No. 4
FEBRERO de 2017

quantil

Documentos de Trabajo

Edición electrónica

Improving transparency of the Colombian Peace Treaty with NLP

Francisco Barreras
Mónica Ribero
Felipe Suárez

quantil | matemáticas aplicadas

Serie Documentos de Trabajo Quantil, 2017-4
Edición electrónica.

FEBRERO de 2017

© 2017, Quantil S.A.S., Estudios Económicos,
Carrera 7 # 77 - 07. Oficina 901, Bogotá, D. C., Colombia
Teléfonos: +57(1) 805 1814
E-mail: info@quantil.com.co
http://www.quantil.com.co

Impreso en Colombia – Printed in Colombia

Improving transparency of the Colombian Peace Treaty with NLP

A Tool for understanding, navigating and summarizing the Colombian Peace Treaty*

Francisco Barreras
Quantil
Bogotá, Colombia
francisco.barreras@quantil.com.co

Mónica Ribero
Quantil
Bogotá, Colombia
monica.ribero@quantil.com.co

Felipe Suárez
Quantil
Bogotá, Colombia
felipe.suarez@quantil.com.co

February 20, 2017

ABSTRACT

Factorization methods and probabilistic models provide useful ways to represent text that can capture properties like sentence relevance, topics in text and even semantic similarity. In general, methods that yield a low-dimensional representation of large volumes of text have become more important and have gained attention in a diversity of fields since they have the potential of assisting in the understanding of vast ammounts of information. The Colombian Peace treaty documented an exhaustive list of agreements between the FARC guerilla and Colombian Armed Forces across six main sections spanning 297 pages. Surprisingly, the final version of the treaty was made public only for 40 days before Colombians had to vote for its approval in a plebiscite. Given that most of the general population probably would not read the document and the political environment (including the media) was highly polarized, there was a growing need for an unbiased and practical way to review the document before the vote. This paper describes the technical details behind the implementation of a tool that analyses the 2016 Colombian peace treaty. By combining Natural Language Processing techniques we were able to provide a web-service that helped increase transparency and unbiased reviewing of each section of the peace treaty.

CCS CONCEPTS

• **Information Systems** → **Data Mining**

KEYWORDS

GloVe, Latent Dirichlet Allocation, Natural Language Processing, NMF, Peace Treaty.

ACM Reference format:

Francisco Barreras, Mónica Ribero, and Felipe Suárez. 2017. Improving transparency of the Colombian Peace Treaty with NLP. In Proceedings of ACM Woodstock conference, Halifax, Nova Scotia - Canada, August 2017 (KDD'17), 6 pages.

1 Introduction

For more than 5 decades Colombia has witnessed a burdensome and cruel armed conflict with the FARC (Colombian Revolutionary Armed Forces), a communist guerrilla. This conflict has been worsened and prolonged by circumstances like drug dealing, international financial support for the guerrilla and challenging topography. In 2016, the two parties finally reached a succesful conclusion to peace negotiations after over 10 failed attempts in the past. The peace treay needed to be approved by the general population in a plebiscite on October the 2nd 2016, only 40 days after its publication.

Several failed attempts of agreement have resulted in disastrous consequences such as increased violence, forced displacements and the consolidation of drug trafficking [10]. Encouraged by the positive disposal of both sides to agree to participate of a peaceful war ending treaty, we propose a novel tool that enables citizens all across the country to extract and summarize the relevant information in the treaty in a free and open web service. We deploy multiple Natural Language Processing methodologies to aid end users capture the most relevant information about a specific topic of their interest within the treaty.

We have seen some applications that targeted similar goals for summarization, but mostly for practical applications in business and industry, this project was a chance to apply such techniques for social good and increased transparency. Since the early days of Natural Language Processing, starting with the works of Mani [6], attempts to extract the most general information of large corpora has improved substantially. Not surprisingly, several authors and newspapers released their own –human version– summaries claiming to be politically unbiased and exhaustive. Even if we trust their claims, drawing up a succinct text that encompasses the key information of a large corpus is a very time consuming task prone to personal biases. It also leads to static texts that may or may not satisfy everyone's personal preferences.

