Individual Assignment by Ian Zhang 54505696

1.Research Question & Intro

1.1 Research Question

In this individual assignment, I did a Russian state image research from different medium sources including RT(Russia Today) and Western medium including(CNN,BBC,NY times,WashingtonPost)

The research question is **"how do RT and Western Medium differ in their reports about Russia".**

1.2 Purposes & Reasons

By comparing the textual differences between different medium sources, this research aims to find how Russian official media and mainstream Western Medium did in reporting topic choices.

The reason I chose this question is recently, many medium researchers criticized RT. Though RT treated "Question More" as their slogan, Western media doubted it is a **"Kremlin-funded English-language television channel"**.If that is true, It's pretty similar with Chinese medium,so I hope to use textual analysis to reason this.

More info: *https://en.wikipedia.org/wiki/RT_(TV_network)*

2.1.1 Get all searching urls and put them into a vector

```
library(rvest)
library(httr)
library(stringr)

cookies = readRDS("cookies.rds")
user_agent = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_1) AppleWebKit/537.36 (KHTML, like Gecko) Chr

url_list = c()
title_list = c()

for(i in seq(10))
{
url = paste0("https://www.rt.com/search?q=Russia&type=&df=&dt=&page=",i)
r = GET(url,set_cookies(cookies),
        user_agent(user_agent))

text = iconv(list(r$content))
text_tree = read_html(text)

text_raw = text_tree%>% html_nodes(css = ".card-list__item") %>% html_text(" ")
text_clean = gsub("[\r\n]", "", text_raw)
text_url = str_extract(text_clean,"https:.+.\\/")
title = text_tree %>% html_nodes(css = ".card__header") %>% html_text("")
title_clean = gsub("[\r\n]", "", title)
title_clean = trimws(title_clean)

url_list = c(url_list,text_url)
title_list = c(title_list,title_clean)
}
```

2.1.2 Pageniations & Get contents

```
for(url in seq(length(url_list)))
{
print(url)
url_u = url_list[url]
r2 = GET(url = url_u)
text_2 = iconv(list(r2$content))
text_2_tree = read_html(text_2)

text_raw_2 = text_2_tree %>% html_nodes(css = ".article__text") %>% html_text(" ")
text_clean_2 = gsub("[\r\n]", "", text_raw_2)
if(identical(text_clean_2, character(0)))
{
  text_clean_2 = "none"
}
text_list = c(text_list,text_clean_2)
}

head(text_list)
saveRDS(text_list,"RT_text_list.rds")
```

2.2 Scraping Western Medium

As I mentioned above. It's pretty hard to scrape Western medium pages as they use JSP to generate pages and have well-built anti-crwaler machinism. So I have to use Facebook API to scrape their posts in their official Facebook pages. By applying Regex, I extracted those posts which contain the word "Russia".

```
library(Rfacebook)
app_id = "848961181910429"
app_secret = "36ea66c85eea3f1b14b53111ffde4d86"
token = Rfacebook::fbOAuth(app_id,app_secret)

token = readRDS("token.rds")
posts = getPage("cnninternational", token = token, n = 1000)
posts2 = getPage("bbcnews",token,n = 1000)
posts3 = getPage("nytimes",token,n = 1000)
posts4 = getPage("washingtonpost",token,n = 1000)

posts_a = subset(posts, !is.na(posts$message[(str_extract(posts$message,"Russia")) == "Russia"]))
posts_b = subset(posts2, !is.na(posts2$message[(str_extract(posts2$message,"Russia")) == "Russia"]))
posts_c = subset(posts3, !is.na(posts3$message[(str_extract(posts3$message,"Russia")) == "Russia"]))
posts_d = subset(posts4, !is.na(posts4$message[(str_extract(posts4$message,"Russia")) == "Russia"]))

posts_a["source"] = "cnninternationa"
posts_b["source"] = "bbcnews"
posts_c["source"] = "nytimes"
posts_d["source"] = "washingtonpost"
post_russia_total = rbind(posts_a,posts_b,posts_c,posts_d)
saveRDS(post_russia_total,"Western_Russia.rds")

post_russia_total = readRDS("Western_Russia.rds")
```

