

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal)

Libraries at University of Nebraska-Lincoln

10-22-2020

Tapping Twitter Data for Analyzing and Visualizing Public Sentiments on Censorship


Naveen Kumar Yadav

CFEES, Defence Research Development Organisation, New Delhi, India., naveen.yadav0074@gmail.com

Akhilesh K.S. Yadav

Centre for Library and Information Management Studies (CLIMS), Tata Institute of Social Sciences, Mumbai, Maharashtra, India., akhilesh.yadav@tiss.edu

Follow this and additional works at: <https://digitalcommons.unl.edu/libphilprac>

 Part of the [Data Science Commons](#), and the [Library and Information Science Commons](#)

Yadav, Naveen Kumar and Yadav, Akhilesh K.S., "Tapping Twitter Data for Analyzing and Visualizing Public Sentiments on Censorship" (2020). *Library Philosophy and Practice (e-journal)*. 4468.
<https://digitalcommons.unl.edu/libphilprac/4468>

Tapping Twitter Data for Analyzing and Visualizing Public Sentiments on Censorship

Naveen Kumar Yadav¹ & Akhilesh K.S. Yadav²

¹Junior Research Fellow, CFEES, Defence Research Development Organisation, New Delhi, India. Email: naveen.yadav0074@gmail.com

²Assistant Professor, Centre for Library and Information Management Studies (CLIMS), Tata Institute of Social Sciences, Mumbai, Maharashtra, India.
Email: akhilesh.yadav@tiss.edu

Abstract

The main objective of this research study is to analyse and visualize Twitter data with tags “#Censorship”. A connection was established with twitter using Twitter API, and receiving the tweets on Google Spreadsheets. Data visualization was performed using various tools such as Voyant Tools, Tableau, Google Spreadsheet and Orange in order to generate different visualizations based upon, language, geographical areas, retweets etc. The sentiment analysis was performed for the sentiments that were attached to the given set of data by the public in their respective tweets. The 23680 tweets were retrieved during the data collection time and there were 13,771 retweets out of these tweets. The most popular application for using twitter by the users was Twitter Web Client which constituted of the 33.67% (7972 tweets); the second most popular app was Android with 23.61% (5592 Tweets) and Twitter for iPhone stayed at third place with a share of 20.07% (4753 Tweets). The most frequent used Hashtags (#) in the tweets were #Twitter, #Facebook, #Google, #YouTube etc. Results show that negative tweets are enormously higher than the neutral sentiments.

Keywords: Censorship, sentiment analysis, public sentiments, text mining, data visualization, text visualization

Introduction

Social Media is not a strange topic in the present-day environment. Social Media has been a frequently used term in the recent past on various issues around the world, where lives of common people are mingled with the computer and internet. The picture that comes into our mind when we hear the term social media is of different applications such as Facebook, Twitter, WhatsApp, YouTube etc. Basically, the word describes the interaction among people or groups of people where they are able to create, exchange or share information and ideas in the web environment (Ahlqvist, Back, Halonen & Heinonen, 2008). The new generation media and networking has given rise to the Social Media. The most widely used social media applications includes WhatsApp, Facebook, Twitter, Instagram, YouTube etc. which are both computer and mobile/tablet-based. People who are well versed with them can access these sites through Internet at home, school, offices etc.

The computational treatment of opinion, sentiment and subjectivity through text is known as Sentiment Analysis. (Pang and Lee, 2008). Kharde and Sonawane (2016) defined Sentiment Analysis as, “the process of the mining of attitudes, opinions, views and emotions automatically from text, speech, tweets and database sources using the techniques of natural language processing”. It includes classifying of someone’s opinions present within the text into different sentiments such as, “positive”, “negative” or “neutral”.

It is often regarded as analysis of subjectivity and also opinion mining. There are many tasks performed under the name of sentiment analysis which includes sentiment extraction, sentiment classification, summarization of opinions etc. Sentiment Analysis mainly focusses in analyzing the sentiments, attitudes, opinions and emotions of people towards different products, individuals, topics, organizations and services.

