Impact Measurement of Tweets

Student(s)

Abdullah Usame Gunay
Egemen Yigit Komurcu
Mert Kosan

Faculty Member(s)

Yucel Saygin
Inanc Arin

Sabancı Üniversitesi
FACULTY OF ENGINEERING AND NATURAL SCIENCES

PURE

Abstract

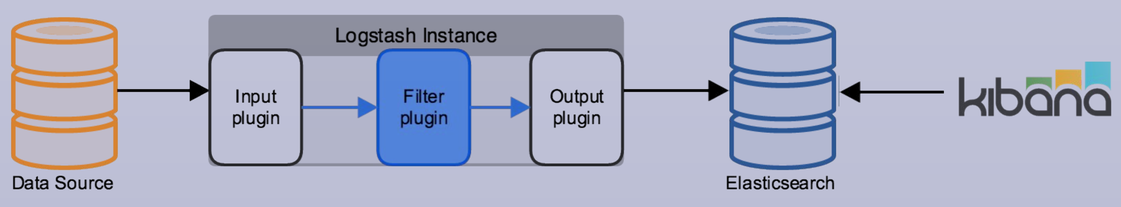
This research project is focusing on the impact of tweets on Twitter. There are some factors in Twitter -retweets, likes or quotes- that can show the impact. However, this is giving missing information about the impact of tweets. People can copy or change tweet a little bit with some comment and post as her/his tweet. They are called as "**Hidden Retweets**" (Arin, 2017) by which could help to get more accurate analysis performance. Also during the process of finding hidden retweets, we need to define some threshold values and make preprocessing to increase accuracy and speed of the process.

Twitter Usage Statistics (2008-2013). Approximately 500 million tweets are tweeted per day. This is huge data and is worth research.

Materials

We used ELK stack and Twitter streaming API for getting, processing and visualizing the tweets. ELK stack stands for the programs Elasticsearch, Logstash and Kibana which work perfectly together in analyzing big data.

Elasticsearch is a very powerful tool to solve data extraction problems. It is also much faster than SQL queries which makes it a great solution for full-text searching. Logstash is a data processing pipeline that allows the user to collect, process and load data from many different sources. We collected the tweets via Logstash and transferred them to Elasticsearch in this project. Kibana is an open-source data exploration and visualization tool that allows us to visualize some statistics about our collected data which includes millions of tweets.

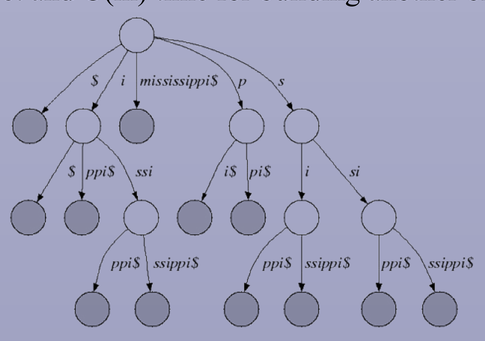


Algorithms and Data Structures

To measure the similarity between tweets, our first approach was to use the most known method- longest common subsequence (LCS). However, the time complexity of the solution of LCS is O(m*n), given two strings A and B with sizes m and n since it creates a matrix of two strings and finds the result by incrementally. Therefore, LCS was not a feasible approach to this problem.

$$LCS(X_i, Y_j) = \begin{cases} \emptyset & \text{if } i = 0 \text{ or } j = 0 \\ (LCS(X_{i-1}, Y_{j-1}), x_i) & \text{if } x_i = y_j \\ \text{longest}(LCS(X_i, Y_{j-1}), LCS(X_{i-1}, Y_j)) & \text{if } x_i \neq y_j \end{cases}$$

Our second approach was to create a generalized suffix tree (GST) of the strings that we wanted to compare. By creating a GST from the set of the tweets, we can perform a fast searching using lexical similarity of tweets to find the hidden retweets. The time complexity of this approach would be O(m+n) since it takes O(n) time for building a suffix tree for a tweet and O(m) time for building another one where m and n are the sizes of the tweets.

