The Colombian Signed Peace Agreement: A Text Mining Analysis of its Comprehension Difficulty

Juan C. Correa, María del Pilar García-Chitiva and Gustavo R. García-Vargas

Abstract

Citation: Juan C. Correa, María del Pilar García-Chitiva and Gustavo R. García-Vargas The Colombian Signed Peace Agreement: A Text Mining Analysis of its Comprehension Difficulty. **protocols.io**

dx.doi.org/10.17504/protocols.io.h8db9s6

Published: 31 May 2017

Protocol

Querying Latin American Intra-State armed conflicts

Step 1.

Let's see what peace agreements have occurred in Latin America as a result of Intra-State armed conflicts. This query can be accessed here. We immediately see that the query shows a total of 104 documents. With the exception of Canada, 102 manuscripts correspond to nine Latin American countries.

Download the Colombian Signed Peace Agreement

Step 2.

Since the Colombian peace agreement was signed on August 25, 2016, let's download this version here. If you have problems in downloading this version, click here

What other peace agreements can be compared with the Colombian one?

Step 3.

Among these nations, only three of them have gone through internal armed conflicts involving fights between guerrilla movements and local governments and have been solved with signed agreements by both parties. These countries are Colombia (1964-2016), Guatemala (1960-1996) and El Salvador (1980-1994). This shows us the relevance of comparing these agreements because of their socio-cultural similarities.

Download the Guatemalan peace agreement

Step 4.

These documents can be downloaded here and here.

Download the peace agreement of El Salvador

Step 5.

These documents can be downloaded in the following web pages

1. Chapultepec agreements

2. New York acts
3. The Mexico Agreement
4. The Geneva agreement

Preprocessing the documents

Step 6.

The preprocessing of these documents involved their conversion from pdf files to UTF-8 text-formatted data without headers and footers since these items present no semantic relevant information. Additional non-semantic features like badges, scanned signatures, vignettes, tables and the like were all deleted. In addition, we removed all numbers, unnecessary whitespaces, punctuation and special characters as well as Spanish stopwords.

These preprocessed documents can be downloaded here. Note that we have both English and Spanish versions of these documents.

If you are using Windows and these documents were downloaded in your "downloads folder", then your files locations should have the following address

C:://Downloads

Running the analysis (with an R script)

