Understanding User Behavior on Social Network During COVID-19: Twitter

Zhou RuiMeng
Macau University of Science and Technology
Macau
China
zhou.ruimeng@gmail.com

Gao HaoWei
University of Southern California
California
America
malago@affiliation.org

Cheng Chi
The Afflicted High-School of Hangzhou Normal University
Hangzhou
China
1624236504@qq.com

Xun XueCheng
Xuzhou No.1 Middle School
Xuzhou
China
2487128518@qq.com

Xin Ran
Wuhan Britain-China School
Wuhan
China
xinryanryan@outlook.com

Abstract. *A pandemic COVID-19 brought a huge influence among who human beings, and the lock down privacy had to be taken. Meanwhile, the number of people who are chatting on the internet has rapidly increased. So it comes to a right period to analyze users' behavior based on Twitter and to see what online surfers like to discuss. In this report, similarity matrix is first to be calculated via different users' tweets and then using the Affinity Propagation to cluster different users. Last step is to analyze why these high-frequency keywords represented the behavior on twitter during COVID-19. However, the result comes with lots of non-useful words. So we build a frequency table to see the outcome of research. It worth noting that the most frequency word is 'covid-19' which exactly focuses on the hottest topic at the moment. And the second keyword is 'realdonaldtrump', maybe it is because the selection will be held in November, so the U.S. President is appealing people to vote for him. However, we cannot analyzing to many data at same time because of the hardware facilities, but we still believe the result represents the topic in the present moment precisely.*

Keywords: *High-frequency words; Similarity matrix; Affinity Propagation; Covid-19; Twitter*

1. **Project Overview**

In December 2019, an outbreak of COVID-19 rapidly leading to a global pandemic[1]. As of right now in July, we have 215 countries affected, over 11 million confirmed, and over 500,000 deaths[2]. During crisis events, social media platforms like Twitter tend to see an increase in user activities[3], and analyzing users' behaviors during crises like the COVID-19 pandemic can make us have a better understanding of the current situation.

By looking into different users' tweets and compared relevant, high-frequency keywords, targeting to analyze and evaluate users' social network behavior on twitter during COVID-19. We applied for a twitter developer account and went through more than 13,000 tweets in our analysis. we used Affinity Propagation in python and obtained three keywords groups. Furthermore, the different frequencies of the words and emojis were counted so that their emotions and thoughts can be shown. In the end, we selected the most common words from the three groups separately and the combination as well to try to identify characteristics of different groups of clusters of people during the global pandemic. We made a Venn diagram to illustrate and we found that words like "president" "trump" "news" "black" "fight" which were consistent with current issues are common. However, there were also some popular words like "Dghisham" "glennjaconstn".

Haowei Gao and Ruimeng Zhou wrote code and wrote analysis as well as literature review and technical section writing.

Chi Cheng administered the part of Summary, while Ran Xin was responsible for the part of the Evaluation. Xuecheng Sun wrote up the part of Lessons Learned.

2. Literature Review

During the research, we took knowledge from different research papers and materials in the areas of Twitter, the clustering method, etc. In this section, we will break down the research papers we took ideas from and how these ideas helped shape our research.

2.1 Twitter

In the current social media era, people often expressed their opinions and sentiments on different social media platforms. Many researchers used twitter because the rich of opinions and experiences on the Twitter platform validate Twitter as a "real-time content, sentiment, and public attention trend tracking tool."[4] Especially during crisis events, people's tweets changed systematically in content and sentiment, which made Twitter a sense-making tool especially in the beginning and late phases of the crisis [3]. Most researchers used Twitter's developer API and tweepy to collect data from twitter [3][4][5][6]. For example, Chen, Lerman, and Ferrara followed specific keywords and trending accounts to collect data from Twitter [5]. In their research on Twitter during three crisis events in 2011, Gupta and Kumaraguru used spectral clustering and degree centrality to analyze the central users of each community and concluded that the top central people represent what the community is saying with 81% accuracy[6].

2.2 Affinity Propagation Algorithm

To find similarities among different subjects, many researchers used Affinity Propagation. Yu-lin and Zhang improved the similarity measurement of the Affinity Propagation algorithm based on the local random walk and local Naive Bayes model, which reduced

running time and had better modularity on complex networks[7]. In a different research, Zhang and Song compared several principle similarity methods like Euclidean distance, Mahalanobis distance and etc., and used Silhouette_score to show the quality of clusters made by Affinity Propagation, FCM, and K-means. Both results verified the high effectiveness and effectiveness and efficiency of Affinity Propagation[8].

