

# Natural Language Processing (NLP)

## Plan du chapitre

1. Aperçu sur le NLP
2. Avantages / inconvénients du NLP
3. Domaines d'application du NLP
4. Fonctionnement d'un système NLP
5. Les principaux défis du traitement du langage naturel

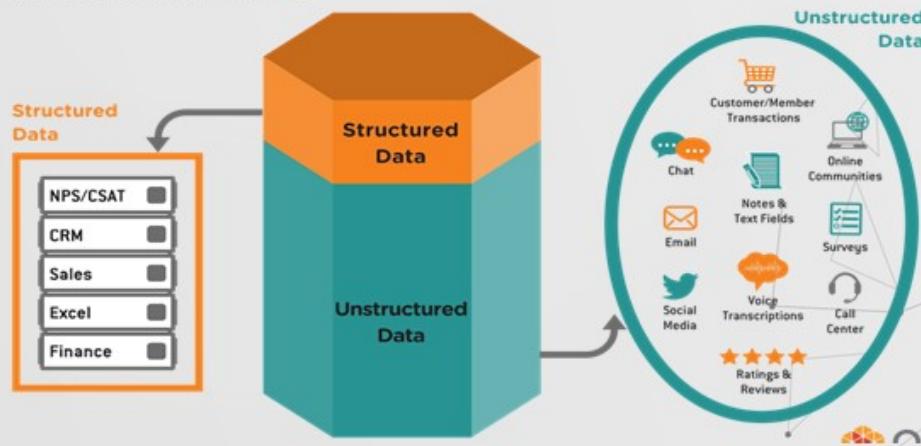
# 01

## Aperçu sur le NLP



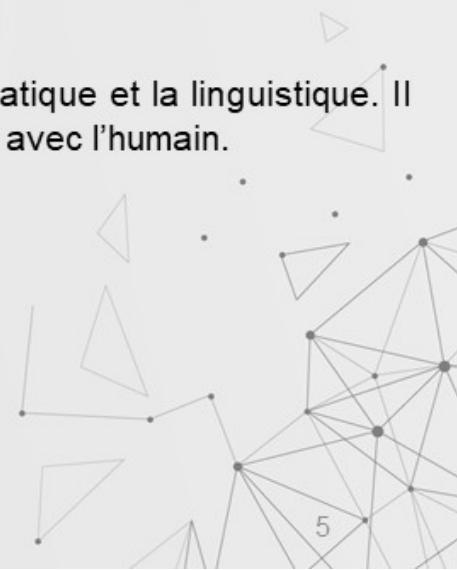
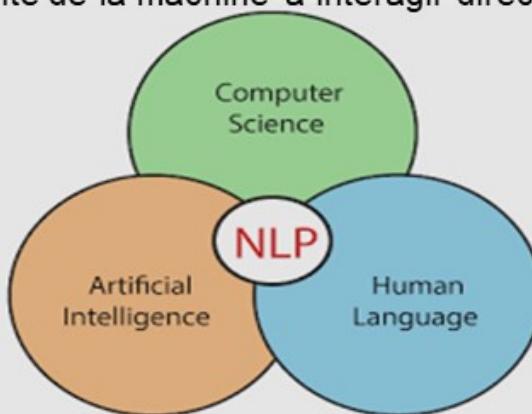
## Introduction

- Selon les estimations de l'industrie, seulement 20% des données disponibles sont présentes sous la forme structurée, les 80% restants représentent quant à eux des données sous forme non structurée.
- Aujourd'hui, la majorité des données existe sous forme textuelle, non structurée, c'est pourquoi il est important de se familiariser avec les techniques d'analyse de texte et le **traitement du langage naturel**, afin de produire des informations significatives et exploitables à partir de ces données.



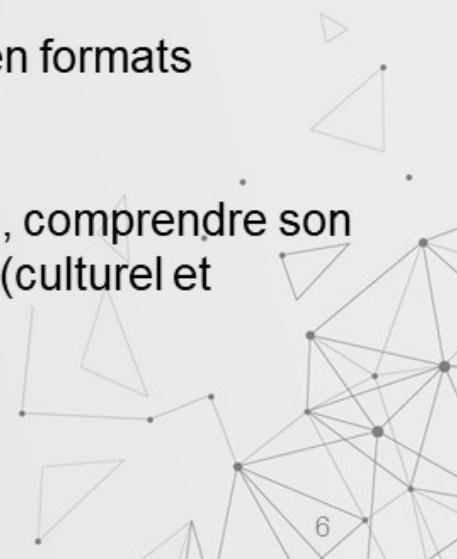
# Qu'est-ce que le traitement de langage naturel ?

- Le traitement du langage naturel (**NLP**) ou Traitement Automatisé du Langage Naturel est le domaine de l'IA et une branche de la Data Science.
- Elle 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.



## Petite précision sur le langage naturel en informatique

- En informatique, le langage naturel se réfère au **langage utilisé quotidiennement par les humains** pour communiquer (par opposition aux langages de programmation, qui impliquent des lignes de code).
- Le NLP traite ce langage naturel en le convertissant en formats compréhensibles par les machines.
- Il utilise diverses techniques pour analyser le langage, comprendre son sens, sa syntaxe, sa sémantique et même son contexte (culturel et émotionnel).



# Petite précision sur le langage naturel en informatique

- Le traitement du langage naturel permet de créer des logiciels qui peuvent:
  - Analyser et interpréter du texte dans des documents, des e-mails et d'autres sources.
  - Interpréter le langage parlé et synthétiser les réponses vocales.
  - Traduire automatiquement des phrases parlées ou écrites entre différentes langues.
  - Interpréter des commandes et déterminer des actions appropriées.
- Le NLP ouvre donc un monde de possibilités pour les **interactions homme-machine**, en rapprochant les ordinateurs de la compréhension du langage humain dans toute sa complexité.

