UFO Report

Takahiro Yamada

12/5/2021

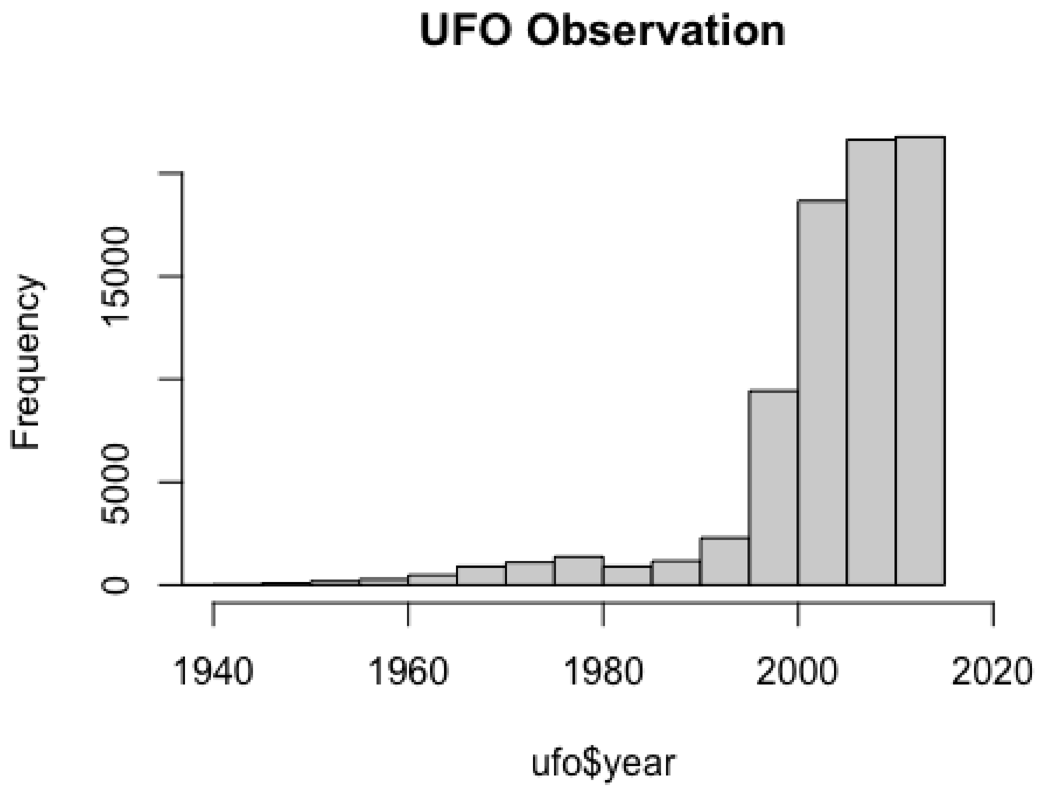
Introduction

X-file is one of my favorite American dramas. When I was a child, I watched its series of videos and wondered if the UFO and Ariens really exist!? There is no evidence to prove they exist, but at the same time, there is no clue that they do not exist. While I was clowning on the website, I found a UFO observation report. In this assignment, I tried to figure out the observation report contents and compare its description with the famous Roswell Report created by U.S. Airforce.

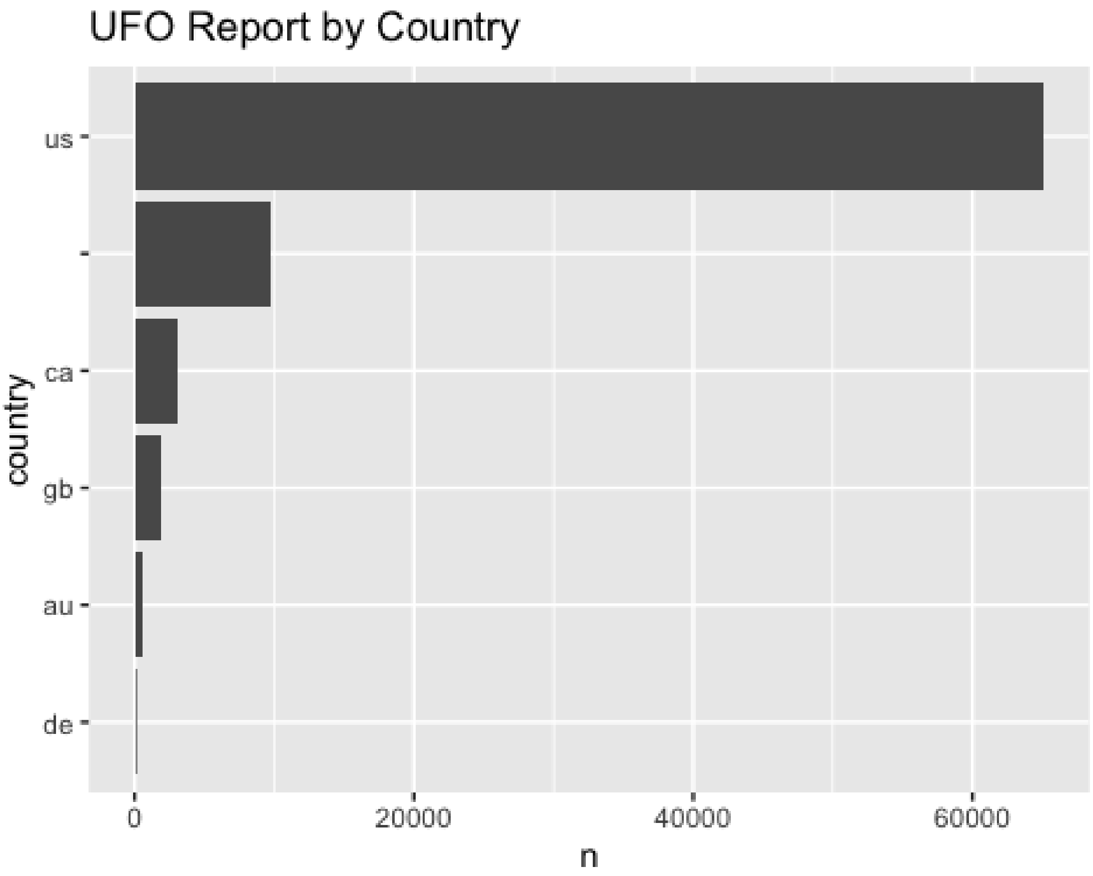
UFO Observation Report

Observation number shift According to this dataset, the UFO report numbers have been skyrocketing from the second half of the 1900s. The timing overlaps with the famous science fiction movies like E.T. (1982,) STAR WARS (1972~2019), Men in Black (1997~2019). The American movie culture would contribute to the UFO report numbers. Country and State Frequent word analysis by country Looking at the UFO observed location, the U.S. is the most. Canada, the U.K., and Australia are the following. However, this is because the dataset is collecting reports written in only English, and if it is possible to collect and combine the multi-language Report, the result could be changed.

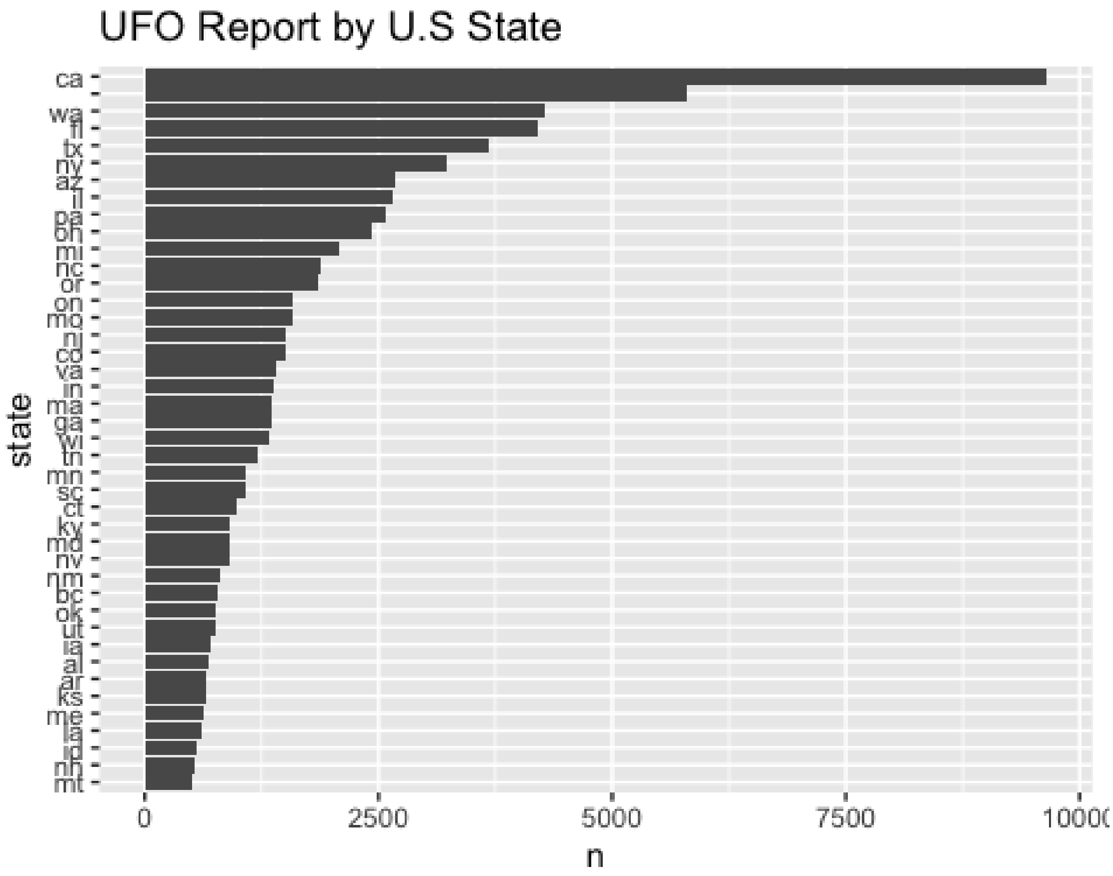
```r
##########################################
## Overview World / US
##########################################
hist(ufo$year,
     main = "UFO Observation",
     xlim=c(1940,2020),
     breaks = 20)
```



```r
ufo_w <- ufo %>%
  select(country) %>%
  count(country) %>%
  mutate(country = reorder(country ,n ))

ggplot(ufo_w ,aes(country, n))+
  geom_col()+
  ggtitle("UFO Report by Country")+
  coord_flip()
```



```
ufo_us <- ufo %>%
  select(state) %>%
  count(state) %>%
  mutate(state = reorder(state ,n )) %>%
  filter(n>500)

ggplot(ufo_us ,aes(state, n))+
  geom_col()+
   ggtitle("UFO Report by U.S State")+
  coord_flip()
```

