Smoking Cessation Community Analysis

Julia Castrén   
Degree Programme in Computer Science and Engineer  
University of Oulu[julia.castren@student.oulu.fi](mailto:julia.castren@student.oulu.fi) Sara Lehto  
Degree Programme in Computer Science and Engineer  
University of Oulu[sara.lehto@student.oulu.fi](mailto:sara.lehto@student.oulu.fi)

*Abstract*—This paper is a project report for the course Social Network Analysis, Spring 2022. Our project topic was the analysis of smoking cessation community in Twitter.

The key task in our project was to identify two hashtags, or key words, that can be utilized to find different communities on social platforms. By these hashtags, we were able to pull tweets and other information from users with the Twitter API. We used this information in our analysis.

We found data of popular influencers in this community, mentions-data and sub-hashtags. We created graphs of the mentions-data and hashtag-data. We constructed graphs of communities inside these networks.

Keywords—social network analysis, network, communities, Twitter)

# Group Information

Our project number is 15. We worked as a team of two people. Our project title was Analysis of Smoking Cessation.

# Introduction

Twitter is a popular social media platform. Its technology is based on so-called microblogging: Users share short posts that can include text, pictures etc. [1]. Twitter is a good platform for building communities: communities can be identified via hashtags, that are keywords for the posts, users and so on. Users can follow, retweet and aswer to each other’s tweets. Users can also have private discussions via direct messages.

Twitter has its own developer application, that can be used to pull data from the social media platform. A developer account can be requested. Developers are provided with consumer and access secrets and tokens, that are utilized to acquire the needed data. Then this data can be used for the developer’s needs with programming.

Users’ interactions form connections [2]. These connections show how content is shared. Network structures

# Problem description

The aim of our project was to analyze a smoking cessation community in Twitter. This is achieved by collecting tweets with two different hashtags that can be linked to these communities. Then the acquired data is manipulated to gather information that is used for the data analysis tasks. Data collection was done via Twitter API.

# Dataset description

We were able to find around a thousand tweets using the hashtags #stopsmoking and #quitsmoking. According to our online research, these are some of the most used hashtags in the smoking cessation Twitter community [3][4]. Collecting tweets turned out to be quite challenging: We were required to collect a few thousand tweets but unfortunately, smoking cessation seems to be quite an unpopular current topic on Twitter. This can also be caused by users not using hashtags in their posts. Due to the short time window for this project, we decided to start working with this amount of data and modify our tasks to comply to our datasets. We managed to complete our tasks quite well considering these challenges.

The data was collected with Tweepy-python library that enables connection to the Twitter API. Then data was saved in a CSV-file. Our data frame had eight columns on each row that are listed below:

1. Username
2. User ID
3. Location
4. Followers
5. Original tweeter if retweet
6. Time Stamp
7. Text
8. Hashtags

Data was collected in a pandas data frame. Tweets were collected with a very straight-forward api.find\_tweets-parameter and tweepy.cursor-command.

# Methodology

The completion of our project is based on Python programming language and its libraries such as pandas, numpy and matplotlib. Networks for analysis and visualization purposes were constructed with NetworkX.

# Detailed methodology

In the project assignment, there were nine tasks that were listed separately. In this chapter, these tasks will be presented and the solutions described. Two of the tasks – identifying popular hashtags and collecting tweets – have already been described.

## Task 3

In task 3, top ten influencers in each hashtag needed to be identified. Then, average and standard deviation of the number of tweets were determined.

Top ten influencers were determined by first counting values for each username in data. Then, ten most common usernames were viewed. First figure is information that was extracted from the #stopsmoking-data frame and second from #quitsmoking-data frame. These calculations show usernames of users that have been most active within the two hashtags that were chosen for this project.

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti

Figure

Average and standard deviation is calculated with python commands and by hand as seen in next images. Figure 3 #stopsmoking and figure 4 #quitsmoking. These calculations describe statistical properties of top ten influencers that were earlier identified.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

## Task 4

In task 4, time stamps of the tweets were analyzed. That was achieved by extracting the timestamp column of our data frame. Data was cleaned to only contain the dates of tweet posts. Then a histogram was plotted for each hashtag. Histograms describe the amount of tweets posted in each hashtag per day.