## Composantes de la NLP

Il existe deux composants du traitement du langage naturel :

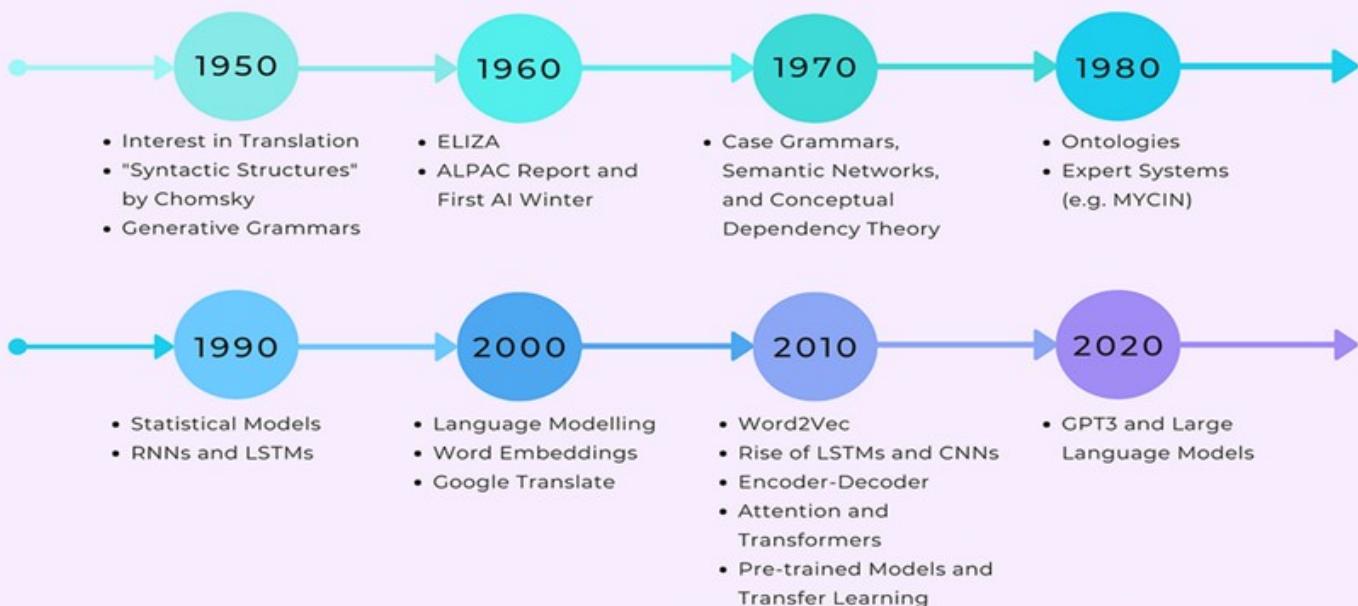
- **Compréhension du langage naturel (NLU)** La NLU permet aux machines de comprendre et d'interpréter le langage humain tel qu'il est parlé ou écrit.
- **Génération de langage naturel (NLG)**: Dans le NLG se concentre la génération d'un langage naturel à partir de données structurées



# NLP vs NLU vs NLG

Characteristic	NLP	NLU	NLG
<b>Presentation</b>	NLP is computers <b>reading</b> language	NLU is computers <b>understanding</b> language	NLG is computers <b>writing</b> language
<b>Focus</b>	Processing and analyzing language data	Interpreting and understanding language input.	Producing coherent and contextually appropriate text or speech.
<b>Input</b>	Text or speech data	Text or speech data	Structured data or instructions
<b>Output</b>	It converts unstructured data to Structured data	It reads data and to structured data	NLG writes structured data
<b>Application</b>	<ul style="list-style-type: none"> <li>- Smart assistance</li> <li>- Language translation</li> <li>- Text analysis</li> </ul>	<ul style="list-style-type: none"> <li>- Speech recognition</li> <li>- Sentiment analysis</li> </ul>	<ul style="list-style-type: none"> <li>- Chatbots</li> <li>- Voice assistance</li> </ul>

## Histoire de la NLP





# 02

## Avantages / inconvénients du NLP

11

### Avantages du NLP

#### Amélioration de l'efficacité:

- **Automatisation des tâches:** Le NLP peut automatiser des tâches répétitives et chronophages, telles que la synthèse de texte, la classification de documents, la traduction automatique, et la recherche d'informations. Cela permet aux humains de se concentrer sur des tâches plus complexes et à plus forte valeur ajoutée.
- **Gain de temps et d'argent:** L'automatisation des tâches permet de réduire les coûts et d'améliorer la productivité.

#### Meilleure compréhension des données:

- **Extraction d'informations:** Le NLP permet d'extraire des informations importantes de grandes quantités de données textuelles, telles que des avis clients, des articles de presse, des documents juridiques, etc.

12

# Avantages du NLP

- **Analyse des sentiments:** Le NLP permet d'analyser le sentiment d'un texte, c'est-à-dire l'opinion ou l'émotion qui s'en dégage. Cela peut être utile pour les entreprises qui veulent comprendre la perception de leur marque par les clients.
- **Détection des anomalies:** Le NLP peut être utilisé pour détecter des anomalies dans les données textuelles, comme des fraudes ou des erreurs.  
Amélioration de la communication et de l'interaction:
- **Chatbots et assistants virtuels:** Le NLP permet de créer des chatbots et des assistants virtuels qui peuvent dialoguer avec les humains de manière naturelle et fluide.
- **Personnalisation des interactions:** Le NLP peut être utilisé pour personnaliser les interactions avec les clients en fonction de leurs besoins et de leurs préférences.

