



Module :  
Recherche d'information Et  
Indexation



Dans ce projet j'ai essayé de réaliser une application de l'indexation sémantique, et la recherche d'information, à l'aide d'une source externe : « Word net »

Dans ce guide je vais vous donner les outils et les méthodes pour le bon fonctionnement de l'application :

Au début je vous conseille d'utiliser ANACONDA, est une distribution libre et open source du langage de programmation Python. C'est un logiciel qui facilite plusieurs taches au niveau d'installation des packages.

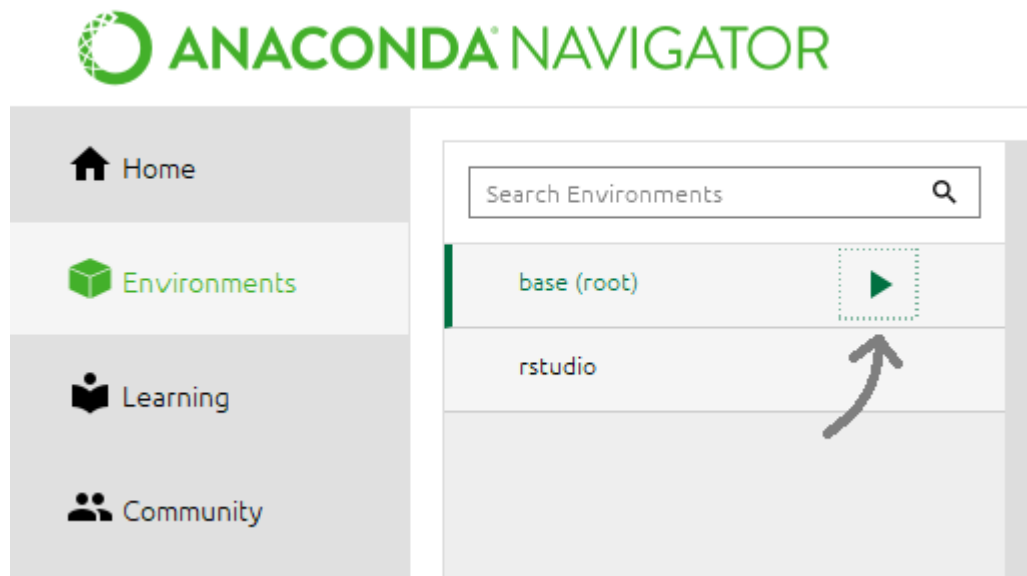
### *1<sup>ère</sup> étape : Installation ANACONDA :*

Selon votre machine et votre système d'exploitation choisissez laquelle la distribution qui vous convient, attention vous devez télécharger la version 3.7. Voilà le lien qui vous amène:

<https://www.anaconda.com/distribution/#download-section>

### *2<sup>ème</sup> étape : Installation NLTK :*

Pour cela il faut d'abord ouvrir ANACONDA et cliquez sur le bouton Environnement qui se trouve à gauche. Après vous ouvrirez le terminal : en cliquant sur le triangle indiqué dans la figure ci-dessous choisissez « Open Terminal ».



Quand le terminal est ouvert coller la commande suivante :

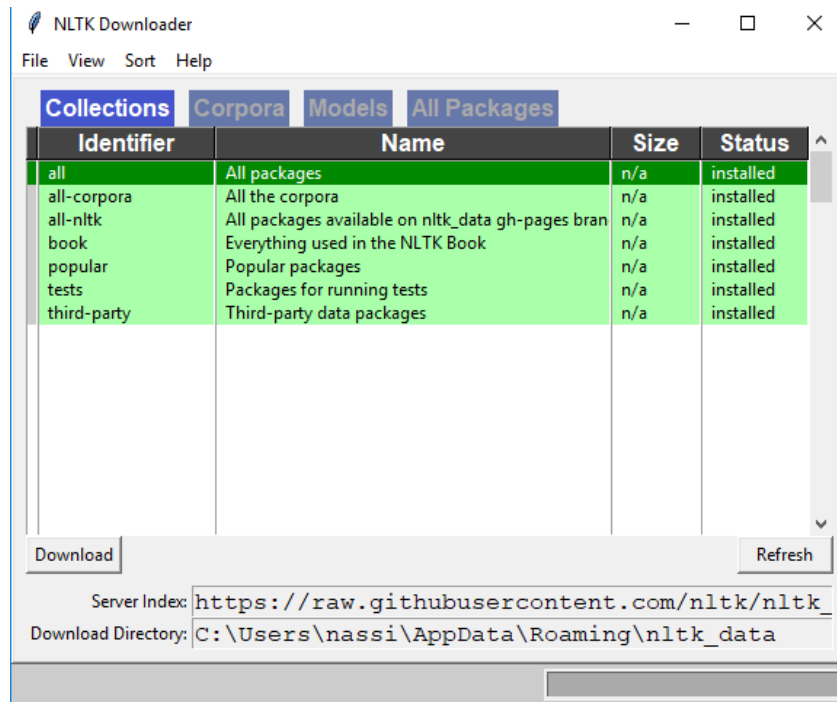
```
conda install -c anaconda nltk
```

À un moment le programme va vous afficher une question de vérification de téléchargement " Proceed ([y]/n)". Vous répondez par "y" (ça veut dire yes).

