W266 Project Milestone
Yi Jin, Ashley Levato, Umber Singh

Text Reviews Rating Prediction

Overview:

We will be using NLP processing techniques to predict a particular review's star rating from the associated text. Specifically, we will be implementing a convolutional neural network (CNN) implementation to accomplish this request.

While convolutional neural networks are widely used in image recognition, many suggest that the importance of invariance to shifts and rotations in objects of images loses value with NLP, a domain in which the order of words can be instrumental, emphasizing the way in which humans process words. This notion leads to the use of Long Short-Term Memory and other RNN's as the preferred NLP technique since it sequences and maintains an internal representation of words one at a time.

However, our thought behind leveraging CNN's is that although the fixed "memory" may pose an issue with extracting semantic information where long-range context is important, the context of Yelp's text reviews are better suited with a model that weighs word and character patterns seen repeatedly pieced together to create meaning. In our project this meaning will be extracted from sentiment scale/polarity analysis. Moreover, CNN's are much easier to design and train making them faster to implement.

Current Status:

So far, we have completed the following initial steps:

1. **Download Dataset:** Downloaded Yelp dataset file (https://www.yelp.com/dataset_challenge), which consists of 2.7M reviews from 687K users for 85K businesses
2. **Data Loading:** reading in data from csv file
3. **Baseline Model:** Applied bags of words model and multinomial NB. The accuracy is 0.568 and the f1 score is 0.564.

Link to notebook: https://github.com/ashley-dsci/266-Yelp/blob/master/W266_Final_Project_Baseline.ipynb

Next Steps:

With the baseline model completed, we can now begin to experiment with CNN implementations and NLP pre-processing techniques to minimize our output errors.

Some CNN hyperparameters we will have to assess:

- Stride Size
- Pooling Layers
- Channels