Lab 5. Networks on Texts

Dmitry Zaytsev

February 11, 2019

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

Introduction to Network Analysis, Spring 2019

Seminar 5, 15/02/2019

Dmitry Zaytsev, PhD and Valentina Kuskova, PhD

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

Welcome to the fifth seminar! As the main problem to start using network analysis is to formulate network-related research question, and collect network data, we will focus today on the one of the most developing area of network analysis - networks on texts.

Contents of today's seminar:

The idea is to collect text data about anything, and prepare text data for network analysis. A possible research task can be to understand alternative discources about a topic.

First, we will try to scrap GOOGLE. But you can produce XML files for further text mining using other search engines (YANDEX) and codes (Rcrawler). Also, it is possible to do everything in Piton.

Your assignment is due on Friday, February 22, at 23.59.

1. Parsing and Text Mining for Russian texts: example of the USE:

1.1. Web scrapping:

To recieve html pages we dicided to use scrapping, not crawling. Rcrawler (https://github.com/salimk/Rcrawler) is a very usefull tool to recieve html pages and files recorded in the working directory. But for Google search results it is not very usefull because it is crawl forever :). I decided to use scraping. For GOOGLE, better to search in News folder, and to put in the Google search settings display by 100 results to get first 100 links similteneously:

```
#install.packages("rvest")
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=%22%D0%B5%D0%B4%D0%B8%D0%BD%
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] "http://www.ugorizont.ru/2019/02/07/okolo-90-tyisyach-shkolnikov-sdadut-
edinyiy-gosudarstvennyiy-ekzamen-v-2019-godu/"
## [2] "http://www.penza-press.ru/lenta-novostey/140088/penzenskie-shkolniki-prinyali-
uchastie-v-aprobacii-ege-po-fizike"
## [3] "http://www.obltv.ru/news/education/nazvany-samye-populyarnye-predmety-po-
vyboru-na-ege-2019-v-sverdlovskoy-oblasti/"
## [4] "https://korolevriamo.ru/article/237934/glavnoe-shkolnoe-ispytanie-sem-mifov-o-
ege-i-novovvedeniya-2019-goda.xl"
## [5] "http://zyuzinomedia.ru/news/moskovskie-novosti/pochti-300-chelovek-sdadut-
edinyy-gosudarstvennyy-ekzamen-po-kitayskomu-yazyku/"
## [6] "http://vechorka.ru/article/poznakomimsya-s-raspisaniem-ege-2019/"
## [7] "https://cjmonitor.com/kak-blokchejn-budet-ispolzovatsya-na-ege/"
## [8] "https://life.ru/t/%25D0%25BD%25D0%25BE%25D0%25B2%25D0%25BE%25D1%2581%25D1%
## [9] "https://www.mos.ru/news/item/50942073/"
## [10] "https://region29.ru/2019/02/05/5c598cbb12f17b5a336f4532.html"
## [11] "https://tass.ru/obschestvo/6078950"
## [12] "https://www.donetsk.kp.ru/online/news/3366740/"
## [13] "https://dostup1.ru/interview/Elena-Tyurina-EGE--eto-po-bolshomu-schetu-
sudba-cheloveka_113563.html"
## [14] "https://lubertsyriamo.ru/article/241426/kolichestvo-sdayuschih-ege-v-podmoskove-
vyrastet-na-1-tysyachu-chelovek-v-2019-godu.xl"
## [15] "https://regnum.ru/news/2562241.html"
## [16] "https://ria.ru/20190204/1550347035.html"
## [17] "https://www.novayagazeta.ru/articles/2019/01/20/79252-vernym-kratkim-
kursom"
