Project

Twitter Text Mining

The Team

Guide: Dr. Mary Saira Bhanu

Members:

Lenoy Jacob	106108041
Ashok Kumar	106108052
Anirudh S	106108053

Motivation

- Micro-blogs are a challenging new source of information for data mining techniques.
- They are social tools and therefore, are a source of valuable real-time data.
- Twitter is a very popular microblogging service where users create status messages (called "tweets"), which reflect the personality of the user.
- Analyzing these tweets can help us build a personality profile of the user, which would help other users follow the people they want.

Aim

This project aims to:

- Analyze a user's tweets
- Mine through them to find specific keywords
- Associate a sentiment (+ve/-ve/neutral) to tweets with these keywords by different sentiment analysis methods
- Build a profile from the information obtained for the user

Challenges

- Differences between Twitter and previous web/document text mining research:

  1. Length. The maximum length of a Twitter message is 140 characters.

- 2. Available data. Another difference is the sheer magnitude of data.
- 3. Language model. Twitter users post messages from many different mediums, including their cell phones. The frequency of misspellings and slang in tweets is much higher than other domains.

Procedure

- Data Collection
- Keyword Classification
- Sentiment Analysis
- Association of sentiments with keywords
- Building the profile of the user
- e.g. If a user tweets positively about Android and negatively about iPhone, output would be:

  <username> (Android +) (iPhone -)

Implementation – Twitter API

- The Twitter API currently provides a Streaming API and two discrete REST APIs. The API is HTTP based, and GET, POST, and DELETE requests can be used to access the data.
- The API uses basic HTTP authentication and requires a valid Twitter account. Data can be retrieved as XML or the more succinct JSON format. The format of the JSON data is very simple and it can be parsed very easily because every line, terminated by a carriage return, contains one object.

References

- Sentiment Knowledge Discovery in Twitter Streaming Data - Albert Bifet and Eibe Frank

  (http://www.cs.waikato.ac.nz/~eibe/pubs/Twitter-crc.pdf)
- Twitter Sentiment Analysis - Alec Go, Lei Huang, Richa Bhayani

  http://twittersentiment.appspot.com/

Thank you