Rapport SAÉ Traiter des données

INTRODUCTION

Dans cette SAÉ nous avons pu acquérir divers compétences en informatique comme l'analyse des données textuelle la collecte et pré-traitement des données textuelles qui nous permet d'etre capable de traiter des données textuelle qui proviennent des systémes d'informations.

SPÉCIFICATION

Compte_lignes_mots(nom_fichier):

La fonctionnalité de cette fonction est de compter le nombre de lignes et de mots du fichier nom_fichier.

Compte_dans_fichiers(liste_fichiers):

La fonctionnalité de cette fonction est de compter les lignes et les mots dans chacun des fichiers de la liste liste_fichiers. Cette fonction prend en entrée une liste de fichiers

mots_fichier(nom_fichier):

La fonctionnalité de cette fonction est de recenser tous les mots utilisés dans le fichier prit en entré et compte le nombre d'utilisations de chacun de ces mots

mots_dans_fichiers(liste_fichiers):

La fonctionnalité de cette fonction est de recenser tous les mots utilisés dans l'ensemble des fichiers de la liste prise en entré et compte le nombre d'utilisations de chacun de ces mots dans chacun des fichiers

apparition_mots(liste_fichiers):

La fonctionnalité de cette fonction est de calculer la fréquence d'apparition des 15 mots apparaissant le plus fréquemment dans chacun des fichiers de la liste prise en entré

mots_dans_fichiers_rep(rep):

La fonctionnalité de cette fonction est de recenser tous les fichiers contenus dans le répertoire pris en entré dans une liste liste_fichiers_rep. Cette fonction recense également tous les mots utilisés dans l'ensemble des fichiers de la liste liste_fichiers_repet compter le nombre d'utilisations de chacun de cesmotsdans chacun des fichiers

apparition_mots_rep(rep):

La fonctionnalité de cette fonction est de calculer la fréquence d'apparition des 15 mots apparaissant le plus fréquemment dans chacun des fichiers contenus dans le répertoire pris en entré

compte_mot_rep(mot, rep):

La fonctionnalité de cette fonction est de calculer la fréquence d'apparition d'un mot donné en entrée par l'utilisateur dans chacun des fichiers contenus dans le répertoire pris en entré.La fonction prendra en entrée le mot donné par l'utilisateur motainsi que le nom du répertoire rep

lecture_tweet(rep):

La fonctionnalité de cette fonction est de lire tous les tweets contenus dans tous les fichiers contenus dans le répertoire pris en entrée

compte_tweet(rep):

La fonctionnalité de cette fonction est de compter le nombre de tweet par journée et renvoit un graphique représentant le nombre de tweet par jour entre le 27/11 et 07/12.

compte_mot_tweet_rep(mot, rep):

La fonctionnalité de cette fonction est de calculer la fréquence d'apparition d'un mot donné en entrée par l'utilisateur par jour sur l'ensemble des tweets contenus dans les fichiers contenus dans le répertoire pris en entrée

RÉALISATION

compte_lignes_mots(nom_fichier):

Afin de réaliser cette toute première fonction , j'ai exploiter les connaissances acquises dans le ressources R107 : utilisation des boucles ,manipulation des fichiers , utilisation des listes , etc. Par exemple l'utilisation des boucles pour la

lecture de ligne . Et on obtient une fonction qui permet de compter le nombre de mots et le nombre de chaque lignes du fichier prit en entrée

compte_dans_fichiers(liste_fichiers):

Cette fonction est la meme que la fonction précedente mais cette fois ci il faut compter dans chaque fichiers d'une liste , donc j'ai reprit la fonction précédente et j'ai fais une boucle qui parcourt tous les fichiers de la liste de fichiers prit en entrées

mots_fichier(nom_fichier) :

J'ai avant tout créer deux listes , la première sert à renvoyer le résultat , la deuxième sert à recenser le nombre d'utilisations de chaque mots et sera ajoutée dans la premiere liste au fur et à mesure par conséquent on obtient une liste dont chaque élément est une liste de 2 éléments

mots_dans_fichiers(liste_fichiers):

