TWEETVIZ: TWITTER DATA VISUALIZATION

Dario Stojanovski, Ivica Dimitrovski, Gjorgji Madjarov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
e-mail: stojanovski.dario@gmail.com,
{ivica.dimitrovski, gjorgji.madjarov}@finki.ukim.mk

ABSTRACT

Twitter is the leading micro-blogging and social network service and is attracting an enormous amount of attention in recent years. Users on Twitter generate an abundance of information every day, establishing Twitter as the focal point for analyzing and visualizing social media data.

In this paper, we present a web tool for visualizing Twitter data, TweetViz. TweetViz offers several different kinds of visualizations that can pertain to a Twitter user or any keyword or hashtag entered through the interface. TweetViz also includes a so called Streamgraph visualization that represents topic distribution in a set of tweets. The topic distributions are created using LDA (Latent Dirichlet Allocation).

1 INTRODUCTION

Increasing popularity of social media has led to micro-blogging services becoming one of the most popular methods of information consumption. Twitter is a social networking and micro-blogging service that enables users to send and read short messages. These messages, also known as tweets, are maximum 140 characters long. As of this year, Twitter has over 271 million monthly active users and over 500 million tweets sent per day. As a result of the massive amount of data generated on a daily basis, Twitter has become the main focus of many researchers involved in data mining.

The Twitter community uses the service as a way of sharing personal thoughts and ideas, posting news and discussing popular topics. It is also used in marketing purposes by companies, institutions, politicians etc.

Some of the research goals so far have been to extract knowledge about user interests and behavior, detect trends amongst group of users, analyze information dissemination in the network etc.

Most of the work focuses on analyzing words, word pairs and hashtags, with little attempts made to leverage some more advance Natural Language Processing (NLP) techniques such as LDA (Latent Dirichlet Allocation) or LSA (Latent Semantic Analysis).[1]

In this short paper, we present our web tool TweetViz for Twitter data analysis and visualizations. TweetViz incorporates several user-orientated and hashtag or search term visualizations. In addition, we explore an approach where tweets are presented as a mixture of topic distributions over some time interval using LDA.

2 RELATED WORK

There has been significant work done in the field of visualizing and analyzing Twitter activity. Many scientific papers and web tools explore different approaches on visualizing data generated from Twitter. Approaches range from visualizing temporal and spatial data to representing network data. Great portion of the conducted research studies trending topics and general Twitter activity about some subject.

When it comes to visualizing tweets from a single user, the tool proposed in [2] offers visualizations that can aid to understanding the user behavior. As in our approach they explore the frequency of the user’s activity in separate days and times of the day. They provide a timeline visualization that presents tweets on a graphic, where the x-axis represents days and the y-axis represents time. Users can also classify tweets by subject, clustering tweets that contain a set of user-defined tags in the category that holds these tags. Their tool DeepTwitter also offers a tag cloud visualization, but it only presents tags specified by the user as opposed to our approach which visualizes all frequent words the user tweeted.

Some of the proposed tools analyze and visualize topic distribution in a collection of tweets. In [1], they attempt to achieve topic alignment between sets of tweets over time. As this is still an open issue in NLP, they aim at solving this problem by using their visualization tool TopicFlow. TopicFlow is an extension of NodeXL, a network visualization tool that offers tweets retrieval. This approach analyses topics at discrete time slices separately. The LDA algorithm is used to provide scaffolding for temporal analysis of Twitter trends. In this approach, similarity between topics is calculated using cosine similarity metric in order to achieve topic alignment. They too provide information for the topics, specifically the words that the topic is consisted of and their statistical importance for the respective topic.

The Streamgraph visualization technique used in TweetViz is also explored in [3] and [4]. ViralViz [3] is another tool that uses LDA and extends NodeXL. This approach differs from the one mentioned before [1] in that way that it takes a more network orientated aspect of topic evolution analysis. ViralViz uses GraphML files generated by NodeXL and then presents topic distribution as a Streamgraph. In addition to

LDA, they provide an approach that extracts keywords based on their statistical significance. In [4], the Streamgraph is not build on data from Twitter and they don't use LDA, but the same principles for creating the visualization apply. The utilization of the Streamgraph and LDA in the presented related work confirms our motivation to use this technique to visualize topic changes in a set of tweets in our tool as well.