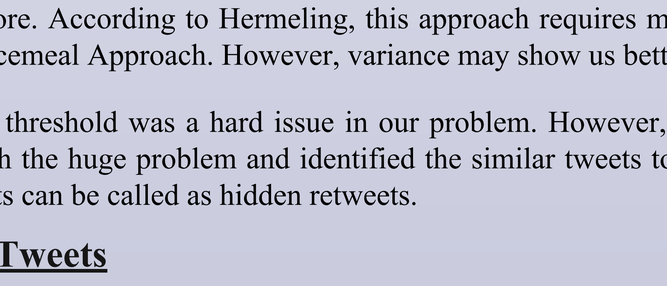


Impact of Tweets

There are three steps which are followed:

Deciding Threshold / Sensitivity Analysis

Sensitivity Analysis: "The study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." (Saltelli, 2002)



We are creating blocks for each similarity scores after first elimination (below 0.69). These blocks are in range [score-0.01, score+0.01]. Every similarity score has LCS or Suffix Tree result with them. Blocks that we created will consist these results. Then we applied two sensitivity approach which have advantages against each other.

The Piecemeal Approach:

We found max difference between values in block, and this max value refers to sensitivity of this similarity score. If sensitivity is too high, we may not select this score as a threshold value. Accuracy level is not good, however it is computationally fast.

$$\Delta = \max_{a_i, a_j \in \{a_1, \ldots, a_M\}} |h(a_i) - h(a_j)|$$

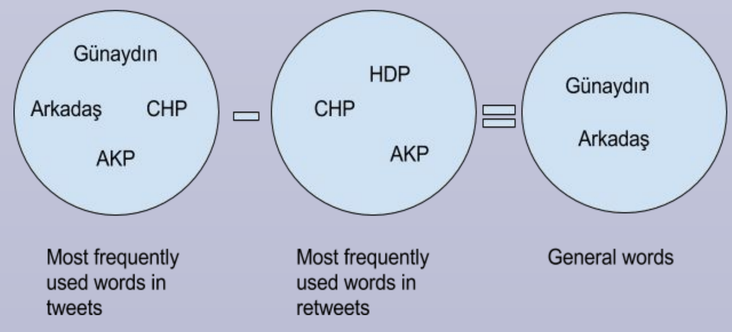
The Monte-Carlo Approach

This method is finding variance of the block and this value refers to sensitivity of the similarity score. According to Hermeling, this approach requires more computational power than The Piecemeal Approach. However, variance may show us better result.

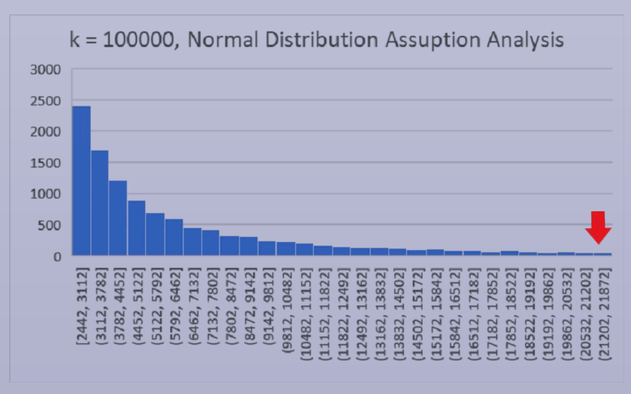
Deciding threshold was a hard issue in our problem. However, with sensitivity analysis, we dealt with the huge problem and identified the similar tweets to an original tweet. These similar tweets can be called as hidden retweets.

General Tweets

General tweets are tweets that are tweeted every day. For example, "Günaydın" (Good morning in Turkish) is a common phrase and is being tweeted every day. So when we are trying to impact of the tweet, we need to eliminate these general tweets. We are interested in impact rather than similarity.



To find general words, we proposed a method that shown on left diagram.



In this experiment, we selected lower bound as 2442 (number of words / 100000), so words with lower frequencies are discarded, and upper bound as 21872 that got from normal distribution assumption analysis with two-sigma value threshold.

Future Works

- An experiment of combining finding similarity with given threshold values and discarding general tweets.
- Turkish words contain a lot of suffixes in words. It should be handled.

References

- Arin, Inanc. (2017), Impact Assessment & Prediction of Tweets and Topics, PhD Thesis, Sabancı University
- Figure of a suffix tree for the word "mississippi" retrieved from https://seqan.readthedocs.io/en/seqan-v2.0.2/Glossary/SuffixTree.html
- Hermeling, C., Mennel, T. (2008). Sensitivity analysis in economic simulations - a systematic approach. Center for European Economic Research
- Internet Live Stats. Twitter usage statistics, 2018. Retrieved from http://www.internetlivestats.com/twitter-statistics/
- Saltelli A. (2002). Sensitivity Analysis for Importance Assessment, Risk Analysis, 22 (3), 1-12.