Document Text Analysis

Magdiel Ablan

27 de abril de 2019

Introduction

The following document describes an algorithm that given a text (in pdf format) and a list of keywords perform the following:

- Search for keywords
- Count the their frequency
- Highlighting the paragraphs where the keyword was found

The process will be explained in detail with the first keyword. For each keyword a table is produced with the page number (original document), line number (text data frame), and fragment text where it is produced. At the end a table with all the keywords is presented.

Data preparation

First, we need to install all the required libraries:

```
# Set of packages for data analysis:
library(tidyverse)
# tidytext library from the book of Silge and Robinson:
library(tidytext)
# text minning package
library(tm)
# to print nice tables
library(xtable)
```

Then, read the document and extract its content. It is importat to switch comments here for the variable `src_encoding`: **If it is running in Linux, it should be "UTF-8". If running in Mac or Windows, it should be "ISO8859-1".**

```
read <- readPDF(engine="xpdf")
document <- Corpus(URISource("./data/Saudi_Vision2030_EN.pdf"),
                   readerControl = list(reader = read))
doco <- content(document[[1]])
#src_enconding <-"ISO8859-1"
src_enconding <-"UTF-8"
end_enconding <- "UTF-8"
doco <- iconv(doco,src_enconding ,end_enconding, sub="")
head(doco)
```

```
## [1] "\f\f\fMY FIRST OBJECTIVE IS FOR" "OUR COUNTRY TO BE A"
## [3] "PIONEERING AND SUCCESSFUL"       "GLOBAL MODEL OF"
## [5] "EXCELLENCE, ON ALL FRONTS,"      "AND I WILL WORK WITH YOU TO"
```

`doco` is a vector of char strings. Each element of the vector is a line of text. Page breaks are given by the symbol: "\f". It will be useful later to know where the page breaks are located:

```
page_breaks <- grep("\\f", doco)
doco[page_breaks[1]]
```

```
## [1] "\f\f\fMY FIRST OBJECTIVE IS FOR"
```

The first three pages do not have any text, since the consecutives "\f".

Now, we convert all the text to lower case and eliminate page breaks and empty strings. The resulting vector doc is where all the analysis are run:

```
# doc0 everything lower case
doc0 =tolower(doco)

# doc1 replaces page breaks
doc1 <-str_replace_all(doc0,"\\f"," ")

# doc2 eliminates isolated characters of length 1
doc2 <- keep(.x = doc1, .p = function(x){str_length(x) > 2})

# everything is saved in doc
doc <-doc2
head(doc)
```

```
## [1] "   my first objective is for" "our country to be a"
## [3] "pioneering and successful"    "global model of"
## [5] "excellence, on all fronts,"   "and i will work with you to"
```

First keyword: God

Locate the keyword in the text

Convert the keyword to lower case:

```
keyword1 <- tolower(params$keyword1)
keyword1
```

```
## [1] "god"
```

First we tokenize the text by word, excluding uninteresting words:

```
text_df <- tibble(text = doc)
text_tidy <-text_df %>%
      unnest_tokens(word, text) %>%
      anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
head(text_tidy)
```

```
## # A tibble: 6 x 1
##   word
##   <chr>
## 1 objective
## 2 country
## 3 pioneering
## 4 successful
## 5 global
## 6 model
```

Then, count the frequency of each word in the text:

```
word_frequency <-text_tidy %>%
  count(word, sort = TRUE)
#head(word_frequency)
```

Locate the keyword in the text

To avoid matches where the keyword is embedded in another word, we need the following regular expression:

```
pat1 <- paste0("\\b",keyword1,"\\b")
```

Which lines contains the keyword?

```
sentences1 <- text_df[str_detect(text_df$text,pat1),]
head(sentences1)
```

```
## # A tibble: 2 x 1
##   text
##   <chr>
## 1 sisters, is one of huge promise and great potential, god
## 2 we are confident that, god
```

How many times does the keyword appear?

```
times1 <- nrow(text_df[str_detect(text_df$text,pat1),])
times1
```

```
## [1] 2
```

Where are they located?

```
lines1 = which(str_detect(text_df$text,pat1))
lines1
```

```
## [1]  66 279
```

What page numbers in the original document?

```
pages1 = findInterval(which(str_detect(doc0,pat1)),page_breaks) + 3
pages1
```

```
## [1]  6 16
```

