STA 9760

Juna Iafelice

Midterm Project /Spring 2020

Summary

YELP REVIEWS

The question is: **What are the top ten most used words grouped by user star rating reviews?** The python code that resolves this question is attached on the Zip folder. After a various try and failed attempts, I finally was able to answer the question. The problems that I faced are:

1. The review dataset was too big (6.3GB) and the business dataset was around 150MB. Every time I calculated for the review dataset, I faced an error of memory out of bound of JVM. What I did is I randomly split the reviews dataset and took only 5 % of it around (300MB) with the random split method provided by the Pyspark data frame. Another issue I faced was saving the randomly split datasets because winutil.exe library was missing from Hadoop bin folder which it took me a long time to figure it out.
2. As a new python developer, I faced many challenges as using lambdas, filtering, and mapping. The problem I wanted to solve was to concatenate each text review into a big text. I managed to do it with the flatMap function.
3. I run the code and observed that the top words were almost the same for all the stars like (the, off, to, and, an, is etc). What I did is I created a file with the 100 most used words in English that I found online. I filtered my results to not contain these most used words.
4. I also found the total numbers of business by each star rating.

Please find attached the screenshot with the question answer, tested with 300MB review data:

