**TALLER 1**

**EVELYN NATALIA GÓMEZ CALDERON**

**UNIVERSIDAD EL BOSQUE**

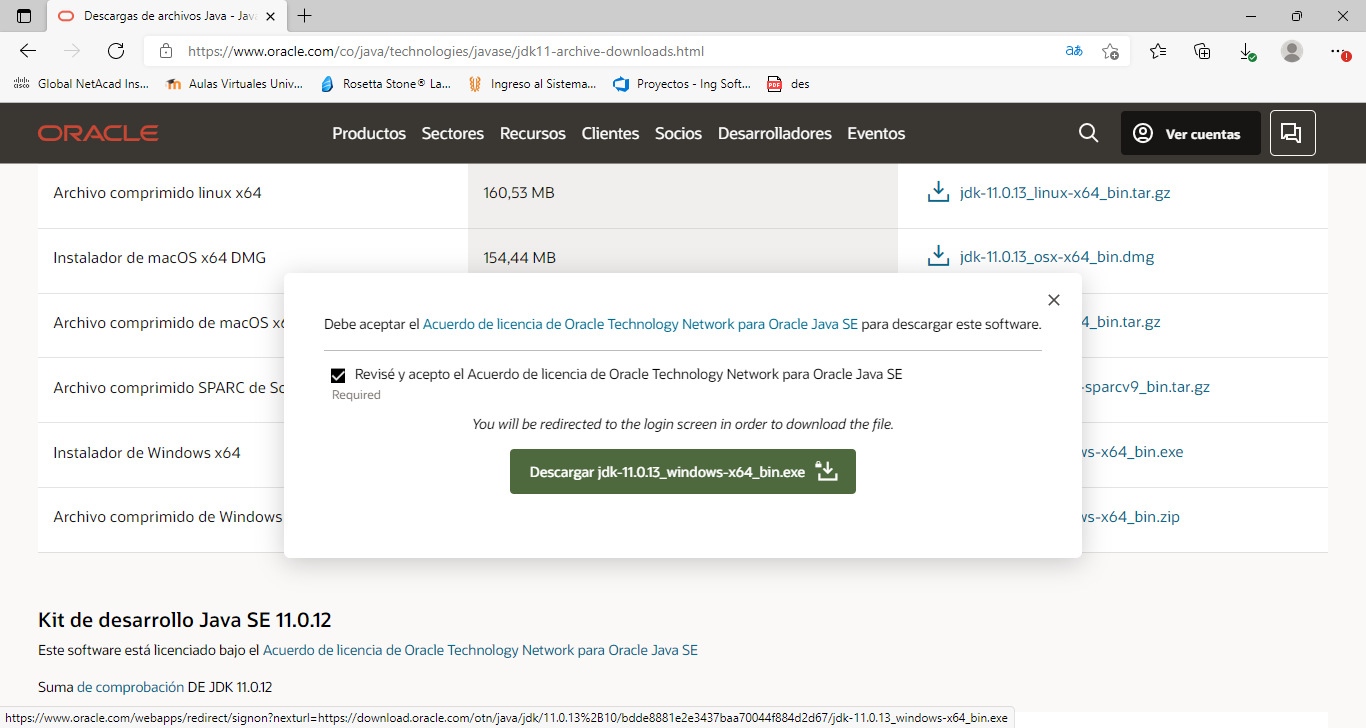
**BIG DATA ANALYTICS**

**BOGOTA D.C.**

**2022**

Descarga e instalación de PySpark.

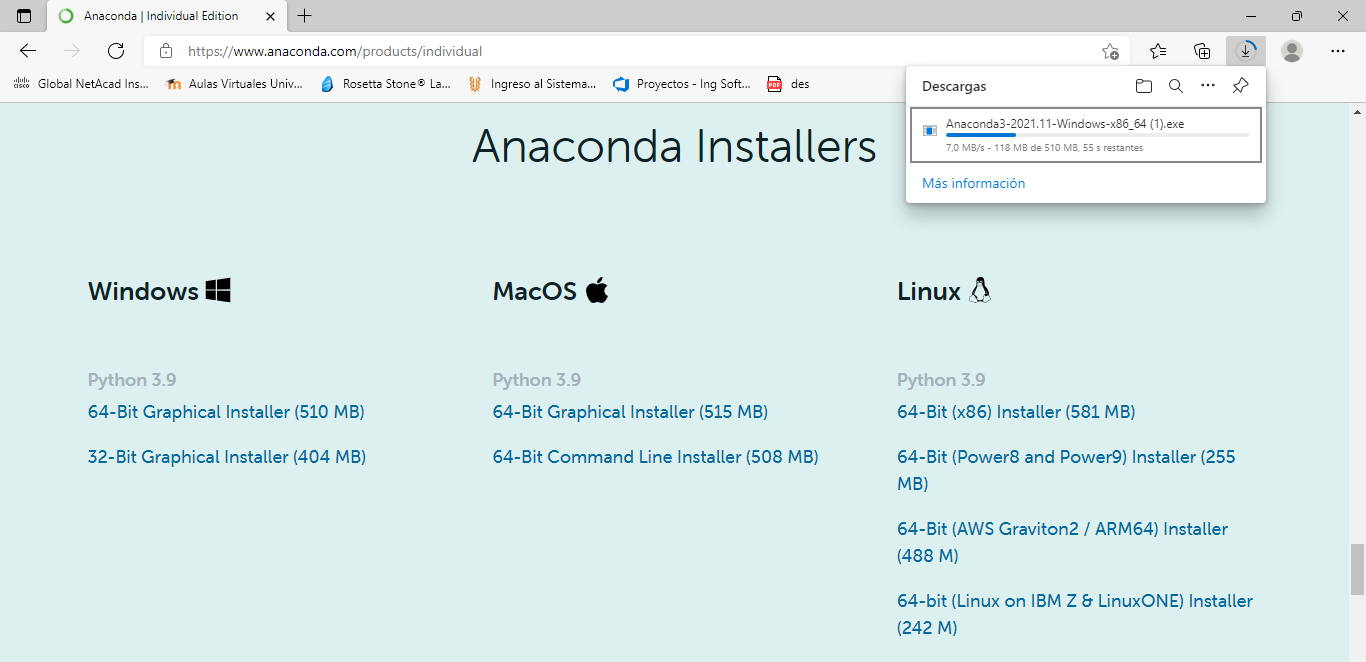
1. Descarga e instalación de Java JDK SE 8

