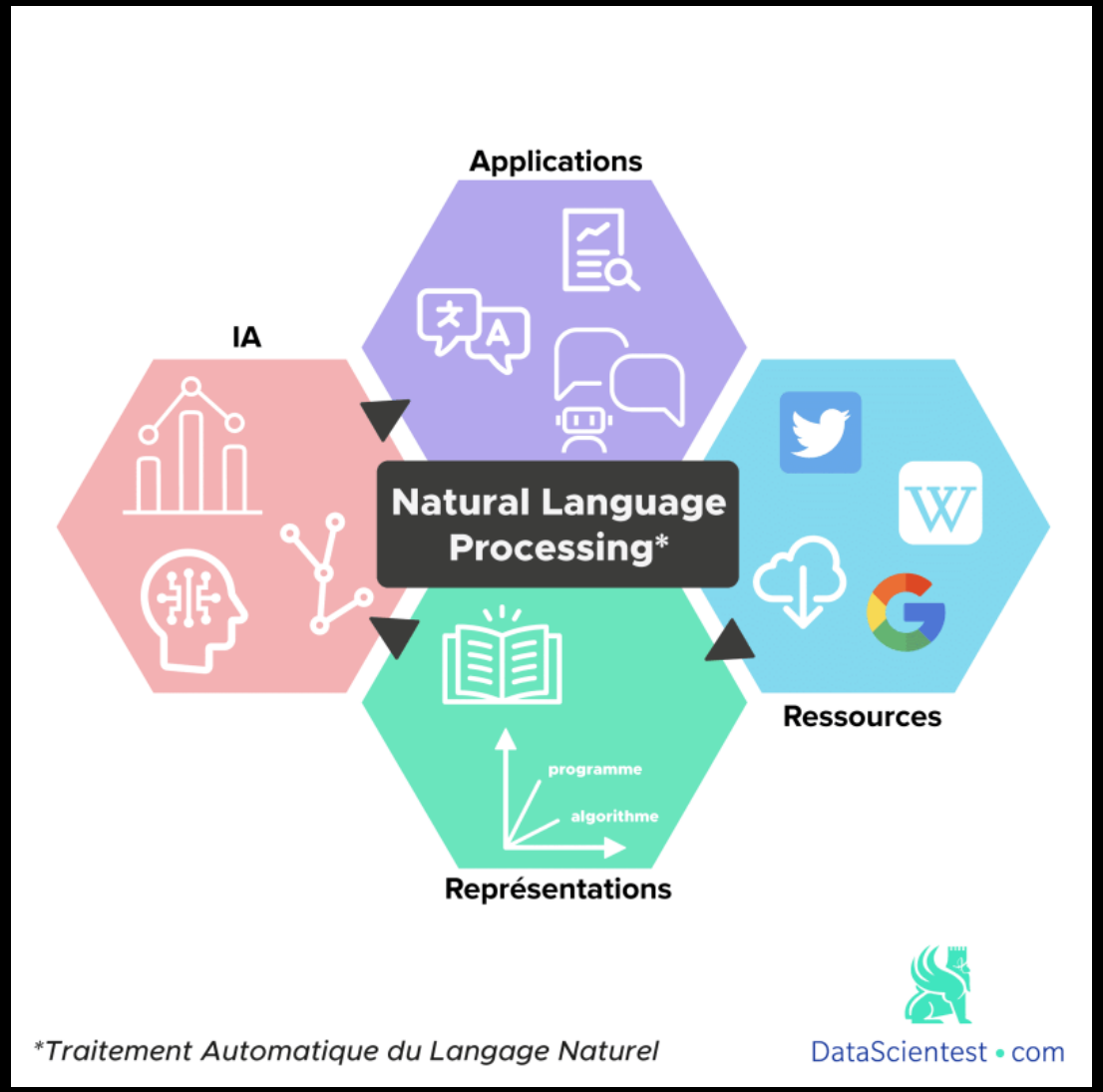
NLP : DÉFINITION

Le NLP pour Natural Language Processing ou Traitement du Langage Naturel est une discipline qui porte essentiellement sur la compréhension, la manipulation et la génération du langage naturel par les machines. Ainsi, le NLP est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.

NLP : DÉFINITION

À QUELLES PROBLÉMATIQUES RÉPOND LE NLP ?