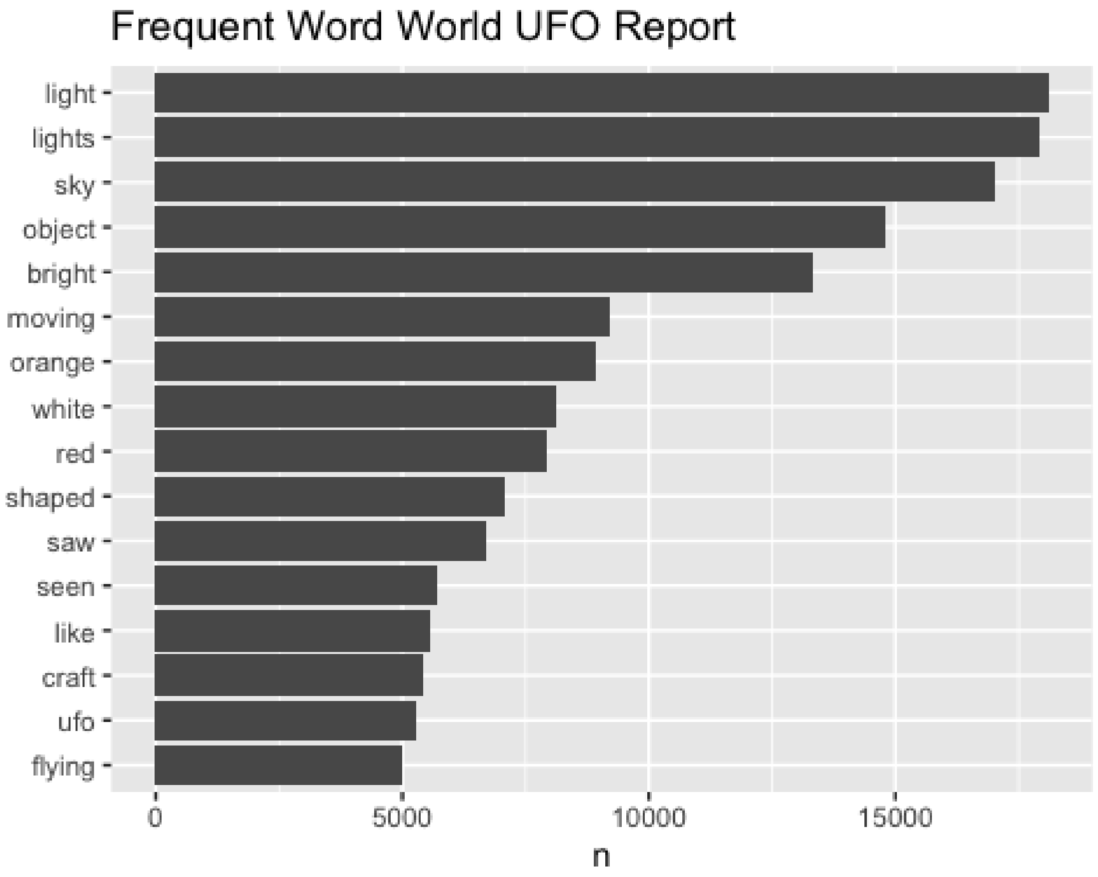


Frequent Words

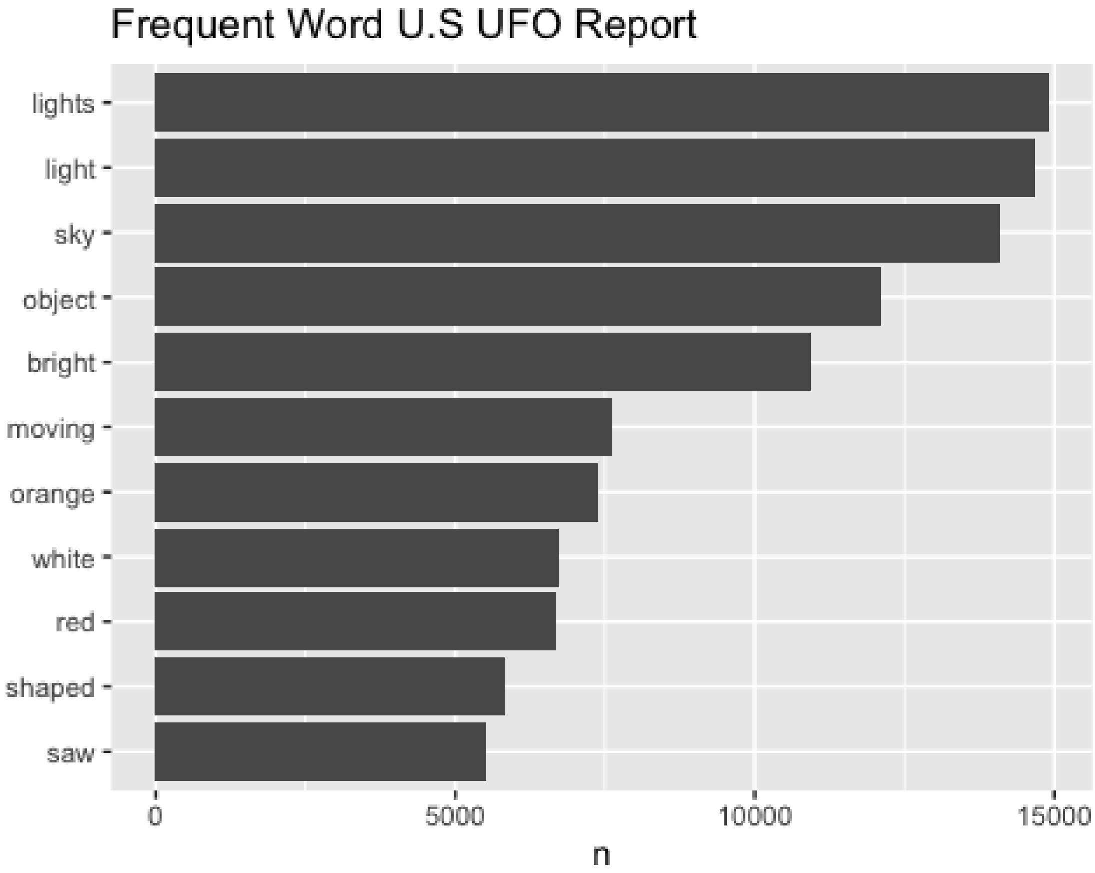
First of all, I tried to run the word frequency analysis. However, not surprisingly, the result is all about the very common words of describing the UFO-like color, shape, move, etc.

```
##########################################
## Tidy countries
##########################################
tidy_us <- ufo %>%
  filter(country== "us") %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)

tidy_gb <- ufo %>%
  filter(country== "gb") %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)

tidy_ca <- ufo %>%
  filter(country== "ca") %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

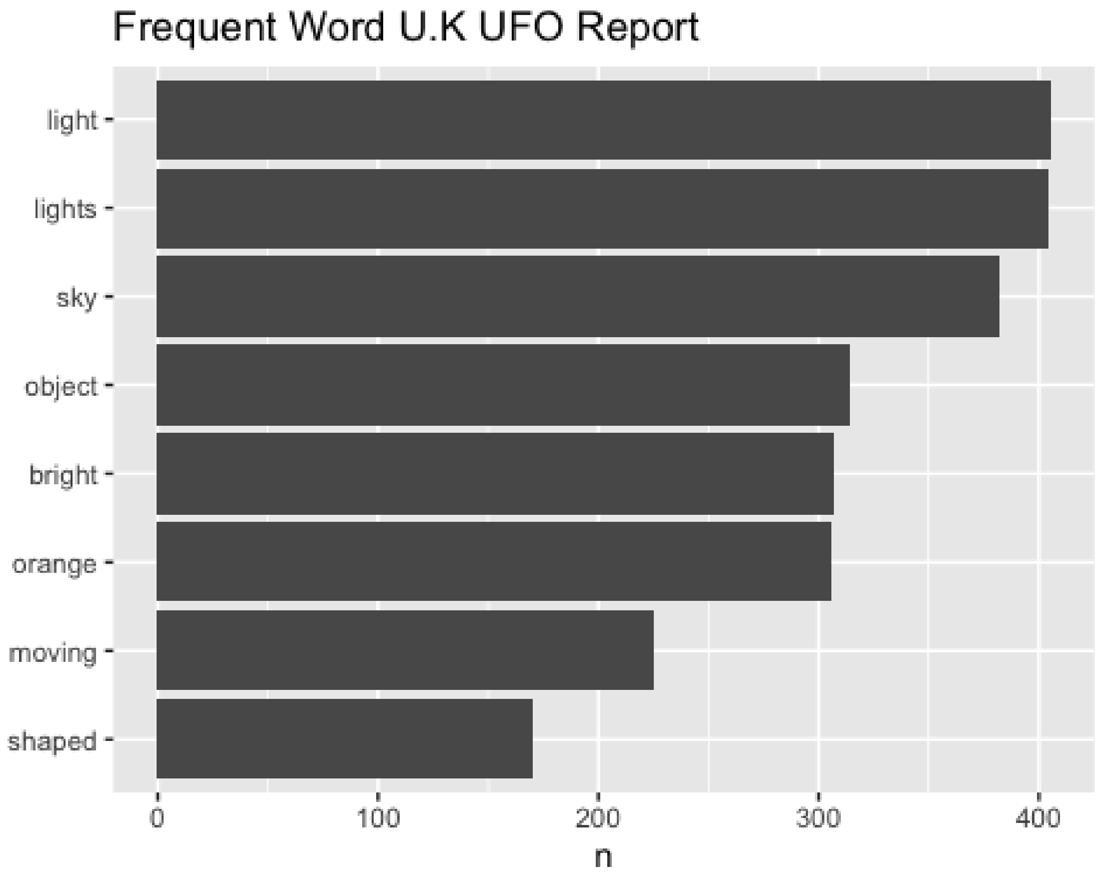
```r
tidy_au <- ufo %>%
  filter(country== "au") %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)


