SimplonSearch : Le nouveau moteur de recherche

Contexte :

La société Simplon, organisme de formation réputé, a mis à disposition de ses chargés de projet et de ses formateurs un ensemble de documents contenant toutes les informations nécessaires pour mener à bien une formation. On y trouve des documents expliquant comment mettre en place une alternance, comment composer un jury, ou préparer une bonne rentré.

Le seul soucis, c'est qu'avec tous ces documents les salariés de Simplon n'arrive plus à s'y retrouver et à savoir quel est le document qui contient les informations les plus pertinentes vis-à-vis de leurs questionnement.

Vous avez donc été mandaté pour créer un moteur de recherche : « SimplonSearch ».

L'objectif :

Trouver dans un texte non pas des phrases exactes, mais des phrases ressemblant autant que possible à un contenu recherché et d'évaluer par un score la pertinence des différentes suggestions.

Le but de ce breif est de vous faire travailler l'usage des expressions régulières (très utilisé dans le traitement des données) dans un contexte de mise en situation réelle.

La pertinence des mots clés n'est qu'un critère d'un moteur de recherche efficaces, mais n'en demeure pas moins primordiale. Ici c'est ce seul critère qui sera abordé.

1. Proposer une fonction *almost(mot, s)* qui trouve dans un texte s toute les occurrence d'un mot dont une lettre a pu éventuellement être enlevée. Exemple : si *mot* vaut « alphonse », une réponse acceptable vaut est « alponse ».

2. Proposez une fonction *pluslarge(mot, s)* qui trouve dans un texte *s* toutes les occurrences d'un mot dont une lettre a pu éventuellement être enlevée, ajouté ou remplacé par une autre.

3. Proposez une fonction *score(p, s)* qui prend en argument une phrase *p* et lui attribue un score en fonction de la présence des mots qu'elle contient dans *s* : cinq points par mot exact, un point par mot approché.

4. Modifiez les fonctions précédentes de façons à ce qu'elles donnent un bonus de 20 points si deux mots successifs dans la phrase initial sont également successif dans le texte parcouru (les mots exacts).
   Exemple : Si on cherche « Le petit bonhomme en mouse » dans « Ce superbe matelas en mousse naturelle ».

Facultatif :

Vous pouvez utiliser la fonction nouvellement créé sur un ensemble de fichier txt et voir quel est le fichier qui correspond le plus à une phrase recherchée.

Pour se faire :

1. Importé « os » et utilisez la fonction *os.listdir()* afin de récupérer dans une liste des fichier présent dans dossiers.
2. Vous pouvez lire chacun de ces fichiers :
   a. Ouvre le fichier avec la fonction *open()* (pensez à spécifier l'encodage, « encoding="utf-8" ») et assigné le à une variables.
   b. Utilisez la méthodes *.read()* et assigné l'objet retourné à une nouvelle variables. Cette variable contiendra maintenant le contenu du fichier txt sous forme de chaîne de caractères.
   c. Utilisez la fonction *score2()* sur cette chaîne de caractères et stocké le résultat dans un dictionnaire avec le nom du fichier pour clé et le score pour valeur. Vous pouvez utiliser une boucle for pour répéter cela pour tous les fichiers.
   d. Afficher le contenu du dictionnaire en ordonnant les résultats par score descendant. Utiliser la fonction *sorted()*.