Passo-a-passo

Aula 5. Estatística Descritiva usando Texto

Os dados, comandos e gráficos desse documento foram produzidos por Clarice Barbosa, na Iniciação Científica: "Mídia, política e China: um estudo do caso queniano".

Prof. Pedro Feliú

Instituto de Relações Internacionais
Universidade de São Paulo

INTRODUÇÃO

Nesta aula vamos aprender a produzir e analisar estatísticas descritivas a partir de texto, uma variável qualitativa nominal. Nas aulas anteriores utilizamos muitas variáveis quantitativas contínuas, realizando gráficos e análises descritivas com variáveis como PIB, expectativa de vida, cobertura florestas, fluxo de refugiados, índice de democracia, entre outras. Nessa aula, vamos utilizar uma base de dados apenas com variáveis qualitativas, com especial atenção para a análise quantitativa de texto. O uso de texto nas pesquisas de relações internacionais é muito comum, especialmente os discursos dos tomadores de decisão em política externa. O discurso na diplomacia é uma ação política e a sua análise permite averiguar a influência de variáveis subjetivas, como as ideias, ideologia e identidade, no comportamento dos países. Uma forma bastante útil de fazer análise de texto é por meio do uso de estatística descritiva e análise de sentimento das palavras.

Na presente aula, utilizaremos como banco de dados os discursos de legisladores e legisladoras do Quênia sobre a China pronunciados no parlamento entre 2006 e 2018. Esse banco de dados, assim como os comandos do R e gráficos que faremos a seguir, foi produzido pela estudante de graduação do IRI, Clarice Barbosa, em iniciação científica (IC) ainda em andamento. Um aspecto de grande interesse nas relações internacionais é a projeção chinesa no terceiro mundo, particularmente sobre o continente africano. O peso das trocas comerciais, investimentos em infraestrutura e serviços, instalação de fábricas e compra de terras são aspectos notáveis das relações entre a China e países da África Austral. Em particular, o setor de mídia e telecomunicações tem sido um alvo estratégico dos investimentos chineses, refletidos tanto no aperfeiçoamento da tecnologia, formação e capacitação de profissionais e exportação de plataformas de comunicação chinesas como canais de televisão, telenovelas e jornais diários. A presença de investimentos chineses no setor de faz parte de uma estratégia do governo para, através de mudanças no conteúdo publicado pela mídia, construir uma percepção positiva sobre a China e, por consequência, garantir que a opinião pública e elite política sejam favoráveis a sua presença nos países africanos. A inserção chinesa em países do terceiro mundo pode suscitar distintas reações na elite política local, podendo polarizar a avaliação dos partidos políticos sobre o crescente engajamento das relações exteriores com o gigante asiático.

A IC da Clarice tem como objetivo central justamente compreender o impacto da presença chinesa, especialmente do setor de telecomunicações, na percepção dos legisladores e legisladoras do Quênia sobre o país asiático. O caso queniano é de elevado interesse pois em 2012 foi construída no país uma sede da CCTV (China Central Television - emissora estatal chinesa) em Nairóbi como um dos principais elementos da presença chinesa na África. Além de transmitir notícias sobre a China e a Ásia em geral, a emissora cobre de maneira bastante completa as notícias africanas e os debates contemporâneos. O fato de a CCTV ser uma emissora estatal atesta o alinhamento ao Partido Comunista Chinês, o que torna ainda mais evidente o uso da mídia como ferramenta de diplomacia pública por parte do governo. Além disso, passou a ser distribuído, também em Nairóbi, o China Daily, maior jornal nacional em inglês da China. Essas iniciativas como parte de um processo de construção de uma imagem positiva da China para os africanos, seguindo sempre a narrativa da amizade e cooperação, reforçada por um discurso anticolonialista e, muitas vezes, antiocidental.

Além disso, a CCTV já ocupa uma posição no mercado similar à da Al-Jazeera e da BBC no Quênia, dois canais que têm dominado o cenário da mídia africana. Assim, em suma, a escolha do caso queniano permite aferir claramente o impacto dos investimentos chineses em telecomunicações nas atitudes dos parlamentares do país, permitindo compreender os efeitos da ascensão chinesa na conformação dos atores e coalizões domésticas na África austral.

A série temporal do banco de dados, de 2006 a 2018, permite comparar seis anos de discursos antes da instalação da estatal chinesa de telecomunicações no país e seis anos depois, viabilizando a análise do impacto deste investimento nas atitudes de importante parcela da elite politica do Quênia. Percebam que vamos nesta aula quantificar a percepção dos legisladores do Quênia, um elemento subjetivo das relações exteriores do país. O banco de dados é denominado kenya.xls e está disponível na pasta de base de dados do moodle. Vamos iniciar os para instalar os pacotes necessários, importar os dados em excel para o Rstudio e preparar os dados para execução dos gráficos.

PASSO 1: Importar, preparar e inspecionar os dados
Iniciamos com os pacotes requeridos:

