

Mining data from social media with R

A very brief introduction

Twitter

1. Social Media with 330 millions active monthly users in 2019
2. Differently from the other social media, in Twitter all the posts are public and visible also for no-registered users
3. Tweets can be freely mined using Twitter's application programming interface (with some temporal limitations)



Benefits

1. Tweets analysis is a powerful tool for studying collective people behavior in real social contexts;
2. Support for explorative purpose and hypotheses generations;
3. May permits and test theoretical integration between Sociological and Cognitive Models.

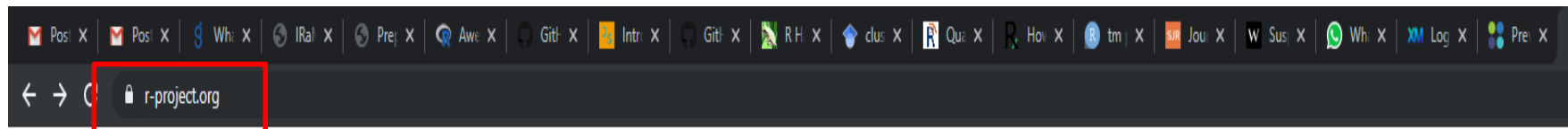
BUT

The Big Data Challenge

The creation of tweets is faster than their analysis, they have a lot of diverse and noisy information (text, images, audios, videos, which are hard to interpret).

How to mine data from twitter with R a step-by-step wannabe guide

STEP 1 → DOWNLOAD R



[Home]

Download

CRAN



The R Project for Statistical Computing

← → ↻ cran.r-project.org/mirrors.html

<https://ftp.heanet.ie/mirrors/cran.r-project.org/>

Italy

<http://cran.mirror.garr.it/mirrors/CRAN/>

<https://cran.stat.unipd.it/>

How to mine data from twitter with R

a step-by-step wannabe guide

STEP 1 → DOWNLOAD R (and then install)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

