Project Proposal

Hemeng Maggie Li, Chudan Ivy Liu, Jiawen Jasmine Zhu

I. Overview

Yelp has become a more and more popular app in our daily life. When we do not know where to go for dinner, we often go on Yelp to check out the restaurants nearby or expect Yelp to recommend some good restaurants based on our past reviews. Our project is about the Yelp Challenge. Although the deadline has already passed, we can still access the data. There are a lot of things we would do with the dataset, but we would like to further explore the recommender system. For a given user and the ratings he/she had in the past, what would be the his/her rating for another business? We are interested in this project not only because we use Yelp a lot, but this project also relates a lot to machine learning, which is a really important topic now in Computer Science. Therefore, we would like to get our hands on some machine learning experience with the Yelp dataset.

II. Specific Python libraries

Python libraries/resources we'd like to investigate/use including:

- Sklearn
- Numpy
- Pandas
- Tensorflow
- TextBlob

Online resources from which we get started including

- Samples and basic set-up come from Yelp GitHub https://github.com/Yelp/dataset-examples

III. Timeline

Week 1 (this week):
We look into different python libraries and decide the ones we would like to use in this project. The candidates are Sklearn, Numpy, Pandas and Tensorflow. We follow the tutorials of these libraries and run some provided demos on our laptop to get familiar with them.

We sketch out the general plan of our project. There are two paths. One is the basic path, to train a recommender system using linear regression model, with the help from Sklearn library. The other is a more in-depth path. We want to train the system using deep learning. Because each of the word in review depends on previous context, recurrent neural network with attention mechanism could help us process the reviews from a broad perspective. Moreover, multiple layers of recurrent neural network could improve the performance.

Week 2 (by Apr 16th):
The goal for week 2 is to process the dataset and fetch the data we want, and start the basic path hopefully. We only want to text and rating parts of the Yelp Data Challenge dataset. First we sort the restaurant reviews into three groups, favorable (reviews with 4 or 5 stars), medium (reviews with 3 stars) and unfavorable (reviews with 1 or 2 stars). For the basic path, we don't want to do the training on raw text, since there's a lot uncertainty to train on text. We will transform each text review to a score depending on its positivity/ negativity. The linear regression model will take in the score of a restaurant review as input, and give predicted label.

Week 3 (by Apr 23rd):
The goal for week 3 is to finish the basic path mentioned in week 2 and start on the other in-depth path. For week 3, we want to train on text instead of score on the recurrent neural network. We convert the text review to word vectors using TextBlob library, and feed these word vectors to the hidden layer of network. Other specific parameters of the network are still to be decided. But we want to use softmax to index the result as a probability vector during training and find a way to build an attention mechanism based on that.

Week 4 (by Apr 30th):
The goal for week 4 is to improve the performance and wrap up the project. For example, we want to adjust the learning rate, change weight initialization methods or try other optimization methods (such as adding momentum or pocket algorithm). We will conduct the result analysis and give a final conclusion on the recommender system we've trained. If we could implement the deep learning model, hopefully, we would like to compare the pros and cons of these two different approaches.

Reference:
https://www.slideshare.net/MarkLevy/efficient-slides
https://cs224d.stanford.edu/reports/LiuSingh.pdf