Step 7.

```
# The aim of this Script is to provide the reader an easy-to-use
# procedure that allows the replication of our findings
# This script works for the original accords which were written in Spanish.

###################################
### INITIALIZING THE ANALYSIS ###
###################################
if(!require(tm)){
install.packages('tm')
}
# Load the required packages for Preprocesing our documents
library(tm)


# Create an object that defines the local directory in
# which you saved the documents of the Colombian Signed Peace Agreement

#######################
###### COLOMBIA ######
#######################



#replace the address where you downloaded the documents
documents <- file.path('/home/lenovo/investigacion/Documents/Colombia (es)')
dir(documents)
```

```
# Create another object as a corpus containing all parts of the Colombian
# Agreement
ColombianDocs <- Corpus(DirSource(documents))
# You can check the contents of this last object like this
summary(ColombianDocs)
#Now lets begin with the preprocessing
#Remove punctuation
ColombianDocs <- tm_map(ColombianDocs, removePunctuation)
#Remove Numbers
ColombianDocs <- tm_map(ColombianDocs, removeNumbers)
#Convert all words in lowercase
ColombianDocs <- tm_map(ColombianDocs, tolower)
#Remove Stopwords
ColombianDocs <- tm_map(ColombianDocs, removeWords, stopwords('spanish'))
#Remove additional white spaces between words
ColombianDocs <- tm_map(ColombianDocs, stripWhitespace)
#Save The Colombian Signed Peace Agreement as Plain Document
ColombianDocs <- tm_map(ColombianDocs, PlainTextDocument)
#Begin the Analysis by creating a Document-Term Matrix
dtm <- DocumentTermMatrix(ColombianDocs)
#You can check the dtm by typing
dtm
# You can transpose this matrix, If you want
tdm <- TermDocumentMatrix(ColombianDocs)
# Let's explore terms frequencies
freq <- colSums(as.matrix(dtm))
# Now let's order these terms from least frequents to most frequents ones. This 'ord' object
# will be used later when computing wordcloud.
ord <- order(freq)
# Let's remove sparse terms from this Term-Document Matrix with, say 10% empty space
NSTDM <- removeSparseTerms(tdm, 0.1)
# Let's see most and least frequent words in all documents
freq[tail(ord)]
freq[head(ord)]
# Do you want to export this info as a table in Excel?
FreqTerms <- as.matrix(tdm)
dim(FreqTerms)
write.csv(FreqTerms, file = 'ColombianFrequentTerms.csv')
# Note that terms are in rows, and columns represent each document
# Let's open this new matrix
if(!require(readr)){
install.packages('readr')
}
library(readr)
ColombianFrequentTerms <- read_csv('/ColombianFrequentTerms.csv')
names(ColombianFrequentTerms) <-c('Words', 'Addendum', 'Point 1', 'Point 2', 'Point 3', 'Point 4',
'Point 5', 'Point 6')
ColombianFrequentTerms$WordTotalFrequency <- rowSums(ColombianFrequentTerms[,2:8])
# Let's take a look to the Wordcloud of the Colombian Agreement
```

```
# that includes most frequent words (those that appear at least
# 25 times in the whole document)
if(!require(wordcloud)){
install.packages('wordcloud')
}
library(wordcloud)
set.seed(142)
wordcloud(names(freq), freq, min.freq=25)

if(!require(koRpus)){
install.packages('koRpus')
}
# Now, let's tokenize our documents
library(koRpus)
Addendum <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/protocolos.txt', lang =
'es')
Point1 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto1.txt', lang = 'es')
Point2 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto2.txt', lang = 'es')
Point3 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto3.txt', lang = 'es')
Point4 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto4.txt', lang = 'es')
Point5 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto5.txt', lang = 'es')
Point6 <- tokenize('/home/lenovo/investigacion/Documents/Colombia (es)/Punto6.txt', lang = 'es')
# Let's calculate the SMOG grading for all these documents
SMOG(Addendum)
SMOG(Point1)
SMOG(Point2)
SMOG(Point3)
SMOG(Point4)
SMOG(Point5)
SMOG(Point6)
# Since the SMOG algorithm provides a double output (i.e., the grading instruction and the age of a
person),
# we can build a dataset that contains these values for all the documents of Colombia, Guatemala
and El Salvador.
# To do that, we need to repeat the same procedures for the other countries.

#######################
###### SALVADOR ######
#######################
library(tm)
Salvador <- file.path('/home/lenovo/investigacion/Documents/Salvador')
dir(Salvador)
SalvadorDocs <- Corpus(DirSource(Salvador))