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("gridExtra")
install.packages("tidytext")
install.packages("wordcloud2")
install.packages("readxl")
install.packages("openxlsx")
install.packages("textdata")
install.packages("tidyr")
install.packages("knitr")
install.packages("kableExtra")
install.packages("widyr")
install.packages("igraph")
install.packages("ggraph")
install.packages("ggrepel")
library(dplyr)
library(ggplot2)
library(gridExtra)
library(tidytext)
library(wordcloud2)
library(readxl)
library(openxlsx)
library(textdata)
library(tidyr)
library(knitr)
library(kableExtra)
library(dplyr)
library(widyr)
library(igraph)
library(ggraph)
library(ggrepel)
```

Feita a instalação e convocação dos muitos pacotes necessários, vamos importar a base de dados kenya.xls pra o Rstudio, denominado o objeto de discurso:

discurso <- read_excel("C:/Users/Paulo/Documents/Documents/Estatistica I/Análise de Texto/kenya.xls", sheet = 1)

Vocês podem importar os dados das outras formas que aprendemos também. O comando que eu utilizei, caso queiram trabalhar várias vezes com o mesmo banco de dados, como no meu caso, ele é bem mais rápido que as outras alternativas. O próximo comando dirá ao R qual o nome da coluna que contém os discursos dos legisladores. No nosso caso, a última coluna denominada texto:

speeches <- discurso$texto

Os próximos comandos criam dois objetos que separam os discursos dos legisladores de direita e esquerda no paramento queniano, utilizando a variável ideologia do banco de dados. Assim podemos fazer análises para cada grupo de parlamentares e partidos políticos em relação aos discursos sobre a China.

speeches_right <- discurso$texto[discurso$ideologia == 'Direita']
speeches_left <- discurso$texto[discurso$ideologia == 'Centro-esquerda']

Vamos inspecionar os dados e ver os nomes das nossas variáveis todas, a matriz invertida dos dados e quantas linhas e colunas temos, respectivamente:

names(discurso)
glimpse(discurso)
dim(discurso)

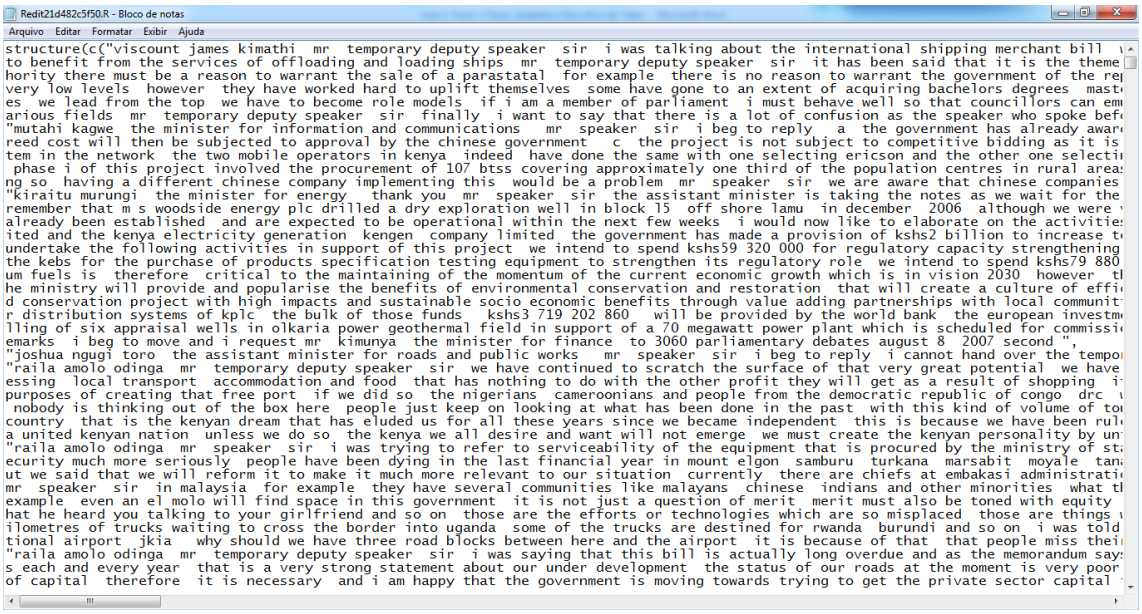
```
> discurso <- read_excel("C:/Users/Paulo/Documents/Documents/Estatistica I/Anál
  de Texto/kenya.xls", sheet = 1)
> speeches <- discurso$texto
> speeches_right <- discurso$texto[discurso$ideologia == 'Direita']
> speeches_left <- discurso$texto[discurso$ideologia == 'Centro-esquerda']
> names(discurso)
 [1] "id_discurso"     "id_legislador"   "nome_legislador" "genero"
 [5] "partido"         "ideologia"       "provincia"       "data_discurso"
 [9] "ano_discurso"    "id_ano"          "casa"            "texto"
> glimpse(discurso)
Observations: 1,029
Variables: 12
$ id_discurso     <dbl> 1155, 1156, 1139, 1102, 1103, 1126, 1117, 1158, 11...
$ id_legislador   <dbl> 104, 42, 2, 57, 62, 62, 62, 2, 108, 18, 63, 103, 9...
$ nome_legislador <chr> "Viscount Kimathi", "Mutahi Kagwe", "Kiraitu Murun...
$ genero          <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", ...
$ partido         <chr> "KANU", "NARC", "PNU", "NARC", "ODM", "ODM", "ODM"...
$ ideologia       <chr> "Direita", "Centro-esquerda", "Direita", "Centro-e...
$ provincia       <chr> "Central", "Central", "Central", "Central", "Centr...
$ data_discurso   <chr> "39211", "39242", "39302", "39390", "39390", "3939...
$ ano_discurso    <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 20...
$ id_ano          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0,...
$ casa            <chr> "National Assembly", "National Assembly", "Nationa...
$ texto           <chr> "Viscount James Kimathi \nMr. Temporary Deputy Spe...
> dim(discurso)
[1] 1029   12
```

Percebam que temos vaiáveis como o partido do legislador, a província que ele representa, a ideologia do partido político, a data do discurso, o gênero do legislador ou

legisladora e a casa legislativa onde o discurso foi proferido. Lembrando que a variável texto é o discurso legislativo sobre a China. Nós temos 1029 discursos e 12 variáveis no total. Vejam que todas são qualitativas nominais (chr), cm exceção dos números de identificação dos legisladores e discurso e o ano e data dos mesmos.

Após verificar a base de dados, vamos ajeitar os dados de texto, os discursos dos parlamentares quenianos. Vamos dar uma olhada no nosso objeto speeches criado anteriormente.

fix(speeches)



O objeto speeches acima contém todos os textos dos discursos de todos os parlamentares na nossa série temporal, como podem observar na janela que apareceu acima. Podem fechar a janela depois. O que vamos fazer agora é limpar o texto, transformar todas as palavras em minúsculas (para que o R não as leia como diferentes) e retirar caracteres indesejados como pontuação e números.

speeches <- sapply(speeches, tolower)
removeSpecialChars <- function(x) gsub("[^a-zA-Z0-9]", " ", x)
speeches <- sapply(speeches, removeSpecialChars)

Com os discursos “limpos”, vamos também usar o comando summary que conhecemos na aula 3 para ver um resumo das variáveis.

summary(discurso)

```
> speeches <- sapply(speeches, tolower)
> removeSpecialChars <- function(x) gsub("[^a-zA-Z0-9 ]", " ", x)
> speeches <- sapply(speeches, removeSpecialChars)
> summary(discurso)
  id_discurso   id_legislador nome_legislador       genero
 Min.   : 138   Min.   :  1   Length:1029        Length:1029
 1st Qu.: 395   1st Qu.: 42   Class :character   Class :character
 Median : 652   Median :134   Mode  :character   Mode  :character
 Mean   : 652   Mean   :134
 3rd Qu.: 909   3rd Qu.:208
 Max.   :1166   Max.   :313
   partido           ideologia          provincia         data_discurso
 Length:1029        Length:1029        Length:1029        Length:1029
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


