Mining Twitter Using R

Presented to the St. Louis R Users Group
By Mathew Woodyard
April 3, 2012

1 Motivation

We set out to answer the questions: Why aren't people buying my widgets? Are people saying bad things about me? Why are the students of Southern Illinois University Edwardsville (SIUE) unhappy?

These are classic questions concerning not only product marketers, but statisticians. What convinces people to tell us what they think, what they own, what they will buy, where they are, and who they are? How do we get honest answers? How do we get these answers in aggregate?

There is some good news yet because...

1.1 300 Million People Are Here To Chat!

For those who know little about Twitter: it is a "microblogging" service that facilitates communication. Each message, a "tweet" contains no more than 140 characters. 300 million people use Twitter. But how do they use it? And why?

Twitter has some use cases and features that make for a nice corpus. Twitter is often used for consumption of news stories and participation in current events. Many people also use twitter to express emotional states: often complaining about and praising places, products, and services. Since tweets contain no more than 140 characters, finding relevant tweets is normally a fairly easy task. Twitter's automatic grouping and "hashtags", tags that group things and start with a hash mark (#), enable us find what we with relative ease.

In my example, I will use tweets about SIUE. Tweets were selected from the stream based on the occurrence of the terms "SIUE" and "#onlyatsiue".

Figure 1. A Tweet.

Now that we know people are generating data we can use, how can we provide analysis that makes these tweets useful?

2 Harnessing Tweets

Our strategy will consist of the following steps.

1. Sip from the stream of tweets using twitteR and ROAuth to build a corpus.
2. Discover your new corpus.
3. Select a sentiment lexicon. Download it and load it into R.
4. Score tweets using the function provided or using your own classification scheme.
5. Summarize the results.

2.1 Building a Corpus

2.2 Exploring Your Hot New Corpus

2.3 Bags of Words

2.4 Scoring Tweets

2.5 Summarizing Your Results

3 Limitations

3.1 Bags of Words Approach

- Irony and function used. Give example.

3.2 P*R*oblems

3.3 Requirements

4 License