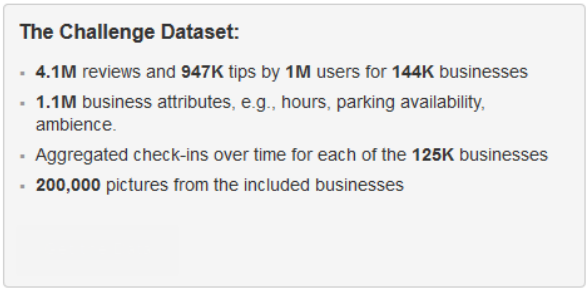
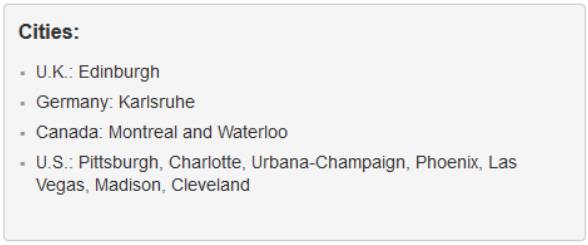
CS9223 Programming for Big Data – Assignment 2

Due Date: March 10th **11PM EST**

Details

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





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

1. Summarize the number of reviews by US city, by business category.

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

3. What is the average rank (# stars) for businesses within 10 miles of the University of Wisconsin - Madison, by type of business?

Center: University of Wisconsin - Madison
Latitude: 43 04’ 30” N, Longitude: 89 25’ 2” W
Decimal Degrees: Latitude: 43.0766, Longitude: -89.4125

The bounding box for this problem is ~10 miles, which we will loosely define as 10 minutes. So the bounding box is a square box, 20 minutes long each side (of longitude and latitude), with UWM at the center.

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

5. For the top 10 and bottom 10 food business near UWM (in terms of 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 (50 extra points)

1. 20 points: complete the assignment in Apache Spark and review the difference in approaches (**you must still complete the original exercise**).
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/