Marco de Referencia

Dentro de las investigaciones hechas en este proyecto, se generaron unas de gran valor, por el contenido de enriquecimiento que estas han aportado a los participantes de estas, inicialmente, debemos expresar que este fue producido desde Databricks, que es una plataforma web en la cual se puede trabajar en Spark, este provee desde la misma web la creación de clusters, y la carga de datos, y claramente, la posibilidad de creación y manipulación de notebooks, que es justamente donde se enfoca nuestro trabajo para este proyecto, dentro de los grandes objetivos cabe señalar 3, la ingesta de datos, la limpieza de los mismos y el análisis de data por medio de algunas técnicas propuestas.

Dentro de las etapas iniciales, claramente se encontraba la ingesta de los datos, estos ya previamente cargados a la web, se debían llevar al notebook, claramente cada comando debía llevar un conocimiento de como h2acerlo basados en PySpark, que es la API de Python con la cual este lenguaje soporta Spark, esto de forma obvia nos llevó a la documentación de Apache Spark, acudiendo en este caso a los módulos de **pyspark.sql**, topándonos entonces con la utilización de queries con algo de similitud con SQL, con el contenido de lo queries notamos que íbamos a interactuar con una gran cantidad de funcionalidades que PySpark nos podía brindar, ya dentro de las investigaciones para la ingesta de los datos, se generó satisfactoriamente, escogiendo en el caso de este proyecto, un solo archivo de noticias.

Seguido de esto, el tener la ingesta y los datos, poder mostrarlos, manejar y ver las columnas que tenía este archivo, no era realmente lo más retador, pero terminó siendo un punto más de estudio, para el manejo óptimo de estos, pero ya se entraba a un tema donde nos íbamos con una parte fundamental para e estudio de este archivo, el estudio que por obvias razones queríamos aplicar para cumplir el reto de Big Data, y era el tema de la limpieza de datos, esto es realmente fundamental ya que dentro de un archivo nos podemos encontrar con palabras o símbolos que nos van a impedir el posterior estudio o también un estudio poco efectivo del texto en sí, llegamos encontrándonos con los “Stopwords”, estos caracteres son los que se debían eliminar, primeramente debimos estudiar el uso debido de la tokenización, cómo dentro de Spark podemos manejar un texto, llevarlo a un array y separar cada palabra como si fuera un token, llevando claramente a tener que limpiar esto, ya que hay palabras como verbos comunes, y los típicos caracteres para la coherencia y debida construcción de un texto que para el ser humano deben estar contenidos, como lo son las “,”, “;”, “:”, “.”, etc, que claramente si van a estar contenidas en

ciertas palabras, con lo que el análisis no los tomaría como los mismos datos, si no distintos, para dejar un ejemplo pequeño podríamos dejar una palabra simple “horrible” y “horrible,”, nuestro análisis de datos va a identificar estas dos como un par de tokens distintos, generando entonces pérdida de data que podría llegar a influenciar en un estudio adecuado, es por eso que llegamos al punto de remover estos “Stopwords”, para la creación de data completamente limpia, claramente dentro de nuestra investigación nos topamos con librerías que nos brindaban un apoyo directo y muy eficaz.

A partir de una limpieza de los tokens previamente hecha, se inicia ahora con un proceso de analítica, en el cuál hay una gran cantidad para aplicar y todo depende de la necesidad que se tenga, pero también nos dimos a la tarea, bajo opinión personal, más compleja dentro de la realización de todo el proyecto, debido a que en este punto ya encontramos una gran influencia en Python en absolutamente todos los métodos de analítica, ya que su composición ya es realmente más compleja, comprometiendo entonces librerías de Python directamente pero que trabajan de forma óptima con Spark, de esta forma componemos las funciones necesarias para llegar al objetivo de la analítica de texto.

https://en.wikipedia.org/wiki/Databrickshttps://en.wikipedia.org/wiki/Databricks

https://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html

https://www.gangboard.com/blog/what-is-pyspark/

https://openwebinars.net/blog/que-es-apache-spark/

https://books.google.es/books?hl=es&lr=&id=HVQoDwAAQBAJ&oi=fnd&pg=PP1&dq=pyspark&ots=tLMoIpKkaP&sig=aG-Cx3yEtmnIIbWH6gLV1DiOcLU#v=onepage&q&f=false