RER – NLP

Contexte :

Se familiariser avec le NLP et la bibliothèque NLTK pour le prétraitement des données textuelles.

Problématique(s) :

Comment utiliser NLTK et Spacy pour effectuer un pré-traitement des données textuelles ?

Comment normaliser les données textuelles ?

Comment transformer un texte en données numériques ?

Mots clés :

- **NLP** : *Natural Langage Processing*. Discipline informatique qui porte sur la compréhension, la manipulation et la génération de langage naturel par les machines.

  Les deux étapes principales de développement sont :

  1. Le pré-traitement des données.

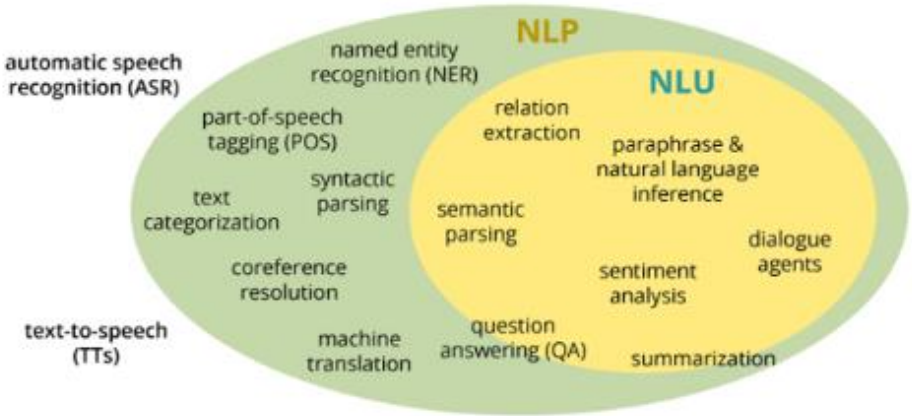
  2. Le développement du model.

- **NLU** : *Natural Language Understanding*. C'est une branche de la NLP qui se concentre sur la compréhension du langage naturel. Consiste à trouver la relation sémantique entre les mots.

  La compréhension du langage naturel peut se faire en deux parties :

  1. Une analyse syntaxique.

  2. Une analyse sémantique.



- **ASR** : *Automatic Speech Recognition*. C'est une technologie qui permet à un ordinateur de transcrire automatiquement la parole en texte. elle permet aux clients de parler aux ordinateurs le plus simplement possible.

- **Normalisation** : Traitement initial sur les données textuelles qui permet de réduire leur variabilité afin de réaliser des traitements plus adaptés. Cela regroupe la mise en minuscule, la suppression de la ponctuation et des accents, la tokenisation, la lemmatisation ou le stemming. La normalisation peut également inclure le remplacement des synonymes ou mots similaires par un mot standard pour améliorer la cohérence des données textuelles.
- **Tokenisation** : La tokenisation cherche à transformer un texte en une série de tokens individuels tels que chaque token représente un mot, groupe de mots, paragraphe, caractères, sous-mots ou une ponctuation.
- **Stemming** : *Désuffixation.* C'est l'action de couper la fin des mots pour extraire le radical.

  Exemple : « trouverez » -> « trouv ».

  La racine ne correspond pas toujours à terme issu de l'usage commun.

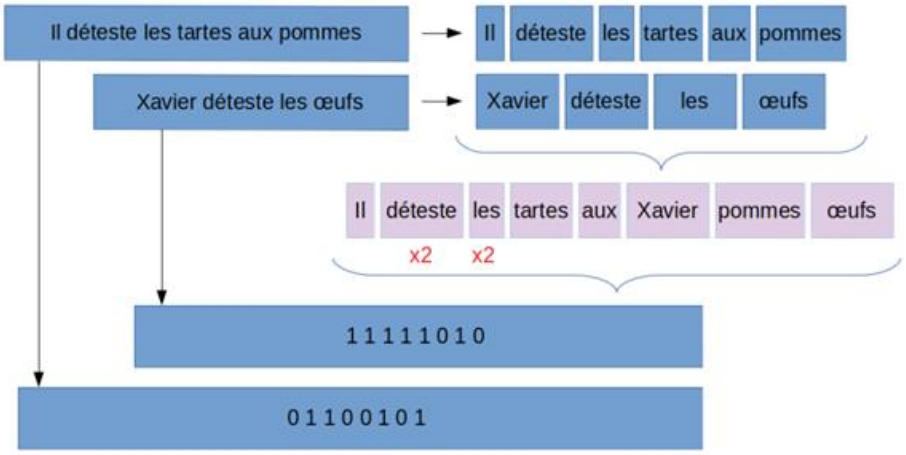
- **Lemmatisation** : C'est réduire un mot à sa forme canonique tel qu'il est défini dans le dictionnaire. Elle est plus précise que le stemming car elle prend en compte le contexte et la signification du mot.