  ano_discurso      id_ano           casa              texto
 Min.   :2007   Min.   :0.0000   Length:1029        Length:1029
 1st Qu.:2011   1st Qu.:0.0000   Class :character   Class :character
 Median :2013   Median :1.0000   Mode  :character   Mode  :character
 Mean   :2013   Mean   :0.6948
 3rd Qu.:2015   3rd Qu.:1.0000
 Max.   :2016   Max.   :1.0000
```

PASSO 2: Gráficos Descritivos

Para personalizar os gráficos, criaremos uma lista exclusiva de cores por meio dos códigos das cores, conformando o objeto my_color que será utilizado adiante.

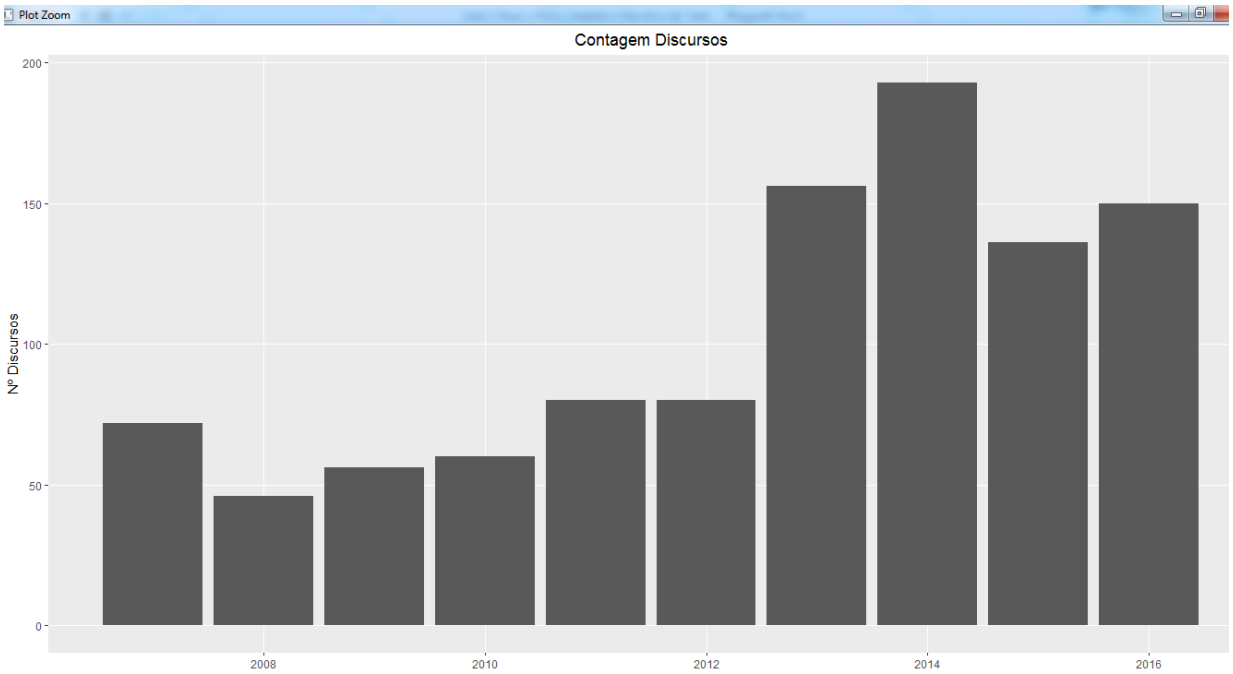
```
my_colors <- c("#E69F00", "#56B4E9", "#009E73", "#CC79A7", "#D55E00")
theme_discurso <- function()
{ theme(plot.title = element_text(hjust = 0.5),
      axis.text.x = element_blank(),
      axis.ticks = element_blank(),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      legend.position = "none") }
```

Eu quero saber quantos são os discursos sobre a China por ano no parlamento do Quênia. Para fazer esse gráfico descritivo, vamos criar o objeto speeches_over_time, por ano (group_by), contabilizando a frequência absoluta de discursos por ano (number_of_texto = n()). Segue o comando, o execute todo de uma vez no Rstudio.

```
speeches_over_time <- discurso %>%
 group_by(ano_discurso) %>%
 summarise(number_of_texto = n())
```

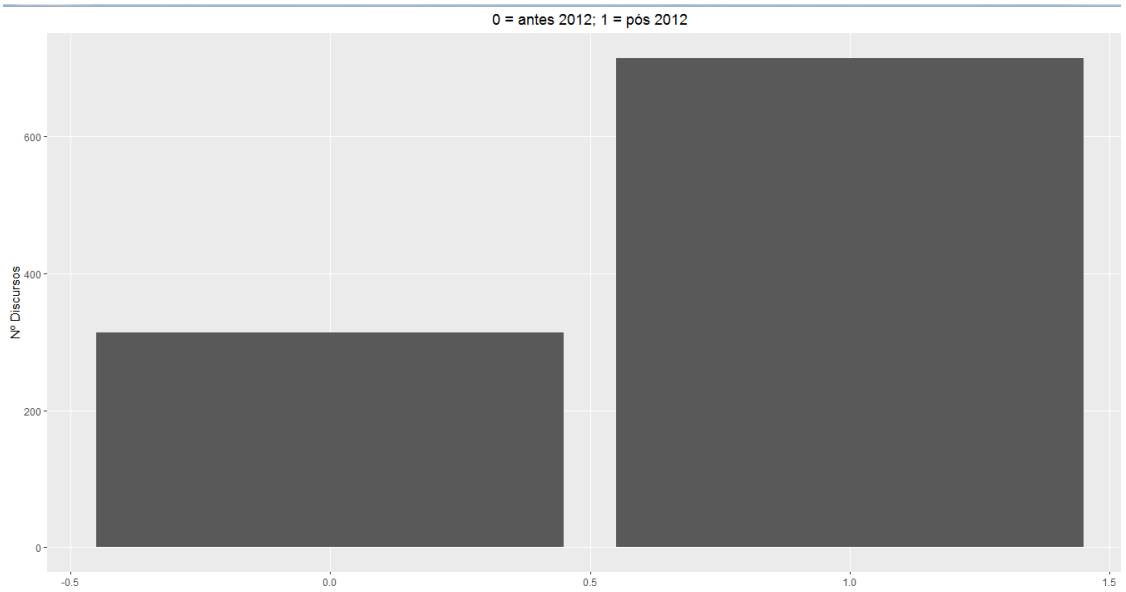
Vamos fazer um gráfico de barras (geom_bar) com o objeto speeches_over_time, usando a função ggplot(), nomeando os eixos e definindo os parâmetros de legenda:

```
speeches_over_time %>%
 ggplot() +
 geom_bar(aes(x = ano_discurso, y = number_of_texto), stat = "identity") +
 theme(plot.title = element_text(hjust = 0.5),
      legend.title = element_blank(),
      panel.grid.minor = element_blank()) +
 labs(x = NULL, y = "Nº Discursos") +
 ggtitle("Contagem Discursos")
```



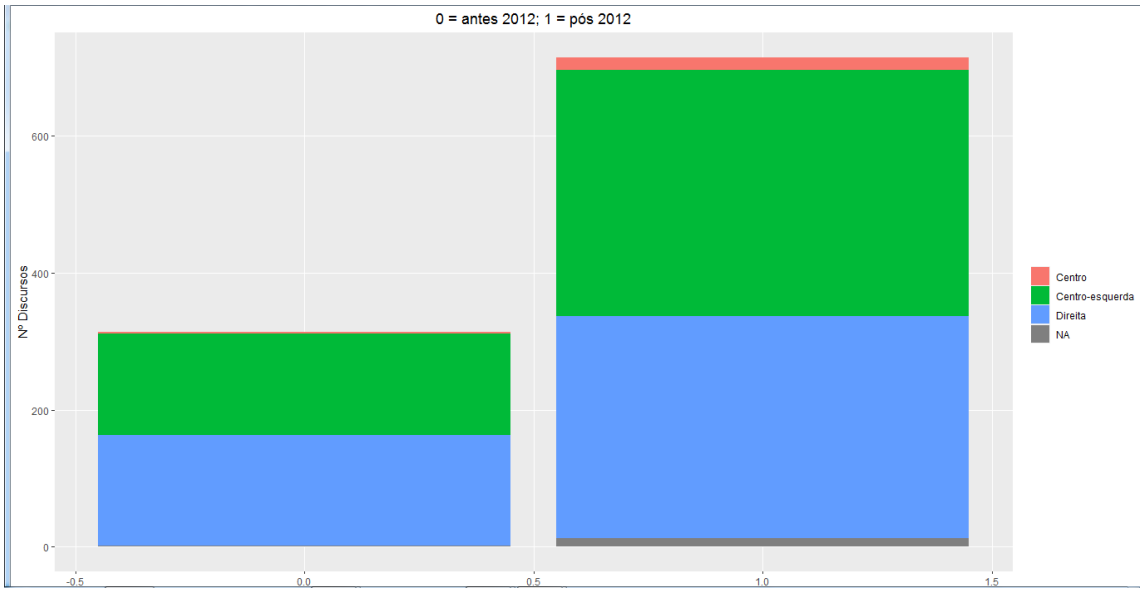
Percebemos pelo gráfico acima o crescimento da quantidade de discursos sobre a China ao longo do tempo no parlamento queniano. Isso evidencia o aumento da importância do país asiático para o país africano, além de já ofertar um indício que pode ter uma relação com a criação da TV chinesa em 2012. Vejam que a partir de 2012 o aumento é bastante nítido. Contudo, não é possível estabelecer uma relação causal apenas com essa descrição dos dados, ou seja, dizer que a TV estatal chinesa causou esse aumento. Muitos fatores de confusão podem desempenhar um papel relevante, como investimentos e a própria ascensão global chinesa no período. Ainda assim, vamos fazer um gráfico de barras que quantificam os discursos antes e depois de 2012 como indicado nos comandos abaixo:

```
speeches_over_time <- discurso %>%
  group_by(id_ano) %>%
  summarise(number_of_texto = n())

speeches_over_time %>%
  ggplot() +
  geom_bar(aes(x = id_ano, y = number_of_texto), stat = "identity") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.title = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = NULL, y = "Nº Discursos") +
  ggtitle("0 = antes 2012; 1 = pós 2012")
```



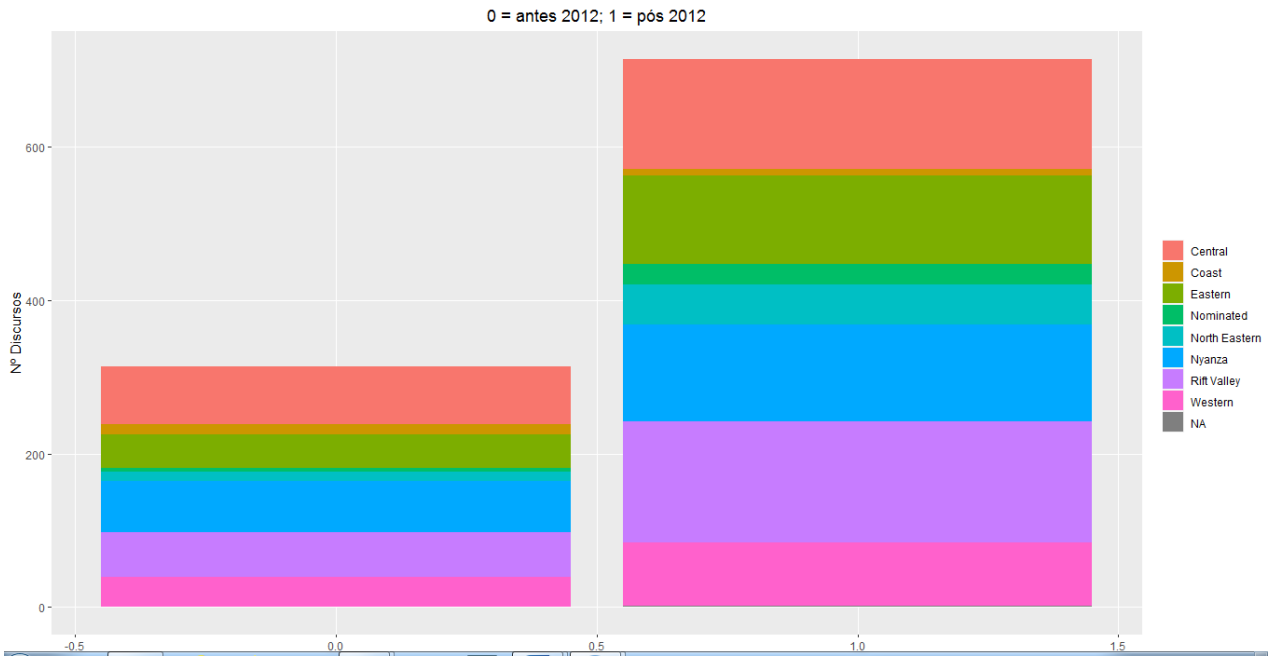
Podemos ver com mais clareza ainda o aumento de discursos sobre a China no parlamento queniano após 2012, barra a direita no gráfico acima. A seguir, vejamos se faz diferença a ideologia dos partidos políticos dos legisladores na quantidade de discursos sobre a China, utilizando um gráfico de barras empilhado. Notem que como estamos usando pacotes específicos, como o ggplot2, os comandos usados aqui, quando comparados aos comandos da aula 4, são um pouco diferentes. Ainda assim, vocês podem notar que a lógica dos comandos é sempre a mesma, conformando uma linguagem do R. Vamos aos comandos:

```
speeches_over_time_ideology <- discurso %>%
  group_by(id_ano, ideologia) %>%
  summarise(number_of_texto = n())

speeches_over_time_ideology %>%
  ggplot() +
  geom_bar(aes(x = id_ano, y = number_of_texto, fill = ideologia), stat = "identity")
+
  theme(plot.title = element_text(hjust = 0.5),
        legend.title = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = NULL, y = "Nº Discursos") +
  ggtitle("0 = antes 2012; 1 = pós 2012")
```



Podemos observar que tanto a direita quanto a centro-esquerda aumentaram a quantidade de discursos sobre a China. Parece bastante proporcional a quantidade de discursos sobre a China em termos de ideologia partidária, ou seja, tanto esquerda quanto direita falam muito da China. Talvez se essa análise fosse feita no Brasil atual, poderia deflagrar uma maior quantidade de menções à China na direita brasileira, principalmente a alinhada ao atual governo que "comprou" a agenda norte-americana de competição com a China. Finalmente, vamos fazer o mesmo gráfico, mas agora por provincial do Quênia para ver se a região que o parlamentar representa revela diferentes quantidades de discurso sobre a China.

```
speeches_over_time_province <- discurso %>%
  group_by(id_ano, provincia) %>%
  summarise(number_of_texto = n())

