projet

Générateur de néologismes

T. NSI

Problématique

Connaissez-vous la signification des mots « ongoscate », « hymone » ou « terouet » ? Contrairement aux apparences, ces mots n'existent pas en français : ils ont été fabriqués par un algorithme qui s'inspire de vrais mots français pour en créer de nouveaux qui « sonnent français » mais n'existent pas dans cette langue.

La direction commerciale d'une entreprise souhaite donner à ses nombreux produits des noms qui « sonnent » français ou anglais et vous demande de créer un algorithme qui permettent de générer un grand nombre de néologismes dans ces deux langues.

Votre projet

Votre chef de projet vous demande de :

1. présenter au client les concepts nécessaires à la compréhension de votre méthode de génération.
2. développer un programme Python capable de générer des néologismes en français et en anglais

Les programmes seront fournis dans un fichier Python à part. Votre analyse de ces programmes et la discussion sur la méthode seront présentés à l'oral en vous appuyant sur le logiciel de présentation de votre choix (diapositives).

Génération de néologismes

Dans un mot en français, la lettre « t » a une plus grande probabilité d'être suivi par un « e » que par un « x ». En anglais, il est plus probable que le « t » soit suivi par un « h » que par un autre « t ».

	a	b	c	d	e	f	g	…	x	y	z	fin
a												
b												
c												
d												
e												
f												
…												
x												
y												
z												

Il vous est donc demandé :

1. de créer un algorithme d'analyse des mots d'un dictionnaire et de générer un tableau statistique comme ci-contre, indiquant pour chaque lettre (représentées par les lignes) le nombre fois où elle est suivi par la lettre de chaque colonne, ainsi que le nombre de fois où elle est la dernière lettre du mot.

   L'analyse d'un dictionnaire français devrait faire apparaître un nombre très faible dans la ligne « t » colonne « x », et bien plus grand dans la ligne « t » colonne « e ».
2. d'utiliser ce tableau pour fabriquer un certain nombre de mots de 4 à 12 lettres, dont les enchaînement de lettres suivent la probabilité décrite par le tableau.
3. d'analyser les mots ainsi fabriqués, avec un dictionnaire français comme avec un dictionnaire anglais, de souligner les éventuelles limites de la méthode, et de sélectionner un certain nombre de mots bien formés dans chaque langue.

Fichiers fournis

Les fichiers *frdic.txt* et *endic.txt* sont encodés au format **utf8** et contiennent les mots des dictionnaires français et anglais issus des fichiers ouverts de Libreoffice, qui ont été traités afin d'enlever les caractères spéciaux et remplacer les caractères accentués par leur équivalent sans accents. Ils peuvent néanmoins contenir des majuscules.

Notation

Attendus techniques

Vous devrez rendre un code Python documenté avec des noms de variable et de fonction bien choisis.

Vous devrez présenter à l'oral :

1. L'interface de votre programme avec notamment :
   - Les préconditions : sous quel format sont attendues les données en entrée, quels sont les attendus implicites de votre programme (données vides acceptées?) etc...
   - Les postconditions : description précise de ce que renvoie votre programme (notamment lorsque les préconditions ne sont pas satisfaites) et sous quel format.
2. Une discussion sur la complexité des algorithmes que vous avez implémenté.

Attendus généraux

Les points suivants sont pris en compte pour la notation :

- Capacité à rechercher des informations, autonomie et initiative.
- Capacité à présenter les résultats obtenus de manière critique, et à analyser la méthode de votre chef de projet, ses limites, et les améliorations que l'on pourrait envisager.
- Capacité à argumenter vos choix :
  - des algorithmes (sous fonction(-alité), complexité…)
  - de validation et de documentation (interface, tests,…)
- La qualité du support (diapositives…) et de la présentation orale.