3 TWEETVIZ

In this paper we present our Twitter analysis and visualization tool TweetViz. TweetViz offers several different interactive visualizations that can provide insight into user interests and activity as well as information about certain keywords and hashtags. TweetViz visualizations can be divided into two separate modules. The first is user-centric and focuses on analyzing user behavior from different aspects. The second module, on the other hand, is more search term orientated, where a user can explore Twitter activity surrounding a specific hashtag or keyword. Moreover, TweetViz incorporates the LDA algorithm for visual representation of topic distribution, from tweets, both from a specified user or tweets containing a search term. Figure 1 depicts the architecture of the web tool. TweetViz is consisted of separate modules for collecting tweets, preprocessing the content of the tweets, and transforming the data to an acceptable format for the visualization modules in the user interface.

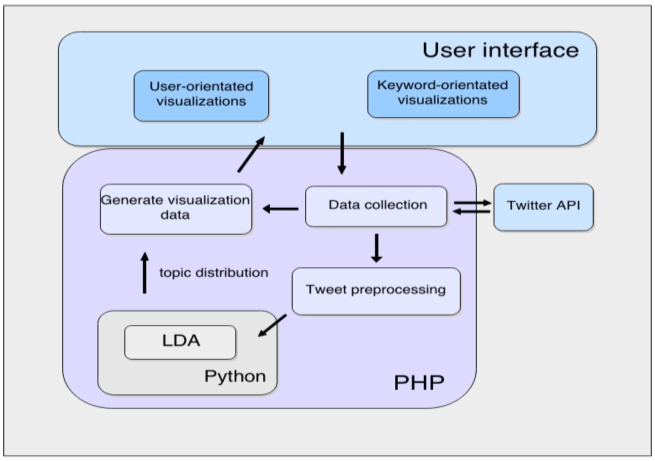

User interface
User-orientated visualizations
Keyword-orientated visualizations
Generate visualization data
Data collection
Twitter API
topic distribution
Tweet preprocessing
LDA
Python
PHP

Figure 1 System architecture

The frontend of TweetViz is built using standard HTML, CSS and JavaScript. The client sends asynchronous request to the PHP based backend which returns the data needed to create the visualizations. In TweetViz we use few third party libraries. Google Charts and d3 (Data Driven Documents) are used to build the different types of visualizations. For generating topic distributions with LDA we used the Python gensim[1] framework.

3.1 Data collection

Twitter enables third party applications and developers to get access to the enormous amount of data generated by users every day. This is done using the Twitter REST API which offers a lot of different endpoints for retrieving this data. Of our interest when building the web tool was the endpoint that allows us to retrieve tweets from a single user. We also leverage the search capabilities offered by the Twitter Search API. This is used to get tweets which contain a given search term, both keyword and hashtag.

Although it offers extensive functionality, the Twitter API has certain limitations. First of all is the rate limit window which limits the number of requests that can be sent. Another deficiency is that the search service provided by Twitter does not index all tweets and as a result, not all tweets are available for retrieval from Twitter Search. Nevertheless, it can provide sufficient number of tweets for the purposes of our web tool.

3.2 User-orientated visualizations

As mentioned before, TweetViz is a web tool that offers different types of visualizations, many of which are focused around a specific user. In order to understand what the user is interested in tweeting about and to provide insight into his behavior on Twitter, TweetViz explores different approaches to creating interactive visualizations.

First of all, TweetViz plots a chart depicting the number of tweets the user posted on a daily basis. Although it is a simple chart, it still can be used to analyze change in user activity and possibly, in combination with other visualizations, why those changes occur.

Twitter enables users to manually label the topic of the tweets they publish with keywords, or also known as hashtags. Many research papers, both in visualizing and analyzing Twitter data revolve around hashtags, proving their usefulness when extracting knowledge from Twitter data.

