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 TD 02

Exercice 1 : Tokenisation Manuelle

1ère Partie (*)

Texte à traiter : "Les étudiants de Master en informatique explorent les dernières avancées en IA."

Questions :
1. Divisez le texte ci-dessus en tokens individuels. Chaque mot et signe de ponctuation doit être considéré comme un token séparé.
2. Expliquez pourquoi "en" et "IA" sont considérés comme des tokens distincts malgré leur brièveté.

2ème Partie (**)

Texte additionnel à traiter : "L'IA, c'est-à-dire l'intelligence artificielle, est l'étude des 'agents intelligents'."

Questions supplémentaires :
1. Comment tokenisez-vous les abréviations et les expressions entre guillemets ?
2. Quelle approche utilisez-vous pour les contractions comme "c'est-à-dire" ?

Exercice 2 : Identification et Suppression de Bruit

1ère Partie (*)

Texte à traiter : "Pour tout renseignement, écrivez-nous à info@univ-constantine.dz ou visitez le site www.univ-constantine.dz. #Renseignement #Université"

Questions :
1. Identifiez les éléments qui constituent du bruit dans le texte ci-dessus.
2. Reformulez le texte en supprimant ces éléments.

2ème Partie (**)

Texte additionnel à traiter : "Nouveautés en WANLP: voir p. 15-22 du sopport de cours; pour plus, abdelkrim.bouramoul@univ-constantine2.dz. #WANLP #Master1SDIA"

Questions supplémentaires :
1. Quels éléments de ce texte considérez-vous comme du bruit et pourquoi ?
2. Comment traitez-vous les références de page (p. 15-22) et les hashtags ?

Exercice 3 : Normalisation Textuelle

1ère Partie (*)

Texte à traiter : "L'INTELLIGENCE ARTIFICIELLE révolutionne le monde !"

Questions :
1. Convertissez le texte en minuscules et expliquez pourquoi cette étape est importante.
2. Quelle serait la forme normalisée de "L'INTELLIGENCE" et "ARTIFICIELLE" après l'application de la lemmatisation ?

2ème Partie (**)

Texte additionnel à traiter : "Dr. BENALI Mohamed, PhD, est spécialisé en TALN: traitement 'automatique' du langage naturel."

Questions supplémentaires :
1. Comment normalisez-vous les titres académiques et les abréviations spécifiques au domaine du TALN ?
2. Quelle est la forme lemmatisée de "spécialisé" et "traitement" dans ce contexte ?

Exercice 4 : Application Complète de Prétraitement

1ère Partie (*)

Texte à traiter : "Prof. ZITOUNI Hamza présentera sa recherche sur l'IA à la conférence internationale sur les avancées de l'Intelligence Artificielle qui se déroulera à Constantine. Contact : hamza.zitouni@univ-constantine2.dz"

Questions :
1. Appliquez la tokenisation au texte.
2. Normalisez le texte tokenisé, y compris la suppression des abréviations et des adresses email.
3. Rédigez le texte final après avoir appliqué tous les processus de prétraitement mentionnés.

2ème Partie (**)

Texte additionnel à traiter : "Conférence, 16h30: 'Impact de l'IA sur l'éco-système numérique'; inscription via le site-web."

Questions supplémentaires :
1. Quelle stratégie de tokenisation appliquez-vous aux termes techniques et aux horaires ?
2. Après avoir normalisé le texte, comment représentez-vous 'éco-système numérique' et gérez-vous le terme 'site-web' ?