3. Natural Language Toolkit (NLTK)

3.1. Introducción

El Natural Language Toolkit (NLTK) es un conjunto de librerías y programas para Python que nos permiten lleva a cabo muchas tareas relacionadas con el Procesamiento del Lenguaje Natural. Muchas de las tareas que necesitaremos hacer ya estarán programadas de manera eficiente en NLTK y las podremos usar directamente en nuestros programas. Además de los programas, se distribuyen también corpus y otros datos lingüísticos. Es una plataforma muy útil tanto para la enseñanza cómo para el desarrollo y la investigación.

Este Toolkit se acompaña de un libro muy interesante que se puede consultar en linea en el siguiente enlace: http://www.nltk.org/book/

3.2. Instalación de NLTK

En el apartado 1.3 explicamos cómo intal·lar Python 3 y ya previmos de instalar una versión totalmente compatible con NLTK. En la página http://www.nltk.org/install.html se explica en detalle cómo instalar NLTK. Reproducimos aquí la información con algún detalle adicional.

3.2.a. Instalación en Windows

Las nuevas versiones de Pyhton (3.5 o superior) incorporan la utilidad **pip** para la instalación de paquetes, librerías, etc. Así que la manera más sencilla de instalar NLTK será usar **pip**, de la siguiente manera:

Antes que nada hay que abrir una pantalla de Símbolo de sistema como administrador, Ve a Inicio y busca **cmd** y cuando aparezca el icono de **cmd** haz clic con el botón derecho del ratón y en el menú que aparece selecciona *Ejecútalo como administrador*. En Símbolo de sistema escribe:

```
pip install nltk
```

y después

```
pip install numpy
```

3.2.b. Linux y Mac

Para instalar NLTK abre un terminal y escribe

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

Opcionalmente podemos instalar NumPy, haciendo desde un terminal

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

Para probar la instalación entra en un terminal, escribe python y en intérprete interactivo escribe

```
import nltk
```

Cuando pongáis sudo, os pedirá la contraseña de administrador, que seguramente será la misma que uséis para entrar al sistema.

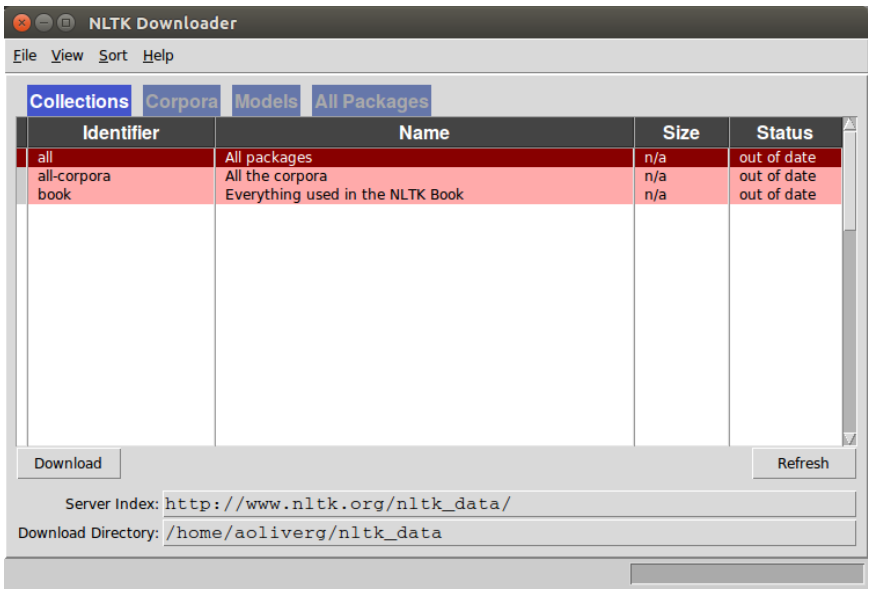
3.3. Instalación de los datos lingüísticos de NLTK

NLTK proporciona muchos datos lingüísticos: listas de palabras, corpus, modelos de lenguaje, etc. Se puede ver una lista completa y actualizada de los datos lingüísticos proporcionados con el NLTK en http://www.nltk.org/nltk_data/.

Para instalar los datos abrimos un intérprete interactivo de Python y escribimos:

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

Aparecerá una ventana cómo la siguiente:



Aquí podemos seleccionar **All** y **Download**. De este modo descargaremos todos los datos disponibles.

3.4. Ejemplos de uso

En esta sección presentamos unos breves ejemplos, que ejecutaremos desde el intérprete interactivo, y nos servirán para verificar la instalación de NLTK y los datos, y ver algunas funcionalidades.

Ejemplo de tokenització (veremos a fondo qué es a la sección 4.4):

```
>>> import nltk
>>> texto="This is a sentence. This is another sentence."
>>> nltk.tokenize.word_tokenize(texto)
['This', 'is', 'a', 'sentence', '.', 'This', 'is', 'another', 'sentence', '.']
```

Un ejemplo de etiquetado morfosintáctico (tema que veremos a fondo a la sección 5.3)

```
>>> tokenized=nltk.word_tokenize(texto)
>>> nltk.pos_tag(tokenized)
[('This', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('sentence', 'NN'), ('.', '.'),
('This', 'DT'), ('is', 'VBZ'), ('another', 'RP'), ('sentence', 'NN'), ('.', '.')]
```

Un ejemplo de acceso a los datos del NLTK, en este caso a un corpus etiquetado del catalán.

```
>>> 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'), ...]
```