Obtaining text data from online sources

Iulia Cioroianu, University of Bath

Intro to Text Analysis ESS 2021

WARNING:

Please do NOT increase the limits for the data collected!

We are all working on a single set of credentials and we'll hit API limits fast if you do.

Thank you!

TOPIC 1. Extracting Twitter data.

Load required packages

```
#load twitter library
library(rtweet)
#plotting
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#text mining library
library(tidytext)
library(rvest)
#library(textreadr)
library(xml2)
```

You can run this short code using my credentials, but you will have to replace the keys below with your Twitter API credentials if you want to collect additional data. Follow the instructions in the lecture.

- the name you gave to your app

```
appname <- "IC_app_3"
```

- api key

```
key <- "ulSLFLtKBBc0JkuFP1x2FxMu5"
```

- api secret

```
secret <- "NZlxcf2cSGYK7WjAg0VQ7j0rBj2JGkYfXhxG1z7JUatijaU3bv"
access_token <- "1904917267-3pSwSwABJ7XDYPKs7xEo4Ltd2TELIgLeTXXIEZv"
access_secret <- "WsSHjrs1ors6zKJbjy54hVSbRUeP8BTCh66dsOXpTTvB3"
```

Create token named "twitter_token"

```
twitter_token <- create_token(
  app = appname,
  consumer_key = key,
  consumer_secret = secret,
  access_token = access_token,
  access_secret = access_secret)
```

Search for 20 tweets using the #textanalysis hashtag

```
result_tweets <- search_tweets(q = "#textanalysis",
                               n = 20)
```

View the results

```
result_tweets
```

```
## # A tibble: 20 x 90
##    user_id   status_id   created_at          screen_name  text           source
##    <chr>     <chr>       <dttm>              <chr>        <chr>          <chr>
##  1 563874119 1414553120~ 2021-07-12 11:51:59 stenysolitu~ "Join our web~ Twitte~
##  2 33893047  1414549994~ 2021-07-12 11:39:34 ontotext     "Join our web~ HubSpot
##  3 478102175 1414128135~ 2021-07-11 07:43:15 AfefBahri    "A combinatio~ Twitte~
##  4 95522197~ 1413851197~ 2021-07-10 13:22:48 liaoyenchieh "A great mixt~ Twitte~
##  5 98771209~ 1413583474~ 2021-07-09 19:38:58 DigtalHuman~ "Check out \"~ K. Whi~
##  6 12986878~ 1413581300~ 2021-07-09 19:30:19 BotLabhd     "Check out \"~ botlab~
##  7 49427287  1413581275~ 2021-07-09 19:30:13 fentnz       "Check out \"~ Twitte~
##  8 10118176~ 1413522305~ 2021-07-09 15:35:54 rstatstweet  "Curious abou~ rstats~
##  9 12889976~ 1413519574~ 2021-07-09 15:25:03 TheNCDS      "Curious abou~ Sprout~
## 10 319916067 1413398834~ 2021-07-09 07:25:16 c_kibaki     "https://t.co~ Twitte~
## 11 49254270~ 1413131384~ 2021-07-08 13:42:31 3rdienterpr~ "Download 3RD~ Hootsu~
## 12 49254270~ 1412422323~ 2021-07-06 14:44:58 3rdienterpr~ "3RDi Search ~ Hootsu~
## 13 73714220~ 1412406563~ 2021-07-06 13:42:20 chidambara09 "#GraphAnalyt~ Twitte~
## 14 73714220~ 1412345733~ 2021-07-06 09:40:37 chidambara09 "#GraphAnalyt~ Twitte~
## 15 82099048~ 1412389214~ 2021-07-06 12:33:24 word_nerdy   "What do you ~ HubSpot
## 16 11613389~ 1412345882~ 2021-07-06 09:41:13 SciphicsTech "#GraphAnalyt~ Twitte~
## 17 13133285~ 1412344096~ 2021-07-06 09:34:07 Discovertec~ "#GraphAnalyt~ tech-t~
## 18 11514944~ 1412337757~ 2021-07-06 09:08:55 PythonExper~ "#GraphAnalyt~ python~
## 19 11750111~ 1412337147~ 2021-07-06 09:06:30 HubPrefer    "#GraphAnalyt~ Twitte~
## 20 29948988  1412335513~ 2021-07-06 09:00:00 payoda       "Eliminate th~ Twitte~
## # ... with 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>, ext_media_expanded_url <list>,
## #   ext_media_type <chr>, mentions_user_id <list>, mentions_screen_name <list>,
## #   lang <chr>, quoted_status_id <chr>, quoted_text <chr>,
```

```
## #   quoted_created_at <dttm>, quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>, quoted_screen_name <chr>,
## #   quoted_name <chr>, quoted_followers_count <int>,
## #   quoted_friends_count <int>, quoted_statuses_count <int>,
## #   quoted_location <chr>, quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>, retweet_created_at <dttm>,
## #   retweet_source <chr>, retweet_favorite_count <int>,
## #   retweet_retweet_count <int>, retweet_user_id <chr>,
## #   retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>, country <chr>,
## #   country_code <chr>, geo_coords <list>, coords_coords <list>,
## #   bbox_coords <list>, status_url <chr>, name <chr>, location <chr>,
## #   description <chr>, url <chr>, protected <lgl>, followers_count <int>,
## #   friends_count <int>, listed_count <int>, statuses_count <int>,
## #   favourites_count <int>, account_created_at <dttm>, verified <lgl>,
## #   profile_url <chr>, profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>
```

