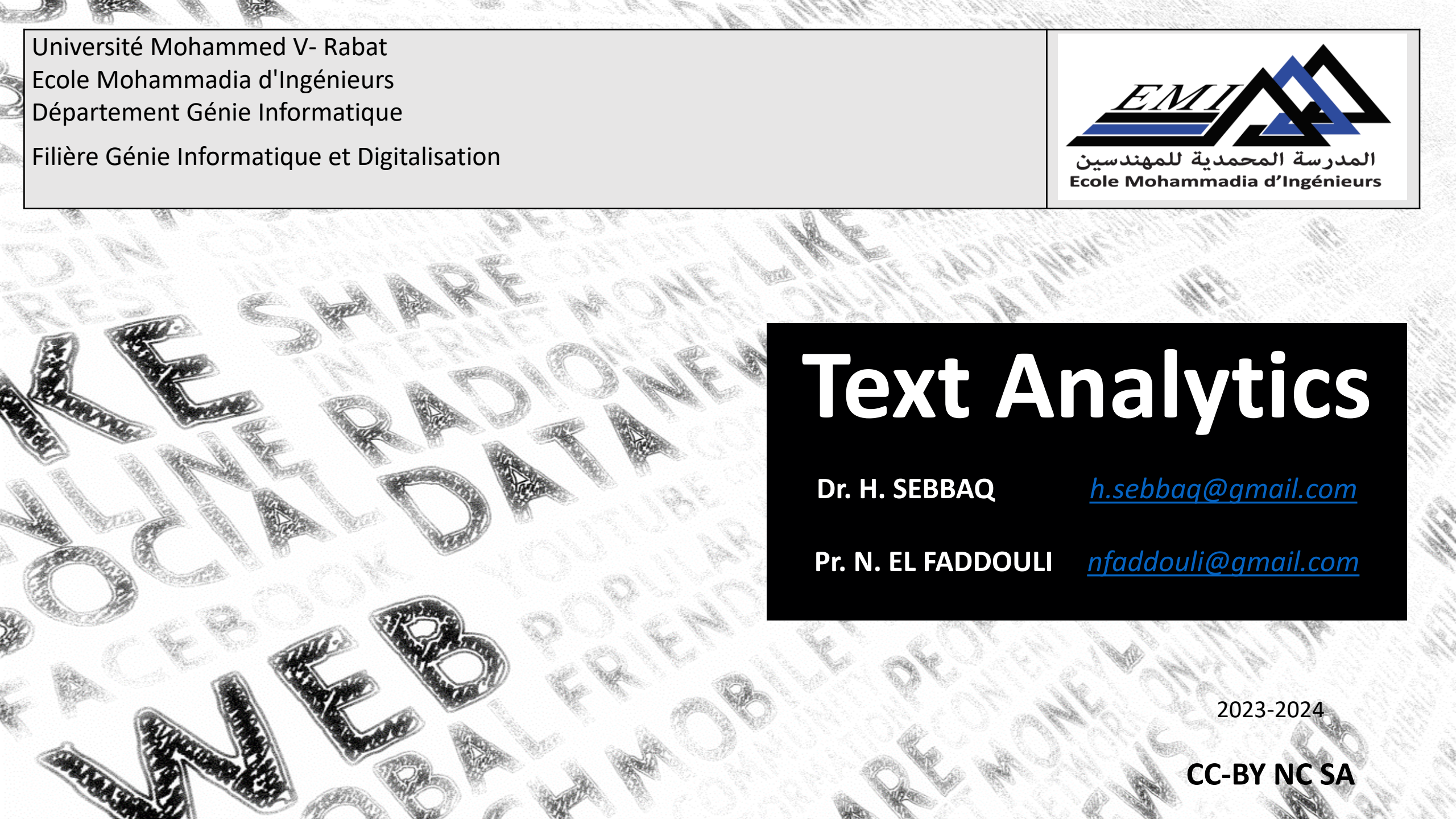
Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation



Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

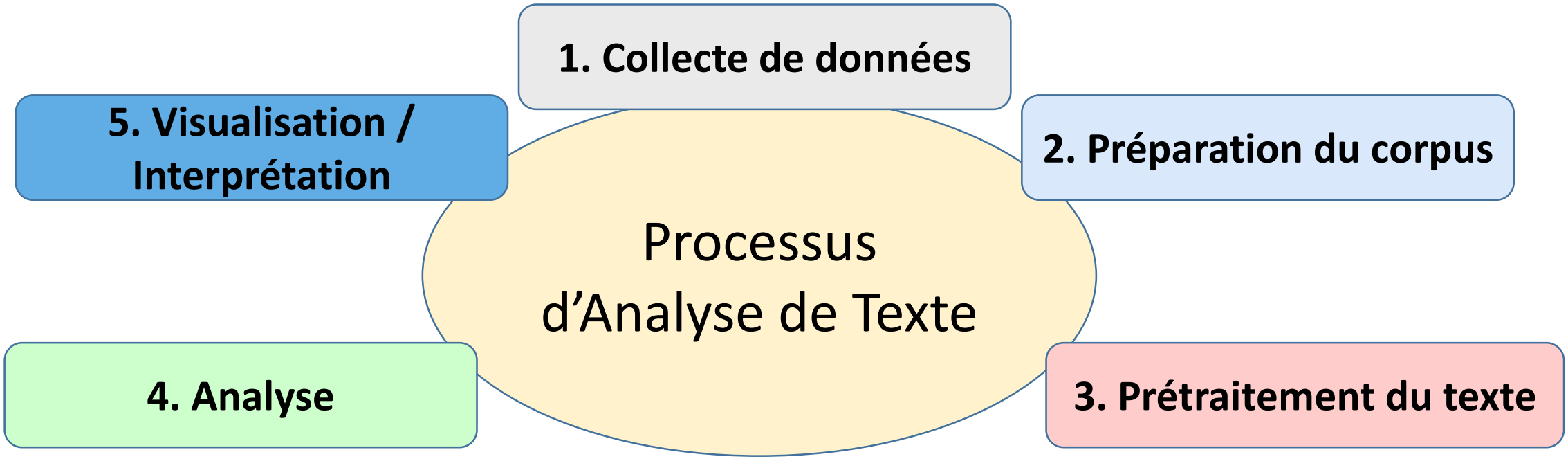
Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

Plan Processus d'analyse de texte

- ❑ **Aperçu du processus d'analyse de texte**
- ❑ **Collecte de données**
- ❑ **Préparation du corpus**
- ❑ **Prétraitement du texte**
- ❑ **Analyse**
- ❑ **Visualisation et interprétation**

Aperçu sur l'Analyse de Texte : Aperçu du processus d'analyse de texte



Aperçu sur l'Analyse de Texte : Collecte de données

- **Objectif de la Collecte de Données Textuelles**
  - **Clarifier l'objectif de l'analyse :** Identifier des tendances, répondre à des questions spécifiques ou comprendre les opinions des utilisateurs.
  - **Exemple :** Analyser les commentaires des clients pour évaluer leur satisfaction concernant la couverture réseau d'un opérateur télécom.
- **Sélection des Sources de Données**
  - **Diversité des sources :** Médias sociaux (Twitter, Facebook), sites web, bases de données, forums, enquêtes, etc.
  - **Exemple :** Collecte des commentaires sur la couverture réseau via Twitter, forums dédiés aux opérateurs télécom.

Aperçu sur l'Analyse de Texte : Collecte de données

- **Méthodes de Collecte de Données**
  - **API (Application Programming Interface) :** Accès structuré via les API pour récupérer des données textuelles.
  - **Web Scraping :** Utilisation d'outils automatisés (BeautifulSoup, Scrapy) pour extraire des commentaires depuis des sites web.
  - **Enquêtes :** Collecte directe de feedbacks via des enquêtes ou des questionnaires.
- **Évaluation de la Qualité :** Vérification et Nettoyage pour Identifier les données manquantes, incohérentes et assurer la qualité des données collectées.
- **Considérations Éthiques :** Respect des Normes Éthiques et Juridiques pour garantir la confidentialité et le consentement éclairé des utilisateurs.

