Twitter Data Stream Sentiment Analysis

By

Group 5

Minor project Software Design and Application 2017-2018

Tim Al 4439848  
Casper Spronk 4369475  
Michiel O’Herne 4387783

Course Professor: Dr. Georgios Gousios

The following report covers our minor project in big data analysis. Detailing the challenges faced and the progress made over the span of several weeks.

**Table of contents:**

1. Project start
   1. Objectives
   2. Motivation
   3. Tools used
2. In-depth project setup
   1. Obtaining the data
      1. Data Sources
      2. Converting obtained data
   2. Analysing the data
   3. Visualisation of analysis
3. Results
4. Addendum
   1. Similar projects
   2. Alternate possibilities

**Project start**

The main objective we set out to accomplish was performing sentiment analysis on the content of tweets, and to present this in several directly understandable manners.

This task can be split into three distinct sections: gathering and sorting the data, performing analysis on the data and lastly portraying this analysis in a useful medium.

**Objectives: (general description)**

* obtaining the data
* analysis
* Visualisation of the analysis

**Motivation:**

Description of the use of what we did  
Description of made choices? Vs include in in-depth description

**Tools used:.**

1. **Python:** Programming language with an aptitude for data analytics tasks.
   1. **NLTK**: Natural Language Toolkit, a suite of libraries focused on the analysis of language.
      1. **Vader Sentiment Analyser:** Part of NLTK focused solely on sentiment analysis.
   2. **Plotly:** a graphing library built to provide a multitude of graphing solutions.
   3. **Tkinter:** A graphic user interface package.
2. **Twitter API:** an api supplied by twitter to allow streaming of their data.
3. **Github:** Code development platform using a web-based git version control repository.
4. **PyCharm:** an IDE tailored to the Python language

**In-depth project setup:**

**Obtaining the data:**

We have chosen to concurrently use two methods of data gathering, i.e. already created data sets obtained from a database, along with live streamed data directly from twitter itself. Our focus is directed more towards the streaming of tweets, as this data opens possibilities to provide real time insight into trends and public opinions. The pre-made datasets however form a way for us to test our analysing tools in a consistent method.

**Data sources:**

Twitter has set up its own api that allows for live streaming of tweets in real-time, as such our data stream is built around this api. The output of Twitter their api is encoded in JSON, containing not only the tweets themselves, but also all possible related data such as number of like, retweets and reactions tweets have received.

**Converting obtained data:**

Tools used to obtain data + convert said data into a usable format:  
Used environment / packages  
from-to formats

**Analysing the data:**

* Divide the data into positive - negative tweets
* Give the overall opinion of the public (i.e. twitter users) of the topic
* Give the popularity of given subject

Tools used to sentiment analyse the data  
used environment / packages

**Visualisation of analysis:**

Description of used manners of visualisation

Detail on tools used:  
used environment / packages

**Results:**

Couple of graphs + part of tweet data set?

**Addendum:**

**Similar project / methods:**

Multiple project setup to do twitter analysis

Mainly business focused models renting their services for research and development purposes

**Alternate methods we could have used:**

Spark streaming s