2.3 User Behavior Analysis

When trying to characterize people through words they used, it was important to understand the meaning behind the word.

For the word “YouTube”, after extensive surveys amongst teenagers, Bhattacharyya analyzed and compared young people’s behavior on different social platforms. In the survey, 85% of teens said they used Youtube, topping all other platforms like Facebook, Twitter, Snapchat, and Instagram. Another survey asking teenagers which platform they visited most frequently, Youtube came in a close second at 32% while twitter only came in at less than 10 percent[9].

For the clapping emoji “👏”, Milott concluded that the clap emoji “👏” had been used by pop icons for the current generation to draw attention or emphasize an idea like an exclamation mark[10].

3. Technical Theory and Methodology

Our research process could be divided into data collection, similarity matrix, and Affinity Propagation. We first collected the data using Twitter’s developer API and tweepy. Then we constructed the similarity matrix of the users based on their twitter content, and finally,

used Affinity Propagation to cluster these users into communities and tried to analyze each group's characteristics by finding and analyzing the words that were most frequently used by each community.

3.1 Data Collection

The majority of work had focused on how to collect the dataset and analyze the similarities of communities based on different users at the same time. We used the Twitter API and tweepy to gather information about tweets that mentioned keywords that were directly related to the COVID-19 pandemic[3][4][5][6]. In order to build a dataset with diversity and cover all the appropriate time as much as possible, we chose two weekdays and one weekend on 50 occasions throughout the morning, afternoon, and evening and collected 136645 tweets during that time from 34016 users. Then, we selected the 100 users that were the most active during this time as our sample to represent the behavior of the community and collected the most 50 tweets from each user[5].

3.2 Similarity Matrix

For calculating the similarity of content and retweet, we used the library called difflib[11]. This module provided classes and functions for comparing sequences. It could be used for example, for comparing files, and could produce different information in various formats, including HTML, context, and unified diffs. The idea was to find the longest contiguous matching subsequence that contained no "junk" elements. We chose to use difflib.SequenceMatcher because it could compare any pair of sequences of any type, so

long as the sequence elements were hashable. Here is the core code of the SequenceMatcher.

```
In [6]: d = pd.read_csv('tweet2.csv',encoding='latin-1')
```

Fig 3.21

```
In [6]: # create dictionary of results
        d_out = {idx: [ SequenceMatcher(None,i, j,).ratio()] \
                       for idx, (i, j) in enumerate(permutations(d, 2))}
```

Fig 3.22

We compared the content and got all the similarities between different users. The problem was that some content contained emoji and other characters instead of ASCII, so we set Latin-1. Latin-1 is an extension of the ASCII code. It added the corresponding characters to the byte code that had not been used by the ASCII code, so it could represent more characters. For a single byte, there was nothing it could not decode, that was why it was suitable for us. In other words, when multi-byte code errors in the content were no need be taken seriously, there would be no decoding abnormalities when using the Latin-1 character set.

	G1	G2	G3	G4
G1	1	0.83	0	0
G2	0.83	1	0	0
G3	0	0	1	0.32
G4	0	0	0.32	1

Fig 3.23 [13]

The data from Twitter API contained a few correct locations like Washington DC, California, or Yellowstone National Park. But most people just located at a non-exist area or the system just showed the Latitude and longitude. It would lead to a huge inaccuracy in our algorithm thus we only used content-similarity.

3.3 Affinity Propagation

After collecting data from Twitter API and constructing a similarity matrix, the next step was to cluster all the users into different groups by using Affinity Propagation and analyze the characteristics of each cluster. We used the results in the similarity matrix as the input of Affinity Propagation. Affinity Propagation was a hierarchical clustering algorithm[7][8]. This algorithm could be used in the area of Face Image Recognition, Gene Expressed Region, and search for the shortest path.