Applications
IA
Natural Language Processing*
Ressources
programme
algorithme
Représentations
*Traitement Automatique du Langage Naturel
DataScientest • com

LES APPLICATIONS DU NLP

Le NLP est terme assez générique qui recouvre un champ d'application très vaste. Voici les applications les plus populaires :

- Traduction automatique
- Sentiment analysis
- Marketing
- Chatbots

NLP & TRADUCTION AUTOMATIQUE

Le développement d'algorithmes de traduction automatique a réellement révolutionné la manière dont les textes sont traduits aujourd'hui. Des applications, telles que Google Translator, sont capables de traduire des textes entiers sans aucune intervention humaine.

Le langage naturel étant par nature ambigu et variable, ces applications ne reposent pas sur un travail de remplacement mot à mot, mais nécessitent une véritable analyse et modélisation de texte, connue sous le nom de Traduction automatique statistique (Statistical Machine Translation en anglais).

NLP & SENTIMENT ANALYSIS

Aussi connue sous le nom de « Opinion Mining », l'analyse des sentiments consiste à identifier les informations subjectives d'un texte pour extraire l'opinion de l'auteur.

À titre exemple, lorsqu'une marque lance un nouveau produit, elle peut exploiter les commentaires recueillis sur les réseaux sociaux pour identifier le sentiment positif ou négatif globalement partagé par les clients.

NLP & SENTIMENT ANALYSIS

De manière générale, l'analyse des sentiments permet de mesurer le niveau de satisfaction des clients vis-à-vis des produits ou services fournis par une entreprise ou un organisme. Elle peut même s'avérer bien plus efficace que des méthodes classiques comme les sondages.

En effet, si l'on rechigne souvent à passer du temps à compléter de longs questionnaires, une partie croissante des consommateurs partage aujourd'hui fréquemment leurs opinions sur les réseaux sociaux. Ainsi, la recherche de textes négatifs et l'identification des principales plaintes permettent d'améliorer les produits, d'adapter la publicité et de réduire le niveau d'insatisfaction des clients.

NLP & MARKETING

Les spécialistes du marketing utilisent également le NLP pour rechercher des personnes étant susceptible d'effectuer un achat.

Ils s'appuient pour cela sur le comportement des internautes sur les sites, les réseaux sociaux et les requêtes aux moteurs de recherche. C'est grâce à ce type d'analyse que Google génère un profit non négligeable en proposant la bonne publicité aux bons internautes. Chaque fois qu'un visiteur clique sur une annonce, l'annonceur reverse jusqu'à 50 dollars !

NLP & MARKETING

De manière plus générale, les méthodes de NLP peuvent être exploitées pour dresser un portrait riche et complet du marché existant, des clients, des problèmes, de la concurrence et du potentiel de croissance des nouveaux produits et services de l'entreprise.

Les sources de données brutes pour cette analyse comprennent les journaux de ventes, les enquêtes et les médias sociaux…

