Université Constantine2 – Abdelhamid Mehri

Faculté des Nouvelles Technologies de l'Information et de la Communication
Département Informatique Fondamentale et ses Applications

Module : WANLP | M1-SDIA

Enoncés du TP 03

Exercice 01 : Prétraitement de Texte en NLP

Texte Original à Utiliser :
En cette belle journée ensoleillée du 5 juillet, la fête de l'Indépendance de l'Algérie a été célébrée avec éclat sur la Place des Martyrs : c'est-à-dire concerts par l'Orchestre Symphonique National, feux d'artifice au-dessus de la Baie d'Alger, et discours patriotiques par le Président de la république Mr. Abdelmadjid Tebboune. Détails sur notre page Facebook 'Festivités-DZ'! Pour plus d'informations, visitez-nous sur: www.festivités-dz.com. #FêteNationale #1erNovembre.

Travail Demandé

1) Tokenisation :
   – Séparez le texte en mots et signes de ponctuation, y compris les apostrophes, les traits d'union et les autres caractères spéciaux.
2) Normalisation :
   – Convertissez le texte en minuscules et retirez la ponctuation, en faisant attention aux éléments comme "c'est-à-dire" qui doivent être traités spécifiquement.
3) Suppression du Bruit :
   – Éliminez les URLs, les hashtags, les noms propres des pages, et les références aux réseaux sociaux, ainsi que les stop words.
4) Traitement des Mots Concaténés et des Guillemets :
   – Détachez les mots concaténés comme "c'est-à-dire" et traitez les expressions entre guillemets avec soin.
5) Analyse des Parties du Discours (POS) :
   – Appliquez le tagging POS pour déterminer les catégories grammaticales de chaque mot.
6) Reconnaissance d'Entités Nommées (NER) :
   – Identifiez les entités nommées telles que les noms de personnes, d'organisations ou de lieux.

Consignes :

- Importez les modules nécessaires de la bibliothèque NLTK.
- Téléchargez des ressources NLTK supplémentaires pour cette tâche en utilisant les commandes `nltk.download('averaged_perceptron_tagger')`, nltk.download('stopwords'), `nltk.download('punkt')`, nltk.download('maxent_ne_chunker')` et `nltk.download('words')`.
- Implémentez des méthodes de NLTK comme nltk.tokenize.word_tokenize pour la tokenisation, et utilisez nltk.corpus.stopwords pour filtrer les mots vides.
- Pour le POS et le NER, utilisez nltk.pos_tag et nltk.ne_chunk après avoir téléchargé les ressources nécessaires.

Astuces :

- Pour les mots concaténés et les expressions entre guillemets, envisagez l'utilisation d'expressions régulières (re module en Python) pour un traitement personnalisé.