Maintenant vous cliquez sur le même triangle ci-dessus mais cette fois-ci choisissez "Open with Python" quand le terminal est ouvert vous tapez ou collez les commandes suivantes:

```
>>> import nltk
>>> nltk.download()
```

Une fenêtre s'affichera, choisissez "all" et cliquez sur le bouton download qui se trouve en bas à gauche.



A ce niveau-là vous êtes normalement capable d'exécuter l'application sans aucun problème

### *3<sup>ème</sup> étape : exécution de l'application :*

Vous trouverez l'application dans le dossier que je vous ai envoyé, le nom du fichier est "app\_rech\_sem.py". Ouvrez-le via "Spyder" un outil qui se trouve dans ANACONDA et vous exécutez le programme (sinon vous pouvez exécuter l'application en faisant double clics sur le programme).

Une fenêtre s'affichera avec un champ où vous spécifiez le chemin vers le dossier qui contient vos fichiers ".txt", quand vous validez les boutons d'indexation seront activés avec cet ordre : quand vous cliquez sur le bouton "tokens" il activera le bouton "tokens nettoyés" et ainsi de suite jusqu'à le dernier bouton "synset". Pour chaque bouton cliqué les résultats seront affichés.

Quand vous validez le chemin le texte de chaque fichier sera affiché:

chemin fichier :

valider

chercher mot :

recherche

tokens

tokens  
nettoyé

occurence

lemmatisation

synset

E\dossier\A.txt  
stars name hello friends start

E\dossier\B.txt  
The Original original original Series began in 1864 and continues to the present. In 1867 an Extra Series was started, of texts already printed but not i  
n satisfactory or readily obtainable editions. In 1921 the Extra Series was discontinued and all publications were subsequently listed and numbered a  
s part of the Original Series. In 1970 the Supplementary Series was initiated: volumes in this series are issued only occasionally and as suitable texts  
become available.

E\dossier\C.txt  
the wall is great I don't know if it is going to save us

E\dossier\D.txt  
original original original original original original original original original original original original the wall is great I don't know if it is going to save us

Quand vous cliqué sur toukens:

tokens

tokens  
nettoyé

occurence

lemmatisation

synset

A.txt:  
stars  
name  
hello  
friends  
start  
B.txt:  
The  
Original  
original  
original  
Series  
began  
in  
1864  
and  
continues  
to  
the  
present  
.  
In  
1867

toukens nettoyés:

tokens

tokens  
nettoyé

occurence

lemmatisation

synset

A.txt:  
stars  
name  
hello  
friends  
start  
B.txt:  
original  
original  
original  
series  
began  
continues  
present  
extra  
series  
started  
texts  
already  
printed  
satisfactory  
readily  
obtainable  
editions  
extra  
series  
discontinued  
publications

Occurrences :

tokens

tokens  
nettoyé

occurence

lemmatisation

synset

A.txt:  
stars-->1  
name-->1  
hello-->1  
friends-->1  
start-->1  
B.txt:  
original-->4  
series-->6  
began-->1  
continues-->1  
present-->1  
extra-->2  
started-->1  
texts-->2  
already-->1  
printed-->1  
satisfactory-->1  
readily-->1  
obtainable-->1  
editions-->1  
discontinued-->1  
publications-->1  
subsequently-->1  
listed-->1

## Lemmatisation et Synset :

	A.txt: star name hello friend start B.txt: original series begin continue present extra start text already print satisfactory readily obtainable edition discontinue
tokens	
tokens nettoyé	
occurrence	
lemmatisation	
synset	

	A.txt: star---star name---name hello---hello friend---friend start---start B.txt: original---master series---series begin---Begin continue---continue present---present extra---supernumerary start---start text---text already---already print---print satisfactory---satisfactory readily---readily obtainable---gettable edition---edition discontinue---discontinue publication---publication subsequently---subsequently list---list numbered---total part---part
tokens	
tokens nettoyé	
occurrence	
lemmatisation	
synset	

Après, le bouton recherche sera activé, maintenant vous pouvez faire des recherches sur les fichiers du dossier désigné au début.

chemin fichier :	<input type="text" value="E:\dossier"/>	<input type="button" value="valider"/>
chercher mot :	<input type="text" value="start"/>	<input type="button" value="recherche"/>

tokens

tokens  
nettoyé

occurrence

lemmatisation

synset

requete : start  
  
le fichier :A.txt    contient :start  
le fichier :B.txt    contient :start  
le fichier :C.txt    contient :go  
le fichier :D.txt    contient :go