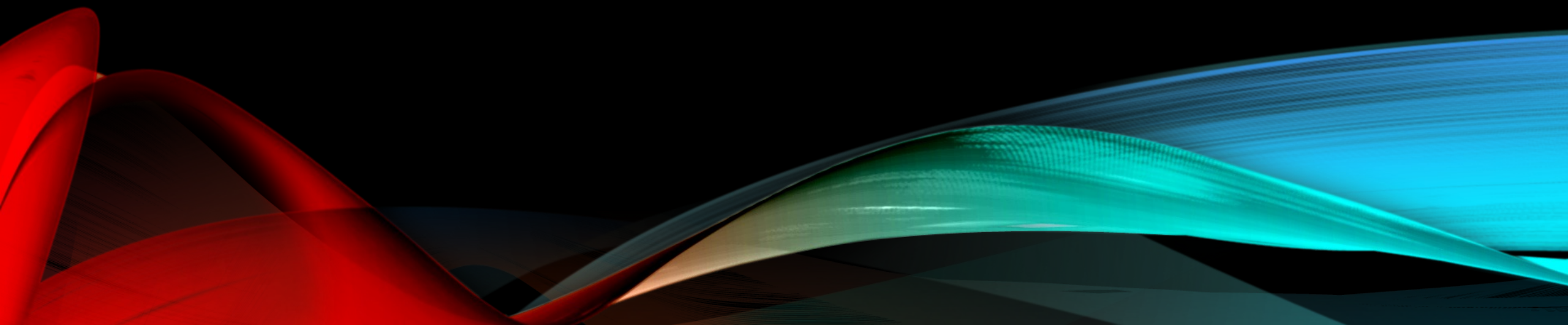
NLP & CHATBOTS

Les méthodes NLP sont au cœur du fonctionnement des Chatbots actuels. Bien que ces systèmes ne soient pas totalement parfaits, ils peuvent aujourd'hui facilement gérer des tâches standards telles renseigner des clients sur des produits ou services, répondre à leurs questions, etc. Ils sont utilisés par plusieurs canaux, dont l'Internet, les applications et les plateformes de messagerie. L'ouverture de la plateforme Facebook Messenger aux chatbots en 2016 a contribué à leur développement.

NLP : AUTRES APPLICATIONS

- Classification de texte : cela consiste à attribuer un ensemble de catégories prédéfinies à un texte donné. Les classificateurs de texte peuvent être utilisés pour organiser, structurer et catégoriser à ensemble de textes.
- **Reconnaissance de caractères** : Cela permet d'extraire, à partir de la reconnaissance des caractères, les principales informations des reçus, des factures, des chèques, des documents de facturation légaux, etc.
- **Correction automatique** : la plupart des éditeurs de texte sont aujourd'hui muni d'un correcteur orthographique qui permet de vérifier si le texte contient des fautes d'orthographe.
- **Résumé automatique** : les méthodes NLP sont également utilisées pour produire des résumés courts, précis et fluides d'un document texte plus long

QUELLES SONT LES PRINCIPALES MÉTHODES UTILISÉES EN NLP ?



NLP : MÉTHODOLOGIES

Globalement, nous pouvons distinguer deux aspects essentiels à tout problème de NLP :

- **La partie « linguistique »**, qui consiste à prétraiter et transformer les informations en entrée en un jeu de données exploitable.
- **La partie « apprentissage automatique »** ou « Data Science », qui porte sur l'application de modèles de Machine LearningouDeep Learningà ce jeu de données.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

Supposons que vous vouliez être capable de déterminer si un mail est un spam ou non, uniquement à partir de son contenu. À cette fin, il est indispensable de transformer les données brutes (le texte du mail) en des données exploitables.

Parmi les principales étapes, on retrouve :

- **Nettoyage** : Variable selon la source des données, cette phase consiste à réaliser des tâches telles que la suppression d'urls, d'emoji, etc.
- **Normalisation des données**

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

…

- **Normalisation des données** :
  - **Tokenisation**, ou découpage du texte en plusieurs pièces appelés tokens.

  Exemple : « Vous trouverez en pièce jointe le document en question » ; « Vous », « trouverez », « en pièce jointe », « le document », « en question ».

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

…

- **Normalisation des données** :
  - **Stemming** : un même mot peut se retrouver sous différentes formes en fonction du genre (masculin féminin), du nombre (singulier, pluriel), la personne (moi, toi, eux…) etc. Le stemming désigne généralement le processus heuristique brut qui consiste à découper la fin des mots dans afin de ne conserver que la racine du mot.

  Exemple : « trouverez » -> « trouv »

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

…

- **Normalisation des données** :
  - **Lemmatisation** : cela consiste à réaliser la même tâche mais en utilisant un vocabulaire et une analyse fine de la construction des mots. La lemmatisation permet donc de supprimer uniquement les terminaisons inflexibles et donc à isoler la forme canonique du mot, connue sous le nom de lemme.

  Exemple : « trouverez » -> trouvez

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

…

- **Normalisation des données** :
  - **Autres opérations** : suppression des chiffres, ponctuation, symboles et stopwords, passage en minuscule.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

