Université Constantine 2
جامعة قسنطينة 2

Web Analytics and Natural Language Processing (WANLP)

Chapitre 02

Concepts avancée du NLP : Prétraitement de texte

Professeur BOURAMOUL Abdelkrim

Département IFA, Faculté NTIC

abdelkrim.bouramoul@univ-constantine2.dz

www.bouramoul.com

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
NTIC	IFA	Master 1	SDIA

Plan du Cours

Section 1 : Importance et défits du NLP

- Importance du TALN dans le domaine de l'IA.
- Principaux défis du TAL : ambiguïté, polysémie, Variabilité linguistique, Compréhension du contexte, Traitement de l'information non structurée
- Exemples illustratifs

Section 2 : Prétraitement de texte en NLP

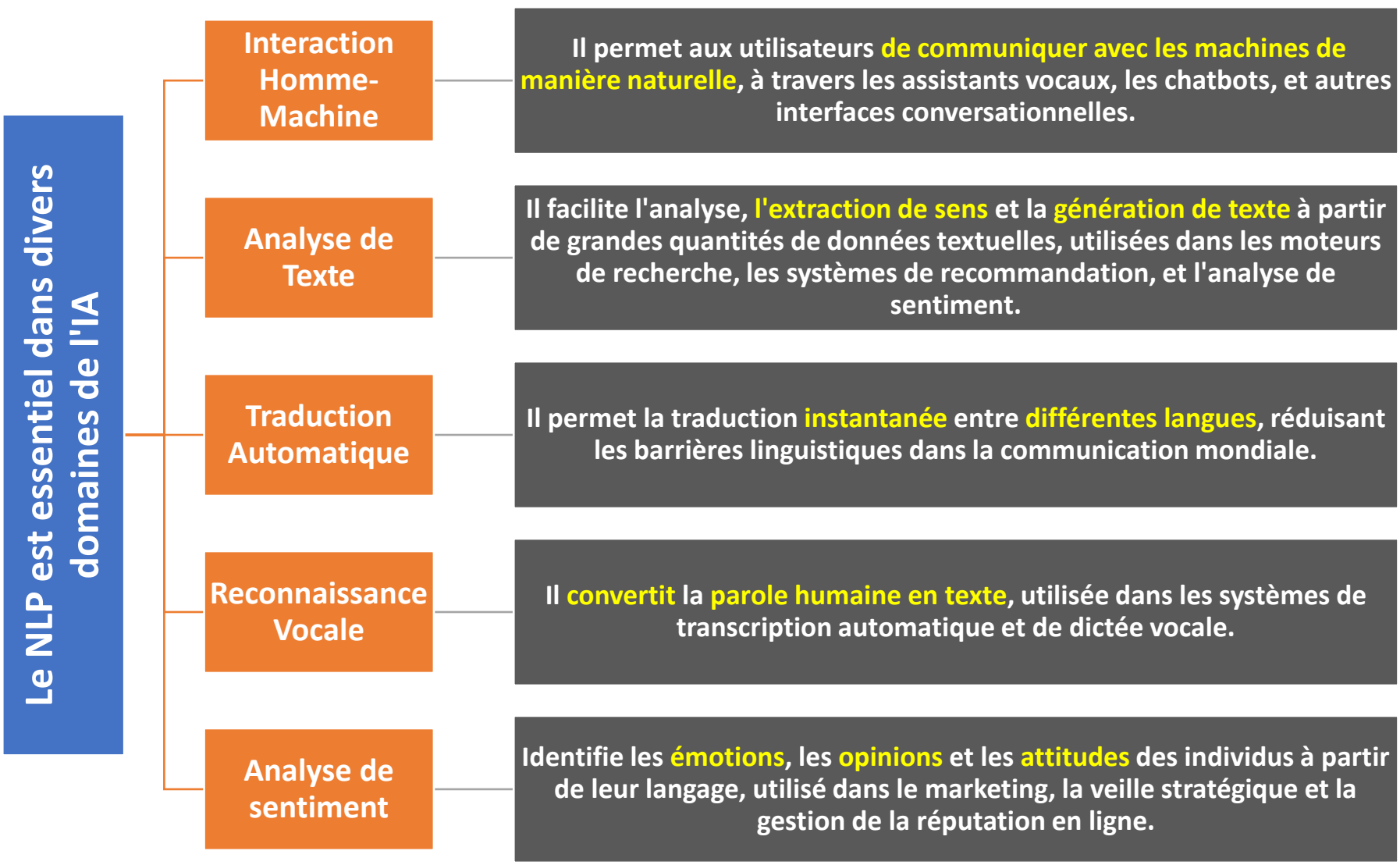
- Processus de NLP
- Notion du prétraitement de texte dans le NLP.
- Importance du prétraitement de texte dans le NLP.
- Tokenisation, Normalisation et Suppression du bruit.
- Illustration du prétraitement sur l'exemple
- Techniques de prétraitement avancées : Stemming, Lemmatisation.