# You can check the contents of this last object like this
summary(SalvadorDocs)
#Now lets begin with the preprocessing
#Remove punctuation
SalvadorDocs <- tm_map(SalvadorDocs, removePunctuation)
#Remove Numbers
```

```
SalvadorDocs <- tm_map(SalvadorDocs, removeNumbers)
#Convert all words in lowercase
SalvadorDocs <- tm_map(SalvadorDocs, tolower)
#Remove Stopwords
SalvadorDocs <- tm_map(SalvadorDocs, removeWords, stopwords('spanish'))
#Remove additional white spaces between words
SalvadorDocs <- tm_map(SalvadorDocs, stripWhitespace)
#Save The Colombian Signed Peace Agreement as Plain Document
SalvadorDocs <- tm_map(SalvadorDocs, PlainTextDocument)
#Begin the Analysis by creating a Document-Term Matrix
dtmSalvador <- DocumentTermMatrix(SalvadorDocs)
#You can check the dtm by typing
dtmSalvador
# You can transpose this matrix, If you want
tdmSalvador <- TermDocumentMatrix(SalvadorDocs)
# Let's explore terms frequencies
freqSalvador <- colSums(as.matrix(dtmSalvador))
# Now let's order these terms from least frequents to most frequents ones. This 'ord' object
# will be used later when computing wordcloud.
ordSalvador <- order(freqSalvador)
# Let's remove sparse terms from this Term-Document Matrix with, say 10% empty space
NSTDMsalvador <- removeSparseTerms(tdmSalvador, 0.1)
# Let's see most and least frequent words in all documents
freqSalvador[tail(ordSalvador)]
freqSalvador[head(ordSalvador)]
# Do you want to export this info as a table in Excel?
FreqTermsSalvador <- as.matrix(tdmSalvador)
dim(FreqTermsSalvador)
write.csv(FreqTermsSalvador, file = 'FrequentTermsSalvador.csv')
# Note that terms are in rows, and columns represent each document
# Let's open this new matrix
library(readr)
FrequentTermsSalvador <- read_csv('/FrequentTermsSalvador.csv')
names(FrequentTermsSalvador) <-c('Words', 'Act', 'Cap1', 'Cap2', 'Cap3', 'Cap4', 'Cap5', 'Cap6',
'Cap7', 'Cap8', 'Cap9', 'DecFin', 'Intro')
FrequentTermsSalvador$WordTotalFrequency <- rowSums(FrequentTermsSalvador[,2:13])
# Let's take a look to the Wordcloud of the Salvadoran Agreement
# that includes most frequent words (those that appear at least
# 25 times in the whole document)
library(wordcloud)
set.seed(142)
wordcloud(names(freqSalvador), freqSalvador, min.freq=25)

# Now, let's tokenize our documents
library(koRpus)
Acts <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Actas.txt', lang ='es')
Cap1 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap1.txt', lang = 'es')
Cap2 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap2.txt', lang = 'es')
Cap3 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap3.txt', lang = 'es')
```

```
Cap4 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap4.txt', lang = 'es')
Cap5 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap5.txt', lang = 'es')
Cap6 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap6.txt', lang = 'es')
Cap7 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap7.txt', lang = 'es')
Cap8 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap8.txt', lang = 'es')
Cap9 <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Cap9.txt', lang = 'es')
DecFin <- tokenize('/home/lenovo/investigacion/Documents/Salvador/DecFin.txt', lang = 'es')
Intro <- tokenize('/home/lenovo/investigacion/Documents/Salvador/Intro.txt', lang = 'es')
# Let's calculate the SMOG grading for all these documents
SMOG(Acts)
SMOG(Cap1)
SMOG(Cap2)
SMOG(Cap3)
SMOG(Cap4)
SMOG(Cap5)
SMOG(Cap6)
SMOG(Cap7)
SMOG(Cap8)
SMOG(Cap9)
SMOG(DecFin)
SMOG(Intro)

########################
###### GUATEMALA ######
########################
library(tm)
Guatemala <- file.path('/home/lenovo/investigacion/Documents/Guatemala')
dir(Guatemala)
GuatemalaDocs <- Corpus(DirSource(Guatemala))
summary(GuatemalaDocs)
#Now lets begin with the preprocessing
#Remove punctuation
GuatemalaDocs <- tm_map(GuatemalaDocs, removePunctuation)
#Remove Numbers
GuatemalaDocs <- tm_map(GuatemalaDocs, removeNumbers)
#Convert all words in lowercase
GuatemalaDocs <- tm_map(GuatemalaDocs, tolower)
#Remove Stopwords
GuatemalaDocs <- tm_map(GuatemalaDocs, removeWords, stopwords('spanish'))
#Remove additional white spaces between words
GuatemalaDocs <- tm_map(GuatemalaDocs, stripWhitespace)
#Save The Colombian Signed Peace Agreement as Plain Document
GuatemalaDocs <- tm_map(GuatemalaDocs, PlainTextDocument)
#Begin the Analysis by creating a Document-Term Matrix
dtmGuatemala <- DocumentTermMatrix(GuatemalaDocs)
#You can check the dtm by typing