# Avantages du NLP

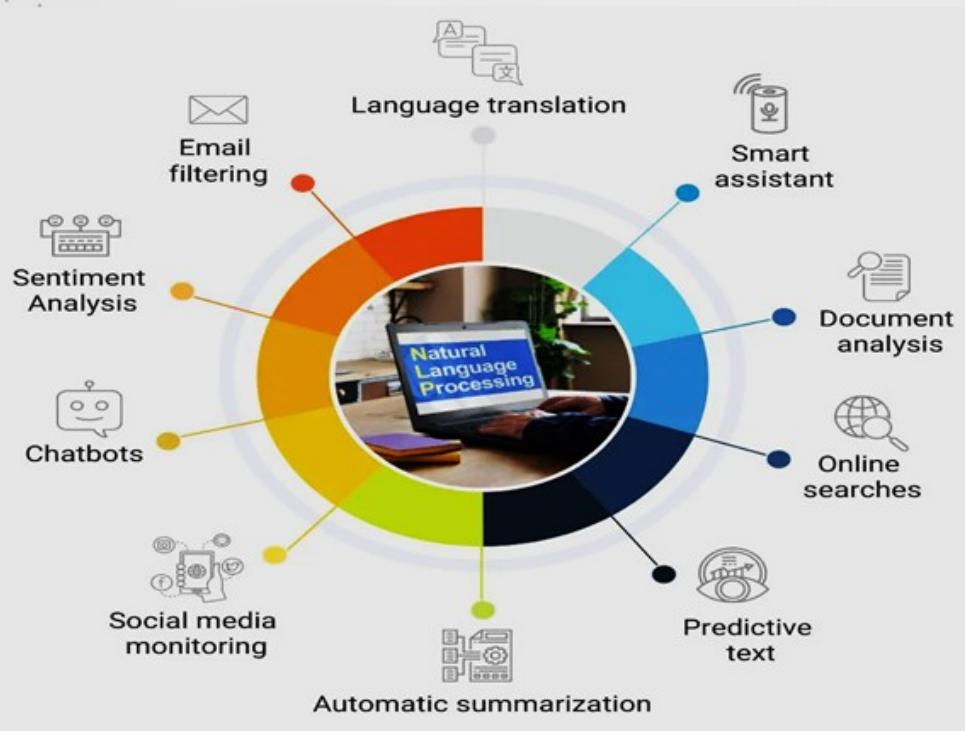
- **Amélioration de l'accessibilité:** Le NLP peut être utilisé pour rendre les informations et les services plus accessibles aux personnes handicapées.  
Innovation et développement de nouveaux produits et services:
- **Développement de nouveaux produits:** Le NLP peut être utilisé pour développer de nouveaux produits et services qui répondent aux besoins des clients.
- **Amélioration des produits existants:** Le NLP peut être utilisé pour améliorer les produits existants en les rendant plus intelligents et plus faciles à utiliser.
- **Développement de nouvelles technologies:** Le NLP est un domaine de recherche en pleine expansion qui permet de développer de nouvelles technologies révolutionnaires.

## Inconvénients de la NLP

- ❑ Pour l'apprentissage du modèle NLP, de nombreuses données et calculs sont nécessaires.
- ❑ De nombreux problèmes se posent en NLP lorsqu'il s'agit d'expressions informelles, de dialectes et de jargon culturel.
- ❑ Les résultats de la NLP ne sont parfois pas précis et l'exactitude est directement proportionnelle à l'exactitude des données.
- ❑ La NLP est conçue pour un travail unique et restreint car elle ne peut pas s'adapter à de nouveaux domaines et a une fonction limitée.

## 03 Domaines d'application du NLP

# Applications du NLP



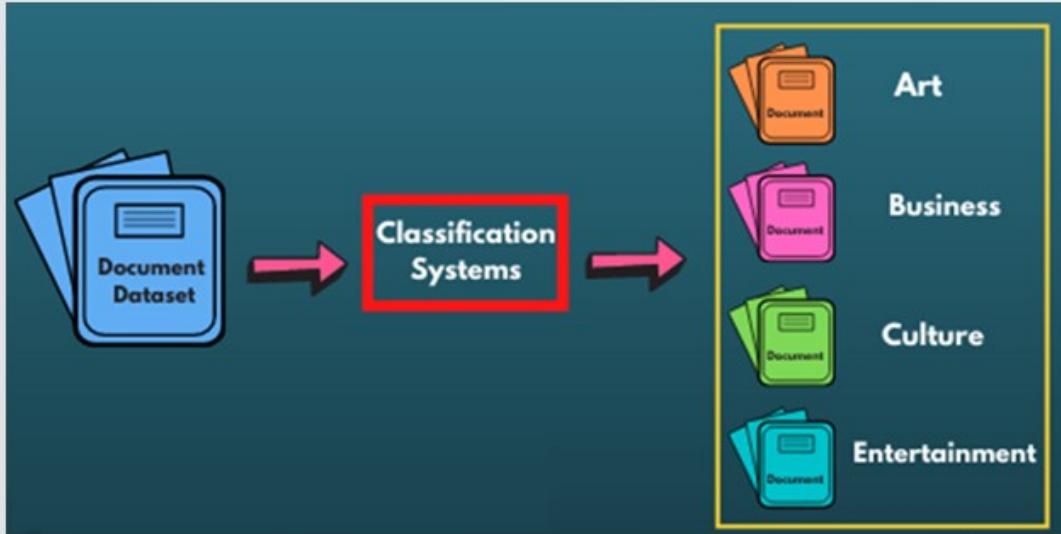
## Traitement du texte et de la parole

- Traitement du texte et de la parole comme les assistants vocaux – **Alexa, Siri, etc.**



# Classification de texte

- Classification de texte comme **Grammarly**, **Voyant Tools**, **KH Coder** et **Google Docs**



