Team 28 DS4A

Text Analysis of Conpes Documents

Contents

1. Team 28 members

[redacted]

2. Executive Summary

A Conpes document is written in collaboration by multiple public entities of Colombia and published by the National Planning Department, a technical entity which supervises this documents and proposes public policy recommendations. These documents are important since they are a foundation for policy to happen with the intention of developing a better country. For example, in the year 2018, the Conpes 3920 was released, which was the first public policy document in Latin America regulating the application of Big Data services, and in the present year, the Conpes 3975 of Digital Transformation and Artificial Intelligence was also published.

For Colombian Government officials, who constantly search within Conpes documents, to support public policies design and propose new projects, Team 28 Dashboard is an all-purpose information provider for Conpes understanding and statistics based on NLP methods.

It is a product that offers a Conpes insights in both, time and topics, a tool that helps government officials to search for multiple words in all Conpes documents, and provides visualizations to explain the evolution of policies.

Although, the project revolves around public policy, the potential users are not restricted only to public officials. Researchers, NGOs, advocacy groups and the general public can benefit from it.

The document is structured in the following way. First, a description of the project purpose and its main users is presented. Second, the technical framework applied in order to solve the challenge. Third, the main results identified are described and the immediate next steps that can be implemented in order to extend the project's reach. Finally, the applications to other fields that the general structure of the project can have.

3. Project purpose

The main purpose of the project is to **support strategic decision making** within government regarding long term public policy design.

Other purposes of the project are:

- To **discover** insights from the overview of topics treated in Conpes documents

- Identify **patterns** of public policy topics within Conpes documents across time and presidential periods.
- Establish the **main topics** of interest in public policy over time

In this way, unattended topics, follow public policy compliance, focus resources, connect despair public policies, among others decisions, can be addressed.

4. Scoping

Dataset sourced:

CONPES documents (around 4000) may be found at:

https://www.dnp.gov.co/CONPES/Paginas/conpes.aspx

Check the link named “Listado en csv” this will download a csv file that contains all the URLs of each of the CONPES documents since 1967.

We think that the data process design has a broad spectrum of use in other fields where text processing is crucial, such as in criminal law, contract analysis, legislation and fraud detection among others.

Version 1:

MVP of an assortment of newer documents that are PDF readable from either economic or social planning and store its information as semi-structured data.

Version 2:

Analyze the semi-structured data and categorize it into a few major clusters without a visual clustering interface for starters, instead stored in an online database.

Version 3:

Add historic and typewritten documents via OCR and machine learning to the processed clusters.

Version 4:

Apply dimensionality reduction techniques to categorize and classify the documents and clusters into the relevant categories, such as innovation, criminal law etc.

Version 5:

Process and automatically analyze the datasets after processing the CONPES documents content (PDFs) in order to identify patterns to classify the main topics stated in CONPES documents providing a visual interface for a non-technical user to easily extract insights from it.

Version 6:

Providing advanced insights via links, filters, features such as zooming, coloring and labeling, graphic plots, correlation analysis and analysis reports in the visual interface.

Version 7:

Add another type of document called National Development Plan to compare against the CONPES documents.

5. Project plan

Week 1:

- Create git accounts and repository ([redacted]) - GOVERNMENT AI, create and share versioning policies
- Create final report document and be responsible for updating it after each iteration
- Download documents - Economic and social conpes

Week 2:

- Create master tables for all the Conpes and its information
- Read the documents

Week 3:

- Identify and research the most sensible way to read as many documents as possible [redacted])
- Local (OCR (Tesseract), or PDF to text)
- Cloud Services (Textract (Amazon), Google Cloud, Microsoft Azure, etc)
- Categorize the documents per quality (Images, text, tables, and so on)
- Exploratory Data Analysis based on the below steps

Week 4:

- Have the document corpus (Text mining)
- Text cleaning:
- Remove special symbols and punctuation symbols.
- Stop Words, numbers
- Try lemmatization (Use the dictionary: https://github.com/michmech/lemmatization-lists)
- Create bag of words (Word to vector)
- Set a weight to the words according to the context

Week 5:

- Start creating graphs to get statistical information from the dataframe
- Normalize the data

Week 6:

- Dimensionality Reduction - Use reduction techniques (Factorial with PCA)
- Apply clustering techniques

Week 7:

- Get advanced statistical information (Intercluster distances, etc)
- Store the cleaned and analyzed data in a database
- Answer the hypothesis

Week 8:

- FRONTEND work
- Connect backend with the frontend

Week 9:

- Testing
- Build insights and correlation reports from the data and its visualization
- Prepare the final presentation for the course

Week 10:

- Add further features to the visualization and graphical interface together with the links to the reports and further resources
- Correct mistakes and errors

6. Technical Framework

In this section, the technical framework implemented is presented in order to design and deploy the project.

6.1. Data Wrangling and cleaning

3.958 Conpes Documents were gathered from the National Planning Department (DNP) website.

Those documents were processed in two complementary ways:

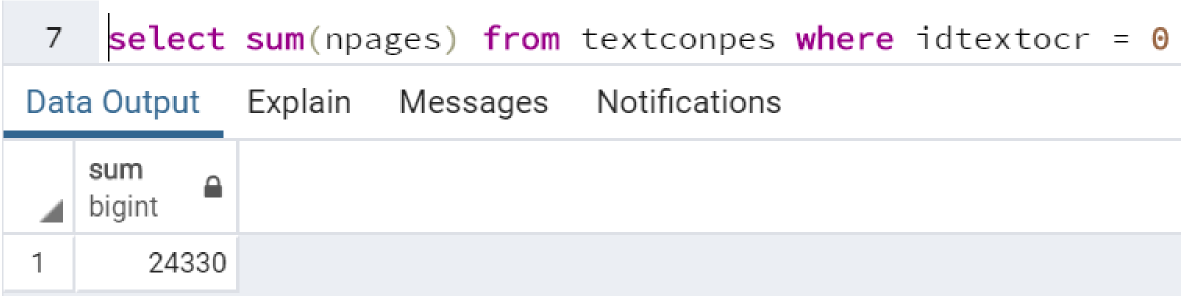
First, a built-in function (PyPDF2.PdfFileReader) was run to extract the text within the documents and count the number of pages.

After that, a series of text cleaning functions were performed (delete special characters, etc.) and a metric of at least 25 words per page was established in order to accept the processed text.

Finally, a manual revision of the top 400 documents ordered by number of pages was performed. The idea was to detect documents that passed the threshold defined before but where the text extraction from the pdf was still not performed in a correct way.

As a result, 994 documents, corresponding to 25% of all documents gathered, were accepted. These documents sum up **24330** pages.

Figure 1 - Documents processed by PyPDF2



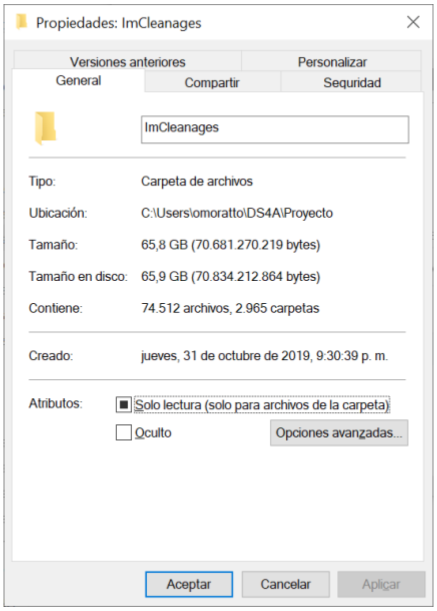
```
select sum(npages) from textconpes where idtextocr = 0
```

```
   sum
   bigint
1     24330
```

For the remaining documents, an OCR process, using Tesseract, was performed.

First, each page of the remaining pdf files was converted to an image. In this process, near **75.000** pages were converted to images. The disk space of these group of images is around 70 GB, as shown below.