Highlight the keyword with a color

This function color the keyword in the sentence fragment where it appears. It uses the function `colFm` specified at the beginning.

```
color_key <-function(sentence,key,debug=FALSE) {
  pat = paste0("\\b",key,"\\b")
  nk <- str_count(sentence,pat)
  if (debug) cat("nk= ",nk,"\n")
  if (nk > 0) {
    #if (debug) print(sentence, "\n")
    index <-str_locate(sentence,pat)
    if (debug) cat(index, "\n")
    p1 <- str_sub(sentence,1,index[1]-1)
    p2 <- str_sub(sentence,index)
    p3 <- str_sub(sentence,index[2]+1,nchar(sentence))
    if (debug) cat("p1= ",p1," p2= ",p2, " p3= ",p3,"\n")
    pall <-paste0(p1,colFmt(p2,'blue'),p3,collapse=" ")
```

```
  } else sentence
}
```

```
get_sentence <-function(charvec,location,key,span=5,debug=FALSE) {
  # charvec: vector of char strings
  # location: index to keyword location in charvec
  # ley: keyword
  # span: number of lines to go above and below the current line to get a
  #       sentence

  # pattern to look for keyword
  pat = paste0("\\b",key,"\\b")
  # Go location +- span lines trying to find a whole sentence
  begin=location-span
  end = location+span
  if (debug) cat("begin,end: ", begin,end,"\n")

  # paste all lines in just one string:
  together = paste(charvec[begin:end,]$text,collapse=" ")
  if (debug) cat("together=", together,"\n")

  # make sentences out of this string looking for "."
  sentences =strsplit(together, "\\.")
  #sentences =strsplit(together, "\\.?!\\.")
  if (debug) { cat("sentences: \n")
    print(sentences)}

  # index is the index of the sentence given by location
  index <-str_which(sentences[[1]],charvec[location,]$text)

  # if index has length 0, it means that the fragment is between two sentences
  # and we go back to just finding the keyword.
  # However, the keyword may appear more than once in the selected sentences.
  # Let's pick the first

  if (length(index)==0) index <- min(str_which(sentences[[1]],pat))

  # Color the keyword in the sentence
  sent.aprox <-color_key(sentences[[1]][index],key,debug=FALSE)

  # i is the index of the sentence where the key is located
  #nk <- sum(str_detect(sentences[[1]],pat))
  #pall<-numeric(nk)
  #for (j in 1:nk) {
  #  sev <- str_which(sentences[[1]],pat)
  #  pall[j]<-color_key(sentences[[1]][sev[j]],pat)
  #  if (debug) {cat("i= \n")
  #    print(i) }
  #}
  sent.aprox
}
```

Results

Produces a table with the page and sentence with the ocurrences of the keyword:

```
context1 <-data.frame(p=numeric(times1),location=lines1,line=numeric(times1))

 for (i in 1:times1) {
   context1$p[i] <- pages1[i]
   context1$line[i] <- color_key(sentences1[i,],keyword1)
   #context1$line[i] <- get_sentence(text_df,lines1[i],keyword1)
 }

 if (is.null(outputFormat)) context1 else
   print(xtable(context1,auto=TRUE),comment=FALSE,
         sanitize.text.function = identity,type=outputFormat)
```

	p	location	line
1	6	66	sisters, is one of huge promise and great potential, god
2	16	279	we are confident that, god

Second keyword: Holy

Repeat the previous steps with the second keyword:

```
keyword2 <- tolower(params$keyword2)
pat2 <- paste0("\\b",keyword2,"\\b")
sentences2 <- text_df[str_detect(text_df$text,pat2),]
times2 <- nrow(text_df[str_detect(text_df$text,pat2),])
lines2 = which(str_detect(text_df$text,pat2))
pages2 = findInterval(which(str_detect(doc0,pat2)),page_breaks) + 3

context2 <-data.frame(p=numeric(times2),location=lines2,line=numeric(times2))

 for (i in 1:times2) {
   #print(i)
   context2$p[i] <- pages2[i]
   context2$line[i] <- color_key(sentences2[i,],keyword2)
   #context2$line[i] <- get_sentence(text_df,lines2[i],keyword2)
 }
```

Print the results:

```
 if (is.null(outputFormat)) context2 else
   print(xtable(context2,auto=TRUE),comment=FALSE,
         sanitize.text.function = identity,type=outputFormat)
```

