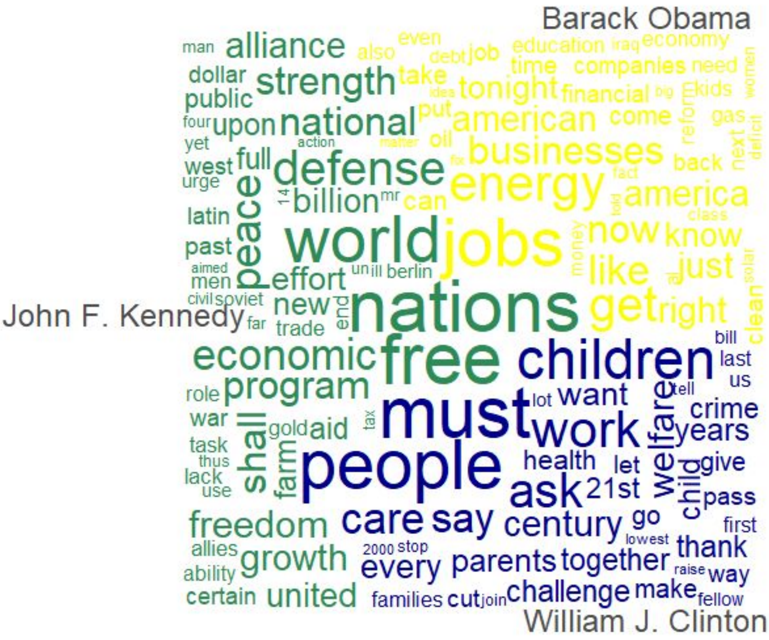
Team 5
Yifan Feng: 2671027
Yunxiang Li: 2674844

1a)

```
> url <- 'https://bit.ly/2QoqUQS'
> (d <- read_csv(url))
> (corp <- corpus(d, text_field = 'text'))
> corpus_subset(corp,President %in% c("Barack Obama", "John F.
Kennedy","William J. Clinton")) %>% dfm(groups = "President", remove =
stopwords("english"),remove_punct = TRUE) %>%dfm_trim(min_termfreq = 5,
verbose = FALSE) %>% textplot_wordcloud(comparison = TRUE, color=c("yellow",
"seagreen","blue4"))
```



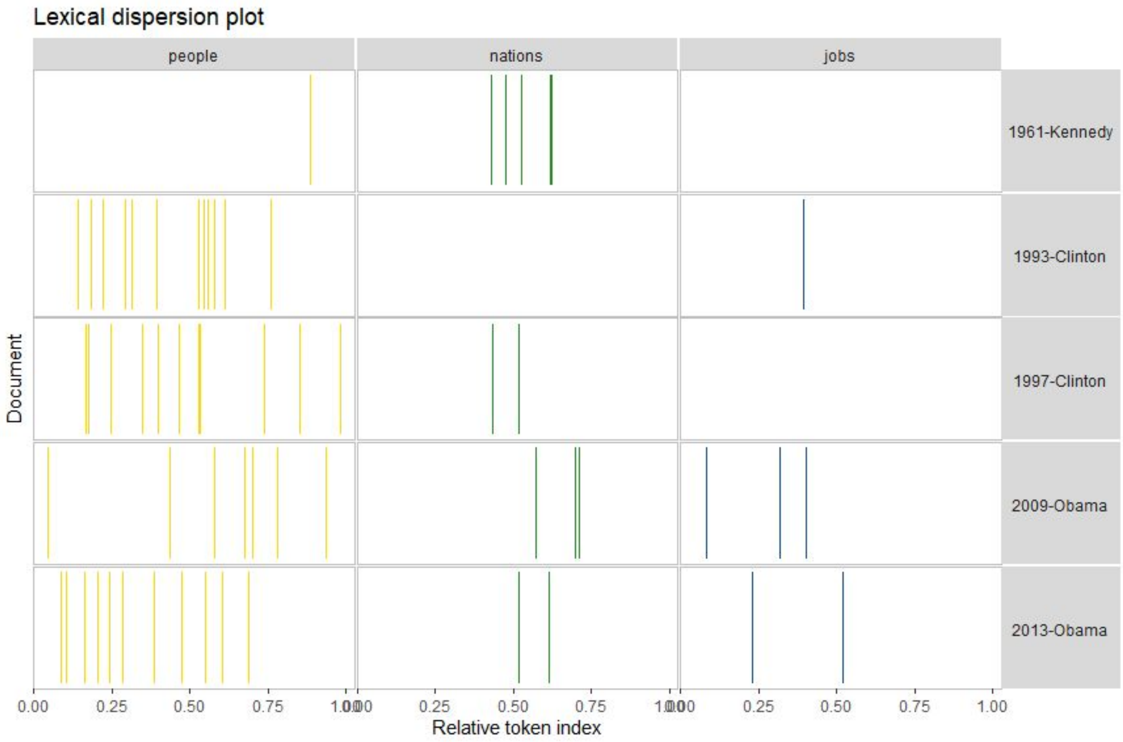
1b)