Figure 5 shows extraction and cleaning of #stopsmoking-data and figure 6 shows the same operations done for #quitsmoking-data.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

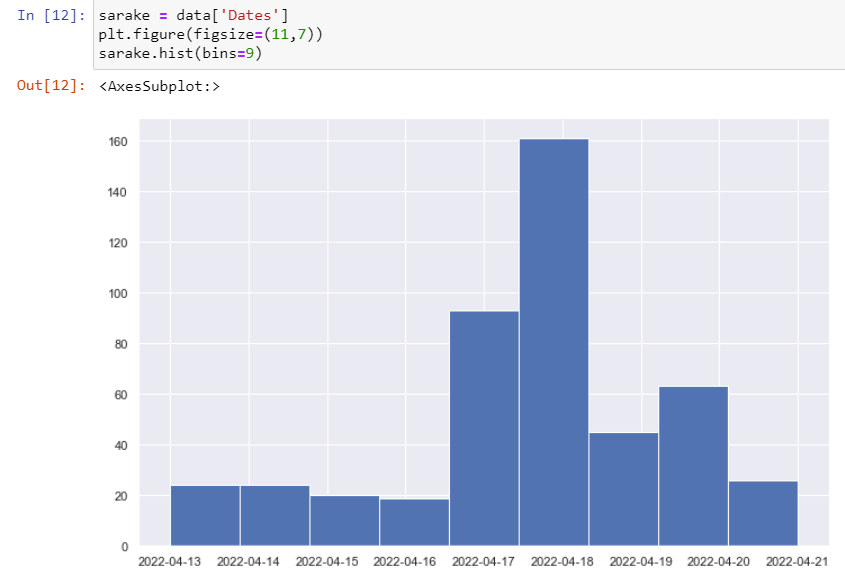
Figure

Kuva, joka sisältää kohteen teksti

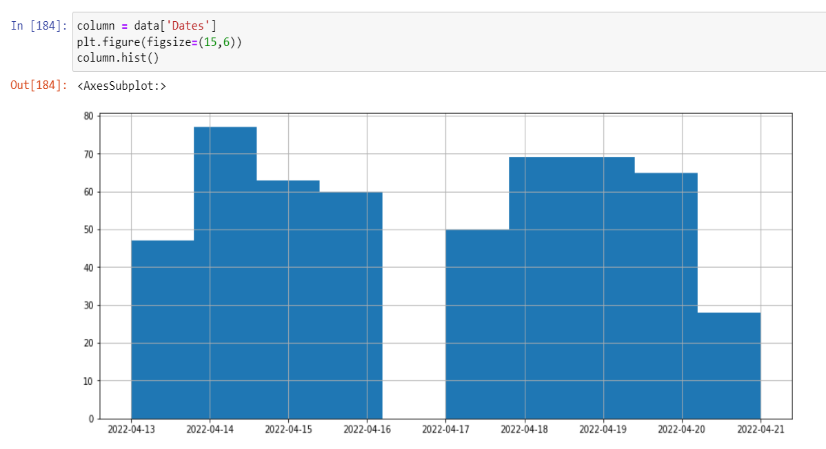
Kuvaus luotu automaattisesti

Figure

Figure 7 shows histogram for #stopsmoking-data frame and figure 8 for #quitsmoking-data frame. Histograms were plotted with daily data to limit the size of the figures to make them easier to review. Data was read as columns and plotted with the hist()-function.



Figure



Figure

From these figures we can determine, that daily variation is larger with #stopsmoking-data.

## Task 5

In task 5, assignment was to find instances of Twitter users that might be linked to pharmaceutical products or adverts. This data was extracted by identifying keywords that might be linked to such adverts or products. Keywords were identified from text column of each data frame.

After tweet texts that included these keywords were identified, usernames that posted these tweets were determined. Usernames were then printed in a list.

By reviewing this list of usernames it can be concluded that many of them are linked to some kind of product or service that is linked to health or pharmaceuticals. For example, NHS (National Health Services) has its own user [5]. Strings that were used in analysis of tweets are listed below.

1. medic
2. treatment
3. pharma
4. therapy
5. hypnosis
6. health

Strings were found with a simple find()-function. If string was found, username list was appended. Process can be seen in figures 9 (#stopsmoking) and 10 (#quitsmoking).