Leaving out retweets

```
result_tweets <- search_tweets(q = "#textanalysis",
                               n = 20,
                               include_rts = FALSE)
```

Who tweets about textanalysis?

```
result_tweets$screen_name
```

```
##  [1] "ontotext"       "liaoyenchieh"   "fentnz"         "TheNCDS"
##  [5] "3rdienterprise" "3rdienterprise" "word_nerdy"     "SciphicsTech"
##  [9] "HubPrefer"      "payoda"
```

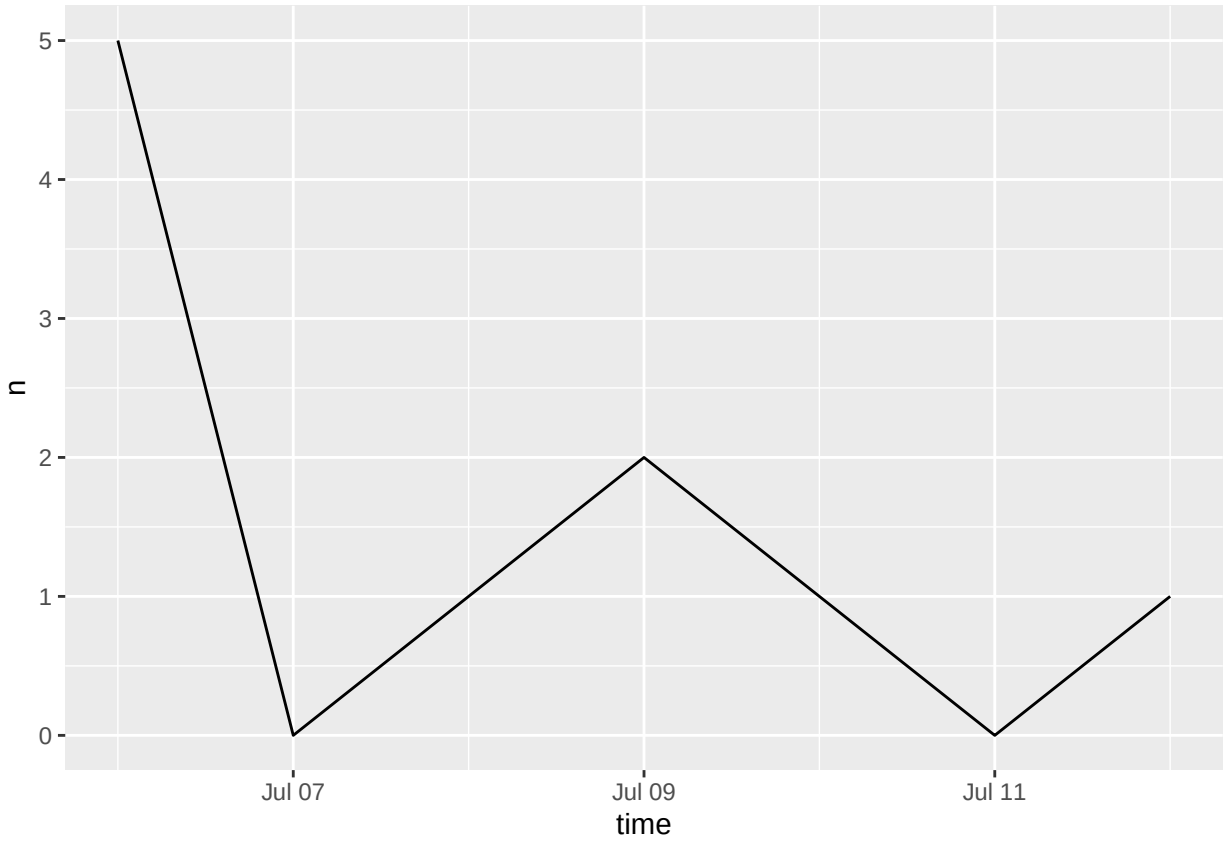
What do they tweet?

```
result_tweets$text
```

```
##  [1] "Join our webinar next week! We will demonstrate you how to discover, promote, deliver and reus
##  [2] "A great mixture of the panel discusses the political text quantitatively and qualitatively wit
##  [3] "Check out \"Unlocking the archives together - text recognition and extraction\" online event Tu
##  [4] "Curious about how you can clean and analyze #textualdata using R? During our #DataMatters cours
##  [5] "Download 3RDi Search Brochure https://t.co/e5d0Jy0rMW. Explore elevated #search technology for
##  [6] "3RDi Search is the ideal platform for analyzing enterprise data. Request Demo https://t.co/O9Mj
##  [7] "What do you do with a ton of text data? From humans to word clouds or doing bot all - we break
##  [8] "#GraphAnalytics make use of #Algorithms to discover the relationships among entries in a graph
##  [9] "#GraphAnalytics make use of #Algorithms to discover the relationships among entries in a graph
## [10] "Eliminate the need for time and money-consuming customer surveys with our comprehensive #TextAn
```