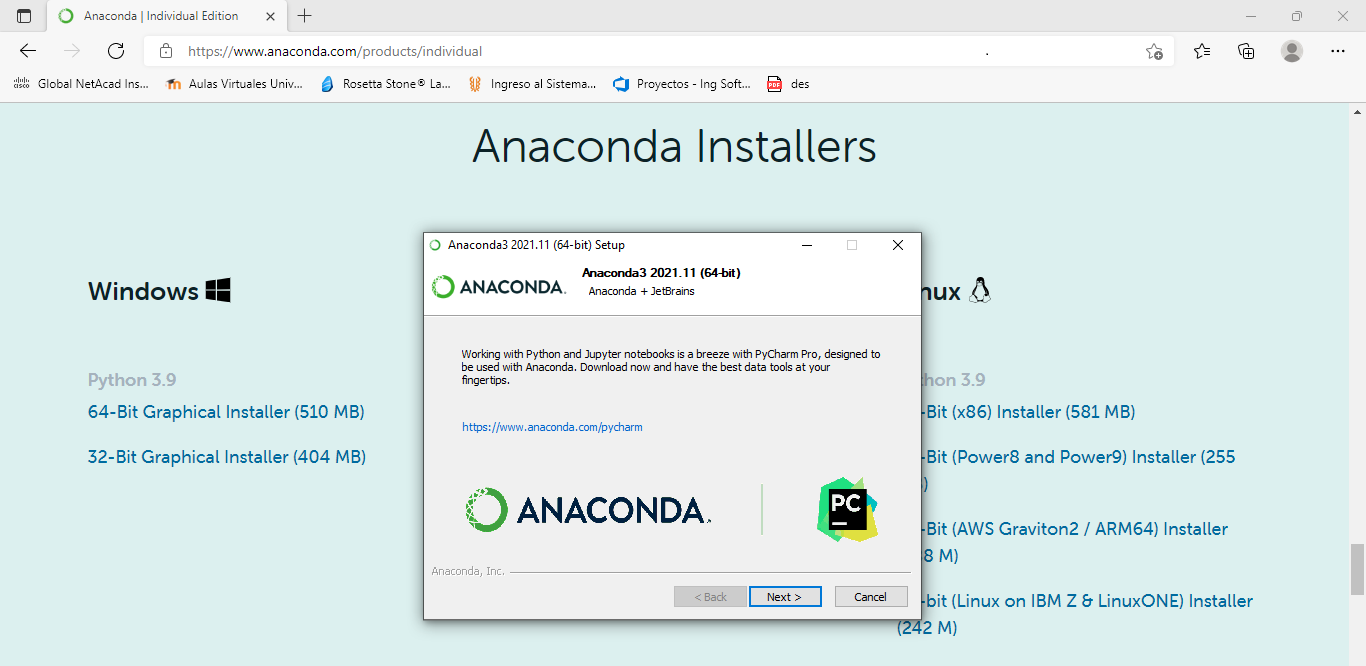


Texto

Descripción generada automáticamente

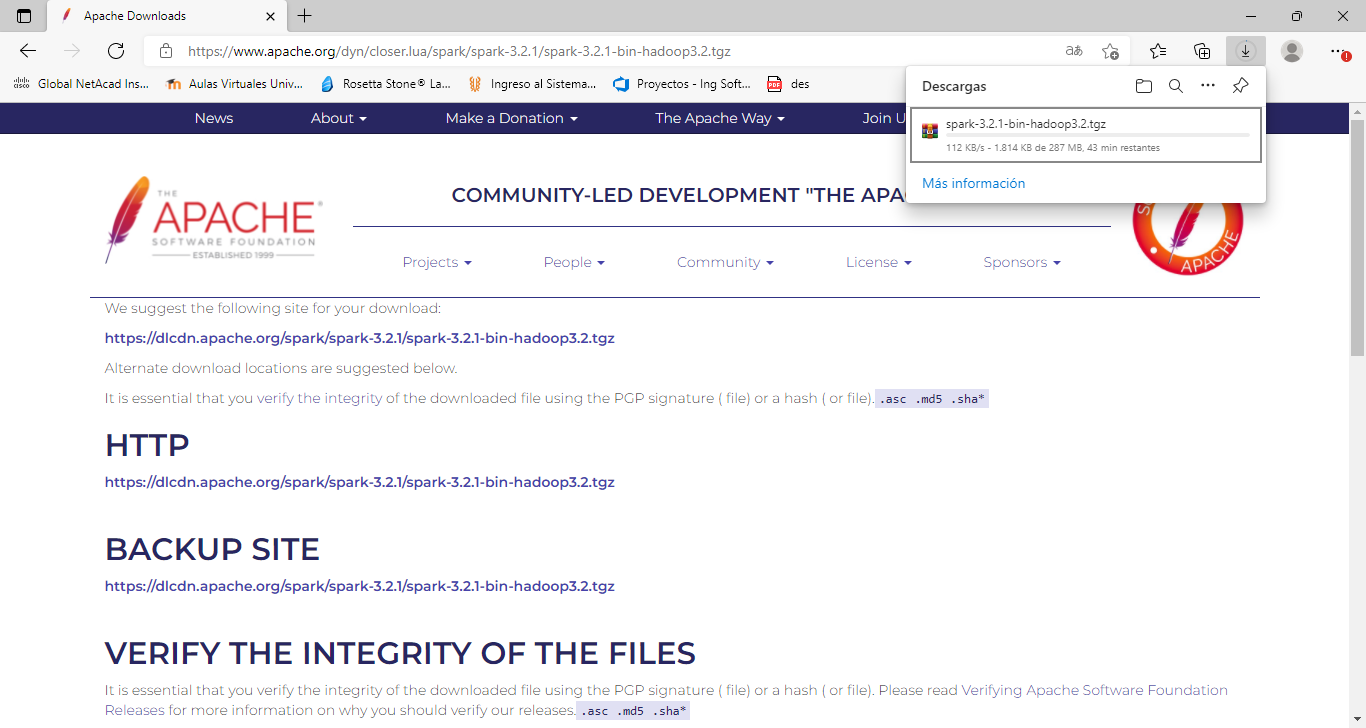
1. Descargar e instalar Anaconda.