## [18] "https://www.gazeta.ru/social/2019/02/06/12167491.shtml"
## [19] "https://vtambove.ru/news/society/182224/"
## [20] "http://yamal-region.tv/news/35198/"
## [21] "https://www.ridus.ru/news/291015"
## [22] "https://tass.ru/obschestvo/6000478"
## [23] "http://www.gazetaingush.ru/news/zayavlenie-na-sdachu-ege-neobhodimo-podat-
do-1-fevralya"
## [24] "https://amurmedia.ru/news/784356/"
## [25] "https://news.rambler.ru/education/41674018-shkolniki-smogut-napisat-itogovoe-
sochinenie-v-dopolnitelnyy-den/"
## [26] "https://www.mr-info.ru/22682-esche-odna-popytkanachalas-registracija-na-
uchastie-v-egje-v-2019-godu.html"
## [27] "http://sakhaday.ru/news/shamaevy-o-shkole-ajyy-kyhata/"
## [28] "https://na.ria.ru/20190204/1550336222.html"
## [29] "https://rosregistr.ru/interesnoe/218384.html"
## [30] "https://rosregistr.ru/interesnoe/c222244.html"
## [31] "https://tass.ru/info/5240176"
## [32] "https://www.mvestnik.ru/newslent/v-2019-godu-edinaya-rossiya-prokontroliruet-
vyplatu-uchitelyam-kompensacij-za-gia/"
## [33] "https://www.souzveche.ru/articles/tribune_deputy/45757/"
## [34] "https://www.ridus.ru/news/287818"
## [35] "https://tass.ru/obschestvo/5954091"
## [36] "https://news.rambler.ru/education/41411963-solovev-sravnil-edinyy-gosudarstvennyy-
ekzamen-s-bludom/"
## [37] "https://snob.ru/entry/163153"
## [38] "http://www.tv21.ru/news/2018/06/06/v-murmanskoy-oblasti-proshel-edinyy-
gosudarstvennyy-ekzamen-po-russkomu-yazyku"
## [39] "https://iz.ru/789104/2018-09-14/rosobrnadzor-ne-budet-vnosit-sereznye-
izmeneniia-v-ege-v-2019-godu"
## [40] "https://tass.ru/obschestvo/5795056"
## [41] "https://life.ru/t/%25D0%25BD%25D0%25BE%25D0%25B2%25D0%25BE%25D1%2581%25D1
## [42] "https://kudago.com/all/news/edinyij-gosudarstvennyij-ekzamen/"
## [43] "https://iz.ru/811685/2018-11-13/ege-dlia-roditelei-organizovali-v-orenburgskoi-
oblasti"
## [44] "https://www.riadagestan.ru/news/nogayskiy_rayon/edinyy_gosudarstvennyy_ekzamen_po_
## [45] "https://iz.ru/798949/2018-10-10/nazvany-samye-slozhnye-ege-dlia-zakliuchennykh-
v-2018-godu"
## [46] "https://news.rambler.ru/education/40236978-esche-odin-obyazatelnyy-ege-
vvedut-v-rossii/"
## [47] "https://regnum.ru/news/2456750.html"
## [48] "https://iz.ru/762644/2018-07-03/v-rossii-vvedut-obiazatelnyi-ege-po-inostrannomu-
iazyku-v-2022-godu"
## [49] "https://ria.ru/20180927/1529479714.html"
## [50] "https://newizv.ru/news/society/15-06-2018/kolledzh-kak-lazeyka-do-10-
abiturientov-postupayut-v-vuzy-v-obhod-ege"
## [51] "https://www.pnp.ru/social/v-ldpr-predlozhili-otmenit-edinyy-gosudarstvennyy-
ekzamen.html"
## [52] "https://www.rbc.ru/rbcfreenews/5c0b968d9a7947150261d921"
## [53] "https://www.currenttime.tv/a/29341975.html"
## [54] "https://www.pnp.ru/social/stali-izvestny-daty-provedeniya-gia-i-ege-v-2019-
godu.html"
## [55] "https://regnum.ru/news/2433477.html"
## [56] "https://regnum.ru/news/2529427.html"
## [57] "https://www.vladtime.ru/shou_biznes/648907"
## [58] "https://www.mk.ru/social/2018/11/06/vasileva-vyskazalas-protiv-dobrovolnosti-
ege.html"
## [59] "https://kirovpravda.ru/%25D0%25B5%25D0%25B4%25D0%25B8%25D0%25BD%25D1%258B
%25D0%25B3%25D0%25BE%25D1%2581%25D1%2583%25D0%25B4%25D0%25B0%25D1%2580%25D1
```

```
