Guía de instalación y uso del programa

1. Requerimientos previos

En este apartado se explicarán los distintos pasos que hay que seguir para poder ejecutar el programa:

- Tener un ordenador con un editor de código fuente que permita trabajar en Python. A la hora de ejecutar el programa no será necesario estar conectado a internet.

- El ordenador debe tener instalado Python. En el momento de realizar este documento, la versión más nueva tiene que ser Python 3.8.10, ya que es la última para la que todas las librerías están actualizadas.

- Hay que instalar las siguientes librerías:

  - pdf2images: se instala con el comando "*pip install pdf2image*". Una vez se ha instalado la librería, se deberá instalar la biblioteca poppler. Para su instalación la mejor opción es descargar la última versión disponible en el enlace https://github.com/oschwartz10612/poppler-windows/releases/. En el momento que finalice la descarga, se deberá acceder a la carpeta descargada y llegar hasta la carpeta /bin. En este momento se deberá copiar la ruta y pegarla en el parámetro *poppler_path* que se encuentra en la función *pdftoimage.*

  - pytesseract: se instala con el comando "*pip install pytesseract*". Después habrá que descargar pytesseract en el siguiente enlace https://github.com/UB-Mannheim/tesseract/wiki y una vez se ha descargado e instalado se tendrá que entrar en la carpeta instalada y copiar la ruta del archivo *tesseract.exe* y pegarla en los parámetros *pt.pytesseract.tesseract_cmd* que se encuentra en la función *imagetotxt.*

  - PDFMiner: se instala con el comando "*pip install pdfminer*".

  - nltk: se instala con el comando "*pip install --user -U nltk*". Cuando se ha completado la instalación, se debe entrar en la Shell de Python e introducir los comandos *import nltk* y *nltk.download().* Una vez se han introducido estos comandos se abrirá la siguiente ventana:

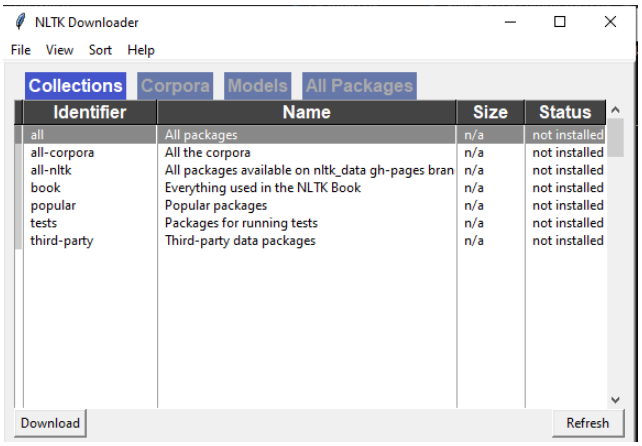

Figura 1: NLTK Downloader antes de la descarga

Cuando la ventana esté abierta, se debe seleccionar la opción *all* y después *download* hasta que todos los campos estén en verde:

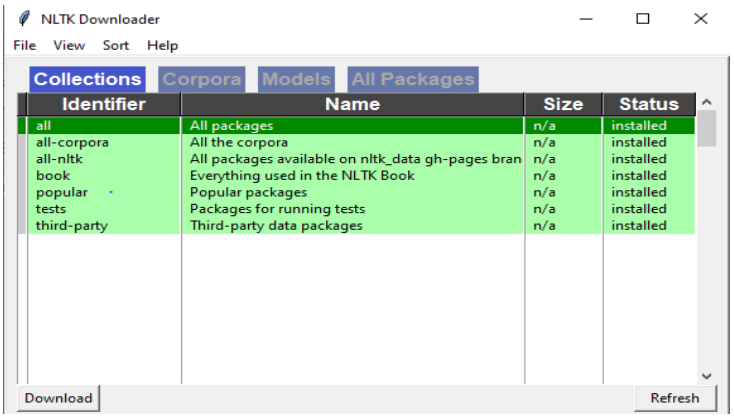

Figura 2: NLTK Downloader después de la descarga

- Numpy: se instala con el comando "*pip install numpy*"

- gensim: se instala con el comando "*pip install gensim==3.8.1*".

- Para el funcionamiento del programa se debe especificar la ruta donde se encuentran los documentos PDF con los que vamos a trabajar, por lo que es preferible agruparlos en la misma carpeta.

- Para el funcionamiento del programa se debe especificar la ruta donde se guardará el documento JSON y los documentos opcionales; por lo que es preferible crear una carpeta donde se guardarán estos documentos.

2. Manual de uso

El usuario debe seguir los siguientes pasos para iniciar el programa:

1) Introducir el siguiente comando en la consola:

   "python nombredelprograma.py -rutaPDF A -Destino B -Eleccion C -Opcionales D",

   donde los parámetros son:

   - A: ruta donde se encuentran los documentos PDF.
   - B: ruta donde se creará una carpeta donde se guardará el documento JSON y los documentos opcionales.
   - C: valor numérico que puede tomar los siguientes valores:
     - 1: si el usuario quiere elegir el documento PDF con el que trabajar.
     - 2: si el usuario quiere trabajar con todos los documentos PDF que se encuentran en la ruta proporcionada
   - D: valor numérico que puede tomar los siguientes valores acerca de los documentos opcionales:
     - 0: el usuario no quiere ningún documento opcional
     - 1: el usuario quiere un documento cuyo contenido es únicamente el texto que no proviene de imágenes.
     - 2: el usuario quiere un documento cuyo contenido es únicamente el texto que proviene de imágenes.
     - 3: el usuario quiere un documento cuyo contenido es el texto completo.
     - 4: el usuario quiere un documento cuyo contenido es el texto completo, etiquetando el texto que proviene de imágenes.
     - 5: el usuario quiere todos los documentos que se acaban de nombrar.

2) En el caso de que el usuario haya elegido en el parámetro *Eleccion* el valor 1, le aparecerá por pantalla una lista numerada de todos los documentos PDF que hay en la ruta proporcionada en el parámetro *rutaPDF*. De esta lista el usuario debe elegir uno de los documentos PDF escribiendo en la consola el número con el que aparece en la lista.

Después de estos dos pasos el usuario solo tiene que esperar a que termine el programa de trabajar con el documento PDF. Para finalizar el programa hay dos opciones:

- Si el usuario escribió en el parámetro *Eleccion* el valor 1, el programa finalizará cuando escriba en la consola un "0" en el momento que el programa le pregunte qué documento quiere seleccionar.

- Si el usuario escribió en el parámetro *Eleccion* el valor 2, el programa finalizará cuando se haya completado el trabajo con todos los documentos PDF de la ruta proporcionada en el parámetro *rutaPDF*.