MSAN 692: Group Project

Maria Daltayanni

September 16, 2015

- **Due**: Sat Sep 30. Please submit your file project.py through canvas.
- **Related background reading**: Course material from weeks 1-6.

Description

The project will be an application of what we learnt (will learn) in this class. You are expected to do the following:

1. Come up with an interesting source of data that you can crawl/scrape. Examples:
   - Yelp reviews (http://yelp.com)
   - Zocdoc reviews (http://www.zocdoc.com/)
   - Amazon reviews (http://www.amazon.com/)
   - Yahoo news (http://news.yahoo.com/)
   - ...
2. Collect the data by writing a python script (advised) or by using wget. You are expected to collect a descent amount of data. For example:
   - For the Zocdoc scenario: more than 500 doctors and their information and all their reviews.
   - For the Yelp and Amazon scenario: more than 500 reviewers, or more than 500 items/products and their respective reviews.
   - For the Yahoo news, 30 days of daily news in all categories (thousands of articles and related information).
3. Based on the data you collect, create an ER diagram that best describes your problem, transform it to the respective relational model and create the respective database (in PostgreSQL). It is optional to submit the ER diagram, but you should submit the relational model.
4. Load the data in the database and run at least 10 meaningful SQL queries (including queries with joins, updates and deletions). Create indexes to make your queries more efficient.
5. Analyze the results of these queries, and visualize at least 5 interesting insights that you learned from this data analysis.

Groups

You can choose to work individually or in groups of 2-4. If you want to form a group but you don't have a partner, please send me an email. In case you form groups, you will be asked to privately evaluate your partner's contribution in the project.

Timeline

- September 21^{th}, group formation: Complete this form by September 21^{th}.
- September 23^{th}, project proposal: You are expected to write down an one page proposal of your project. Include the website you will crawl, what data you will collect and why this is an interesting dataset to analyze.
- September 30^{th}: You have to be done with the data scraping/crawling part by the end of September, so that you can have time to come up with the SQL queries and analyze the data.
- October 7^{th}, in-class project presentation.
- October 14^{th}, project report (one report for each group).