# Extraction d'informations

- Un logiciel d'extraction de données aide les entreprises à collecter des informations depuis des sites web, des fichiers PDF et des fichiers texte sur des disques locaux. Le type d'informations qui peut être extrait varie du texte aux images, en passant par le résumé de contenu. Un logiciel pour extraire des données permet souvent d'organiser les données extraites dans un document Word formaté ou un tableau Excel.
- Exemple: les moteurs de recherche comme **DuckDuckGo**, **Google**



## Outils et logiciels d'extraction de données

Nom de l'outil	Intégrations	Format de fichier	Essai gratuit
👉 Bright Data	AdsPower, PhantomBuster, SessionBox, Apify, etc.	CSV, Email, HTML, JSON et API	7 Jours
Apifier	PHP, Node.js et proxy	JSON, XML, Excel et CSV	30 Jours
ScrapingBee	Feuilles Google, Google Drive, Airtable, Slack, Bot télégramme	CSV, PDF, etc.	1000 appels API
ScraperAPI	Gratterbox, Marionnettiste NodeJS, Sélénium, etc.	HTML, XML ou JSON, etc.	7 Jours
DocParser	Feuilles de calcul Google et Salesforce	JSON, CSV ou XML	21 Jours

## Chatbots et assistants virtuels

- ❑ Les utilisateurs peuvent avoir des conversations avec leur système. Ce sont des outils de service client courants. Ils peuvent également guider les utilisateurs à travers des workflows compliqués ou les aider à naviguer sur un site ou une solution.



# Chatbots vs assistants virtuels

Paramètres de comparaison	Chatbot	Assistant virtuel
Fonctionnalité	Fonctionnalité limitée. Ils sont généralement conçus pour effectuer une série de tâches spécifiques, telles que répondre aux questions fréquemment posées, faciliter des transactions simples ou fournir des informations de base.	fonctionnalité étendue. Les assistants virtuels, tels que Millie, Siri, Alexa ou Google Assistant, ont des capacités plus larges et peuvent effectuer une variété de tâches, du réglage des alarmes à la lecture de musique en passant par la réponse aux questions générales.
Interaction	interaction textuelle. Bien que certains chatbots puissent prendre en charge les commandes vocales, votre principale interaction se fait généralement par le texte.	interaction vocale. Bien qu'ils puissent également interagir par le texte, les assistants virtuels seraient principalement conçus pour être contrôlés par la voix.
Contexte	contexte limité. Les chatbots peuvent avoir une compréhension limitée du contexte d'une conversation.	compréhension contextuelle. Les assistants virtuels ont souvent une meilleure compréhension du contexte et peuvent gérer des commandes et des questions plus complexes.
Intégration	Intégration dans des canaux spécifiques: Les chatbots se trouvent souvent sur des sites Web, des applications mobiles ou des plateformes de messagerie comme Slack ou Facebook Messenger.	intégration multiplateforme. Les assistants virtuels sont souvent intégrés à divers appareils et plates-formes, tels que smartphones, haut-parleurs intelligents et ordinateurs.
Application pratique	Les chatbots assistent les entreprises et les clients.	Les assistants virtuels aident leurs utilisateurs à effectuer des tâches telles que planifier des réunions, répondre aux e-mails, définir des rappels, etc.

## Les FAQs dynamiques

- Ces modèles de Foire Aux Questions fonctionnent comme des moteurs de recherche. Les utilisateurs tapent une question ou un mot clé, puis l'outil détecte automatiquement les réponses.
- les meilleurs outils et logiciels de questions fréquentes

- Zendesk Sell
- Wix Answers
- HappyFox
- Zoho Desk
- SupportBee
- HelpScout
- GrooveHQ



# Traduction automatique

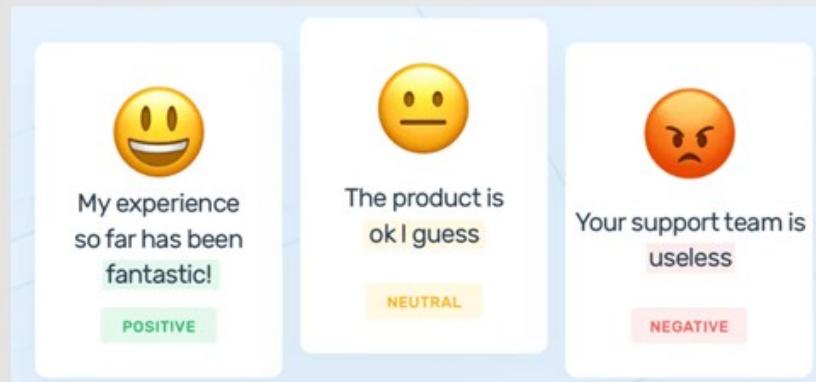
- Le développement d'**algorithmes de traduction automatique** a réellement révolutionné la manière dont les textes sont traduits aujourd'hui. Des applications, telles que **Google Translator**, sont capables de **traduire des textes entiers** sans aucune intervention humaine.
- Le langage naturel étant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



# Analyse des sentiments

- 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.



# Analyse des sentiments

- 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 se plaint 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.