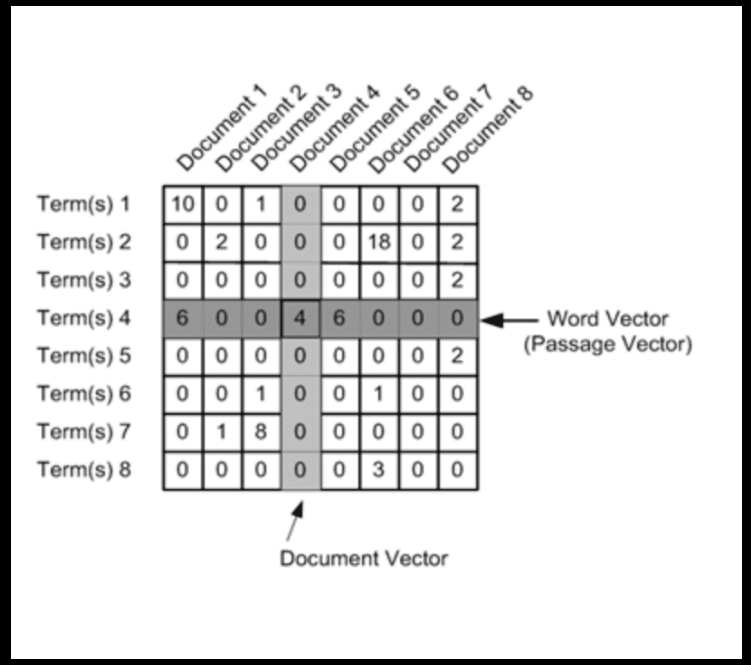
Afin de pouvoir appliquer les méthodes de Machine Learning aux problèmes relatifs au langage naturel, il est indispensable de transformer les données textuelles en données numériques.

Il existe plusieurs approches dont les principales sont les suivantes :

- **Term-Frequency (TF)**
- **Term Frequency-Inverse Document Frequency (TF-IDF)**

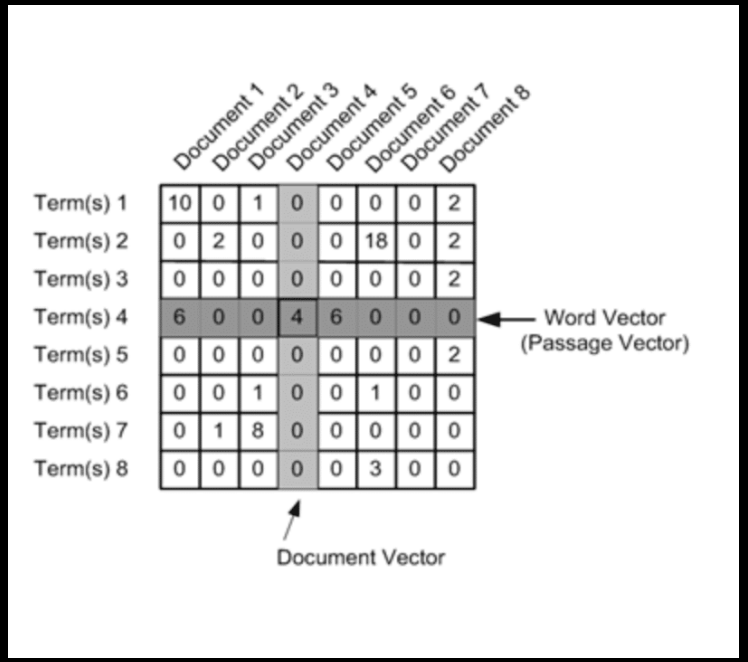
LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term-Frequency (TF)** : cette méthode consiste à compter le nombre d'occurrences des tokens présents dans le corpus pour chaque texte. Chaque texte est alors représenté par un vecteur d'occurrences. On parle généralement de **Bag-Of-Word**, ou sac de mots en français.



LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term-Frequency (TF)** : cette méthode consiste à compter le nombre d'occurrences des tokens présents dans le corpus pour chaque texte. Chaque texte est alors représenté par un vecteur d'occurrences. On parle généralement de **Bag-Of-Word**, ou sac de mots en français.



Néanmoins, cette approche présente un inconvénient majeur : certains mots sont par nature plus utilisés que d'autres, ce qui peut conduire le modèle à des résultats erronés.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term Frequency-Inverse Document Frequency (TF-IDF)** : cette méthode consiste à compter le nombre d'occurrences des tokens présents dans le corpus pour chaque texte, que l'on divise ensuite par le nombre d'occurrences total de ces même tokens dans tout le corpus.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term Frequency-Inverse Document Frequency (TF-IDF)** : …

Pour le terme x présent dans le document y, on peut définir son poids par la relation suivante :

$$w_{x,y} = tf_{x,y} \cdot log(\frac{N}{df_x})$$

ou,

- $tf_{x,y}$ est la fréquence du terme x dans y ;
- df_x est le nombre de documents contenant x ;
- N est le total de documents.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term Frequency-Inverse Document Frequency (TF-IDF)** : …

Pour le terme x présent dans le document y, on peut définir son poids par la relation suivante :

$$w_{x,y} = tf_{x,y} \cdot log(\frac{N}{df_x})$$

ou,

- $tf_{x,y}$ est la fréquence du terme x dans y ;
- df_x est le nombre de documents contenant x ;
- N est le total de documents.