##########################################
## Frequent words world / us/ uk/ gb/ ca/ au
##########################################
tidy_ufo <- ufo %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort = T) %>%
  filter(n>5000) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Frequent Word World UFO Report")+
  xlab(NULL)+
  coord_flip()
print(tidy_ufo)
```



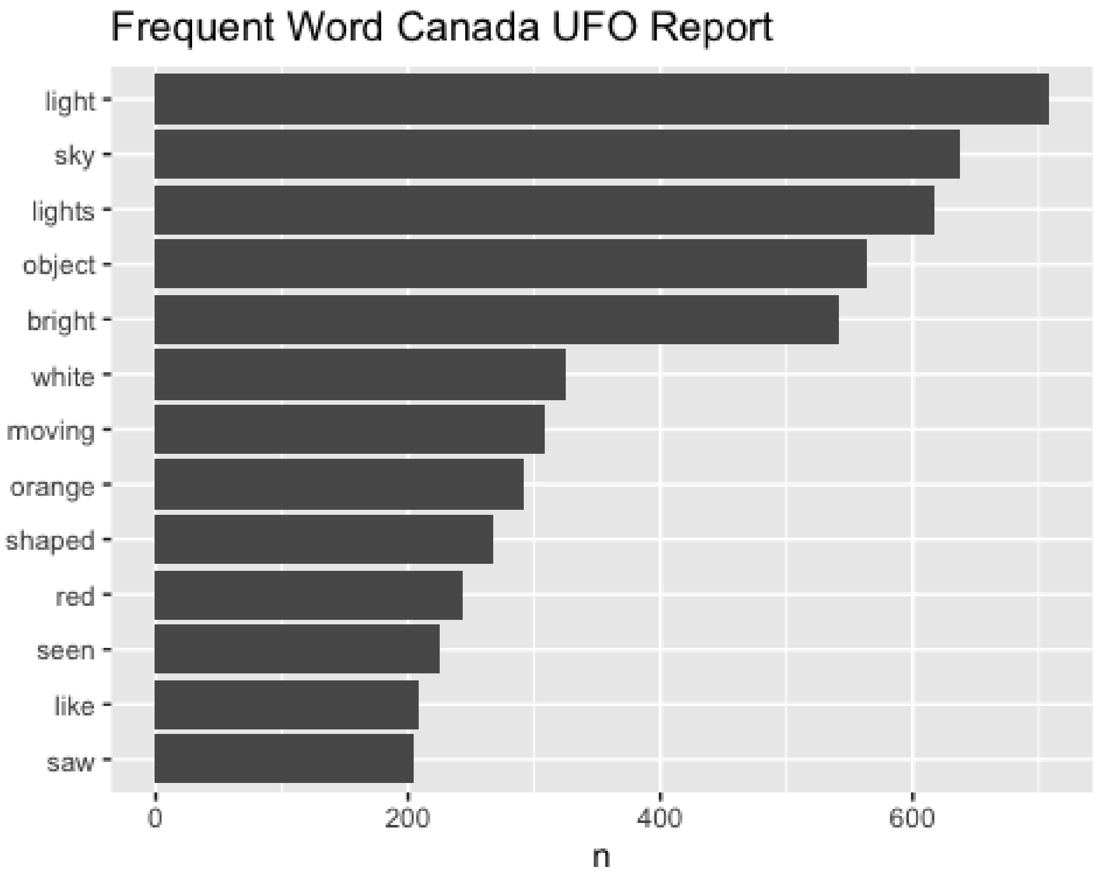
```
tidy_us_fr <- tidy_us %>%
  count(word, sort = T) %>%
  filter(n>5000) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Frequent Word U.S UFO Report")+
  xlab(NULL)+
  coord_flip()
print(tidy_us_fr)
```



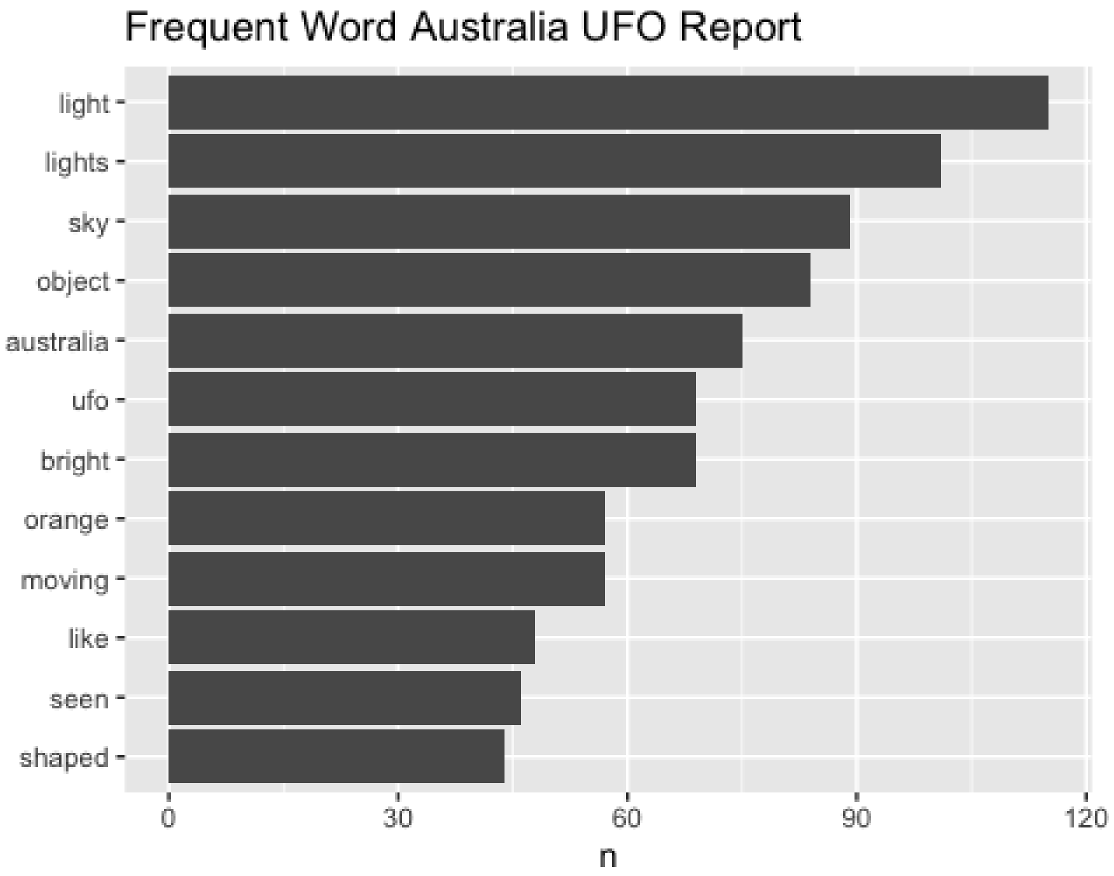
```
tidy_gb_fr <- tidy_gb %>%
  count(word, sort = T) %>%
  filter(n>150) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Frequent Word U.K UFO Report")+
  xlab(NULL)+
  coord_flip()
print(tidy_gb_fr)
```