How did the number of tweets mentioning text analysis evolve over time?

```
ts_plot(result_tweets)
```



Get user IDs of accounts followed by EssexSumSchool:

```
ess_friends <- get_friends("EssexSumSchool")
print(ess_friends)
```

```
## # A tibble: 804 x 2
##    user           user_id
##    <chr>          <chr>
##  1 EssexSumSchool 3964884800
##  2 EssexSumSchool 1230782260934250498
##  3 EssexSumSchool 3096613518
##  4 EssexSumSchool 4055124016
##  5 EssexSumSchool 999029280741974017
##  6 EssexSumSchool 848826023470878720
##  7 EssexSumSchool 55241759
##  8 EssexSumSchool 4491008295
##  9 EssexSumSchool 1043120783638323200
## 10 EssexSumSchool 52693210
## # ... with 794 more rows
```

Get user IDs of accounts following me:

```
ic_followers <- get_followers("iuliacioroianu", n = 10)
ic_followers
```

```
## # A tibble: 10 x 1
##    user_id
```

```
##    <chr>
##  1 3323037854
##  2 1377987857793626112
##  3 49668806
##  4 1017846579003183104
##  5 1043967121397567488
##  6 248647655
##  7 979261158
##  8 820038719033970688
##  9 44602492
## 10 897384666516389888
```

Get ESS timeline:

```
ESS_timeline <- get_timelines("EssexSumSchool", n = 10)
```

What are they saying?

```
#ESS_timeline$text
```

Save collected tweets to csv file.

```
save_as_csv(ESS_timeline, "ESS_timeline.csv", prepend_ids = TRUE, fileEncoding = "UTF-8")
```

TOPIC 2: Web scraping

Loading the packages:

```
library(rvest) # Used for webscraping
library(stringr) # Dealing with strings and cleaning up data
library(dplyr) # For the data_frame function
```

Example 2: Extracting the text of the budget speech 2020

Extract the whole page:

```
speech <- read_html("https://www.gov.uk/government/speeches/budget-speech-2020")
speech #What does it look like? Go back to the browser, compare and inspect it.
```

```
## {html_document}
## <html lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=utf-8 ...
## [2] <body>\n    <script>document.body.className = ((document.body.className)  ...
```

