***Reporte técnico y Reporte de Avance, fases, por:***

***Carlos Mario Blanco Pérez***

***Santiago Vásquez***

***ETL****:* En el proceso de extracción de los datos se procederá a descargar el *dataset* de la biblioteca de Gubenberg el cual incluye una colección de documentos tanto en inglés como en español. (Dado que el *dataset* se encuentra en la carpeta raíz su avance es del 100%)

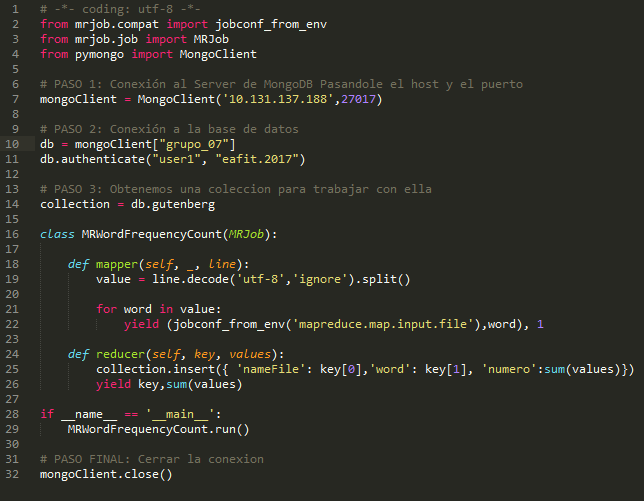
En el proceso de transformación se llevaran todos los elementos que integran el *dataset* a un solo formato .txt (dado que los documentos dentro del *dataset* tienen todos una extensión .txt su avance es del 100%)

En el proceso de carga se llevara el *dataset* a HDFS (Hadoop Distributed File System, cuyo entorno ya está configurado por defecto con un master en la dirección 10.131.137.188 y sus respectivos slaves en los cuales estará distribuida la carga), avance 100%

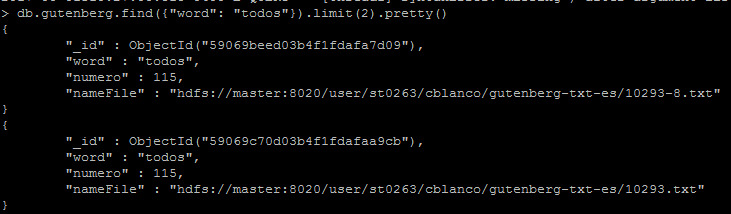
***Procesamiento:*** El procesamiento se realizara bajo una arquitectura de datos en *batch* donde se integrara al lenguaje de programación *Python* junto con una librería *mrjob* (*map/reduce)* para el manejo y procesamiento del *dataset* bajo un *índice invertido* para clasificar, ordenar y consultas posteriores. Avance 100%

***Aplicación:*** Se desarrollara una aplicación en *Ruby on Rails (RoR),* la cual será la que tendrá interacción directa con el usuario que en forma de búsqueda el agregara en un *text área* una serie de palabras las cuales retornaran como resultado el documento(s) .txt los cuales tengan la mayor ocurrencia de las palabras ingresadas por el usuario. Avance 100%.

***Algoritmo Map/reduce, programado en Python***

******

***La estructura y outputs en la base de datos Mongodb son:***

******

***Comandos usados***

*python doc\_mapreduce.py -r hadoop hdfs:///user/st0263/cblanco/gutenberg-txt-es/\**

*// Correr el algoritmo Map/reduce pasandole como parametros todos los textos planos dentro de la carpeta Gutenberg-txt-es*

*mongo grupo\_07 -u user1 –p*

*// Conectarse a la base de datos MongoDB*

*hadoop fs -ls "hdfs://master:8020/user/st0263/cblanco/"*

*//Inspectionar*

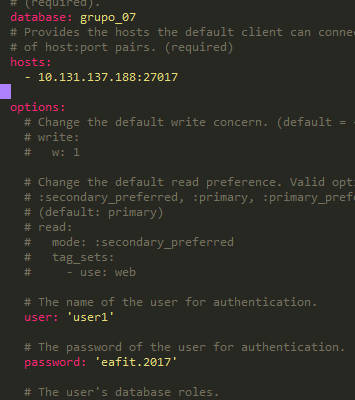
*hadoop fs -put /home/cblanco/gutenberg-txt-es /user/st0263/cblanco*

*//agregar*

*hadoop fs -cat "hdfs://master:8020/user/st0263/cblanco/gutenberg-txt-es/10293.txt"*

*//Mostrar en pantalla el contenido de “10293.txt”*

***Configuración – MongoDB – Ruby on Rails – mongoid.yml***

**