%25D1%258D%25D0%25BA%25D0%25B7%25D0%25B0%25D0%25BC%25D0%25B5%25D0%25BD-%25D0%25B2-%25D0%25BA%25D0%25B8%25D1%2580/"
## [60] "https://na.ria.ru/20181022/1531069799.html"
## [61] "https://life.ru/t/%25D0%25BD%25D0%25BE%25D0%25B2%25D0%25BE%25D1%2581%25D1
## [62] "https://www.pnp.ru/social/rezultaty-ege-po-matematike-profilnogo-urovnya-obyavyat-do-13-iyunya.html"
## [63] "https://tsargrad.tv/news/vasileva-ocenka-egje-javljaetsja-neobosnovannoj-v-otsutstvie-standartov-obrazovanija_142151"
## [64] "https://russian.rt.com/russia/news/498050-ege-angliiskii-uproschenie"
## [65] "https://tass.ru/obschestvo/5719857"
## [66] "http://www.amur.info/news/2018/10/18/144832"
## [67] "https://www.gorodche.ru/news/society/115126/"
## [68] "https://ria.ru/20180604/1522045397.html"
## [69] "https://utv.ru/material/uchitelnicu-iz-bashkirii-posadili-v-tyurmu-za-pomosh-pri-sdache-ege/"
## [70] "https://www.m24.ru/news/obrazovanie/26112018/55636"
## [71] "https://ria.ru/20180404/1517900317.html"
## [72] "https://novostivolgograda.ru/news/society/20-09-2018/osenyu-v-rossiyskih-shkolah-protestiruyut-ege-po-kitayskomu-yazyku"
## [73] "https://www.novayagazeta.ru/articles/2018/02/11/75479-ekzamen-dlya-pamyati"
## [74] "https://www.kommersant.ru/doc/3592580"
## [75] "https://www.alt.kp.ru/daily/26818/3854525/"
## [76] "http://dubna.ru/article/2018/02/probnyy-edinyy-gosudarstvennyy-ekzamen-ege-po-russkomu-yazyku"
## [77] "https://www.eg.ru/culture/494449/"
## [78] "https://meduza.io/news/2018/08/13/polnuyu-proverku-ege-doveryat-iskusstvennomu-intellektu-k-2030-godu"
## [79] "https://regnum.ru/news/2434416.html"
## [80] "https://life.ru/t/%25D0%25BD%25D0%25BE%25D0%25B2%25D0%25BE%25D1%2581%25D1
## [81] "https://kikonline.ru/2018/10/19/v-2020-godu-devyatiklassniki-budut-sdavat-ekzamenyi-po-novomu/"
## [82] "https://tass.ru/obschestvo/5268308"
## [83] "https://www.pnp.ru/social/rosobrnadzor-usovershenstvuet-edinyy-gosudarstvennyy-ekzamen-po-obrashheniyam-grazhdan.html"
## [84] "https://www.mos.ru/news/item/40822073/"
## [85] "https://tass.ru/obschestvo/5023405"
## [86] "https://www.nakanune.ru/articles/114017/"
## [87] "https://newizv.ru/news/society/15-08-2018/s-2019-goda-prinimat-ege-po-informatike-budut-kompyutery"
## [88] "https://ria.ru/20181206/1547492473.html"
## [89] "https://iz.ru/763367/2018-07-05/bolee-75-rossiian-obviniaiut-ege-v-ukhudshenii-kachestva-znanii-shkolnikov"
## [90] "http://tvolk.ru/news/society/ege-2018-vybor-budushchego-/"
## [91] "http://ysia.ru/v-yakutii-uluchshilis-rezultaty-ege-po-biologii/"
##  [92] "https://regnum.ru/news/2439178.html"
##       [93]    "https://news.rambler.ru/education/39758144-ege-2018-kogda-nachinaetsya-
osnovnoy-etap-ekzamenov/"
##  [94] "https://regnum.ru/news/2527059.html"
##    [95]  "http://dontr.ru/novosti/samym-slozhnym-bylo-sochinenie-donskie-vypuskniki-
sdavali-ege-po-russkomu-yazyku/"
##  [96] "https://www.nakanune.ru/news/2018/03/27/22502499/"
##      [97]    "https://vesti22.tv/video/shkolnik-s-bioprotezami-yaroslav-shishov-na-domu-
sdaval-ege"
##    [98]  "http://in-reutov.ru/novosti/obrazovanie/edinyy-gosudarstvennyy-ekzamen-ege-
nachinaetsya-v-podmoskove"
##  [99] "https://www.hab.kp.ru/daily/26857.4/3898804/"
## [100] "https://www.ural56.ru/news/578572/"
```

