Web Science Coursework Report

Tawan Thampipattanakul

24251421T

Repo Link

Data Link

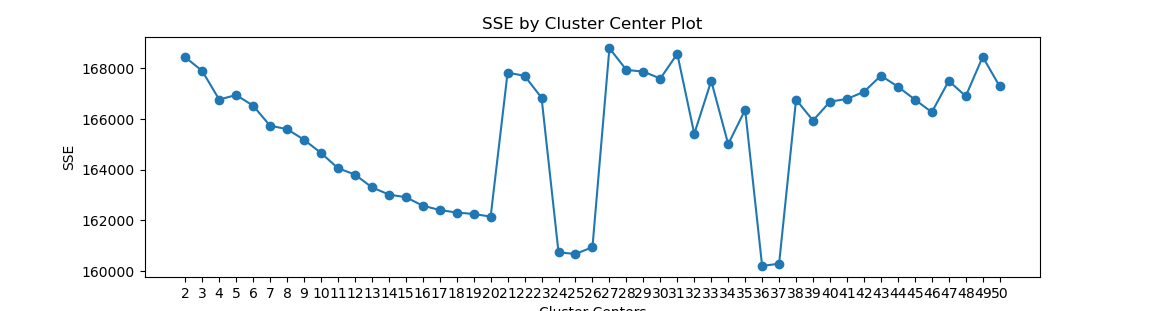
Section 1 : Introduction

1. <describe the software> <specify which code part was adopted from else where> The crawler and data analysis software were designed using hybrid architecture utilizing both Stream and REST Twitter’s API
2. The Stream API and REST API crawler scripts were run parallelly for 3 hours on Saturday night of 29th February 2019

Section 2 : Data Crawl

1. Use Twitter Streaming API for collecting 1% data
   1. <code appendix>
   2. Basic filter Streaming API was used to crawl realtime data on specific topics [“corona virus”, “Brexit”, “sonic movie”, “stock price”]. Initially I had chosen to get all realtime tweets without filtering any topics but then data grouping script in section 3 could not meaningfully cluster tweets with considerate level of square sum error. I believe that was because all tweets from everywhere were too sparse. That was why I later decided to scope down to only interesting topics. Those topics are what I thought topical at the time so I could gather more tweets compared to outdated topics. The retrieved data were stored in a Mongo DB collection using tweet IDs as unique indices so there would not be any duplicate documents in the document preserving storage usage and increasing analysis quality in later sections. There is no restriction for using Streaming API that the code needed to conform but it had limitation of 1% of data that is accessible.
2. Enhance the crawling using the architecture of Twitter Streaming & REST API
   1. <rest crawler appendix>
   2. The addition crawler loads tweet documents from Mongo DB collection and then analyses which users were most mentioned in the crawled tweets. Then, it uses the user\_timeline REST API to gather more tweets from those users repetitively with newly crawled tweets taken into account. Since the has usage limit of 1500 requests per 15 minutes window, the script gives 0.67 seconds sleeping interval between each request guaranteeing that the limit of roughly 1350 request in a window. I intentionally left some quota buffer for unexpected reasons.

Section 3 : Grouping of tweets: Group the tweets based on content analysis, You can collect the data and then cluster them using any off-the-shelf software; or use any locality sensitive hashing software; or build a content index and group them

1. I have thought about several ways to cluster the crawled tweets including sentiment analysis, K-Mean of geolocation, K-Mean of vectorized tweets’ content and N-bins hashing. First of all, randomly hashing would not create much semantic relation nor user relation among a group while sentiment analysis can only produce limited number of groups so I put away those two ideas. Clustering geolocations is interesting but very low proportion of tweets has geolocation enabled. Therefore, I finally decided to cluster texts of tweets using TFIDF vectorization and K-Mean clustering because it should yield decent number of clusters(groups) and it should be straight forward to look at texts of a same cluster and verify the result.  
     
