**# 11/10/2024**

Mot clef et résumé d'un texte

Plusieurs textes .txt en anglais, qui traitent de l'éthique de l'IA.

|  |  |
| --- | --- |
| Collecter les pdf/html | OK |
| Parser | Problème avec HTML, supprime certains espaces 🡪 balek |
| CLeaning | Reste quelques docs dégueu (CSS). |
| TF-IDF sur chaque phrase pour trouver les phrases les plus importantes en fonction des mots. | Fait |
| Trouver les mots clefs dans le corpus (mots les plus fréquents ? Mots donnés par le prof ? TF-IDF sur chaque document puis mots les plus fréquents ?) | Fait |
| TF-IDF sur les N-Grammes | A faire, Alex |
|  |  |
|  |  |

**# 16/10/2024**

Résultats du TF-IDF :

* Somme sur les phrases : résultats bons !
* Moyenne uniquement sur les phrases contenant le mot : Résultats dégueulasses. La moyenne est biaisée par les mots n’apparaissant que dans une seule phrase qui ont donc une moyenne haute alors que leur TF-IDF est bas.

Graphe

Deux mots sont reliés s’ils apparaissent dans une certaine distance. Distance à préciser.

Plus court chemin entre deux mots donnés par la bibliothèque scikit-network.

<https://scikit-network.readthedocs.io/en/latest/tutorials/ranking/pagerank.html>

Document par document

Sur l’ensemble du corpus

Avec des bigrammes

Mercredi prochain on parle des résumés

15 min de présentation. Présenter notre méthodologie plus que les résultats.