Problem Set 1

Text Mining: Models and Algorithms

Björn Komander ∘ Mathias Schindler ∘ Cristina Susanu ∘ Simón Caicedo

Data Science for Decision Making, Class of '22
Barcelona School of Economics

January 17, 2021

---

1. Queries Using `twarc2`

Exercise 1

The problem with this request is the date format, it should be in the format "yyyy-mm-dd" and not "yyyy-dd-mm" as in the exercise. Also, there are many spaces where they should not be. In the same way, the `archive` option is missing, as search alone only retrieves tweets from the last 7 days of activity. The correct form would be:

```
twarc2 search --archive --start-time "2020-05-25" --end-time "2020-12-31" "(#
    georgefloyd OR #justiceforgeorgefloyd OR #ICantBreathe OR #icantbreathe OR
     #blm OR #BLM OR #BlackLivesMatter OR #blacklivesmatter) lang:en -is:
    retweet" Exercise1.jsonl
```

Exercise 2

There are 2 ways of knowing in advance how many tweets will be produced, that is by using the command searches followed by counts (for queries as txt files) or directly the counts command.
A query for obtaining in advance the amount of tweets that will be produced in Exercise 1 is:

```
twarc2 counts --archive --csv --start-time "2020-05-25" --end-time
    "2020-12-31" "(#georgefloyd OR #justiceforgeorgefloyd OR #ICantBreathe OR
    #icantbreathe OR #blm OR #BLM OR #BlackLivesMatter OR #blacklivesmatter)
    lang:en -is:retweet" Exercise2.csv
```

After running this query we will obtain a *.csv* file, which can be uploaded in R to sum the column *hour_count* in the following way:

```
counts_df <- read.csv('Hasthags.csv')
sum(counts_df$hour_count)
```

This search results in 8,516,020 tweets. The Figure 1 presents the series of the number of tweets obtained from the request in this exercise, while the Figure 2 shows the number of tweets per hour. From the plots, we can clearly see that the most quantity of tweets was produced during May and June, and that the users were more active around this topic on Twitter between 3 p.m. to 7 p.m.