In this paper we use a plain text version of the original agreement[1] together with a large corpus consisting of articles from the press, twitter feeds, books of public access and legislative texts. In section 2 we detail the theoretical concepts that support the algorithmic implementations that we explain later in section 3. In section 4 we discuss the results of our implementations under multiple queries. Our conclusions of the work and possible future improvements are presented in section 5.

* The deployment of the tool is found in the website: www.acuerdosdepaz.co

[1] The PDF version can be found in http://www.acuerdodepaz.gov.co/acuerdos/acuerdo-final.

2 Theoretical Framework

We use several topic discovery techniques that vectorize words and documents as Latent Dirichlet Allocation and GloVe. These methods were used to filter the sentences in the treaty by semantic similarity to a user's query. A factorization method (NMF) was then applied to the resulting subset from which a relevance score was computed for each sentence and a chart was produced to visualize the proportion of different topics. The relevance score was further used to filter the resulting sentences and produce a summary of varying length, displaying sentences above a relevance threshold in order of appearance. We will now elaborate further on each step of the process.

2.1 Latent Dirichlet Allocation

Consider a collection of documents $\mathcal{D} = \{d_1, d_2, ..., d_n\}$ where each document d is a sequence of N_d words $d = (w_{d_1}, ..., w_{d_{N_d}})$ from a vocabulary $V = \{w_1, ..., w_N\}$. Latent Dirichlet Allocation is an NLP technique proposed by [2] to reveal latent dimensions in corpus $\mathcal{D}$, called topics, or in general on any collection of discrete data. In turn, each topic is a distribution over the words in corpus's vocabulary.

The main idea is that document d is generated by a distribution $\theta_d \in \Delta_{k-1} \sim \text{Dirichlet}(\alpha)$ of topics and each topic is a distribution of words $\phi_j \in \Delta_{N-1}$ for $j = 1, ..., k$ assuming the following procedure:

- Choose a topic assignment $z_{i,d} \sim \text{Multinomial}(\theta)$ for each word $i = 1, ..., N_d$ in document d
- Choose a word $w_k \sim \text{Multinomial}(\phi_{z_{i,d}})$

Consequently, topic j is discussed in document d with probability $\theta_{j,d}$ and each word i is discussed in topic j with probability $\phi_{i,j}$ for $i = 1, ..., w$ and $j = 1, ..., k$. These distributions are learned from data using unsupervised learning; we used the `lda` package in R that performs Collapsed Gibbs Sampling [3].

$$p(\mathcal{T}|\theta, \phi) = \prod_{d \in \mathcal{D}} \prod_{j=1}^{N_d} \phi_{j, z_{j,d}} \theta_{z_{j,d}, d} \tag{1}$$

These probabilities distributions can be, however, seen as vectorizations of words and documents. This will be relevant later when we present the way we matched user query's to relevant sentences.

2.2 Semantic word embeddings

GloVe is a word embedding model proposed by [7]. The model combines global matrix factorization and local context window methods to capture semantic and syntactic properties of words. However, the approach is different from LDA since it is based in word co-occurrence windows and not in the whole document.

GloVe is based on probabilities of word co-ocurrence. More specifically, let i, j, k three different words, then $\frac{p_{ik}}{p_{jk}}$ should be close to one if and only if both words i, and j coocur in the same proportion with word k. Otherwise, this ratio will be either close to zero or bigger than one. Based on these facts, [7] propose a least squares regression model with cost function

$$J(W) = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_k + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{2}$$

where

- V is the size of the vocabulary
- X is the co-occurrence word matrix
- f is a weighting function that allows to deal with rare or too frequent co-occurrences.
- $w_i, \tilde{w} \in \mathbb{R}^d$ are word and context word embeddings respectively.
- b and $\tilde{b}$ are bias terms.

2.3 Sentence Ranking

Many summarization techniques are based on a sentence ranking algorithm and then by chosing the top M sentences where M is a longitude parameter chosen by the *user*. Here, we used a Matrix factorization technique proposed by [5] to uncover dimensions of sentences representing semantic features that are later used to compute a relevance measure to rank sentences. The algorithm is as follows.