The installation is super easy, so I decided to skip that part

How to mine data from twitter with R a step-by-step wannabe guide

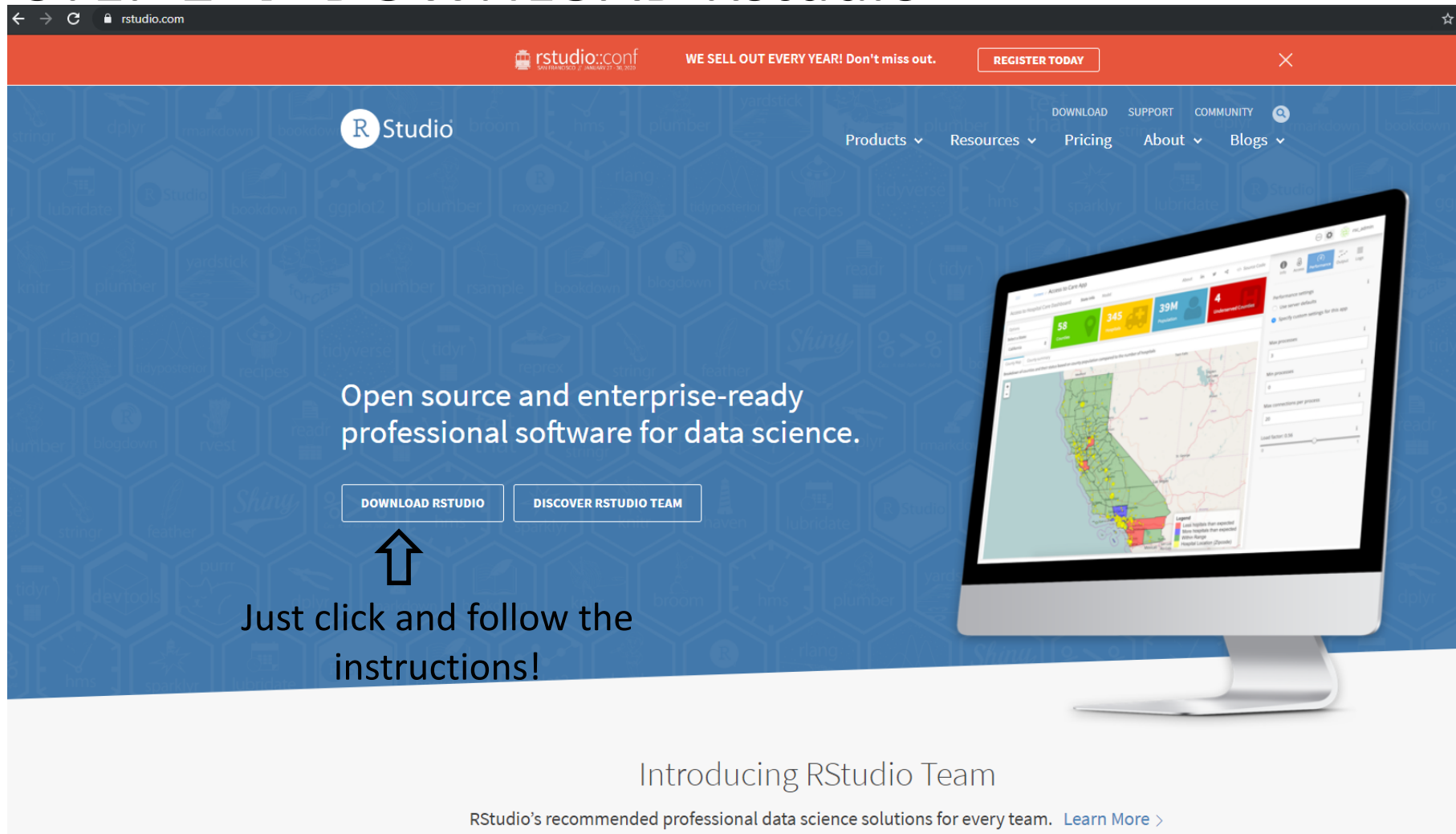
STEP 2 → DOWNLOAD Rstudio

Not really necessary but Rstudio is great and I
use Rstudio.

You should do the same!

How to mine data from twitter with R a step-by-step wannabe guide

STEP 2 → DOWNLOAD Rstudio



The image shows a screenshot of the RStudio website homepage. The browser address bar shows 'rstudio.com'. The website has a blue background with a hexagonal pattern. At the top, there is a navigation bar with 'rstudio:conf' and a 'REGISTER TODAY' button. Below that, the 'R Studio' logo is on the left, and navigation links for 'Products', 'Resources', 'Pricing', 'About', and 'Blogs' are on the right. The main content area features the text 'Open source and enterprise-ready professional software for data science.' and two buttons: 'DOWNLOAD RSTUDIO' and 'DISCOVER RSTUDIO TEAM'. An arrow points from the 'DOWNLOAD RSTUDIO' button to the text 'Just click and follow the instructions!'. On the right side, there is a computer monitor displaying a data visualization interface with a map of California and various data points. At the bottom, there is a section titled 'Introducing RStudio Team' with a 'Learn More >' link.

← → ↻ 🔒 rstudio.com ☆

rstudio:conf WE SELL OUT EVERY YEAR! Don't miss out. REGISTER TODAY

R Studio

DOWNLOAD SUPPORT COMMUNITY

Products ▾ Resources ▾ Pricing About ▾ Blogs ▾

Open source and enterprise-ready professional software for data science.

DOWNLOAD RSTUDIO DISCOVER RSTUDIO TEAM

Just click and follow the instructions!

Introducing RStudio Team

RStudio's recommended professional data science solutions for every team. [Learn More >](#)

How to mine data from twitter with R a step-by-step wannabe guide

STEP 3 → install and load rtweet package

The screenshot displays the RStudio interface during the installation of the `rtweet` package. The 'Install Packages' dialog box is the central focus, with a red circle highlighting the 'Install' button. The dialog shows the following configuration:

- Repository (CRAN)
- Packages (separate multiple with space or comma): `rtweet`
- Install to Library: `C:/Users/gabriel/Documents/R/R-3.5.2/library [Default]`
- Install dependencies

The background shows a data table with columns: `id`, `favorite_count`, `retweet_count`, `Tweet`, `id.net.file.con.tutti.i.23000.tweets`, `screen_name`, and `Formale.informale.1.10.`. The console window shows the following output:

```
[workspace loaded from ~/.RData]
Loading required package: BayesFactor
Loading required package: coda
Loading required package: Matrix
*****
Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richardmorey@gmail.com).
Type ?manual() to open the manual.
*****
Loading required package: lmerTest
Loading required package: lme4
Attaching package: 'lmerTest'
The following object is masked by '_by_'.GlobalEnv':
  ham
The following object is masked from 'package:lme4':
  lmer
The following object is masked from 'package:stats':
  step
```

How to mine data from twitter with R a step-by-step wannabe guide

STEP 3 → install and load rtweet package

The screenshot shows the RStudio interface with several panels:

- Environment:** Lists variables like `immig`, `Ant1.Immigrazione`, `Pro.Immigrazione`, `pwr`, and `rtweet`.
- Console:** Shows the execution of `library(rtweet)` and a warning message: "Warning message: package 'rtweet' was built under R version 3.5.3". A large blue arrow points to this console output, and a red circle highlights the `library(rtweet)` command.
- Documentation:** Shows the help page for `pwr.f2.test`, titled "Power calculations for the general linear model".