Social Media Platforms, and especially microblogging in the recent years have enabled a sudden rise of unstructured content including opinions with regard to a variety of topics and have provided a cheap and valuable source for analysis of opinions about different organizations and individuals. It is not feasible to analyze all the data without taking assistance of automated computation. For an average person, it is very challenging to identify and summarize important information in huge amount of data (Wani & Jabin, 2018), along with this, the amount of data created presently over Social Media everyday makes it an impractical task for human processing (Bello-Organ, Jung & Camacho, 2016).

Sentiment Analysis may be performed at diverse level such as document, sentence and aspect levels. The most common approach is document level classification. It classifies the whole document by assuming that there is single entity opinion expressed in the whole document. For instance, customer reviews (Zhang & Liu, 2017). However, it becomes harsh for some documents which might include opinions with regards to diverse entities or even the applications may require classification regarding various aspects of the entity.

A lot of work has been done in the field of “Sentiment Analysis on Twitter” in the past decade or so by researchers around the globe. Park, Ko, Kim, Liu & Song (2011) and Bakliwal, et al. (2012) tried to experiment opinion mining using news stories or tweets by users respectively to identify political views and opinions. In the approach towards twitter, it was also considered if the tweet contained sarcastic remarks or not.

Tweets are short texts of length of maximum 280 characters. Every day, millions of tweets are generated on different kind of issues around the world to share and discuss the thoughts and views of the users. Sentiment analysis on Twitter has been used for several applications. Mostafa (2013) in a recent work identified 4 major areas in which sentiment analysis was used which includes “political orientation extraction, stock market predictions, product reviews and movie reviews”. The terms, ‘sentiment analysis’ and ‘opinion mining’ are used identically, where the later one is used in the area of industries. The extent of Sentiment Analysis is to ascertain the views, emotions, judgement or attitude of a person towards a particular topic of concern. The biggest problem that researchers face is that almost every document that they want to harvest is written in Natural Language Text, which is often delineated as unstructured data. Other problem that rely is the complication to distinguish between the nuanced usages of words and it also becomes extremely problematic to detect sarcasm or noisy data. Most of the approaches that are taken in order to classify a document after properly appropriating some positive/negative classifiers and then collate their results. However, in this case, Sentiment Analysis is a categorization problem in which the documents are categorized into two categories, positive and negative.

Objectives of the study

- 1) To understand the public discussions on censorship on Twitter microblogging.
- 2) To identify the locations, trends and applications used by public for frequent posts on Twitter.
- 3) To apply text mining techniques to analyze unstructured text content on the topic of censorship.
- 4) To find out the pattern and context with censorship.

Research Methods

The main motive of this phase of the research work was mainly to set up the webserver and test the Twitter API to pull in specific tweets in real time using Google Spreadsheets. A spreadsheet with tweets was created from January 5, 2018 to February 9, 2018. The tweets with the hashtag, “#Censorship”, that were later used for data analysis and visualization. A number of tools were used in order to achieve the desired results, as mentioned below: a) Twitter microblog is used to comment, respond and amplify on real time events. Microblogging became very popular since Twitter and Jaiku came into existence. b) Voyant Tools: Voyant Tools is a web-based software developed for text reading and text analysis. c) NodeXL: It is an application for network analysis on general purpose and supports network overview, discovery and exploration. d) Google Spreadsheets: Google spreadsheet is a web-based application that allows users to create, renovate and alter spreadsheets and share the data live online. The Ajax-based program can work both with Microsoft Excel and CSV (comma-separated values) file formats.

Data analysis

The total number of tweets archived during the data collection time was 23,680 and there were 13,771 Retweets. The amount of unique archived tweets was 23,081 which is used to monitor the quality of the archived data. The first tweet was retrieved on 5th January’2018 and the last recorded one was on 9th February’2018. The average tweet rate count was approximately 0.1 Tweets per minute (this is calculated after each time archive has been run). There were around 1819 IDs which were mentioned in replies.

