SODA 501 project

Shipi Kankane, Claire Kelling, So Young Park, Xiaoran Sun

Problem Statement

For our group project for SODA 501, we will be analyzing a dataset of geo-tagged tweets from Twitter. The time window for these tweets is December 16-31, 2015. Overall, this dataset has over 4.8 million geo-tagged tweets and is stored in JSON format. Since this dataset is collected over the holiday season, we are interested in conducting a spatial analysis of holiday travel patterns. Investigating this data will reveal important patterns about travel and can illuminate some clusters of origin and destinations. For example, we are interested to see if we can see if the population of a college town is leaving the college town during this time period. This analysis will have many challenges. First, we will have to distinguish tweets from organizations/campaigns/celebrities from individuals. We will have to identify tweets from which we can infer potential travel. This will be accomplished by filtering individuals who made a minimum of two tweets in two distinct places to illustrate that they have traveled. This will be followed by further understanding the the content of tweets and how they relate to their travel. We then can develop questions about what people are saying in these tweets, perhaps relating to where they are traveling.