

Lab n°2 Word Embedding n°2 part 2

Data Preparation

1. Téléchargez les textes à cette adresse:

<https://github.com/AttilaDSA/IntilaqDSAcademy/blob/master/Word%20Embedding%20Labs/Word%20Embedding%20Lab%20n%C2%B03/Word%20Embedding%20Lab%20n%C2%B03.rar>

2. Préparez les textes avant de lancer le model de word embedding :

- a. Séparez les phrases les une des autres
- b. Décomposez chaque phrase en une liste de mots
- c. Afin d'éviter que les mêmes mots conjugué de manière différentes ou bien accordés de manières différentes (selon le genre ou la pluralité) ne soient considérés comme des mots différents, trouvez un moyen de les relier entre eux. (Vous pouvez vous reposer sur le dictionnaire disponible à cette adresse :
<https://github.com/AttilaDSA/IntilaqDSAcademy/blob/master/Word%20Embedding%20Labs/Word%20Embedding%20Lab%20n%C2%B02/dictionary.rar>

)

3. Lancer les deux models de word embedding (Wor2Vec et FastText) sur les textes traités.

Data use

4. En vous basant sur la proximité des mots obtenus à travers les models de word embedding, essayer d'écrire un algorithme d'écriture d'un poème aléatoire avec chaque ligne étant composée d'un sujet, verbe et complément (nom comun).
5. En vous basant sur les modèles développés dans la première partie (les phrases, les expressions et les mots) produisez une version écrite du poème.