Compte rendu TP1

Objectif :

Obtenir les catégories des aliments les plus évoqués dans les 5 premiers épisodes de la saison 3 de The Big Bang Theory.

Type d’analyses :

Pour ce faire nous utiliserons les fichiers txt comprenant les sous-titres des 5 épisodes.

Pour notre projet nous aurons besoin de segmatiser nos mots.

Puis nous utiliserons un étiquetage morpho-syntaxique ou une liste de mots vides pour filtrer nos textes (par exemple : <https://www.ranks.nl/stopwords>).

Et enfin une lemmatisation de nos mots filtrés.

Méthode :

Tout d’abord, nous allons regrouper tous les textes des 5 épisodes en un seul texte.

Puis nous tokeniserons nos mots avec des espaces ou des tirets en excluant les mots non segmentables.

Ensuite il est important de traiter nos tokens, en supprimant les guillemets, les tirets, les astérisques (dans le texte « \* \* \* \* \* ») et en transformant les doubles espaces en espaces simples pour garder une structure cohérente.

Puis nous utiliserons un étiquetage morpho-syntaxique sur nos tokens. En général les mots en rapport avec la nourriture sont principalement des noms et des adjectifs. Le but sera de filtrer nos textes pour ne laisser que les noms et les adjectifs.

En parallèle, nous procèderons aux mêmes opérations sur des recettes de cuisine trouvées sur internet regroupées en un seul grand texte.

Pour finir nous utiliserons une lemmatisation sur notre texte de recettes de cuisine et de sous-titres pour transformer nos mots pluriels aux singuliers.

Ensuite nous filtrerons les mots dans notre texte pour ne garder que les mots qui apparaissent aussi dans nos recettes de cuisine.

Nous finirons par classer la nourriture des protagonistes dans les quatre épisodes de The Big Bang théorie en différentes catégories. A partir de Wikipédia il est possible de récupérer des catégories alimentaires : <https://en.wikipedia.org/wiki/Category:Foods>

Evaluation :

Pour vérifier nos résultats nous demanderons à un cobaye d’annoter tous les mots en rapport avec l’alimentation dans nos sous-titres et les compareront avec ceux obtenus avec notre méthode. Nous pourrons alors expliciter le nombre de faux positifs et faux négatifs de mots alimentaires obtenus dans nos résultats. On pourra alors calculer le pourcentage de réussite de notre méthode.