Team Mount Antero

Unstructured Data

Team Project

4/25/2020

Project Design

Major Components:

* S3 storage bucket
* Pyspak Jupyter notebooks
  + Libraries: String, pyspark, pyspark.sql (SparkSession), pyspark.sql.functions, pyspark.sql.types
* AWS EMR cluster

Component Interfaces:

Basic structure: The dataset will be stored in the S3 bucket in a CSV format and called into pyspark via a URL. Once the dataset has been loaded into pyspark it will be converted into a pyspark DataFrame for data manipulation.

The pyspark code will be written in jupyter notebooks and deployed to the Leeds EMR cluster to be executed.

Model Descriptions:

· Article word count: The article word count function is a user defined function in pyspark that was used to find the number of unique words in each article. This was then used to generate a new column in the data set that displays the number of unique words in each article. The culmination of this function was to create a histogram displaying the distribution of the number of unique words used in covid-19 articles. The result was as expected with a very right skewed distribution showing that the majority of articles tend to use fewer unique words and very few using many unique words.

o The function to calculate the number of unique words in an article is specified by the following line of code: *sum([i.strip(string.punctuation).isalpha() for i in wordString.split()]) \*requires the string library to be imported*

§ The function simply takes a string variable as an input, splits and strips whitespace, and sums the alphabetical letter groups.

Furthermore, in order to count the specific unique words and determine the main sentiment of the media, the Leeds School of Business EMR cluster is used to run the python code and take care of the data. A definition function is used to set the rule of word counts. The useless prepositions and verbs will be removed before the results come out. The results will show only the top 10 hot words the media is using for now.

· The DateTime\_Cleaning code leverage the col, hour, minute, and second functions from pyspark.sql to convert the timestamp column to separate columns of time parameters. Using the df.withColumn() function, date was extracted from the timestamp in YYYY-MM-DD format. Within the DateTime\_Cleaning notebook, the data frame is filtered to exclude null hour values as well as blank authors.