Overview of the most common relevant words in the book *The Scandal of Petroleum*

Tiago de Almeida Silva

2022-04-21

Introduction

This project aims to discover the most relevant common words in one of the classic books of Brazilian Literature. The Scandal of Petroleum ("O Escandalo do Petróleo" in Portuguese) was written by Monteiro Lobato who is considered to be one of the greatest writers in Brazilian history. Despite he is known for his children's books which are very famous in the country, he used to be a visionary and successful businessman and in 1936 he wrote the polemical book entitled *The Scandal of Petroleum* in which he criticized the way Brazilian oligarchs such as politicians and businessmen were working together with some geologists from the USA to hide from the public the oil and gas reserves in the country. Monteiro Lobato was one of the first people in Brazil who brought this issue to the surface.

I chose this book because I am a petroleum engineer and I managed to buy and read the first version of the book from 1936 and I got astonished by the way he depicted this corruption scandal in his book. It is very interesting to notice how power is involved in the oil and gas industry even from the beginning. It was also curious to see the use of the archaic Portuguese in the first version because the writing was very different from the current one. I am using the new version released in 2011 to make the construction of the Word Cloud easier as some of the words in the first version do not exist anymore and that could cause some mismatches when using the "stop words library" and the stemming process which aims to reduce words to their basic syntax.

Part 1 - Bag of Words and Word Cloud

Installing and loading the packages that will be used in this project:

```
pacotes <- c("tidytext","ggplot2","dplyr","tibble","wordcloud","stringr",
             "SnowballC","widyr","janeaustenr", "stringi", "stopwords")

if(sum(as.numeric(!pacotes %in% installed.packages())) != 0){
  instalador <- pacotes[!pacotes %in% installed.packages()]
  for(i in 1:length(instalador)) {
    install.packages(instalador, dependencies = T)
    break()}
  sapply(pacotes, require, character = T)
} else {
  sapply(pacotes, require, character = T)
}
```

Assigning The Scandal of Petroleum Book to an object named "book":

I needed to remove some unuseful pages from the original file to not affect the word count (bag of words) and word cloud. Pages like summary and merchandising from the publishing company were removed.

```
book <- read.delim("TheScandalOfPetroleum.txt", header = F, encoding = "UTF-8") %>%
  rename("text" = "V1")

book %>%
  head(15) %>%
  knitr::kable()
```

text
O ESCÂNDALO DO PETRÓLEO
Primeira Parte
Introdução
O caso do petróleo brasileiro prende-se ao caso do petróleo em geral. Esse produto é o sangue da terra, é a alma da indústria moderna, é a eficiência do poder militar, é a soberania, é a dominação. Tê-lo, é ter o sésamo abridor de todas as portas. Não té-lo é ser escravo. Daí a fúria moderna na luta pelo petróleo. O livro de Essad Bey revela tudo isso do modo mais impressionante.
A base do poder dos Estados Unidos está, sobretudo, no petróleo. Arrancam do seio da terra quase um bilhão de barris por ano, na maior parte consumidos lá — e nossa imaginação tonteia ao calcular o que tamanha onda de óleo, transfeita em energia mecânica, representa para a economia daquele povo.
Qui aura le pétrole aura 1'Empire, escreveu Henri Bérenger na nota diplomática que em 1928 endereçou a Clemenceaa nas vésperas da conferência franco-brita"nica sobre o futuro do mundo.
"Império dos mares, por meio das essências leves; irripério dos continentes, por meio da gasolina. E império do mundo, por meio do poder financeiro desse produto, mais precioso, mais envolvente e mais dominador do planeta do que o próprio ouro."

Separating and assigning each word (token) to a different row in the dataframe

```
#removing numbers and unuseful letters from the text before the "tokenization"

nums <- book %>% filter(str_detect(text, c("0", "1", "2", "3", "4",
                                           "5", "6", "7", "8", "9",
                                           "la", "las", "lo", "los"))) %>% select(text)

book <- book %>%  anti_join(nums, by = "text")

#"tokenization"

book <- book %>% unnest_tokens(word, text)

book %>%
  head(15) %>%
  knitr::kable()
```

Excluding the stop words (like articles and prepositions) in Portuguese. I chose the current version of the book for this reason as the first one contains many words that do not exist anymore in modern Portuguese.