"Mon expérience est fantastique"	Positif
'Le produit est correct, je suppose'	Neutre
"Votre équipe d'assistance est inutile"	Négatif

## Correction automatique

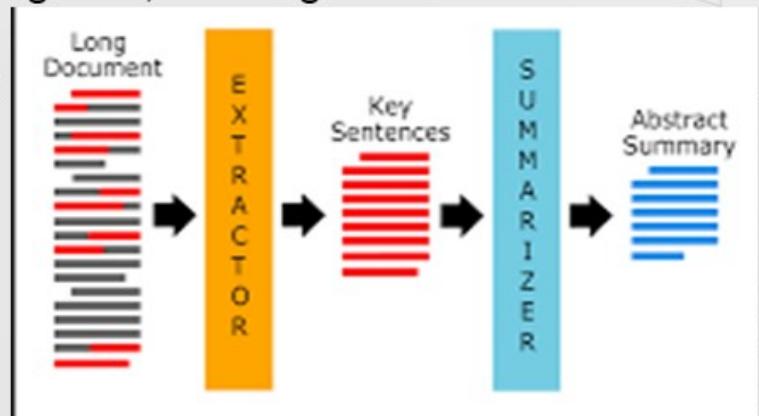
- La correction automatique est devenue courante dans la plupart des éditeurs de texte (on l'utilise tous les jours).
- Les outils de correction orthographique proposent des suggestions de corrections en temps réel pendant la rédaction, améliorant ainsi la qualité et l'exactitude du texte final.
- Les 5 meilleurs correcteurs d'orthographe en ligne à utiliser quotidiennement

- ANTIDOTE
- BONPATRON
- CORDIAL-ENLIGNE
- SCRIBENS
- REVERSO



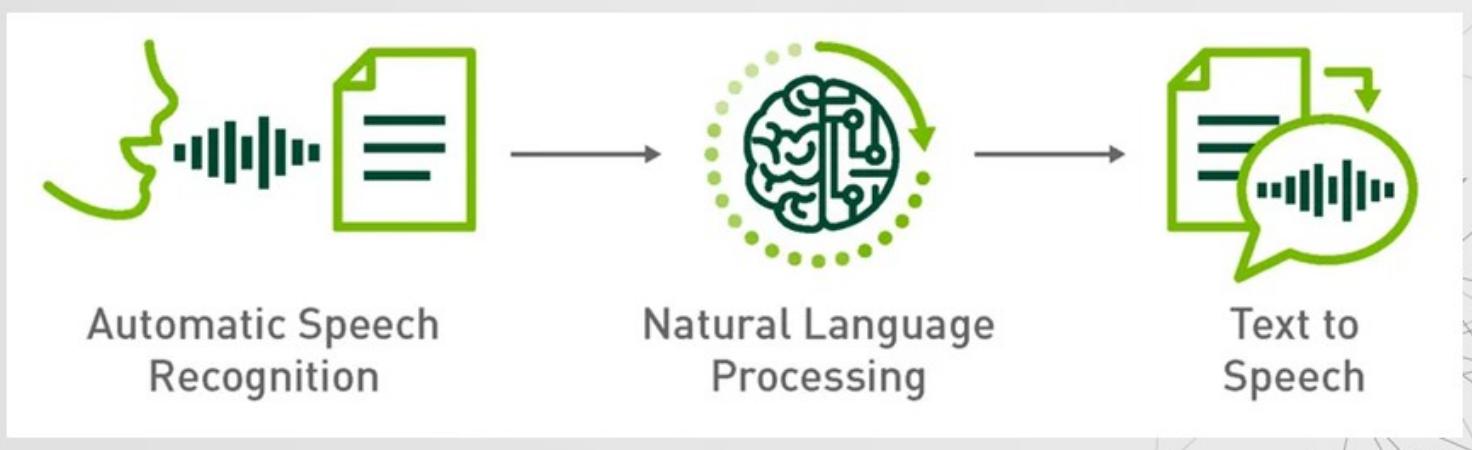
# Résumé automatique

- Les technologies de NLP sont également exploitées pour générer des résumés concis de documents texte plus longs. Cette fonctionnalité est très utile pour les outils de veille et de recherche documentaire, permettant aux utilisateurs de gagner du temps en obtenant des informations essentielles rapidement.
- Cinq outils pour résumer automatiquement un texte
  - Summarize Bot : dans Slack et Facebook Messenger, multilingue
  - Resoomer : en ligne + extension navigateur, multilingue
  - SMMRY : en ligne, anglais
  - Text Summarizer : en ligne, anglais
  - Text Compactor : en ligne, anglais



# Assistance virtuelle des agents ou des employés

- L'assistance virtuelle dans l'exploitation des communications écrites ou orales. Par exemple, en offrant un modèle de libre-service, un routage intelligent, une aide à la recherche dans les documents, etc.



# Identification de courriers indésirables

- Le NLP joue un rôle central dans la création de systèmes de détection du **spam sophistiqués et efficaces**. Il permet de reconnaître les signaux subtils qui distinguent les courriels légitimes des indésirables. Les algorithmes examinent le contenu textuel pour rechercher des schémas linguistiques suspects.
- En se basant sur les modèles de machine learning, le **NLP effectue une classification automatique permettant leur identification**. L'analyse sémantique aide à repérer les formulations typiques du spam.



