Project Proposal - CS410 - NVK

Nikhil Khandekar (nikhilk5) (captain)
Vineet Chinthakindi (vineetc2)
Kyr Nastahunin (knasta2)

1. **What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?**

For our free topic, we would like to perform sentiment analysis of businesses' Yelp reviews by keywords. Our goal is that whenever a user enters a link to some yelp business, they are presented with a list of things that this business does well and a list of things that are bad about this business. See table 1 for examples of our inputs and outputs.

Input	**Output Positives**	**Output Negatives**
"We had to wait for so long until we were finally able to get a table. The waiter was in a rush and brought us the wrong order. The beef was great, but the service was disappointing."	Beef	Long wait Wrong order Disappointing service
"They fixed my car so fast!"	Fast fix	N/A

This will help the users see generalized information about what goes into a particular business' star rating and make their own judgment if this business provides the services they currently need. For example, some places can be good at service #1 and bad at service #2 and therefore have a bad rating. However, if a user is only interested in service #1, they might still want to deal with that business despite its low overall rating. On the other hand, this project can also help business owners analyze what they're doing right and what needs improvements based on reviews.

We will use the Sentiment140 dataset to train our model. A potential challenge that we might face is extracting opinions about specific keywords rather than the whole document. We will also use Yelp API to get business reviews to be analyzed (our input). Our plan is to perform keyword extraction first. This will identify potential positives and negatives in the review. Then we will perform sentiment analysis for each of those keywords and categorize them as shown in

table 1. We believe we can achieve keyword analysis by performing a sentiment analysis only on the sentences that mention the keyword and ignoring all other sentences in the review.

If we use all 60 hours to create and train our model and achieve reasonable results, we will deliver a Jupyter Notebook with all the training we had and predicting what we had to perform. If we make good progress fast enough, we will try to expose this model as a Django API (Python) and provide a simple web UI that allows the users to paste the link to a business on yelp and then displays the results of running our analysis in a table (JavaScript).

We will evaluate our success based on a testing dataset we plan on building from random businesses on Yelp. It is hard to define a mathematical measure of success since it is a subjective topic, so we will be making explicit judgments ourselves during the evaluation process to see if our model extracts keywords and categorizes them as we expect.

2. **Which programming language do you plan to use?**

We plan to use Python to gather data and build our model. We plan on using such Python packages as NLTK, SKLearn (scikit-learn), and MeTa. In addition, if we decide to build a wrapper for our prediction model, we will use Python's Django and JavaScript to build an API and UI, respectively.

3. **Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**

We expect to spend about 20-30 hours coding a keyword extraction algorithm. 20-30 hours for coding a sentiment analysis algorithm using those keywords. And about 10 hours for coding a wrapper for our model.