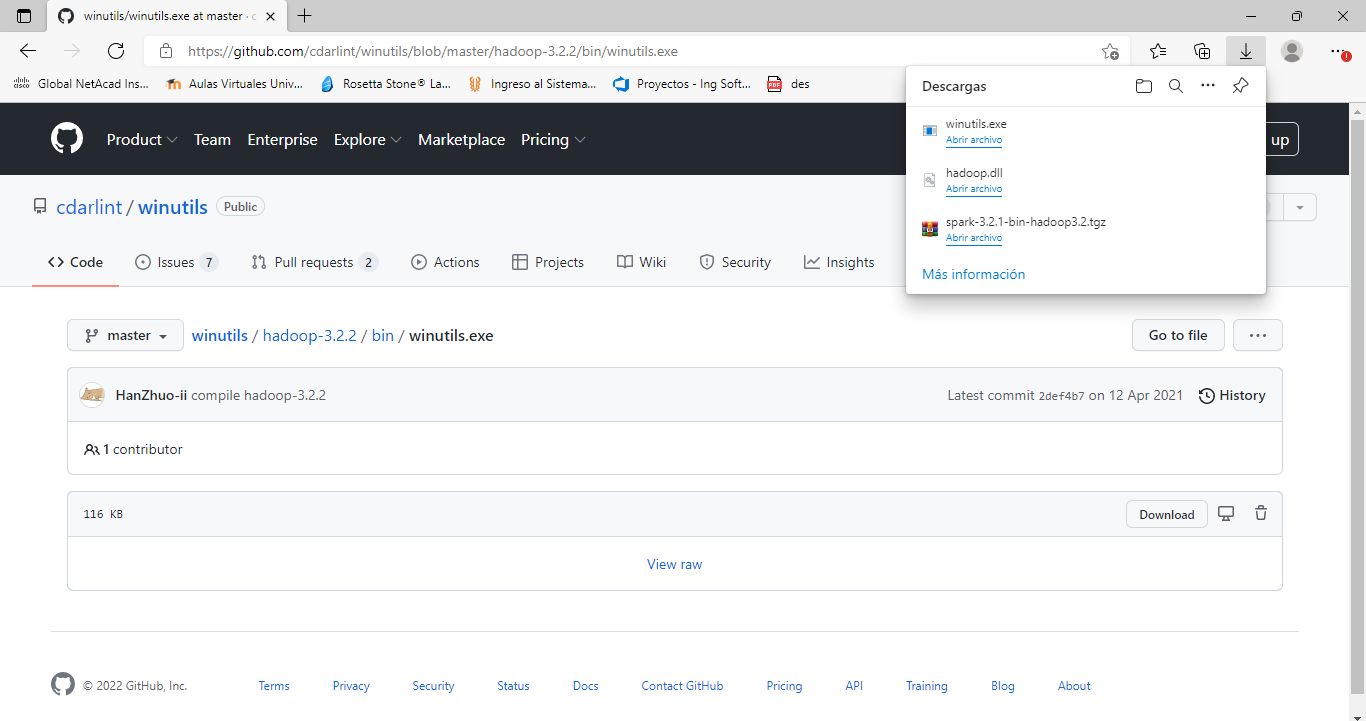


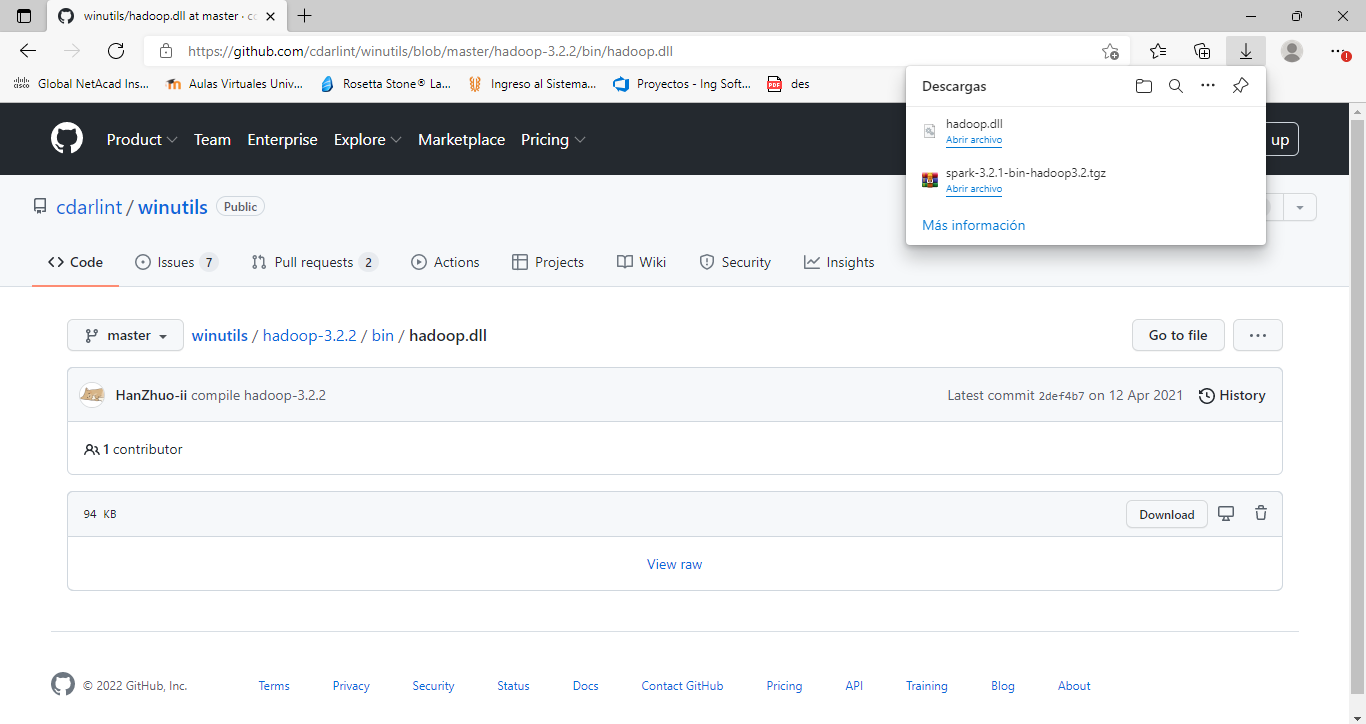


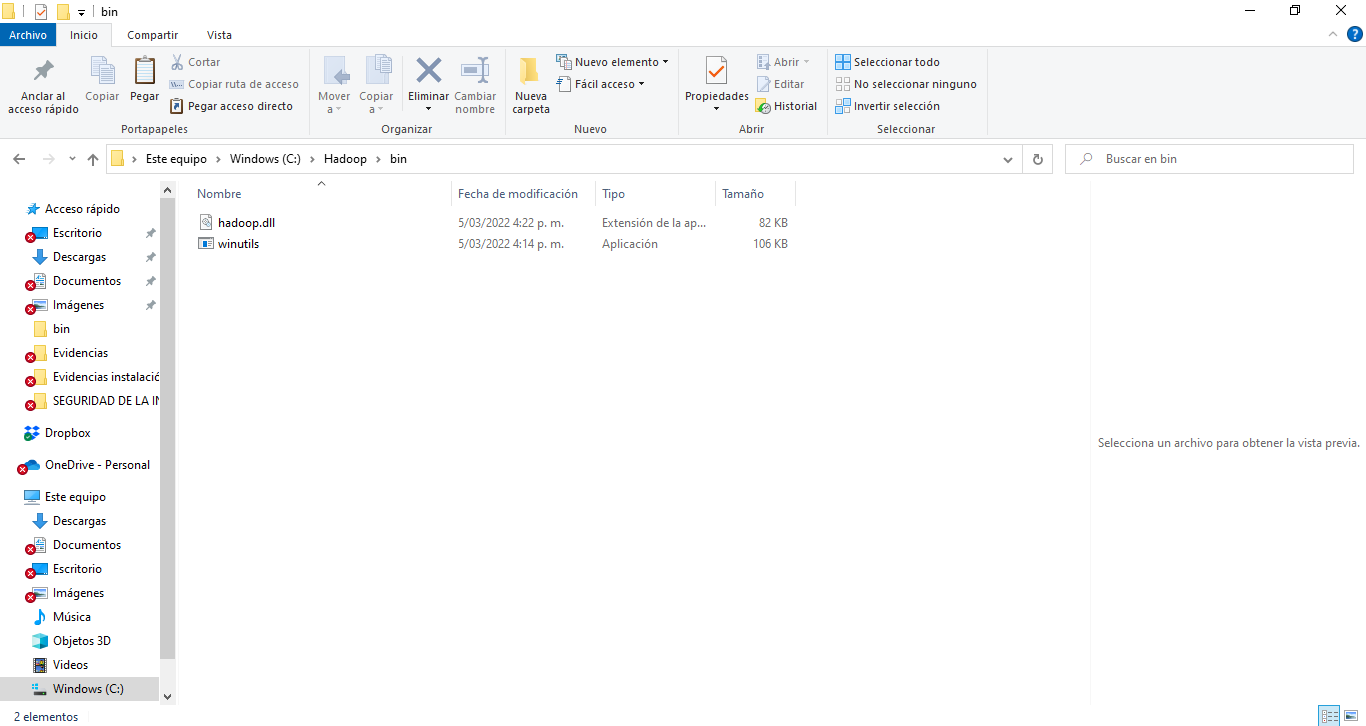
1. Descarga de Spark



1. Descarga de Winutils.exe y Hadoop.dll







1. Creación de variables de entorno

* JAVA\_HOME

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* HADOOP\_HOME

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* SPARK\_HOME

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* PYSPARK\_PYTHON

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* PYSPARK\_DRIVER\_PYTHON

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* PYSPARK\_DRIVER\_PYTHON\_OPTS

