SI 618 Fall 2021 Homework 5 (100 points)

Data to be used in this homework: On the Hadoop cluster, I have put the following file in HDFS:

hdfs:///var/umsi618f21/hw5/yelp_academic_dataset_business.json

This file was downloaded from the Kaggle Yelp dataset.

Note that you do not need to download the Yelp dataset yourself as it is already put into HDFS on the Hadoop cluster.

Business Statistics by City

The goal is to compute the number of businesses, average rating, number of businesses with wheelchair access and number of businesses with some form of parking (garage, lot or street) for each category in each city considering only businesses that have at least one review. If a business has multiple categories, its review count and rating should be attributed to all of the categories. If the category list is empty, then we will use 'Unknown' as the name of the category.

Your final result should be a TSV file that is the same as the provided si618_hw5_desired_output.tsv file (differences in ordering due to ties are acceptable)

In this desired output file, each row contains 6 columns, which are separated by a tab. For example, consider this following row:

Atlanta	Event Planning & Services	944	3.643008475	241	344

This means the category of “Event Planning & Services” in the city of “Atlanta” has 944 businesses, their average rating is 3.643, 241 of them have wheelchair access and 344 have some form of parking.

The rows in the output file should be sorted in alphabetical order of the city names, and the categories in each city are sorted by the number of businesses in decreasing order. Note that there are some cities with names that may be data entry errors (such as “51 Richard Beall Hwy 17-92”) as well as differently capitalized names for the same city (e.g.: ATLANTA, atlanta, Atlanta); to match the desired output you should leave these cities in your output and not perform any filtering or formatting to remove cities with strange/incorrect names or letter casing.

You MUST use Spark to do this homework. A non-Spark solution will not get any credit.

HINT: You can modify from the provided example code spark_total_reviews_per_category.py. Save it as “youruniquename_si618_hw5.py”

What to submit:

Submit a zip file named si618_hw5_youruniqname.zip containing:

- Your Python source code: youruniquename_si618_hw5.py
- The tsv outpout file: youruniquename_si618_hw5_output.tsv