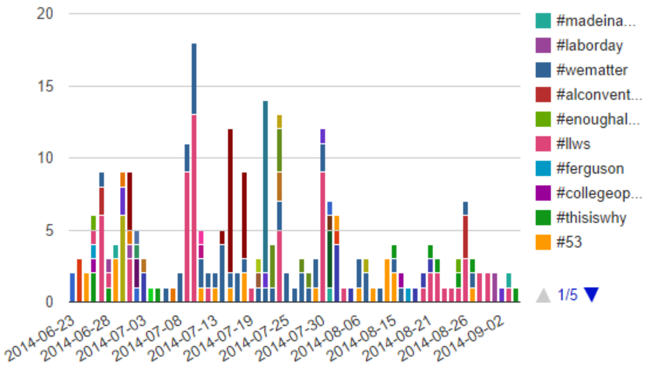


Figure 2: User-hashtag distribution

In Figure 2 is presented a stacked column chart that displays the different hashtags the user tweeted about over some time interval. This provides a nice visual way of seeing what the user is interested in and even detect what sort of topics he tends to combine.

Another interesting approach is visualizing user activity in different parts of the day and even analyzing if this affects the topics he tweets about. This approach to analyzing user activity is certainly not well explored, and this chart can bring

[1] http://radimrehurek.com/gensim/

some information as to its usefulness. Although overwhelming at first, the interactive characteristic of the chart enables users to easily get around this visualization.

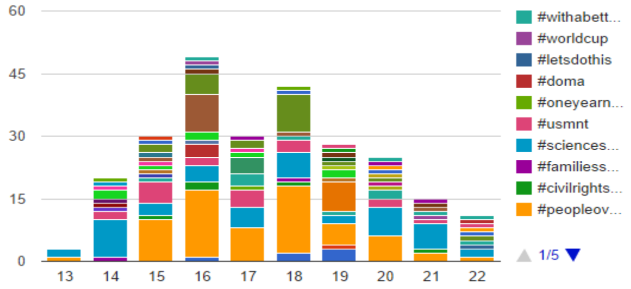


Figure 3: User-hashtag distribution in different time of day

There are some other types of visualizations where one can see the popularity of the user by observing the number of retweets and favorites his tweets get. This is a simple visualization that can give information about the user's influence in the Twitter community.

Generating these visualizations, as well as those similar to these in the keyword-orientated module does not require any special processing of the data.

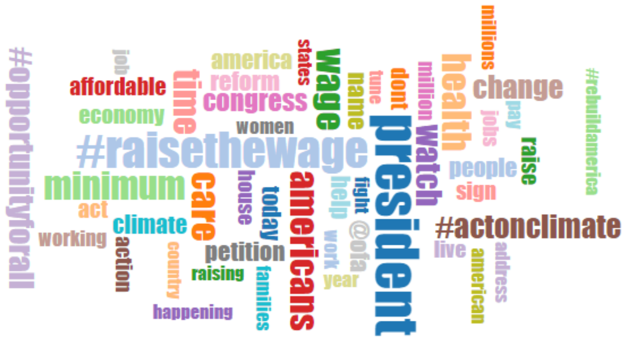


Figure 4: Word cloud

One common way of visualizing frequent keywords in any type of text analysis is creating a so called *word cloud* or *tag cloud*. Before proceeding with generating this *word cloud* (Figure 4), some preprocessing steps need to be made. Almost mandatory, when processing natural language, we need to remove stopwords from the text. Because of the specific domain, the list of stopwords needs to be extended with some specific Twitter words and abbreviations such as "RT", "retweet", "cc" etc. The rest is a simple weighting process, where more common words get larger dimensions in the *word cloud* as opposed to less frequent ones. This is a nice way of observing what a user tweets about that is not concentrated to hashtags only.