1. Construct the term - sentence matrix $A \in \mathbb{R}^{m \times n}$
2. The matrix is factorized to uncover k semantic features for each sentence $A = WH$ where $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$
3. The "Generic Relevance for Sentences" (GRS) is computed for each sentence where

$$GRS(j) = \sum_{i=1}^{k} H_{ij} weight(H_{i*})$$

$$weight(H_{i*}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{k} \sum_{q=1}^{n} H_{pq}}$$

4. Chose the M sentences with higher GRS.

Here the GRS computes how much relevant topics are discussed in each sentence. k is a parameter chosen by the modeler.

3 Implementation

In this section we discuss the approach we took in dealing with the two proposed objectives. The peace treaty is divided into six main sections regarding different aspects of the termination of the armed conflict together with several other sections of appendixes comprising 297 pages. The most relevant information is, thus, contained within the 6 main sections and we extract our corpus from it. We also gathered a large corpus consisting of articles from the press, twitter feeds, books of free access and legislative texts to extract the vectorial representation for words.

After properly preprocessing the texts in order to remove unwanted punctuation or non-ascii characters, unwanted numbering, hashtags, links and repeated spaces, we performed the following procedures:

1. Trained word representation using LDA and GloVe.
2. Vectorized the Peacy treaty by section.

3. Computed pretrained topics.

The first step of the data processing was training the vector representation of words. For LDA, we constructed a Document-Term matrix (DTM) from individual propositions as documents. Propositions were parsed by splitting paragraphs in separate sentences divided by period or semicolon. Additionally, we removed stopwords and swept the number of topics, k, in the range $\{30, 50, 100, 300\}$.

For GloVe representation, we constructed the Term Coocurrence Matrix (TCM), X, using windows of length k in the range $\{3, 5, 7, 9\}$. The corpus from which we trained these algorithms was sought to be spanish texts regarding topics about politics, war and narcotraffic; the complete corpus consisted of the peace treaty, 12 books, 9800 paragraphs from Constitutional Court sentences, 2000 paragraphs on FARC related news, and 50000 tweets.

Calibration of these parameters was done empirically by evaluating the coherence of the group of similar words for each word in a test set. Our test set consists of the words: *campo*, *conflicto*, *derechos*, *farc*, *internacional*, *justicia*, *lesa*, *militares*, *paramilitares*, *participación*, *patria*, *paz*, *verdad*, *víctimas*. Neighboring words are calculated using the cosine distance.

The second step of the processing was to vectorize each section of the treaty. This helped us relate each part of the text with a point in $\mathbb{R}^k$ and consequently be able to retrieve all related parts of the text by finding its cosine-neighborhood. Sentences were vectorized by averaging the vectors corresponding to each word in the sentence.

$$\phi(p) = \frac{1}{|p|} \sum_{w \in p} \phi(w). \tag{3}$$

For each possible section $s \in \{1, \ldots, 6\}$ we will call $\mathbf{W}_s$ to the vectorized texts within section s,

$$\mathbf{W}_s = \{\phi(p),\ p \in s\} \tag{4}$$

Following these processing procedures, we included two algorithms to retrieve specific queries based on particular preferences. There are two type of queries –topic and summary– to which we perform the algorithms **??** and **??**.

The last data processing procedure was to compute a list of pretrained topics. This enabled us to retrieve the most searched summaries and distributions instantly without needing to do any vectorization or neighborhood calculation. We will call predefined topics as $\mathbf{P}$ and the relevances to the section s as $\mathbf{r}_{l,s}$:

$$\mathbf{P} = \{\phi(q_1), \ldots, \phi(q_l)\},$$

$$\mathbf{r}_{i,s} = \textsc{Summarize}(q_i, s, \varepsilon, d).$$

Identification of semantic relation of an incoming query within $\mathbf{P}$ is done using cosine similarity. If,

$$\max_{\mathbf{p} \in \mathbf{P}} \cos(\phi(q), \mathbf{p}) > \eta,$$

then $\mathbf{r}_{i,s}$ is returned as summary of $\mathbf{q}$ instead of $\textsc{Summarize}(q, s, \varepsilon, d)$.