Extracting the text of the speech: We pass the nodes in html_nodes and extract the text:

```
speech_text <- speech %>%
  html_nodes(xpath = "//div[@class='govspeak']")  %>%
  html_text()
```

Have a look at the text colected:

```
speech_text
```

```
## [1] "\nMadam Deputy Speaker,\n\nI want to get straight to the issue most on everyone's mind- coronavi
```

Extract the data the speech was published. Again, use xpath.

```
speech_date <- speech %>%
  html_nodes(xpath = '//*[@id="content"]/div[2]/div/div[1]/div/dl/dd[2]')  %>%
  html_text()
# Have a look at the text collected:
speech_date
```

```
## [1] "11 March 2020"
```

Example 3: Extracting article headlines from Google News

Extract the whole page

```
google <- read_html("https://news.google.com/")
google #What does it look like? Go back to the browser, compare and inspect it.
```

```
## {html_document}
## <html lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body id="yDmH0d" jscontroller="pjICDe" jsaction="rcuQ6b:npT2md; click:FA ...
```

Extracting the sources We pass the nodes in html_nodes and extract the text We use stringr to delete string elements that are not important

```
sources_all <- google %>%
  html_nodes("c-wiz div div div div main c-wiz div div div article div div a") %>%
  html_text()
```

The xpath is: "//*[@id="yDmH0d"]/c-wiz/div/div[2]/div[2]/div/main/c-wiz/div[1]/div[3]/div/article/div[2]/div/a" # Edit to "c-wiz div div div div main c-wiz div div div article div div a"

Inspecting the results:

```
sources_all[1:10] # look at the first ten
```

```
##  [1] "NBC News"                 "CNN "
##  [3] "KXAN.com"                 "Fort Worth Star-Telegram"
##  [5] "U.S. News & World Report" "Fox News"
##  [7] "Fox News"                 "CBS News"
##  [9] "Miami Herald"             "The Washington Post"
```

TOPIC 3. Extracting news article

Example 4: Extracting headlines from Google News

Loading the packages:

```
library(rvest) #Used for webscraping
library(stringr) #Dealing with strings and cleaning up data
library(dplyr) #For the data_frame function
```

Extract the whole page

```
google <- read_html("https://news.google.com/")
google #What does it look like? Go back to the browser, compare and inspect it.
```

```
## {html_document}
## <html lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
```

```
## [2] <body id="yDmH0d" jscontroller="pjICDe" jsaction="rcuQ6b:npT2md; click:FA ...
```

Extracting the sources We pass the nodes in html_nodes and extract the text We use stringr to delete string elements that are not important

```
sources_all <- google %>%
  html_nodes("c-wiz div div div div main c-wiz div div div article div div a") %>%
  html_text()
```

Inspecting the results:

```
sources_all[1:10] #look at the first ten
```

```
##  [1] "NBC News"                 "CNN "
##  [3] "KXAN.com"                 "Fort Worth Star-Telegram"
##  [5] "U.S. News & World Report" "Fox News"
##  [7] "Fox News"                 "CBS News"
##  [9] "Miami Herald"             "The Washington Post"
```

Extracting the headlines and using stringr for cleaning

```
headline_all <- google %>% html_nodes("article") %>% html_text("span") %>%
  str_split("(?<=[a-z0-9!?\\.])(?=[A-Z])")

headline_all <- sapply(headline_all, function(x) x[1]) #extract only the first elements

headline_all[1:10] #look at the first ten
```

```
##  [1] "Texas Democrats flee state in effort to block GOP-backed voting restrictionsamp"
##  [2] "Texas House Democrats leave state to block Republicans from passing voting restrictionsamp"
##  [3] "Texas Legislature: What is a quorum?amp"
##  [4] "Texas Democrats can't beat election bill, so they're running to Washington for bailoutamp"
##  [5] "Editorial Roundup: Texas | Texas News | US News - U."
##  [6] "Candidate Biden supported restoring diplomatic relations with communist Cubaamp"
##  [7] "Democratic socialists ignore Cuban protesters railing against communist dictatorshipamp"
##  [8] "Thousands protest in Cuba over food shortages and rising pricesamp"
##  [9] "What do Cubans who were shot at for staging unprecedented protests want? The right to be | Opi
## [10] "Opinion | Cubans are losing their fear. We want change.amp"
```