Section 3 : Analyse lexicale, syntaxique et sémantique en NLP

Section 4 : Tâches fondamentales de l'NLP

Section 1 :
Importance et Défis du Traitement Automatique du Langage

Importance du NLP dans le domaine de l'IA



Principaux Défis du NLP

Ambiguïté

- Les mots et les phrases peuvent avoir plusieurs significations différentes, rendant la compréhension difficile.

Polysémie

- Certains mots peuvent avoir plusieurs sens, ajoutant une complexité à l'interprétation du langage.

Variabilité Linguistique

- Les langues naturelles sont diverses, avec des dialectes, des jargons et des idiomes propres à chaque région.

Compréhension du Contexte

- La capacité à comprendre le contexte est essentielle pour une interprétation précise du langage, mais souvent difficile à réaliser pour les machines.

Traitement de l'Information Non Structurée

- La plupart des données textuelles sont non structurées, ce qui rend leur analyse et leur manipulation complexes.

Principaux Défis du NLP (Exemple explicatifs)

Ambiguïté

Il a vu le match avec des jumelles.	J'ai vu cet homme avec une longue-vue.	Elle a touché le verre avec douceur.
Le mot "jumelles" peut se référer à des instruments optiques pour voir de loin ou à des sœurs jumelles.	Le mot "longue-vue" peut se référer à un instrument optique pour voir de loin ou à une vision prolongée d'un homme.	Le mot "verre" peut se référer à un récipient en verre ou à un matériau transparent.

Principaux Défis du NLP (Exemple explicatifs)

Polysémie

Il a joué une note grave;	Lors de la conférence, il a parlé du Java.	Le coureur a pris son pied pendant la course.
"Grave" peut signifier une note musicale basse ou une situation sérieuse.	"Java" peut faire référence à un langage de programmation ou à l'île indonésienne.	"Pied" peut faire référence à la partie du corps ou à l'expression de prendre plaisir.

Principaux Défis du NLP (Exemple explicatifs)

Variabilité Linguistique

Elle a piqué une crise.	**Il est parti en flèche.**	**Il a vendu la mèche.**
Expression pour dire qu'elle s'est **mise en colère** sans rapport avec la crise.	Expression pour indiquer qu'il est **parti très rapidement** sans rapport avec la flèche	Expression signifiant qu'il a **révélé un secret**, sans rapport avec une vente réelle.

Principaux Défis du NLP (Exemple explicatifs)

Compréhension du Contexte

Le gardien a arrêté les tirs.	**Il a tapé dans la caisse.**	**Elle a coulé le projet.**
Peut signifier qu'il a **stoppé des balles** de football ou des **coups de feu**.	Peut indiquer qu'il a **volé de l'argent** ou simplement **frappé physiquement** sur une caisse.	Peut signifier qu'elle a **fait échouer** le projet ou qu'elle l'a **submergé de travail**.