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

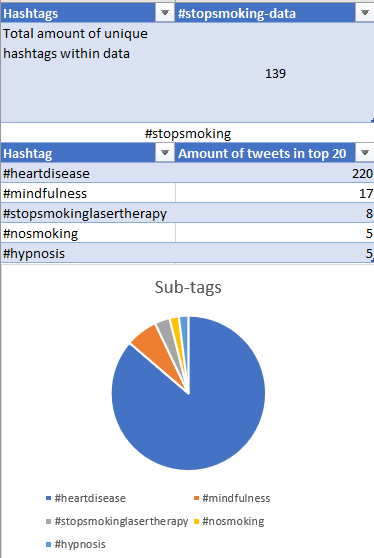
## Task 6

Assignment in task 6 was to identify sub hashtags for both data frames. A simple string search in tweet messages was completed to find these sub hashtags.

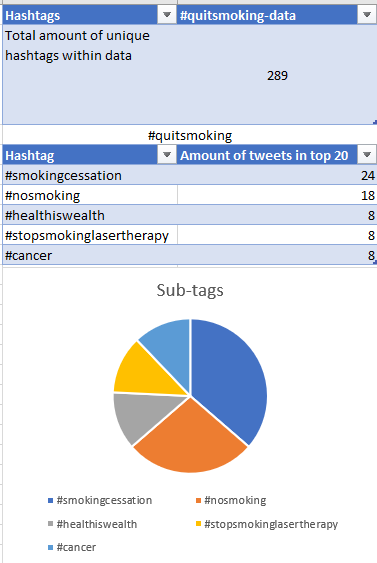
First, data column for text was searched for hashtags. It was done with string search for certain symbols with findall()-function. After that, head(20)-function was used to find twenty most common combinations of hashtags from the counted values.

A list of most common hashtags was printed. Sub hashtags were identified by reviewing this list and counting hashtags that appeared most on the top 20 list. Statistically we can conclude, that these are the most common hashtags in the whole data frame, even if the full list of hashtags is not reviewed.

Proportions for sub hashtags were summarized in tables for both cases. Tables can be seen in figure 11 and 12.



Figure



Figure

It can be concluded that identified sub hashtags differed a lot depending on the data that was reviewed. Data frame of #stopsmoking-tweets contained a very large amount of single sub hashtag. Internal differences were stronger in #quitsmoking-data.

Process of gathering data that was used to build these tables can be seen in figures 13 and 14. Data was collected from #stopsmoking-tweets.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen teksti

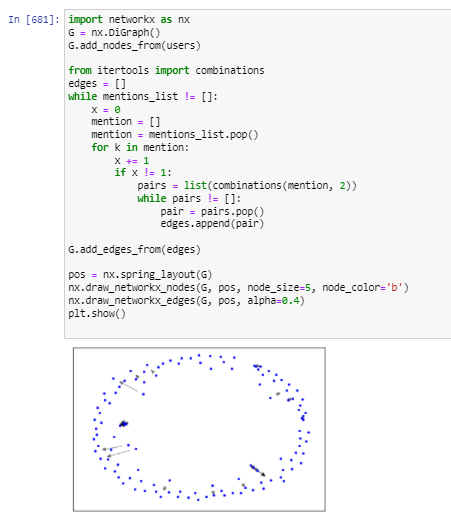
Kuvaus luotu automaattisesti

Figure

## Task 7

In task 7 so-called mentions were identified in each tweet. Mention is a string that begins with an @-sign. The code to find the mentions in the tweets was performed by using findall() function and searching for all the strings that started with the symbol @.

After the search, network graph was constructed, where the nodes correspond to user and edges between two nodes indicate that those two nodes share a mention. The code constructing the graph can be seen in figure 15.