1.2. Saving scraped multiple html links into the html files in R:

Let's create the vector of first 10 scraped web pages about единый государственный экзамен:

```
links <- c(
"https://regnum.ru/news/2569977.html",
"http://www.ugorizont.ru/2019/02/07/okolo-90-tyisyach-shkolnikov-sdadut-edinyiy-gosudarstvennyiy-ekz
"https://vm.ru/news/590807.html",
"https://www.mos.ru/news/item/50942073/",
"https://spbvedomosti.ru/news/country_and_world/nazvan_samyy_populyarnyy_predmet_dlya_sdac
"http://zyuzinomedia.ru/news/moskovskie-novosti/pochti-300-chelovek-sdadut-edinyy-gosudarstvennyy-e
"https://korolevriamo.ru/article/237934/glavnoe-shkolnoe-ispytanie-sem-mifov-o-ege-i-novovvedeniya-201
"https://life.ru/t/%25D0%25BD%25D0%25BE%25D0%25B2%25D0%25BE%25D1%2581%25D1%2582%2
"http://vechorka.ru/article/poznakomimsya-s-raspisaniem-ege-2019/",
"https://dostup1.ru/interview/Elena-Tyurina-EGE--eto-po-bolshomu-schetu-sudba-cheloveka_113563.htm
```

Using vector links we now can save html.files of our scraped data into the working directory:

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

1.3. Create a Corpus from many html files in R:

1.3.1. First variant, using htmlToText function:

```
#setwd() # this folder has your HTML files
html <- list.files(pattern="\\.(htm|html)$") # get just .htm and .html files
```

```
# load packages
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
```

```
# get some code from github to convert HTML to text
#writeChar(con="htmlToText.R", (getURL(ssl.verifypeer = FALSE, "https://raw.github.com/tonybreya
#source("htmlToText.R")
```

```
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="cacer
      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::noscript)][
    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)
}
```

```
# convert HTML to text
html2txt <- lapply(html, htmlToText)
```

```
# clean out non-ASCII characters
html2txtclean <- sapply(html2txt, function(x) iconv(x, "latin1", "ASCII", sub=""))
```

```
# make corpus for text mining
corpus <- Corpus(VectorSource(html2txtclean))
```

Let's check the corpus we recived:

```
dtm_matrix<-DocumentTermMatrix(corpus)
use_dtm<-as.matrix(dtm_matrix)
```

It happenes that this htmlToText function produce not enouth clean results. with a lot of unreadable symbols, that prevent R from reading cyrilics. That is why I found other codes, that produce more clean data.