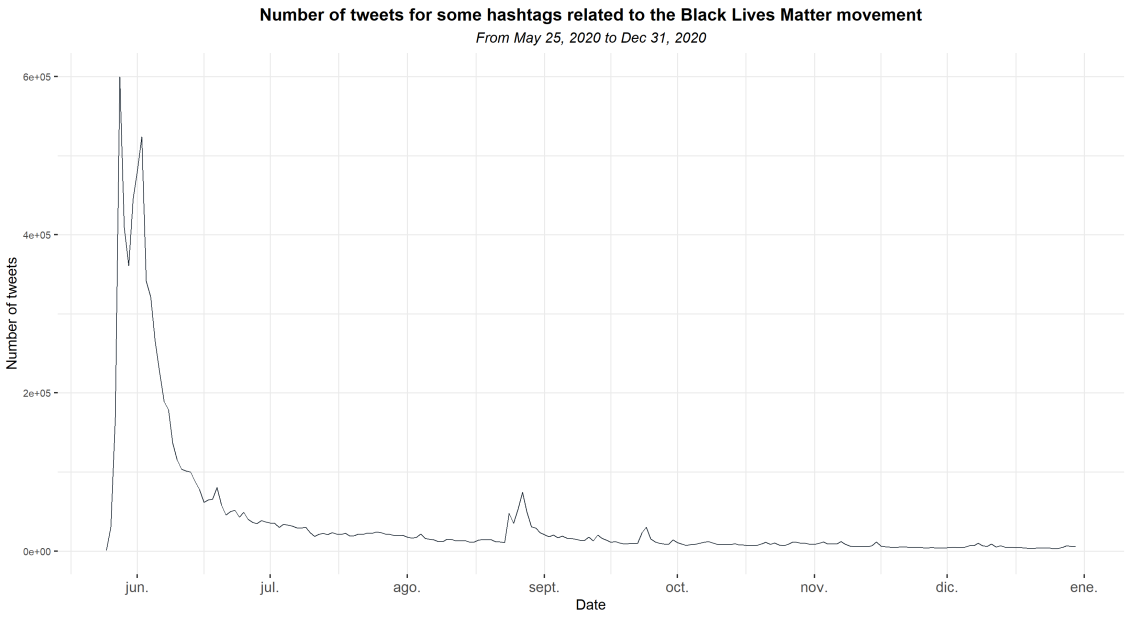

Figure 1: Series of Number of Tweets Related to *Black Lives Matter* Movement

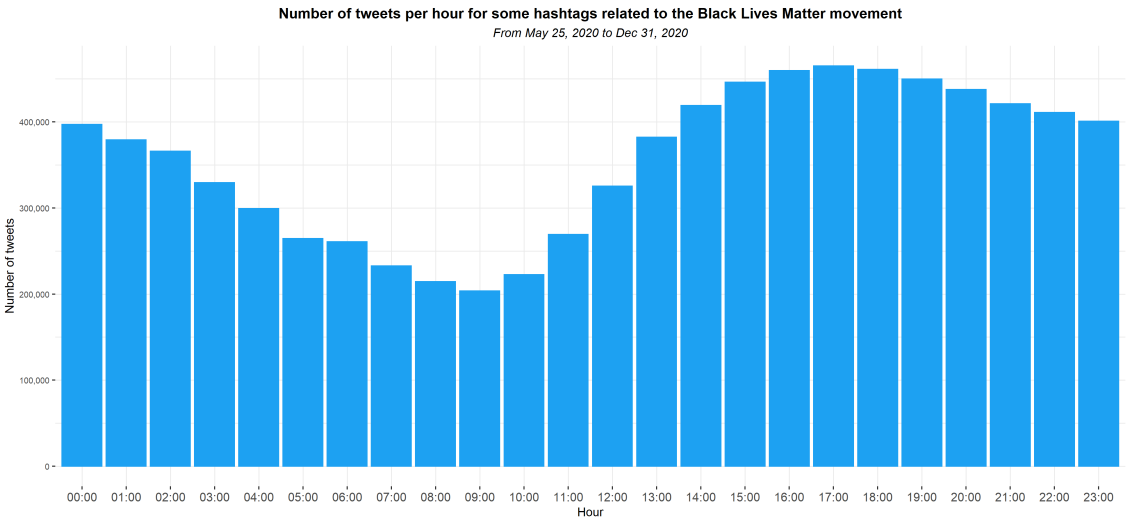

Figure 2: Number of Tweets per Hour Related to *Black Lives Matter* Movement

Exercise 3

The problem with this request is that `--archive` is not included and the coordinates for Miami Beach are not the best, because drawing a circle with the 3.11 radius we find an area that is mainly located in the sea (see Figure 3). Also, when using the point radius one has to switch the order of longitude and latitude coordinates

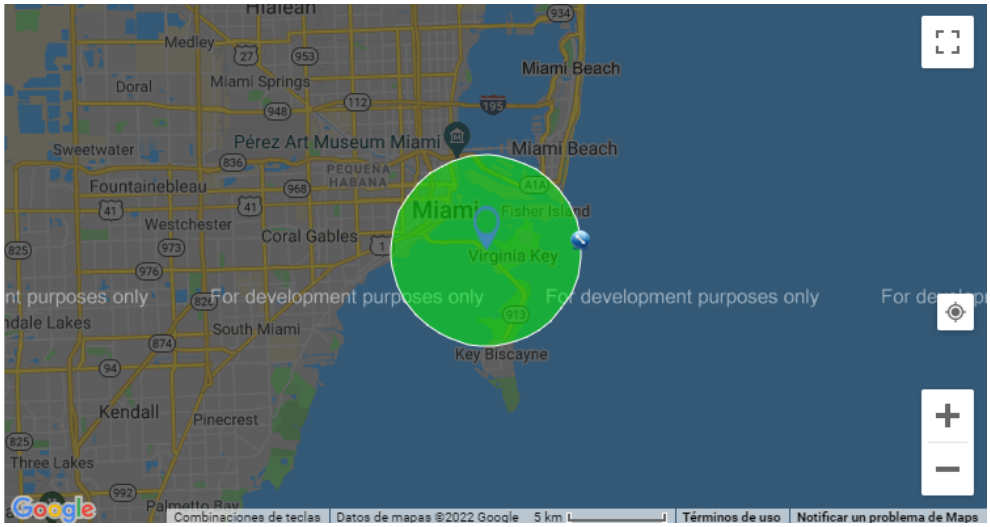

Figure 3: Coordinates Location with 3.11 Radius Circle

Exercise 4

```
twarc2 search --archive --start-time "2019-03-29" --end-time "2019-03-31" "
    place":"Miami Beach" > miami.json
```

Or another option is to use bounding box

```
twarc2 search --archive --start-time "2019-03-29" --end-time "2019-03-31"
    bounding_box: []" > miami.json
```

So just by construction, a rectangle will certainly give a different amount of tweets than a circle-similar for the place search where the spatial boundaries seems to be quite flexible (in terms of geometric shapes).