Figure 2 - Information processed



ImCleanages

Tipo: Carpeta de archivos

Ubicación: C:\Users\omoratto\DS4A\Proyecto

Tamaño: 65,8 GB (70.681.270.219 bytes)

Tamaño en disco: 65,9 GB (70.834.212.864 bytes)

Contiene: 74.512 archivos, 2.965 carpetas

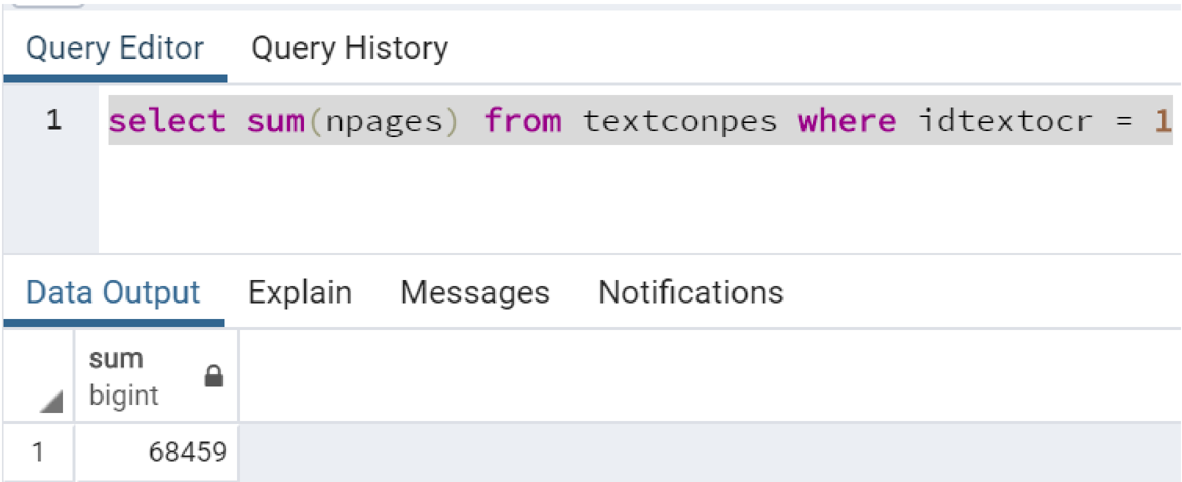
Creado: jueves, 31 de octubre de 2019, 9:30:39 p. m.

Atributos: ■ Solo lectura (solo para archivos de la carpeta)

□ Oculto

After that, an OCR process was conducted in order to extract the text within each image. 2863 Conpes documents were processed in this way, corresponding to **68459** pages.

Figure 3 - Documents processed by Tesseract



```
select sum(npages) from textconpes where idtextocr = 1
```

```
     sum
     bigint
1    68459
```

The time consumed in the OCR process was significant.

At the end, **3877** Conpes documents were processed, containing **18 million words**

The text from the Conpes documents was cleaned performing different kinds of operations:

- Stop-words in Spanish
- Removing numbers and special characters (i.e. dollar signs, space characters, new line characters, etc.)

- Removing short characters left over from the previous cleaning.
- Removing more than one white space between characters and leftovers at the end of sentences.
- An analysis of most frequent words was performed in order to identify additional words that should be removed. Different kinds of words were removed, for example planeacion (the National Planning Department is the author of all the Conpes documents, so planning – planeacion is a very frequent word with no specific context), US (one subcategory of Conpes documents is related with financing, usually in USD, so appears very frequently in the documents), words wrongly processed the reading process, such as multiple repeated characters, such as "nnnn" or "aaaaa".

All the information regarding the Conpes documents was stored in a PostgreSQL database (conpes number, conpes type – Economic or Social, conpes title, conpes text, conpes clean text, whether the text was processed by reading the pdf file directly or by OCR and the conpes issue date among other characteristics). Another information presented was the presidential period from the date that the Conpes documents have been created. The database was normalized in order to ensure data integrity.

6.2. Postgres Database

All the results have been persisted in a postgres database (conpes number, conpes type – Economic or Social, conpes title, conpes text, if the text was processed by reading the pdf file directly or by OCR, conpes issued date, among others). Another information persisted is the presidential periods from the date that the conpes documents have been created. The database was normalized in order to ensure data integrity.

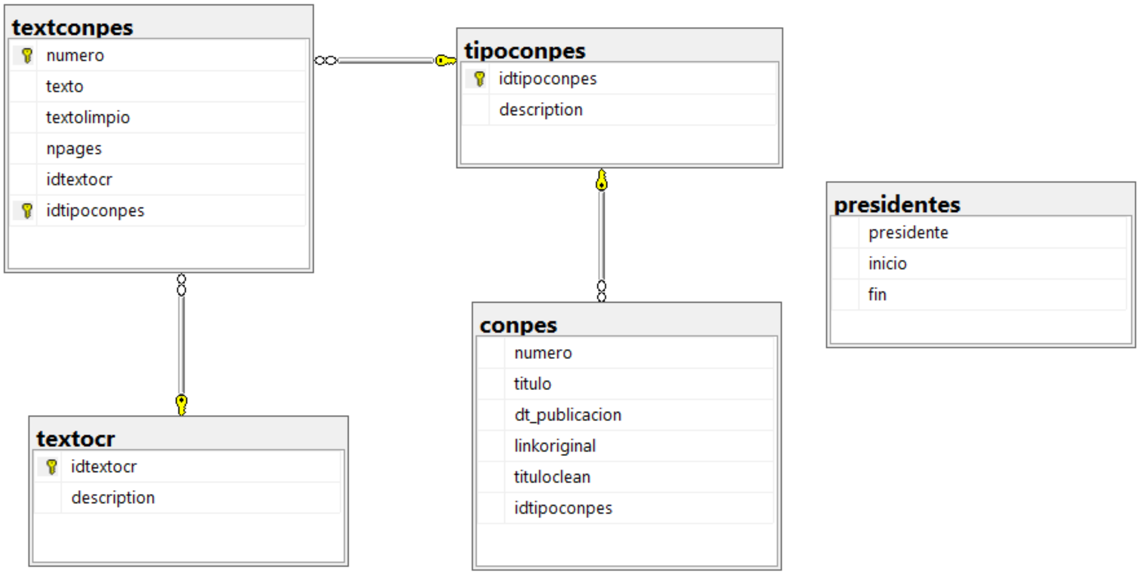
6.2.1. Conection parameters

[redacted]

6.2.2. Entity Relationship Diagram

The database design was based on the E-R Model. The primary source of information contains information about Conpes documents (title, original link, type, date of publication), registered on the table Conpes. There are two types of Conpes documents, Social and Economic, inserted into the table tipoConpes. The Conpes documents were processed and persisted into the table textconpes. In this table, information related with the Conpes number, type of Conpes, original processed text, clean text, type o text processing (OCR or other) is stored. Additionally, a table with the beginning and end of the presidential periods since the Conpes documents are implemented was desing (table Presidentes)

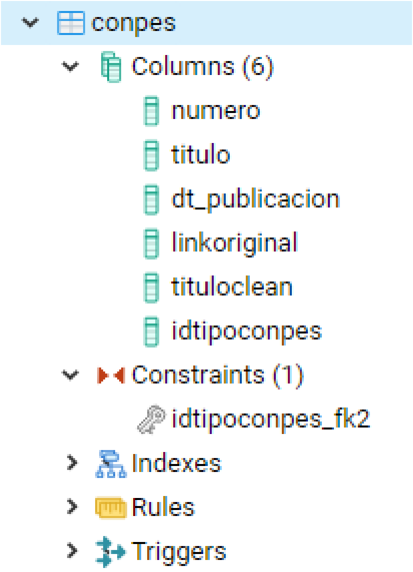
Figure 4 - E-R Diagram



6.2.3. Tables Schema

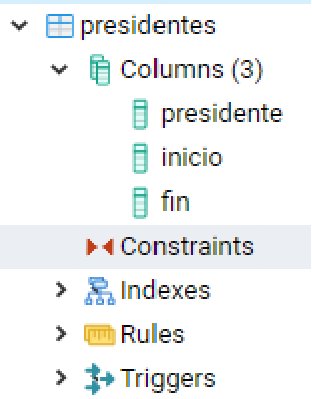
The schemas of the tables as shown below.