The primary conclusion can draw upon word size. Topics differ per president. Kennedy talks much about the nation that represents through words like “defense”, “peace”. Obama stresses economy-related topics via “jobs”, “businesses” and so on. Clinton mentions society’s development in the 21st century in words such as “welfare”, “crime”. Overall, we can see that Kennedy focuses on the national level whilst Obama and Clinton pay attention to individual(citizen) level. However, we fail to distinguish nuances due to the limited information.

```
# To compare certain values against each other, it is suggested to comparison charts
such as line charts or bar charts. In bar chart, we can read the differences in
different categories. To make data more readable and concise, it is recommended to
only use labels and titles for the graph. In this case, the category should be labeled
by four president’s names and bar can be word frequency or the weight by
percentage. Or one can use the lexical dispersion plot, which shows the
importance of one word by calculating the dispersion’s weight in a corpus.
```

1c)

```
> g <- corpus_subset(data_corpus_inaugural,President %in% c("Kennedy", "Clinton",
"Obama"))
> tplot <- textplot_xray( kwic(g, pattern = "people"),kwic(g, pattern =
"nations"),kwic(g, pattern = "jobs"))
> tplot + aes(color = keyword) + scale_color_manual(values = c("gold", "forestgreen",
"dodgerblue4")) + theme(legend.position = "none")
```



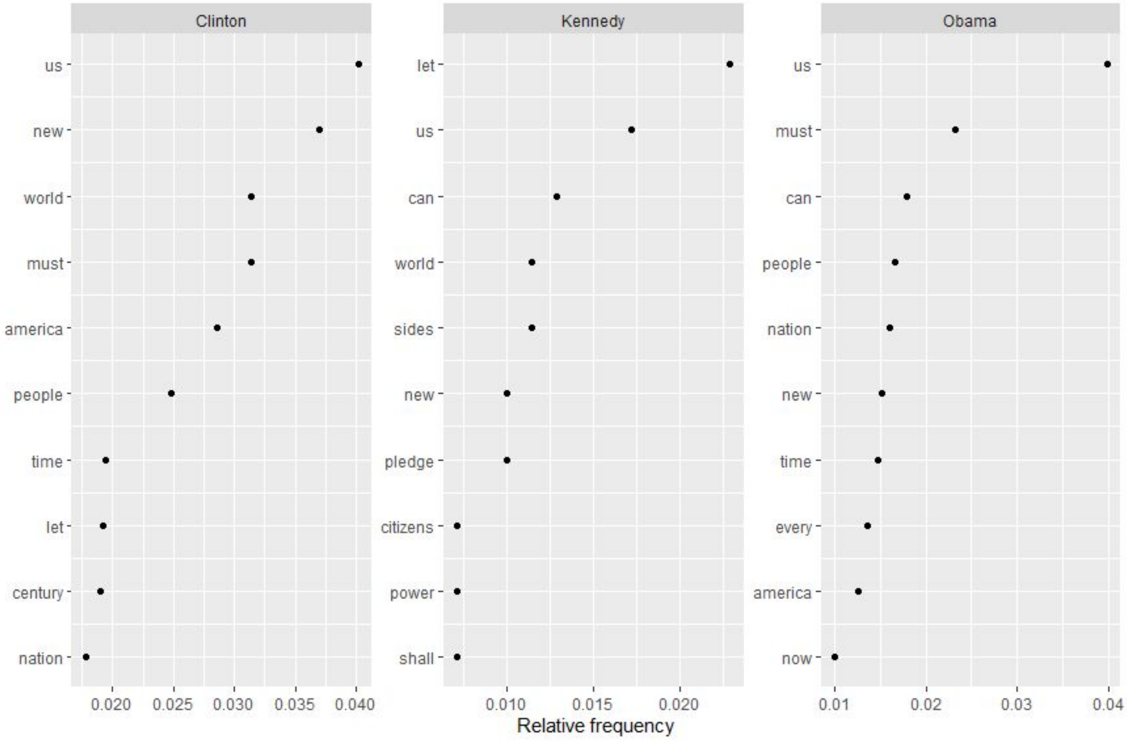
1d)

```
#The x-ray plot shows the frequency (rate range from 0 to 1) of the commonly-used
words (i.e. people, nations and jobs) by three presidents in different years.  It shows
the index of one word, meaning that a word’s location in the speech is also seen
(beginning, middle or end). X-ray plot, compared to word cloud is more readable for
users because one can overview the occurrence rate in numbers rather than in word
```

