Bureau d'étude

Clustering automatique de texte

Description du projet

Le but de ce projet est d'analyser automatiquement les contributions libres du grand débat 2019 afin de construire des catégorie de contributions :

https://www.irit.fr/~Philippe.Muller/grand_debat.7z

Description des données: les données sont constituées d'un ensemble de textes libres. Les textes sont divisés en 4 grands thèmes.

L'approche est fondée sur la recherche des mots importants du texte. Ces mots importants sont déterminés en fonction de leur fréquence d'apparition dans le corpus (l'ensemble des textes considérés). Une fois que l'ensemble des *n* mots les plus pertinents est défini, chaque texte sera alors décrit comme le vecteur des fréquences d'apparition des mots pertinents. Le but sera de construire des catégories homogènes de contribution en utilisant la méthode des *k-means*.

Modalités :

- Groupe de 3-4 étudiants
- Une recherche personnelle devra être effectué sur les notions de *TF-IDF, distance cosinus, lemmatisation et K-means* nécessaires pour le projet
- Ponctuellement les encadrants prévoiront des séances tutoriels pour l'ensemble des groupes (dates à déterminer).
- Vous devrez fournir un dossier intermédiaire avec (pour la partie 1, cf la suite du sujet):
    - Une description des outils théoriques utilisés
    - Les choix de conception et la description des algorithmes
    - Des statistiques sur l'ensemble des textes (temps de chargement, nombre de mots total, moyenne de mot par texte, etc)
    - Des exemples d'utilisation de TF-IDF sur quelques textes
    - Etude de la pertinence de TF-IDF pour la détection des mots courants de la langue (les mots très fréquents dans tous les textes).
- Fournir un dossier final avec :
    - Une description des outils théoriques utilisés
    - Les choix de conception et la description des algorithmes
    - Une présentation des résultats

- Fournir le code du projet en python avec
    - Le code commenté
    - Un fichier de code différent pour chaque partie du projet et un pour le programme principal
- Le groupe devra rencontrer autant de fois que nécessaire son tuteur (Tom Bourgeade ou Philippe Muller)
- La notation sera fondée sur :
    - Le dossier de conception
    - Le programme
    - Une soutenance orale de présentation du projet
- Date de retour des dossier et programme : à déterminer (fin du semestre)

Partie 1 : Traitement des fichiers

La base de données contient actuellement un ensembles de textes, dans les quatre catégories.
On appelle ici texte chaque contribution faite par une personne différente sur une des questions posées ou la réponse attendue est de type “libre”.
Il faut :
Lemmatiser et nettoyer chaques texte (enlever la ponctuation, retirer les mots trop fréquents, etc)
Sauvegarder les données nettoyées.

Partie 2 : Analyse de texte

Dans cette partie vous déterminerez un ensemble de fonction pour :

- Lire et récupérer l’ensemble des mots d’un texte
- Calculer la fréquence d’apparition d’un mot dans un texte
- Calculer et stocker pour chaque texte les mots et leur fréquence. Stocker toutes ces informations pour un ensemble de textes disponibles dans un répertoire
- Calculer la fréquence d’apparition d’un mot dans un ensemble de texte (i.e. le pourcentage de fois où le mot apparaît dans un texte du corpus).
- Calculer l’indice TF-IDF d’un mot d’un texte par rapport au corpus

Partie 3 : Nuage de mots

- Faire une fonction pour visualiser les mots importants d’un texte ou d’un ensemble de texte avec une représentation type nuage de mots (word cloud).
    - Utilisez https://github.com/amueller/word_cloud
- Utiliser cette représentation pour chacun des quatres thèmes.

Partie 4 : Distance entre texte

Dans cette partie vous déterminerez un ensemble de fonctions pour :

- Déterminer l'ensemble des *n* mots les plus pertinents dans le corpus (*n* est un paramètre, la notion de pertinence est basée sur IDF)
- Transformer un texte en un vecteur de fréquences sur les *n* mots les plus pertinents
- Calculer la distance cosinus entre deux textes représentés par leur vecteur de fréquence respectif

Partie 5 : Classification non supervisée

Dans cette partie vous déterminerez un ensemble de fonctions pour :

- Déterminer automatiquement *k* groupes de textes homogènes à l'aide de l'algorithme *k-means*
- Pour chaque groupe de textes donner le vecteur central et visualiser le à l'aide des nuages de mots
- Tester avec 4 clusters et vérifier si on peut retrouver les grands thèmes de départ.
- Tester aussi la méthode à l'intérieur de chaque grand thème pour trouver k sous-thèmes qui vous paraissent faire du sens.