Projet - Détecteur de langue

S2 - Structures de données

Projet : détecteur de la langue d'un texte

Figure 1: Tour de Babel

L'objectif final de ce projet est de produire un programme qui détecte dans quelle langue un texte est écrit, en s'appuyant sur une analyse fréquentielle des lettres du texte.

Il s'agit d'un projet : les consignes sont donc volontairement très limitées et seul le résultat attendu est détaillé (cahier des charges). Il vous revient d'organiser votre travail comme vous le voulez : création de fichiers, de bibliothèques, recherches documentaires.

Vous pouvez exercer votre liberté et choisir le style de programmation qui vous plait : impératif, POO, modulaire …

Des compléments facultatifs sont proposés pour les plus rapides.

Étape 1

Créer un programme qui :

- lit un fichier texte (encodé en UTF-8) ;
- crée un dictionnaire dont les clés sont les lettres de l'alphabet de "a" à "z" ;
- analyse le contenu du fichier caractère par caractère et remplit le dictionnaire avec comme valeur le nombre d'apparitions de chacune des lettres dans le texte. Les lettres autres que les lettres "a" à "z" sont ignorées ;
- crée un nouveau dictionnaire avec les mêmes clés, mais en valeurs les fréquences d'apparition de chaque lettre ;
- affiche un diagramme en barres de cette répartition de fréquences à l'aide de la librairie MatPlotLib.

Remarques :

- il peut être judicieux de décomposer le programme en plusieurs fonctions ;
- vous rechercherez dans la documentation les exemples d'utilisation de MatPlotLib ;
- voici un fichier texte pour tester votre programme : texte.

Étape 2

La fréquence d'apparition des lettres dans les différentes langues qui utilisent l'alphabet latin est différente. C'est pour cette raison que les points attribués aux différentes lettres ainsi que le nombre de ces lettres dans le jeu du Scrabble ne sont pas les mêmes dans tous les pays.

Le tableau disponible sur Wikipédia à cette adresse : -Wikipédia- donne cette fréquence dans les langues les plus courantes.

Nous allons choisir comme *signature* d'une langue la liste des dix lettres les plus utilisées dans cette langue, de la plus utilisée à la moins utilisée.

Créer un dictionnaire `signature` dont les clés sont le nom des langues sous forme de chaîne de caractères et les valeurs sont ces listes de dix lettres.

Par exemple, l'appel `signature["Français"][0]` retournera `"e"`.

On se limitera aux langues suivantes : Français, Anglais, Allemand, Espagnol, Italien, Portugais, Espéranto, Polonais et Néerlandais.

Étape 3

À partir du programme réalisé à l'étape 1, programmer une fonction qui, pour un fichier texte donné, crée une liste `signature_texte` avec pour éléments les dix lettres les plus utilisées dans ce texte, de la plus utilisée à la moins utilisée.

Écrire une fonction `detecte_langue(texte)` qui, à partir des éléments précédents, retourne le nom de la langue dans laquelle le `texte` donné en paramètre est le plus probablement écrit. Il vous faudra notamment choisir un moyen de comparer la liste `signature_texte` avec l'ensemble des listes du dictionnaire `signature` afin de trouver cette qui est la plus "proche".

Pour tester votre programme, voici des fichiers texte dans les différentes langues concernées : textes (fichier .zip à décompresser).

Proposer des améliorations possibles.

Complément facultatif

Programmer une fonction `detecte_langue(url)` qui prend en argument l'adresse d'une page web et qui retourne le nom de la langue dans laquelle la page est le plus probablement écrite. Vous pourrez utiliser la bibliothèque Python BeautifulSoup qui permet d'extraire le texte présent dans les balises HTML d'une page web (voir la documentation).