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

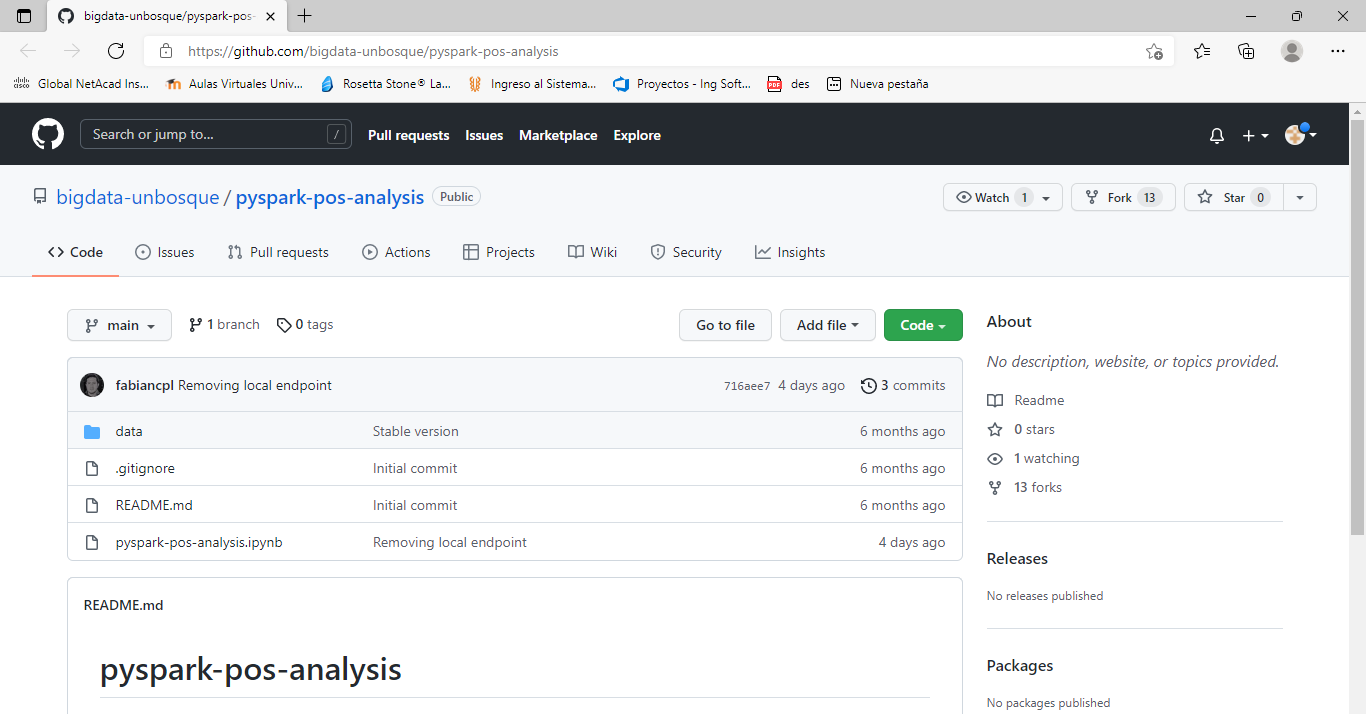
Descripción generada automáticamente

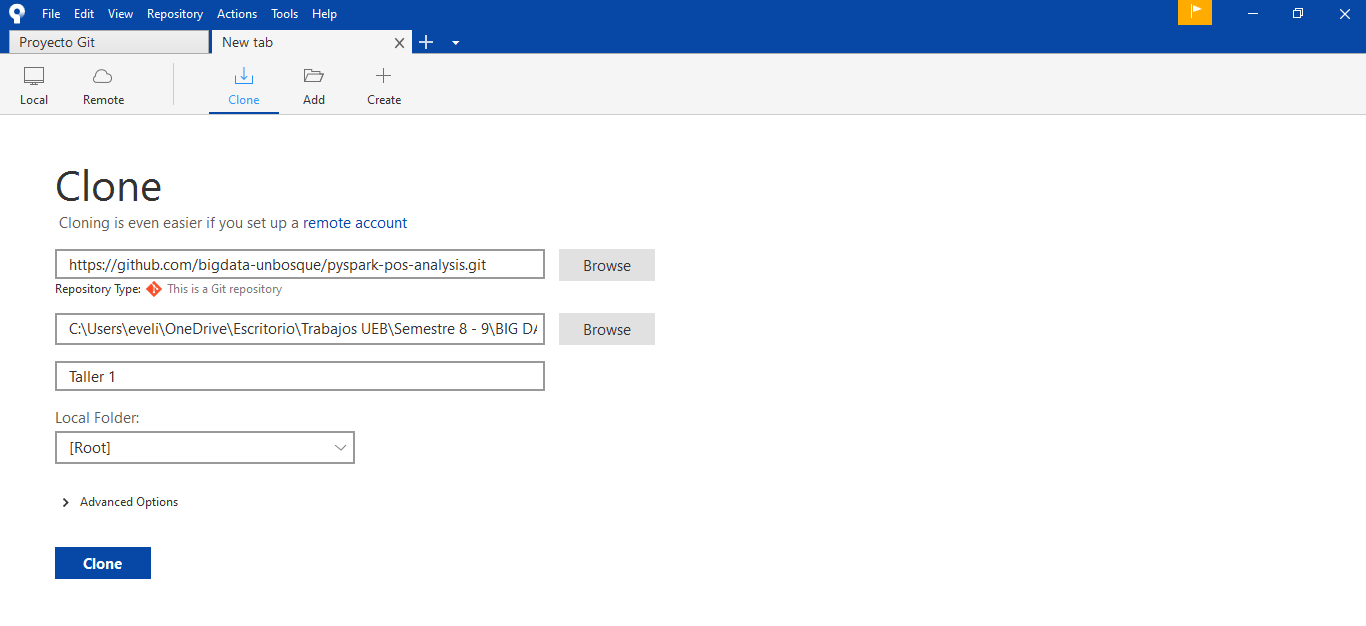
* Modificación variable de entorno Path

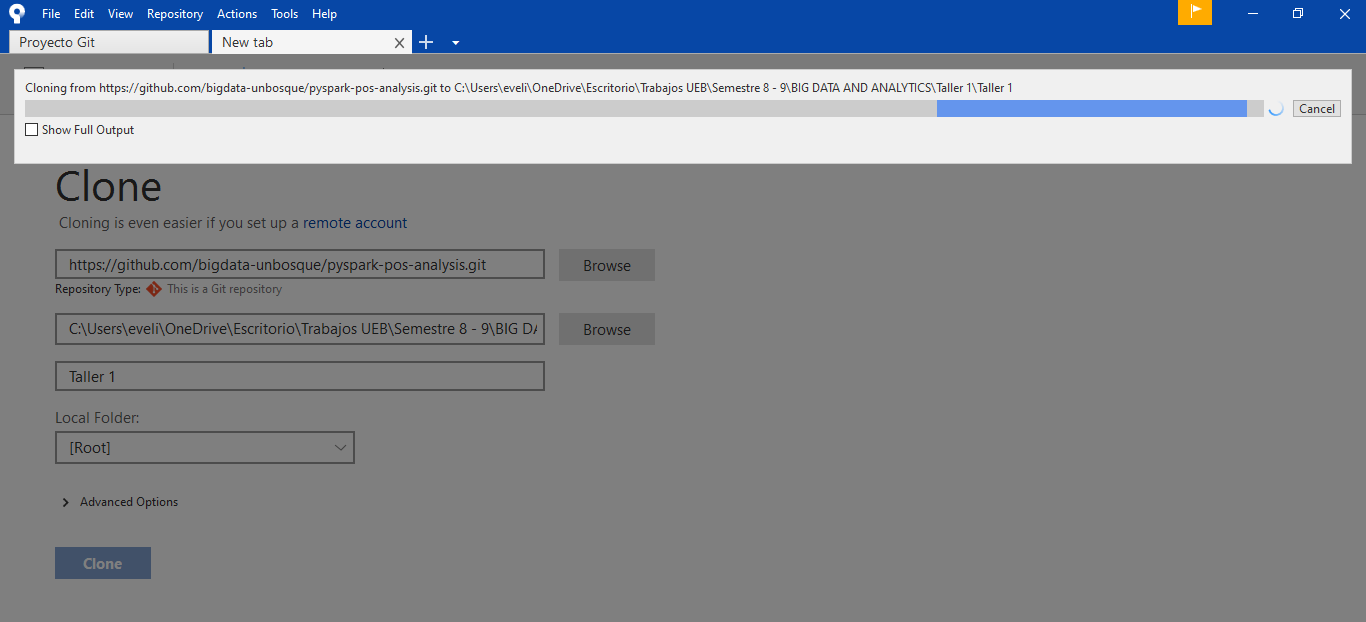
Interfaz de usuario gráfica

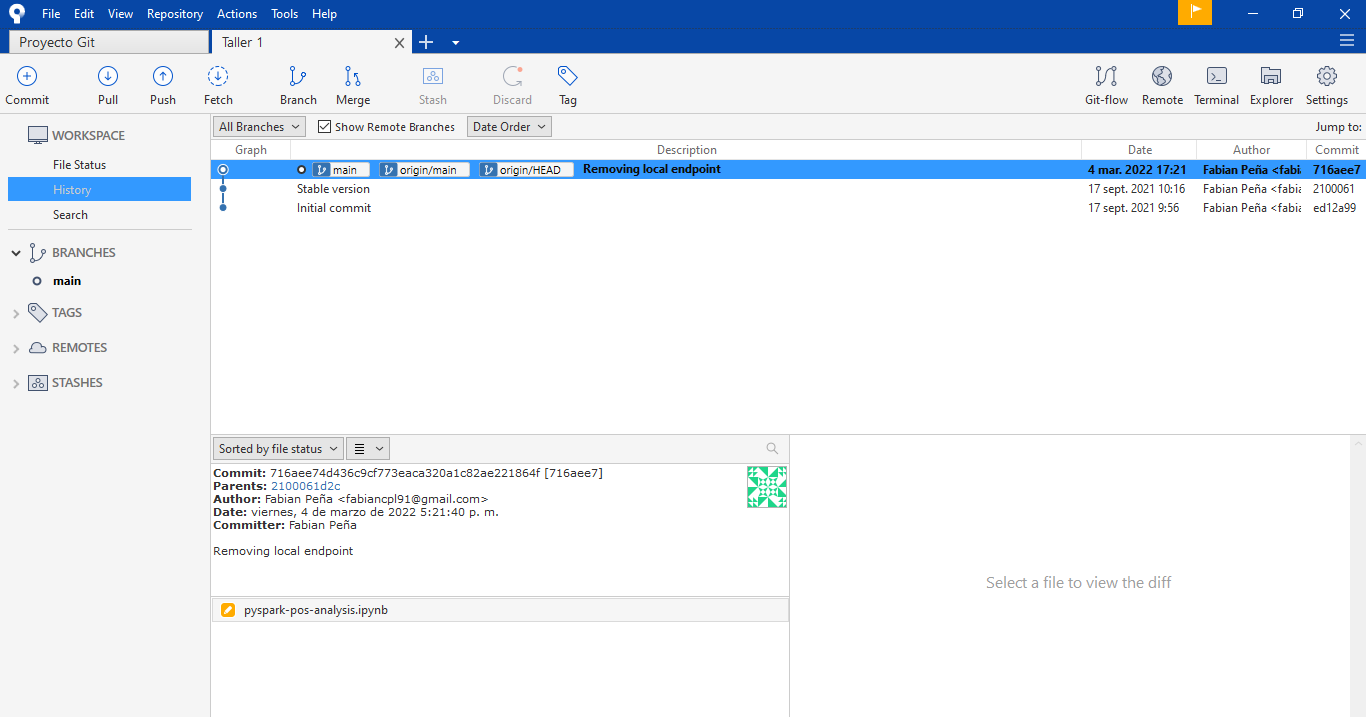
