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

Professor Hwa

CS 1671

1 March 2016

**Categorization of Words to Determine Seasonal Relevance**

**Program 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 would be interesting to see if certain words and phrases could be categorized as particularly relevant during certain weather conditions. For example, during a cold wintry season people might include nouns such as coffee, hot chocolate, fireplaces, and use adjectives like “warm” or “cozy” more often in positive reviews. Conversely, during the hot summer season, reviewers might mention lemonade, the outdoors, animals, and “cool” or “refreshing” refreshments. This type of information could be useful for a business to determine which goods to offer their customers at different times of the year, or what aspects of their business to promote during each season.

**Development and Validation Process**

To accomplish our goal of categorizing words by seasonal relevance, we will begin by creating a training corpus from the Yelp Data Challenge review dataset. The training corpus will be collected from a particular city, and using the dates available with each review, separated into seasons (e.g. winter, summer, spring, and fall). We will then run categorization algorithms on the datasets to identify words that occurred significantly more frequently in certain seasons as opposed to others. Once the model is trained, we can then attempt to use the model to predict the seasons in which the remaining reviews were made. Since all Yelp reviews contain a date, this approach provides us with a way to quickly determine the accuracy of the model. We can then repeat the development and validation steps to improve program accuracy. 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 are particularly relevant for the seasons they were categorized in, meaning those words were more relevant to customer experiences during that particular time of the year.

A notable advantage of this project is the ability to quickly determine the accuracy of the resulting model using the data available in reviews. This allows some flexibility in regards to the creation of the model; some ways in which the implementation may vary include tagging the words in the reviews and only considering 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). Although there are many terms we anticipate being affiliated with specific seasons, this program will allow for the potential identification of unexpected terms that are frequently discussed by reviewers. Furthermore, if the model proves relatively accurate when applied to seasons, finer granularity could be applied for an examination across individual months. Special consideration will also be required for cities in climates that vary relatively little. One way our program will accommodate for climate differences among cities is by utilizing different training datasets for each individual city. *[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. =* ***Not sure if this sentence is necessary, or if we should have an extra “conclusion”-type sentence****.]*