UNIVERSITÉ DE LORRAINE

Traitement de corpus

TD 5 : Fichiers

1 Traitement de fichiers

1. Écrire une fonction `nettoyage` qui prend en paramètre un texte et qui renvoie le texte après les pré-traitements suivants :
   — Mettre le texte en minuscules.
   — Retirer les ponctuations suivante : ,.;!?
   — Remplacer les sauts de lignes \n par un espace avec la fonction `replace()`.

   Note : En cours, nous avons vu la fonction `strip()` qui permet de retirer les sauts de lignes et les espaces mais **seulement** en début et en fin de chaîne de caractères. Pour supprimer les sauts de lignes au sein du texte, on utilise donc la fonction `replace()`.
2. Ouvrir le fichier `texte_extrait.txt` dans une variable appelée `fichier`.
3. Stocker le contenu du texte dans une variable appelée `contenu` et l'afficher.
4. Appliquer la fonction `nettoyage` et stocker le texte nettoyé dans une variable appelée `texte_propre`.
5. Couper le texte nettoyé en mots
6. Afficher le nombre total de mots.
7. Récupérer les mots sans doublons dans une variable appelée `vocab` avec la fonction `set()`.

   Note : La fonction `set()` permet de retirer les doublons dans une liste. Elle renvoie un objet appelé **ensemble** où chaque élément est unique.

```python
mots = ["hello", "hola", "bonjour", "bonjour", "hello", "hello"]
chiffres = [3, 3, 2, 1, 2, 5, 2, 3]

print(set(mots))
print(set(chiffres))
```

```
{'bonjour', 'hello', 'hola'}
{1, 2, 3, 5}
```

8. Afficher la taille du vocabulaire (soit le nombre de mots du vocabulaire).
9. Écrire dans un nouveau fichier appelé `infos_texte.txt` le texte nettoyé, la liste des mots du texte, le nombre total de mots, les mots du vocabulaire, le nombre de mots du vocabulaire.
   Votre fichier final doit ressembler à ça :

```
Le texte nettoyé : denise était venue à pied...
Les mots du texte sont : ['denise', 'était', 'venue', 'à', 'pied', ...]
Le nombre total de mots est de : 171
Les mots du vocabulaire sont : {'nuit', 'gare', 'côté', '... }
Le nombre de mots du vocabulaire est de : 127
```

   Attention ! N'oubliez pas d'ajouter des sauts de lignes entre chaque ligne sinon tout sera écrit à la suite.