3 Analysing the Data

3.1 Cleaning up and combine the two data sets

Combining those data to a dataframe.

```
Western_russia = readRDS("Western_Russia.rds")
RT_russia = readRDS("RT_text_list.rds")

RT_russia_data = data.frame(RT_russia)
RT_russia_data["source"] = "RT"

Western_russia_data = Western_russia[c("message","source")]
colnames(RT_russia_data) = c("message","source")
combined = rbind(RT_russia_data,Western_russia_data)

text_from_RT = combined$message[combined$source == "RT"]
text_from_Western = combined$message[combined$source != "RT"]
as.character(text_from_RT[1]) # Example from RT
```

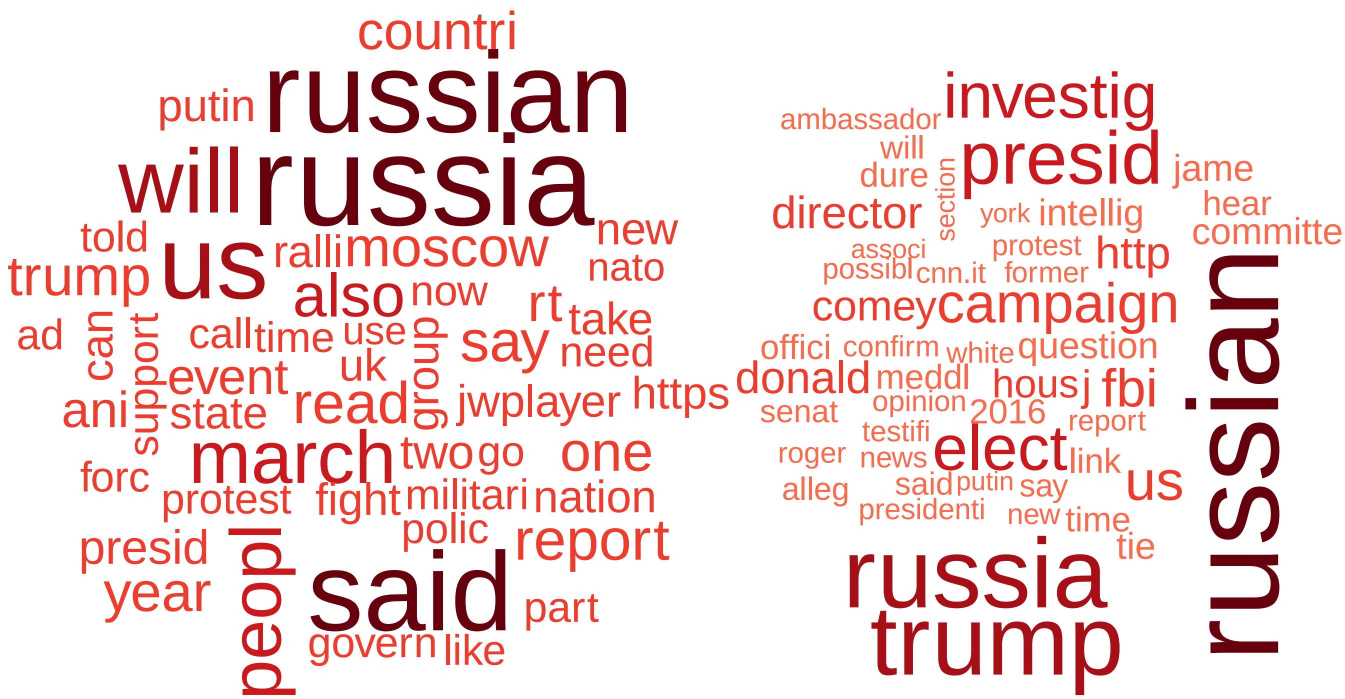
```
## [1] "‘Thanks for separating Crimea and Ukraine’The forum’s moderator Geoff Cutmore of CNBC couldn’t refrai
IN THE NOW (@IntheNow_tweet) March 30, 2017"Once Reagan, discussing taxes, addressed Americans, saying ‘Read
No’!" the Russian leader said.‘Hot Finnish guys’Finland’s leader, Sauli Niinisto, addressed the issue of hea
Ruptly (@Ruptly) March 30, 2017"Need help?" the Russian leader asked, jokingly offering military assistance
RT (@RT_com) March 30, 2017"
```

```
as.character(text_from_Western[1]) # Example from Western Medium
```

```
## [1] "Neither of Jared Kushner's meetings - with the Russian ambassador and with a Russian banker -
were about sanctions, a source says."
```