```
tidy_ca_fr <- tidy_ca %>%
  count(word, sort = T) %>%
  filter(n>200) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Frequent Word Canada UFO Report")+
  xlab(NULL)+
  coord_flip()
print(tidy_ca_fr)
```



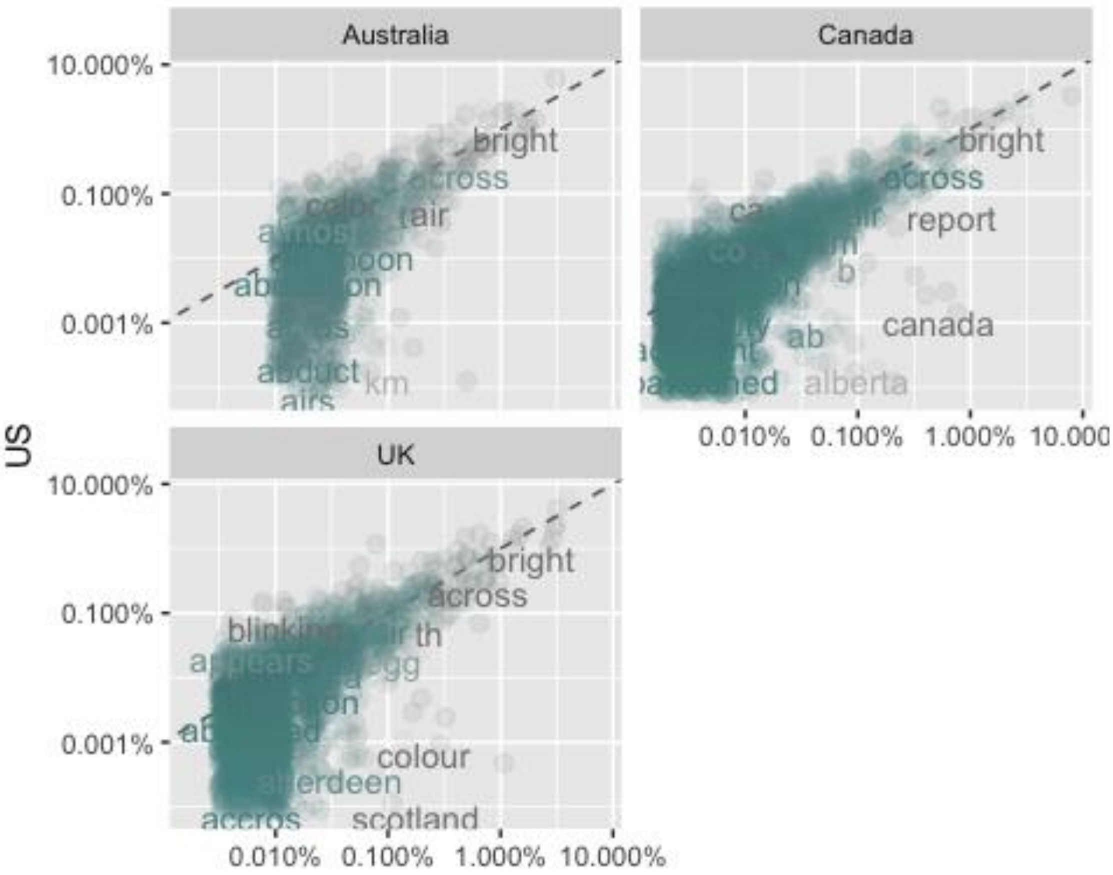
```
tidy_au_fr <- tidy_au %>%
  count(word, sort = T) %>%
  filter(n>40) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Frequent Word Australia UFO Report")+
  xlab(NULL)+
  coord_flip()
print(tidy_au_fr)
```



Word correlation vs. the U.S.

Figuring out if there is a unique difference between countries, I created the Word correlation matrix. This matrix shows the U.S. as a Y-axis and Australia, Canada, and the U.K. as X-axis. The Report from each country shows quite identical frequent word construction. I see some unique words for each country, but most of them are the name of cities or specific locations. Just for confirmation, I run the correlation test, but each correlation, U.S. vs. the UK, Canada, Australia, shows a very high correlation over 0.96~0.98. It means that the UFOs observed in each country are quite identical.

```
##########################################
## Country correlation
##########################################
frequency <- bind_rows(mutate(tidy_us, country="US"),
                       mutate(tidy_gb, country="UK"),
                       mutate(tidy_ca, country="Canada"),
                       mutate(tidy_au, country="Australia"))%>%
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(country, word) %>%
  group_by(country) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(country, proportion) %>%
  gather(country, proportion, `UK`, `Canada`, `Australia`)


ggplot(frequency, aes(x=proportion, y=`US`,
                      color = abs(`US`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "g
ray75")+
  facet_wrap(~country, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "US", x=NULL)
```



```
cor.test(data=frequency[frequency$country == "UK",],
         ~proportion + `US`)

##
##  Pearson's product-moment correlation
##
## data:  proportion and US
## t = 227.3, df = 2339, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9762774 0.9797922
## sample estimates:
##       cor
## 0.9781044
```

```
cor.test(data=frequency[frequency$country == "Canada",],
         ~proportion + `US`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and US
## t = 286.32, df = 3263, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9793151 0.9819442
## sample estimates:
##       cor
## 0.9806739
```

```
cor.test(data=frequency[frequency$country == "Australia",],
         ~proportion + `US`)
```

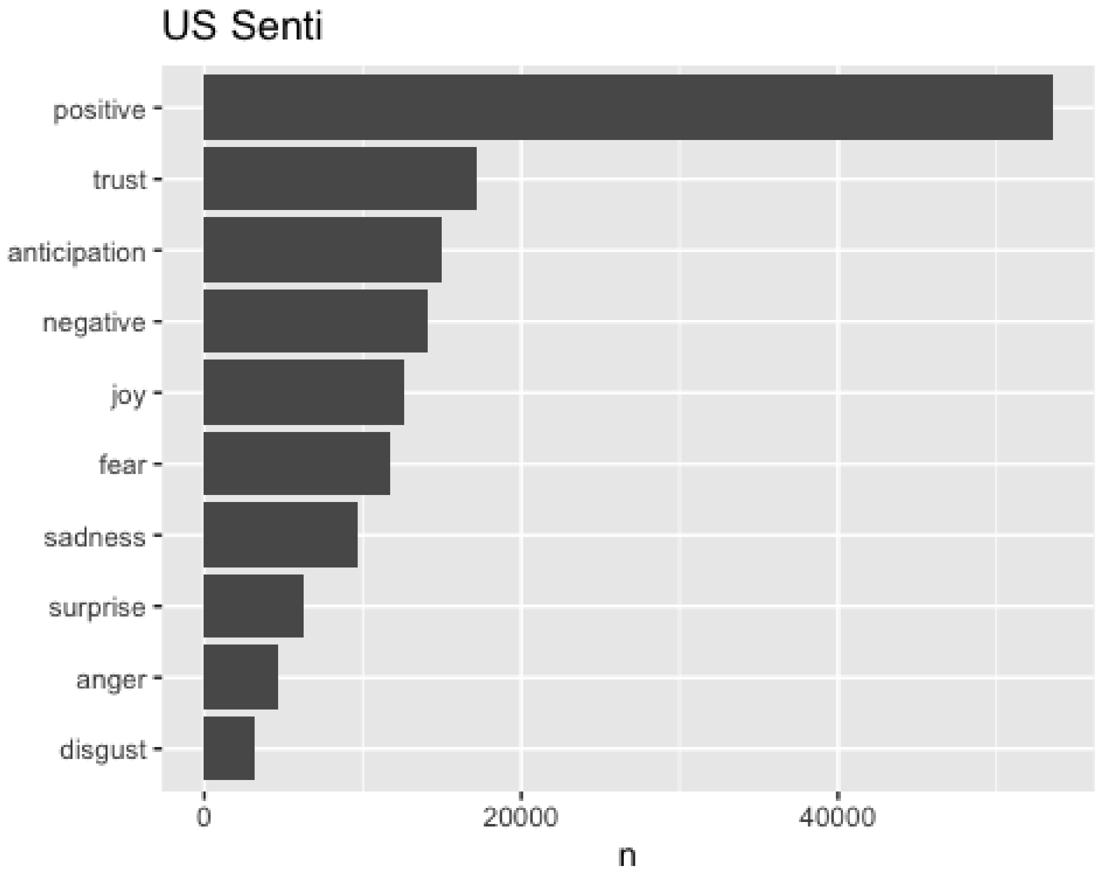
```
##
##  Pearson's product-moment correlation
##
## data:  proportion and US
## t = 148.2, df = 1321, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9679912 0.9741235
## sample estimates:
##       cor
## 0.9712179
```

Sentiment Analysis by country

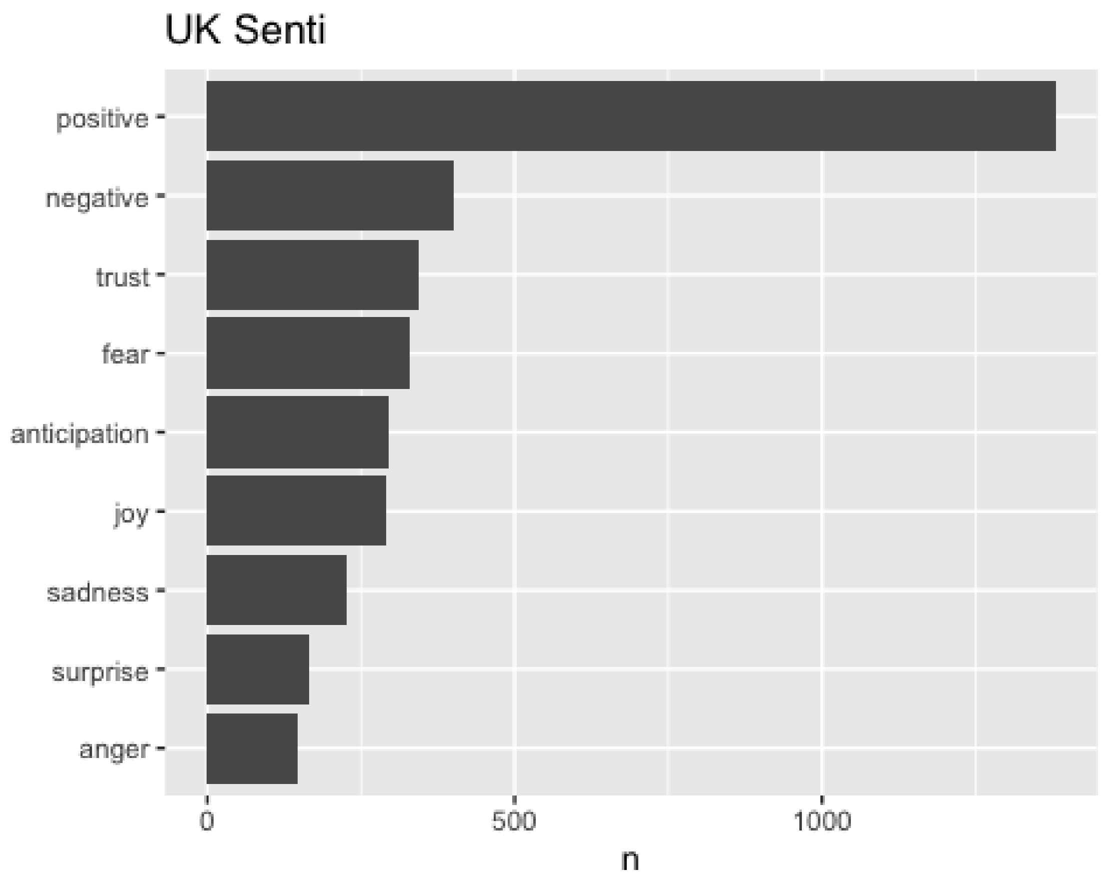
U.S., U.K, Canada are positive, Australia has more negative. Then how do people feel when they see the UFO. To see the impression difference by each country, I ran sentiment analysis with the sentiment data set “NRC.” The positive sentiment is common in all countries; however, only Australia shows negative sentiment in the second place. I tried to find some references from a website but could not find anything that explains its negative sentiment against the UFO. I assume these differences come from the number of the Report. Australian UFO report is quite limited compared with other countries.