Aperçu sur l'Analyse de Texte : Préparation du corpus

C'est quoi un Document? Un document est une pièce unique d'information enregistrée, se référant généralement à un texte écrit ou numérique unique (Un livre, un article, un rapport, un courrier ou de toute autre unité d'information.

C'est quoi un Corpus ? Un corpus est un ensemble structuré de textes ou de données orales utilisé pour des analyses ou des études linguistiques.

C'est quoi un Corpora? (Pluriel de corpus) sont simplement des collections multiples de textes ou de données linguistiques utilisées à des fins d'analyse linguistique ou de recherche. Ces collections peuvent être regroupées en fonction de divers critères tels que le domaine, la langue ou le genre, et sont utilisées pour étudier les modèles linguistiques, la grammaire ou d'autres phénomènes linguistiques.

Document **Corpus** **Corpora**

Aperçu sur l'Analyse de Texte : Préparation du corpus

- **Importance d'un Corpus Représentatif**
  - **Diversité des Sources :** Sélection de divers médias, genres et styles pour une vue globale du sujet.
  - **Représentation et Adéquation :** Le corpus doit refléter fidèlement les aspects et variations du sujet étudié.
  - **Fiabilité des Résultats :** La qualité du corpus affecte directement la fiabilité des résultats analytiques.
  - **Évaluation et Raffinement :** Ajustements réguliers pour intégrer de nouvelles tendances ou informations pertinentes.

Aperçu sur l'Analyse de Texte : Prétraitement du texte

La phase de prétraitement des textes est cruciale dans le **pipeline d'analyse de texte**. Après avoir constitué notre corpus, les données textuelles subissent une série de transformations pour s'assurer qu'elles sont affinées et prêtes à être analysées. Cela implique diverses étapes pour traiter le bruit, les incohérences et les informations non pertinentes dans le texte →

- **Nettoyage des données :** Suppression des balises HTML, des caractères spéciaux, de la ponctuation…
- **Tokenisation :** Décomposition du texte en unités plus petites (tokens ou jetons).
- **Mise en minuscules :** Conversion de l'ensemble du texte en minuscules afin d'en assurer la cohérence.
- **Élimination du bruit (Noise Removal):** Élimination des caractères non pertinents, des symboles ou des caractères spéciaux.
- **Suppression des mots vides (Stopwords Removal) :** Filtrage des mots courants (par exemple, "et", "est", "le") qui n'ajoutent pas de signification significative.

Aperçu sur l'Analyse de Texte : Prétraitement du texte

- **Dérivation (Stemming) :** Réduire les mots à leur forme de base ou racine (par exemple, « running » devient « run », "facilement" devient "facile")..
- **Lemmatisation (Lemmatization) :** Semblable au stemming, elle réduit les mots en fonction de leur lemme ou de leur forme dans le dictionnaire (par exemple, "better" à "good").
- **Vérification orthographique (Spell Checking):** Correction des fautes d'orthographe pour améliorer la qualité du texte.
- **Étiquetage grammatical (POS Tagging) :** Attribution des étiquettes grammaticales à chaque mot (ex : nom, verbe, adjectif, etc.).
- **Analyse syntaxique (Parsing) :** Analyse de la structure grammaticale pour comprendre les relations entre les mots.

Aperçu sur l'Analyse de Texte : Prétraitement du texte

Texte brut → **"La c0uv3rture réso est tellement m3diocre !! #frustré #sos"**

Tokenisation ["La", "c0uv3rture", "réso", "est", "tellement", "m3diocre", "frustré", "sos"]

↓

Mise en minuscules «la c0uv3rture réso est tellement m3diocre !! #frustré #sos »

↓

Noise Removal « la couverture réso est tellement médiocre frustré sos »

↓

Stopwords Removal « couverture réso est tellement médiocre frustré sos »

↓

Stemming « couvertur réso est tel médiocre frustré sos »