3.2 Conduct a textual analysis of at least one of the data sets

Getting the topic words and wordclouds




```
##  russia russian    said      us    will   march   peopl    also  report
##     274     243     236     216     180     143     136     113     107
##     say
##     105
```

```
##   russian   russia    trump   presid investig    elect       us campaign
##        55       40       40       29       23       23       20       19
##       fbi   donald
##        18       15
```

From two Wordclouds and Topfeatures words we could know, first one the reports from RT also mentioned the "US","march","protest"(happened in Russia recently) and so on. But from Western medium we could see it emphasized the "Trump", "President","Investigate", which means that maybe most reports are involved around the US-Russia relationship or Trump's business with Russia.

Based on the above hypothesis, I conduct a the exploratory analysis and topic model analysis which will prove this point.

3.3 Simple Exploratory Analysis

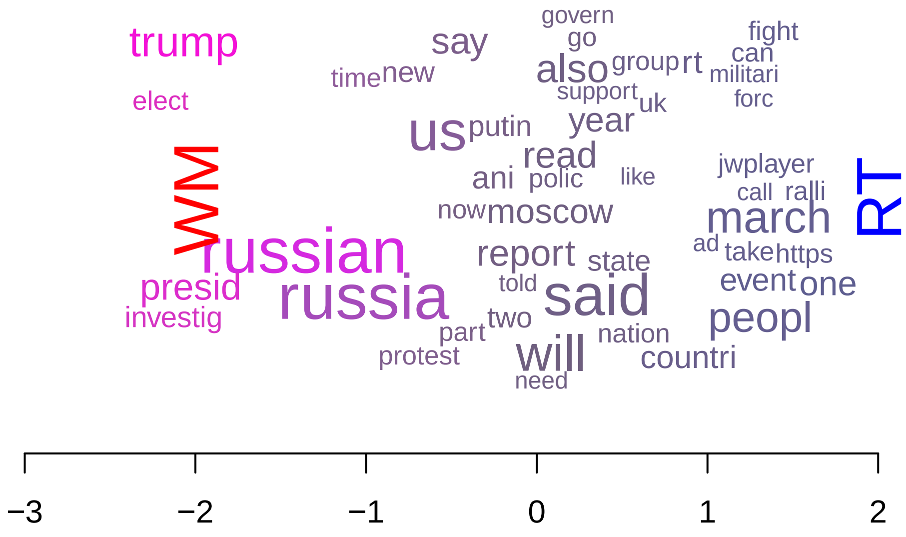
Using corpora.compare we could get what are overrepresentted in both data frame,then using dictionary analysis to prove what has been emphasized in both corpus(number of items per source or per target).

```
cmp = corpustools::corpora.compare(dtm_text_RT,dtm_text_Western)
cmp = cmp[order(cmp$over, decreasing = F), ]

h = rescale(log(cmp$over), c(1, .6666))
s = rescale(sqrt(cmp$chi), c(.25,1))
cmp$col = hsv(h, s, .33 + .67*s)

cmp = arrange(cmp, -termfreq)
with(head(cmp, 50), plotWords(x=log(over), words=term, wordfreq=termfreq, random.y = T, col=col, scale=
```

```r
text(-2, 0, "WM", srt=90, col="red", cex=2)
text(2, 0, "RT", srt=90, col="blue", cex=2)
title(xlab="Overrepresentation")
```




```r
d = dictionary(list(trump=c("investig*", "trump"), protest=c("protest*", "putin")))
dfm_2 = dfm(corpus(as.character(text_from_RT)),dictionary = d)
dfm_3 = dfm(corpus(as.character(text_from_Western)),dictionary = d)

k1 = data.frame(dfm_2)
k2 = data.frame(dfm_3)

sum(k1$trump)
```

```
## [1] 128
```

```r
sum(k1$protest)
```

```
## [1] 135
```

```r
sum(k2$trump)
```

```
## [1] 53
```

```r
sum(k2$protest)
```

```
## [1] 11
```

```r
rt_trump_protest_ratio = sum(k1$trump)/sum(k1$protest)
western_trump_protest_ratio = sum(k2$trump)/sum(k2$protest)

print(rt_trump_protest_ratio)
```

```
## [1] 0.9481481
```

```r
print(western_trump_protest_ratio)
```

```
## [1] 4.818182
```