   Please be noted that tweets of retweets was filtered out before the process started to reduce great amount of duplications which would alter the grouping result.  
     
   In brief details, TFIDF vectorization transforms texts into points in a vector space of |Vocab| dimensions. Then K-Mean algorithm tries to find points of similar semantic meaning. K-Mean starts with randomly generate K points in vector space and put vector points near to the same pivot as the same cluster. Average point of each cluster will be estimated and treated as a new pivot of the cluster. The process continues repetitively until the sum of distances from each vector point to its pivot (Square Sum Error or SSE) converged. At the final stage, each cluster represent a group with similar text meaning. The value K must be specified before the algorithm started and different K values will lead to different grouping result and SSE. To choose a proper K, elbow method was used to identify which amount of K best reduce SSE without splitting a meaningful group into multiple clusters. In the case of this coursework K=20 could be the ideal value but K=8 were chosen because it would be easier to analyse.  
     
   

Overall statistics are as followed.

|  |  |  |  |
| --- | --- | --- | --- |
| Num of Groups | Max Size | Min Size | Average Size |
| 20 | 129056 | 971 | 21992 |

Numerical statistics for general group and clustered group (retweet excluded for clustered groups)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Size | Unique Posters | Unique Hashtags | Unique Mentions |
| General | 207269 | 112100 | 20648 | 75034 |
| 0 | 4125 | 3689 | 239 | 2169 |
| 1 | 129056 | 67594 | 11388 | 58891 |
| 2 | 2599 | 2341 | 839 | 1198 |
| 3 | 22627 | 17263 | 8921 | 9365 |
| 4 | 8532 | 7303 | 860 | 4441 |
| 5 | 3570 | 3241 | 443 | 2643 |
| 6 | 4455 | 4044 | 335 | 2521 |
| 7 | 971 | 908 | 162 | 155 |