dtmGuatemala
# You can transpose this matrix, If you want
tdmGuatemala <- TermDocumentMatrix(GuatemalaDocs)
```

```
# Let's explore terms frequencies
freqGuatemala <- colSums(as.matrix(dtmGuatemala))
# Now let's order these terms from least frequents to most frequents ones. This 'ord' object
# will be used later when computing wordcloud.
ordGuatemala <- order(freqGuatemala)
# Let's remove sparse terms from this Term-Document Matrix with, say 10% empty space
NSTDMguatemala <- removeSparseTerms(tdmGuatemala, 0.1)
# Let's see most and least frequent words in all documents
freqGuatemala[tail(ordGuatemala)]
freqGuatemala[head(ordGuatemala)]
# Do you want to export this info as a table in Excel?
FreqTermsGuatemala <- as.matrix(tdmGuatemala)
dim(FreqTermsGuatemala)
write.csv(FreqTermsGuatemala, file = 'FrequentTermsGuatemala.csv')
# Note that terms are in rows, and columns represent each document
# Let's open this new matrix
library(readr)
FrequentTermsGuatemala <- read_csv('/FrequentTermsGuatemala.csv')
names(FrequentTermsGuatemala) <-c('Words', 'Caps1', 'Caps2', 'Caps3', 'Caps4', 'Caps5', 'Caps6',
'Caps7', 'Caps8', 'Caps9', 'Caps10', 'Caps11', 'Caps12', 'ProtGua')
FrequentTermsGuatemala$WordTotalFrequency <- rowSums(FrequentTermsGuatemala[,2:14])
# Let's take a look to the Wordcloud of the Guatemalian Agreement
# that includes most frequent words (those that appear at least
# 25 times in the whole document)
library(wordcloud)
set.seed(142)
wordcloud(names(freqGuatemala), freqGuatemala, min.freq=25)

# Now, let's tokenize our documents
library(koRpus)
Caps1 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps1.txt', lang = 'es')
Caps2 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps2.txt', lang = 'es')
Caps3 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps3.txt', lang = 'es')
Caps4 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps4.txt', lang = 'es')
Caps5 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps5.txt', lang = 'es')
Caps6 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps6.txt', lang = 'es')
Caps7 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps7.txt', lang = 'es')
Caps8 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps8.txt', lang = 'es')
Caps9 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps9.txt', lang = 'es')
Caps10 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps10.txt', lang = 'es')
Caps11 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps11.txt', lang = 'es')
Caps12 <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/Caps12.txt', lang = 'es')
ProtGua <- tokenize('/home/lenovo/investigacion/Documents/Guatemala/ProtocoloGuatemala.txt',
lang = 'es')
# Let's calculate the SMOG grading for all these documents
SMOG(Caps1)
SMOG(Caps2)
SMOG(Caps3)
SMOG(Caps4)
```

```
SMOG(Caps5)
SMOG(Caps6)
SMOG(Caps7)
SMOG(Caps8)
SMOG(Caps9)
SMOG(Caps10)
SMOG(Caps11)
SMOG(Caps12)
SMOG(ProtGua)

##############################################
### ANALYZING ALL PEACE AGREEMENTS AT ONCE #####
##############################################

# We can build a new dataset with the values resulting from the SMOG analyses that we
# just applied to all documents
Country <- c('Colombia', 'Colombia', 'Colombia', 'Colombia', 'Colombia', 'Colombia', 'Colombia',
'Salvador', 'Salvador', 'Salvador', 'Salvador', 'Salvador', 'Salvador', 'Salvador', 'Salvador', 'Salvador',
'Salvador', 'Salvador', 'Salvador', 'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala',
'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala', 'Guatemala',
'Guatemala')
TextID <- c('Addendum', 'Point1', 'Point2', 'Point3', 'Point4', 'Point5', 'Point6', 'Acts', 'Cap1', 'Cap2',
'Cap3', 'Cap4', 'Cap5', 'Cap6', 'Cap7', 'Cap8', 'Cap9', 'DecFin', 'Intro', 'Caps1', 'Caps2', 'Caps3', 'Caps4',
'Caps5', 'Caps6', 'Caps7', 'Caps8', 'Caps9', 'Caps10', 'Caps11', 'Caps12', 'ProtGua')
Smog <- c(18.71, 19.19, 21.05, 19.17, 21.1, 21.11, 18.94, 14.8, 18.84, 19.6, 24.85, 21.64, 21.51,
15.9, 17.3, 18.51, 10.14, 14.96, 19.85, 19.29, 18.58, 18.25, 19.29, 20.15, 20.79, 19.71, 17.02, 20.33,
19.16, 18.32, 18.79, 17.11)
Age <- c(23.71, 24.19, 26.05, 24.17, 26.1, 26.11, 23.94, 19.8, 23.84, 24.6, 29.85, 26.64, 26.51, 20.9,
22.3, 23.51, 15.14, 19.96, 24.85, 24.29, 23.58, 23.25, 24.29, 25.15, 25.79, 24.71, 22.02, 25.33, 24.16,
23.32, 23.79, 22.11)
PeaceReadability <- data.frame(Country, TextID, Smog, Age)
# Let's export this dataset to an Excel file
write.csv(PeaceReadability, file = 'PeaceReadability.csv')