The overrepresentation analysis shows us that Western media seems to just emphasize the what Trump did with Russia, but From RT we could find there are lots of topic have been covered, they didn't avoid to report domestic Bad news including mock protest. **From dictionary analysis it would be more clear: Western media mentioned Trump 53 times in their report but just report protest 11 times. However, for RT both Trump and Protest just are treated as common topics.**

3.4 Topic Modeling Analysis

3.4.1 Creating topic models

```
library(corpustools)
d = convert(dfm, to="topicmodels")
m = LDA(d, k = 10, method = "Gibbs", control=list(alpha=.1, iter=100))
saveRDS(m,"topic.rds")

d2 = convert(dtm_text_Western, to="topicmodels")
m2 = LDA(d2, k = 10, method = "Gibbs", control=list(alpha=.1, iter=100))
saveRDS(m2,"topic2.rds")
```

3.4.2 Combining topics into a data frame and analysis

After finding the topics, I set a threshold(0.1) to check whether those articles are related to this topic or not and I make a dataframe to load them.

```
m = readRDS(file = "topic.rds")
m2 = readRDS(file = "topic2.rds")

tpd = posterior(m)$topics
tpd = as.data.frame(tpd)
colnames(tpd) = c("UK","Fighter","Nuclear","Oil","Hockey","Moscow_Protest","Trump","Ukrainian","Putin",


tpd2 = posterior(m2)$topics
tpd2 = as.data.frame(tpd2)
colnames(tpd2) = c("Moscow_Protest","Meddle_Election","America_Russia","America_Russia","Moscow_Protest
print(head((tpd)))
```

```
##               UK      Fighter      Nuclear          Oil       Hockey
## text1 0.16881720 0.0146953405 0.0326164875 0.0003584229 0.0577060932
## text2 0.02720307 0.0003831418 0.8624521073 0.0003831418 0.0003831418
## text3 0.07854478 0.2016791045 0.0151119403 0.0039179104 0.0001865672
## text4 0.14518072 0.0006024096 0.0006024096 0.0126506024 0.0006024096
## text5 0.14464286 0.0017857143 0.0017857143 0.0196428571 0.0017857143
## text6 0.31033210 0.0003690037 0.0003690037 0.5391143911 0.0003690037
##       Moscow_Protest        Trump    Ukrainian      Putin        Syria
## text1   0.0003584229 0.0003584229 0.0003584229 0.69569892 0.0290322581
## text2   0.0157088123 0.0042145594 0.0118773946 0.06934866 0.0080459770
## text3   0.0001865672 0.3751865672 0.0244402985 0.30055970 0.0001865672
## text4   0.0006024096 0.0006024096 0.0006024096 0.25963855 0.5789156627
## text5   0.2339285714 0.0375000000 0.1267857143 0.43035714 0.0017857143
## text6   0.0003690037 0.0003690037 0.0003690037 0.14797048 0.0003690037
```

```
print(head((tpd2)))
```

```
##       Moscow_Protest Meddle_Election America_Russia America_Russia
```

```
## text1    0.341666667     0.008333333    0.258333333    0.008333333
## text2    0.009090909     0.009090909    0.009090909    0.009090909
## text3    0.003571429     0.003571429    0.003571429    0.003571429
## text4    0.005000000     0.005000000    0.905000000    0.005000000
## text5    0.005882353     0.005882353    0.947058824    0.005882353
## text6    0.052380952     0.004761905    0.004761905    0.004761905
##       Moscow_Protest Meddle_Election Moscow_Protest Russia_Hacker
## text1    0.008333333     0.008333333    0.008333333   0.008333333
## text2    0.009090909     0.009090909    0.009090909   0.009090909
## text3    0.003571429     0.003571429    0.003571429   0.003571429
## text4    0.005000000     0.005000000    0.005000000   0.055000000
## text5    0.005882353     0.005882353    0.005882353   0.005882353
## text6    0.004761905     0.004761905    0.290476190   0.004761905
##       America_Russia America_Russia
## text1    0.341666667    0.008333333
## text2    0.009090909    0.918181818
## text3    0.003571429    0.967857143
## text4    0.005000000    0.005000000
## text5    0.005882353    0.005882353
## text6    0.004761905    0.623809524
```

