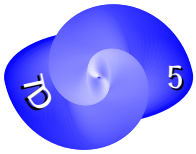

concepts et méthodes d'analyse lexicale

NLTK

Exercice 1

Sites de référence

1. http://nltk.org/. Site de base.
2. http://nltk.org/book3/. C'est le site du livre *Natural Language Processing with Python* aux éditions O'reilly.
3. http://streamhacker.wordpress.com/tag/nltk/ site présentant des exemples intéressants de Web mining.

Fragment 1 – initialisation

```
# on importe les donnees du livre
from nltk.book import *
text1.concordance('whale')
# puis dans une autre cellule
text1.similar('fear')
```

Exercice 2

débuts

On importe la librairie et les données du livre. Le corpus est beaucoup plus vaste que ça mais pour démarrer c'est suffisant. Après on peut importer son propre corpus de textes.

- Importez le corpus si ce n'est pas déjà fait comme dans 1
- Quel est le `text1` ?
- Trouvez les concordances avec le mot *whales*.
- Puis les similarités avec *fear*
- Pour obtenir la liste des mots du texte triés :**sorted**(**set**(text1))

Fragment 2 – Diversité

```
def diversite(text):
        return len(text)/len(set(text))
```

Exercice 3

répartition dans le texte

- Comment interpréter la fonction diversite ? Utilisez la pour évaluer la diversité lexicale dans *MobyDick*.
- Comparez avec d'autres corpus.

Fragment 3 – répartition dans le texte

```
text4.dispersion_plot(["citizens", "democracy", "freedom", "duties", "America"])
```

Exercice 4 — *Répartition*

1. Que fait la commande ci-dessus ?
2. Afficher des dispersions de mots dans d'autres corpus.

Fragment 4 – Génération de texte

```
text4.generate()
```

Exercice 5 — *Generation*

1. Essayez la fonction generate ci-dessus
2. Testez la sur le corpus de *chat* de NLTK

Fragment 5 – Obtention du Vocabulaire

```
sorted(set(text))
```

Exercice 6 — *Generation*

1. Expliquez pourquoi le code ci-dessus permet de récupérer le Vocabulaire associé à un corpus.
2. Quel est le vocabulaire du corpus de *chat* de NLTK ? Comparez le à celui des oeuvres littéraires.

Fragment 6 – Occurences d'un mot

```
text.index('moon')
```

Exercice 7 — *Occurences d'un mot*

Trouver la première occurrence du mot *whale* dans *Moby Dick* ainsi que tous les mots qui l'entourent. Trouver toutes les occurences de ce mot avec `count`

Fragment 7 – Mot fréquents et rares

```
fdist1 = FreqDist(text1)
Vocabulaire=fdist1.keys()
```

Exercice 8 — *Répartition*

1. Que font les commandes ci-dessus ?
2. ajoutez ensuite fdist1.plot(30, cumulative=True)
3. puis testez : fdist1.hapaxes()

Fragment 8 – Filtrages de mots avec les compréhensions

```
V = set(text1)
mots = [m for m in V if len(m) > 12]
```

Exercice 9

Filtrage

1. Que font les commandes ci-dessus ?
2. Pouvez vous filtrer tous les mots de la Genèse comportant entre 3 et 6 lettres et commençant par la lettre g ?