```
###########################################
## Sentiments
###########################################
```

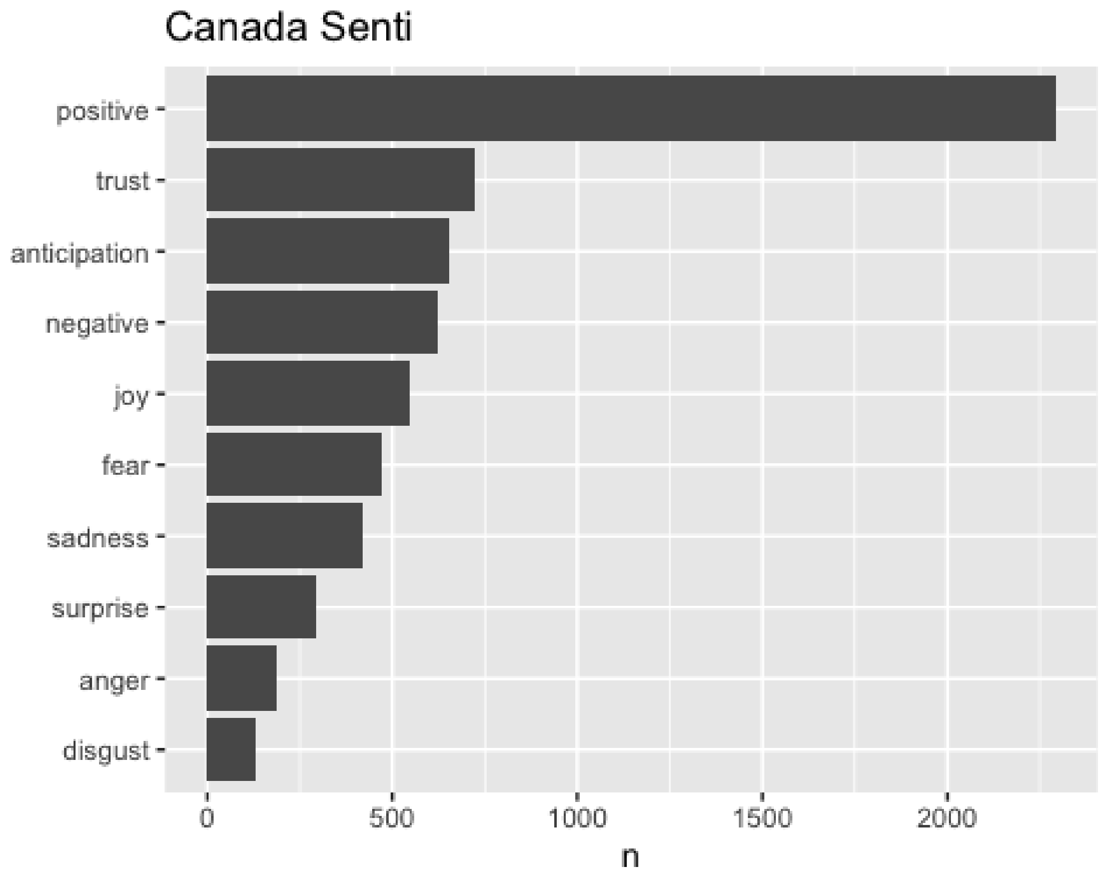
```r
nrc <- get_sentiments("nrc")

us_senti <- tidy_us %>%
  inner_join(nrc) %>%
  count(sentiment, sort = T) %>%
  filter(n>100) %>% # we need this to eliminate all the low count words
  mutate(sentiment = reorder(sentiment,n )) %>%
  ggplot(aes(sentiment, n,))+
  geom_col()+
  ggtitle("US Senti")+
  xlab(NULL)+
  coord_flip()
print(us_senti)
```



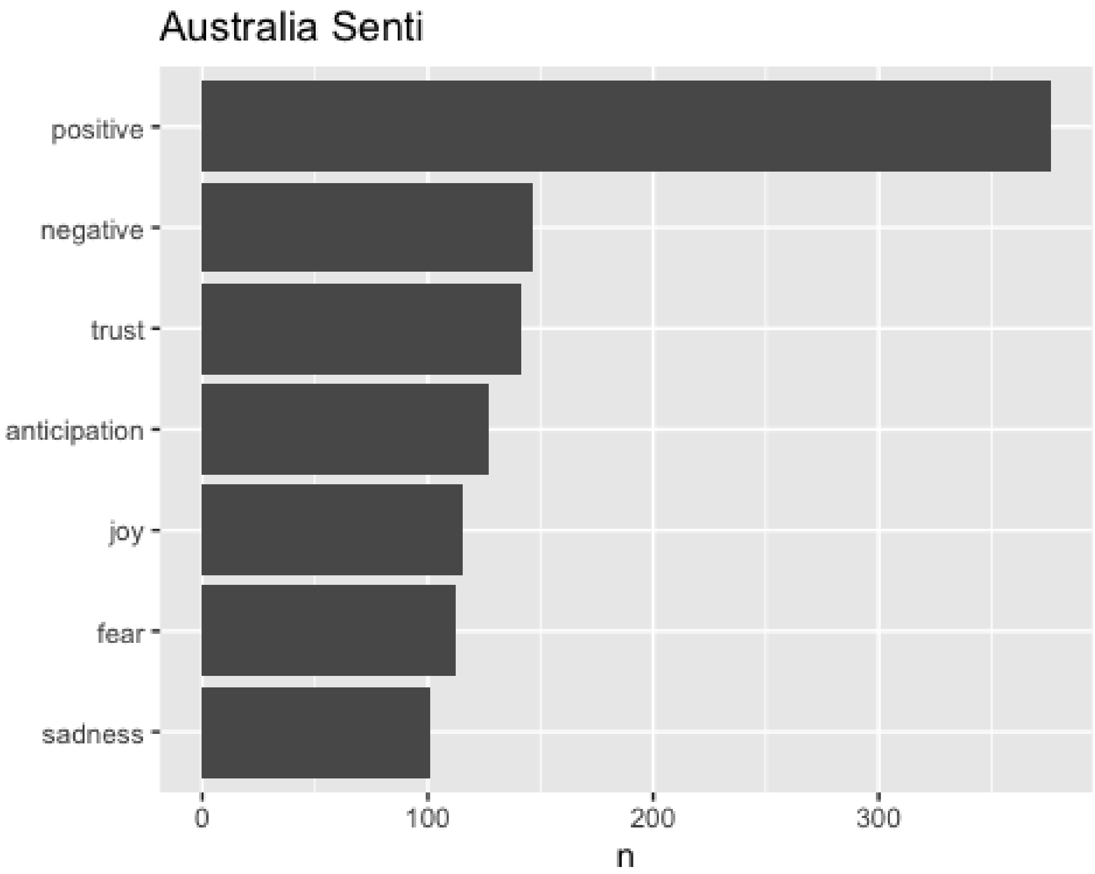
```r
gb_senti <- tidy_gb %>%
  inner_join(nrc) %>%
  count(sentiment, sort = T) %>%
  filter(n>100) %>% # we need this to eliminate all the low count words
  mutate(sentiment = reorder(sentiment,n )) %>%
  ggplot(aes(sentiment, n))+
  geom_col()+
  ggtitle("UK Senti")+
  xlab(NULL)+
  coord_flip()
print(gb_senti)
```



```
ca_senti <- tidy_ca %>%
  inner_join(nrc) %>%
  count(sentiment, sort = T) %>%
  filter(n>100) %>% # we need this to eliminate all the low count words
  mutate(sentiment = reorder(sentiment,n )) %>%
  ggplot(aes(sentiment, n))+
  geom_col()+
  ggtitle("Canada Senti")+
  xlab(NULL)+
  coord_flip()
print(ca_senti)
```