```
topicmodel = function(tpd)
{
  list_freq = c()
  for (i in seq(length(colnames(tpd))))
  {
    list_freq = c(list_freq,colnames(tpd)[i])
    list_freq = c(list_freq,(table(tpd[,i]>=0.1)))
    freq = as.matrix(list_freq)
  }
  return(freq)
}

tpd_overrepresentation = topicmodel(tpd)
tpd2_overrepresentation = topicmodel(tpd2)
df = data.frame(tpd_overrepresentation,tpd2_overrepresentation)
print(df)
```

```
##    tpd_overrepresentation tpd2_overrepresentation
## 1                      UK          Moscow_Protest
## 2                      65                      78
## 3                      35                      14
## 4                 Fighter         Meddle_Election
## 5                      77                      68
## 6                      23                      24
## 7                 Nuclear          America_Russia
## 8                      77                      72
## 9                      23                      20
## 10                    Oil          America_Russia
## 11                     80                      82
## 12                     20                      10
## 13                 Hockey          Moscow_Protest
## 14                     92                      84
## 15                      8                       8
## 16         Moscow_Protest         Meddle_Election
```

```
## 17                        79                 62
## 18                        21                 30
## 19                     Trump     Moscow_Protest
## 20                        74                 82
## 21                        26                 10
## 22                 Ukrainian      Russia_Hacker
## 23                        93                 74
## 24                         7                 18
## 25                     Putin     America_Russia
## 26                        43                 64
## 27                        57                 28
## 28                     Syria     America_Russia
## 29                        76                 71
## 30                        24                 21
```

From the above table,we could know Putin. UK(Brexit),Syria,Moscow Protest are top topics for RT. But for western medium, they just focus on recent America Russia relationship and whether Russia meddles the election of president.

Conclusion

From this individual assignment, By analysing and comparing the text difference between the report of RT and Western Medium, I get the primary answer to the research question **"how do RT and Western Medium differ in their reports about Russia".**.

Generally Speaking, from reports of RT I could see they tend to cover every aspect which has relations with Russia, as their topics and overrepresetation are pretty various, which including domestic issues and other countries issues. They are 2 features I could conclude

- RT likes to report those countries which have **national interest** with its(UK,Syria,US).
- They also like to emphasize their president Putin(in topic analysis 57 articles occured Putin).

So it has the features of a national media agency(like Xinhua Agency in China). However, I have not found any clues from data about **"Kremlin-funded English-language television channel"**, as they also treat their domestic turmoil also as a important topic. However, due to I didn't analyze the sentiment of those articles, so that would be my improvement part to do further analysis.

However,western medium has a bias when reports Russia. From the data, we could know recently they **just focus on the relationship between Trump and Russia**. It deeply explained how does medium in different countries cater to their domestic readers.