Descripción generada automáticamente

* Clonación repositorio





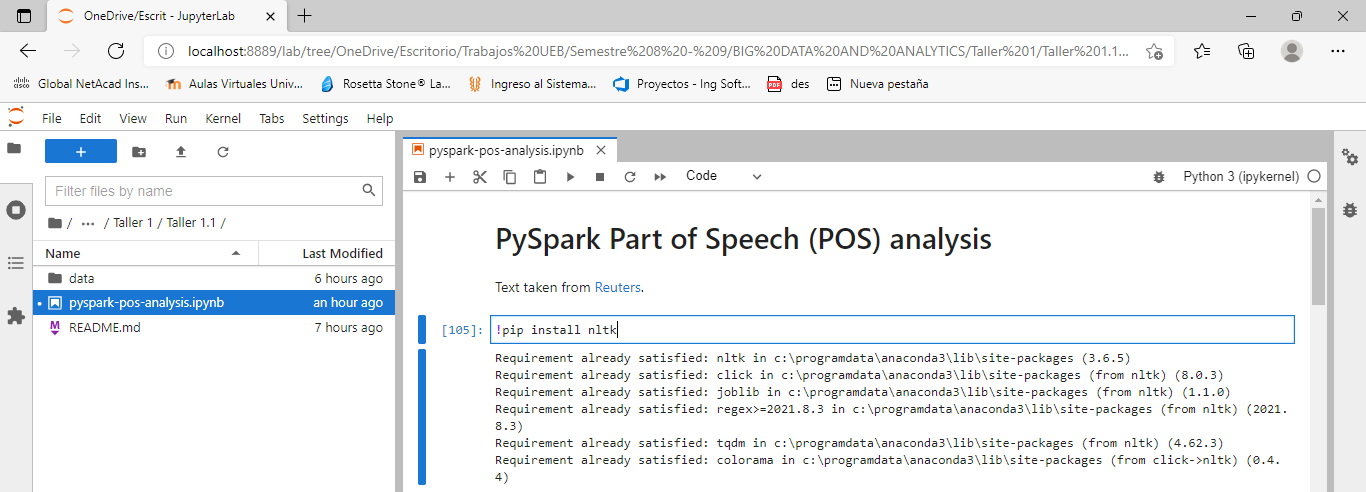




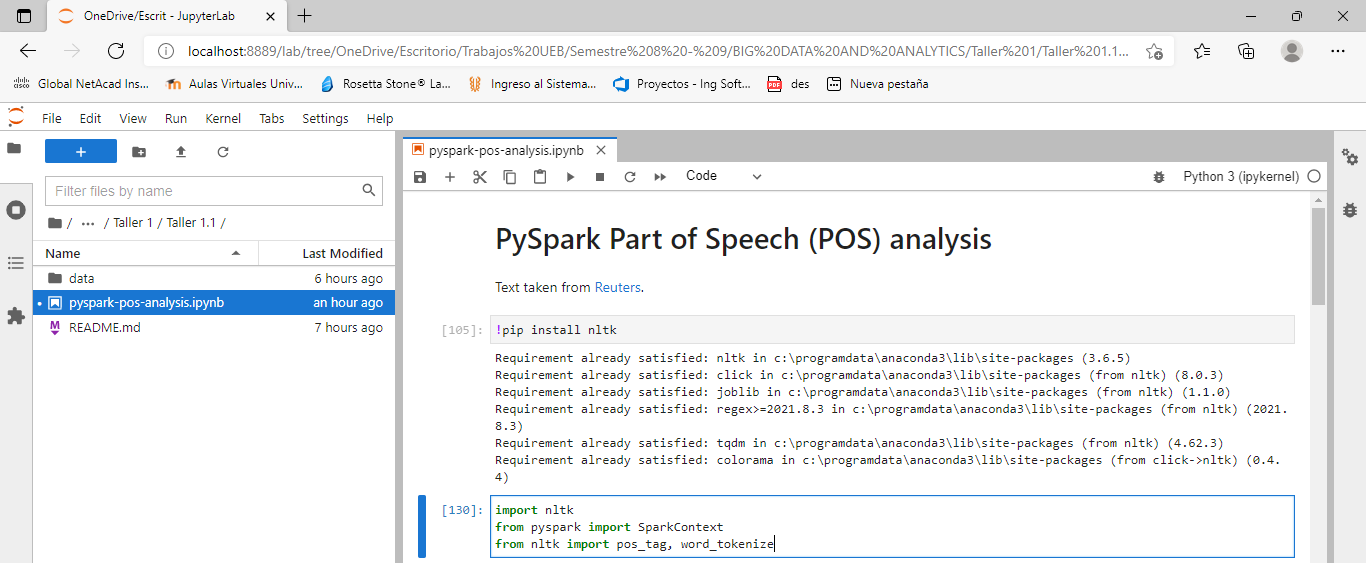
2. Realizar conteo de los diferentes elementos de Parto of Speech (POS) del articulo “Banks beware, Amazon and Walmart are cracking the code for finance.”.