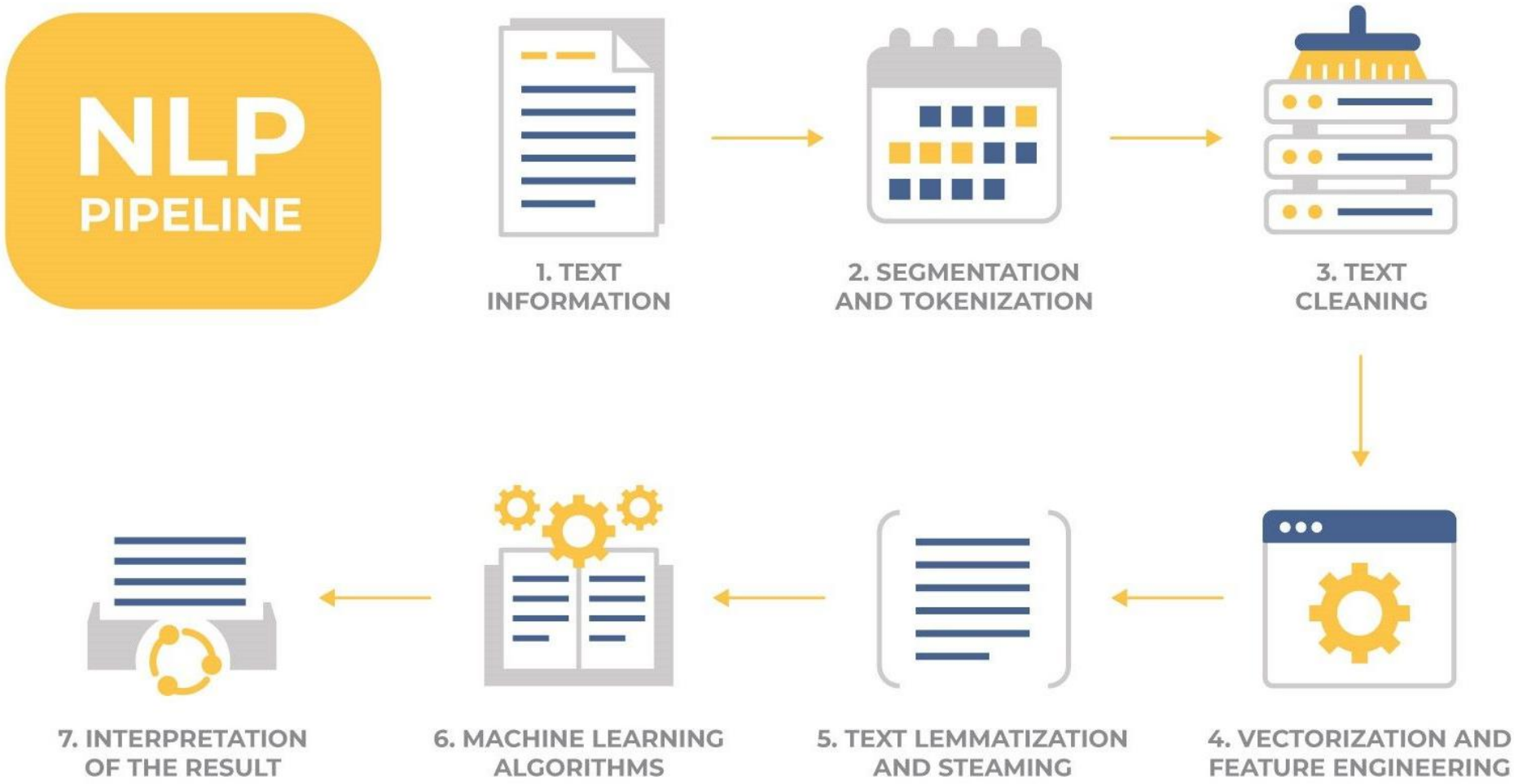
Principaux Défis du NLP (Exemple explicatifs)

Traitement de l'Information Non Structurée

Le pilote a navigué à travers le cloud.	**Les feuilles ont été diffusées sur le réseau.**	**Le sujet a été branché toute la nuit.**
"**Navigué**" et "**cloud**" peuvent évoquer la conduite d'un avion à travers les nuages ou l'utilisation d'Internet.	"**Feuilles**" peut se référer à des documents ou à des feuilles d'arbres, et "**réseau**" peut être un réseau informatique ou social.	"**Branché**" peut vouloir dire connecté à l'électricité ou à la mode, et "**sujet**" peut être une personne ou un thème de discussion.

Section 2 :
Prétraitement de texte en LPN

Processus du NLP



Processus du NLP

Phrase Originale :

"Les touristes visitent souvent la Tour Eiffel, un des symboles emblématiques de Paris."

1. Segmentation et Tokenisation

Division de la phrase en mots et signes de ponctuation.	**Après tokenisation** : ["Les", "touristes", "visitent", "souvent", "la", "Tour", "Eiffel", ",", "un", "des", "symboles", "emblématiques", "de", "Paris", "."]