id	favorite_count	retweet_count	Tweet	id.nel.file.con.tutti.i.23000.tweets	screen_name	Formale.informale..1.10.
1	16479	11060	3527	Come può un Ministro dell'Interno esporre alla gogna ...	16479	lauraboldrini
2	5557	9288	1835	Ho 5 milioni di italiani poveri, quando avrò sfamato loro...	5557	matteosalvinimi
3	5210	7987	1616	Possono darmi anche l'ergastolo! Rs. Grazie al popolo di...	5210	matteosalvinimi
4	16194	7870	1725	Dovevano abolire la povertà e invece si accaniscono con...	16194	lauraboldrini
5	3415	7691	1468	Fantastico!<U>-2764> E se a qualcuno non va bene che...	3415	matteosalvinimi
6	15952	6106	1411	Non abbassare la testa di fronte a chi ogni giorno mette...	15952	lauraboldrini
7	5153	5943	938	Saviano, Cacciari, Benigni e Gad Lerner hanno firmato u...	5153	matteosalvinimi
8	5213	5882	1835	GRAZIE a chi mi sta manifestando il suo affetto twittand...	5213	matteosalvinimi
9	5781	5423	1263	<U>0001F534>Nave Open Arms, Ong e bandiera spagn...	5781	matteosalvinimi
10	5889	5246	1494	ASCOLTATE! Vogliamo il permesso di soggiorno SUBITO, ...	5889	matteosalvinimi
11	18745	5158	1378	Sono fiero del mio amico @davidefaraone, padre di una...	18745	matteoreenzi
12	14648	6063	366	Grazie di cuore a tutte e tutti per i fantastici #Aurivill au...	14648	lauraboldrini

```
> library(rtweet)
Warning message:
package 'rtweet' was built under R version 3.5.3
```

Power calculations for the general linear model

Description

Compute power of test or determine parameters to obtain target power (same as power.anova.test).

Usage

```
pwr.f2.test(u = NULL, v = NULL, f2 = NULL, sig.level = 0.05, power = NULL)
```

Arguments

- `u`: degrees of freedom for numerator
- `v`: degrees of freedom for denominator
- `f2`: effect size
- `sig.level`: Significance level (Type I error probability)
- `power`: Power of test (1 minus Type II error probability)

Details

Exactly one of the parameters 'u', 'v', 'f2', 'power' and 'sig.level' must be passed as NULL, and that parameter is determined from the others. Notice that the last one has non-NULL default so NULL must be explicitly passed if you want to compute it.

Value

How to mine data from twitter with R a step-by-step wannabe guide

STEP 4 → Let's start!

Function 1 → Search tweet!

search for 18000 tweets using the metoo hashtag

```
rt <- search_tweets(  
  "#MeToo", n = 18000, include_rts = FALSE  
)
```

**## search for 250,000 tweets containing the
word data**

```
rt <- search_tweets(  
  "data", n = 250000, retryonratelimit = TRUE  
)
```

How to mine data from twitter with R a step-by-step wannabe guide

STEP 4 → Leeeet's staaaaart!

Function 2 → Plotting geo-coordinates!

```
## search for 10,000 tweets sent from the US
```

```
rt <- search_tweets(  
  "lang:en", geocode = lookup_coords("usa"), n = 10000  
)
```

```
## create lat/lng variables using all available tweet and profile geo-location data
```

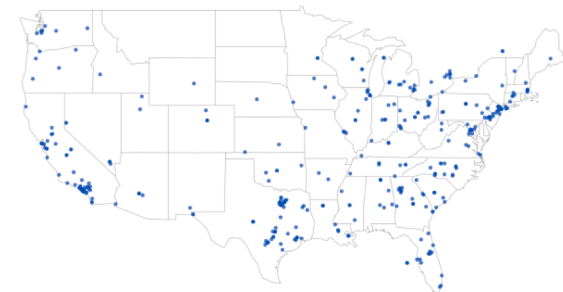
```
rt <- lat_lng(rt)
```

```
## plot state boundaries
```

```
par(mar = c(0, 0, 0, 0))  
maps::map("state", lwd = .25)
```

```
## plot lat and lng points onto state map
```

```
with(rt, points(lng, lat, pch = 20, cex = .75, col = rgb(0, .3, .7, .75)))
```



How to mine data from twitter with R a step-by-step wannabe guide

STEP 4 → Leeeet's staaaaart!

