BUT SCIENCE DES DONNEES

Traitement du langage naturel

2023/2024

**TD Préparation de données textuelles pour le machine learning**

L'objectif du TD est de mettre en œuvre les concepts vus en cours magistral.

Le code doit être réalisé en python.

Les packages préconisés sont les suivantes (ils peuvent orienter vers une solution, mais vous avez l'initiative d'utiliser d'autres approches)

## Packages préconisés

```
import numpy as np
import spacy
from collections import Counter
import unidecode
import string
import pandas as pd
import matplotlib.pyplot as plt

from wordcloud import WordCloud

import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import WordPunctTokenizer
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords

from sklearn.feature_extraction.text import CountVectorizer,TfidfVectorizer
```

## 1. Création d'un jeu de test

Créer une variable « texte » à laquelle vous affectez un contenu de type texte libre (extract d'une page wiki par ex.). Le texte doit contenir plusieurs phrases en langue francaise.
Affichez les 20 mots les plus fréquents et leur fréquence

## 2. Lowering / Suppression d'accents

Transformez le texte en minuscules et supprimez les accents.
Affichez les 20 mots les plus fréquents et leur fréquence
Faites un wordcloud

## 3. Stop words

Définissez une liste de stop words en francais.

Supprimez les stop words de votre texte.
Affichez les 20 mots les plus fréquents et leur fréquence
Faites un wordcloud

## 4. Suppression de la ponctuation

Supprimez la ponctuation de votre texte.
Affichez les 20 mots les plus fréquents et leur fréquence

## 5. Stemming

Appliquez un traitement de stemming sur vos données.
Affichez les 20 mots les plus fréquents et leur fréquence
Faites un wordcloud

## 6. Lemmatization

Appliquez un traitement de lemmatization sur vos données.
Affichez les 20 mots les plus fréquents et leur fréquence
Faites un wordcloud

## 7. Bag of Words

Construisez un modele BagOfWord (CountVectoriser) à partir de votre texte initial
Affichez les poids de chaque mot de chaque phras
Appliquez ce modèle à une nouvelle phrase

## 8. TF-IDF

Construisez un modele TfIdf (TfIdfVectoriser) à partir de votre texte initial
Affichez les poids de chaque mot de chaque phrase
Appliquez ce modèle à une nouvelle phrase

## 9. Sentiment analysis

Entrainez un modèle de prediction de sentiment à partir du dataset

```
from nltk.corpus import twitter_samples
```