2. Nettoyage de Texte

Suppression des signes de ponctuation et des mots vides (stop words).	**Après nettoyage** : ["touristes", "visitent", "Tour", "Eiffel", "symboles", "emblématiques", "Paris"]

3. Vectorisation et Ingénierie des Caractéristiques

Transformation des mots en vecteurs numériques pour l'analyse, avec des méthodes comme TF-IDF ou Word2Vec.	**Après vectorisation** : Représentation numérique de chaque mot non illustrable ici.

Processus du NLP

4. Lemmatisation et Stemming

Réduction des mots à leur forme de base.	**Après lemmatisation/stemming** : ["touriste", "visiter", "Tour", "Eiffel", "symbole", "emblématique", "Paris"]

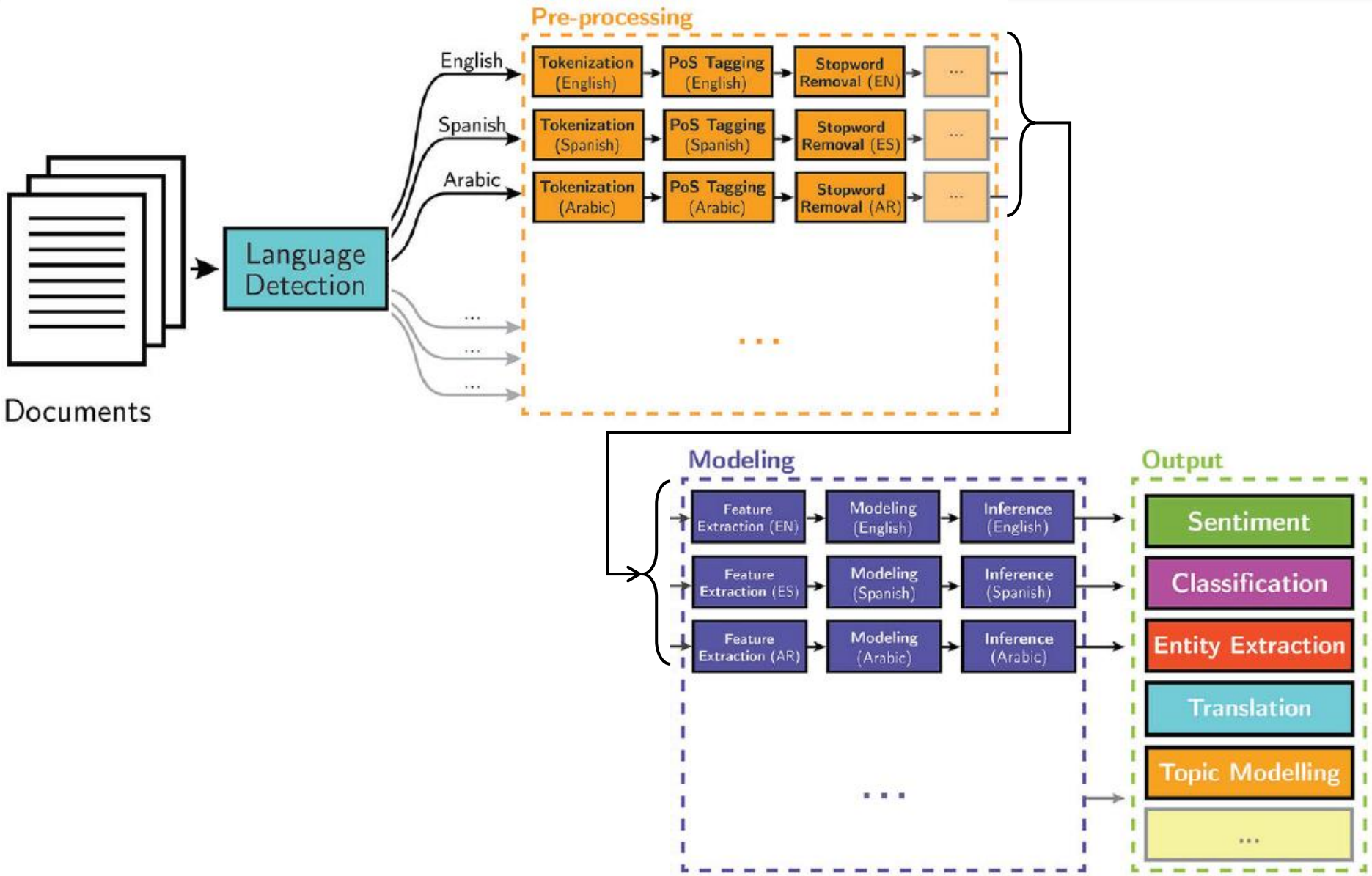
5. Algorithmes d'Apprentissage Automatique

Application d'algorithmes pour la classification, la prédiction, ou autres tâches de NLP.	**Après ML** : Par exemple, la classification de la phrase comme faisant partie d'un texte touristique.

