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 02

Exercice 1 : Introduction à l'analyse syntaxique avec NLTK

Objectif : Se familiariser avec la bibliothèque NLTK pour analyser les parties du discours (part of speech, POS) des mots dans une phrase.

Contexte : En TALN, l'analyse syntaxique consiste à identifier la fonction grammaticale de chaque mot dans une phrase. Cela permet de comprendre la structure d'une phrase et de mieux analyser son sens.

Travail demandé : Écrivez un script Python qui effectue les opérations suivantes :
- Prenez la phrase "*La JS Kabylie a battu le MC Alger dans un match passionnant au stade du 1er Novembre.*" comme input.
- Tokenisez la phrase en mots.
- Utilisez la fonction `**pos_tag**` pour obtenir et afficher les parties du discours de chaque mot.

Consignes :
1. Importez les modules nécessaires de la bibliothèque NLTK.
2. Utilisez la fonction `**word_tokenize**` pour diviser une phrase en mots, ou "**tokens**".
3. Appliquez la fonction `**pos_tag**` sur les tokens pour obtenir leurs parties du discours.
4. Affichez les mots avec leurs étiquettes de parties du discours correspondantes.

Notes :
- Les parties du discours incluent les noms, les verbes, les adjectifs, etc.
- NLTK peut nécessiter le téléchargement de ressources supplémentaires telles que les ensembles de données de tokenisation et les modèles POS, que vous pouvez télécharger en utilisant `**nltk.download('averaged_perceptron_tagger')**` et `**nltk.download('punkt')**`.

Exercice 2 : Reconnaissance d'entités nommées avec NLTK

Objectif : Utilisez NLTK pour identifier les entités nommées dans un texte, telles que les noms de personnes, d'organisations ou de lieux.

Contexte : La reconnaissance d'entités nommées (Named Entity Recognition, NER) est une tâche importante en TALN qui permet d'extraire des informations spécifiques à partir du texte. Les entités nommées sont souvent des noms propres qui font référence à des objets spécifiques dans le monde réel.

Travail demandé : Écrivez un script Python qui réalise les actions suivantes :
- Prenez la phrase "*Le pont Sidi M'Cid, situé à Constantine, offre une vue imprenable sur les gorges du Rhummel et attire de nombreux touristes chaque année.*" comme input.
- Tokenisez et étiquetez la phrase avec les parties du discours.
- Appliquez la reconnaissance d'entités nommées pour extraire et afficher les entités avec leurs catégories spécifiques.

Consignes :
1. Importez les modules nécessaires de NLTK.
2. Utilisez la fonction `word_tokenize` pour diviser une phrase en tokens.
3. Appliquez la fonction `pos_tag` pour étiqueter les tokens avec leurs parties du discours.
4. Utilisez la fonction `ne_chunk` pour identifier et étiqueter les entités nommées dans la phrase.

Notes :
- Les catégories d'entités nommées incluent les personnes (PERSON), les organisations (ORGANIZATION), et les lieux (LOCATION).
- Vous devrez peut-être télécharger des ressources NLTK supplémentaires pour cette tâche en utilisant les commandes **`nltk.download('maxent_ne_chunker')`** et **`nltk.download('words')`**.