Exercise:

Identify the time since publication and extract it. Make sure you go back to Chrome and check the source for this.

```
time_all <- google %>%
  html_nodes("div article div div time") %>%
  html_text()
```

Check the results

```
time_all[1:10] #Look at the first ten
```

```
##  [1] "6 hours ago"  "2 hours ago"  "4 hours ago"  "4 hours ago"  "8 hours ago"
##  [6] "3 hours ago"  "14 hours ago" "9 hours ago"  "6 hours ago"  "4 hours ago"
```

Keep the smallest list.

```
min <- min(sapply(list(sources_all, time_all, headline_all), length))
```

Cut to smallest length

```r
sources_all <- sources_all[1:min]
time_all <- time_all[1:min]
headline_all <- headline_all[1:min]
```

Put the data into a new dataframe.

```r
df_news <- tibble(sources_all, time_all, headline_all)
#View(df_news)
```

Save data frame as .csv

```r
write.csv(df_news, "Google_News_headlines.csv")
```

Example 5. Extracting more article info using NewsAPI.org

```r
library(newsanchor)
```

#Let's inspect that the package can do. Uncomment or type "newsanchor::" below. #newsanchor::

Go to newsapi.org, and register for an API key, then paste it here to replace the existing one

```r
api_key="6c24c4efd0544256a3ad3e11273ed9b2"
```

Get headlines from the BBC

```r
results_BBC <- get_headlines(sources = "bbc-news", api_key=api_key)
#results_BBC
```

What kind of object is this?

```r
results_BBC[1]
```

```
## $metadata
##   total_results status_code        request_date
## 1            10         200 2021-07-12 22:59:48
##                                                               request_url
## 1 https://newsapi.org/v2/top-headlines?sources=bbc-news&pageSize=100&page=1
##   code message page page_size
## 1                  1       100
```

```r
#View(results_BBC$results_df)
```

Access one column:

```r
results_BBC$results_df["description"]
```

```
##                                                                                                     descri
## 1               Republicans in the state are proposing some of the most restrictive voting laws in th
## 2       Protesters say they are angry about the government's handling of both coronavirus and the eco
## 3        Thousands rallied on Sunday, angry at the country's economic crisis and curbs on civil liber
## 4                    Users in South Korea's capital are told to limit sweat-splashing and quick-breat
## 5       The special forces commander has handed off control as the 20-year US-led mission nears its c
## 6           Cuba has been plunged into turmoil by unusual protests. We look at the main drivers of un
## 7       Afghan woman Shakila Zareen had to have 22 operations after being shot in the face by her hus
## 8              An original Zelda cartridge and a Mario game set successive records at auction within 
## 9  Critics say the clip of a woman gasping for air is unfair to young people who can't get a vaccine
## 10          It will assess the security situation after last week's attack that triggered unrest in H
```

First element:

```
results_BBC$results_df["description"][1,]
```

```
## [1] "Republicans in the state are proposing some of the most restrictive voting laws in the US."
```

Write results to csv file:

```
write.csv(results_BBC$results_df, "BBC_results.csv")
```