3.3 Keyword-orientated visualizations

TweetViz offers users to visualize Twitter activity surrounding a given term. They can search for a specific hashtag or any given keyword. Keyword-orientated visualizations in TweetViz are somewhat similar to the user-orientated. Users can view a chart showing the number of tweets sent containing the search term per day. This is a simple way of detecting spikes in Twitter activity about that term. There is also a chart that shows popularity of a hashtag in different times of the day, again useful for discovering patterns in activity around a search term. The *word cloud* visualization is also available when a user enters a keyword or a hashtag, contributing to a better understanding of the context surrounding the search term.

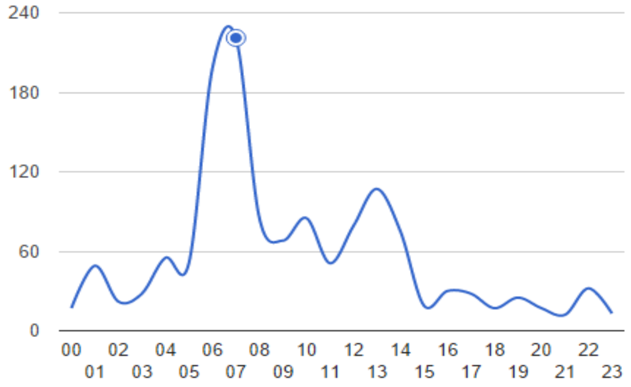


Figure 5 Temporal distribution of a keyword or hashtag

3.4 Visualizing topic distribution

Most of the approaches to analyzing data generated from Twitter revolve around simple techniques such as word count, hashtag count etc. In this paper we leverage a more advance NLP algorithm, LDA. Latent Dirichlet Allocation [5] is a three layer hierarchical Bayesian model in which each text document is modeled as a finite mixture of a set of topics. Every topic is modeled as an infinite mixture over an underlying set of topic probabilities. Simply put, a tweet is represented as a set of topics accompanied with appropriate probabilities, and each topic is made up of words with respective probability distributions. For example, a topic can be represented by a set of words ["mobile", "wear", "watch"]. When generating the models, again we preprocess the textual data, by removing stopwords and additionally, stemming the words.

This interactive visual representation of topic distribution can provide insight into how user interests change over time. An appropriate way of visualizing topic distribution in a time interval is by utilizing a Streamgraph (Figure 6). The Streamgraph visualization technique was proposed by [6] as a more aesthetic alternative to stacked graphs. A Streamgraph is consisted of a finite number of layers, each layer presenting a time series. There a lot of different aspects to be considered when creating Streamgraphs, such as algorithms for generating the graph, coloring and ordering of the layers, all of which are detailed in [6]. For the purposes of this paper, we use an implementation of a Streamgraph in d3[2], which implements the techniques suggested in [6].

In our case, each layer represents a topic, and we track user interest in the topic along the time interval. We tried with a few different color schemes, keeping in mind to use a broader

[2] http://bl.ocks.org/WillTurman/4631136

color range to better distinguish different layers. We use the *silhouette* algorithm for generating the Streamgraph.

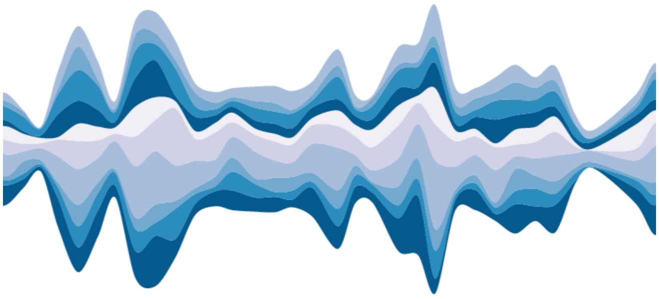
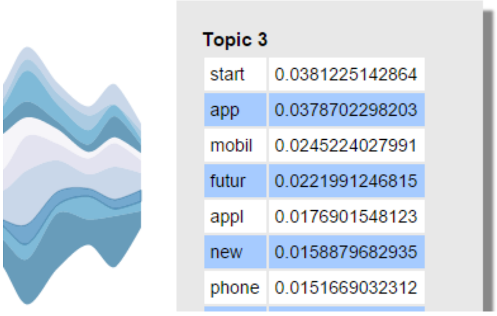