6. Interprétation du Résultat

Analyse des résultats obtenus par les algorithmes d'apprentissage automatique.	**Après interprétation** : Conclusion que la phrase est liée au tourisme et concerne une attraction célèbre à Paris.

Processus du NLP



Notion du prétraitement de texte

Principe

- Le prétraitement de texte fait référence à toutes les étapes de nettoyage et de préparation des données textuelles avant de les utiliser pour l'analyse ou par les algorithmes d'apprentissage automatique.

Rôle

- Les données textuelles brutes sont souvent désordonnées et pleines de caractères inutiles et de variations de mots. Le prétraitement de texte nettoie et organise ces données, les rendant exploitables pour les algorithmes du NLP.
- La maîtrise de ces techniques est essentielle pour les professionnels de l'analyse du langage naturel, pour la création de modèles de NLP efficaces, permettant aux machines de comprendre et d'interagir avec le langage humain de manière intelligente.

Importance du prétraitement de texte

1. Réduction du Bruit

- **Description** : Élimination des éléments inutiles comme les balises HTML, les adresses e-mail, et les mots vides.
- **Exemple** : Suppression des hashtags et des mentions dans un tweet pour ne conserver que le message pertinent.

2. Uniformisation des Données

- **Description** : Standardisation des variantes orthographiques et grammaticales, telles que les majuscules/minuscules et les formes fléchies.
- **Exemple** : Conversion de tout le texte en minuscules pour éviter la distinction entre "Apple" et "apple" dans une analyse de sentiment.

3. Facilitation de l'Extraction de Caractéristiques

- **Description** : Transformation du texte en un format structuré pour simplifier l'extraction des caractéristiques linguistiques.
- **Exemple** : Utilisation de la tokenisation pour diviser un texte en mots, facilitant ainsi la détection des fréquences de mots.

Importance du prétraitement de texte

4. Amélioration de la Performance des Modèles

- **Description** : Amélioration de l'apprentissage des modèles de NLP grâce à des données plus cohérentes et représentatives.
- **Exemple** : Meilleure reconnaissance d'entités nommées après avoir filtré le texte des éléments perturbateurs.

5. Optimisation du Temps de Traitement

- **Description** : Accélération de l'analyse de texte et de l'entraînement des modèles en simplifiant le texte.
- **Exemple** : Diminution du temps d'entraînement d'un modèle en éliminant les parties non essentielles du corpus de texte.

Tokenisation, Normalisation et Suppression du bruit

Tokenisation

- Définition : Séparation d'un texte en unités de base (tokens) pour l'analyse.
- Exemple : "Paris est magnifique." devient ["Paris", "est", "magnifique", "."].

Normalisation

- Définition : Unification du format des tokens pour la cohérence et la précision de l'analyse.
- Exemple : "ÉTÉ" et "été" sont normalisés en "été".

Suppression du Bruit

- Définition : Élimination des éléments textuels non pertinents pour clarifier le sens.
- Exemple : "Contactez-nous à info@example.com ou visitez notre site web!" devient "Contactez-nous ou visitez notre site web!" après suppression des adresses e-mail.

Synthèse

- Ces étapes transforment le texte brut en données structurées, facilitant ainsi les diverses tâches de TALN comme l'analyse sémantique, la classification de texte, ou les systèmes de réponse aux questions.