4 Use examples

Results of sample queries are displayed in detail in this section as well as illustrations of the two resulting vector embeddings implemented. We discuss the differences of each vectorization and suggest additional improvements that we could leverage for future work. We show the final tool released as a website publicly days after the publication of the treaty at the end of the section.

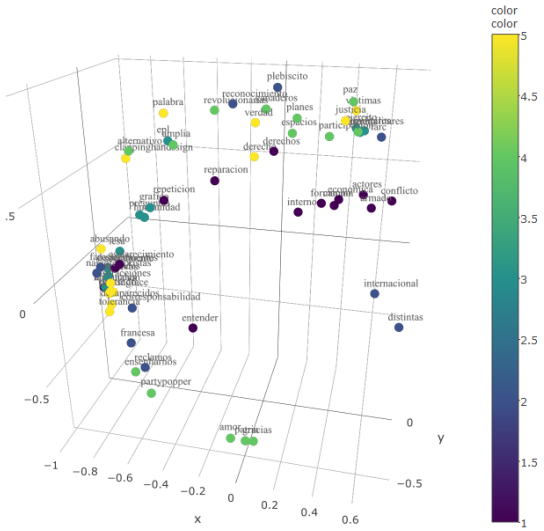

Figure 1: Illustration of LDA vectorization of words projected onto its three strongest components.

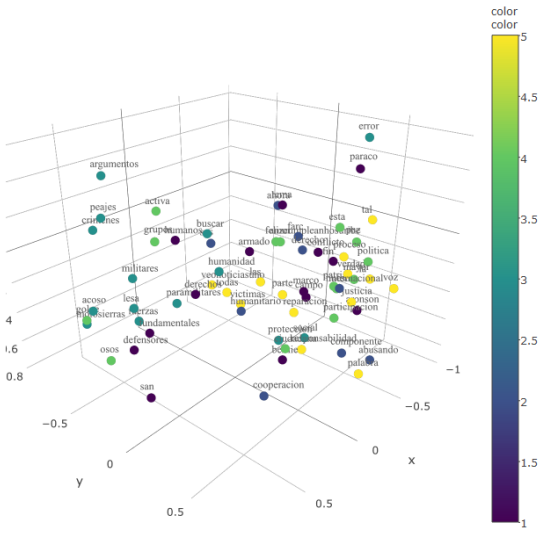

Figure 2: Illustration of GloVe vectorization of words projected onto its three strongest components.

Empirical assesment of the optimal parameters after sweeping the number of topics and window size for the Latent Dirichlet Allocation gave a vectorization that we display projected in a three dimensional afine transformation via Principal Component Analysis, PCA [4]. PCA helps us visualize the axes that maximze indivial variance and minimize cross correlation. We display the words closest to the test set word list in Figure 2. The colors of the points correspond to neighborhoods that we may want to reveal semantic clusters.

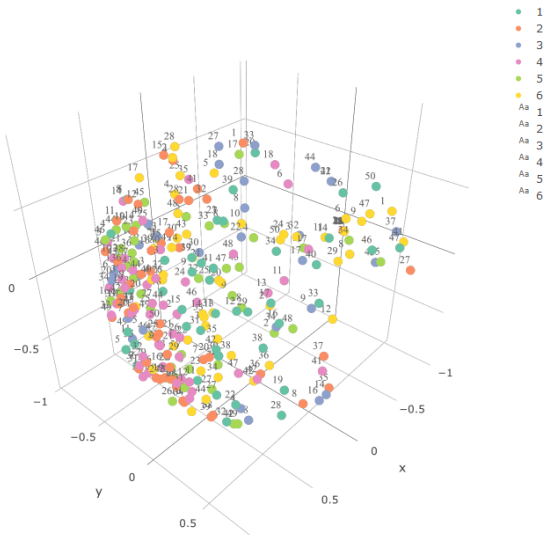

Figure 3: Illustration of the vectorized treaty by section projected onto its three strongest components.

Figure 4: Final interface.

For instance, in figure 3 we illustrate each part of the treaty after the vectorization transformation.