speeches_over_time_province %>%
  ggplot() +
  geom_bar(aes(x = id_ano, y = number_of_texto, fill = provincia), stat =
"identity") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.title = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = NULL, y = "Nº Discursos") +
  ggtitle("0 = antes 2012; 1 = pós 2012")
```



Assim como para ideologia, a província do parlamentar também parece não ter muito efeito na quantidade de discursos sobre a China. Como podem ter províncias que recebem mais investimento, ou o sinal de TV é melhor e mais casas têm acesso, entre outros fatores, seria possível ter boa variabilidade por província. Como no caso dos parlamentares de Roraima, mesmo do DEM e PSDB, votaram a favor da entrada da Venezuela no MERCOSUL em 2007. Algo que ficou claro com os gráficos descritivos iniciais é que geral o aumento de discursos sobre a China no Quênia. Entretanto, alguns legisladores podem falar bem ou mal da China, a quantificação absoluta como fizemos não nos diz quem elogia e quem critica China. Imaginemos uma análise dessas no Congresso brasileiro recente, o simples fato de Carlos Bolsonaro, por exemplo, falar mal da China, pode ocorrer a reação de vários outros parlamentares defendendo a importância da China para o Brasil. Assim, tanto governo quanto oposição aumentariam a quantidade de discursos mas em direção oposta. Para saber a direção dos discursos, favoráveis ou contrários à China, vamos realizar uma análise de sentimento das palavras, classificando-as como positivas ou negativas.

PASSO 3: Text Mining (mineração de dados de texto)

Vamos começar fazendo uma nuvem de palavras para ver quais tipos de palavras são mais utilizadas pelos legisladores nos discursos em que a China aparece. Iniciamos retirando uma lista de palavras que a Clarice criou referente a muitos termos e palavras procedimentais utilizadas no legislativo do Quênia:

undesirable_words <- c("deputy", "speaker", "mr", "mister", "ministry", "minister", "committee", "senator", "senators", "members", "member", "temporary", "laughter", "debates", "debate", "thank you", "house", "parliament", "senate", "editor", "cent", "court", "hansard", "time", "motion", "support", "bill", "report", "county", "country")

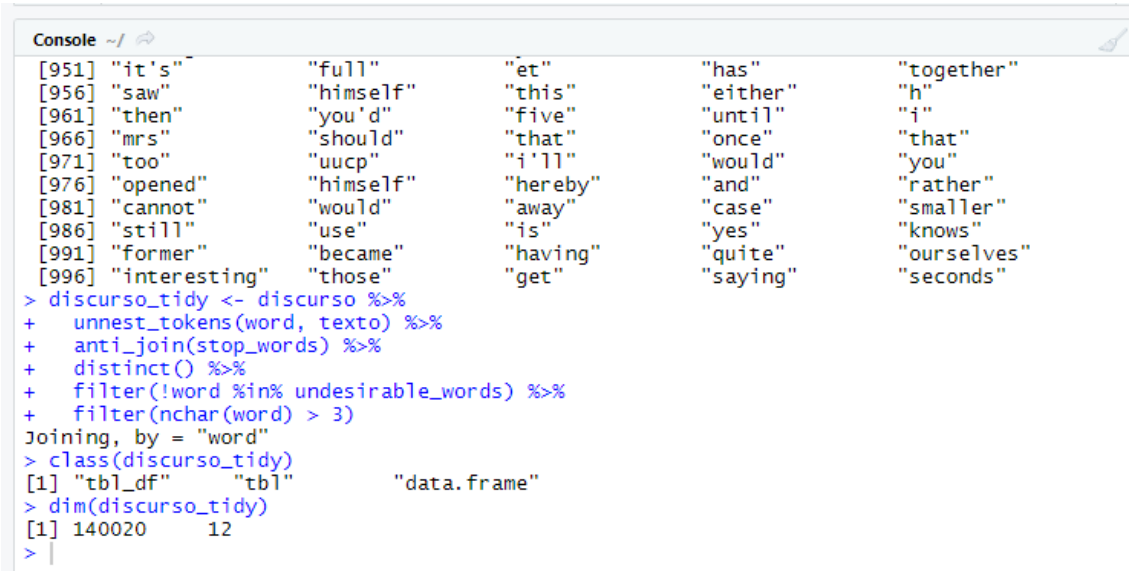
Na sequência, vamos eliminar as chamadas "stop words", palavras muito comuns sem significado algum para a nossa análise. Em português, alguns exemplos seriam: onde, assim, ainda, seguido, normalmente, entre outras. No comando abaixo vamos limitar em 1000 (mil) as stopwords do vocabulário em inglês, aparecerá a lista de palavras para vocês checarem:

```
head(sample(stop_words$word, 1000), 1000)
```

```
discurso_tidy <- discurso %>%
  unnest_tokens(word, texto) %>%
  anti_join(stop_words) %>%
  distinct() %>%
  filter(!word %in% undesirable_words) %>%
  filter(nchar(word) > 3)
```

Vamos utilizar os comandos de resumo dos dados para ver o nosso objeto recém criado dicurso_tidy.

```
class(discurso_tidy)
dim(discurso_tidy)
```



Vejam que esse novo objeto que criamos, discurso_tidy, possui uma característica diferente das demais vistas até o momento. Transformamos o nosso banco de dados de tal forma que cada palavra dita no discurso vira uma linha da nossa matriz de dados. Para visualizar essa descrição vocês podem utilizar o comando fix()

```
fix(discurso_tidy)
```

Agora estamos prontos para fazer a nuvem de palavras com o comando abaixo, estabelecendo 300 palavras, da mais frequente a menos frequente:

```
discurso_words_counts <- discurso_tidy %>%
  count(word, sort = TRUE)
wordcloud2(discurso_words_counts[1:300, ], size = .5)
```

PASSO 4: Análise de Sentimento

Iniciamos agora a análise de sentimento. Temos o banco de dados no formato adequado já, com cada palavra na linha da matriz de dados como vocês puderam notar anteriormente. Precisaremos de mais pacotes no R, aqueles específicos sobre sentimentos das palavras. Basicamente, esses pacotes e comandos de sentimento importam para o R dicionários de palavras da língua inglesa previamente classificadas como negativas, neutras ou positivas. Um trabalho prévio realizado por linguistas, compreendendo enormes listas de palavras que buscam cobrir o máximo de palavras de determinada língua. Para o português, por exemplo, há pacotes, mas em menor abundância e qualidade quando comparado ao inglês. No comando, vai aparecer uma pergunta dos criadores garantindo que se publicarmos algo vamos citar a fonte da base de dados como indicado na mensagem do R. Digite 1, a opção yes, no console do R e ele vai baixar todos os léxicos. Na instalação do pacote textdata, caso apareça uma janela solicitando reiniciar o R, clique em não (no).

install.packages("textdata")

get_sentiments("afinn")

get_sentiments("bing")

get_sentiments("nrc")

O pacote tidytext inclui um conjunto de dados chamado sentimentos que fornece vários léxicos distintos. Esses léxicos são dicionários de palavras com uma categoria ou valor de sentimento atribuído. O tidytext fornece três léxicos de uso geral:

affin: atribui palavras com uma pontuação que varia entre -5 e 5, com pontuações negativas indicando sentimentos negativos e pontuações positivas indicando sentimentos positivos

Bing: atribui palavras a categorias positivas e negativas

nrc: atribui palavras a uma ou mais das dez categorias a seguir: positivo, negativo, raiva, antecipação, nojo, medo, alegria, tristeza, surpresa e confiança

Na realização da nuvem de palavras, acima, observamos unigramas ou palavras únicas. Considerem agora a palavra "China", por exemplo, muito comum na nossa base de dados já que é o próprio objeto da mesma. Quais palavras a precede? Ou segue? Olhar palavras soltas fora de contexto pode não ser muito útil. Formaremos, portanto, alguns bigramas ou pares de palavras. O pacote tidytext oferece essa possibilidade. Convocamos o comando unnest_tokens (), acrescentando o argumento do token para ngrams. Como vamos formar bigrams (duas palavras consecutivas), n = 2.

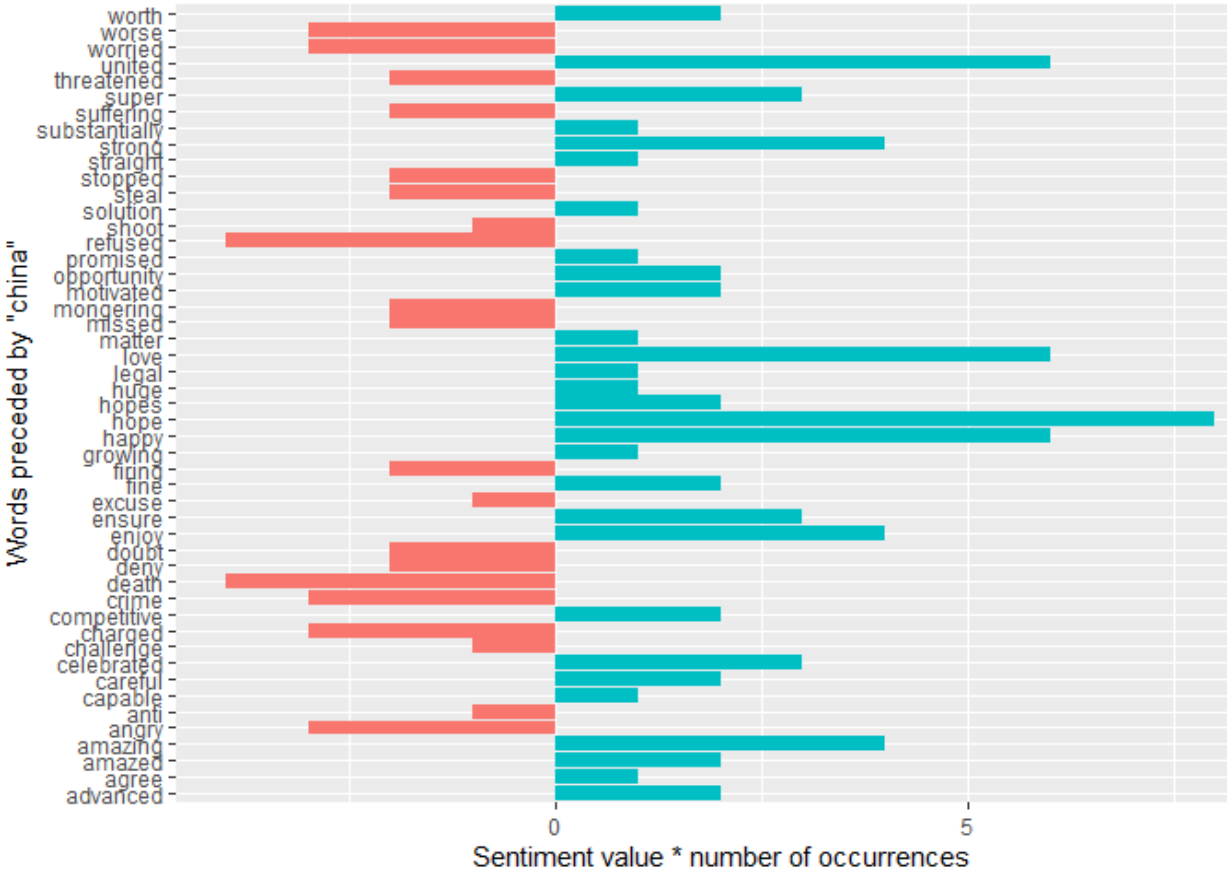
```
library(tidyr)
discurso_bigrams <- discurso_tidy %>%
unnest_tokens(bigram, word, token = "ngrams", n = 2)