size. Besides, one can base index to analyze the distribution of required information (i.e. word use).

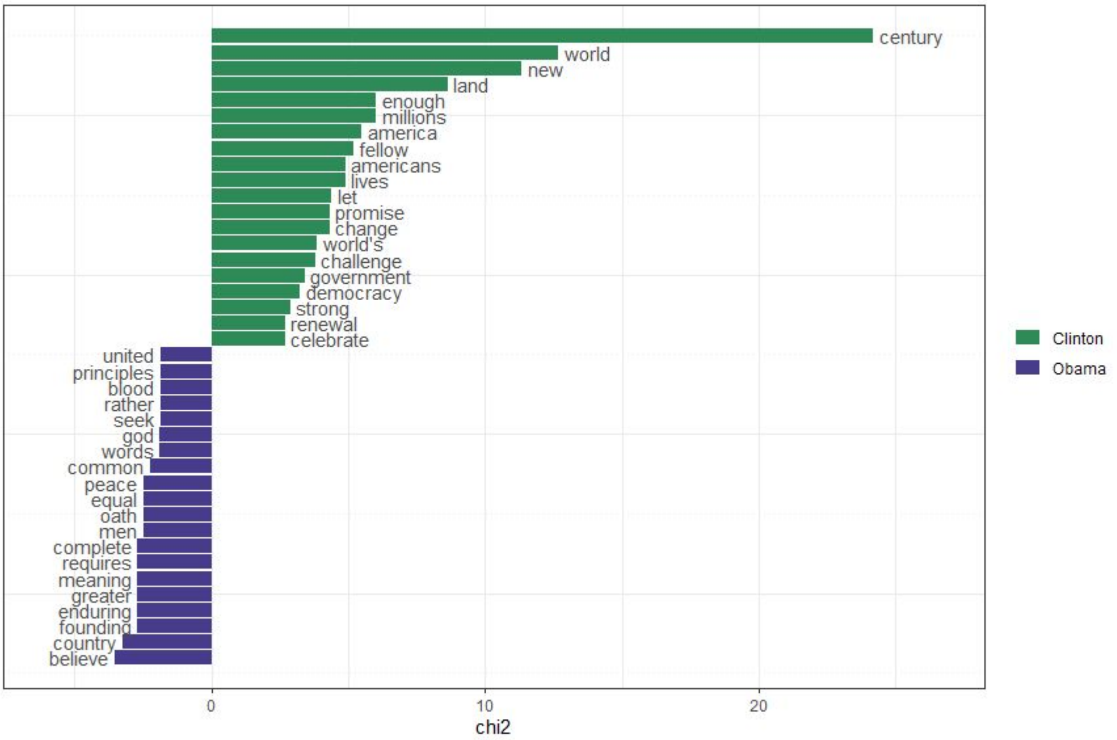
1e)

```
> dfm_weight_pres <- corpus_subset(data_corpus_inaugural,President %in%
c("Kennedy", "Clinton", "Obama")) %>% dfm(remove = stopwords("english"),
remove_punct = TRUE) %>% dfm_weight(scheme = "prop")
> freq_weight <- textstat_frequency(dfm_weight_pres, n = 10, groups = "President")
> ggplot(data = freq_weight, aes(x = nrow(freq_weight):1, y = frequency)) +
geom_point() + facet_wrap(~ group, scales = "free") + coord_flip() +
scale_x_continuous(breaks = nrow(freq_weight):1,labels = freq_weight$feature) +
labs(x = NULL, y = "Relative frequency")
```