Assume there were N data in the sample. Affinity Propagation regards all data as the cluster center at first and based on the information about the similarity of each data, every data would compete for the final cluster center. Assume $X=(x_1, x_2,x_3...x_n)$, the similarity matrix based on Euclidean Distance would be like:

$$S(i,j)=\begin{bmatrix} p & d_{12} & \cdots & d_{1N} \\ d_{21} & p & \cdots & d_{2N} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & p \end{bmatrix}$$

Fig 3.31[8]

By using this method, similar users would cluster together. Since the most active users in each cluster could effectively represent the ideas of the whole community, we needed to monitor and analyze only these top users instead of all the users in a community[6]. Thus, we showed most central people (by degree centrality) represent what users in the cluster were talking about. The reason why we chose Affinity Propagation will be mentioned in the evaluation part. In our code, we used the Sklearn.cluster.AffinityPropagation.[12]

```
# Compute Affinity Propagation
x = preprocessing.scale(x)

clf = AffinityPropagation(preference=-50)
clf.fit(x)
cluster_centers_indices = clf.cluster_centers_indices_
labels = clf.labels_

n_clusters_ = len(cluster_centers_indices)

print('Estimated number of clusters: %d' % n_clusters_)

org_df['group'] = np.nan
for i in range(len(x)):
    org_df['group'].iloc[i] = labels[i]

contents = {}
for i in range(n_clusters_):
    temp_df = org_df[org_df['group'] == float(i)]
    content_cluster = temp_df[(temp_df['Useless'] == 1)]
    content = 1.0 * len(content_cluster) / len(temp_df)
    contents[i] = content
print(contents)
```

Fig 3.32

4. Evaluation

4.1 Results

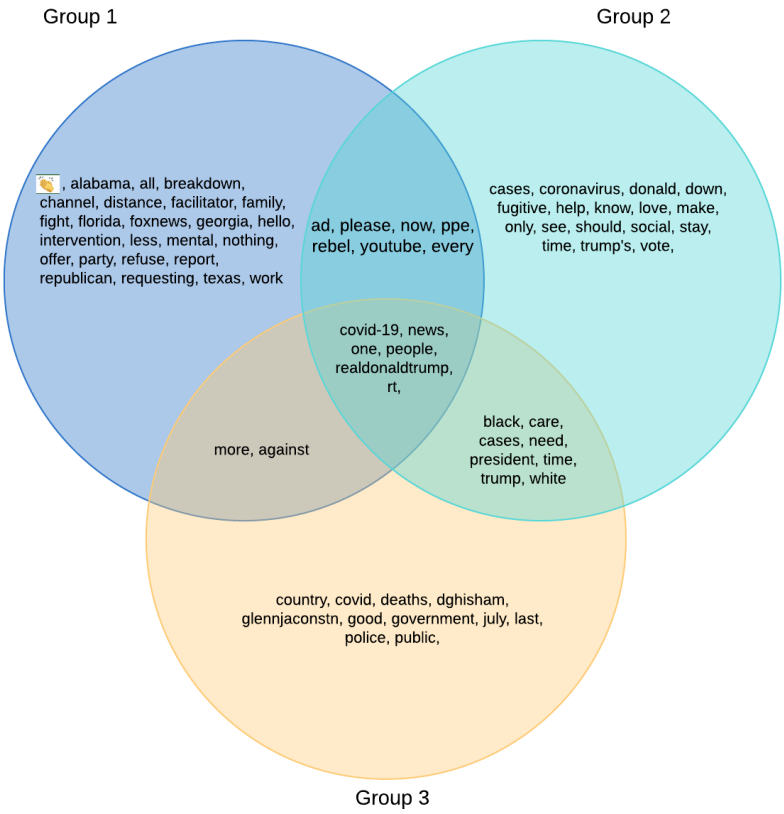

Fig 4.11

The Affinity Propagation clustered the 100 twitter users into 3 groups. Group 1 with 28 users, group 2 with 38 users, and group 3 with 34 users. We assembled the 50 tweets from each user earlier to find the most common words from each group. After deleting the words that do not have meaning like “that”, “a” etc., we organized the words that appeared

most frequently into the Venn diagram above where the top left circle represents group 1, the top right circle represents group 2, and the bottom circle representing Group 3. We are going to analyze this Venn diagram from inside out:

From the intersection part of all three groups, we can draw a few conclusions about the twitter community during the pandemic:

1. Twitter users tend to use the form of retweet instead of original tweets during the pandemic.
2. The center of attention during the pandemic is, not surprisingly, COVID-19 and the twitter handle of President Donald Trump.
3. The word “news” shows that people on twitter frequently reference and discuss news during the pandemic.