- Table Conpes:



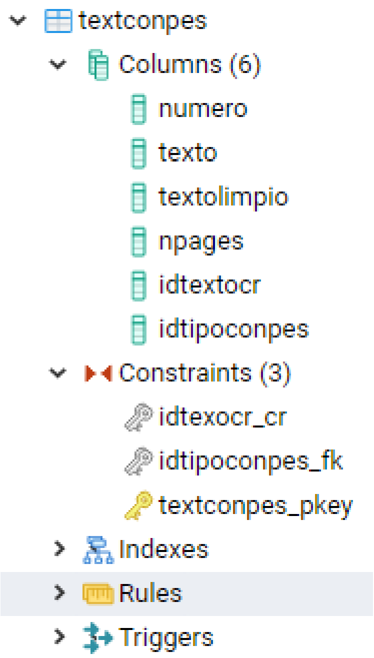
```
conpes
  Columns (6)
    numero
    titulo
    dt_publicacion
    linkoriginal
    tituloclean
    idtipoconpes
  Constraints (1)
    idtipoconpes_fk2
  Indexes
  Rules
  Triggers
```

- Table Presidentes:



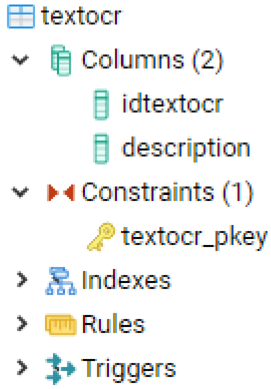
```
presidentes
  Columns (3)
    presidente
    inicio
    fin
  Constraints
  Indexes
  Rules
  Triggers
```

- Table TextConpes



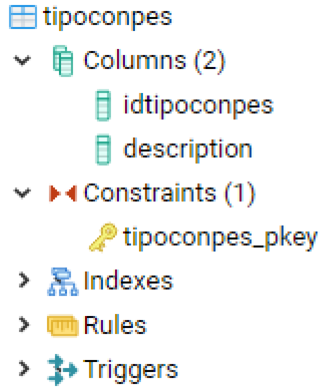
```
textconpes
    Columns (6)
        numero
        texto
        textolimpio
        npages
        idtextocr
        idtipoconpes
    Constraints (3)
        idtexocr_cr
        idtipoconpes_fk
        textconpes_pkey
    Indexes
    Rules
    Triggers
```

- Table TextOcr



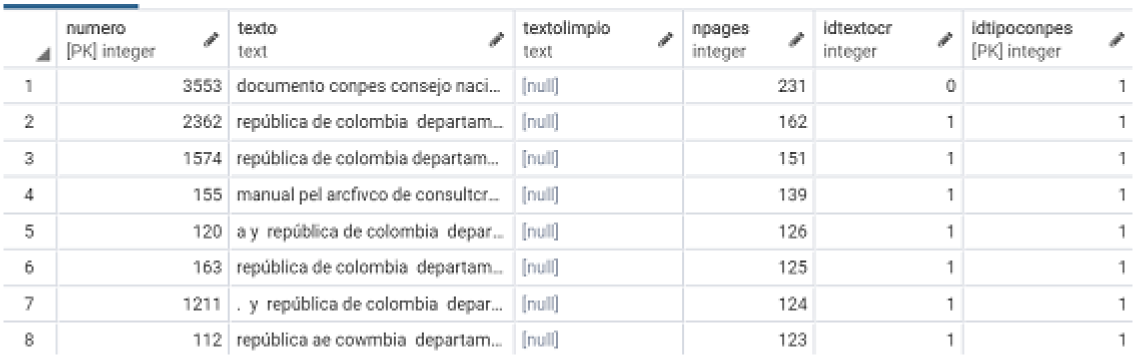
```
textocr
    Columns (2)
        idtextocr
        description
    Constraints (1)
        textocr_pkey
    Indexes
    Rules
    Triggers
```

- Table TipoConpes



```
tipoconpes
    Columns (2)
        idtipoconpes
        description
    Constraints (1)
        tipoconpes_pkey
    Indexes
    Rules
    Triggers
```

It is important to note that there are Conpes documents with significant number of pages



	numero [PK] integer	texto text	textolimpio text	npages integer	idtextocr integer	idtipoconpes [PK] integer
1	3553	documento conpes consejo naci...	[null]	231	0	1
2	2362	república de colombia departam...	[null]	162	1	1
3	1574	república de colombia departam...	[null]	151	1	1
4	155	manual pel arcfivco de consultcr...	[null]	139	1	1
5	120	a y república de colombia depar...	[null]	126	1	1
6	163	república de colombia departam...	[null]	125	1	1
7	1211	. y república de colombia depar...	[null]	124	1	1
8	112	república ae cowmbia departam...	[null]	123	1	1

Meanwhile different queries over the database has been designed in order to analyze the data and select possible visualizations for the final dash. Also, a preliminary version of the dash has been designed.

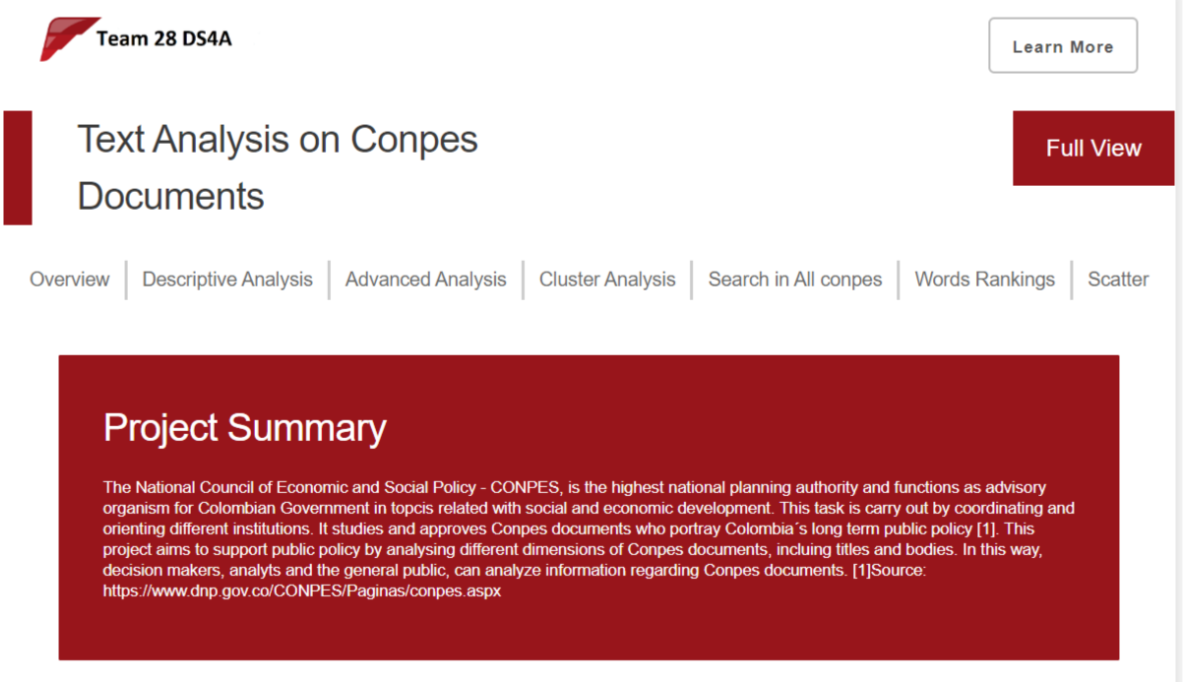
6.3. Front End – Dash Board

A Dash application was developed. The Dash was deployed on a custom URL:

http://colombiagovernmentai.ml:80/

The Dash consists of the seven (7) sections

Figure 5 - Dash structure



Text Analysis on Conpes Documents

Project Summary

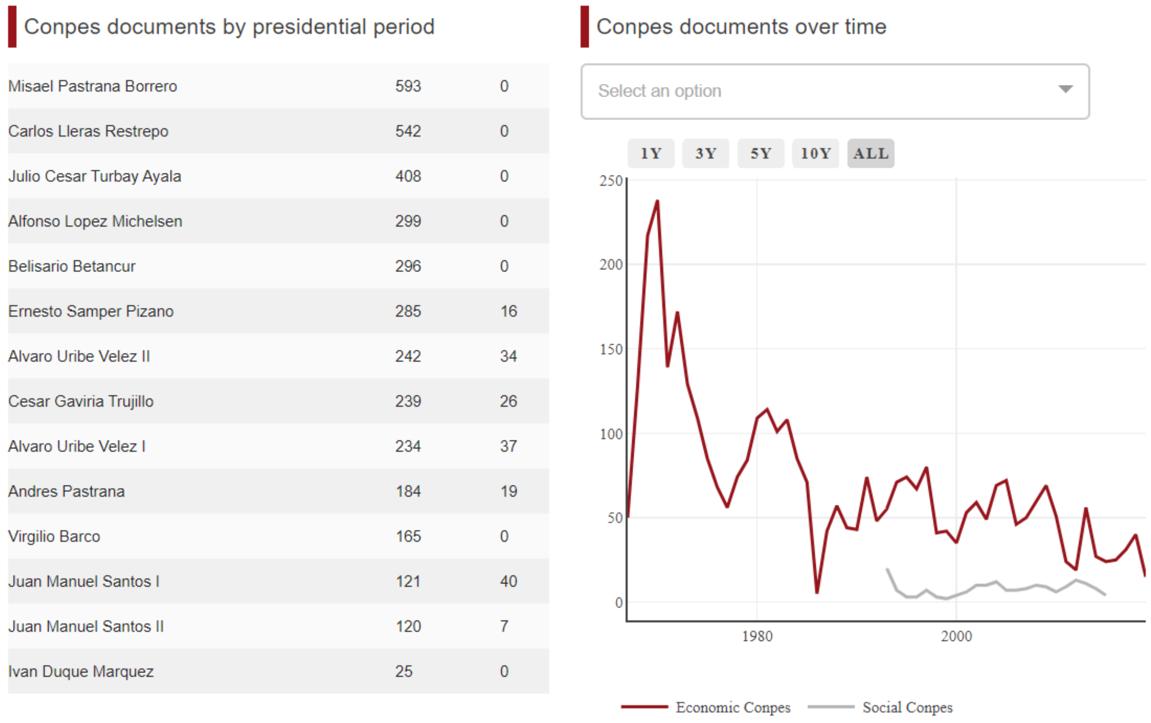
The National Council of Economic and Social Policy - CONPES, is the highest national planning authority and functions as advisory organism for Colombian Government in topcis related with social and economic development. This task is carry out by coordinating and orienting different institutions. It studies and approves Conpes documents who portray Colombia´s long term public policy [1]. This project aims to support public policy by analysing different dimensions of Conpes documents, including titles and bodies. In this way, decision makers, analyts and the general public, can analyze information regarding Conpes documents. [1]Source: https://www.dnp.gov.co/CONPES/Paginas/conpes.aspx

The sections are:

6.3.1. Overview

This section, contains an overview of the project, a broad description of the volume of data processed and some key findings. Some visualizations have been included, such as the number of Conpes documents, by type, approved by presidential period (dynamically updated from a query to the DB) and also, a line chart of the number of documents over time.

Figure 6 - Dash: Overview page



Conpes documents by presidential period

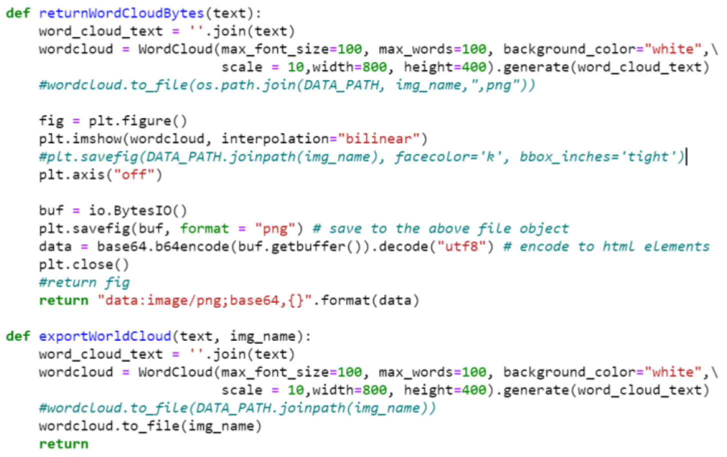
Misael Pastrana Borrero	593	0
Carlos Lleras Restrepo	542	0
Julio Cesar Turbay Ayala	408	0
Alfonso Lopez Michelsen	299	0
Belisario Betancur	296	0
Ernesto Samper Pizano	285	16
Alvaro Uribe Velez II	242	34
Cesar Gaviria Trujillo	239	26
Alvaro Uribe Velez I	234	37
Andres Pastrana	184	19
Virgilio Barco	165	0
Juan Manuel Santos I	121	40
Juan Manuel Santos II	120	7
Ivan Duque Marquez	25	0

Conpes documents over time

6.3.2. Descriptive analysis

It includes a basic analysis of documents, such as volumetric, wordclouds, most frequent words, and so on. Basic visualizations have been performed. A big challenge encountered in this phase was the fact that Matplotlib figures are no longer supported by Dash. Some matplotlib functionalities, such as wordclouds, are not present in plotly. Hence, a workaround was designed, applied to wordclouds, but it can be applied to any matplotlib figure needed to be displayed into Dash. This is to export matplotlib figures as images (both in disk and memory) and then insert into dash. This workaround consumed a significant time. The workaround was tested on a wordcloud from the Conpes titles. Its source code is shown below:

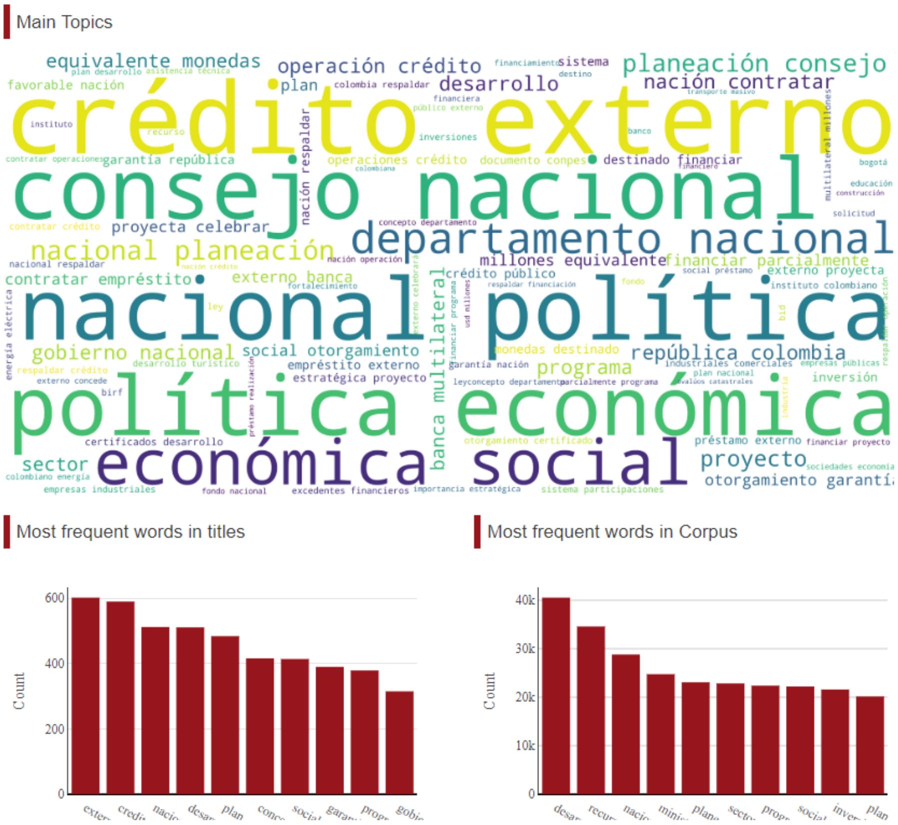
Figure 7 - Workaround to import Matplotlib figures into Dash



