Lab 5. Networks on Texts. Part 2

National Research University Higher School of Economics, Faculty of Social Sciences

Introduction to Network Analysis, Spring 2019

Seminar 5, Part II, 15/02/2019

Dmitry Zaytsev, PhD and Valentina Kuskova, PhD

(with deep appreciation to all used sources; references available in text and upon request)

2. Parsing and Text Mining for English texts: example of the Protests in Venezuela 2019:

First, Get links from the GOOGLE search results protests in venezuela 2019:

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 3.5.2
```

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.5.2
```

```
page_use <- read_html("https://www.google.com/search?q=protests+in+venezuela+2019&num=100&newwindow=1&r
links <- page_use %>% html_nodes(xpath='//h3/a') %>% html_attr('href')
gsub('/url\\?q=','',sapply(strsplit(links[as.vector(grep('url',links))],split='&'),'[',1))
```

```
##   [1] "https://www.theguardian.com/world/2019/feb/12/venezuela-guaido-maduro-protest"
##   [2] "https://www.cnn.com/2019/01/29/americas/venezuela-protests-deaths/index.html"
##   [3] "https://ru.euronews.com/2019/02/12/ru-venezuela-protests"
##   [4] "https://www.cnn.com/americas/live-news/venezuela-protests-2019/index.html"
##   [5] "https://www.cnn.com/2019/02/02/americas/venezuela-unrest/index.html"
##   [6] "https://www.theguardian.com/world/2019/feb/08/venezuela-juan-guaido-central-university-speech
##   [7] "https://www.theguardian.com/world/video/2019/feb/03/tens-of-thousands-protest-in-venezuela-to
##   [8] "https://www.theguardian.com/world/2019/feb/10/venezuela-maduro-chavez-guaido-protests"
##   [9] "https://www.theguardian.com/global-development/2019/feb/14/haiti-disarray-anti-government-pro
##  [10] "https://www.euronews.com/2019/02/15/live-spain-braces-for-snap-election-trump-to-declare-emer
##  [11] "https://www.cnn.com/2019/01/28/americas/venezuela-unrest-maduro-accusations-guaido/index.html
##  [12] "https://www.workers.org/2019/02/14/danes-protest-u-s-coup-attempt-in-venezuela/"
##  [13] "https://www.cnn.com/2019/01/23/americas/venezuela-protests/index.html"
##  [14] "https://www.theguardian.com/world/2019/feb/15/can-juan-guaido-save-venezuela-from-cruel-dicta
##  [15] "https://www.theguardian.com/world/video/2019/jan/23/venezuela-opposition-leader-juan-guaido-d
##  [16] "https://www.washingtonpost.com/opinions/2019/02/14/venezuelas-regime-is-using-death-squads-to
##  [17] "https://www.tagesschau.de/ausland/venezuela-proteste-hilfslieferungen-103.html"
##  [18] "https://www.euronews.com/2019/02/13/us-trying-to-impose-puppet-government-in-venezuela-madrur
##  [19] "https://www.nytimes.com/2019/02/01/world/americas/venezuela-voices-protests.html"
##  [20] "https://www.theguardian.com/world/gallery/2019/jan/24/venezuela-protests-as-two-leaders-vie-t
##  [21] "https://www.hrw.org/news/2019/01/25/venezuela-arrests-killings-anti-government-protests"
##  [22] "https://www.theguardian.com/world/video/2019/jan/23/mike-pence-backs-venezuelan-protesters-se
##  [23] "https://ru.euronews.com/2019/01/23/venezuela-protests"
##  [24] "https://ru.euronews.com/2019/01/31/venezuela-more-protests"
```

```