Function 3 → Collect the stream of tweet in real time!

```
## random sample for 30 seconds (default)
```

```
rt <- stream_tweets("")
```

```
## stream tweets from USA for 60 seconds
```

```
rt <- stream_tweets(lookup_coords("usa"), timeout = 60)
```

```
## stream tweets for a week (60 secs x 60 mins * 24 hours * 7 days)
```

```
Rt<-stream_tweets(  
  "",  
  timeout = 60 * 60 * 24 * 7,  
)
```



You need a good computer that will do just this

How to mine data from twitter with R a step-by-step wannabe guide

STEP 4 → Leeeet's staaaaart!

Function 4 → Extract followers and following!

```
## get user IDs of accounts followed by Matteo Salvini
```

```
Matteo_friends <- get_friends("@matteosalvinimi")
```

```
## lookup data on those accounts
```

```
Matteo_friends_data <- lookup_users(Matteo_friends$user_id)
```

```
## get user IDs of accounts following Matteo Salvini
```

```
Matteo_followers <- get_followers("@matteosalvinimi", n = 75000)
```

```
## lookup data on those accounts
```

```
Matteo_followers_data <- lookup_users(Matteo_followers$user_id)
```

How to mine data from twitter with R a step-by-step wannabe guide

STEP 4 → Leeeet's finish!

Function 4 → Extract users' timeline!

#Timeline of politicians

```
tmls <- get_timelines(c("@GiorgiaMeloni", "@matteosalvinimi",  
"@luigidimaio", "@nzingaretti", "@emmabonino", "@lauraboldrini", "@matteorenzi"),  
n = 32000)
```

And now save your data in a .csv file

```
write.csv2(tmls, "filename.csv")  
#change name if you want different files
```




**NEW CHALLENGER
APPROACHING**



Reddit

1. Social Media with 48 millions active monthly users in 2019
2. Posts are public
3. Comments and discussions can be freely mined using Reddit's application programming interface (with some limitations)
4. Discussions usually are way longer than in Twitter (but less users).



How to mine data from ~~twitter~~ Reddit with R

a step-by-step wannabe guide

Install and load the new library

```
install.packages("RedditExtractorR")  
library(RedditExtractorR)
```

Look for links containing the word Emergency

```
links <- reddit_urls(search_terms = "Emergency", page_threshold = 2,  
cn_threshold= , subreddit =, regex_filter =, sort_by = )
```

Look for contents contained in the links

```
content <- reddit_content(links$URL)
```

How to mine data from ~~twitter~~ Reddit with R

a step-by-step wannabe guide

You can also obtain a network of the discussion!

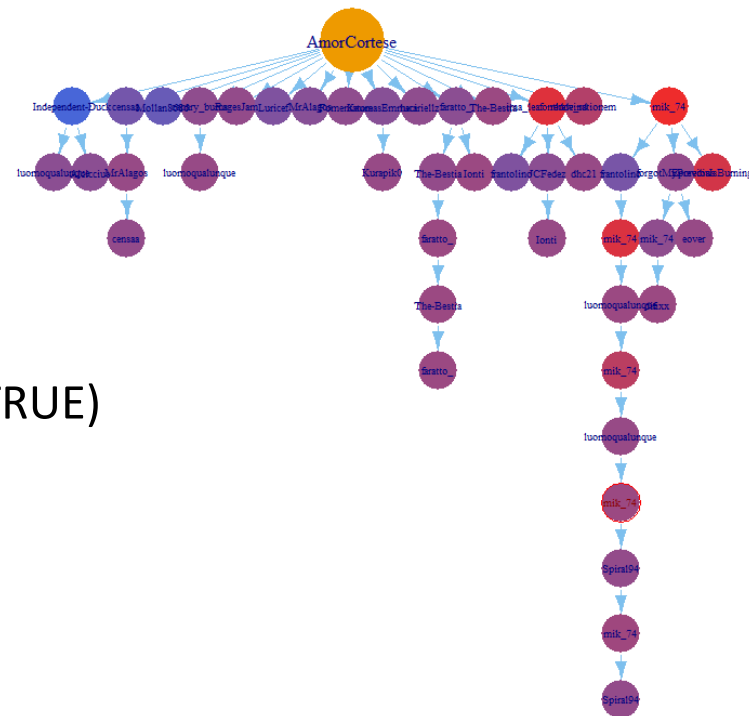
Look for content in one discussion

```
content1 <- reddit_content(links$URL[1])
```

Plot the network

```
graph <- construct_graph(content1, plot = TRUE)
```

Calabria, Gino Strada: "Accordo tra Emergency e Protezione Civile per rispondere all'emergenza sanitaria"



My contacts if you want some help

brunogabriel.salvadorcasara@phd.unipd.it

Brisbane, Queensland, Australia

From January
Office 27, Psicologia 1 via Venezia 12, Padova

Thank you for your attention!

A little help!

https://psicologiapd.fra1.qualtrics.com/jfe/form/SV_aeBH64H2qS2qVNz