Aperçu sur l'Analyse de Texte : Prétraitement du texte

Texte brut → **"La c0uv3rture réso est tellement m3diocre !! #frustré #sos"**

Lemmatization

« couverture réseau être tellement médiocre frustré sos »

Spell Checking

« couverture réseau être tellement médiocre frustré SOS»

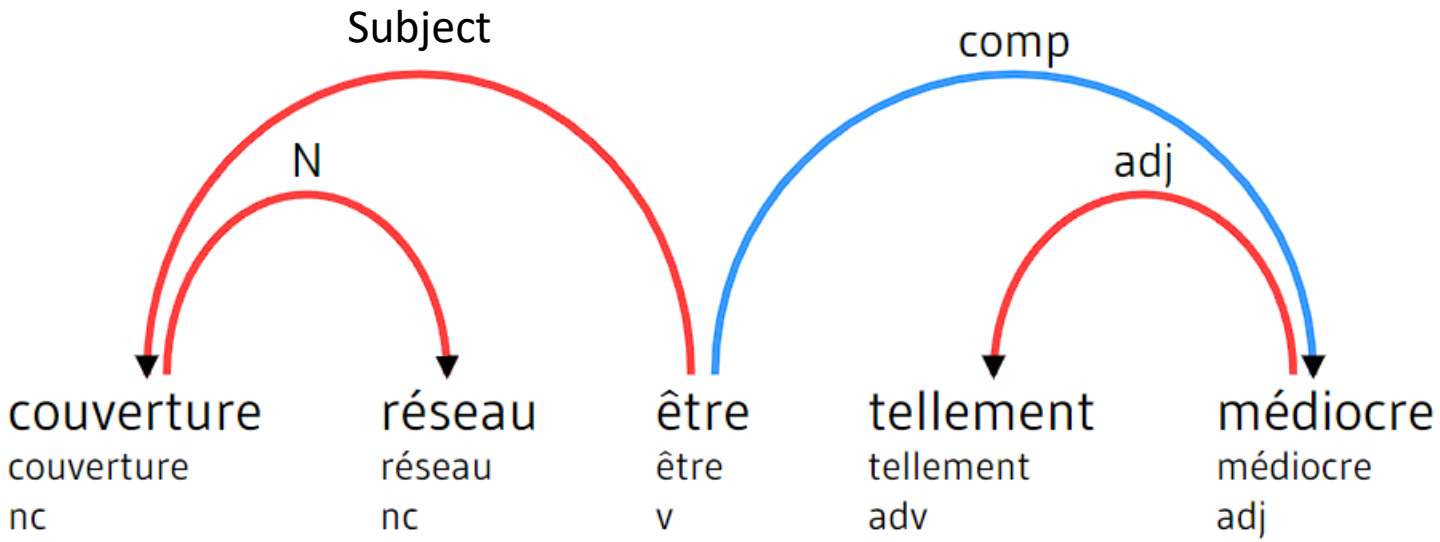
POS Tagging

NN	NN	VB	ADV	ADJ	ADJ	NN	← Tag

"couverture réseau être tellement médiocre frustré SOS "

Aperçu sur l'Analyse de Texte : Prétraitement du texte

Parsing



Aperçu sur l'Analyse de Texte : Analyse

- La phase d'analyse commence après le prétraitement du texte. Elle implique l'application de diverses techniques statistiques et algorithmiques afin d'extraire des informations significatives des données textuelles prétraitées:

- Analyse des sentiments
- Classification de textes
- Modélisation thématique (Topic Modeling)
- Reconnaissance d'entités nommées (Named Entity Recognition - NER)
- Analyse de similitude et regroupement (Clustering)
- Extraction d'informations

Aperçu sur l'Analyse de Texte : Visualisation et interprétation

- Enfin, la visualisation des résultats est souvent utilisée pour rendre plus compréhensibles les informations obtenues grâce à l'analyse de texte. Des graphiques, des nuages de mots ou des représentations visuelles sont utilisés pour présenter les résultats de manière accessible.

- Cette phase d'analyse convertit les données textuelles prétraitées en informations exploitables. Générant ainsi des informations significatives pour la prise de décision et une compréhension globale du contenu textuel