Se realizo la clonación del repositorio de git el cual contiene el Script de base “pyspark-pos-analysis.ipynb” y se ejecutó desde Jupyter lab.

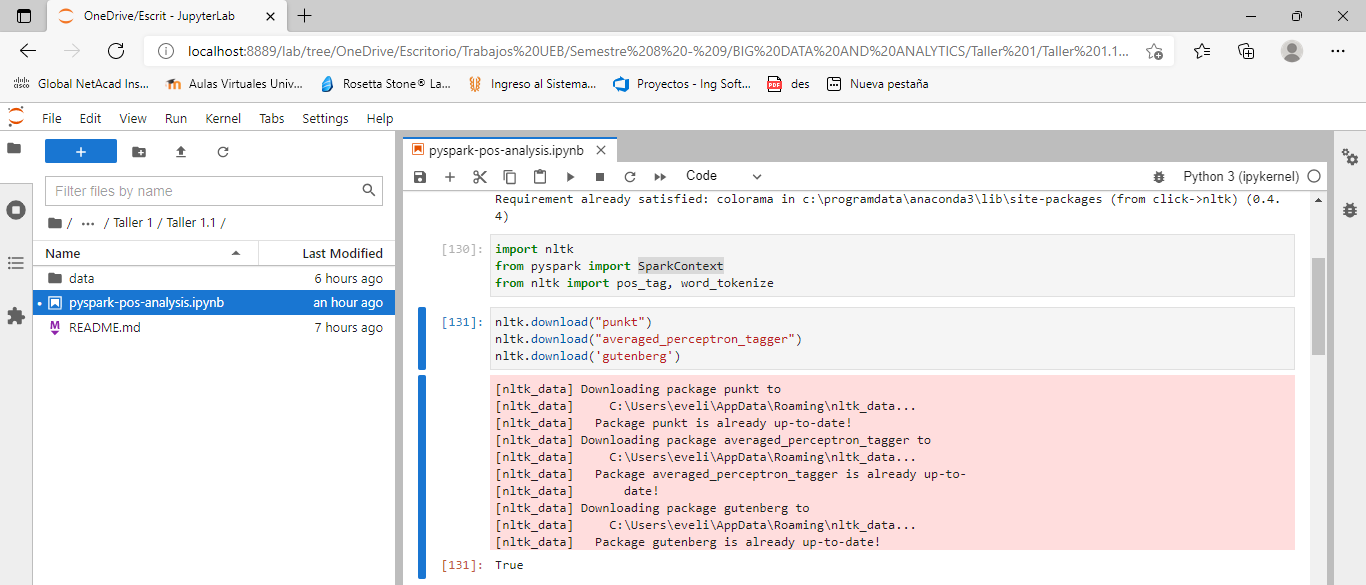
Se instalo el paquete de tokenizador nltk.



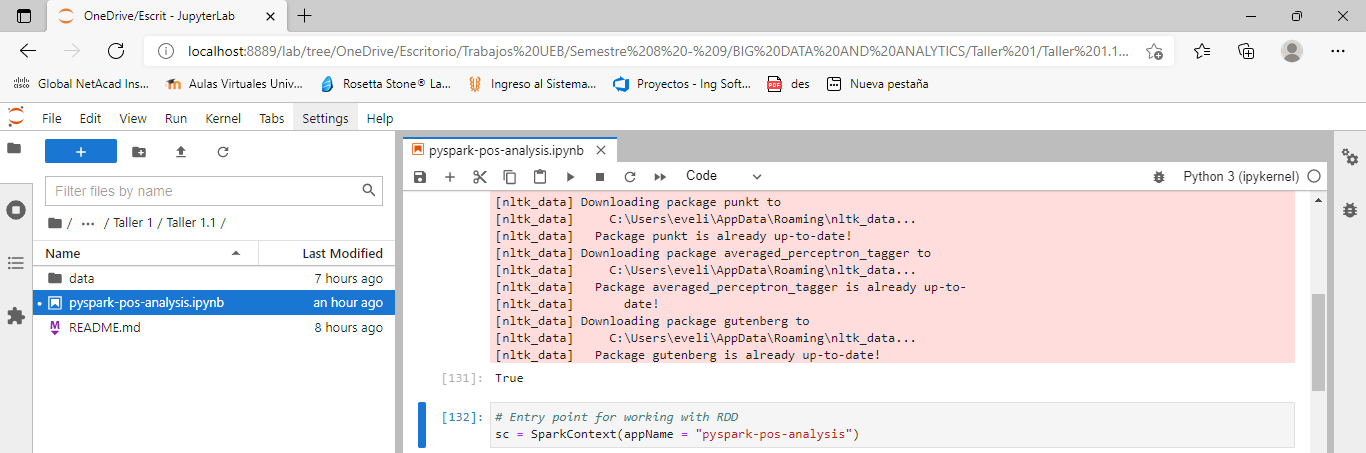
Se importo la librería nltk y se realiza la conexión al clúster de Spark.