Figure 6: Streamgraph displaying topic distribution

The LDA model was trained on a corpus of over 1.4 million tweets with 3 passes so as to get a better representation of the topic distributions. One limitation of the LDA algorithm is the fact that the number of topics has to be predefined. We choose this parameter to be 20, though we only show 10 topics at a time in the visualization. We decided to use a smaller number of topics in order not to overwhelm the users in the Streamgraph visualization. [1]

The Twitter data that is presented on the Streamgraph is separated into time slices. Each time slice is consisted of a set of tweets. As a result, time slices containing more tweets will have larger y-axis values. A layer's height in a certain time interval is dependent on the presence of the related topic in the set of tweets. As was done in [3], we tend to bring topics with greater differences in distribution to the top and bottom of the Streamgraph as oppose to those with lower differences that end up in the middle. This adds to a clearer way of presenting the layers and differentiating between them.



Topic 3

start	0.0381225142864
app	0.0378702298203
mobil	0.0245224027991
futur	0.0221991246815
appl	0.0176901548123
new	0.0158879682935
phone	0.0151669032312

Figure 7 Detail information for a specific layer

This visualization offers interactivity by allowing users to hover over a layer and get some additional information about it. Users are presented with the words that the topic is consisted of and their respective probabilities as shown in Figure 7. The words appear in a stemmed form, but still, they are informative and can be used to understand what the layer is representing.

3.5 User interface

When designing the user interface, we strived for simplicity. Users are not overwhelmed with many controls and options. There is only one input control in which the user can enter a valid Twitter username, keyword or a hashtag. Then, he is presented with all of the available visualizations, which are interactive in order to improve user experience. The users can hover over certain parts of the visualizations to get additional information. Also, when clicking on those parts, users are presented with the tweets relevant to that part of the visualization. For example, in Figure 3, if a user wants to see the tweets that contain hashtag "#letsdothis" posted at 7pm, he only needs clicking on that particular part of the graph. This is a feature present in most of the other visualizations.

4 CONCLUSION

Analyzing data can be greatly simplified by visualizing it first, which is more appealing to the eye. In this paper, we present our web tool for analyzing and visualizing data generated from the micro-blogging service Twitter. TweetViz offers a set of user-orientated and keyword-orientated visualizations. We show how this web tool can be used to understand user behavior and interests from different aspects as well as general Twitter activity connected to some keyword or hashtag.

We also propose a not so well explored approach of visualizing topic distribution in a set of tweets over some time interval. Topic distributions are generated using the LDA algorithm. Our web tool TweetViz can be of use to anyone interested in exploring Twitter activity and provide for a nice visual way of analyzing data from Twitter.

Acknowledgement

We would like to acknowledge the support of the European Commission through the project MAESTRA – Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References:

[1] S. Malik, A. Smith, P. Papadatos, J. Li. TopicFlow: Visualizing Topic Alignment of Twitter Data over Time. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

[2] G. Rotta, V. de Lemos, F. Lammel, I. Manssour, M. Silveira, A. Pase. Visualization Techniques for the Analysis of Twitter Users' Behavior. ICWSM, 2013.

[3] J. Bradley, N. Fung, I. Julien, M. Malu, M. Mauriello. ViralViz: Visualizing Temporal Content Flow in Social Networks.

[4] S. Havre, B. Hetzler, L. Nowell. ThemeRiver: Visualizing Theme Changes over Time. In Proceedings of the IEEE Symposium on Information Vizualization 2000 (INFOVIS '00). IEEE Computer Society, Washington, DC, USA, 115-.

[5] D. M. Blei, A. Ng, M. I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.

[6] L. Byron, M. Wattenberg. Stacked Graphs-Geometry & Aesthetics. IEEE Trans. Vis. Comput. Graph., 14(6), 1245-1252.