discurso_bigrams %>%
count(bigram, sort = TRUE)
```

Vamos tirar a lista de palavras indesejadas

```
bigrams_separated <- discurso_bigrams %>%
separate(bigram, c("word1", "word2"), sep = " ")

 bigrams_filtered <- bigrams_separated %>%
 filter(!word1 %in% stop_words$word) %>%
 filter(!word2 %in% stop_words$word) %>%
 filter(!word1 %in% undesirable_words) %>%
 filter(!word2 %in% undesirable_words)
```

Agora vamos analisar o sentimento das palavras que antecedem a palavra “China” nos discursos dos parlamentares quenianos.

```
AFINN <- get_sentiments("afinn")

china_words <- bigrams_filtered %>%
  filter(word1 == "china") %>%
  inner_join(AFINN, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)

china_words  %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(50) %>%
  ggplot(aes(word2, n * value, fill = n * value >= 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Words preceded by \"china\"") +
  ylab("Sentiment value * number of occurrences") +
  coord_flip()
```



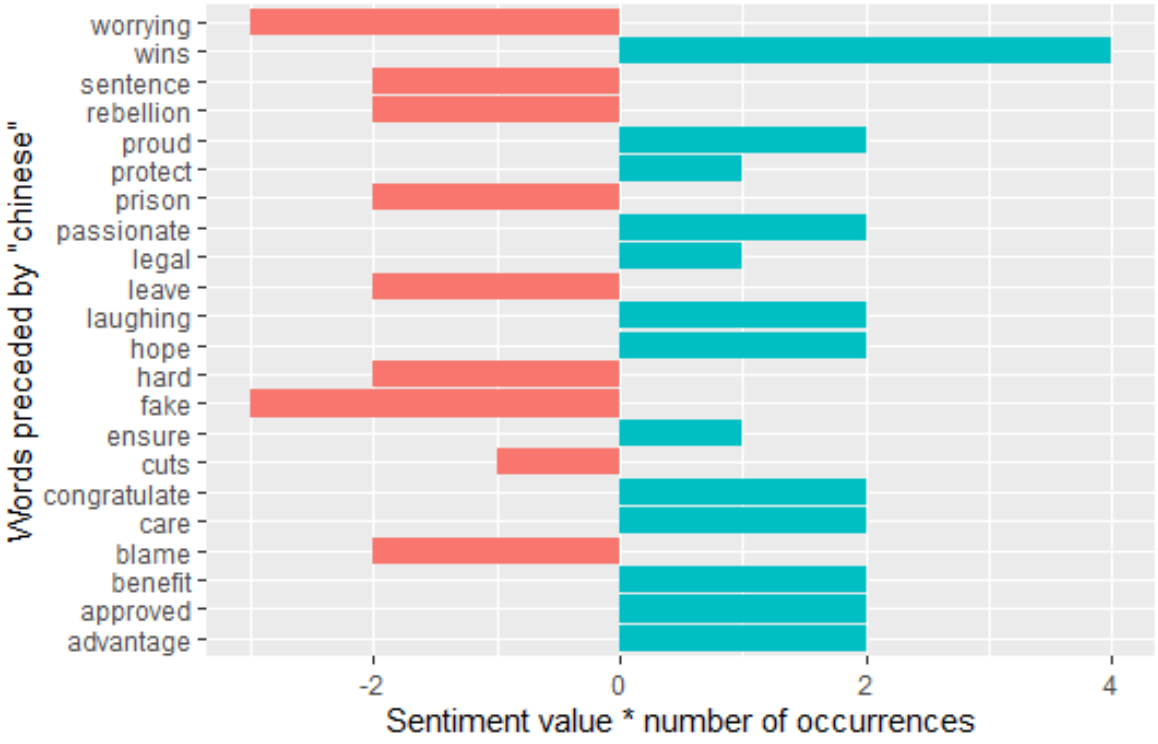
O léxico AFINN, criado no objeto do primeiro comando, atribui às palavras pontuações de -5 a 5, com pontuações negativas indicando sentimento negativo e pontuações positivas indicando sentimento positivo. Filtradas as palavras indesejadas e stopwords, dividimos os textos dos discursos em bigrams, pares de palavras. Em seguida, separamos esses pares em “word1” e “word2”. A partir disso, selecionamos apenas as palavras precedidas por “China” e geramos nosso primeiro gráfico de sentimento, utilizando o léxico AFINN, que atribui às palavras pontuações de -5 a 5.

Observa-se no gráfico acima que a maioria das palavras precedidas pro China são classificadas com um valor maior que zero, ou seja, são palavras cujo sentimento

atribuído é positivo. Podemos dar destaque a algumas delas, como “advanced”, “competitive”, “ensure”, “growing”, “opportunity”, “solution” e “strong”. Quantos às palavras com valor negativo, destacamos “worried”, “threatened”, “worse”, “anti” e “refused”.

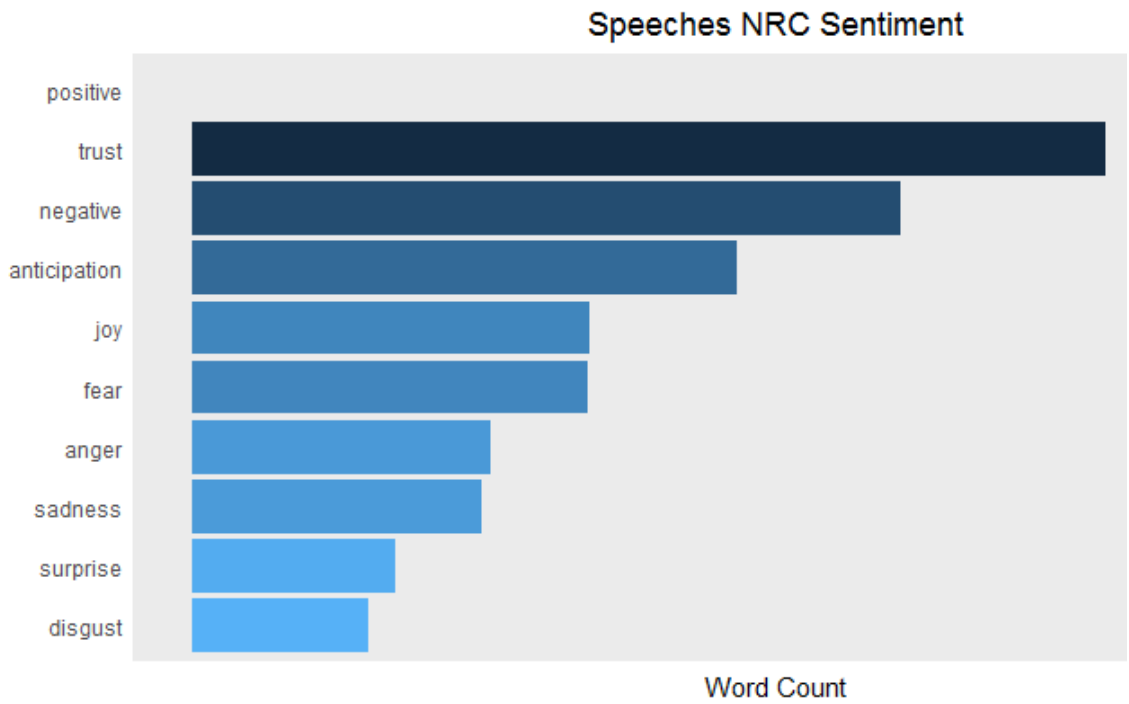
Vamos fazer a mesma coisa, mas agora ao invés de usar a palavras China, utilizaremos a palavra chinese, ou seja, vamos analisar as palavras nos discursos legislativos que precedem chinese.