# 04

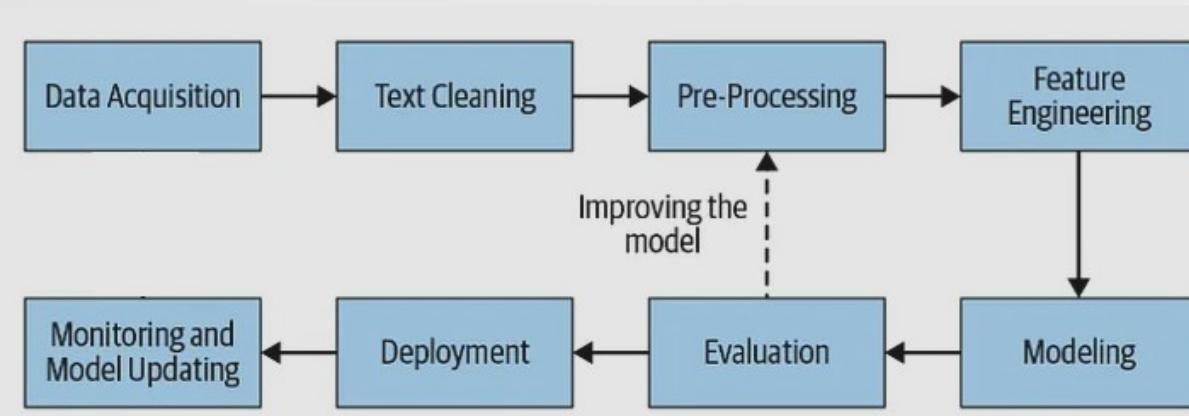
## Fonctionnement d'un système NLP

# Fonctionnement du NLP

- Le traitement du langage naturel fonctionne de plusieurs manières.
  - **Le NLP basé sur l'IA** implique l'utilisation d'algorithmes et de techniques de Machine Learning pour traiter, comprendre et générer le langage humain.
  - **Le NLP basé sur des règles** implique la création d'un ensemble de règles ou de schémas pouvant être utilisés pour analyser et générer des données de langage.
  - **Le NLP statistique** implique l'utilisation de modèles statistiques dérivés de grands ensembles de données pour analyser et faire des prédictions sur le langage.
  - **Le NLP hybride** combine ces trois approches.

# Fonctionnement du NLP

- L'approche du NLP basé sur l'IA est actuellement la plus populaire. Comme pour toute autre approche d'apprentissage pilotée par les données, le développement d'un modèle de NLP nécessite de prétraiter des données textuelles et de sélectionner minutieusement l'algorithme d'apprentissage.



# Etape 1: Acquisition des données

- **La première étape** du processus de développement de tout système PNL consiste à collecter des données pertinentes pour la tâche donnée.
- Supposons que l'on nous demande de développer un système NLP pour identifier si une **requête client entrante** (par exemple, via une interface de chat) est une demande commerciale ou une demande de service client. Selon le type de requête, elle doit être automatiquement acheminée vers la bonne équipe. Comment peut-on construire un tel système ? Eh bien, la réponse dépend du type et de la quantité de données avec lesquelles nous devons travailler.

# Etape 2: Nettoyage du texte

- **Le nettoyage de texte** est une étape essentielle du traitement du langage naturel (NLP) qui consiste à affiner et à préparer les données textuelles pour l'analyse.
- Les données de texte brut provenant de diverses sources peuvent être désordonnées, contenant des caractères non pertinents, des signes de ponctuation et des incohérences de formatage.
- Le nettoyage du texte vise à normaliser et à simplifier le texte, en garantissant qu'il soit prêt pour une analyse plus approfondie.
- Ce processus comprend généralement des tâches telles que la suppression des caractères spéciaux, des chiffres et des espaces blancs en excès, ainsi que la gestion des majuscules et de la radicalisation ou de la lemmatisation pour réduire les mots à leur forme racine. Un nettoyage de texte efficace améliore non seulement la précision des modèles NLP, mais rend également les données plus lisibles et interprétables pour des informations significatives.

HTML Parsing  
and cleanup

Unicode  
normalization

Spelling correction

System specific error  
correction

## Etape 3: Prétraitement des données

- Le prétraitement des données est essentiel pour que les machines puissent en tirer profit. **Elle permet également de mettre en évidence les caractéristiques du texte que l'algorithme peut employer.** Différentes approches sont déployées pour transformer les données textuelles brutes en données exploitables.

**Des langages de programmation comme Python ou R sont souvent utilisés pour ces techniques.** Ils sont en effet dotés de bibliothèques contenant des règles et des méthodes pour classer les segments de texte.

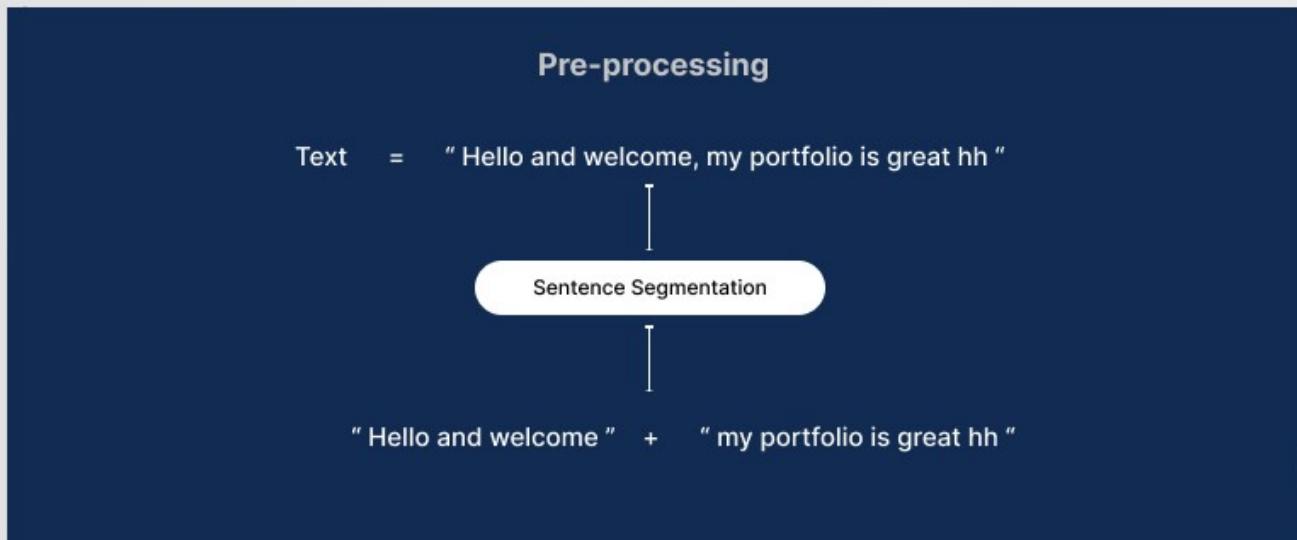
## Etape 3: Prétraitement des données

- Cette partie comprend des sous-étapes importantes : **segmentation des phrases, tokenisation des mots, radicalisation et lemmatisation,...**
  - **la tokenisation** : cette pratique fractionne le texte en phrases ou en mots (**tokens**), éliminant ainsi les caractères indésirables tels que les ponctuations.
  - une fois que nous avons divisé le texte en phrases, nous approfondissons la division de ces phrases en mots individuels.
  - **Stemming**, consiste à réduire les mots à leurs formes racines. C'est comme simplifier des mots en supprimant des parties supplémentaires (les préfixes et suffixes). Par exemple, «running» devient «run» et «jumps» devient «jump». La racine aide l'ordinateur à regrouper des mots similaires, même s'ils ont des terminaisons différentes.