From the intersection between Group 1 and Group 2, we can draw the following conclusion about Group 1 and Group 2:

1. The word “youtube” is more likely used by the younger generation [10]
2. The words “please”, “ppe”, “now” indicates that these users care a lot about the ppe shortage that is caused by the pandemic, and how urgent the situation is.

From the intersection between Group 2 and Group 3, we can draw the following conclusion about Group 2 and Group 3:

1. The words “black” and “white” highlight the Black Lives Matter movement that was going on during the time that we collect the data.

2. The words “cases”, “need”, “care” show that these people are really concerned about the high rising case number in the United States during the time that we collected the data [2].
3. The words “president” and “trump” show that the users in Group 2 and Group3 show more attention to President Trump during the pandemic thru mentioning him in different ways.

From the intersection between Group 1 and Group 2, the only common words are “more” and “against”. We can’t really deduct anything from these words, and we may discover more shared characteristics between Group 1 and Group 2 if we collect more data in the future.

From the Group 1 part of the Venn diagram, we can draw the following conclusions about users from Group 1:

1. The clapping emoji “👏” are used by mainly younger generations to draw attention or emphasize ideas. Paired with the clapping emoji with the word “youtube” appearing in the intersection between Group 1 and Group 2, we may draw the conclusion that users in Group 1 maybe even younger than users in Group 2.
2. The words “Alabama”, “Texas”, “Georgia” are all states that are leaning towards republican than democrats. Accompanied by “Fox News”, a network reporting usually in republican’s favor and the words “republican” “party”, it shows that these younger generations really have a strong feeling towards the republican party.

3. The words “breakdown”, “fight”, “intervention”, “refuse” are all words with an aggressive connotation. It implies that the younger generation may have a stronger and more aggressive response towards the current pandemic situation.

From the Group 2 part of the Venn diagram, we can draw the following conclusions about users from Group 2:

1. Group 2 users’ continuous addressing of President Trump through using different referrals like “Donald”, “Trump’s”, “trump” and “president” shows that they are the groups that talk about President Trump the most.
2. The words “fugitive” and “stay” may refer to the situation of deporting fugitives by ICE that was trending a few months before the time we collected the data, and users in Group 2 clearly still have a strong opinion about that situation.

From the Group 3 part of the Venn diagram, we can draw the following conclusions about users from Group 3:

1. The word “dghisham” refers to the twitter handle of Noor Hisham Abdullah, who is the Director-General of Health of Malaysia, and the “glennjaconstn” refers to Glenn Jacobs, former WWE wrestler and Mayor of Knox County, Tennessee. These two people are all public figures with over 100k followers on Twitter, which means that Group 3 may have stronger opinions on public figures on Twitter.
2. The words “police”, “public” and “government” accompanied by the words “black” and “white” that show up in the intersection of group 2 and group 3 show that Group 3 may have the strongest opinion about the Black Lives Matter movement on Twitter.

While all three groups cared deeply about the pandemic and had a lot of conversations about President Trump, Users from Group 1 were probably the youngest group that may have the most aggressive response to the situation and strong opinions about the democratic party; users from Group 2 were the second youngest group with a broad focus on movements including the deportation of the fugitives and the Black Lives Matter movement; Users from Group 3 were probably the oldest group, and they had strong opinions of public figures on Twitter and were very passionate about the Black Lives Matter movement. Since these were just words out of context, it was impossible for us to capture the specific detailed characteristics of each group, but we could connect the dot of some recurring themes through the words that these groups of users used most frequently.

4.2 Technical Evaluation

For the clustering method, we chose Affinity Propagation instead of Spectral Clustering, since the Affinity Propagation didn't require the previously labeled data and we did not need to know how many groups we had to cluster[7][8]. The number of groups could be defined by several simple parameters. If we use d Spectral Clustering or K-means, it was required to calculate extra degree centrality manually or to know the k value beforehand. In this way, situations of being more complex and getting the wrong division of groups were avoided. However, Affinity Propagation was weak in its complexity of algorithms.

While designing and testing the algorithm, it tended to consume a long time to get results. In the following figure, writers provided the log function on the value of time and it was clear that Affinity Propagation would cost much more time when the sample is

larger than 500 to 1000.