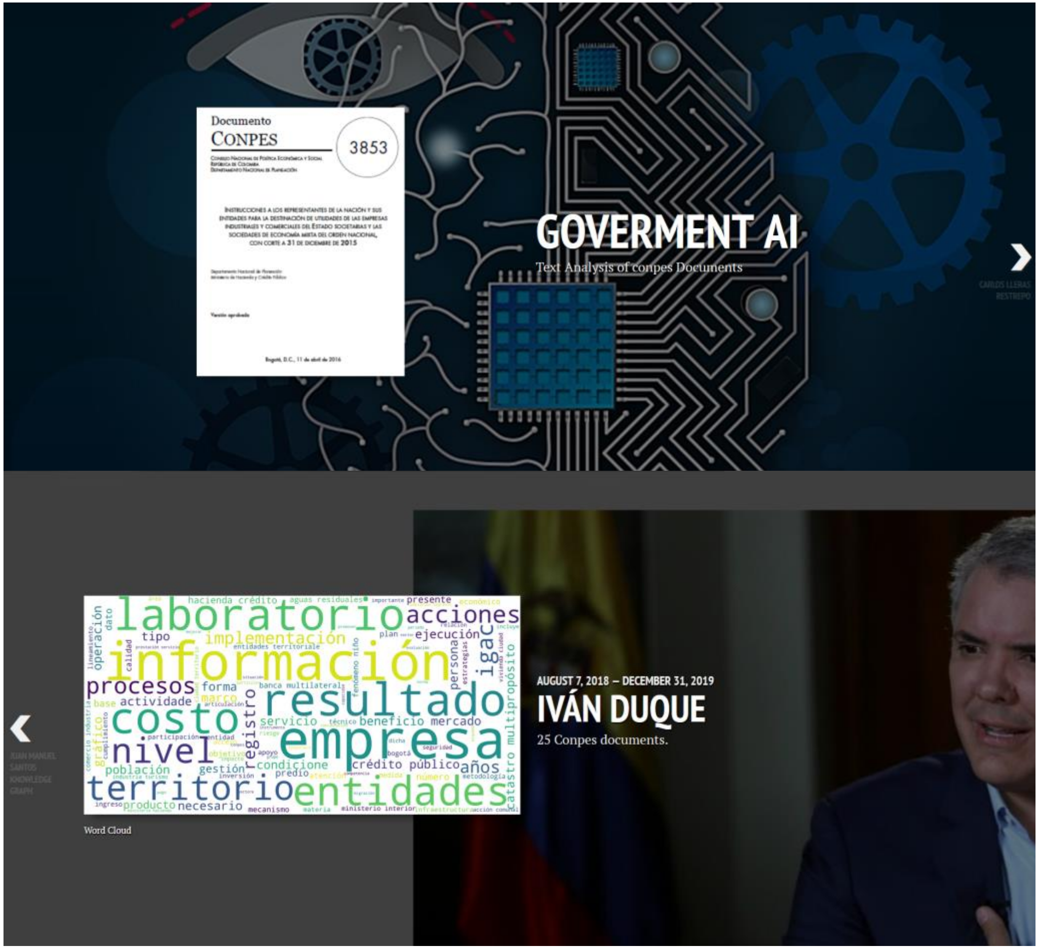
```python
def returnWordCloudBytes(text):
    word_cloud_text = ''.join(text)
    wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white",\
                          scale = 10,width=800, height=400).generate(word_cloud_text)
    #wordcloud.to_file(os.path.join(DATA_PATH, img_name,",png"))

    fig = plt.figure()
    plt.imshow(wordcloud, interpolation="bilinear")
    #plt.savefig(DATA_PATH.joinpath(img_name), facecolor='k', bbox_inches='tight')
    plt.axis("off")

    buf = io.BytesIO()
    plt.savefig(buf, format = "png") # save to the above file object
    data = base64.b64encode(buf.getbuffer()).decode("utf8") # encode to html elements
    plt.close()
    #return fig
    return "data:image/png;base64,{}".format(data)

def exportWorldCloud(text, img_name):
    word_cloud_text = ''.join(text)
    wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white",\
                          scale = 10,width=800, height=400).generate(word_cloud_text)
    #wordcloud.to_file(DATA_PATH.joinpath(img_name))
    wordcloud.to_file(img_name)
    return
```

At the end, due to the high amount of time consumed and the fact that this functionality is not dynamic, WordCloud was developed, presented as image and the image was imported into the Dash, but not dynamically.

Figure 8 - Dash: Descriptive Analysis page



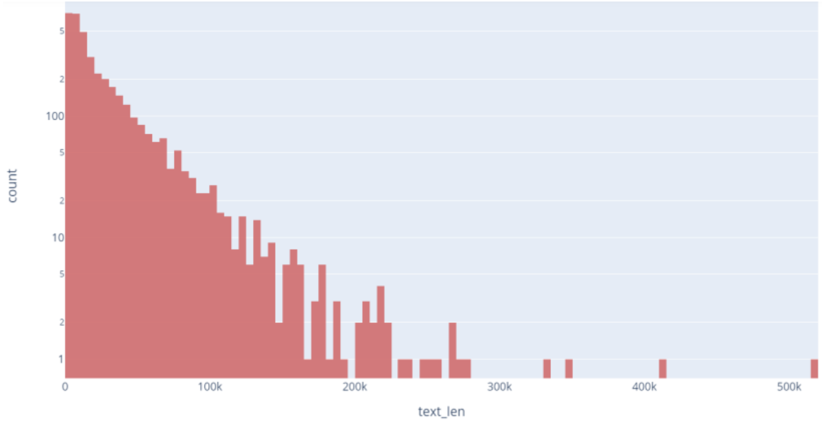
For the purposes of elegantly displaying the wordclouds per presidential period, we built a story telling timeline to show the key aspects of every president in their Conpes documents.



6.3.3. Advanced Analysis

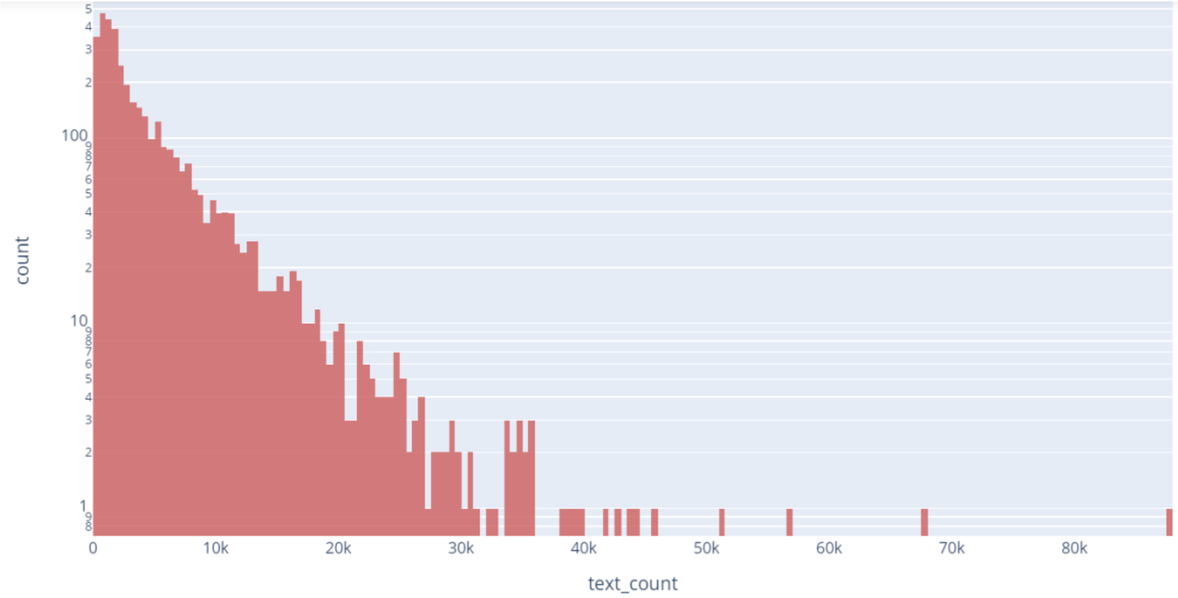
Contains more advanced analysis of the documents and further visualizations, such as, the text longitude and the count of words.

