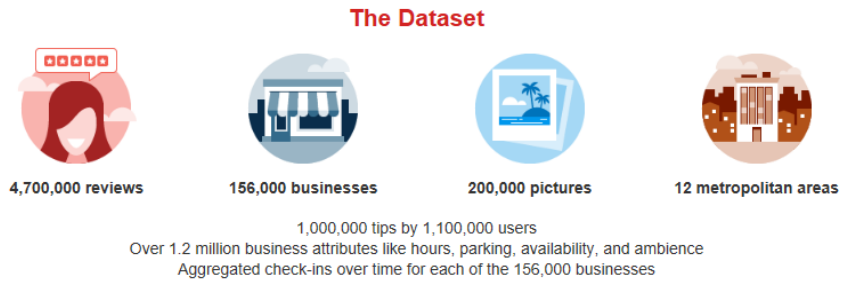
CS9223 Programming for Big Data – Assignment 2

Due Date: December 11, **11PM EST**

You ****MUST** use Spark** (scala, java, python or R bindings) . Extra credit for visualization, and for solving the assignment in Hadoop (Map/Reduce Java or Python, or Pig) to analyze the Yelp 2017 data challenge: https://www.yelp.com/dataset

- You must use the JSON dataset. The purpose of the homework is to parse and analyze JSON data.



Specifically, you ***must*** provide answers **and code** to the 5 following questions:

1. The average # of reviews and the average # of stars grouped by city and business category.

2. Pivot the business categories as columns, and show the average # stars for each category, by (city,state):

e.g.	City	State	Mexican	Chinese	Cigars	Restaurants	…..
	Toronto	CA	4.3	2.3	5	3.4	
	Tucson	AZ	5	3	1.2	4.1	
	etc						

3. What is the average rank (# stars) for businesses that are 'Mexican' category, AND offer takeout: (e.g. "attributes": {"RestaurantsTakeOut": true,…})

4. For businesses within 15km of Toronto center, show the average # stars aand average # reviews by type of business category

Center: Toronto, CA
Latitude: 43.6532° N, 79.3832° W
The bounding circle for this problem is a ~15 km radius. A business falls in the region if it's coordinates are within the circle.

The shortest distance (the geodesic) between two given points $P_1=(lat_1, lon_1)$ and $P_2=(lat_2, lon_2)$ on the surface of a sphere with radius R is the great circle distance. It can be calculated using the formula:

$$dist = \arccos(\sin(lat_1) \cdot \sin(lat_2) + \cos(lat_1) \cdot \cos(lat_2) \cdot \cos(lon_1 - lon_2)) \cdot R \quad (1)$$

For example, the distance between the Statue of Liberty at (40.6892°, -74.0444°)=(0.7102 rad, -1.2923 rad) and the Eiffel Tower at (48.8583°, 2.2945°)=(0.8527 rad, 0.0400 rad) – assuming a spherical approximation[a] of the figure of the Earth with radius R=6371 km – is:

$$dist = \arccos(\sin(0.7102) \cdot \sin(0.8527) + \cos(0.7102) \cdot \cos(0.8527) \cdot \cos(-1.2923 - 0.0400)) \cdot 6371 \text{ km}$$
$$= 5837 \text{ km} \quad (2)$$

Radians = Degrees * PI / 180, Degrees = Radians * 180 / PI

5. For the top 10 and bottom 10 food businesses near Toronto (ranked by stars), summarize star rating for reviews in January through May.

Grading (total 150 points)

This assignment **MUST** be completed on your own. Duplicate assignments will be flagged and failed.

- 25 points each question (1-5) = 125 points
- 15 points for the submission report and presentation quality
- 10 points for code quality

Extra Points 1 (125 extra points)

1. 125 Points: Complete the assignment in Apache Hadoop (MR or Pig) and review the difference in approaches (**you must still complete the original exercise**).

Extra Points 2 (30 extra points)

2. 10 points: provide suitable statistical analysis of your results with R.
3. 20 points: provide visualizations for results (distributions, graphs, maps, in R).

Submission:

In a single zip package, submit:
- report, max 10 pages.
- runnable code for all questions, clearly labeled (no dataset).
- results data for each question.

Hints/References

Apache Spark: http://spark.apache.org/
Pig JSON loader: https://pig.apache.org/docs/r0.10.0/func.html#jsonloadstore
Pig Latin: http://infolab.stanford.edu/~olston/publications/sigmod08.pdf
R maps – leaflet: https://rstudio.github.io/leaflet/
Yelp data samples: https://github.com/Yelp/dataset-examples