1f)

```
> pres_corpus <- corpus_subset(data_corpus_inaugural, President %in%
c("Obama", "Clinton"))
> pres_dfm <- dfm(pres_corpus, groups = "President", remove =
stopwords("english"), remove_punct = TRUE)
> result_keyness <- textstat_keyness(pres_dfm, target = "Clinton")
> textplot_keyness(result_keyness, color = c('seagreen','slateblue4'))
```

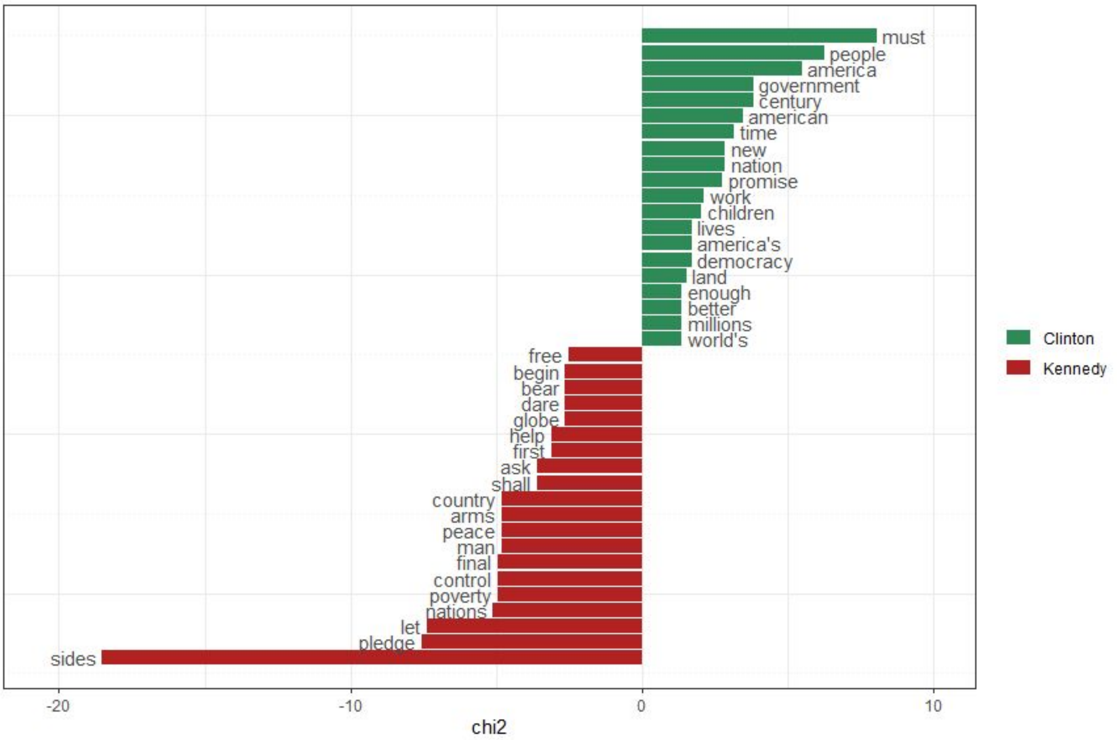


1g)

#Conclusions:
We visualized data from Clinton (in green), Kennedy (in red) and Obama (in yellow). An overall view is that Kennedy, compared to Clinton and Obama has more personal-featured words (shown by keyness). All three president emphasizes the important position of their society members (i.e. citizens) and their country. Both Obama and Clinton are inclined to use the modal verb “must” to stress their speech. Among the three sets of vocabulary, we notice that there is a distinctive focus for each president, namely “social development” for Clinton (via words “children” “promise” “democracy”), “peace and union” for Kennedy (via “nation” “peace” “human”), and “social equality and future” for Obama (via “American” “journey” “future” “woman”).

