Alex Lederer, Colin Swan, William Sheppard-Sage, Kevin Zhang

Yelp Data Project Proposal

CS1671

March 1st, 2016

**Categorization of Words to determine Seasonal Relevance**

**Concept**

While taking a cursory look at the data available through the Yelp data challenge, we noticed that the date on which a review was made is included with each review. We began brainstorming ways to use this information, and we realized that it could be interesting to see if certain words and phrases could be categorized as particularly relevant during certain weather conditions. So for example, during a cold wintery season people might mention things like coffee, hot chocolate, fireplaces, and use adjectives like “warm” or “cozy” more often in positive reviews. And during a hot summer season they might mention lemonade, the outdoors, animals, and “cool” or “refreshing” refreshments. This type of information could be useful for a business to determine what goods to offer customers at different times of the year, or what aspects of their business to highlight throughout the year.

**Development Process**

To accomplish the goal of categorizing words as being more relevant for certain times of the year, we would begin by creating a training corpus from the yelp data challenge set. We would select a particular city to collect reviews from, and then use the dates available with each review to separate the reviews into seasons (e.g. winter, summer, spring, and fall). We would then run categorization algorithms on the data sets to look for words that occurred significantly more often in certain seasons as opposed to others. Once the model is trained we can then attempt to use the model to predict during what seasons other reviews were made. This provides us with a way to quickly determine the accuracy of the model, as we can use the dates for the test reviews to determine if the prediction was correct or not. If the model is able to predict the season in which a review was made with relatively high accuracy then we know the words in each category were particularly relevant for the seasons they were categorized in, meaning the words were more relevant to customers during that particular time of the year.

The biggest advantage for this project is the ability to quickly determine the accuracy of the resulting model using the date data available for reviews. It also allows some flexibility in regards to the creation of the model; some ways in which the implementation could vary could be, for example, tagging the words in the reviews and only looking at nouns and adjectives when looking for words that are relevant to seasons. This would allow us to focus primarily on goods that customers are more interested in during a season (e.g. hot chocolate and fireplaces during the winter), as well as certain characteristics that are more attractive to customers during a season (e.g. cozy, warm, inviting). Additionally, if the model proves relatively accurate when applied to seasons, finer granularity could be applied for an examination across individual months. Furthermore, special consideration will be required for cities in climates that vary relatively little. It may be that a differently tuned model is required for such cities, or that extracting distinct patterns for each season is entirely impossible, but this will be easy to take into account given the approach we have chosen to take.