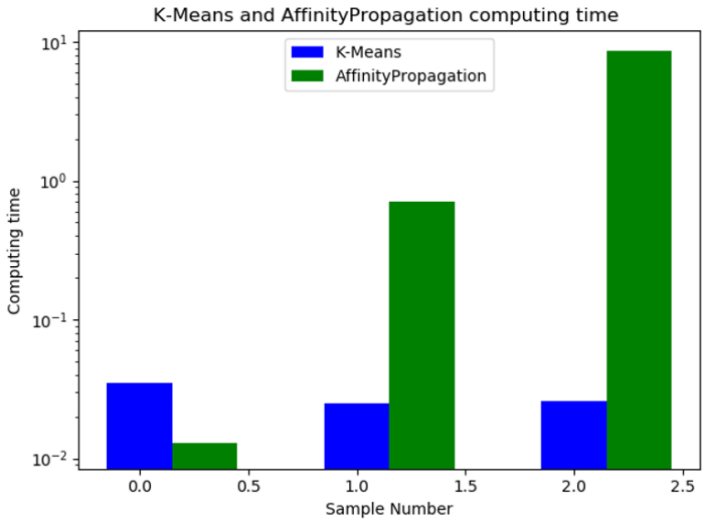

Fig 4.21[14]

Since there was no dataset pre-processed by other institutions or studies, we needed to process and fetch data through APIs in python. The predefined function in the library improved our accuracy in finding similarity significantly since it was a well-tested algorithm. However, a built-in API was used, which was called SequenceMatcher in pandas, to provide analytics in similarity.

It had low efficiency while processing numerous amounts of data. If Euclidean Distance or Cosine Similarity methods were adapted, the efficiency of execution and accuracy of results would be improved. Moreover, there was a loop comparison involved in similar datasets, which would cause one tweet to compare with itself. However, the irregular value was taken out by the methods described later.

For repeating and inappropriate results, an algorithm was performed to delete irregular values such as pairings with similarities which approaches “1” (means nearly exactly the same). Then permutation method was adapted instead of combination to find similarities

of every two tweets in order to increase accuracy.

For results storage, originally, both raw data and process results were stored in the same file, but it caused abundance in file size and the rate of next-step processing was significantly slow; as a result, it was suggested to store processed value in the processed file only, which both improved efficiency, and reduced file size.

However, there were some problems presented. Due to the limitations of Twitter's Developer API, there was no opportunity to obtain exactly complete sentences from the users' tweets. So when the raw data was replaced with a more accurate dataset, the accuracy of the algorithm presented would get improved. As the dataset was limited in the number of tweets, using more datasets could cause more accurate results.

For other practical uses, the algorithm could also be used to perform similar analysis in not only user behavior on other platforms, but also on finding regular patterns of numerous datasets.

5. Administrative

We were fortunate to have a harmonious team and conducted the research efficiently.

Collect a good dataset with the help of the professor	40 hours
Two experiments design totally cost	56 hours
Data Processing	28 hours
Data analysis	8 hours
Report write-up will cost	20 hours

Table 1.

Estimation time cost for each procedure

For data collection, it must be the most difficult part of our group. Unlike other groups, we had no available dataset which could be used directly. We had to apply for the Twitter API first with a long period of approval. Luckily, Haowei Gao had successfully been approved. Then he began to write code to get the dataset. And the first dataset was not perfect at all, there were several problems like incomplete content, the location was longitude and latitude, not the region area, and so on. But we overcame them and finally got a clean dataset.

For building models and processing the dataset, it was not quite easy at all. Ruimeng Zhou considered it easier to use a system called SequenceMatcher, but it cost a lot of time. Then we decided to optimize and cut some of the useless output to make it run faster. Even so, the result of the dataset was still nearly 10G. With the help of a team member's computer, we had the processed dataset.

For clustering, we tried Affinity Propagation instead of Spectral Clustering, which helped us to get the final result more straightforward and clearer. At last, we got a beautiful result and conclusion.

The result analysis was mainly done by Haowei Gao and Ruimeng Zhou where Haowei Gao picked out the most frequently used words by top users and each group and provided analysis on each separate community. The final report and presentation were finished by all of us.

6. Lessons Learned

During this project, all the members in our group learned lots of research skills such as finding high-quality papers, collecting datasets, and so on. Our professor helped us with basic python skills and jupyter notebook because these were the main tools we used in the research. We also learned a lot of expertise about Machine Learning and mathematics such as k-means, Affinity Propagation, and Naïve Bayes Classifier, which we may put in use in the future. It was worth mentioning that there were three high school students in our group, this research experience was also a great chance for them to learn how to write a proposal and a report and how to collaborate in a team with different people at different ages.

