Rapport SAÉ 105 : Traiter des données

MOUIGNI HADJI 1A

I. Introduction

Dans cette SAÉ, le but était, comme son nom l'indique, de traiter des données, plus particulièrement des tweets. Cela va nous permettre de découvrir de nouvelles choses et de buter sur d'autres et de chercher une solution et de nous initier à des problèmes posés à un homme/femme qui est un professionnel des R&T
Afin de parvenir à nos résultats voulus, nous allons utiliser ce que nous avons acquis les modules R107 et R108
Le projet va se scier en deux parties. Pour la première partie, nous allons plus « préparer le terrain » pour pouvoir être efficace dans la partie 2 qui elle va être plus de l'ordre de l'étude de données avec de la lecture et écriture de graphiques, calculs de fréquence etc…

II. Fonctions utilisées

Donc pour parvenir à notre objectif, nous avons utilisé pas mal de fonction que l'on va détailler dans cette partie.

Partie 1 :

1. Environnement : Cette première « fonction » va être utilisée pour créer l'environnement de travail de cette SAÉ.
2. Compte_lignes_mots(nom_fichier) : comme son nom l'indique cette fonction va compter le nombre de lignes et de mots dans un fichier.
3. Compte_dans_fichier(liste_fichiers) : Pareil que le précédent sauf que cela est fait pour une liste de fichier.
4. Mots_fichier(nom_fichier) : Cette fonction va recenser tous les mots présents dans un fichier et compter le nombre d'occurrences de chaque mot.
5. Mots_dans_fichier(liste_fichiers) : Cette fonction ressemble a celle d'avant mais celle-là va déterminer le nombre d'occurrence d'un mot dans tout le fichiers donnés
6. Apparition_mots(liste_fichier) : Cette fonction va elle compter les 15 mots les plus fréquents de chaque fichier donné en paramètres et donne leur fréquence.

Partie 2 :

1. mots_dans_fichiers_rep(rep) : Cette fonction à la même utilité que celle qui référence les mots et leur occurrences dans une liste de fichiers mais cette fonction renvoie ses résultats sous forme de dictionnaire

2. Apparition_mots_rep(rep) : calcule la fréquence d'apparition des 15 mots apparaissant le plus fréquemment dans chacun des fichiers contenus dans le répertoire rep
3. Compte_mot_rep (mot, rep) : Calcule la fréquence d'apparition d'un mot donné par l'utilisateur dans chacun des fichiers présents dans le répertoire rep.
4. Lecture_tweet(rep) : cette fonction permet de lire tous les tweets contenus dans tous les fichiers contenus dans le répertoire rep.
5. Compte_tweet(rep) : compte le nombre de tweet par journée et renvoie un graphique qui représente le nombre de tweets par jour.
6. Compte_mot_tweet_rep(mot, rep) : Calcule 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 rep. Toutes ces fréquences seront affichées dans un graphique

III. Réalisation des fonctions

Partie 1 :

1. Environnement : J'ai réalisé cette fonction avec le module « os » qui m'a permis d'utiliser la fonction commande « os.mkdir » .
2. Compte_lignes_mots(nom_fichier) : Utilisation de la commande fd.readline**s** puis en la manipulant elle renvoie une liste de deux éléments qui sont le nombre de lignes et de mots.
3. Compte_dans_fichier(liste_fichiers) : Utilisation du module « glob » que j'ai très souvent utilisé dans cette SAÉ car elle renvoie liste dont les éléments sont les fichiers d'un répertoire. Et en le combinant avec la fonction précédente j'ai réussi à réaliser cette fonction.
4. Mots_fichier(nom_fichier) : Pour celle-là, j'avais une chaine de caractère qui contenait tous les caractères que je considérais et en balayant les mots du fichiers lorsqu'un caractère n'est pas reconnu, le programme le remplace par un espace avec la commande « replace » pour qu'en suite je puisse utiliser « split » pour mieux compter.
5. Mots_dans_fichier(liste_fichiers) : Pour celle-là j'ai essayé de broder quelque chose avec « glob.glob » mais je ne suis pas parvenu à le faire cela me renvoie pas ce que je veux.
6. Apparition_mots(liste_fichier) : Ø

Partie 2 :

1. mots_dans_fichiers_rep(rep) : Pour réaliser celle la j'ai utiliser le module globe et surtout la grosse « nouveauté » ce sont les dictionnaires qui permettent d'associer le mot à son nombre d'occurrences et c'est très pratique
2. Apparition_mots_rep(rep) : Ø
3. Compte_mot_rep (mot, rep) : Ø
4. Lecture_tweet(rep) : Pour cela j'ai importé le module « json » qui m'a permis de créer des dictionnaires, donc chaque tweet était sous la forme d'un dictionnaire ce qui rendait la manipulation des tweets beaucoup plus simple

5. Compte_tweet(rep) : Là les dictionnaires m'ont aider car j'ai seulement lister les valeurs des clés que je voulais et pour le graphique j'ai importé le module « matplotlib » qui m'a permis de créer mon graphique avec les données désirées
6. Compte_mot_tweet_rep(mot, rep) : Ø

IV. Accomplissements et difficultés

Accomplissement :

Pour ce qui est des accomplissements, sur l'ensemble j'ai réussi à répondre a 7 énoncés plus 1 sur lequel j'étais sur la « voie ». Maintenant sur les accomplissements « technique », ce projet m'a aider a mieux manipuler les listes malgré mes lacunes, la connaissances de nouveaux modules très intéressants comme glob, os, matplotlib mais je pense que la chose que j'ai le plus apprécié est le fait d'avoir découvert les dictionnaires que je trouve particulièrement pratique par rapport aux listes et j'aimerais mieux les maitriser mais ce projet m'a permis de le découvrir et donc à moi de mieux le revoir de plus près.

Difficultés :

Pour ce projet il y a certains énoncés que je n'ai pas réalisés. Souvent ce sont les plus compliqués et puis j'ai eu du mal à réaliser les fonctions qui prenaient en paramètre plusieurs fichiers mais cela s'est arrangé avec le découvert du module « glob ». Il y e a aussi des programmes que je n'ai pas fait car ils dépendaient ou étaient liés à certains que je n'arrivais pas à réaliser et donc je n'ai pas pu voir certaines choses je pense comme le fichier csv.

V. Conclusion

Le but de ce projet était d'analyser des tweets en avançant pas à pas mais avec les circonstances évoquées dans la partie difficultés, il n'a pas pu aboutir pour ma part mais il y a quand même des éléments d'analyse qui permettent de « jouer » avec les tweets.

Pour mon avis sur ce projet je pense que c'est un bon projet malgré les difficultés et certaines exigences dont je n'avais pas le niveau pour les réaliser. Mais quand bien même ce dernier m'a permis d'apprendre de nouvelles choses et d'approfondir certaines choses que je connaissais déjà.

English sum up :

The objective of the project is to analyse datas on text files. On these, we want to count the number of répétitions of each words in all files and know the frequency of occurencies of the fifteen words the most presents. First of all, we have to prepare the working environment and after that, we can start to code the first program which going to process data. For that, we are going to proceed in 2 parts. The first one is to analyse the texts by count words, lines and the second is more to presents the results of the analyse with graphs and frequency calculation. To reach our oubjective we will do it step by step until we can say that it's a real analysis. (I didn't reach the objective but i learn things during this project).