Figure 5: Final interface, wordcloud

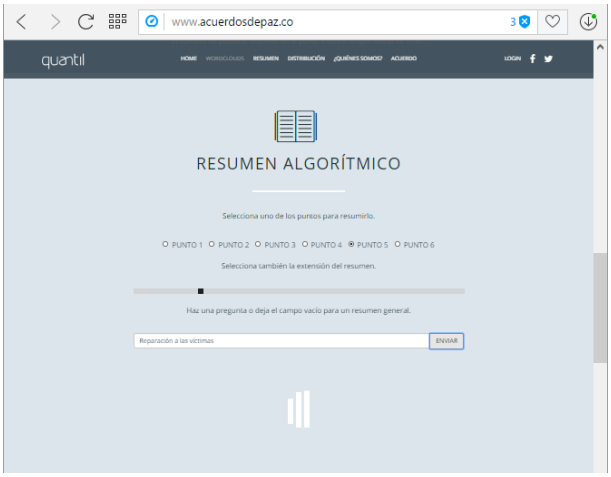

Figure 6: Final interface, summary.

Figure 7: Final interface, distribution.

All algorithms were implemented in R language [8]. Likewise, the final user tool consisted of the website www.acuerdosdepaz.co constructed with the DJango web framework [9].

5 Acknowledgments

We would like to thank Simón Ramírez, Carlos Cortés and Sebastián Terán for helping us develop the front end of the website. We would like to thank Diego Jara for the idea of using NLP techniques to facilitate the reading of the peace agreements. Also, for his comments and support that, along with those of Álvaro Riascos, helped to develop the tool as a user consumable good. Darío Correal pointed out valuable comments during the implentation stage for which we are thankful. Finally, we would like to thank Quantil for providing infrastructure and financial support.

6 Conclusions

We introduced a methodology that allowed citizens to explore the peace agreements between Colombian government and the FARC after a sixty years' conflict. It was a complex and unstructured legal text that people should review in a one month's period before deciding their vote for the plebiscite. Our aim was to ease the reading rather than replacing it; to achieve this we developed tools to contextualize people with principal factors, and to easily and quickly find sections concerning a particular topic and/or section of the documents.

Our methodology used different Natural Language Processing and Data Mining techniques that included GloVe, Latent Dirichlet Allocation and Non-Negative Matrix Factorization. It allowed users the opportunity to visualize and understand documents compositions by using topic distributions and to make automatic summaries.

The tool could generalize to explore other documents providing a friendly way to ensure people is informed. By using these tools in an ethical way one could avoid political interests and media to bias and interfere with people's decisions. Here we presented some examples of the results.

Future directions could include improvement in performance by introducing other classification techniques as Neural Networks. Other summarization techniques that improve coverage could also be explored; in the aforementioned methodology we used a NMF technique for summaries because its computational simplicity and efficiency. However, one could compare results with other techniques; for example, using Differential evolution algorithms to optimize coverage while minimizing redundancy have proven to be successful in some corpora [1].

Other directions that one could explore to enlarge analysis could include the possibility to compare other documents they provide and/or media documents to prove their accuracy. Also, to use Named Entity Recognition techniques to give precise responses to questions concerning amounts or specific actors.

We believe that tools like the one presented are important and should be developed in order to provide people a way to interact directly with documents they usually consult with other people or media that can have biased opinions and may omit information for personal interests. In other scenarios it would also be helpful to advice people that usually does not have access to legal assessment in different contexts.

References

[1] Rasim M Alguliyev, Ramiz M Aliguliyev, and Nijat R Isazade. An unsupervised approach to generating generic summaries of documents. *Applied Soft Computing*, 34:236–250, 2015.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Jonathan Chang. *lda: Collapsed Gibbs Sampling Methods for Topic Models*, 2015. R package version 1.4.2.

[4] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[5] Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34, 2009.

[6] Inderjeet Mani. *Automatic summarization*, volume 3. John Benjamins Publishing, 2001.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[9] R Foundation for Statistical Computing, Lawrence, Kansas. *Django [Computer Software]*, 2016.

[10] Juan Camilo Restrepo and MA Bernal. La cuestión agraria. *Bogotá: Colección Penguin*, 2014.