```
au_senti <- tidy_au %>%
  inner_join(nrc) %>%
  count(sentiment, sort = T) %>%
  filter(n>100) %>% # we need this to eliminate all the low count words
  mutate(sentiment = reorder(sentiment,n )) %>%
  ggplot(aes(sentiment, n))+
  geom_col()+
  ggtitle("Australia Senti")+
  xlab(NULL)+
  coord_flip()
print(au_senti)
```

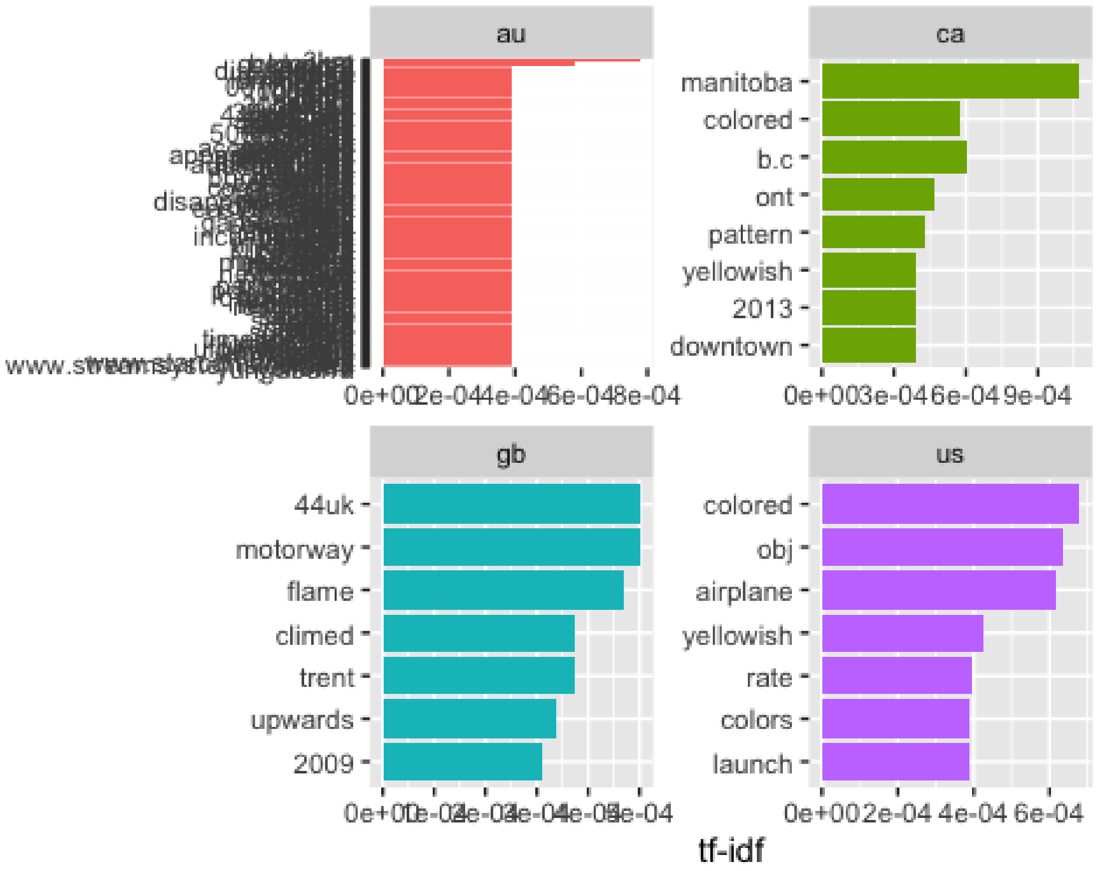


Unique Words U.S, U.K, Canada, Australia (TF-IDF)

The high TF-IDF score words were almost location names, therefore I eliminate location words from the analysis. American report contains words related with color, airplane since TF-IDF score shows word uniqueness, American report might not mention about colors or association with airplane often. Canadian high TF-IDF words contains some location name, color and "2013." This may be related with Canadian Minister of Defense at that time, that he admitted the existence of Aliens. UK has unique words like motorway, flame and 2009. Since Australia has fewer reports, many words get same TF-IDF score, therefore, its graph is not useful to find any insight.

```
##########################################
## TF_IDF
##########################################
ufo_token <- ufo %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words_city_name) %>%
  count(country, word, sort=TRUE) %>%
  ungroup()

total_words <- ufo_token %>%
  group_by(country)
```

```r
ufo_words <- left_join(ufo_token, total_words)%>%
  filter(country %in% c("us", "gb", "ca", "au"))

country_words <- ufo_words %>%
  bind_tf_idf(word, country, n)

uniqueness <- country_words %>%
  arrange(desc(tf_idf))
#what can we say about these words?

#############
# looking at the graphical approach:
country_words %>%
  arrange(desc(tf_idf)) %>%
  mutate(word=factor(word, levels=rev(unique(word)))) %>%
  group_by(country) %>%
  top_n(7) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill=country))+
  geom_col(show.legend=FALSE)+
  labs(x=NULL, y="tf-idf")+
  facet_wrap(~country, ncol=2, scales="free")+
  coord_flip()
```

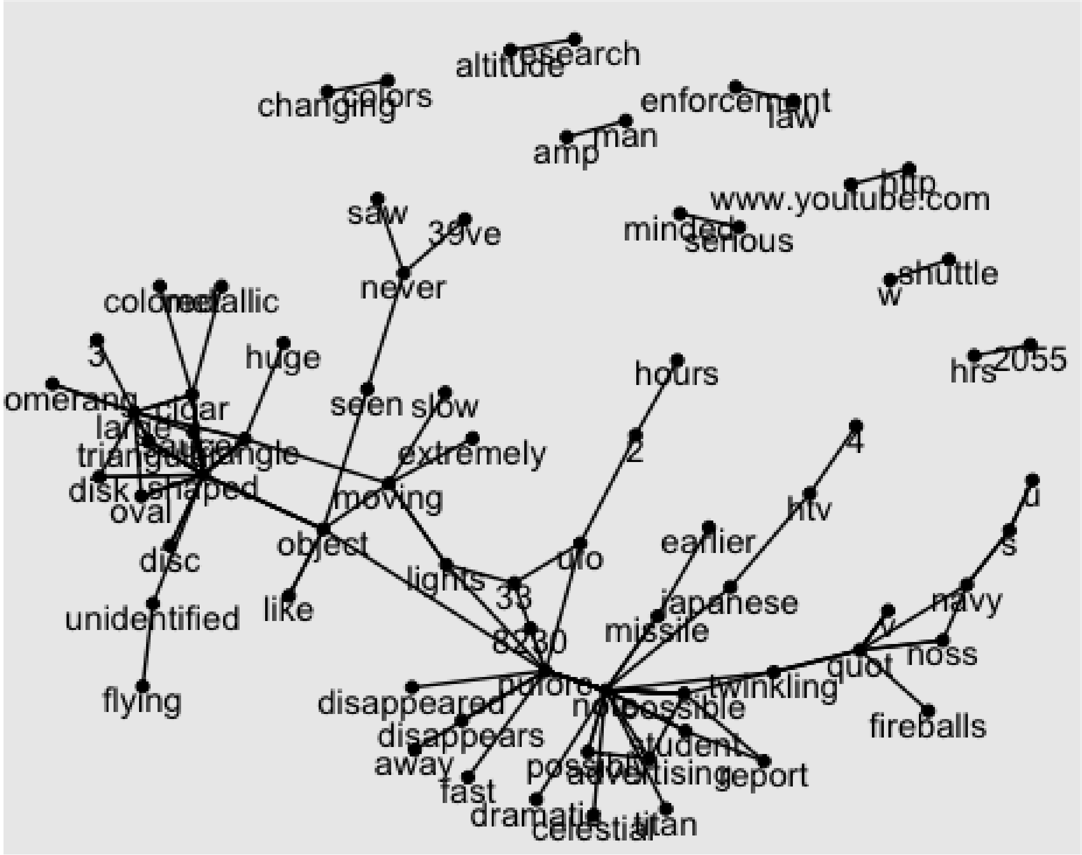


Quad-gram Analysis for entire report

Now, I try to see the structure of UFO observation report. The report seem to be structured with object hape, fling status, object color, speed, reporter's excitement and note of the events which might happen at the same time like rocket or missile launch and celestial event. The most unique thing for me is the convention of "Japanese" + "htv". Htv means H-ll Transfer Vehicle, rocket. From 2003 to 2021, Japanese Aerospace Exploration Agency launched H-ll rocket around 70 times In to the space. The rocket unit fall down after detachment and they seemed to be observed as a UFO.

```
##########################################
## n-gram
##########################################

ufo_quadgrams <- ufo %>%
  unnest_tokens(quadgram, text, token = "ngrams", n=4)

#ufo_quadgrams %>%
  #count(quadgram, sort = TRUE) #this has many stop words, need to remove the
m

#to remove stop words from the bigram data, we need to use the separate funct
```

```
ion:
quadgrams_separated <- ufo_quadgrams %>%
  separate(quadgram, c("word1", "word2", "word3", "word4"), sep = " ")

quadgrams_filtered <- quadgrams_separated %>%
  filter(!word1 %in% stop_words_city_name$word) %>% # ! means "not" all witho
ut !
  filter(!word2 %in% stop_words_city_name$word) %>%
  filter(!word3 %in% stop_words_city_name$word) %>%
  filter(!word4 %in% stop_words_city_name$word)