1. After groups were identified, owning-users/hashtags/mentioned-users of tweets belonging to a group was extracted using the script <appendix to script> by simply accessing [status->user->name], [status->entities->hashtags], and [status->entities->user\_mentions] values of tweets. Frequencies of them were analyzed and the result statistics of all tweets and each groups are as follow.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Size | Important Posters | Important Hashtags | Important Mentions |
| General | 207269 | ('Marvin R. Jeffcoat', 234)  ('.', 221)  ('Chris', 130)  ('quis custodiet', 117)  ('Julie.TrumpsGirl 🇺🇸🇺🇸🇺🇸🇺🇸🇺🇸', 109) | ('AEWRevolution', 894)  ('coronavirus', 249)  ('AEW', 219)  ('SCprimary2020', 200)  ('UFCNorfolk', 198) | ('TySne B. Smith', 10141)  ('Donald J. Trump', 1830)  ('Cristiano', 1349)  ('perneeya', 1025)  ('CJ McCollum', 819) |
| 0 | 4125 | ('My Style', 17)  ('Mawufeasi', 11)  ('Andrew Wong', 10)  ('Daniel B. Orijuela', 10)  ('PeterSweden', 8) | ('coronavirus', 30)  ('Corona', 28)  ('CoronaVirusUpdate', 19)  ('corona', 13)  ('COVID19', 11) | ('Donald J. Trump', 280)  ('Nancy Pelosi', 52)  ('Dana Milbank', 31)  ('President Trump', 24)  ('YouTube', 20) |
| 1 | 129056 | ('.', 133)  ('quis custodiet', 115)  ('Chris', 83)  ('Ankhesenamun 🌊⚛⚜🇸🇪', 74)  ('Portland Police log', 72) | ('AEWRevolution', 660)  ('UFCNorfolk', 157)  ('AEW', 157)  ('SCprimary2020', 153)  ('SouthCarolinaPrimary', 138) | ('Donald J. Trump', 894)  ('Joe Biden (Text Join to 30330)', 607)  ('Bernie Sanders', 386)  ('Elizabeth Warren', 258)  ('Tom Steyer', 141) |
| 2 | 2599 | ('BoyBandStan⁷', 20)  ('Dr.Bones', 18)  ('gallina fiedler', 7)  ('Terry Dinan', 7)  ('Mark Hansen', 7) | ('ON', 22)  ('Sia', 20)  ('NMS', 18)  ('coronavirus', 12)  ('ctbb', 7) | ('방탄소년단', 23)  ('Most Requested Live', 22)  ('Radio Carson PNW', 20)  ('Donald J. Trump', 14)  ('YouTube', 6) |
| 3 | 22627 | ('brandon 🌹🔭', 61)  ('Tickeron', 46)  ('All Elite Wrestling', 26)  ('Mystery Solvent', 20)  ('Marvin R. Jeffcoat', 19) | ('AEWRevolution', 129)  ('MyTwitterAnniversary', 87)  ('AEW', 35)  ('SCprimary2020', 32)  ('LeapDay', 29) | ('YouTube', 144)  ('Bernie Sanders', 92)  ('Donald J. Trump', 72)  ('Joe Biden (Text Join to 30330)', 51)  ('Nashville SC’s MLS Debut is TONIGHT', 33) |
| 4 | 8532 | ('BrexitBot', 14)  ('.', 11)  ('Taylor Rae', 9)  ('💎Diamond Mami 💎', 8)  ('Gouthama Venkata Ramana Raju Chekuri', 8) | ('Endomondo', 75)  ('endorphins', 69)  ('AEWRevolution', 40)  ('WoodenAward', 27)  ('SCprimary2020', 11) | ('Joe Biden (Text Join to 30330)', 55)  ('Donald J. Trump', 53)  ("Wendy's", 28)  ('Bernie Sanders', 26)  ('Elizabeth Warren', 15) |
| 5 | 3570 | ('surendra shetty', 14)  ('Muhammad Farid', 11)  ('Mike', 8)  ('💙✨Traylonda\U0001f9e1✨', 8)  ('Chris', 7) | ('AEWRevolution', 25)  ('UFCNorfolk', 8)  ('SouthCarolinaPrimary', 4)  ('photography', 3)  ('TheInvisbleMan', 3) | ('Joe Biden (Text Join to 30330)', 13)  ('Donald J. Trump', 12)  ('Nashville SC’s MLS Debut is TONIGHT', 6)  ('St. Louis BattleHawks', 5)  ('Elizabeth Warren', 5) |
| 6 | 4455 | ('Chelle Belle ❤ EU🇮🇪🇪🇺', 16)  ('🦋', 5)  ('Pussy fairy \U0001f9da🏽\u200d♀️', 5)  ('Simon Clancy', 4)  ('.', 4) | ('AEWRevolution', 19)  ('loveisblindnetflix', 4)  ('loveisblind', 4)  ('SouthCarolinaPrimary', 3)  ('AEW', 3) | ('Donald J. Trump', 33)  ('Joe Biden (Text Join to 30330)', 19)  ('Elizabeth Warren', 12)  ('Tom Steyer', 12)  ('Mike Bloomberg', 9) |
| 7 | 971 | ('River Levels UK', 27)  ('Tim King', 6)  ('𝘚𝘢𝘴𝘴𝘺𝘴𝘩𝘦𝘳𝘪𝘯𝘢💋', 4)  ('Phil Strickland', 4)  ('☾', 3) | ('March', 7)  ('birthmonth', 3)  ('BirthMonth', 2)  ('minggupagi', 2)  ('Matchmaking', 1) | ('Jon Rothstein', 3)  ('Vince Mathews', 2)  ('Duke Men’s Basketball', 2)  ('LANY', 1)  ('lauv', 1) |

Section 4 : Describe the metod for Capturing & Organising User and hashtag information. Objective is to build a user interactive graph.

1. <code appendix>

TODO In Code

* [Done] Statistic of groups (task 2 c)
* Users occurring together in general and in each group
* occuring together graph
* retweet graph
* quote reply graph