The top Tweeterer in the entire data collection period who posted with maximum number of tweets had the username “schestowitz” and posted a total of 234 tweets. The user received mentions in 70 tweets and got 5% of the total retweets. The second in the list with 174 tweets was username “kosmofilo”. The third place was taken by username “AnonymousVideo” with 96 tweets, 19 mentions and only 1% of retweets. Username “DailyBrian” posted 88 related to censorship and received 352 mentions and 16% Retweets. The fifth top username “BrianBrownNet” with 86 Tweets, 19 mentions and the highest number of Retweets with 99% Retweets.

The figure-1 shows the number of tweets that were recorded by the server on each day. It can be seen that there has been some un-common rise in Tweets based on the Tags ‘Censorship’ on 10th and 11th of January when the number of counts raised up till 2163 in one single day which indicates that there could have been some issue/news event occurred related to ‘Censorship’ at that particular time. The issue that must have achieved so many tweets was the announcement of the controversial Bollywood movie ‘Padmavat’. The least number of tweets were recorded on 8th January with just 320 tweets.

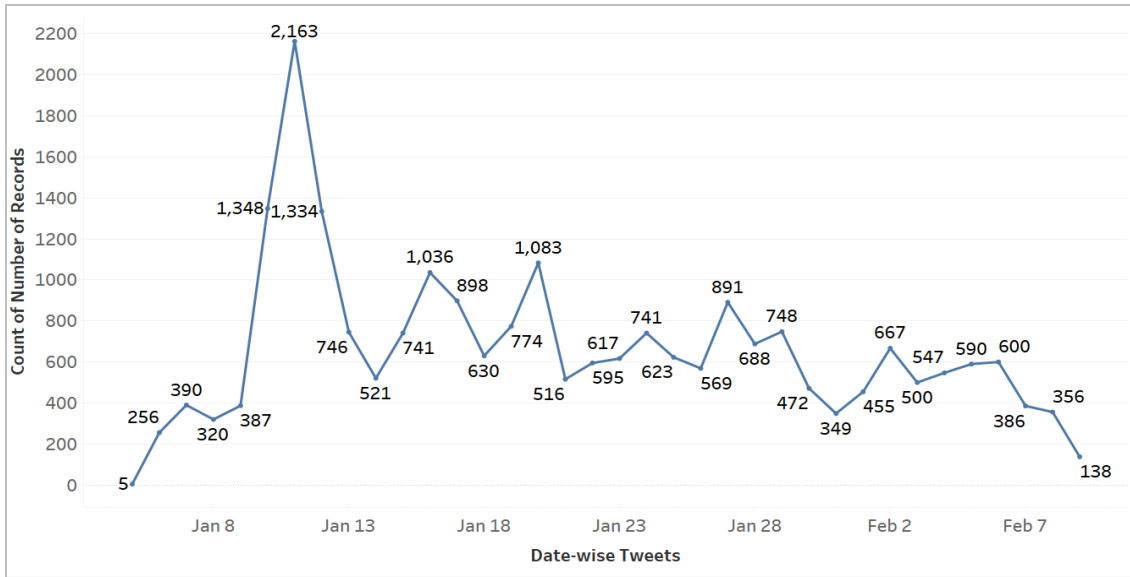


Figure 1: Twitter trend on censorship

The most used application for using twitter by the users was Twitter Web Client which constituted of the 33.67% (7972 tweets) of the total tweets posted. The second most popular app was Android with 23.61% (5592 Tweets) and Twitter for iPhone stayed at third place with a share of 20.07% (4753 Tweets). The other top application in the later order included Twitter for iPad, Twitter Lite, TweetDeck, Hootsuite etc. There were around 243 different number of applications that were used by the users while accessing their twitter accounts and posting tweets.

