3.1. Introducció

El Natural Language Toolkit (NLTK) és un conjunt de llibreries i programes per a Python que ens permet dur a terme moltes tasques relacionades amb el Processament del Llenguatge Natural. Moltes de les tasques que necessitarem fer ja estaran programades eficientment a l'NLTK i les podrem fer servir directament en els nostres programes. A més dels programes, es distribueixen també corpus i altres dades lingüístiques. És una plataforma molt útil tant per a l'ensenyament com per al desenvolupament i la recerca.

Aquest Toolkit s'acompanya d'un llibre molt interessant que es pot consultar en línia en el següent enllaç: http://www.nltk.org/book/

Els arxius disponibles d'aquesta unitat són:

- Aquests materials en format PDF: 03-NLTK-cat.pdf
- El Notebook de Jupter: cap3-cat.ipynb
- Enllaç al notebook a Google Colab: https://colab.research.google.com/github/aoliverg/python/blob/master/notebooks/cap3-cat.ipynb
-

3.2. Instal·lació de l'NLTK

A l'apartat 1.3 vam explicar com intal·lar Python 3 i ja vam preveure d'instal·lar una versió totalment compatible amb NLTK. A la plana http://www.nltk.org/install.html s'explica detalladament com instal·lar NLTK. Reproduim aquí la informació amb algun detall addicional.

3.2.a. Instal·lació a Windows

Les noves versions de Pyhton (a partir de la versió 3.5) incorporen el pip per a la instal·lació de paquets, llibreries, etc. Així que la manera més senzilla d'instal·lar Python serà fer servir pip, de la següent manera:

Primer de tot cal obrir una pantalla de Símbolo de sistema com a administrador, Ves a Inicio i cerca cmd i quan aparegui la icona de cmd fes clic amb el botó dret del ratolí i en el menú que surt selecciona Executa'l com a administrador. En Símbolo de sistema escriu:

```
pip install nltk
```

i després

```
pip install numpy
```

3.2.b. Linux i Mac

Per instal·lar NLTK obre un terminal i fes (recorda que si crides a Python3 fent servir python3 hauràs de canviar pip per pip3):

```
sudo pip install -U nltk
```

Opcionalment podem instal·lar NumPy, fent des d'un terminal

```
sudo pip install -U numpy
```

Per provar la instal·lació entra en un terminal, escriu python i a l'intèrpret interactiu escriu

```
import nltk
```

Quan feu sudo, us demanarà la contrassenya d'aministrador, que segurament serà la mateixa que feu servir per entrar al sistema.

