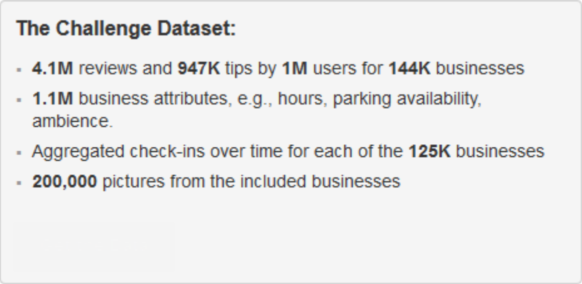
CS9223 Programming for Big Data – Assignment 2

Due Date: March 22th **11PM EST**

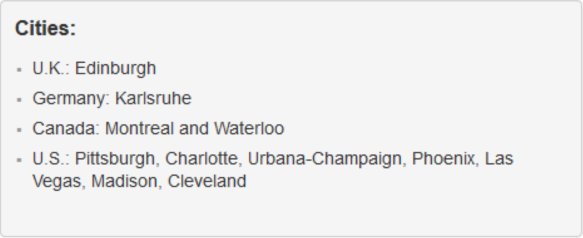
Details

You **must use** Hadoop Map/Reduce (Java or Python) or Pig, (Spark as extra credit) to analyze the Yelp data challenge: https://www.yelp.com/dataset_challenge.



The Challenge Dataset:

- **4.1M** reviews and **947K** tips by **1M** users for **144K** businesses
- **1.1M** business attributes, e.g., hours, parking availability, ambience.
- Aggregated check-ins over time for each of the **125K** businesses
- **200,000** pictures from the included businesses



Cities:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland

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

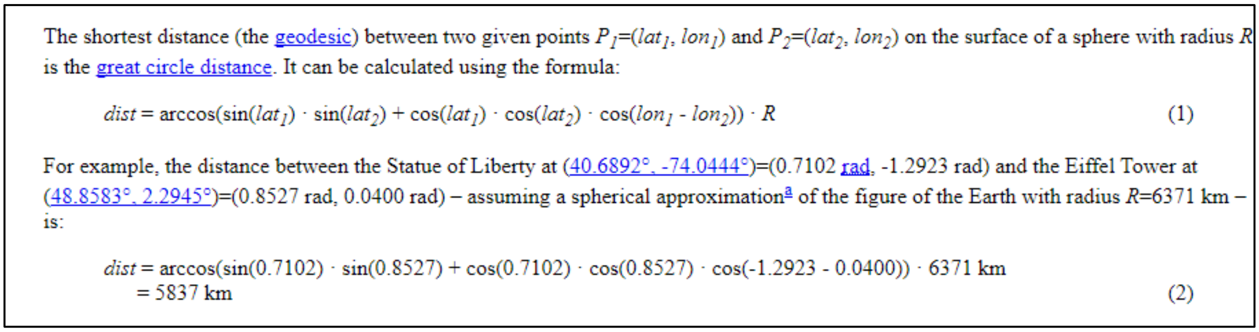
1. Summarize the number of unique ***reviewers*** by US city, by business category. That is, count the unique reviewers by city, by business.

2. Rank all ***cities*** by # of stars descending, for **each category**

3. What is the average rank (# stars) for businesses within 15 km of Edinburgh Castle, Scotland, by type of business (category)? **Note: A business with more than one category will be listed more than once, once per category,**

Center: Edinburg Castle, Scotland, UK
Latitude/Longitude: 55.9469753, -3.2096308

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:

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

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:

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

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

4. Rank reviewers **in Q3** by their number of reviews. For the top 10 reviewers, show their average number of stars, by category.

5. For the top 10 and bottom 10 category ***Food*** businesses **in Q3**, (in terms of stars), summarize star rating for reviews in January through May only.

Grading (total 150 points)

This assignment **MUST** be completed on your own. Duplicate assignments will be flagged and failed. 30 points each question (1-5) = 125 points

Extra Points (100 extra points)

Complete the assignment in Apache Spark (Scala, Java or Python) (**you must still complete the original exercise**). 20 points per question.

Extra Points (20 extra points)

1. Provide visualizations for results (distributions, graphs, maps, in any suitable package).

Submission:

In a single zip package, with your name in the filename, submit:
- runnable code for all questions, clearly labeled (no dataset).
- results data for each question.

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/