There are several ways we can improve in the future:

a. We went through more than 5,000 tweets in our analysis, it was a very small number compared to the number of tweets on the platform. With the data collecting code already set up, we can collect and analyze more data.
b. We can try other methods of clustering like spectral clustering or k-medoids clustering from different packages like scipy or Matlab to cross-validate and obtain a more accurate
c. Although twitter is a huge platform, it will be very helpful if we can obtain data from other social media platforms like Facebook, Instagram, etc. To cross-validate across platforms, so we may potentially have a better understanding of the user's behavior.

7. Summary

In this project, we analyzed and evaluated user behavior on social networks on Twitter during COVID-19. We used a dataset based on Twitter API to construct the content similarity matrix, then applied the Affinity Propagation to aggregate tweets. When analyzing the characteristics of clusters, we only look at the top and center part of the people, which could represent the whole community with high accuracy, to get an understanding of the whole group according to one of the papers we referenced. The most important reason we chose the Affinity Propagation was that it could help us omit the procedure of calculating the central degree and use the similarity matrix directly. We do it with python and yield three keywords groups. By comparing the similar contents between the three keywords groups obtained, we got some interesting results that different types or generations of people actually have different reactions and statements towards the pandemic. Finally, if possible, we will keep optimizing our structure of the dataset and algorithms code, which can enable less time consuming and improve the efficiency of processing data.

8. Acknowledge

We would like to appreciate Professor Nick Feamster who give us advice about giving theoretical support and the CIS team which gave us a chance to improve ourselves, especially the capability of doing research.

9. Reference

1. Polak, S., Van Gool, I., Cohen, D., von der Thüsen, J. and van Paassen, J. (2020). A systematic review of pathological findings in COVID-19: a pathophysiological timeline and possible mechanisms of disease progression. *Mod Pathol*.
2. WHO, (2020). *WHO Coronavirus Disease (COVID-19) Dashboard*. [online] Covid19.who.int. Available at: <https://covid19.who.int/> [Accessed 10 July 2020].
3. Gascó, M., Bayerl, P., Denef, S. and Akhgar, B. (2017). What do citizens communicate about during crises? Analyzing twitter use during the 2011 UK riots. *Government Information Quarterly*, 34(4), pp.635-645.
4. Chew, C. and Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5(11), p.e14118.
5. Chen, E., Lerman, K. and Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, 6(2), p.e19273.
6. Gupta, A., Joshi, A. and Kumaraguru, P. (2012). Identifying and characterizing user communities on Twitter during crisis events. *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media - DUBMMSM '12*, pp.23-26.

7. Yu-Ling, H., & Zhang. (2016). Research on Affinity Propagation algorithm based on common neighbors. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. doi:10.1109/smc.2016.7844776
8. Zhang, H., & Song, K. (2011). Research and experiment on Affinity Propagation clustering algorithm. *2011 Second International Conference on Mechanic Automation and Control Engineering*. doi:10.1109/mace.2011.5988401
9. Bhattacharyya, M. (2020). *Young Users Favor Youtube &Amp; Snapchat, Should Facebook Bother?*. [online] Zacks Investment Research. Available at: <https://www.zacks.com/stock/news/306329/young-users-favor-youtube-amp-snapchat-should-facebook-bother> [Accessed 10 July 2020].
10. Milott, P. (2020). *Emojis And Emoticons In Court 44 Reporter (The) 2017*. [online] Heinonline.org. Available at: <https://heinonline.org/HOL/LandingPage?handle=hein.journals/report44&div=32&id=&page=> [Accessed 10 July 2020].
11. Storchaka, S. (2019, November 12). Python/cpython. Retrieved July 12, 2020, from https://github.com/python/cpython/blob/3.8/Lib/difflib.py
12. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
13. Zarpelão, B. (2020, June 25). Figure 7. Example of similarity matrix. . Retrieved July 12, 2020, from https://www.researchgate.net/figure/Example-of-similarity-matrix_fig3_315628119
14. *聚类算法 Affinity Propagation(AP) | 数据常青藤*. (2015, May 19). 数据常青

藤.https://www.dataivy.cn/blog/%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95affinity-propagation_ap/