1.3.2. Second variant, using htmlToText function:

```
setwd("C:/Users/Dmitry/Dropbox/TEACHING/SNA/Lab 5/Lab 5")
doc.html = htmlTreeParse("filename_1.html", useInternal = TRUE)
doc.text = unlist(xpathApply(doc.html, '//p', xmlValue))
doc.text = gsub('\\n', ' ', doc.text)
doc.text = paste(doc.text, collapse = ' ')
doc.text
```

```
## [1] "<U+041C><U+043E><U+0441><U+043A><U+0432><U+0430>,          11 <U+0444><
80-75          Email: sfo@regnum.ru          <U+0421><U+0431><U+0440><U+043E><U+0441> <U
```

```
doc.html = htmlTreeParse("filename_1.html", useInternal = TRUE)
doc.text = unlist(xpathApply(doc.html, '//p', xmlValue))
doc.text = gsub('\\n', ' ', doc.text)
doc.text = paste(doc.text, collapse = ' ')
write.csv (doc.text, file="1.csv")
```

To make this process automatically let's write the loop:

```
for (i in 1:10){
setwd("C:/Users/Dmitry/Dropbox/TEACHING/SNA/Lab 5/Lab 5")
fl1<-paste("filename_",i,".html", sep="")
doc.html = htmlTreeParse(fl1, useInternal = TRUE)
doc.text = unlist(xpathApply(doc.html, '//p', xmlValue))
doc.text = gsub('\\n', ' ', doc.text)
doc.text = paste(doc.text, collapse = ' ')
fl<-paste(i,".csv")
write.csv (doc.text, fl)
}
```

This loop may not work in RMarkdown, because it change the working directory, so, run it in RScript.

Now, we have to join all csv.files into one:

```
library(data.table)
filenames <- list.files("C:/Users/Dmitry/Dropbox/TEACHING/SNA/Lab 5/Lab 5", pattern="*.csv", fu
data <- rbindlist(lapply(filenames,fread))
write.csv(data, file="1_10.csv")
```

Let's open the file 1_10.csv. It is consist of non-Russian latters. To correct the problem we have to change the code:

```
encoding = "utf-8"
Sys.setlocale("LC_CTYPE", "russian")
```

```
## [1] "Russian_Russia.1251"
```

```
write.csv(data, file="1_10.csv")
```

In the file 1_10.csv Data, From File, From Text/CSV; Import the File again; In File Origin choose **1251: Cyrilic (Windows), press Edit; Keep; Save. May be you need to Save as the file as separate one, and delete needless rows and columns.

1.4. Mining opertions:

First, we create corpus:

```
library(readxl)
encoding = "utf-8"
Sys.setlocale("LC_CTYPE", "russian")
```

```
## [1] "Russian_Russia.1251"
```

```
use <- read_excel("1_10_f.xlsx")
```

```
use_text<-use$Text
library("tm")
library("NLP")
use_vec<-VectorSource(use_text)
use_corpus<-VCorpus(use_vec)
```

Second, we create DocumentToTerm and TermToDocument matrices:

```
dtm_matrix<-DocumentTermMatrix(use_corpus)
use_dtm<-as.matrix(dtm_matrix)
tdm_matrix<-TermDocumentMatrix(use_corpus)
use_tdm<-as.matrix(tdm_matrix)
```

Let's check the matrices:

```
head(use_tdm)
```

```
##                    Docs
## Terms               1 2 3 4 5 6 7 8 9 10
##   —георгий          0 0 0 0 1 0 0 0 0  0
##    отправить        0 0 0 0 0 0 0 0 1  0
##   ""вечерняя        0 0 0 0 2 0 0 0 0  0
##   ""ctrl+enter""    0 0 0 0 0 0 0 1 0  0
##   ""медиа           0 0 0 0 0 0 0 0 1  0
##   ""подписаться"",  0 0 2 0 0 0 0 0 0  0
```

```
dim(use_tdm)
```

```
## [1] 2572   10
```

```
dim(use_dtm)
```

```
## [1]   10 2572
```

Our matrix consist of 2572 terms in 10 documents. It is happened that we have a lot of non-Russian, and non-words symbols, as stopwards, and punctuation, etc. You can clean them up, using different comands in R.