Ps: I will also add some English words to my stopwords list because English is a very used language in this industry and I do not want to see in my word cloud prepositions, articles, and meaningless words.

Code which shows the package with some of the stop words in Portuguese that will be used in this project:

```
head(stopwords::stopwords("portuguese"), 40)
```

```
##  [1] "de"     "a"      "o"      "que"    "e"      "do"     "da"     "em"
##  [9] "um"     "para"   "com"    "não"    "uma"    "os"     "no"     "se"
## [17] "na"     "por"    "mais"   "as"     "dos"    "como"   "mas"    "ao"
## [25] "ele"    "das"    "à"      "seu"    "sua"    "ou"     "quando" "muito"
## [33] "nos"    "já"     "eu"     "também" "só"     "pelo"   "pela"   "até"
```

Removing the stop words from different sources through an anti_joins:

```
book_final <-  book %>%  anti_join(get_stopwords(language = "pt",
                                                 source = "snowball"), by = "word")
book_final <- book_final %>% anti_join(get_stopwords(language = "en",
                                                 source = "snowball"),by = "word")
book_final <- book_final %>% anti_join(get_stopwords(language = "pt",
                                                 source = "nltk"),by = "word")
book_final <- book_final %>% anti_join(get_stopwords(language = "pt",
                                                 source = "stopwords-iso"),by = "word")
```

Appling the stemming process to reduce the words to their basic syntax. That's important to count the words with similar syntaxes such as "Brasileiro" and "Brasileira" which is the gender differentiation of a person who was born in Brazil, male and female respectively but both present the same meaning and syntax.

```
book_stem <- book_final %>%
  mutate(stem = wordStem(word))

book_stem %>%
  head(15) %>%
  knitr::kable()
```

word	stem
escândalo	escândalo
petróleo	petróleo
introdução	introdução
caso	caso
petróleo	petróleo
brasileiro	brasileiro
prende	prend
caso	caso
petróleo	petróleo
produto	produto
sangue	sangu
terra	terra
alma	alma
indústria	indústria
moderna	moderna

As we can see above, the dataframe has now a new column named "stem" where the basic syntax of every single word in the "word" column is shown.

I will start the first step of creating the word cloud and for this reason, I need to count how many different syntaxes there are in the dataframe.

```
book_count <- book_stem %>%
  select(word) %>%
  count(word, sort = T)

book_count %>%
  head(15) %>%
  knitr::kable()
```

word	n
petróleo	229
brasil	86
lobato	57
mundo	50
terra	50
país	49
governo	47
imposto	46
café	42
homem	37
barris	36
standard	30
milhões	29
nacional	29
ministério	28

Finally I will plot the word cloud with the "word" column:

```
pal <- brewer.pal(8,"Dark2")

book_count %>%
  with(wordcloud(word, n, random.order = F, max.words = 100, colors = pal))
```

As seen above, most of the words are related to the title of the book and the main one is the word “Petróleo” which means petroleum in Portuguese.

And the word cloud with the “stem” column:

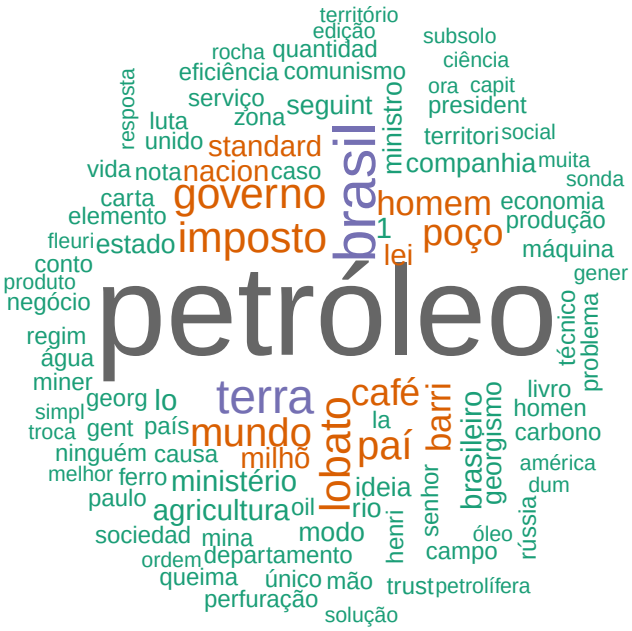
```
#counting the "stem" column

book_count_stem <- book_stem %>%
  select(stem) %>%
  count(stem, sort = T)

#ploting the word cloud

pal <- brewer.pal(8,"Dark2")

book_count_stem %>%
  with(wordcloud(stem, n, random.order = F, max.words = 100, colors = pal))
```



When it comes to the word cloud containing the syntaxes, the outcome is very similar to the first word cloud but it is interesting to notice some differences in the composition of the cloud, like the word syntaxes: *miner, homen, barri, nacion, paí* and so on. All these words are abbreviations (syntaxes) of different words.

Part 2 - Finding Related words in the book through n-grams

In the first part of this project, I got the most common words in The Scandal of Petroleum book but now I will analyze which words (4 words per row in this project) are related to each other in the text. That's something interesting to analyze because we can get four words that appear together many times in the text and perhaps they present a different meaning together in comparison to their single meaning.

Loading some objects created in Part 1 to Part 2:

```
#Loading the book The Scandal of Petroleum

book2 <- read.delim("TheScandalOfPetroleum.txt", header = F, encoding = "UTF-8") %>%
  rename("text" = "V1")

#removing numbers and unuseful letters from the text before the "tokenization"
```

```
nums <- book2 %>% filter(str_detect(text, c("0", "1", "2", "3", "4",
                                            "5", "6", "7", "8", "9",
                                            "la", "las", "lo", "los"))) %>% select(text)

book2 <- book2 %>%  anti_join(nums, by = "text")
```

I will use the *tokenization* process with the n-gram concept which aims to separate words according to an n-number value per row. In my project, I will use as mentioned above the n-number = 4 because we can get very meaningful insights with this amount.

```
book2_token <- book2 %>%
  unnest_tokens(word, text, token = "ngrams", n = 4)

book2_token %>%
  head(15) %>%
  knitr::kable()
```

word
o escândalo do petróleo
NA
NA
o caso do petróleo
caso do petróleo brasileiro
do petróleo brasileiro prende
petróleo brasileiro prende se
brasileiro prende se ao
prende se ao caso
se ao caso do
ao caso do petróleo
caso do petróleo em
do petróleo em geral
petróleo em geral esse
em geral esse produto

Counting the object "book2":

```
book2_count <- book2_token %>% count(word, sort = TRUE)

book2_count %>%
  head(15) %>%
  knitr::kable()
```

word	n
NA	177
conselho nacional do petróleo	7
nota da edição de	7
o valor da terra	6
do conselho nacional do	5
do ministério da agricultura	5

word	n
o petróleo do lobato	5
da edição de 1946	4
matéria prima da máquina	4
senhor fleury da rocha	4
a primeira edição de	3
amor de deus que	3
contra a vontade da	3
da américa do sul	3
da edição de 1946	3

Removing the Nas from the book probably coming from the white spaces at the beginning of each chapter and paragraphs:

```
book2_count <- na.omit(book2_count)
```

Taking a quick look at the word counting:

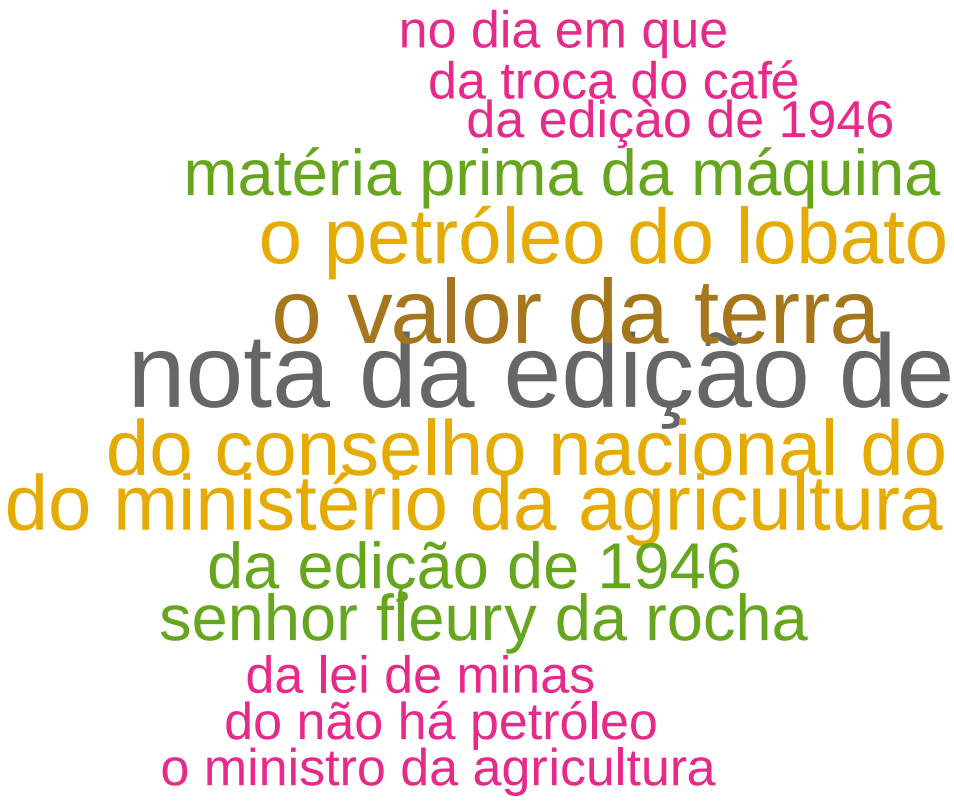
```
book2_count %>%
  head(15) %>%
  knitr::kable()
```

	word	n
2	conselho nacional do petróleo	7
3	nota da edição de	7
4	o valor da terra	6
5	do conselho nacional do	5
6	do ministério da agricultura	5
7	o petróleo do lobato	5
8	da edição de 1946	4
9	matéria prima da máquina	4
10	senhor fleury da rocha	4
11	a primeira edição de	3
12	amor de deus que	3
13	contra a vontade da	3
14	da américa do sul	3
15	da edição de 1946	3
16	da lei de minas	3

Plotting a word cloud of the 15 most common short phrases in the book:

```
pal <- brewer.pal(8,"Dark2")

book2_count %>%
  with(wordcloud(word, n, random.order = F, max.words = 15, colors = pal))
```



It is possible to see very outstanding short phrases such as:

- *amor de Deus que / the love of God*, it shows how religious the society was in the 30s;
- *matéria prima da máquina / machine's raw material*, it appears many times in the book and it refers to the development of new machines throughout human history and how they are fueled;
- *do ministério da agricultura / the ministry of agriculture*, the ministry responsible for the petroleum industry in Brazil in 1936;
- *o valor da terra / the land's value*, it refers to the Americans who were buying lands with very high geological/oil prospection value for very cheap prices;
- *petróleo do lobato / lobato's petroleum*, despite it having a similar name to the book's author, Lobato is a neighborhood of Salvador, the capital of the state of Bahia, where it was drilled the first oil and gas well in Brazil;
- *senhor Fleury da Rocha / Sir Fleury da Rocha*, Fleury da Rocha was the vice-president of the National Council of Petroleum and it was accused by Lobato many times in the book to sabotage the information

of the petroleum existence in Brazil.