#creating the new bigram, "no-stop-words":
quadgrams_counts <- quadgrams_filtered %>%
  count(word1, word2, word3, word4, sort = TRUE)

quadgrams_graph <- quadgrams_counts %>%
  filter(n>4) %>%
  graph_from_data_frame()

ggraph(quadgrams_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=.5)
```



Roswell Report U.S Air Force

Next, I tried to run the text analysis with the famous Roswell Report written by U.S Airforce. This Report is completely excluding the possibility of UFO existence and describes and trying to be proving that the incident was caused or misunderstood with weather balloon project.

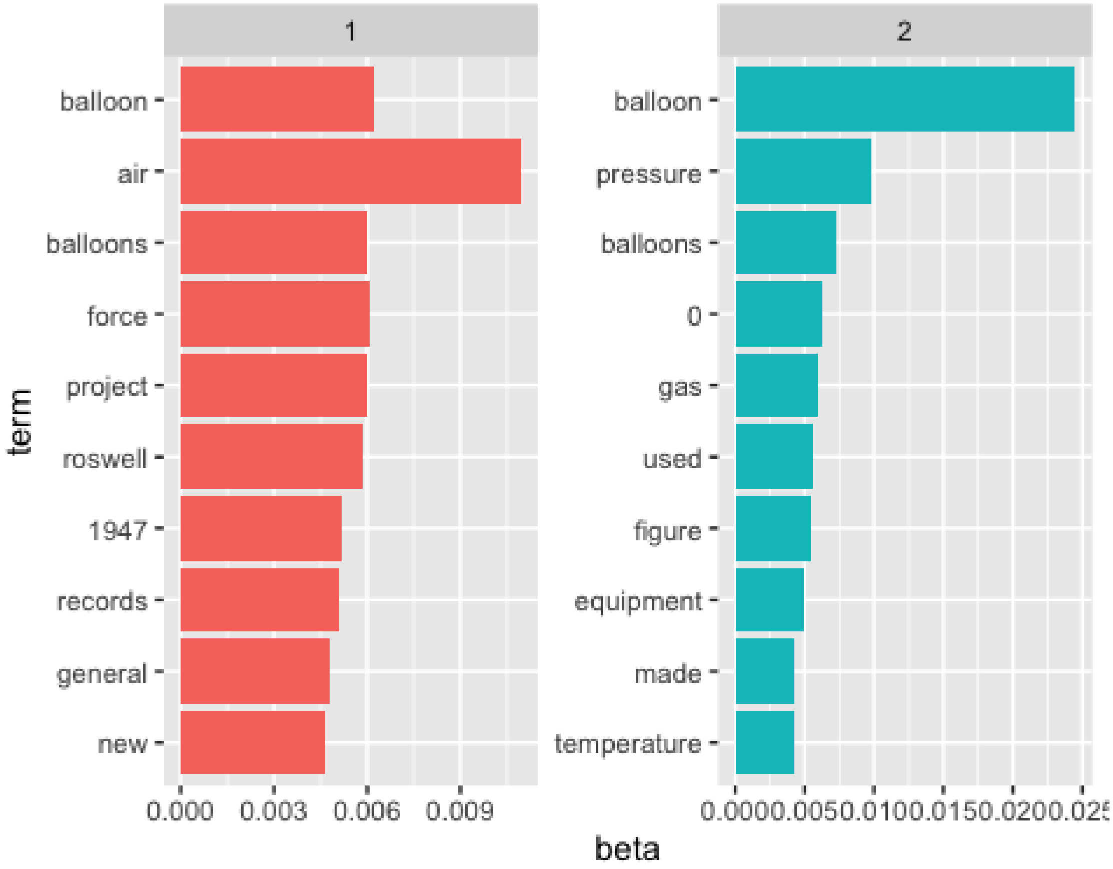
```
#############################################################################
#######
#############################################################################
#######
## Roswell Report
#############################################################################
#######
#############################################################################
#######
setwd("/Users/takahiroyamada/Desktop/MBAN/20211116 Text Analysis/A3 Assingmen
t/Roswell")
nm <- list.files(path="/Users/takahiroyamada/Desktop/MBAN/20211116 Text Analy
sis/A3 Assingment/Roswell")
my_pdf_text <- do.call(rbind, lapply(nm, function(x) pdf_text(x))) %>% t()
colnames(my_pdf_text) <- c("text")
my_pdf <- data.frame(line=1:994, text=my_pdf_text)

custom_lex_ros <- data_frame(word=c("a","b","c","d","e","f","g","h","i","j","
k","l",
                                "m","n","o","p","q","r","s","t","u","v",
                                "w","x","y","z"), lexicon=rep("SMART", each=2
6))
custom_lex_custom_r <- data_frame(word=my_pdf$line, lexicon=rep("SMART", each
=994))
stop_words_custom_r <- rbind(stop_words,  my_pdf$line, custom_lex_ros, custom
_lex_custom_r)


tidy_ros <- my_pdf %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words_custom_r)
```

Topic Analysis

Topic analyses seem to show two topics. 1: Air force balloons project report, 2: Project details temp, press, figure.

```
##########################################
## Topic Analysis
##########################################
ros_dtm <- tidy_ros %>%
  count(line, word) %>%
  cast_dtm(line, word, n)
```

```
#calling the Latent Dirichlet Allocation algorithm
ros_lda <- LDA(ros_dtm, k=2, control=list(seed=123))  ## k=2 - only 2 topics

#now we are looking for the per topic per word probabilities aka. beta
#beta - what is the probability that "this term" will be generated by "this t
opic"

ros_topics <- tidy(ros_lda, matrix="beta")

ros_top_terms <- ros_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

#lets plot the term frequencies by topic
ros_top_terms %>%
  mutate(term=reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic, scales = "free") +
  coord_flip()
```

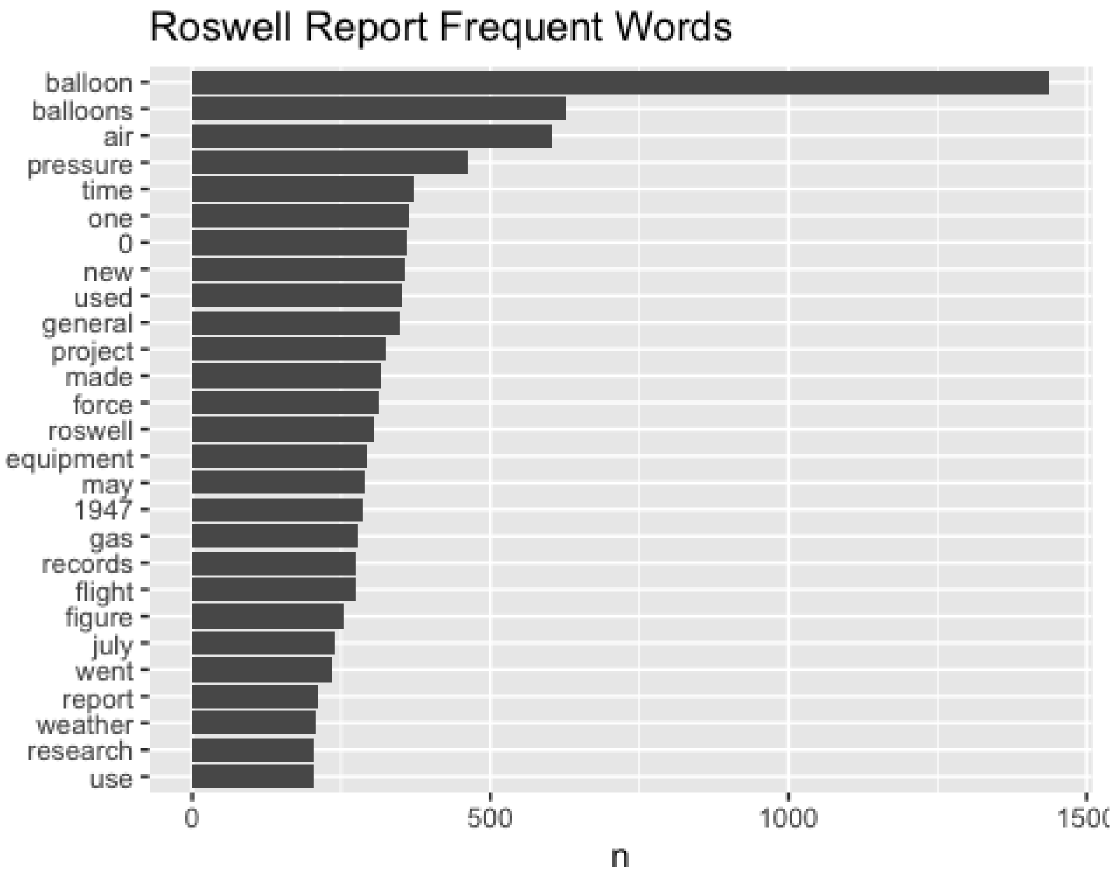


```
#lets calculate the relative difference between the betas for words in topic
1
#and words in topic 2

ros_beta_spread <- ros_topics %>%
  mutate(topic=paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1>.001 | topic2 >.001) %>%
  mutate(log_rate = log2(topic2/topic1))   ## log rate shows the big diff bet
ween topic 1 and topic 2
```

Report frequent words

The word frequency bar chart supports topic analysis that many words related to balloons, projects are used in the Report.