Figure

Next the global properties of the network were determined by using simple Networkx functions. These codes are presented in figures 16 and 17. Unfortunately it was not possible to compute the diameter because the digraph is not strongly connected. The table that contains all the results in task 7 is in figure 18.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen pöytä

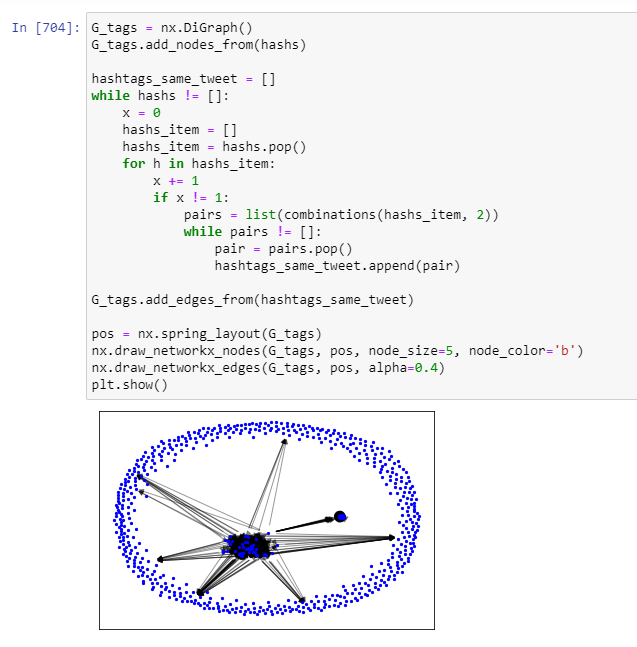
Kuvaus luotu automaattisesti

Figure

## Task 8

In task 8, content of hashtags was explored. Hashtags were determined with a similar program as in task 7 where mentions were explored with findall()-function.

In this assignment all different hashtags were identified. A network graph was constructed, where the hashtags correspond to nodes and egdes between two nodes indicate that the hashtags are contained in the same tweet message. As other graphs, it was constructed with a NetworkX-function. Process can be seen in figure 19.



Figure

The functions used to determine the average degree and number of both weakly and strongly connected components can be seen in figure 20 and 21.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure

Properties of the graph were placed in a table similarly as in task 7. The table includes number of nodes, edges and components. Average degree is also included. Table is presented in figure 22.

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti

Figure

## Task 9

In task 9, network from task 8 was analyzed. Max clique method was used for community discovery. Methods were already implemented in NetworkX. Summary of these communities and their properties are presented in tables. In figure 23, max clique communities key figures can be seen for both data sets.

Kuva, joka sisältää kohteen pöytä

Kuvaus luotu automaattisesti

Figure 23

NetworkX analysis process can be seen in figures 24 and 25. First the graph needed to be transformed into undirected graph so the community functions work. Then the largest clique was searched and made into graph and the properties were determined.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure 24

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure 25

Node amounts were found using two different NetworkX-algorithms: Label propagation and Greedy modularity. The application of these algorithms can be seen in figures 26 and 27. This analysis was done for the #quitsmoking-dataset. Same analysis was done on #stopsmoking-data. Results for label propagation was 63 nodes and for greedy modularity 18 nodes.

Kuva, joka sisältää kohteen teksti

Kuvaus luotu automaattisesti

Figure 26

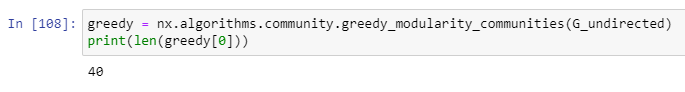


Figure 27

# Results and Discussion

Almost all of the tasks that were assigned were also solved. However, there were many challenges that had to be tackled and some were left unsolved. In this chapter those challenges will be discussed and the results more specifically evaluated. Task-specific graphs and tables were viewed earlier in the “Methodology”-chapter.

First task was to find current popular hashtags for smoking cessation. On a social media platform like Twitter, those popular keywords can change over time, even quite quickly [3][4]. This can also be the reason for variation in the amount of collected tweets. Time variance can also play a massive role in the vastness of networks that can be constructed.

The amount of tweets that were collected in this project was not ideal: communities are easier to find in larger datasets and analyses represent the true network structure more realistically. However, dataset described the current smoking cessation community somewhat well, since Twitter API enables tweet retrieval from the stream of last seven days [6]. Trends change and during the week of our project execution, smoking cessation was not a trending topic. The histogram of one of our chosen hashtag showed a lot of variation. This depicts the change of trends quite well.

Task three taught us about the top influencers within our data frames. From the results we can see that the same user is on top of the list in both hashtag-data frames. By this we can conclude that hashtags of our choice were fitting. They cover the theme of smoking cessation well.