Figure 9 - Text length frequency



At it is shown, the vast majority of Conpes documents has a character longitude smaller than 20K characters. There are outliers, such as documents with more than 400K characters.

Figure 10 - Number of words frequency



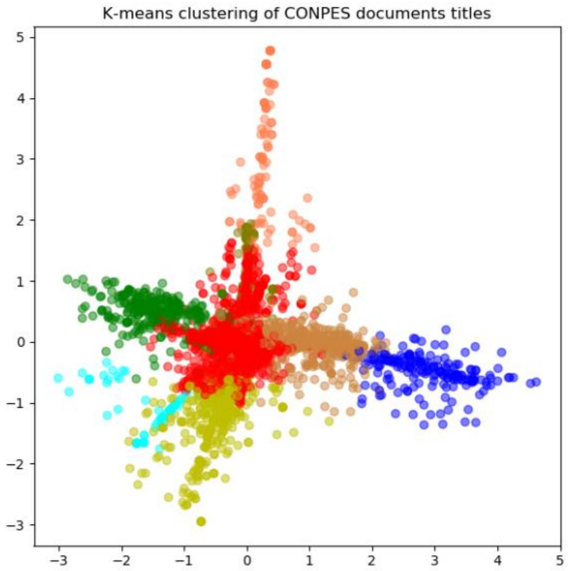
Regarding words, the majority of Conpes documents have fewer than 30K words, with atypical values with more than 60K words per document.

6.3.4. Cluster Analysis

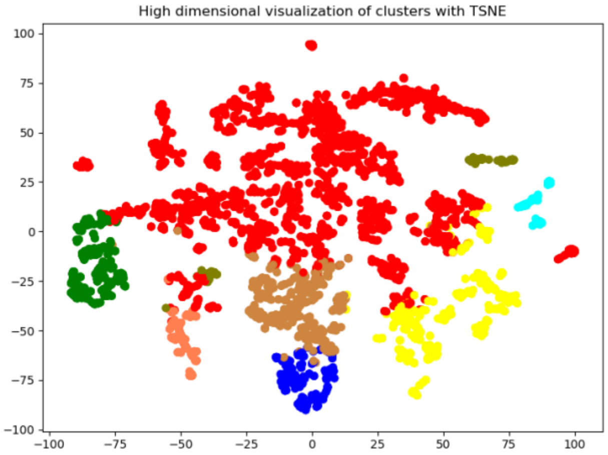
A clustering exercise was done over the 3877 titles of the documents loaded, to find what topics are highlighted among all the diversity of public policies. The process goes as follows:

1. Create a corpus object from the previously cleaned text of the titles.
2. Vectorize the text with the technique called Term Frequency – Inverse Document Frequency.
3. Build a distance matrix using [1 - cosine similarity]
4. Reduce dimensionality using Principal Components Analysis - PCA. This is necessary since the distance matrix is extremely large for the clustering algorithm.
5. Find the optimal number of clusters using 3 methods: Elbow, Silhouette and Within-Cluster-Sum-of-Squares. Decide the optimal number by majority rule.
6. Use the principal components that store the majority of the variance explained to run the K-means algorithm with the chosen number of clusters.
7. Since the plot of these two components show the documents overlapping each other, we used a manifold technique called T-student Stochastic Neighbor Embedding – TSNE, that helps in a cleaner visualization of the clusters.

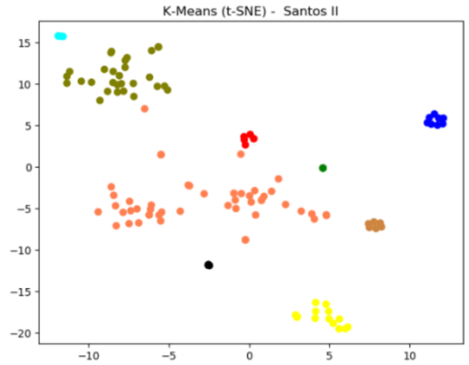
This plot shows the clustering using the first two components of the PCA:

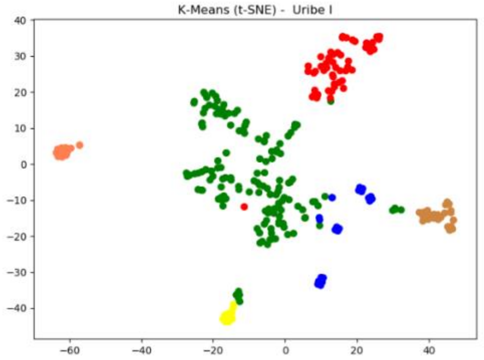


This multidimensional object contains all the titles represented by the color dots in 8 clusters. To visualize it in two dimensions TSNE was used:



The same algorithm was implemented in two presidential periods, to see the differences between the public policies. The presidents chosen were: Álvaro Uribe Vélez (2002-2006) and Juan Manuel Santos Calderón (2014-2018). Their clustering graphs were the following:



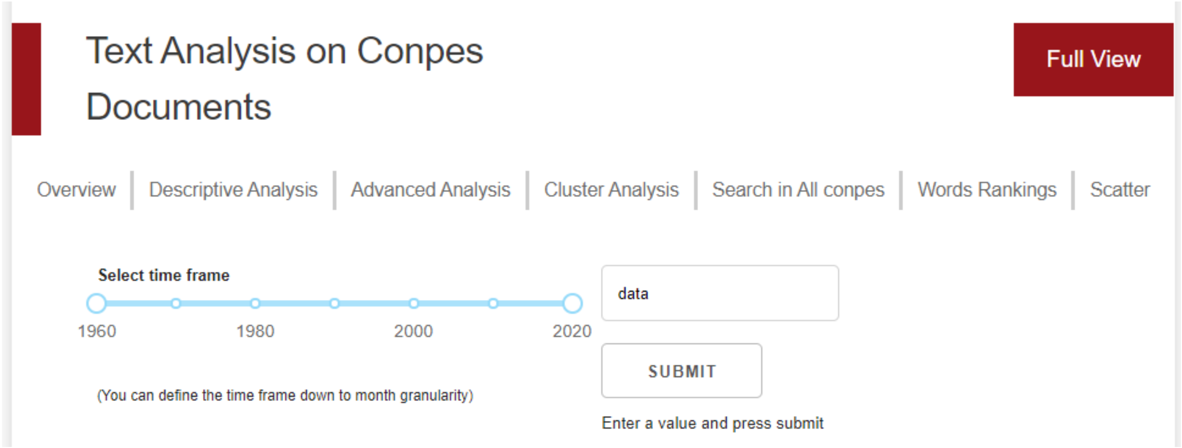


The Conpes documents produced during the Álvaro Uribe period were grouped in 6 clusters, in contrast to the 9 clusters of the Santos Period. The discoveries made are also included in the results section.

6.3.5. Search in all conpes

This is a dynamic functionality included in the project. In this section, the user can search for words within the Conpes Documents.