# Let's take a look at the resulting SMOG grading
library(psych)
describe.by(PeaceReadability$Smog, group = PeaceReadability$country)
describe.by(PeaceReadability$Age, group = PeaceReadability$country)
# Now let's take a look at the total of polysyllables.
library(qdap)
library(tibble)
ColombianPolysyllables <- as_tibble(combo_syllable_sum(ColombianFrequentTerms$Words))
GuatemalanPolysyllables <- as_tibble(combo_syllable_sum(FrequentTermsGuatemala$Words))
SalvadoranPolysyllables <- as_tibble(combo_syllable_sum(FrequentTermsSalvador$Words))
describe(ColombianPolysyllables$polysyllable.count)
describe(GuatemalanPolysyllables$polysyllable.count)
describe(SalvadoranPolysyllables$polysyllable.count)
library(ggplot2)
ggplot(PeaceReadability, aes(Smog, fill = Country, colour = Country)) + geom_density(alpha = 0.1)
# What about the age that the person must be to fully understand these text?
```

```
ggplot(PeaceReadability, aes(Age, fill = Country, colour = Country)) + geom_density(alpha = 0.1)
# Do these graphs indicate significant statistical differences between the accords? Let's
# test it
fit <- aov(Smog Country, data=PeaceReadability)
summary(fit)
# The results show that the differences among these peace accords proved to be not significant.
#
# Now, let's analyze all documents in just one 'big' corpus
LatinAmericans <- file.path('/home/lenovo/investigacion/Documents/LatinAmericanAgreements')
dir(LatinAmericans)
library(tm)
# Create another object as a corpus containing all parts of the Colombian
# Agreement
LatinAmericans <- Corpus(DirSource(LatinAmericans))
# You can check the contents of this last object like this
summary(LatinAmericans)
LatinDocs <- summary(LatinAmericans)
#Now lets begin with the preprocessing
#Remove punctuation
LatinAmericans <- tm_map(LatinAmericans, removePunctuation)
#Remove Numbers
LatinAmericans <- tm_map(LatinAmericans, removeNumbers)
#Convert all words in lowercase
LatinAmericans <- tm_map(LatinAmericans, tolower)
#Remove Stopwords
LatinAmericans <- tm_map(LatinAmericans, removeWords, stopwords('spanish'))
#Remove additional white spaces between words
LatinAmericans <- tm_map(LatinAmericans, stripWhitespace)
#Save The Colombian Signed Peace Agreement as Plain Document
LatinAmericans <- tm_map(LatinAmericans, PlainTextDocument)
# Let's apply stemming procedure for retrieving suffices of words
LatinAmericans <- tm_map(LatinAmericans, stemDocument, language = 'spanish')
#Begin the Analysis by creating a Document-Term Matrix
dtmLatin <- DocumentTermMatrix(LatinAmericans)
#You can check the dtm by typing
dtmLatin
# You can transpose this matrix, If you want
tdmLatin <- TermDocumentMatrix(LatinAmericans)
# Let's explore terms frequencies
freqLatin <- colSums(as.matrix(dtmLatin))
# Now let's order these terms from least frequents to most frequents ones. This 'ord' object
# will be used later when computing wordcloud.
ordLatin <- order(freqLatin)
# Let's remove sparse terms from this Term-Document Matrix with, say 10% empty space
NSTDMLatin <- removeSparseTerms(tdmLatin, 0.1)
# Let's see most and least frequent words in all documents
freqLatin[tail(ordLatin)]
freqLatin[head(ordLatin)]
# Do you want to export this info as a table in Excel?
```

```
FreqTermsLatin <- as.matrix(tdmLatin)
dim(FreqTermsLatin)
write.csv(FreqTermsLatin, file = 'LatinFrequentTerms.csv')
# Note that terms are in rows, and columns represent each document
# Let's open this new matrix
library(readr)
LatinFrequentTerms <- read_csv('/LatinFrequentTerms.csv')
names(LatinFrequentTerms) <-c('Words', 'col1', 'col2', 'col3', 'col4', 'col5', 'col6', 'col7', 'gua1', 'gua2',
'gua3', 'gua4', 'gua5', 'gua6', 'gua7', 'gua8', 'gua9', 'gua10', 'gua11', 'gua12', 'gua13', 'sal1', 'sal2',
'sal3', 'sal4', 'sal5', 'sal6', 'sal7', 'sal8', 'sal9', 'sal10', 'sal11', 'sal12')
LatinFrequentTerms$WordTotalFrequency <- rowSums(LatinFrequentTerms[,2:8])
# Let's take a look to the Wordcloud of the Colombian Agreement
# that includes most frequent words (those that appear at least
# 25 times in the whole document)
library(wordcloud)
set.seed(142)
wordcloud(names(freqLatin), freqLatin, min.freq=50)
# Let's plot the most common frequent words of all three agreements
library(ggplot2)
CommonFrequentTerms <- subset(LatinFrequentTerms, WordTotalFrequency>180)
ggplot(CommonFrequentTerms,aes(x= reorder(Words,-WordTotalFrequency),WordTotalFrequency)) +
geom_bar(stat ='identity') + ylab('Frequency') + xlab('Words') + theme(axis.text.x =
element_text(angle = 90, hjust = 1))
```