Exercise 5

Note that by default duplicates are excluded for the search command.

```
twarc2 search --limit 10 --archive --start-time "2019-03-29" --end-time
    "2019-03-31" "point_radius :[-80.1353006 25.7929198 3.11mi]" > miami.json

pip install twarc-csv
```

Note that csv command has merge-retweets option set as Default.

```
twarc2 csv miami.json tweets.csv
```

Exercise 6

The output is empty as Donald Trump's Twitter account (@realDonaldTrump) was deleted and Twarc does not allow to pull data from banned or deleted Twitter accounts.

Exercise 7

By counting first the estimated results, we obtained a total of 84 tweets.

```
twarc2 counts --granularity day --archive --csv "\"Barcelona School of
    Economics\" -is:retweet" > Exercise7.csv

twarc2 search --archive "\"Barcelona School of Economics\" -is:retweet" >
    Exercise7.json
```

2. Mini-Research Project

Country chosen is Colombia. 59 Colombian politicians and their Twitter accounts have been identified as shown in Table 1 in Appendix A.1. In order to identify the politicians' Twitter accounts, we searched the Twitter accounts of the Colombian Senate, Congress, Presidency and some political parties (see Appendix A.2). We saved the usernames in 2 separate txt files, one containing only their username, one user per line (Colombia,txt), and another file used for counting with From: at the beginning of each line (Colombia2). Then, we analyzed the users that these accounts are following, which normally are important politicians.
From counting the data we found that 13597 tweets were sent one week before and after the elections in 2018 compared to only 8622 tweets sent in the same time period one year before.

```
twarc2 searches --archive --start-time "2018-05-20" --end-time "2018-06-03" --
    counts-only Colombia2.txt count2018.csv

twarc2 searches --archive --start-time "2017-05-20" --end-time "2017-06-03" --
    counts-only Colombia2.txt count2017.csv
```

Using the queries below we obtained the full activity of the politicians:[1]

```
twarc2 timelines --use-search --start-time "2018-05-20" --end-time
    "2018-06-03" Colombia.txt elections2018.json

twarc2 timelines --use-search --start-time "2017-05-20" --end-time
    "2017-06-03" Colombia.txt elections2017.json
```

In Figure 5 we can clearly see that the number of tweets produced during 2018 (elections period) is larger than the quantity produced in 2017.

[1] See https://bit.ly/3rgBmxO for the referenced `.txt`-file.

A Appendices

A.1 Colombian Politicians and Their Twitter Accounts

Politician	Twitter Account
Iván Duque	@IvanDuque
Gustavo Petro	@petrogustavo
Claudia López	@ClaudiaLopez
Juan Manuel Santos	@JuanManSantos
Álvaro Uribe Vélez	@AlvaroUribeVel
Antanas Mockus	@AntanasMockus
Enrique Peñalosa	@EnriquePenalosa
Sergio Fajardo	@sergio_fajardo
Iván Cepeda	@IvanCepedaCast
Jorge Enrique Robledo	@JERobledo
Daniel Samper Ospina	@DanielSamperO
Óscar Iván Zuluaga	@OIZuluaga
Marta Lucía Ramírez	@mluciaramirez
Paloma Valencia	@PalomaValenciaL
Andrés Pastrana	@AndresPastrana_
Maria Fernanda Cabal	@MariaFdaCabal
Alejandro Gaviria	@agaviriau
Humberto de la Calle	@DeLaCalleHum
Daniel Quintero Calle	@QuinteroCalle
Roy Barreras	@RoyBarreras
Ángelica Lozano Correa	@AngelicaLozanoC
Carlos F. Galán	@CarlosFGalan
Antonio Navarro	@navarrowolff
Clara López	@ClaraLopezObre
Horacio Serpa	@HoracioSerpa
Germán Vargas	@German_Vargas
Rafael Pardo	@RafaelPardo
Armando Benedetti	@AABenedetti
Gustavo Bolívar	@GustavoBolivar
Aída Avella	@AidaAvellaE