Illustration du prétraitement sur l'exemple

Phrase Originale

- "Découvrez la magie de la Casbah d'Alger, un joyau historique! Pour plus d'infos, visitez notre site: www.tourisme-dz.com. #DécouverteAlgérie #Patrimoine"

1. Tokenisation

- On divise la phrase en mots et signes de ponctuation.
- **Résultat** : ["Découvrez", "la", "magie", "de", "la", "Casbah", "d'Alger", ",", "un", "joyau", "historique", "!", "Pour", "plus", "d'infos", ",", "visitez", "notre", "site", ":", "www.tourisme-dz.com", ".", "#DécouverteAlgérie", "#Patrimoine"]

2. Normalisation

- On convertit tout le texte en minuscules et on supprime la ponctuation superflue.
- **Résultat** : ["découvrez", "la", "magie", "de", "la", "casbah", "d'alger", "un", "joyau", "historique", "pour", "plus", "d'infos", "visitez", "notre", "site", "www.tourisme-dz.com", "découvertealgérie", "patrimoine"]

Illustration du prétraitement sur l'exemple

3. Suppression du Bruit

- On enlève les URLs, les hashtags et on peut aussi retirer les stop words si nécessaire (non illustré ici car cela dépend de la langue et de la liste des stop words utilisée).
- **Résultat** : ["découvrez", "la", "magie", "de", "la", "casbah", "d'alger", "un", "joyau", "historique", "pour", "plus", "d'infos", "visitez", "notre", "site"]

Phrase prétraitée

- "découvrez la magie de la casbah d'alger un joyau historique pour plus d'infos visitez notre site"

Phrase prête pour l'Analyse

- Le texte est désormais prêt pour être utilisé dans des tâches de NLP, telles que l'identification de thèmes clés ou l'extraction d'entités nommées (par exemple, "Casbah d'Alger" comme une entité de type lieu).

Techniques de prétraitement avancées : Stemming et Lemmatisation.

Stemming

Définition

- Une technique de prétraitement de texte qui consiste à réduire les mots à leur racine ou base en éliminant les affixes (préfixes, suffixes, infixes, circonflexes). Cette opération est généralement effectuée par des algorithmes simples qui tronquent les mots sans tenir compte du contexte.

Avantages

- Rapide et facile à implémenter.
- Utile pour les systèmes de RI où la correspondance exacte des mots n'est pas nécessaire.

Inconvénients

- Peut produire des racines qui ne sont pas des mots valides de la langue.
- Ne tient pas compte du contexte, ce qui peut mener à des erreurs lorsque les mots ont différentes racines en fonction de leur signification.

Exemple

- Phrase : "Les feuilles tombent des arbres en automne."
- Stemming : ["Les", "feuill", "tomb", "des", "arbr", "en", "automn"].

Techniques de prétraitement avancées : Stemming et Lemmatisation.

Lemmatisation

Définition

- Un processus de réduction des mots à leur lemme, c'est-à-dire la forme canonique ou dictionnaire d'un mot. La lemmatisation tient compte du contexte et utilise une analyse morphologique complète pour retourner le mot à sa forme de base.

Avantages

- Plus précis que le stemming car il tient compte de la signification du mot dans le contexte.
- Produit des résultats qui sont des mots valides et peuvent être utilisés pour une analyse sémantique plus précise.

Inconvénients

- Généralement plus lent que le stemming car il nécessite une analyse linguistique plus complexe.
- Peut nécessiter plus de ressources comme les dictionnaires et mes BD morphologiques.

Exemple

- Phrase : "Les chats jouent avec les souris qu'ils ont attrapées."
- Lemmatisation : ["Les", "chat", "jouer", "avec", "les", "souris", "que", "ils", "avoir", "attraper"].

Techniques de prétraitement avancées : Stemming et Lemmatisation.

Quand utiliser l'un plutôt que l'autre ?

Utilisez le stemming

- Lorsque la vitesse est essentielle et que vous travaillez avec des données textuelles où la précision des mots individuels n'est pas critique. C'est souvent le cas dans les systèmes de recherche d'informations où l'objectif est de faire correspondre des variantes d'un même mot.

Optez pour la lemmatisation

- Lorsque vous avez besoin d'une analyse précise et que le contexte et la signification complète des mots sont importants pour la tâche. Cela est particulièrement vrai pour les tâches de TALN telles que l'analyse de sentiments, la traduction automatique, ou l'interaction homme-machine où comprendre l'intention exacte est crucial.

Synthèse

- Choisir entre le stemming et la lemmatisation dépend donc des objectifs spécifiques de votre projet de TALN et des contraintes de performance de votre système.

Fin de Chapitre 02