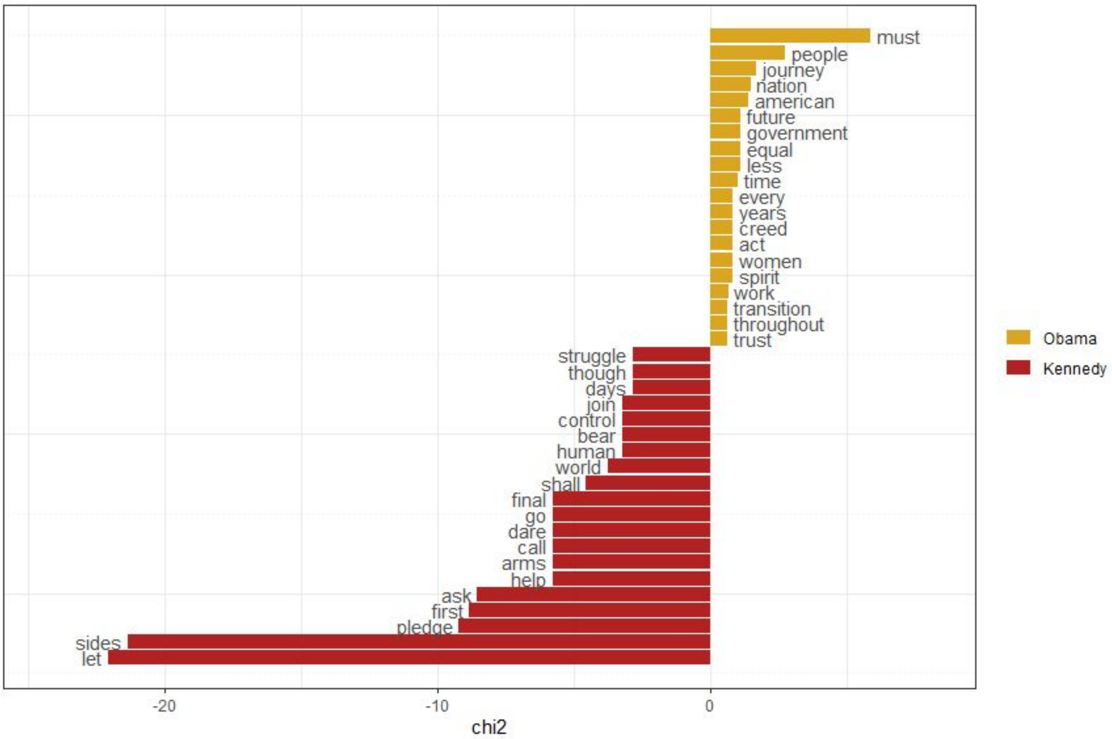
```
#Kennedy v.s. Clinton

> pres_corpus <- corpus_subset(data_corpus_inaugural, President %in%
c("Kennedy", "Clinton"))
> pres_dfm <- dfm(pres_corpus, groups = "President", remove =
stopwords("english"), remove_punct = TRUE)
> result_keyness <- textstat_keyness(pres_dfm, target = "Clinton")
> textplot_keyness(result_keyness, color = c('seagreen','firebrick'))
```



```
#Kennedy v.s. Obama

> pres_corpus <- corpus_subset(data_corpus_inaugural, President %in%
c("Kennedy", "Obama"))
> pres_dfm <- dfm(pres_corpus, groups = "President", remove =
stopwords("english"), remove_punct = TRUE)
> result_keyness <- textstat_keyness(pres_dfm, target = "Obama")
> textplot_keyness(result_keyness, color = c('goldenrod','firebrick'))
```


must
people
journey
nation
american
future
government
equal
less
time
every
years
creed
act
women
spirit
work
transition
throughout
trust
struggle
though
days
join
control
bear
human
world
shall
final
go
dare
call
arms
help
ask
first
pledge
sides
let
-20
-10
0
chi2
Obama
Kennedy