Politician	Twitter Account
Miguel Uribe	@MiguelUribeT
Carlos Alberto Baena	@Baena
Federico Gutiérrez	@FicoGutierrez
Alejandro Char	@AlejandroChar
Alejandro Ordóñez	@A_OrdonezM
Alejandro Eder	@alejoeder
Luis Alfredo Ramos	@LuisAlfreRamos
Alicia Arango	@AliciaArango
Piedad Córdoba	@piedadcordoba
Ángela María Robledo	@angelamrobledo
Ernesto Macías	@ernestomaciast
Luis Ernesto Gómez	@LuisErnestoGL
Ana Paola Agudelo	@AnaPaolaAgudelo
Luis Alfredo Ramos	@LuisAlfreRamos
Carlos Eduardo Guevara	@carlos_guevara
Juan Diego Gómez	@Juandiegogj
Arturo Char	@ArturoCharC
Carlos Felipe Mejía	@CARLOSFMEJIA
Rodrigo Lara	@Rodrigo_Lara_
Fernando Nicolás Araujo	@FNAraujoR
Carlos Caicedo	@carlosecaicedo
Luis Fernando Velasco	@velascoluisf
Aydée Lizarazo	@aydeelizarazoc
Efrain Cepeda	@EfrainCepeda
Clara Luz Roldán	@ClaraLuzRoldan
Wilson Arias	@wilsonariasc
Rodolfo Hernández	@ingrodolfohdez
Pacho Santos	@PachoSantosC
Juan Carlos Pinzón	@PinzonBueno

Table 1: List of Colombian Politicians and Their Twitter Accounts

A.2 Useful Twitter Accounts

Figure 4: Twitter Accounts Useful for Finding Colombian Politicians

A.3 Activity Counts, 2017 vs. 2018

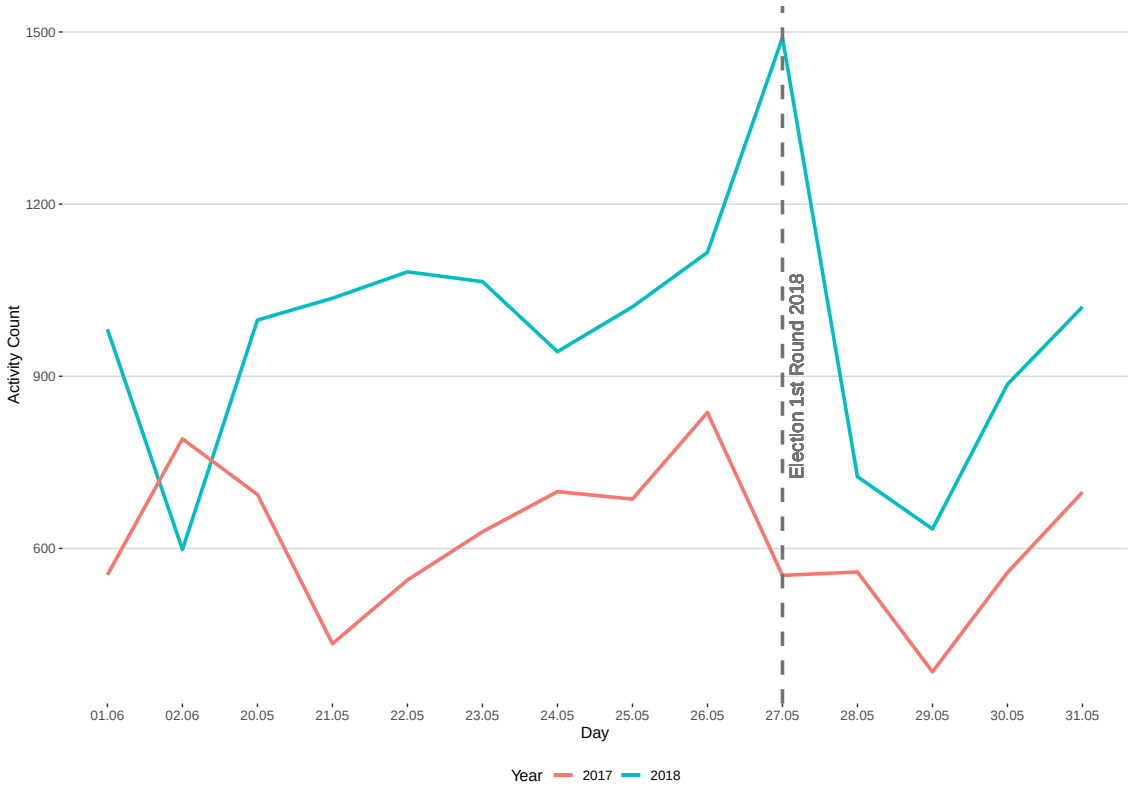

Figure 5: The Activity Count of above Listed Politicians, 2017 vs. 2018