Analyses d'un fichier texte

Ces exercices vont vous guider pour obtenir quelques statistiques générales sur les mots d'un document texte.

Ils sont à faire individuellement, et vous préparent au travail qui sera à faire sur le projet en groupe.

Ici, on suppose que le document texte est déjà téléchargé et nettoyé de tout code HTML. On travaillera sur le fichier `candide.txt` disponible sur le dépôt du cours dans le dossier `docs`.

L'objectif général est d'obtenir la liste des mots et des bigrammes de mots les plus fréquents du texte.

Exercice 1 Préparation du fichier

1. Récupérer le fichier (un simple `pull` du dépôt devrait suffire) et copier le vers votre dossier de travail
2. Pour faciliter le comptage, transformer le texte pour obtenir un mot par ligne. Utiliser la commande `grep` est une bonne stratégie !
3. S'assurer que le texte est bien nettoyé : pas de ponctuation et tout en minuscule. La commande `tr` pourra vous être utile.
4. Faire un script qui effectue le traitement souhaité. Il doit prendre le chemin vers le fichier à traiter en argument et écrire le résultat sur la sortie standard.

envoyez votre réponse sur votre git !

Exercice 2 Édition d'une liste de fréquences

En vous inspirant du travail réalisé les semaines passées (où nous comptions les entités). Écrire un second script qui donne les mots les plus fréquents d'un texte. Ce script doit

1. faire appel au script de l'exercice 1 (et non pas recopier son contenu)
2. prendre le nom du fichier texte en premier argument
3. prendre le nombre de mots à afficher en second argument **optionnel**. Si ce nombre de mot n'est pas donné, afficher les 25 plus fréquents par défaut.

envoyez votre réponse sur votre git !

Exercice 3 Liste de fréquences de bigrammes (plus dur)

Pour cet exercice, vous serez moins guidé et il vous faudra faire preuve d'imagination.

On cherche à obtenir la liste de fréquences des bigrammes de mots. C'est à dire des suites de deux mots consécutifs.

Par exemple pour la phrase précédente, les bigrammes à obtenir seraient :
c est
est à
à dire
dire des
des suites
suites de
de deux
deux mots
mots consécutifs

Proposez un script semblable à celui produit à l'exercice 2, mais qui donne des résultats pour les bigrammes

Quelques remarques :

1. La commande `paste` vous sera utile.
2. On peut arriver au résultat sans utiliser de boucle dans notre script.
3. Vous pouvez générer des fichiers intermédiaires et aller vers le résultat en plusieurs étapes.

envoyez votre réponse sur votre git !