Cette approche permet donc d'obtenir pour chaque texte une représentation vectorielle qui comporte des vecteurs de poids et non plus d'occurrences.

LA PHASE DE PRÉTRAITEMENT : DU TEXTE AUX DONNÉES

- **Term Frequency-Inverse Document Frequency (TF-IDF)**
- **Term Frequency-Inverse Document Frequency (TF-IDF)**

L'efficacité de ces méthodes diffère selon le cas d'application. Toutefois, elles présentent deux principales limites :

- Plus le vocabulaire du corpus est riche, plus la taille des vecteurs est grande, ce qui peut représenter un problème pour les modèles d'apprentissage utilisées dans l'étape suivante.
- Le comptage d'occurrence des mots ne permet pas de rendre compte de leur agencement et donc du sens des phrases.

LA PHASE D'APPRENTISSAGE : DES DONNÉES AU MODÈLE

De manière globale, on peut distinguer 3 principales approches NLP :

- les méthodes basées sur des **règles**,
- les méthodes basées sur des **modèles classiques de Machine Learning**
- les méthodes basées sur des **modèles de Deep Learning.**

LA PHASE D'APPRENTISSAGE : DES DONNÉES AU MODÈLE

- Les méthodes basées sur des règles :

Les méthodes fondées sur des règles reposent en grande partie sur l'élaboration de règles spécifiques à un domaine (par exemple, les expressions régulières). Elles peuvent être utilisées pour résoudre des problèmes simples tels que l'extraction de données structurées à partir de données non structurées (par exemple, les pages web).

Dans le cas de la détection de spams, cela pourrait consister à considérer comme e-mails indésirables, ceux qui comportent des buzz words tels que « promotion », « offre limitée », etc.

Néanmoins, ces méthodes simples peuvent être rapidement dépassées par la complexité du langage naturel et s'avérer être inefficace.

LA PHASE D'APPRENTISSAGE : DES DONNÉES AU MODÈLE

- Les méthodes basées sur des modèles classiques de Machine Learning

Les approches classiques d'apprentissage automatique peuvent être utilisées pour résoudre des problèmes plus difficiles. Contrairement aux méthodes fondées sur des règles prédéfinies, elles reposent sur des méthodes qui portent réellement sur la compréhension du langage. Elles exploitent les données obtenues à partir des textes bruts prétraités via une des méthodes décrites en haut par exemple. Elles peuvent également utiliser des données relatives à la longueur des phrases, à l'occurrence de mots spécifiques, etc. Elles mettent généralement en œuvre un modèle statistique d'apprentissage automatique tels que ceux de **Naive Bayes**, de **Régression Logistique**, etc.

LA PHASE D'APPRENTISSAGE : DES DONNÉES AU MODÈLE

- Les méthodes basées sur des modèles de Deep Learning.

L'utilisation de modèles **d'apprentissage en profondeur** pour les problématiques NLP fait l'objet de nombreuses recherches actuellement. Ces modèles se généralisent encore mieux que les approches classiques d'apprentissage car ils nécessitent une phase de prétraitement du texte moins sophistiquée : les couches de neurones peuvent être perçues comme des extracteurs automatiques de features.

LA PHASE D'APPRENTISSAGE : DES DONNÉES AU MODÈLE

- Les méthodes basées sur des modèles de Deep Learning.

...

Cela permet alors de construire des modèles de bout en bout avec peu de prétraitement des données. En dehors de la partie feature engineering, les capacités d'apprentissage des algorithmes de Deep Learning sont généralement plus puissantes que celles de Machine Learning classique, ce qui permet d'obtenir de meilleurs scores sur différentes tâches complexes de NLP dures telles que la traduction.