```
chinese_words <- bigrams_filtered %>%
  filter(word1 == "chinese") %>%
  inner_join(AFINN, by = c(word2 = "word")) %>%
  count(word2, value, sort = TRUE)

chinese_words  %>%
  mutate(contribution = n * value) %>%
  arrange(desc(abs(contribution))) %>%
  head(50) %>%
  ggplot(aes(word2, n * value, fill = n * value > 0)) +
  geom_col(show.legend = FALSE) +
  xlab("Words preceded by \"chinese\"") +
  ylab("Sentiment value * number of occurrences") +
  coord_flip()
```



A maioria das palavras tem valor positivo e podemos destacar as palavras “protect”, “ensure”, “benefit” e “advantage”. Quanto às palavras negativas, destacamos “worrying” e “fake”. Em seguida, vamos utilizar o léxico nrc, que categoriza as palavras em positivo, negativo, raiva, antecipação, nojo, medo, alegria, tristeza, surpresa e confiança para gerar um gráfico indicando os sentimentos predominantes nos discursos:

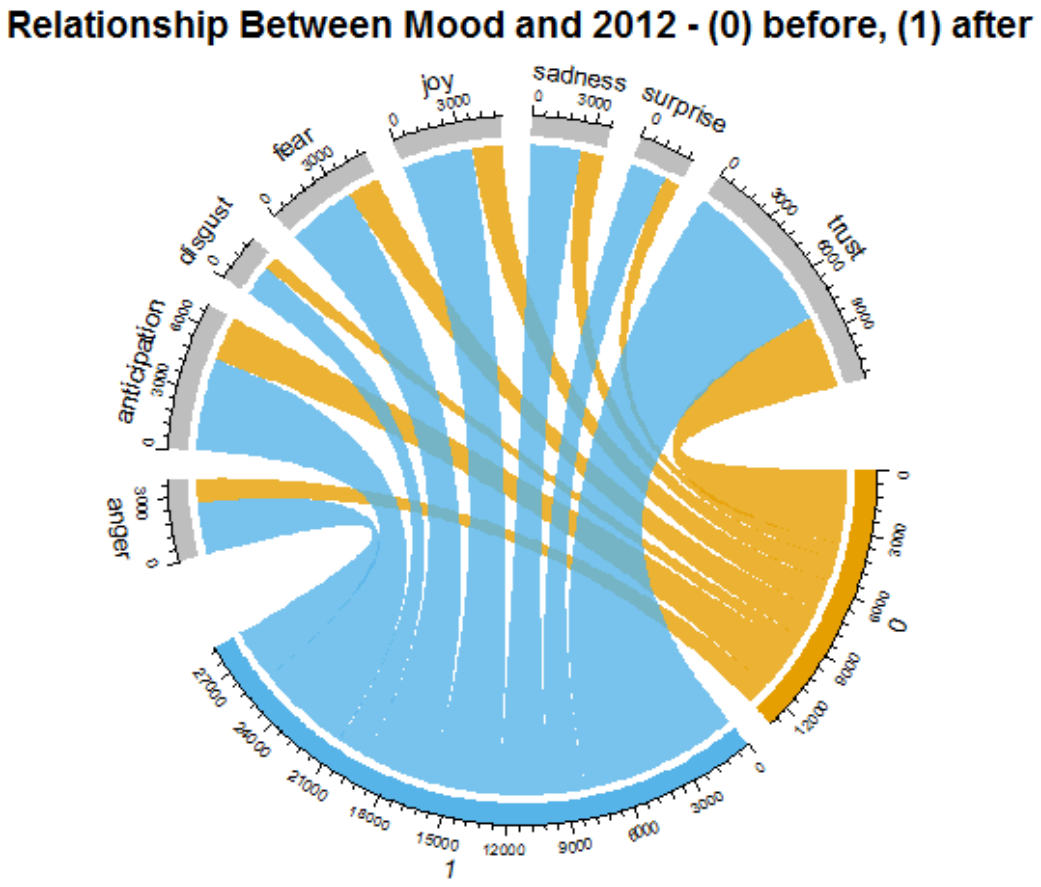
```
discurso_bing <- discurso_tidy %>%
  inner_join(get_sentiments("bing"))
discurso_nrc <- discurso_tidy %>%
  inner_join(get_sentiments("nrc"))
discurso_nrc_sub <- discurso_tidy %>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!sentiment %in% c("positive", "negative"))