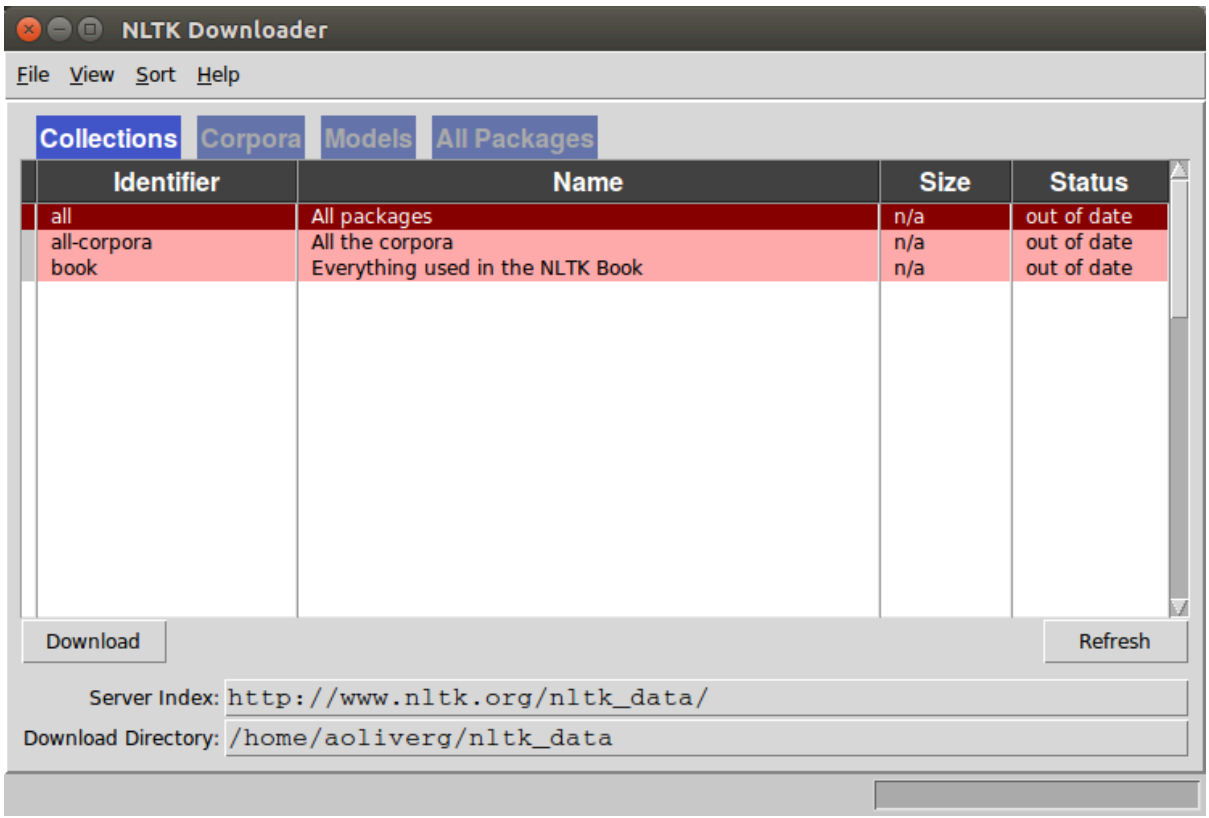
3.3. Instal·lació de les dades lingüístiques de l'NLTK

NLTK proporciona moltes dades lingüístiques: llistes de paraules, corpus, models de llenguatge, etc. Es pot veure una llista completa i actualitzada de les dades lingüístiques proporcionades amb l'NLTK a http://www.nltk.org/nltk_data/.

Per instal·lar les dades obrim un intèrpret interactiu de Python i escrivim:

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

Apareixerà una finestra com la següent



Aquí podem seleccionar all i Download. D'aquesta manera descarregarem totes les dades disponibles.

Probablement si treballes amb Linux o Mac us apareixerà un menú de text enlloc d'una finestra gràfica:

```
    >>> import nltk
>>> nltk.download()
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader>
```

Amb l'opció d seleccionem Download:

Downloader> d

Download which package (l=list; x=cancel)?
Identifier>

Amb l ens mostrarà la llista de recursos disponibles. Podem escriure all per descarregar-los tot.

3.4. Exemples d'ús

En aquesta secció presentem uns breus exemples, que executarem des de l'intèrpret interactiu, i ens serviran per verificar la instal·lació de l'NLTK i les dades, i veure algunes funcionalitats.

Exemple de tokenització (veurem a fons què és a la secció 4.4):

>>> import nltk

>>> text="This is a sentence. This is another sentence."

>>> nltk.tokenize.word_tokenize(text)

['This', 'is', 'a', 'sentence', '.', 'This', 'is', 'another', 'sentence', '.']

Un exemple d'etiquetatge morfosintàctic (tema que veurem a fons a la secció 5.3)

>>> tokenized=nltk.word_tokenize(text)

>>> nltk.pos_tag(tokenized)

[('This', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('sentence', 'NN'), ('.', '.'), ('This', 'DT'), ('is', 'VBZ'), ('another', 'RP'), ('sentence', 'NN'), ('.', '.')]

Un exemple d'accés a les dades de l'NLTK, en aquest cas a un corpus etiquetat del català.

>>> from nltk.corpus import cess_cat

>>> cess_cat.words()

['El', 'Tribunal_Suprem', '-Fpa-', 'TS', '-Fpt-', 'ha', ...]

>>> cess_cat.tagged_words()

[('El', 'da0ms0'), ('Tribunal_Suprem', 'np0000o'), ...]