Figure 11 - Dash: Search in all Conpes



data

(You can define the time frame down to month granularity)

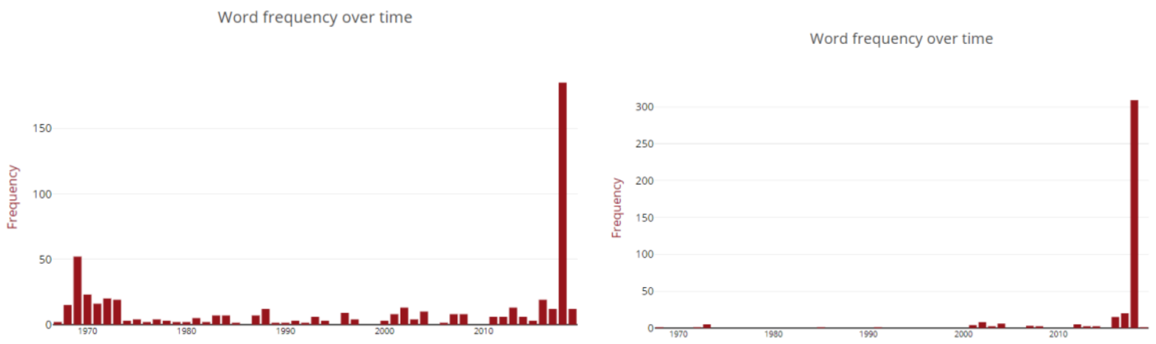
Enter a value and press submit

The Dash shows the frequency of the searched term along months and years. In this way, the user can see the tendency, along time of the searched term, if the is concentration of Conpes documents through time related with the topic. Also, how public policy matches/mismatches political juncture.

Figure 12 - Example of word search

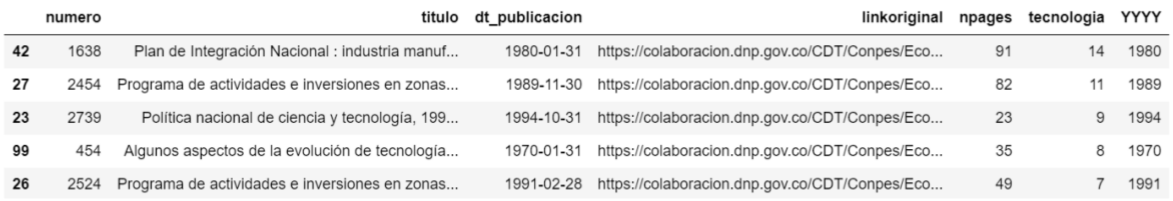
Frequency of the term data

Frequency of the term Farc



Also, a table with the specific Conpes documents with highest frequency of the term of interest are shown. In that way, the user can access the original resource for more information.

Figure 13 - Results: Conpes documents with highest frequency of the searched term

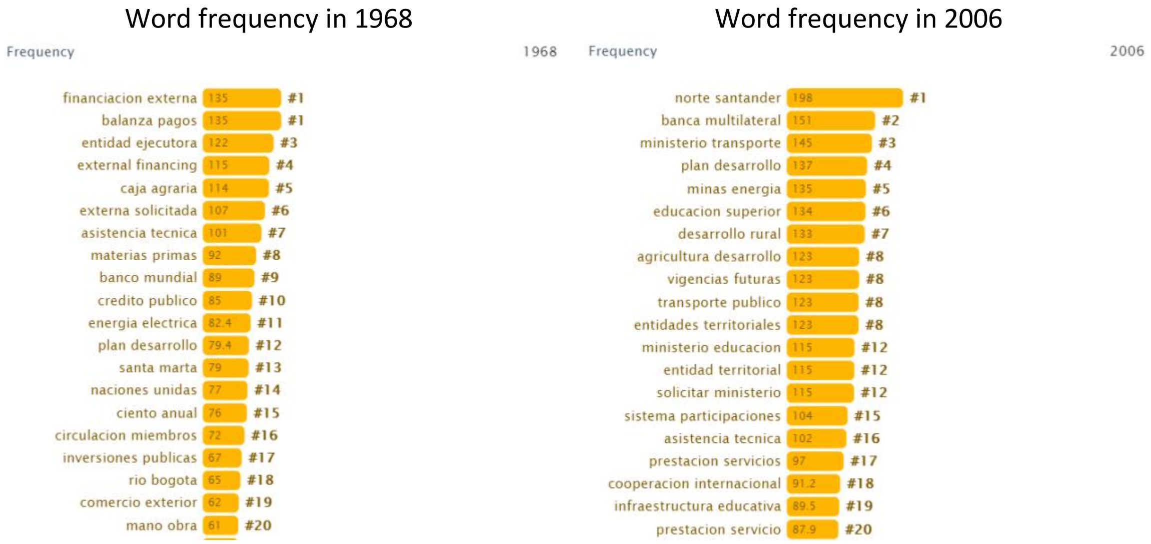


	numero	titulo	dt_publicacion	linkoriginal	npages	tecnologia	YYYY
42	1638	Plan de Integración Nacional : industria manuf...	1980-01-31	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	91	14	1980
27	2454	Programa de actividades e inversiones en zonas...	1989-11-30	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	82	11	1989
23	2739	Política nacional de ciencia y tecnología, 199...	1994-10-31	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	23	9	1994
99	454	Algunos aspectos de la evolución de tecnología...	1970-01-31	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	35	8	1970
26	2524	Programa de actividades e inversiones en zonas...	1991-02-28	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	49	7	1991

6.3.6. Word Rankings

The idea of public policy changing over time led to the development of this functionality. The main purpose of this section is to show how the most frequent words in Conpes documents have changed over time. An animated barplot was designed and incorporated into the Dash in order to show this.

Figure 14 - Word frequency over time

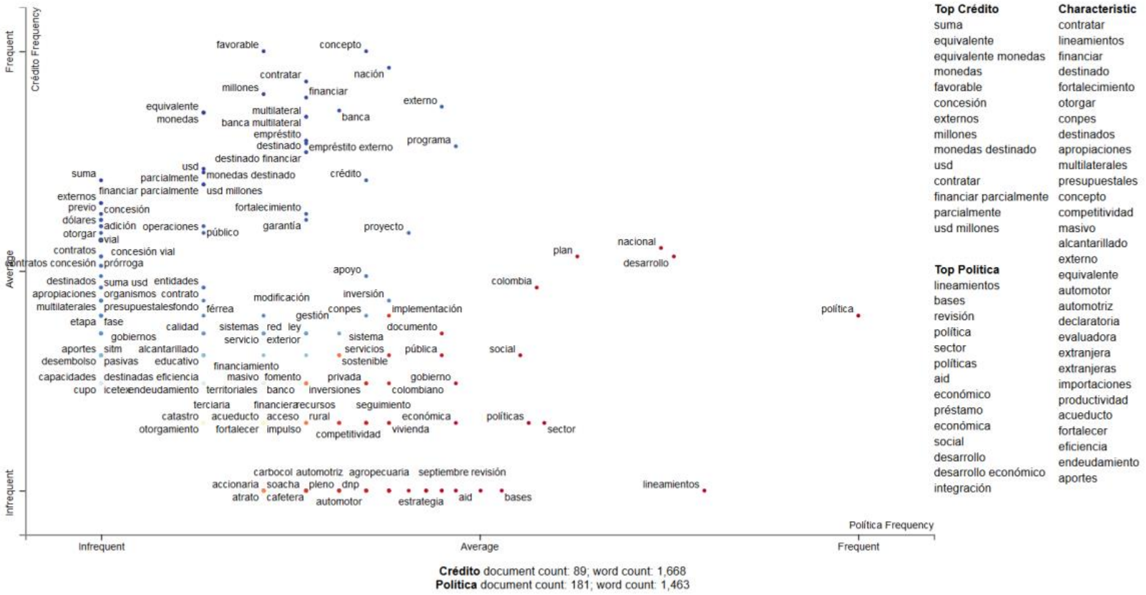


This functionality was found to be so useful in terms of insight that a unique Dash section was devoted to it.