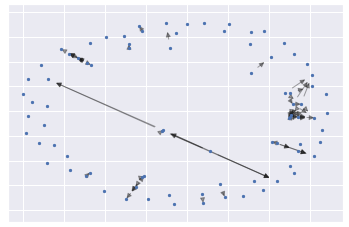
The top influencer in both cases was called “PreventiveBot.” The word “Bot” suggests that this user is a bot as well. Numbers of tweets were substantially higher than from other users in our datasets. Motive behind these bot-accounts can be to further spread information of a health-related issue like smoking cessation. Discussion of these issues can be viewed as a meaningful way of utilizing social media platforms [1].

In task 5, however, users linked to pharmaceutical products and health services were identified. By analyzing some keywords that can be linked to this industry were used to find these users and many were found.

This shows that smoking cessation communities in Twitter contain some marketing as well. Some users were linked to a public service, whose agenda is to prevent unhealthy habits for the benefit of the public service itself [5].

Task 6 results indicate that most hashtags that are linked to this community are health-related, which is expected. This seems logical also considering results in task 5.

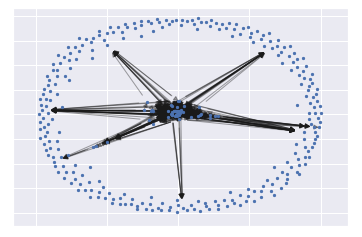
Analysis of mentions in task 7 provided more information about links between users. Links were found even in smaller datasets that were used in this project. Graph presentation of #stopsmoking-data can be seen in figure 28.



Figure

Results indicate that sense of community is important for users that are a part of this community. Research is being conducted about support groups regarding heavy smokers [7]. Studies show that it is more difficult to reach smokers of lower socio-economic groups. Social media is more accessible than ever before. Social media platforms could offer a solution for accessibility issues concerning support groups. Costs are also a challenge for both service providers and smokers. Social media platforms could be an affordable solution for providers. This could bring down costs for users and result in better achievements.

Process of analysis in task 8 was very similar as in task 7. In this task however, hashtags of tweets were analyzed. Results show that many hashtags share tweet messages. This shows a trend of using specific combinations of tags is present in the community. This also supports the claim that chosen hashtags are applicable for analyzing smoking cessation communities. Graph presentation of #stopsmoking-data is seen in figure 29.



Figure

Communities were searched in task 9. Methods for community search are already implemented in Networkx-functions [8]. Several communities were found despite the narrow datasets. Results support claims of prior tasks.

# Conclusions and Perspectives

Analyzing the smoking cessation communities in social media has provided valuable insight of a community that is strongly linked to health values. This type of study could be very applicable in the field of health science. Also marketing of pharmaceutical products or health services could benefit from this study.

Social media communities have an opportunity of preventing many health issues. Results of this project show that linking users within these communities is possible.

This study could have been conducted more efficiently with more data. As mentioned before, due to a narrow time window, collecting data was limited to tweets from the last 7 days. If time window was larger, more data could have been gathered. Larger datasets provide more insight about social networks. Studies of larger datasets have more statistical value.

In conclusion, this study was quite successful despite some challenges and lack of certain results in the tasks. We managed to provide answers to most of the assignments.

##### References

1. P. E. Walck, “Book Review, Twitter: Social Communication in the Twitter Age,” International Journal of Interactive Communication Systems and Technologies, 2013, pp. 66-69.
2. Himelboim I, Smith MA, Rainie L, Shneiderman B, Espina C. Classifying Twitter Topic-Networks Using Social Network Analysis. Social Media + Society. January 2017. doi:10.1177/2056305117691545
3. “Best #quitsmoking-hashtags,” URL: https://displaypurposes.com/hashtags/hashtag/quitsmoking
4. “Popular hashtags for smokefree on Twitter and Instagram,” URL: <https://ritetag.com/best-hashtags-for/smokefree>
5. NHS, URL: <https://www.nhs.uk/>
6. Twitter API, URL: <https://developer.twitter.com/en/products/twitter-api>
7. Thompson B, Hopp HP. Community-based programs for smoking cessation. *Clin Chest Med*. 1991;12(4):801-818.
8. NetworkX, URL: https://networkx.org/documentation/stable/reference/algorithms/community.html