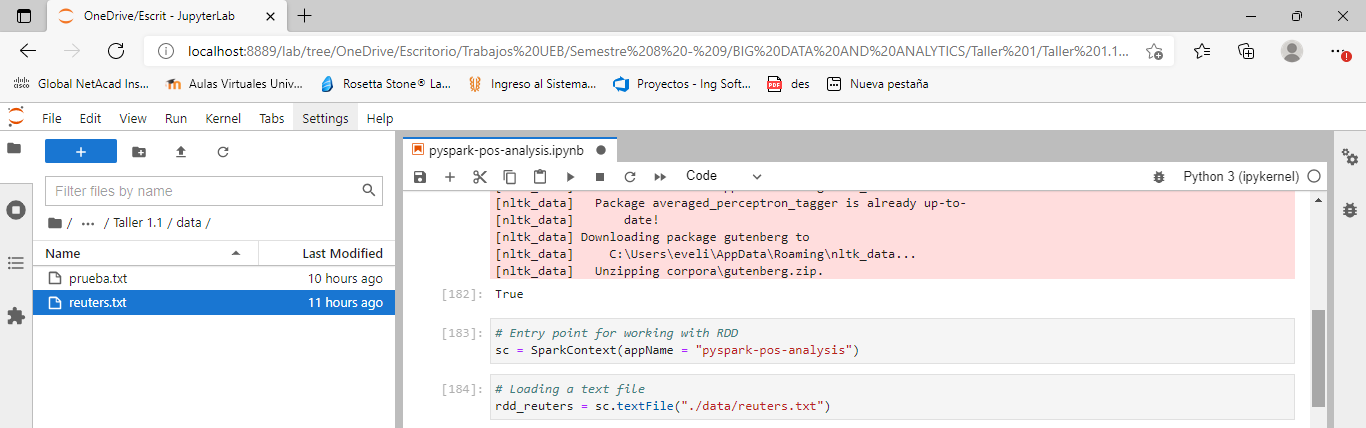
Se descargan los paquetes punkt, averaged\_perceptron\_tagger y Gutenberg.



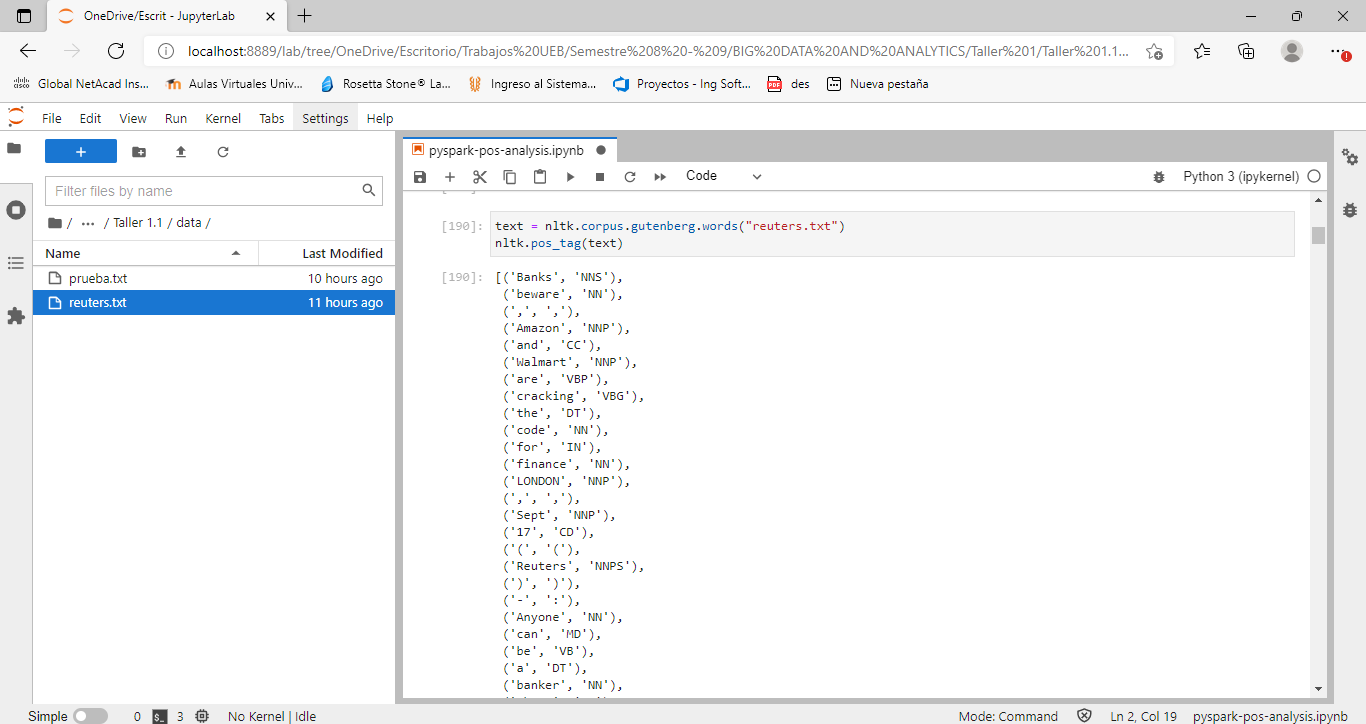
Se crea el punto de entrada para trabajar con RDD.



Se realiza el cargue del archivo “reuters.txt” con los datos.



Se realiza la clasificación de las palabras que se encuentran en el archivo en nombres, adjetivos, verbos y preposiciones.



Por último, se realizó el cargue de la información a GitHub.

