Exploring machine-learning ranking systems through the Yelp dataset.

SI-650 / EECS-549 Project Proposal

Yash Bhalgat, *uniqname: yashsb* , Brad Schwartz, *uniqname: baschwa*

Objective

- Train a machine learning algorithm able to *score* restaurants based on the user-generated unstructured reviews data.
- Establish a ranking of the restaurants based on the scores generated by our algorithm.

Motivation

With the increase in review websites, businesses are hugely affected by online user reviews. It is critical that we understand the social behaviour and give businesses a more proper understanding of the user market and *vice versa*.

Dataset

We will be pulling our datasets directly from the Yelp website. Yelp is very public with their datsets, providing them in clean and sanitized files, as either JSON or SQL format, including a related photographs dataset. This means no ugly and invasive web-scraping will be necessary.

Timeline

- Create and load database/easily-queried format.
- Explore machine learning algorithms - unsupervised vs. supervised.
  - Unsupervised algorithms: k-means clustering, artificial neural networks
  - Supervised algorithms: Support vector machines, Decision Trees
  - Reinforcement learning will most likely not be explored, due to it primarily focusing on learning in dynamic environments.
- Begin scoring and ranking of restaurants, exploring incorporating different attributes of business.
  - *Reach*: Incorporate photograph dataset.
- *Reach*: Web Interface for exploring datasets.

Deliverables

Our deliverable will be a report, that will include our process, techniques explored, and analysis of our scoring and ranking system.

Plan of Action

Our plan of action is essentially laid out in our timeline. We will first investigate the best format for storing and querying our data. Next, we will explore machine learning algorithms, with the main focus here being looking into different metrics for ranking the unlabeled data. From here, we can see how different businesses rank against each other based on their reviews and scores.

Testing and Evaluation

A large portion of working with machine learning algorithms is choosing loss functions and hyper-parameters. A common technique for this is cross-validation. Of course, this in large part depends on how we score (choose our loss-function). This is difficult for unlabeled problems, especially depending on what we choose for a ”ground-truth”. One possible evaluation would be looking at the tf-idf for ”positive” words, along with how ”useful”, ”funny”, or ”cool” (Yelp attributes of reviews) other users found some review and how many ”stars” a company has.