Devoir libre en NLP
Deuxième Année Ingénierie des Données

Pr. Tarik BOUDAA

April 18, 2023

Ce devoir sera noté et sera évalué par des entretiens individuels avant la fin du semestre. Le travail à rendre et à présenter le jour de l'évaluation doit être constitué d'un seul dossier contenant:

- 3 jupyter notebook (un par exercice) contenant les explications des méthodes et des algorithmes implémentés ainsi que le code source et les résultats des exécutions.
- Les autres fichiers et ressources utilisés
- Pour chaque exercice il y a des étapes de pré-traitement à réaliser, ces étapes doivent êtres détaillées dans vos réalisations.

1 Exercice 1

1. Réaliser un correcteur d'orthographe en python pour la langue anglaise en se basant sur l'algorithme implémenté dans les TPs.
2. Implémenter une variante du programme précédent qui utilise un modèle de langue pour ordonner les corrections possibles par leurs probabilités selon le modèle de langue et selon la distance minimale d'édition.
3. Implémenter une troisième variante qui utilise un modèle de langue qui modélise la distribution de séquences de lettres au lieu des mots.
4. Implémenter une quatrième variante en utilisant *noisy channel model*

2 Exercice 2

En utilisant le même algorithme implémenté en TP (avec éventuellement quelques adaptions) réaliser un système d'étiquetage morphosyntaxique pour la langue française en utilisant des données annotées de votre choix ou les données indiquées dans le lien suivant:
https://github.com/qanastek/ANTILLES/tree/main/ANTILLES

3 Exercice 3

1. Écrire un programme qui calcule la similarité entre deux documents avec deux approches différentes :

   - Uniquement en utilisant la comparaison lexicale des mots des deux documents et en utilisant WodNet
   - En utilisant word embedding et la distance Word Mover's Distance (vous pouvez utiliser la bibliothèque gensim et les word embedding de votre choix)

2. En utilisant un corpus de votre choix (de taille convenable aux capacités de votre machine) générer des word embedding spécifiques destinées au domaine de l'informatique avec deux méthodes différentes:

   - en implémentant votre code de génération des word emebdding *from scratch*
   - en utilisant la bibliothèque Gensim.

3. Utiliser un classificateur basé sur votre implémentation de la régression logistique pour implémenter un système d'analyse de sentiment:

   - Utiliser un jeu de données de *training* et de teste de votre choix
   - L'entrée du système est le texte à analyser qui sera présenté sous forme d'un vecteur égal à la moyenne des vecteurs de ses mots (utiliser un word embedding de votre choix)