Third keyword: Nation

Repeat the previous steps with the second keyword:

```
keyword3 <- tolower(params$keyword3)
pat3 <- paste0("\\b",keyword3,"\\b")
sentences3 <- text_df[str_detect(text_df$text,pat3),]
```

	p	location	line
1	4	9	custodian of the two holy mosques
2	5	35	holy mosques, the most sacred sites on earth, and the
3	6	70	that muslims from around the world can visit the holy
4	6	95	two holy mosques, king salman bin abdulaziz al-saud,
5	15	265	we have been given the privilege to serve the two holy
6	15	266	mosques, the pilgrims and all visitors to the blessed holy
7	15	274	the two holy mosques, as well as modernizing and
8	15	277	train projects that will serve visitors to the holy mosques
9	16	291	and holy sites. we have reinforced the network of our
10	20	340	our expansion of the two holy mosques has led to a

```
times3 <- nrow(text_df[str_detect(text_df$text,pat3),])
lines3 = which(str_detect(text_df$text,pat3))
pages3 = findInterval(which(str_detect(doc0,pat3)),page_breaks) + 3

context3 <-data.frame(p=numeric(times3),location=lines3,line=numeric(times3))

 for (i in 1:times3) {
   context3$p[i] <- pages3[i]
   context3$line[i] <- color_key(sentences3[i,],keyword3)
   #context3$line[i] <- get_sentence(text_df,lines3[i],keyword3)
 }
```

Print the results:

```
 if (is.null(outputFormat)) context3 else
   print(xtable(context3,auto=TRUE),comment=FALSE,
         sanitize.text.function = identity,type=outputFormat)
```

	p	location	line
1	5	39	become a global investment powerhouse. our nation
2	5	54	younger generation. they are our nation's pride and the
3	6	56	tougher circumstances than today, our nation was
4	6	61	the blessings allah has bestowed on our nation, we
5	6	106	therefore, we will not rest until our nation is a leader in
6	6	141	position as a great nation in which we should all feel an
7	10	165	an ambitious nation
8	10	166	an ambitious nation.. effectively governed
9	10	167	an ambitious nation.. responsibly enabled
10	12	188	society, a thriving economy and an ambitious nation.
11	12	211	our nation is ambitious in what we want to achieve. we
12	15	240	what makes our nation
13	15	243	nation is the core of the arab
14	28	611	diabetes and cancer that threaten our nation's health.
15	61	1332	nation
16	63	1334	an ambitious nation..
17	65	1450	an ambitious nation..
18	71	1552	an ambitious nation..
19	71	1554	the nation we aspire to build will
20	71	1571	excellence for our nation, our society, our families
21	73	1641	an ambitious nation..

All keywords together

Group the results of the different keywords and sort it for order of apperance in the text

```
context <- bind_rows(context1,context2,context3)
context <-arrange(context,location)
```

```
if (is.null(outputFormat)) context else
  print(xtable(context,auto=TRUE),comment=FALSE,
        sanitize.text.function = identity,type=outputFormat)
```

	p	location	line
1	4	9	custodian of the two holy mosques
2	5	35	holy mosques, the most sacred sites on earth, and the
3	5	39	become a global investment powerhouse. our nation
4	5	54	younger generation. they are our nation's pride and the
5	6	56	tougher circumstances than today, our nation was
6	6	61	the blessings allah has bestowed on our nation, we
7	6	66	sisters, is one of huge promise and great potential, god
8	6	70	that muslims from around the world can visit the holy
9	6	95	two holy mosques, king salman bin abdulaziz al-saud,
10	6	106	therefore, we will not rest until our nation is a leader in
11	6	141	position as a great nation in which we should all feel an
12	10	165	an ambitious nation
13	10	166	an ambitious nation.. effectively governed
14	10	167	an ambitious nation.. responsibly enabled
15	12	188	society, a thriving economy and an ambitious nation.
16	12	211	our nation is ambitious in what we want to achieve. we
17	15	240	what makes our nation
18	15	243	nation is the core of the arab
19	15	265	we have been given the privilege to serve the two holy
20	15	266	mosques, the pilgrims and all visitors to the blessed holy
21	15	274	the two holy mosques, as well as modernizing and
22	15	277	train projects that will serve visitors to the holy mosques
23	16	279	we are confident that, god
24	16	291	and holy sites. we have reinforced the network of our
25	20	340	our expansion of the two holy mosques has led to a
26	28	611	diabetes and cancer that threaten our nation's health.
27	61	1332	nation
28	63	1334	an ambitious nation..
29	65	1450	an ambitious nation..
30	71	1552	an ambitious nation..
31	71	1554	the nation we aspire to build will
32	71	1571	excellence for our nation, our society, our families
33	73	1641	an ambitious nation..