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

- **Stop words** : Ce sont les mots vides tel que les articles, pronoms, prépositions (un, une, le, etc.). Opération intégrée à la normalisation des données en NLP qui consiste à supprimer les mots considérés comme inutile à la compréhension de la phrase ou au groupe de mots.
- **Bag of words** : Son principe se résume en trois phrases :

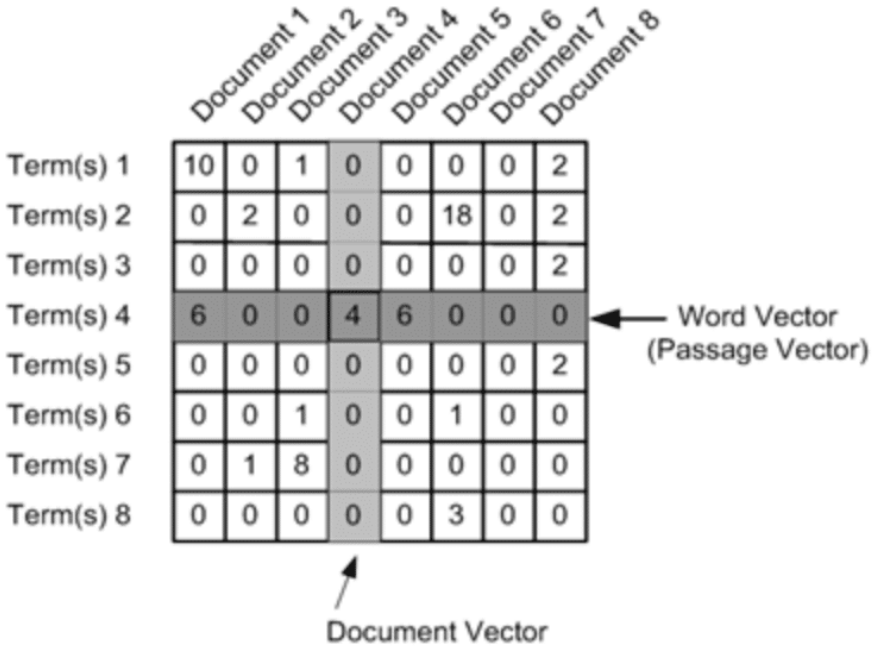
1. La tokenisation
2. La constitution d'un dictionnaire global.
3. L'encodage des chaines de caractères par rapport au vocabulaire constitué précédemment

La sortie est alors numérique.



Les codes sont inversés ci-dessus.

- **TF** : *Term Frequency*. 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**



$$Tf = \frac{n}{N}$$

Avec

n : Nombre d'occurrence du terme

N : Nombre de mots

- **TF-IDF** : *Term Frequency-Inverse Document Frequency* : 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.

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

où

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

- **N-grams** : Ce sont des séquences de n mot voisins d'éléments dans un document. Se réfèrent à des séquences continues de n mots consécutifs dans un texte.

  Exemple : bi-grams, séquence de deux mots. Tri-grams , séquence de trois mots.

  "On utilise ces n-grams en …"

  ```
  ["On utilise", "utilise ces", "ces n", "n grams", "grams en", ...]
  ```

- **NLTK** : *Natural Langage ToolKit*. Bibliothèque pour faire du NLP sur des données textuelles.
- **Spacy** : Bibliothèque pour faire du NLP sur des données textuelles. Spacy est développé dans un but de production contrairement à NLTK. La différence fondamentale entre NLTK et Spacy est le fait que NLTK contient une grande variété d'algorithmes pour résoudre un problème alors que spaCy n'en contient qu'un, mais le meilleur algorithme pour résoudre un problème.

Spacy est écrite en cython. Equivalent du numpy pour le NLP.

	NLTK	SAPCY
Avantages	• Il est très connu et possède une librairie complète pour le NLP	• Il est rapide • Utilise les réseaux de neurones pour le training • Spacy est développé dans un but de

	• Il supporte différents langages et idiomes • Facile à prendre en main et à utiliser	production contrairement à NLTK
Inconvénients	• Il ignore souvent le contexte des mots • Il est lent • Il ne repose pas sur un modèle de réseau de neurones	• Manque de flexibilité • Pas de Stemming • Moins facile à utiliser

Hypothèses :

Le choix entre Stemming et Lemmatisation se fait selon les données textuelles - Adeline

Vrai, cela dépend du projet.

Il est obligatoire d'utiliser en même temps Stemming, Lematisation et tokenisation lors du pré-traitement des données - Loïc

Faux

Le traitement par NLP n'est pas suffisant pour réaliser du topic extraction - Étienne

Faux

On peut utiliser le NLP pour détecter des cas de plagiat – Axel

Vrai

Le NLP n'a pas de limites dans le traitement des données textuelles – Adeline II

Faux

Spacy est plus complet que NLTK pour réalisation d'un projet– Adrien

Vrai

On peut utiliser le NLP pour extraire du texte à partir d'images - Jean-Paul SOSSAH

Faux

On peut combiner le NLP et les règles d'associations pour améliorer la compréhension d'un texte – Seydou

Vrai

On utilise du NLP pour générer des interaction homme-machine qui passe le test de turing – Briand

Vrai mais pas suffisant pour passer le test de turing.

La normalisation des données peut être fatale pour le futur modèle NLP – Tetyana

Vrai

Le NLP peut être utilisé avec le computer vision – Osman

Vrai

Les regex font parties du NLP – Ode à la joie

Vrai

La matrice de Gram est le seul outil mathématique permettant de calculer la fréquence des mots dans un document – Solenn

Faux

On peut utiliser du NLP dans une barre de recherche – Nicolas

Vrai, moteur de recherche sémantique.

On peut transformer des données textuelles en numériques sans normaliser les données - Loïc II

Vrai mais pas intéressant.

Est-ce que le pré-traitement dépend de l'application que l'on souhaite faire – Bassam

Vrai

Plan d'action :

Comparaison entre NLTK et Spacy

Workshop : NLTK puis Spacy