Software Requirements Specification

for

Trends Analysis

Prepared By

Group 1,

IT 7th Semester

Contents

1) Introduction

1.1) Purpose

Extract trending topics from a continuous stream of tweets from Twitter

1.2) Product Scope

Knowledge of trend patterns can be used to predict human behavior. This has a variety of applications. For example, trend patterns of financial news can be used to make profitable predictions about stocks.

1.3) References

Twitter Data Analytics – Shamanth Kumar, Fred Morstatter, Huan Liu

NewsInEssence: Summarizing Online News Topics – Dragomir Radev, Jahna Otterbacher, Adam Winkel, Sasha Blair-Goldensohn

Streaming Trend Detection in Twitter – James Benhardus

Trend Analysis of News Topics on Twitter – Rong Lu, Qing Yang

2) Overall Description

2.1) Product Functions

➢ Backend

- **Tweet Extraction –** A process runs in the background which uses the Twitter Streaming API to continuously extract tweets, pre-process them, and store in a persistent storage.
- **Tweet Clustering –** The tweets received are put through a clustering process, which helps us aggregating similar tweets.
- **Topic Extraction –** When we have enough information about a cluster, we can safely extract the corresponding topic.

➢ Frontend

- **Top trending topics –** The clusters are sorted based on size and the top results are displayed.
- **Trend Graph –** Each topic has an associated graph that shows the trend pattern in the time domain.
- **Related Resources –** Each topic has a list of related URLs.
- **Related Topics –** List of topics which may have close associations with a given topic.

2.2) Additional Features

Search - User is given an option to search for trends related to specified keyword(s).

Limitations:

- We have access to only a maximum of 1-week old tweets.
- There may be a processing overhead involved.

2.3) User Classes

- **The individual** – The common users of the website. They are the ones who want to stay updated with current affairs.
- **The corporation** – News agencies and financial corporations keep track of specific news topics for decision making.
- **The academia** – The data collected and the analysis provided can be used for a plethora of research projects.

3) External Interface Requirements

3.1) User Interfaces

The user is provided with a website which displays the trending topics.

The home page consists of the top 10 trending topics and an option to choose category.

There will be a search box in which the user enters keywords for searching.

Each topic has its dedicated webpage which displays related links, related topics, time-based trend graph etc.

3.2) Software Interfaces

- **Twitter Streaming API** – API provided by Twitter that gives the product access to 5% of the latest tweets which can be filtered according to given key words.
- **Twitter Search API** – API provided by Twitter that gives product access to Twitter's search results.
- **MongoDB** – The proposed NoSQL DBMS tool for persistent storage.
- **Django Framework** – A python-based web framework for web development.

4) System Features

4.1) Tweet Extraction

The Twitter Streaming API provides access to a continuous stream of tweets. Each tweet is represented by a JSON object that contains a lot of unnecessary information. This has to be cleaned, stemmed and the stop words need to be removed. The output of this pre-processing is stored in the database.

4.2) Tweet Clustering

Relevant clustering algorithms like k-means can be used to calculate the closeness of a given set of tweets. This gives us a set of clusters which is self-adaptive and can then be used for topic extraction.

4.3) Topic Extraction

Using named entity recognition and graph-based algorithms, topic can be extracted from each individual cluster.

4.4) Top Trending Topics

Once topic extraction is done, we can keep track of the number of mentions for each extracted topic. Using data structures like max-heap, we can dynamically come up with the top trending topics.

4.5) Trend Graph

For each topic, we maintain controlled historical information upto a week. This helps us plot a smooth time vs frequency graph which will provide a visual analysis of the trend pattern.

4.6) Related Resources

Frequently twitter users share URLs of related news articles, blogs, etc. The frequency of a URL can be used as a metric for authenticity of the resource. Based on a threshold, we can guarantee authenticity and display it on the website.

4.7) Related Topics

Distance between clusters can be compared against a experimentally calculated threshold to determine closeness and relevancy.

4.8) Search

The Twitter Search API provides with a result set for given keywords. Optimization has to be performed to ensure real-time search functionality.

5) Glossary

NoSQL – A NoSQL (originally referring to "non SQL" or "non relational) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.

k-means – k-means clustering is a method of vector quantization that is popular for cluster analysis in data mining.

JSON – JavaScript Object Notation is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.