- *não há petróleo no / there is not petroleum in Brazil*, Monteiro Lobato used to say it many times because Fleury da Rocha and the American geologists had rejected the idea of profitable oil reserves in the country.

Part 3 - Analyzing the correlation between pairs of words in the book

```
book2_sections <-  book2 %>%
  mutate(section = row_number() %/% 10) %>%
  filter(section > 0) %>%
  unnest_tokens(word, text) %>%
  filter(!word %in% stop_words$word)

#Removing the stop words from different sources through an anti_joins:

book2_final <-  book2_sections %>%  anti_join(get_stopwords(language = "pt",
                                                source = "snowball"), by = "word")
book2_final <- book2_final %>% anti_join(get_stopwords(language = "en",
                                                source = "snowball"),by = "word")
book2_final <- book2_final %>% anti_join(get_stopwords(language = "pt",
                                                source = "nltk"),by = "word")
book2_final <- book2_final %>% anti_join(get_stopwords(language = "pt",
                                                source = "stopwords-iso"),by = "word")
```

Counting how many times the words correlate to each other and checking the top 15 words with the highest correlation:

```
word_pairs <- book2_final %>%
  pairwise_count(word, section, sort =  T)

word_pairs %>%
  head(15) %>%
  knitr::kable()
```

We can see that government and petroleum-related words compose the majority of the most common correlations in the book.

Getting the correlation rate of each pair of words from the book and checking the top 25 words with the highest correlation rate:

```
word_cor <- book2_final %>%
  group_by(word) %>%
  filter(n() >= 20) %>%
  pairwise_cor(word, section, sort = T)

word_cor %>%
  head(25) %>%
  knitr::kable()
```

item1	item2	correlation
agricultura	ministério	0.7785684
ministério	agricultura	0.7785684

item1	item2	correlation
lobato	poço	0.4145661
poço	lobato	0.4145661
georgismo	imposto	0.4048986
imposto	georgismo	0.4048986
barris	1	0.4032642
1	barris	0.4032642
lobato	petróleo	0.3592069
petróleo	lobato	0.3592069
imposto	terra	0.3317473
terra	imposto	0.3317473
poço	petróleo	0.3114454
petróleo	poço	0.3114454
café	milhões	0.2758008
milhões	café	0.2758008
brasileiro	governo	0.2743171
governo	brasileiro	0.2743171
standard	petróleo	0.2685150
petróleo	standard	0.2685150
café	governo	0.2612313
governo	café	0.2612313
café	país	0.2476796
país	café	0.2476796
rio	petróleo	0.2408369

Regarding the correlation rate of all pairs of words in the book, *agricultura* (agriculture) and *ministério* (ministry) got, by far, the highest score (78%) and that's easily explained by the fact in 1936 there was no government-owned institution related to the petroleum industry and the Ministry of Agriculture was in charge to prospect and look for oil and gas in the Brazilian territory, and as mentioned before, lots of its staff were involved in corruption scandals with American geologists who were trying to convince the population and press there was no petroleum in the Brazilian territory due to American interests in energy resources from overseas.

Another fascinating outcome of the correlation rate is that we can get a glimpse of the Brazilian economy by that time with word correlations like *governo / país / café* (coffee was the backbone of the Brazilian economy by that time) and *georgismo / terra* (Georgism is an economic ideology holding that, although people should own the value they produce themselves, the economic rent derived from land—including from all natural resources, the commons, and urban locations—should belong equally to all members of society).

Conclusion

It has been 86 years since Monteiro Lobato wrote the fascinating book *The Scandal of Petroleum* about the game of lies and corruption between the Brazilian oligarchs and USA geologists and it is interesting and sad to notice some of the bad features of the society from that time are still visible nowadays. With the word cloud, we were able to see the most common words in the book and the short phrases more used by the author as well. The correlation between pairs of words was something very interesting to analyze due to it is possible to see through numbers how and how many times the words in the book were connected.