##  [25] "https://www.theguardian.com/world/video/2019/feb/10/protesters-in-madrid-call-for-their-prime
##  [26] "https://www.amnesty.org/en/latest/news/2019/01/venezuela-more-than-a-dozen-people-killed-in-pr
##  [27] "https://www.theguardian.com/world/video/2019/feb/01/why-is-venezuela-in-crisis-video-explainer
##  [28] "https://www.wtvq.com/2019/02/12/watch-protests-in-venezuela/"
##  [29] "https://ru.euronews.com/2019/01/29/at-least-40-dead-during-protests-in-venezuela"
##  [30] "https://www.660citynews.com/video/2019/02/12/venezuela-protesters-call-for-emergency-aid/"
##  [31] "https://www.ctvnews.ca/canada/violent-protests-in-haiti-trap-more-than-100-canadian-tourists-a
##  [32] "https://www.miamiherald.com/news/nation-world/world/americas/venezuela/article224893920.html"
##  [33] "https://www.aljazeera.com/news/2019/01/venezuela-detains-foreign-journalists-crackdown-protest
##  [34] "https://www.aljazeera.com/news/2019/01/venezuela-shuts-internet-protests-190124124829727.html"
##  [35] "https://www.nytimes.com/2019/01/30/world/americas/venezuela-maduro-protests-faes.html"
##  [36] "https://www.local10.com/espanol/noticias/venezuela/venezuelan-protesters-already-in-mourning-m
##  [37] "https://www.washingtonpost.com/world/as-maduro-is-threatened-his-government-becomes-more-dange
##  [38] "https://kpfa.org/episode/flashpoints-february-13-2019/"
##  [39] "https://www.miamiherald.com/news/nation-world/world/americas/venezuela/article225158290.html"
##  [40] "https://www.cbsnews.com/news/protest-in-venezuela-millions-of-demonstrators-expected-to-flood-
##  [41] "https://www.thestar.com/news/world/americas/2019/02/12/venezuelan-opposition-banking-on-protes
##  [42] "https://www.theguardian.com/world/2019/feb/06/venezuela-faes-special-forces-nicolas-maduro-ban
##  [43] "http://www.ipsnews.net/2019/02/bullets-pots-pans-crackdown-venezuelas-protests-brutal/"
##  [44] "https://www.nytimes.com/2019/02/04/world/americas/venezuela-maduro-guaido-legitimate.html"
##  [45] "https://www.breitbart.com/national-security/2019/02/13/venezuelan-priests-protest-against-madu
##  [46] "https://toronto.citynews.ca/video/2019/02/02/mass-protests-held-in-support-of-venezuelas-oppos
##  [47] "https://www.latimes.com/nation/la-na-pol-abrams-congress-venezuela-20190213-story.html"
##  [48] "https://www.jpost.com/International/Jewish-journalist-on-the-frontlines-of-protest-in-Venezuel
##  [49] "https://www.aljazeera.com/news/2019/02/anti-maduro-protests-venezuelans-overseas-support-guaid
##  [50] "https://www.euronews.com/2019/01/25/venezuela-unrest-and-brussels-climate-change-protests-no-c
##  [51] "https://www.bbc.com/news/world-latin-america-47193837"
##  [52] "https://www.cnn.com/2019/02/11/americas/venezuela-migrant-women-prostitution-intl/index.html"
##  [53] "https://www.miamiherald.com/news/nation-world/world/americas/haiti/article225931055.html"
##  [54] "https://www.euronews.com/2019/02/14/live-brexit-debate-venezuela-crisis-spain-faces-early-elec
##  [55] "https://www.theguardian.com/world/2019/feb/06/this-man-plotted-guaidos-rise-and-still-dreams-o
##  [56] "https://www.bloomberg.com/news/articles/2019-01-23/venezuelan-protests-turn-violent-in-parts-o
##  [57] "http://www.chroniclet.com/national-news/2019/01/31/Scenes-from-a-protest-Venezuelans-fill-stre
##  [58] "https://theintercept.com/2019/02/10/intercepted-podcast-vijay-prashad-venezuela-india/"
##  [59] "https://www.bbc.co.uk/news/world-latin-america-47211509"
##  [60] "https://barbadostoday.bb/2019/02/10/protest-near-us-embassy-in-support-of-president-nicolas-ma
##  [61] "https://www.ozy.com/ain-navigation/pdb-92395/out-in-force-92446"
##  [62] "https://www.washingtonpost.com/national/ap-photos-editor-selections-from-latin-america-caribbe
##  [63] "https://www.thestar.com.my/news/world/2019/02/02/venezuela-opposition-rally-to-keep-up-pressur
##  [64] "https://psmag.com/news/viewfinder-demonstrations-in-venezuela"
##  [65] "https://www.ndtv.com/india-news/venezuela-oil-exports-us-warns-nations-as-sanctions-hit-venezu
##  [66] "https://www.miamiherald.com/news/nation-world/world/americas/venezuela/article224952860.html"
##  [67] "https://www.pymnts.com/news/international/2019/venezuela-money-supply-currency-inflation/"
##  [68] "https://www.japantimes.co.jp/news/2019/02/04/world/venezuela-slums-maduros-camp-now-said-victi
##  [69] "https://reliefweb.int/report/venezuela-bolivarian-republic/iachr-alarmed-arrests-context-prote
##  [70] "https://www.aysor.am/en/news/2019/01/24/venezuela-protests/1516660"
##  [71] "https://www.stltoday.com/news/world/photos-venezuelans-protest-in-st-louis/collection_4fd1bd0b
##  [72] "https://www.bbc.co.uk/news/world-latin-america-46864864"
##  [73] "https://cruxnow.com/church-in-the-americas/2019/01/25/venezuelan-army-besieged-hundreds-of-pro
##  [74] "https://www.caracaschronicles.com/2019/02/02/a-next-generation-protest-movement-swarms-the-str
##  [75] "https://www.foxnews.com/politics/us-special-envoy-for-venezuela-clashes-with-rep-omar"
##  [76] "https://cpj.org/2019/01/raids-media-shutdowns-and-internet-disruptions-ami.php"
##  [77] "https://www.yahoo.com/news/see-revived-2019-ram-1500-045800896.html"
##  [78] "https://katc.com/news/around-acadiana/lafayette-parish/2019/01/23/venezuelans-in-acadiana-prot
```

```
##  [79] "https://www.miamiherald.com/news/business/article226053360.html"
##  [80] "https://www.usnews.com/news/world/articles/2019-01-24/back-to-the-streets-venezuelan-protests
##  [81] "http://www.dailystar.com.lb/News/World/2019/Jan-29/475244-guaido-calls-for-fresh-venezuela-pr
##  [82] "https://www.breakingbelizenews.com/2019/01/18/haitian-opposition-plans-to-protest-vote-on-madu
##  [83] "https://www.miamiherald.com/news/nation-world/world/americas/venezuela/article224959780.html"
##  [84] "https://www.ktvz.com/news/national-world/venezuela-a-day-of-dueling-protests/1000127274"
##  [85] "https://www.miamiherald.com/news/nation-world/world/americas/haiti/article225999135.html"
##  [86] "https://www.wtvq.com/2019/01/24/watch-surfing-seniors-ice-fishing-and-protests-in-venezuela-w
##  [87] "https://www.dailypioneer.com/2019/world/over-350-protesters-detained-in-venezuela-this-week--
##  [88] "https://www.catholicnews.com/services/englishnews/2019/venezuelan-bishops-say-fresh-calls-for
##  [89] "https://www.caracaschronicles.com/2019/02/03/the-regime-didnt-dare-to-rough-up-yesterdays-pro
##  [90] "https://www.telegraph.co.uk/news/2019/01/29/pentagon-refuses-rule-us-military-deployment-venez
##  [91] "https://www.democracynow.org/2019/2/11/headlines/nyc_protesters_blast_guggenheims_ties_to_sack
##  [92] "https://www.democracynow.org/2019/2/8/headlines/report_us_based_plane_caught_bringing_arms_in
##  [93] "https://www.news.com.au/lifestyle/health/health-problems/venezuelas-disastrous-government-send
##  [94] "https://www.miamiherald.com/news/nation-world/world/americas/venezuela/article223796500.html"
##  [95] "https://www.aljazeera.com/news/2019/01/maduro-opponents-hold-protest-call-venezuela-elections
##  [96] "https://www.burnabynow.com/opinion/blogs/burnaby-candidates-singh-robinson-under-fire-from-ve
##  [97] "https://www.miamiherald.com/news/local/community/miami-dade/article224190025.html"
##  [98] "https://www.zeit.de/2019/07/berlin-i-love-you-film-kino-klischees"
##  [99] "http://time.com/5495476/venezuela-supreme-court-justice-flees-us/"
## [100] "https://www.local10.com/espanol/noticias/venezuela/maduro-s-new-opponent-calls-for-jan-23-pro
```