What other sources?

```
newsanchor::terms_sources
```

```
##                         sources
## 1                       abc-news
## 2                    abc-news-au
## 3                    aftenposten
## 4             al-jazeera-english
## 5                           ansa
## 6                         argaam
## 7                   ars-technica
## 8                       ary-news
## 9                associated-press
## 10  australian-financial-review
## 11                          axios
## 12                       bbc-news
## 13                      bbc-sport
## 14                           bild
## 15               blasting-news-br
## 16                bleacher-report
## 17                      bloomberg
## 18                breitbart-news
## 19               business-insider
## 20            business-insider-uk
## 21                       buzzfeed
## 22                       cbc-news
## 23                       cbs-news
## 24                           cnbc
## 25                            cnn
## 26                         cnn-es
## 27              crypto-coins-news
## 28                     daily-mail
## 29               der-tagesspiegel
## 30                       die-zeit
## 31                       el-mundo
## 32                       engadget
## 33           entertainment-weekly
## 34                           espn
## 35                 espn-cric-info
## 36                 financial-post
## 37                financial-times
## 38                          focus
## 39                football-italia
## 40                        fortune
## 41                  four-four-two
## 42                       fox-news
## 43                     fox-sports
```

```
## 44                      globo
## 45                 google-news
## 46               google-news-ar
## 47               google-news-au
## 48               google-news-br
## 49               google-news-ca
## 50               google-news-fr
## 51               google-news-in
## 52               google-news-is
## 53               google-news-it
## 54               google-news-ru
## 55               google-news-sa
## 56               google-news-uk
## 57             goteborgs-posten
## 58                gruenderszene
## 59                  hacker-news
## 60                handelsblatt
## 61                         ign
## 62               il-sole-24-ore
## 63                  independent
## 64                      infobae
## 65                   info-money
## 66                    la-gaceta
## 67                    la-nacion
## 68                la-repubblica
## 69                     le-monde
## 70                        lenta
## 71                      lequipe
## 72                    les-echos
## 73                   liberation
## 74                        marca
## 75                     mashable
## 76           medical-news-today
## 77                        metro
## 78                       mirror
## 79                        msnbc
## 80                     mtv-news
## 81                  mtv-news-uk
## 82           national-geographic
## 83               national-review
## 84                      nbc-news
## 85                        news24
## 86                 new-scientist
## 87                   news-com-au
## 88                      newsweek
## 89             new-york-magazine
## 90               next-big-future
## 91                      nfl-news
## 92                      nhl-news
## 93                           nrk
## 94                      politico
## 95                       polygon
## 96                           rbc
## 97                        recode
```

```
## 98                 reddit-r-all
## 99                      reuters
## 100                          rt
## 101                         rte
## 102                   rtl-nieuws
## 103                         sabq
## 104              spiegel-online
## 105           svenska-dagbladet
## 106                          t3n
## 107                   talksport
## 108                  techcrunch
## 109                techcrunch-cn
## 110                   techradar
## 111   the-american-conservative
## 112               the-economist
## 113          the-globe-and-mail
## 114             the-guardian-au
## 115             the-guardian-uk
## 116                    the-hill
## 117                   the-hindu
## 118         the-huffington-post
## 119             the-irish-times
## 120          the-jerusalem-post
## 121               the-lad-bible
## 122          the-new-york-times
## 123                the-next-web
## 124             the-sport-bible
## 125               the-telegraph
## 126          the-times-of-india
## 127                   the-verge
## 128     the-wall-street-journal
## 129         the-washington-post
## 130        the-washington-times
## 131                        time
## 132                   usa-today
## 133                   vice-news
## 134                       wired
## 135                    wired-de
## 136           wirtschafts-woche
## 137                  xinhua-net
## 138                        ynet
```

Get headlines published in the category sports

```
results_sports <- get_headlines(category = "sports", country="it", api_key=api_key)
#View(results_sports$results_df)
```

What other categories are available? Check terms_category to see:

```
newsanchor::terms_category
```

```
##        category
## 1      business
## 2 entertainment
## 3       general
## 4        health
```

```
## 5       science
## 6        sports
## 7    technology
```

get headlines published in Germany

```
results <- get_headlines(country = "de",  api_key=api_key)
#results
```

Exercise:

Get headlines about COVID-19 published in the UK. Save the results in a .csv file.

```
results_C19_GB <- get_headlines(country = "gb",  query = "COVID-19", api_key=api_key)
write.csv(results_C19_GB$results_df, "UK_headlines_C19.csv")
```

On the free NewsAPI.org account you can get all results up to the limit of 1000 free articles and only 1 month back in time. *Using your own NewsAPI.org credentials*, - read about the get_everything function in the newsanchor package. - get everything published about "wave" and Covid-19 but without mentions of the NHS - feel free to use any available resources to help you solve this exercise.

```
results_search <- get_everything(query = "+wave +Covid -NHS",
                                 domains = "bbc.com",
                                 from = "2021-07-05",  api_key=api_key)
#View(results_search$results_df)
```