## Etape 3: Prétraitement des données

- **La lemmatisation**, revient à trouver la forme d'un dictionnaire d'un mot. Il s'agit de simplifier les mots en trouvant leur forme de base et originale. Par exemple, il transforme « running » en « run » et « better » reste « better ». La lemmatisation aide les ordinateurs à comprendre le sens principal des mots.
- **La suppression des Stop Words** : les pronoms, les articles et les prépositions sont écartés pour accélérer le traitement.

## Segmentation des phrases



# Tokenisation des mots

" my portfolio is great hh "

Word Tokenization

my + portfolio + is + great + hh

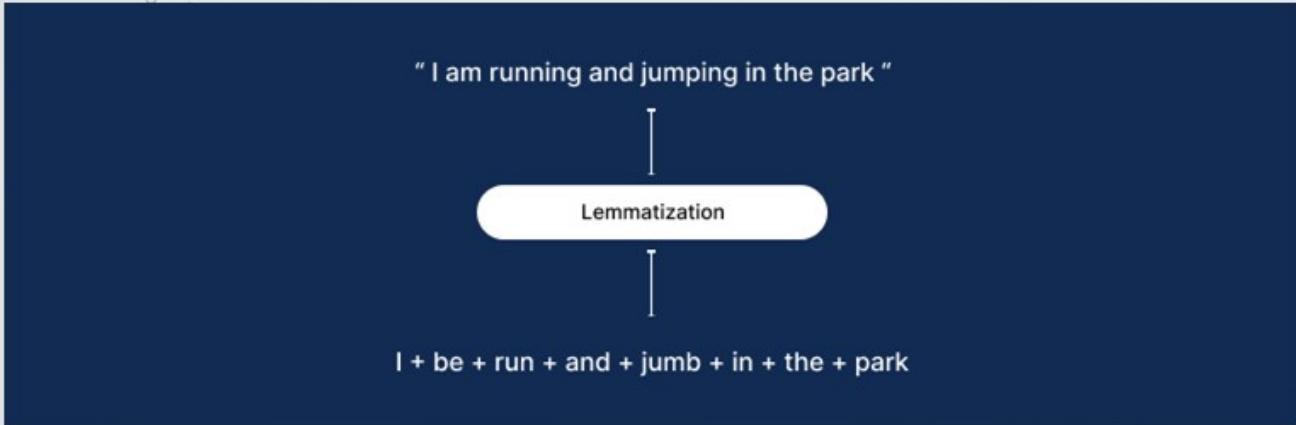
# Stemming

" Reading my blog posts while eating breakfast "

Stemming

read + my + blog + post + while + eat + breakfast

# Lemmatization



## Etape 4: Feature Engineering (Ingénierie des fonctionnalités)

- ❑ Lorsque nous utiliserons des méthodes ML pour effectuer notre étape de modélisation, nous aurons toujours besoin d'un moyen d'introduire ce texte prétraité dans un algorithme ML.
- ❑ Feature Engineering fait référence à l'ensemble des méthodes qui permettront d'accomplir cette tâche.
- ❑ L'objectif de feature engineering est de capturer les caractéristiques du texte dans un vecteur numérique pouvant être compris par les algorithmes ML.
- ❑ Ces méthodes incluent le **one-hot encoding**, qui convertit les variables catégorielles en colonnes binaires ; **label encoding**, attribution d'étiquettes numériques aux catégories ; **Bag of Words (BOW)**, qui compte la fréquence des mots dans le texte ; **Bag of N-grams**, capturant des séquences de mots ; TF-IDF (Term Frequency-Inverse Document Frequency), mesurant l'importance des mots ; et l'intégration de mots, représentant les mots comme des vecteurs dans un espace continu.

## Etape 5: Modélisation

- La construction du modèle dépend de votre problème ; choisir le meilleur modèle nécessite de bonnes connaissances en ML et en deep learning.
- Il existe trois approches principales.
  - **Le Machine Learning** : il implique l'utilisation de données d'entraînement pour permettre à un système un apprentissage automatique.
  - **Le Deep Learning** : il recourt à des réseaux de neurones avec plusieurs couches. Cette méthode traite des tâches complexes comme la traduction grâce à l'apprentissage itératif.
  - **Les règles linguistiques** : leur fonction est de résoudre des défis élémentaires, tels que l'élaboration de données structurées à partir de contenus non structurés.

