Spatial And Temporal Sentiment Analysis For Tweets Team 049

Song Luan(slaun) Xiaokai Sun(xiaokais)

**Description**

Natural Language Processing is a popular topic currently, and we want to apply it to analyze how tweets reflect people’s mood based on space and time. We will use Stream API to get the all of the tweets in a period of time. Since every tweet contains information of location and time. We will use analyze how cities and date influence the score of every tweet by applying natural language processing technique. Then we can know in which city or which day (e.g. weekend) people feel happy. Somehow tweets can’t reflect people’s mood depending on the location and time, because somebody post some official information daily which can’t tell their mood exactly. Or when people feel happy, they won’t spend time on posting tweets. We will delve how the data reflect people’s mood by getting the analysis result.

**Platform and Techniques**

We will use Stream API to get the tweets from twitter by specifying some dates. And we plan to use the Stanford NLP API to do the sentiment analysis. We will analyze the data on local computer sequentially, using local hadoop and using AWS EMR. By these three methods, we can understand better the advantage hadoop on processing big data and disadvantage on a single machine. Two team members will be in charge of sequential analysis and hadoop MapReduce code on city and date separately.

**Expectation**

We think people in small cities and weekend should be happier than those in big cities and weekdays. Accordingly, the tweets in small cities and weekend should have higher score which means more positive attitude based on the benchmark of Stanford NLP. We think the speed should be like this: EMR > sequential analysis > local hadoop. We think local hadoop is running mapreduce on a local machine and needs configuration and assign jobs which takes too much time and decrease the advantage of hadoop framework. EMR is the fastest speed because there will be a couple of machines working together which can get the work done much faster than sequential analysis.

**Project Timeline**

Before April 7th, we will study how to use Stream API and get the dataset from Twitter.

Before April 15th, we will calculate the average score of all tweets on every day and some cities for one-week tweets using sequential analysis and local hadoop.

Before April 23th, we will calculate the above score using AWS EMR on more tweets.