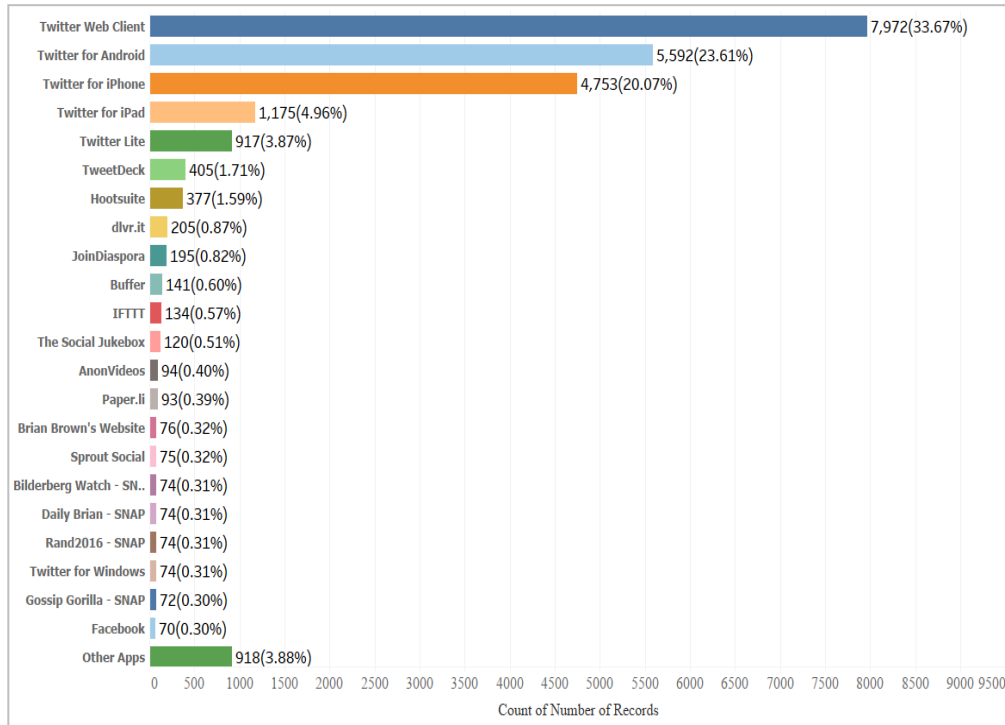


Figure 2: Applications used for tweets

The figure 3 shows the count of the most frequently used language by the users while posting tweets. As observed, the most widely used language was English in the Big Blue Bubble with a total of 21,497 tweets posted alone. Most of people use English as a language to post the content so that others are able to understand, with English being the most widely used language in the world. In order to reach the global audience, users on twitter use English language while posting their content. The second most widely used language turned out to be British English (en-gb). We know that there is a difference between the American English and the British English and the British English, mostly spoken/used by the people belonging to United Kingdom has been the second most used language for posting content. There was a little difference between the languages used on 3rd and 4th place. The other languages were the European Languages German (de) and French (Fr). The fifth language that was used for posting content by users was Spanish (es). The figure shows that European Languages somehow dominate the Twitter usage count after English.