Third, Data cleaning:

Let's first have a look on the most frequent words:

```
term_frequency <- colSums(use_dtm)
term_frequency <- sort(term_frequency, decreasing = TRUE)
term_frequency[1:20]
```

```
##     что     егэ     для     это экзамен    тыс.     как    года   будет
##      44      37      33      32      25      22      21      20      19
##    2019    если   нужно     все     или    есть   июня,    году     при
##      18      17      16      15      15      14      14      13      13
##     уже   чтобы
##      13      12
```

Then, let's write function to remove whitespaces, punctuation, numbers, stopwords, and tolower:

```
clean_corpus <- function(corpus){
    corpus <- tm_map(corpus, stripWhitespace)
    corpus <- tm_map(corpus, removePunctuation)
    corpus <- tm_map(corpus, content_transformer(tolower))
    corpus <- tm_map(corpus, removeNumbers)
    corpus <- tm_map(corpus, removeWords, c(stopwords("ru")))
    return(corpus)
}
clean_corp <- clean_corpus(use_corpus)
dtm_matrix_cl2<-DocumentTermMatrix(clean_corp)
use_dtm_cl2<-as.matrix(dtm_matrix_cl2)
tdm_matrix_cl2<-TermDocumentMatrix(clean_corp)
use_tdm_cl2<-as.matrix(tdm_matrix_cl2)
head(use_tdm_cl2)
```

```
##             Docs
## Terms        1 2 3 4 5 6 7 8 9 10
##   —георгий   0 0 0 0 1 0 0 0 0  0
##   «“мэш”     0 0 0 0 0 1 0 0 0  0
##   «вечерняя  0 0 0 0 3 0 0 0 0  0
##   «базу»     0 0 0 0 0 0 0 0 0  1
##   «единой    0 0 0 0 0 0 0 0 0  1
##   «двойкой»  0 0 1 0 0 0 0 0 0  0
```

```
dim(use_tdm_cl2)
```

```
## [1] 2070   10
```

```
dim(use_dtm_cl2)
```

```
## [1]   10 2070
```

```
term_frequency_cl2 <- colSums(use_dtm_cl2)
term_frequency_cl2 <- sort(term_frequency, decreasing = TRUE)
term_frequency_cl2[1:20]
```

```
##     что     егэ     для     это экзамен    тыс.     как    года   будет
##      44      37      33      32      25      22      21      20      19
##    2019    если   нужно     все     или    есть   июня,    году     при
##      18      17      16      15      15      14      14      13      13
##     уже   чтобы
##      13      12
```

It seems that there are a lot of cleaning things are remaining to do, but I will leave this up to you to find appropriate methods. Feel free to use GOOGLE.

Fourth, we also can make n-grams for our text analysis:

```
#install.packages("RWeka")
library("RWeka") # for this library you nedd earliest version of 64-bit Java version
```

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

```
tokenizer<-function(x)
NGramTokenizer(x, Weka_control(min=2, max=2))
bigram_tdm<-TermDocumentMatrix(clean_corp,control=list(tokenize=tokenizer))
use_tdm_bi<-as.matrix(bigram_tdm)
bigram_dtm<-DocumentTermMatrix(clean_corp,control=list(tokenize=tokenizer))
use_dtm_bi<-as.matrix(bigram_dtm)
```

1.5. Network Analysis of Text Data:

```
c<-use_tdm_cl2 %*% use_dtm_cl2
#dim(c)
#library(igraph)
#NET <- graph_from_adjacency_matrix(c)
#par(mar=c(0,0,0,0))
#plot(NET)
```

It takes a lot of time for R to produce the plot. So better to use Pajek.

```
encoding = "utf-8"
Sys.setlocale("LC_CTYPE", "russian")
```

```
## [1] "Russian_Russia.1251"
```

```
write.csv(x=c, file="c.csv")
```

Use .csv file to create .net file for Pajek. Network - Create new network - Transform - Remove - Loops, Lines with value. Experiment with partitions: Network - Create partition - k-Core, etc. Draw - Network+Partition. Experiment with Layouts.