Fall 2017

Big data in media technology DA2210

Project Proposal

Eysteinn Gunnlaugsson

Carlo Rapisada

Steinn Elliði Pétursson

Yunus Koçyiğit

23. september 2017

Project description

In today's world social media has become a large part of society. People express their opinions on the internet every day through sources like twitter and facebook. The data posted every day contains information about peoples' opinions on day to day events. Our theory is that this data expresses the opinion shift in society on known entities throughout this world and it is possible to analyse the data and formulate this opinion shift.

Project goal

The goal of this project is to do sentimental analysis on the data posted on twitter every day in order to formulate how events in the world affect peoples' opinion on known entities related to these events. For example, if a sports club just lost a match it should have won we hope to see a difference in peoples' opinion before and after the loss.

Project work plan

The first thing that needs to be done is gathering data for training. The dataset used for training in this project will be the VADER dataset which is an open source dataset and sentiment analysis tool. This will be used to evaluate different sentiment evaluation classifiers. In this project at least three different approaches will be explored, a Naïve Bayes classifier, Support vector machine and a neural network. On top of that various methods that can help with information extraction and sentimental analysis will be explored, such as tf-idf, stemming, word and sentence tokenization and more. After the training phase data will be gathered for testing. The data gathered will be related to the project description and goal outlined above. Certain hashtags will be used to gather data from the Twitter api to show correlation between events and peoples' opinions on the entities related to them.

Materials, references & libraries

VADER dataset
Twitter API
tf-idf
tokenization
word stemming
Various scholarly papers can be found about tf-idf, tokenization and stemming. These will be used for reference in the project.
The project will be written in python, various modules will be used, most notably the researchers will rely a lot on the modules provided by scikit-learn and nltk