Second, Download in working environment the vector of first 10 links:

```
links2 <- c ("https://www.theguardian.com/world/2019/feb/12/venezuela-guaido-maduro-protest",
"https://ru.euronews.com/2019/02/12/ru-venezuela-protests",
"https://www.cnn.com/americas/live-news/venezuela-protests-2019/index.html",
"https://www.theguardian.com/world/2019/feb/08/venezuela-juan-guaido-central-university-speech-maduro-p
"https://www.cnn.com/2019/02/02/americas/venezuela-unrest/index.html",
"https://www.cnn.com/2019/01/29/americas/venezuela-protests-deaths/index.html",
"https://www.theguardian.com/world/video/2019/feb/03/tens-of-thousands-protest-in-venezuela-to-urge-nic
"https://www.theguardian.com/world/2019/feb/10/venezuela-maduro-chavez-guaido-protests",
"https://www.cnn.com/2019/01/28/americas/venezuela-unrest-maduro-accusations-guaido/index.html",
"https://montreal.citynews.ca/video/2019/02/12/venezuela-protesters-call-for-emergency-aid/")
```

Third, Download .html files in the working directory:

```
for(i in 1:length(links2)){
  html_object  <- read_html(links2[i])
  somefilename <- paste0("filenameVE_", i, ".html")
  write_xml(html_object, file = somefilename)
}
```

Fourth, convert html to text

```
html <- list.files(pattern="\\.(htm|html)$") # get just .htm and .html files
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.5.2
```

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 3.5.2
```

```
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.5.2
```

```
## Loading required package: bitops
```

```
library(XML)
```

```
## Warning: package 'XML' was built under R version 3.5.2
```

```
##
## Attaching package: 'XML'
```

```
## The following object is masked from 'package:rvest':
##
##     xml
```

```
htmlToText <- function(input, ...) {
  ###---PACKAGES ---###
  require(RCurl)
  require(XML)


  ###--- LOCAL FUNCTIONS ---###
  # Determine how to grab html for a single input element
  evaluate_input <- function(input) {
    # if input is a .html file
    if(file.exists(input)) {
      char.vec <- readLines(input, warn = FALSE)
      return(paste(char.vec, collapse = ""))
    }

    # if input is html text
    if(grepl("</html>", input, fixed = TRUE)) return(input)

    # if input is a URL, probably should use a regex here instead?
    if(!grepl(" ", input)) {
      # downolad SSL certificate in case of https problem
      if(!file.exists("cacert.perm")) download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="
      return(getURL(input, followlocation = TRUE, cainfo = "cacert.perm"))
    }

    # return NULL if none of the conditions above apply
    return(NULL)
  }

  # convert HTML to plain text
  convert_html_to_text <- function(html) {
    doc <- htmlParse(html, asText = TRUE)
    text <- xpathSApply(doc, "//text()[not(ancestor::script)][not(ancestor::style)][not(ancestor::noscri
    return(text)
  }

  # format text vector into one character string
```

```
  collapse_text <- function(txt) {
    return(paste(txt, collapse = " "))
  }

  ###--- MAIN ---###
  # STEP 1: Evaluate input
  html.list <- lapply(input, evaluate_input)

  # STEP 2: Extract text from HTML
  text.list <- lapply(html.list, convert_html_to_text)

  # STEP 3: Return text
  text.vector <- sapply(text.list, collapse_text)
  return(text.vector)
}
html2txt <- lapply(html, htmlToText)
```

Fifth, clean out non-ASCII characters

```
html2txtclean <- sapply(html2txt, function(x) iconv(x, "latin1", "ASCII", sub=""))
```

Sixth, make corpus for text mining

```
corpus <- Corpus(VectorSource(html2txtclean))
```

Seventh, Let's check the corpus we recived:

```
tdm_matrix<-TermDocumentMatrix(corpus)
VE_tdm<-as.matrix(tdm_matrix)
dim(VE_tdm)
```

```
## [1] 3707   10
```

Eighth, Let's clean put data:

Better to do it in excel:

```
write.csv(VE_tdm, file = "VE.csv")
```

In Excel you can clean the data, and cut not important terms (I chose to cut words with frequencies less than 9, so I leave with 217 most frequent words out of 3677). I cut terms - can not read col.names - I cut them, by multiplication of matrices create TermToTerm matrix. Write it in working directory.

```
library(readxl)
VE_noterms <- read_excel("VE_noterms.xlsx")
VE<-as.matrix(VE_noterms)%*%t(as.matrix(VE_noterms))
dim(VE)
```

```
## [1] 217 217
```

```
write.csv(VE, file = "VE_matrixwithNOterms.csv")
```

Then, I add terms back, and read it again:

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
##
## Attaching package: 'readr'
```

```
## The following object is masked from 'package:rvest':
##
##     guess_encoding
```

```
VE_matrixwithterms <- read_csv("VE_matrixWITHterms.csv")
```

```
## Warning: Duplicated column names deduplicated: 'said' => 'said_1' [76],
## 'maduro' => 'maduro_1' [85], 'venezuela' => 'venezuela_1' [88], 'TRUE' =>
## 'TRUE_1' [187], 'guaido' => 'guaido_1' [207]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
dim(VE_matrixwithterms)
```

```
## [1] 217 217
```

```
library(network)
```

```
## network: Classes for Relational Data
## Version 1.13.0.1 created on 2015-08-31.
## copyright (c) 2005, Carter T. Butts, University of California-Irvine
##                     Mark S. Handcock, University of California -- Los Angeles
##                     David R. Hunter, Penn State University
##                     Martina Morris, University of Washington
##                     Skye Bender-deMoll, University of Washington
##  For citation information, type citation("network").
##  Type help("network-package") to get started.
```

```
net <- as.network(VE_matrixwithterms)
plot(net)
```