Pour écrire cette fonction il faut utiliser plusieurs boucles (pour parcourir tous les fichiers , lecture de lignes ,etc) et on crée une liste pour contenir les mots puis on parcourt dans tous les fichiers si le mot apparait alors on incrémente si le mot n'est pas dans la liste alors on l'ajoute .À la fin on obtient une liste contenant plusieurs éléments dont chaque éléments est une liste qui recense le nombre d'utilisation de chaque mots dans tous les fichiers

apparition_mots(liste_fichiers):

Dans le script de cette fonction on appel aux fonctions définis précédemment selon le contexte :mots_dans_fichiers(liste_fichiers) s'il y a plusieurs fichiers et mots_fichier(liste_fichiers) s'il ya qu'un seul fichier, ces fonctions permet de compter le nombre d'utilisations de chaque mots, ensuite j'ai utilisé list.sort afin de mettre dans l'ordre décroissant le nombre d'utilisations des mots et enfin j'ai crée une boucle in range 0 15 pour former une liste des 15 mots apparaissant le plus fréquemment. Et pour finir j'ai utilisé fd.write pour écrire le résultat dans un fichier csv

mots_dans_fichiers_rep(rep) :

Dans cette deuxième partie on devait prendre en entrée des repertoires ainsi j'ai importé le module os pour l'utilisation de la méthode listdir qui permet de renvoyer un repertoire en une liste des fichiers ensuite j'ai construit une liste de mots puis on la parcourt si le mot n'est pas dans la liste alors on ajoute une

nouvelle paire dans le dictionnaire , si le mot est déjà dans le dictionnaire alors incrémenté sa valeur associée de 1

apparition_mots_rep(rep):

Dans le script de cette fonction , j'ai repris le résultat de la fonction précédente , puis j'ai extrait les informations utiles à cette partie c'est à dire les dictionnaires contenant le nombre d'utilisations de chaque mots puis j'ai utilisé la méthode sorted qui a la meme fonction que sort mais pour un dictionnaire , pour les mettre dans l'ordre décroissant et enfin j'ai fait une boucle pour prendre seulement les 15 paires des mots qui apparaissent le plus fréquemment

compte_mot_rep(mot, rep):

Dans ce script j'ai compté calculer la fréquence d'apparition d'un mot donné en entrée par l'utilisateur dans chacun des fichiers contenus dans le répertoire puis j'ai construit un graphique grace à la fonction matplotlib.pyplot.bar qu'il faut importer

lecture_tweet(rep):

Dans ce script j'ai fait des boucles pour lire tous les fichiers contenus dans le repertoire puis j'ai extrait la date de publication et le texte du tweet correspondant et je les ai ajouté dans une liste au fur et à mesure et qui sera ell aussi ajouté dans la liste des résultats

compte_tweet(rep):

Dans ce script j'ai repris le résultat de la fonction précédente puis j'extrait seulement les dates et je les ajoute dans le dictionnaire si la date n'y est pas encore , au contraire j'incrémente de 1 la valeur de cette date et pour finir jai déssiné un graphique grace aux fonctions importées : plt.figure , plt.bar

compte_mot_tweet_rep(mot, rep):

Dans ce script , j'ai tout d'abord crée une liste qui contient toutes les dates ensuite repris le résultat de la fonction lecture_tweet puis on parcourt uniquement les dates si la date correspond à la date entrée par l'utilisateur puis on appel à la fonction compte_mot_rep

Accomplissements et difficultés

Accomplissements : J'ai appris à utiliser des fonctions prédéfinies que je ne connaissais pas , j'ai étudier plus profondément l'utilisation et le fonctionnement des dictionnaires. Difficultés : difficultés à trouver et utiliser les nouvelles méthodes et solutions afin de pouvoir réussir

Conclusion

SAÉ intéressant qui permet d'acquérir plusieurs nouvelles notions mais qui nécessite énormément de recherche et de travail pour réussir

During the realization of this sae I acquired skills and new basic notions in programming which allows me to extract, analyze textual data but it was rather difficult to find methods, solutions, it took a lot of time but finally I was able to learn very useful notions to process data and it also allowed me to progress by writing several scripts which are very useful, in all these scripts that I wrote, I I reused several notions seen in the R107 resources