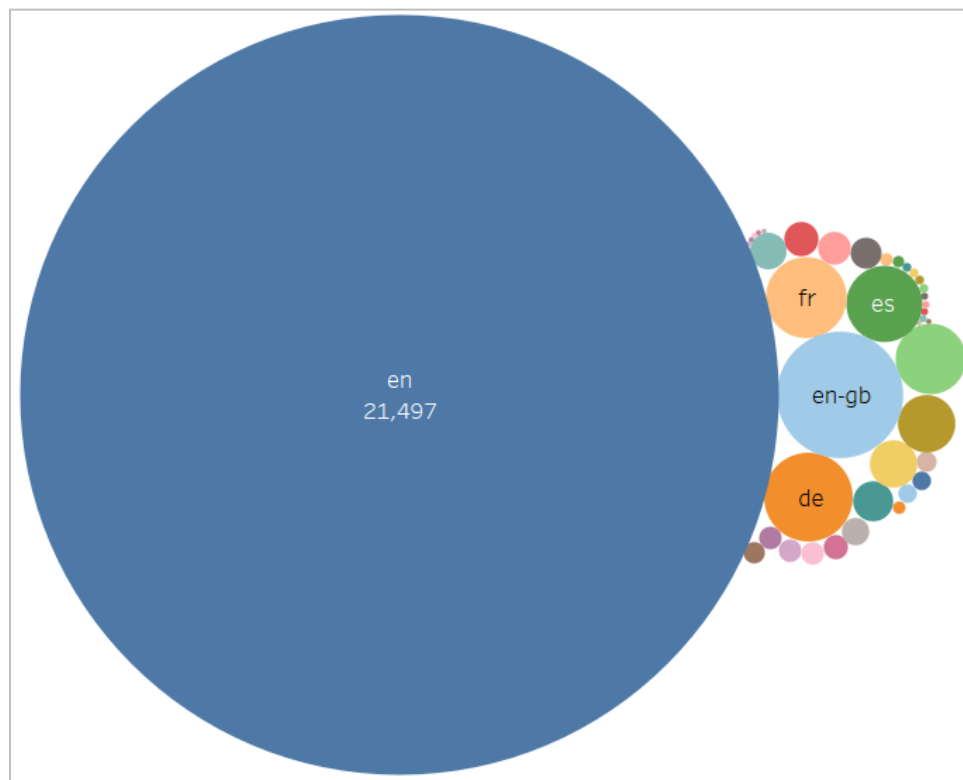


Figure 3: Bubble Graph of the Languages used in Tweets

It was found that the most of the people who tweets on #censorship are from various countries e.g. India, Australia, New Zealand, South Africa etc. The least posting countries turned out to be the South American Countries of Brazil, Paraguay, Argentina along with Canada, Some of the Middle East Nations and South East Asian Countries. Major technologically advanced countries such as United States of America and China stayed on the second to last spot in the list with an average of 2000 tweet counts.



Figure 4: User's Location

It depends on the users if they wish to allow access to their location or not on Twitter. Sometimes, this data is not clear, as it cannot be justified with many users not letting the applications to access their location. However, it can be articulated by cleaning up the data and taking out the separate data with location coordinated in it. By using Google Maps feature, the region-based information can be achieved for doing research on the location of the users. It can be observed in the figure that the most located regions by the locator is United States of America, United Kingdom and India. Along with this, there are few locations traced from Australia, Africa and South America as well. The least recorded regions turned out to be China and Russia with restrictions imposed on the users by the Government authorities for not allowing location access to the applications or may be because of users are not allowed to post anything related to censorship.

Retweets is a re-posting of a tweet. The retweet feature in twitter helps users to quickly share the Tweets with their followers, which can be from them or tweets from other users as well. Users mention RT before the tweet to state that it is a retweet. In the present day, retweets signify the importance or popularity of a particular tweet. The software collected the retweet counts within 1-2 days of the posting of the tweets. The most retweeted tweet was posted by the username “@GrrrGraphics” and received 1458 retweets. The tweet was, “What happens when left-wing authoritarians run the most powerful companies”. The tweet suggests that it is related to Politics. The second most retweeted tweet was by the same username “@GrrrGraphics” and received half of the retweets as compared to the first one and got 713 retweets. The tweet read, “#Google Runs over #FreedomOfSpeech #Censorship of #Conservatives #BenGarrison #cartoon #JamesDamore”. Username “@clivebushjd” got 507 retweets with ‘#Censorship’. The tweet read, “#Censorship Twitter Suspends Ohio #Republican Congressional Candidate- Chris Depizzo- After Tweeting About #Democrat Rival”, which again is something related to Politics in United States of America.

