Progress report - CSE 5334

• Introduction/Motivation:

Yelp data challenge is regularly undertaken by many different universities data science/ database classes. In our project, we'll be looking into the development of intelligent algorithms to extract useful information for a vast dataset. The dataset volume poses many problems in itself such as loading data, generating queries and handling data to perform various arithmetic operations. In order to compare our work with other counterparts, we are closely looking at the group from University of Washington and Georgia Tech to notice there way of handling this unique problem.

• Problem definition

We have divided the project in 3 steps.

1) Storing the json data in a relational databases such as Postgres and Hue.

2) Running queries to obtain some useful facts about the reviews.

3) Develop unsupervised learning models to further analyze the relationship in data. [1] We are looking to find some labeled data and then apply semi supervised learning as in [1]. As detailed out in [1], we develop our own neural net architecture that can handle this massive dataset. Alternatively we try to use algorithm as in [2] where it uses behavioral footprints, an unsupervised method to generate data, then train again with normal supervised, similar to an auto-encoder.

• Preliminary Results

Currently we are the first step of the project:

1) We have parsed the json data to populate the database using simple json package. We have worked on to avoid duplicates and preprocess the data for query and analysis. for example in json format each business has a category attribute which stores a list of categories the business belongs to. Each category is repeated many times leading to duplicity. we created three tables called business, business category and category to reduce the duplicate data.

2) We are successfully able to use Hue to find the best restra out of the database using hive tables and an SQL query.

Challenges:

We are having some problem in using postgres when it uses a JDBC connection and currently working on it.

References

[1] Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien. "Semi-Supervised Learning."
[2] Mukherjee, Arjun, et al. "Spotting opinion spammers using behavioral footprints." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.