```
##########################################
## Word Frequency
##########################################
tidy_ros_fr <- tidy_ros %>%
  count(word, sort = T) %>%
  filter(n>200) %>% # we need this to eliminate all the low count words
  mutate(word = reorder(word,n )) %>%
  ggplot(aes(word, n))+
  geom_col()+
  ggtitle("Roswell Report Frequent Words")+
  xlab(NULL)+
  coord_flip()
print(tidy_ros_fr)
```

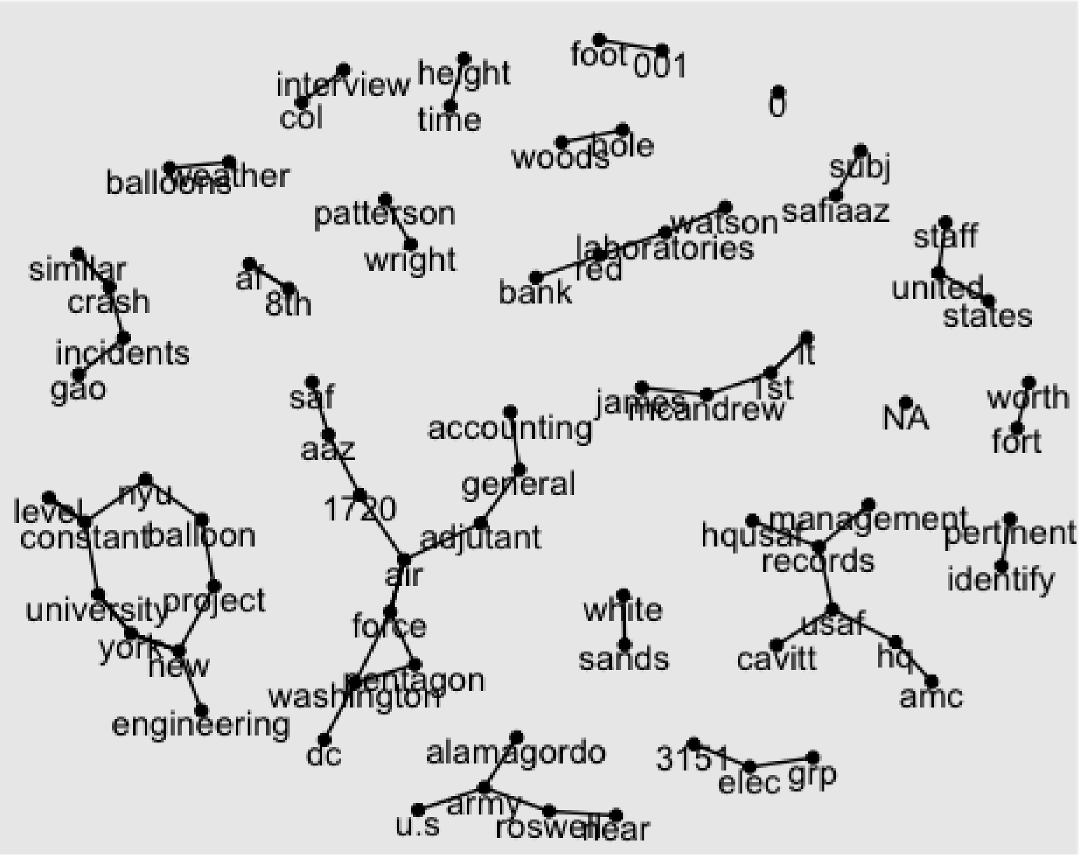


Quad-gram Analysis for entire Report

Quad-gram shows that the word balloon connected with New York, University, constant, project, and engineering. The incident happened in California, but Report seems to be saying that the incident was related to New York Univ project.

```
##########################################
## n-gram
##########################################
ros_quadgrams <- my_pdf %>%
  unnest_tokens(quadgram, text, token = "ngrams", n=4)

#ros_quadgrams %>%
  #count(quadgram, sort = TRUE) #this has many stop words, need to remove the
m
```

```r
#to remove stop words from the bigram data, we need to use the separate function:
ros_quadgrams_separated <- ros_quadgrams %>%
  separate(quadgram, c("word1", "word2", "word3", "word4"), sep = " ")

ros_quadgrams_filtered <- ros_quadgrams_separated %>%
  filter(!word1 %in% stop_words_custom_r$word) %>% # ! means "not" all without !
  filter(!word2 %in% stop_words_custom_r$word) %>%
  filter(!word3 %in% stop_words_custom_r$word) %>%
  filter(!word4 %in% stop_words_custom_r$word)

#creating the new bigram, "no-stop-words":
ros_quadgrams_counts <- ros_quadgrams_filtered %>%
  count(word1, word2, word3, word4, sort = TRUE)

ros_quadgrams_graph <- ros_quadgrams_counts %>%
  filter(n>3) %>%
  graph_from_data_frame()

ggraph(ros_quadgrams_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=.5)
```

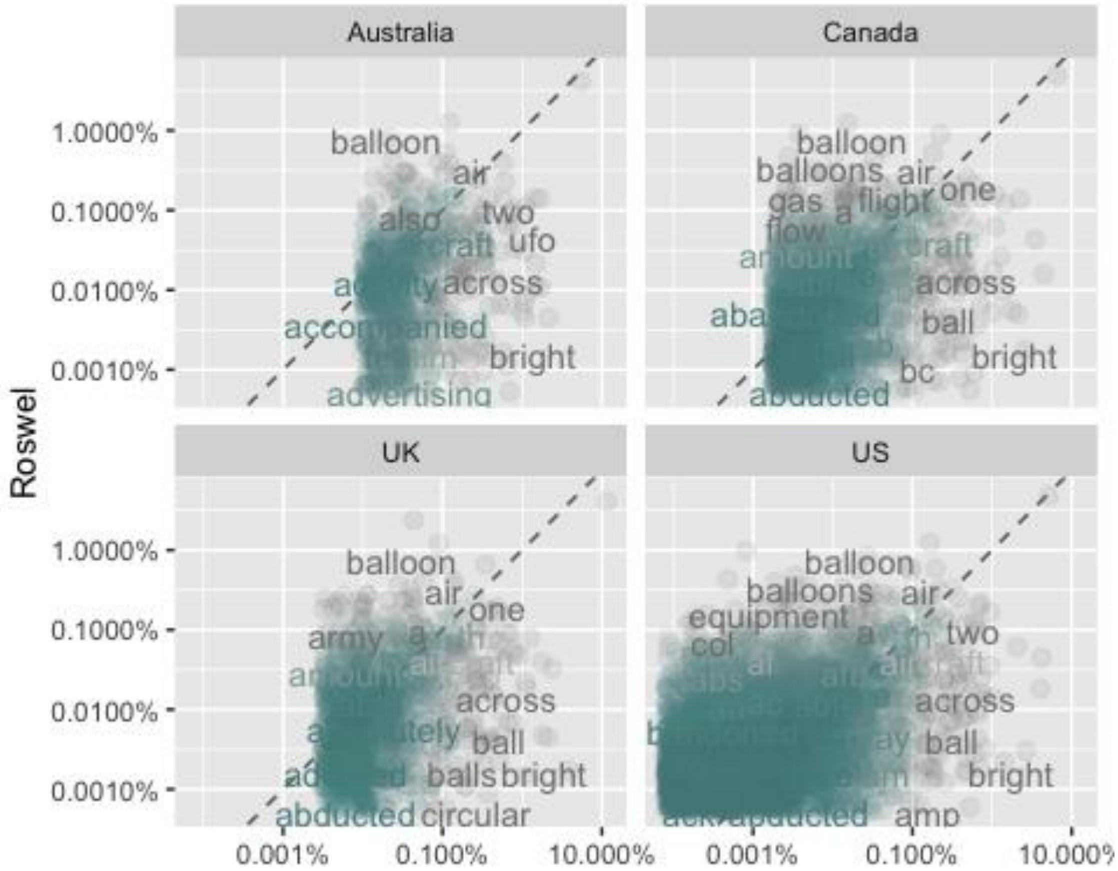


Word Correlation Roswell vs. U.S, U.K, Canada, Australia

Now, I run the frequency word comparison with Roswell Report vs. the U.S, U.K, Canada, Australia observation report. Each document the obsecration report from each country is trying to describing the shape, color, situation, and its activity, the Report is sticking to the balloon. However, the correlations of the Report and countries are around 0.65, and they look moderate.

```
##########################################
## Report vs Country correlation
##########################################
ros_frequency <- bind_rows(mutate(tidy_ros, country="Roswell"),
                           mutate(tidy_us, country="US"),
                           mutate(tidy_gb, country="UK"),
                           mutate(tidy_ca, country="Canada"),
                           mutate(tidy_au, country="Australia"))%>%
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(country, word) %>%
  group_by(country) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(country, proportion) %>%
  gather(country, proportion, `US`, `UK`, `Canada`, `Australia`)
```

```
ggplot(ros_frequency, aes(x=proportion, y=`Roswell`,
                          color = abs(`Roswell`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "g
ray75")+
  facet_wrap(~country, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "Roswel", x=NULL)
```



```
cor.test(data=ros_frequency[ros_frequency$country == "US",],
         ~proportion + `Roswell`)

##
##  Pearson's product-moment correlation
##
## data:  proportion and Roswell
## t = 58.985, df = 5029, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6228477 0.6555250
## sample estimates:
##       cor
## 0.6394751

cor.test(data=ros_frequency[ros_frequency$country == "UK",],
         ~proportion + `Roswell`)

##
##  Pearson's product-moment correlation
##
## data:  proportion and Roswell
## t = 31.78, df = 1406, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6150999 0.6759735
## sample estimates:
##       cor
## 0.6465647

cor.test(data=ros_frequency[ros_frequency$country == "Canada",],
         ~proportion + `Roswell`)

##
##  Pearson's product-moment correlation
##
## data:  proportion and Roswell
## t = 34.794, df = 1904, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5952157 0.6501528
## sample estimates:
##       cor
## 0.6234531

cor.test(data=ros_frequency[ros_frequency$country == "Australia",],
         ~proportion + `Roswell`)

##
##  Pearson's product-moment correlation
##
## data:  proportion and Roswell
## t = 25.238, df = 882, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6075623 0.6842774
## sample estimates:
##       cor
## 0.6475576
```

Conclusion

The UFO observation is raising recently however, each of them are quite identical. No difference word characteristic and sentiment between countries. This might mean that what people seeing is really same and the reports are very reliable. Or, they are all biased by American movie culture which planted identical and fixed UFO image to people all over the world. The most famous UFO incident, Roswell, which was the origin of the UFO booming happened in U.S 1947. However, the American Airforce report denied its existence and trying to relate the UFO to the Air Balloons, instead.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.