IFT6285 – Devoir 3

Correction de mots

Maxime MONRAT
Juan Felipe DURAN

Université de Montréal
Automne 2021

a) Librairies utilisées dans le devoir:

1. **Jaro-Winkler** (*import jaro*), pour calculer les distances de Jaro-Winkler,
2. Librairies python natives : *time, urllib* et *operator*.

Les distances d'édition et cosinus ont été implantées à la main. La distance d'édition a été inspirée du code de **PySpellChecker** basé sur le blog de Peter Norvig, et la distance cosinus de ce fil de StackOverflow.

b) Le programme **eval** prend en entrée la sortie du correcteur (*devoir3-sortie.txt*), puis vérifie pour chaque ligne si le mot à corriger correspond à un mot du fichier d'entraînement. Si c'est le cas, il vérifie si la correction du fichier d'entraînement est présente parmi les corrections suggérées par **corrige**. Le programme présente deux tests différents : un test lourd (*hard test*) et un test léger (*soft test*)

1. **Hard Test**
   Ce test donne 5 points si la solution se trouve à la 1ère place dans les suggestions, 4 si elle est à la 2nde etc. Il effectue ensuite une moyenne pondérée des scores et renvoie le pourcentage de réussite.

2. **Soft Test**
   Le test léger vérifie simplement si la solution se trouve parmi les propositions. Si c'est le cas, il attribue un point. On effectue ensuite également la moyenne pondérée des résultats pour obtenir un pourcentage.

c)

distance	Hard Test	Soft Test	temps de traitement (s)
Jaro-Winkler	**67.95%**	84.88%	15057.64
Édition	67.87%	**90.90%**	**1835.72**
Cosinus	47.69%	64.59%	7247.88

Table 1 Impact des différentes distances sur les performances et le temps de correction

d) D'après les résultats obtenus (voir en c. Table 1), la meilleure distance à considérer est ici la distance d'**édition** : bien plus rapide à calculer, c'est elle qui offre le meilleur résultat de test léger (ce qui veut dire que 9 fois sur 10 la solution se trouve dans les 5 propositions de correction).

Le code de **corrige** est inspiré de **PySpellChecker** :
Il prend en entrée un mot mal écrit et le compare à toutes les chaînes à 1 ou 2 distances d'édition. Parmi ces chaînes on sélecte les 5 qui correspondent aux mots d'un corpus de texte, classés par ordre décroissant de fréquence.
Bien que cette méthode soit efficace pour les fautes d'orthographe et les fautes de frappe, elle reste moins pertinente pour les mots d'emprunts (tels que des mots d'anglais dans un contexte en Français). Par ailleurs les mots étant très différents de leur version correcte auront également du mal à être corrigés.