## Etape 6: Évaluation

- Dans cette étape, nous déterminons essentiellement les performances du modèle sur des données invisibles.
- Cela dépend beaucoup de la métrique sélectionnée pour l'évaluation et du processus d'évaluation lui-même.
- Quelques évaluations effectuées pour le modèle PNL sont répertoriées ci-dessous :

Accuracy

Precision

Recall

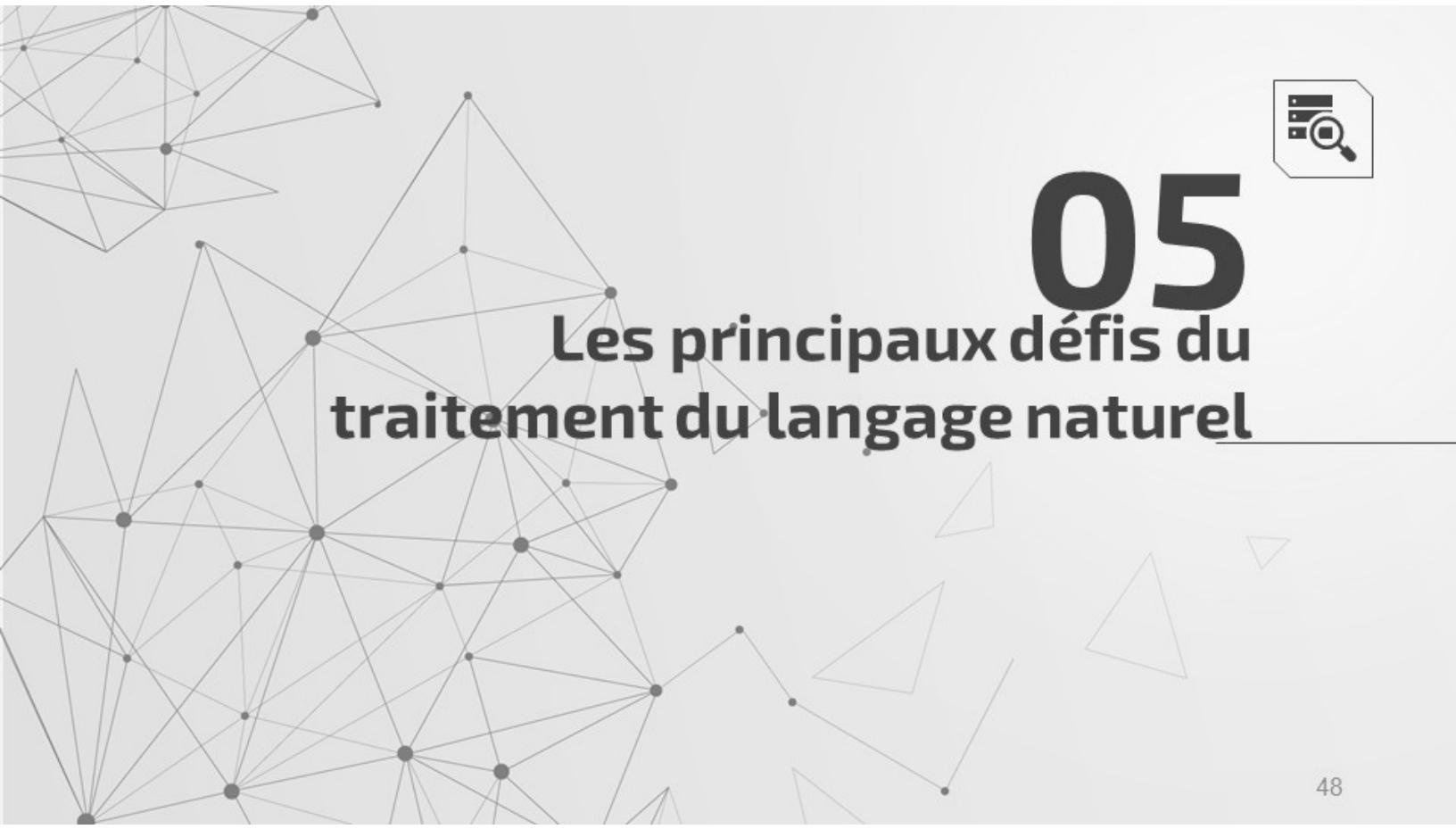
F1 Score

AUC

Mean  
Reciprocal  
Rank

## Etape 7-8: Déploiement et surveillance

- Le déploiement d'un logiciel NLP se fait généralement sous la forme d'un service **Web** ou **REST** pouvant être consommé par les utilisateurs ou d'autres services.
- Les performances d'un modèle déployé sont surveillées en permanence. Ce type de surveillance est différent des outils de surveillance logiciels traditionnels.
- Une fois qu'on commence à collecter davantage de données, on doit mettre à jour le modèle déployé afin que les données les plus récentes soient également prises en compte par le modèle lors des prédictions.



**05**  
**Les principaux défis du  
traitement du langage naturel**



# les principaux défis

- Le NLP présente plusieurs défis qui reflètent la complexité du langage humain et la technologie sous-jacente du programme.
  - **La diversité du langage humain** forme l'un des principaux obstacles. Les langues ont des structures et des règles variées, avec des dialectes, des idiomes et des niveaux de formalité différents. Adapter les modèles NLP à ces variations exige des efforts considérables.
  - **La subjectivité du langage** incarne un défi majeur. Les nuances, les métaphores et les connotations de chaque langue rendent difficile la compréhension du contexte émotionnel et intentionnel des phrases. Cela provoque parfois des interprétations erronées.

# les principaux défis

- **Le manque de données** constitue une entrave pour les langues moins courantes et les domaines de spécialisation. Les modèles NLP requièrent des ensembles de données volumineux pour l'apprentissage automatique. Cette exigence limite leur efficacité pour les langues moins fréquemment utilisées. La complexité des systèmes NLP est un autre défi.
- **Les systèmes NLP combinent diverses techniques** telles que l'analyse syntaxique, sémantique et la génération de langage naturel. Elles rendent leur développement et leur maintenance délicats.
- **Le coût des systèmes NLP** est également un obstacle important. L'entraînement et le déploiement de modèles NLP sophistiqués réclament des ressources informatiques considérables et des compétences techniques spécialisées dans les programmes informatiques. Le budget nécessaire peut rendre l'accès à ces technologies difficile pour certaines organisations.