nrc_plot <- discurso_nrc %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  ggplot(aes(sentiment, word_count, fill = -word_count)) +
  geom_col() +
  guides(fill = FALSE) + #Turn off the legend
  theme_discurso() +
  labs(x = NULL, y = "Word Count") +
  scale_y_continuous(limits = c(0, 15000)) + #Hard code the axis limit
  ggtitle("Speeches NRC Sentiment") +
  coord_flip()
plot(nrc_plot)
```



Observamos que tanto o sentimento positivo quanto o negativo são frequentes, destacando-se também o sentimento de confiança. No gráfico acima, ocorreu algum erro que não consegui consertar, pois a barra de positivo, a maior, não aparece. Anda assim, é muito clara predominância de sentimentos positivos em relação à China por parte das legisladoras e legisladores do Quênia.

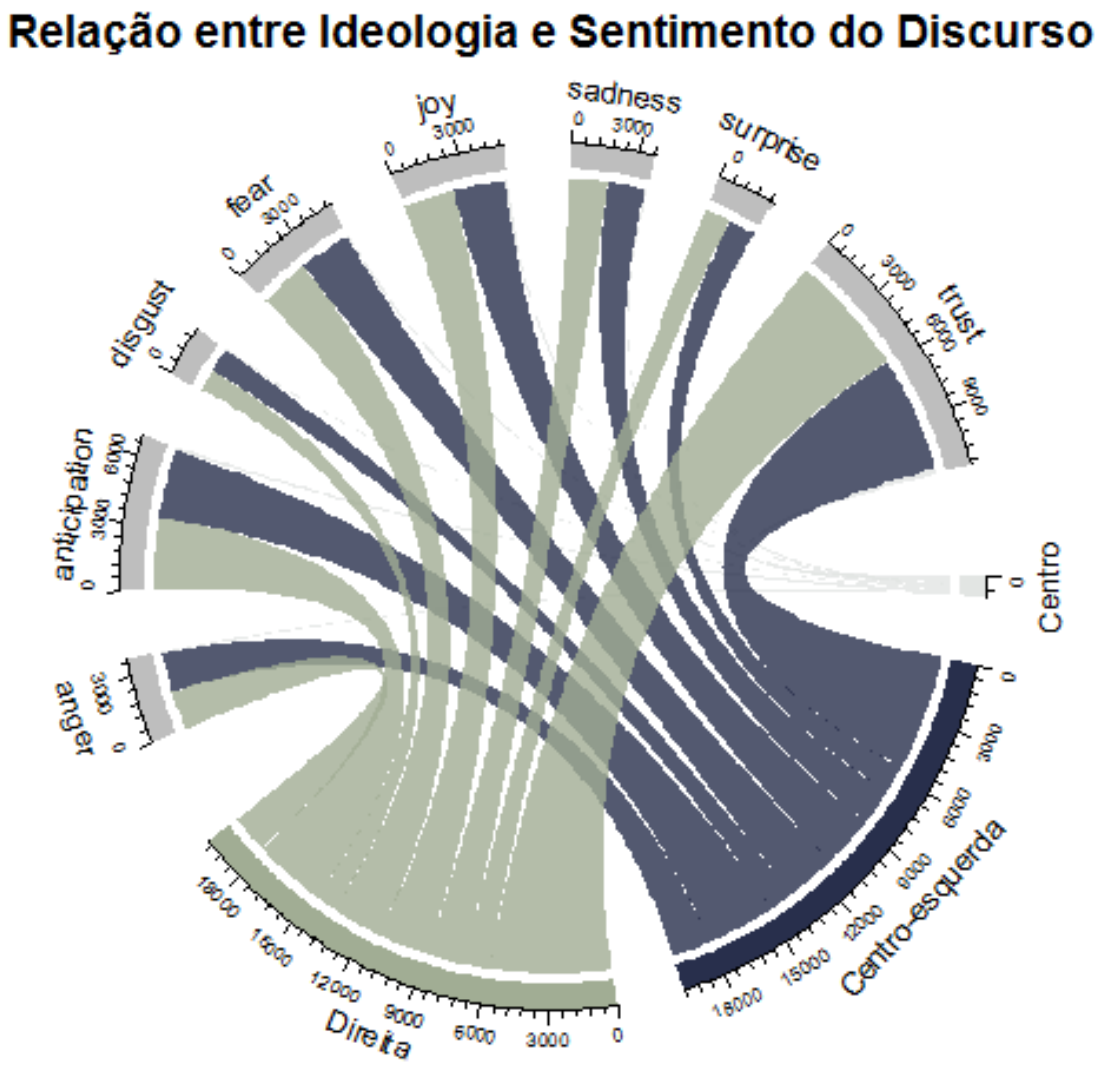
Agora iremos utilizar o léxico bing e faremos um gráfico em forma de círculo, indicando a relação entre o sentimento relacionado às palavras e o ano em que o TV estatal chinesa foi instalada no Quênia, tendo o antes (2006-2011 = 0) e o depois (2012-2018-1). Seguem os comandos:

```
install.packages("circlize")
library(circlize)

grid.col = c("0" = my_colors[1], "1" = my_colors[2], "anger" = "grey",
"anticipation" = "grey", "disgust" = "grey", "fear" = "grey", "joy" = "grey",
"sadness" = "grey", "surprise" = "grey", "trust" = "grey")

decade_mood <-  discurso_nrc %>%
  filter(id_ano != "NA" & !sentiment %in% c("positive", "negative")) %>%
  count(sentiment, id_ano) %>%
  group_by(id_ano, sentiment) %>%
  summarise(sentiment_sum = sum(n)) %>%
  ungroup()

circos.clear()

circos.par(gap.after = c(rep(5, length(unique(decade_mood[[1]])) - 1), 15,
rep(5, length(unique(decade_mood[[2]])) - 1), 15))
chordDiagram(decade_mood, grid.col = grid.col, transparency = .2)
title("Relationship Between Mood and 2012 - (0) before, (1) after")
```



É possível observar que aumentam tanto os discursos com sentimentos positivos como negativos. Embora os gráficos anteriores revelem uma maior quantidade de palavras positivas, percebemos que o sentimento positivo em relação à China está longe de ser unânime entre a elite política do Quênia. Para finalizar essa aula, vamos realizar o mesmo gráfico, mas desta vez utilizaremos a ideologia dos partidos políticos dos legisladores para averiguar se há diferença no sentimento.

```
grid.col = c("Direita" = my_colors[1], "Centro-esquerda" = my_colors[2], "Centro" = my_colors[3], anger" = "grey", "anticipation" = "grey", "disgust" = "grey", "fear" = "grey", "joy" = "grey", "sadness" = "grey", "surprise" = "grey", "trust" = "grey")

decade_mood <-  discurso_nrc %>%
  filter(ideologia != "NA" & !sentiment %in% c("positive", "negative")) %>%
  count(sentiment, ideologia) %>%
  group_by(ideologia, sentiment) %>%
  summarise(sentiment_sum = sum(n)) %>%
  ungroup()

circos.clear()

circos.par(gap.after = c(rep(9, length(unique(decade_mood[[1]])) - 1), 15,
                         rep(9, length(unique(decade_mood[[2]])) - 1), 15))
chordDiagram(decade_mood, grid.col = grid.col, transparency = .2)
title("Relação entre Ideologia e Sentimento do Discurso")
```



Podemos perceber que ideologia também não tem um impacto muito forte no sentimento ser negativo ou positivo, as duas ideologias predominantes, centro-esquerda e direita, distribuem sentimentos negativos e positivos sobre a China. Não há um corte ideológico na aliança do Quênia com a China. Imagino que o mesmo gráfico para o Brasil já poderia revelar uma diferença maior, principal nos partidários do atual presidente da República. A China tem sido bem sucedida em criar uma imagem positiva na elite politica do Quênia, ainda que não seja consensual.

Finalizamos essa aula que introduziu vocês à utilização de estatística descritiva a partir de discursos políticos.