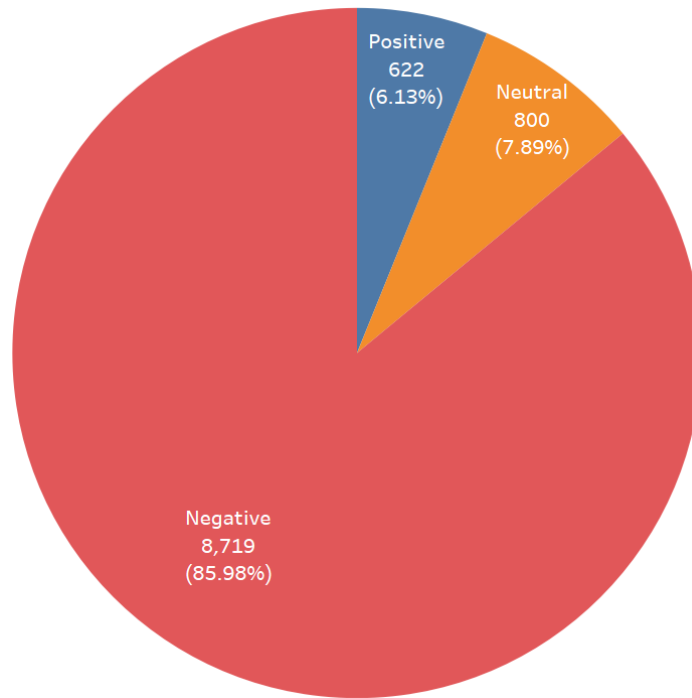


Figure 6: Sentiment Analysis

Conclusion

Millions of Tweets are created everyday by users around the world on the social networking site Twitter to contribute their ideas, opinion, facts and news online. To understand this huge amount of data present in texts through reading all or a part of it is very time taking and hectic. To utilize the human's visual interpretation and recognize easier ways for exploring the textual data, information visualization tools are developed. These tools allow the users to understand data visually on Twitter. While most of the existing tools are directed towards targeting few topics or even a specific group of people, there has been a great demand for tools that are pointed at any user who is interested in learning about the activities around the world through visualization over Twitter.

Sentiment Analysis is growing at enormous rate due to the significance of automation in the fields of mining, extracting and processing information in order to discover the general opinion of a person. Sentiment Analysis is one way by which the researchers can explore Social Media conversations along with other systematic viewpoint available for Knowledge Discovery. For these reasons only, there is a need for further research to guide the researchers in order to select and match various text analysis approaches using other social media data sources to generate pinpoint and error-free outcomes. The use of sentiment analysis among various data analysis method techniques is growing at a good pace. In this paper, we tried to investigate various approaches towards sentiment analyses and data and text visualization.

There are some limitations to this research study. Firstly, the time period that was selected for the research consisted of only one month, which led to the collection of a limited number of tweets by the researchers. Secondly, the Tags used only one term, “#Censorship” and no related terms. Censorship is a very broad topic which can be sub-divided into Book Censorship, Movie Censorship, and News Censorship etc. There are related terms also such as Ban, Censor, and Restrict etc. This study can be further enhanced by including related terms, increasing time span etc.

References

- Ahlqvist, T., Back, A., Halonen, M., & Heinonen, S. (2008). *Social Media Roadmaps: Exploring the futures triggered by social media*. Finland: Julkaisija. Available at: <http://www.vtt.fi/inf/pdf/tiedotteet/2008/t2454.pdf>
- Bakliwal, A, et al. (2012). Mining sentiments from tweets. 3rd Workshop on Sentiment and Subjectivity Analysis (WASSA) in Conjunction with 50th annual meeting of Association for Computational Linguistics (ACL) 2012. Jeju, Island, Republic of Korea. Available at: http://web2py.iit.ac.in/research_centres/publications/download/inproceedings.pdf.a2ef40d7e7aceb14.33325f50617065722e706466.pdf
- Balusamy, B., Varma, V. T. S., & Grandhi, S. S. M. Y. (2017). Social Network Web Mining: Web Mining Techniques for Online Social Network Analysis. In *Web Data Mining and the Development of Knowledge-Based Decision Support Systems* (pp. 284-310). IGI Global.
- Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, 28, 45-59.
- Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. *International Journal of Computer Applications*, 139(11), 5-15. Available at: <http://www.ijca.org/research/volume139/number11/kharde-2016-ijca-908625.pdf>
- Mostafa, M.M. (2013). More than words: Social networks’ text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10), 4241-4251. Available at: <https://doi.org/10.1016/j.eswa.2013.01.019>.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1–135. Available at: <http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>
- Park, S., Ko, M., Kim, J. Liu, Y. & Song, J. (2011). The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW '11). ACM: New York, 113-122. Available at: <https://doi.org/10.1145/1958824.1958842>
- Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36, 1-22.
- Wani, M. A., & Jabin, S. (2018). Big Data: Issues, Challenges, and Techniques in Business Intelligence. In *Big Data Analytics* (pp. 613-628). Springer, Singapore.
- Zhang L., Liu B. (2017) Sentiment Analysis and Opinion Mining. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, MA. Available at: https://doi.org/10.1007/978-1-4899-7687-1_907