Presentación Taller Arranque

Andrés Leonardo Araque, Rafael Alexander Bermúdez, Andrés Mauricio Pérez
Universidad de los Andes, Bogotá, Colombia
{al.araque, ra.bermudez, am.perezv}@uniandes.edu.co
Fecha de presentación: febrero 3 de 2020

1 Introducción

El presente documento describe el desarrollo del procesamiento de las noticias de Reuters-21578, realizando conteos y frecuencia de palabras de archivos de noticias en las cuales el usuario tenga acceso a estas estadísticas a través de una aplicación de web.

2 Desarrollo de retos

Para dar solución a la serie de retos propuestos, se hace uso de expresiones regulares sobre UNIX, empleando una Maquina Virtual con sistema operativo Centos 6.10, allí se actualiza la maquina y se instalan las dependencias necesarias. Inicialmente se establece de forma conjunta el lenguaje y framework a utilizar, como lo es Python 3.8.5 y Flask 1.1.2 respectivamente, seguidamente se hace uso del dataset suministrado que contiene grandes cantidades de noticias en diferentes archivos, para su caso se eligieron los siguiente archivos reut2-000.sgm , reut2-001.sgm , reut2-003.sgm , reut2-004.sgm.A partir de ahí se definen dos parámetros de búsqueda para el usuario: el número de palabras más frecuentes del texto y el archivo al cual se desee procesar.

3 Desarrollo de aplicación web

Para la aplicación web se usó Python y el framework Flask. Se crearon rutas para cada uno de los retos y para la página inicial. En la página inicial se muestran los retos con cuadros de texto y listas desplegables para ingresar los parámetros. Una vez se ingresa la información y se hace click en enviar, se llaman las rutas correspondientes al reto escogido y se despliegan los datos.

4. Conclusiones

Para obtener información de diversas fuentes de datos, es fundamental conocer el tipo de archivos que se manejan, su formato y organización. Esto con el fin de diseñar estrategias de pre procesamiento que nos permitan extraer datos válidos, es decir datos que nos permitan inferir conocimiento importante para el estudio o investigación que estemos realizando. Es necesario profundizar en este aspecto para obtener mejores resultados.

4 Bibliografía

1. *flask.palletsprojects.com/en/1.1.x/.* [En línea] [Citado el: 2 de Febrero de 2021.] https://flask.palletsprojects.com/en/1.1.x/.

2. García, Salvador & Ramírez-Gallego, Sergio & Luengo, Julián & Benítez, José & Herrera, Francisco. (2016). Big data preprocessing: methods and prospects. Big Data Analytics. 1. 10.1186/s41044-016-0014-0. [Citado el: 2 de Febrero de 2021.]