6.3.7. Scattertext

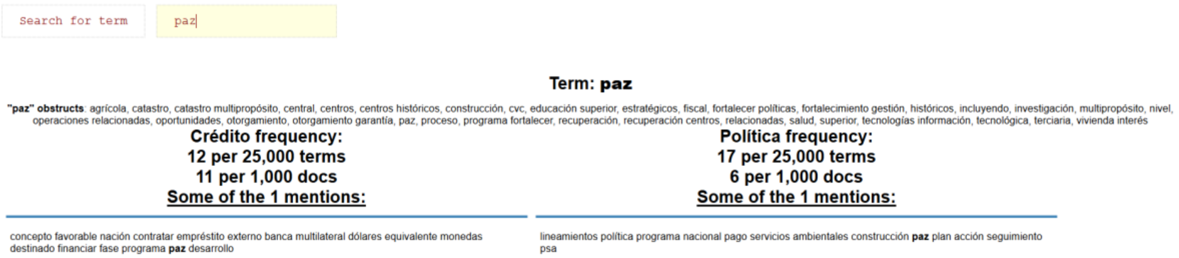
Another dynamic functionality incorporated into the Dash. In this section the user can see how language differs among document types, and in this specific example, see the difference between two types of Conpes documents. The documents are called ‘Policy’ and ‘Credit’ Conpes, the first one focuses on policies that public entities in collaboration with private ones should follow to reach certain economic and social objectives, while the second type deals with financing projects that could be beneficial to our society.

The following graph shows a sample of the two kinds of documents since with a large number of text requires a long loading time.



Crédito document count: 89; word count: 1,668
Politica document count: 181; word count: 1,463

In addition to this, the user can search for specific words and find their location in the text, like for example, searching for the word ‘paz’ or peace in english returns information about related words.



Term: paz

"paz" obstructs: agrícola, catastro, catastro multipropósito, central, centros, centros históricos, construcción, cvc, educación superior, estratégicos, fiscal, fortalecer políticas, fortalecimiento gestión, históricos, incluyendo, investigación, multipropósito, nivel, operaciones relacionadas, oportunidades, otorgamiento, otorgamiento garantía, paz, proceso, programa fortalecer, recuperación, recuperación centros, relacionadas, salud, superior, tecnologías información, tecnológica, terciaria, vivienda interés

Crédito frequency:
12 per 25,000 terms
11 per 1,000 docs
Some of the 1 mentions:

concepto favorable nación contratar empréstito externo banca multilateral dólares equivalente monedas destinado financiar fase programa **paz** desarrollo

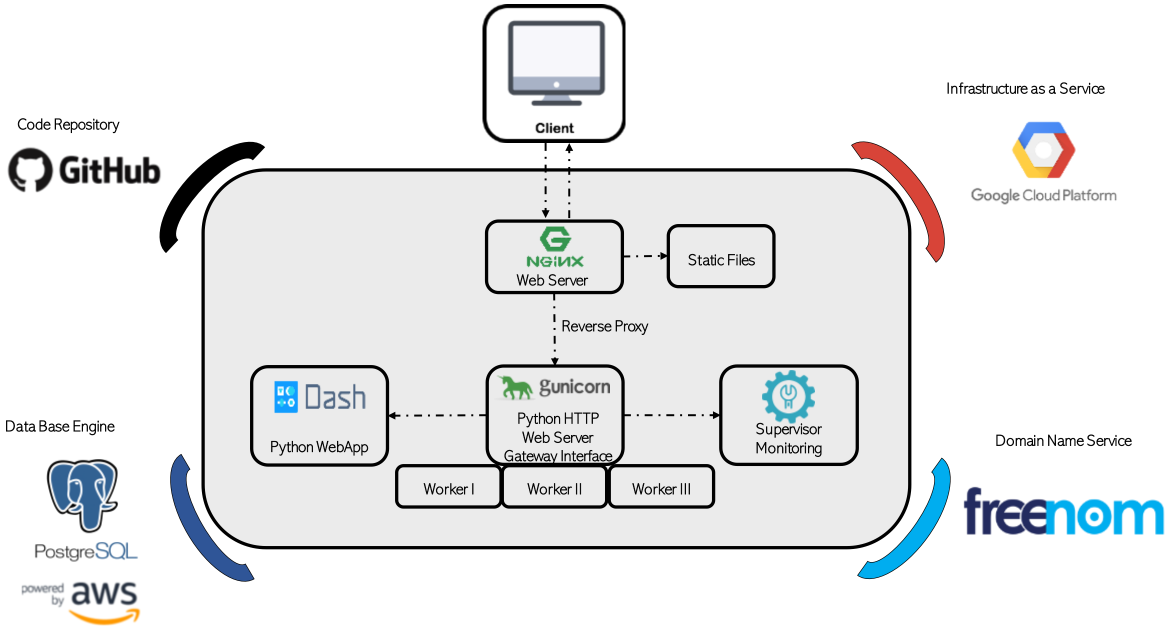
Política frequency:
17 per 25,000 terms
6 per 1,000 docs
Some of the 1 mentions:

lineamientos política programa nacional pago servicios ambientales construcción **paz** plan acción seguimiento psa

6.4. AWS Infrastructure

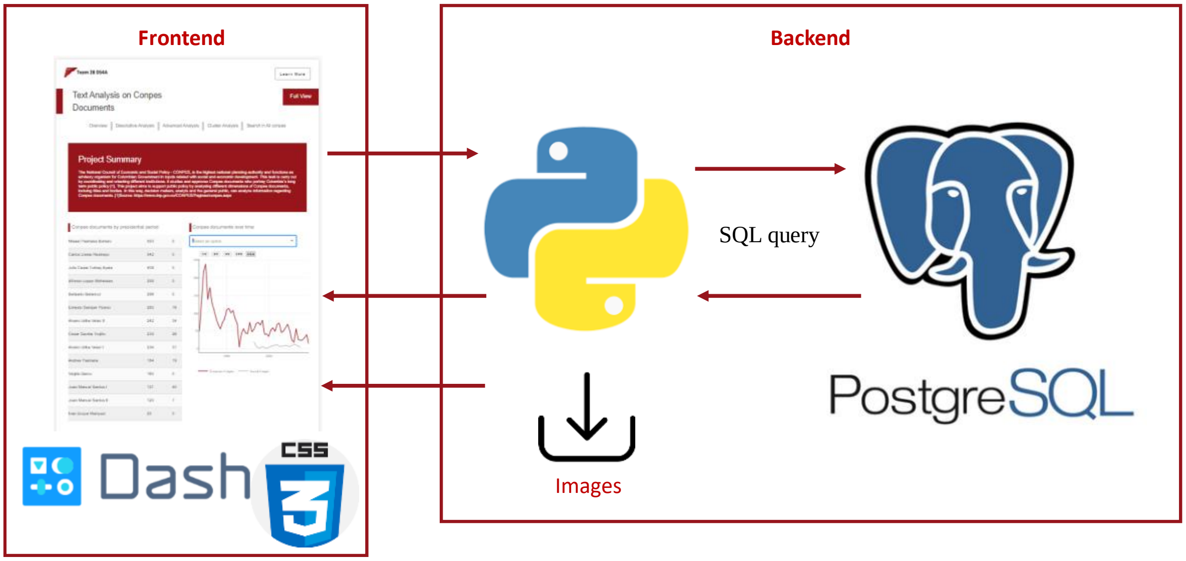
As mentioned before, the data was stored in a PostgreSQL database in AWS. The front end is deployed in a Google Cloud Platform. The website is working on a production environment. The domain name was obtained from Freenom. The Dash web app was deployed on a gunicorn Python HTTP Web Server, as shown below.

Figure 15 - Solution Architecture



The next figure shows the interrelation of the different components at run time. Petitions over the front end are carried out by the python engine. The dynamic features of the interface interact through SQL querys with the PostgreSQL Data base. The static components, such as cluster visualizations, due to the high processing requirements and long response time, are loaded into the application statically (images) that are persisted on the application server.

Figure 16 - Application calls

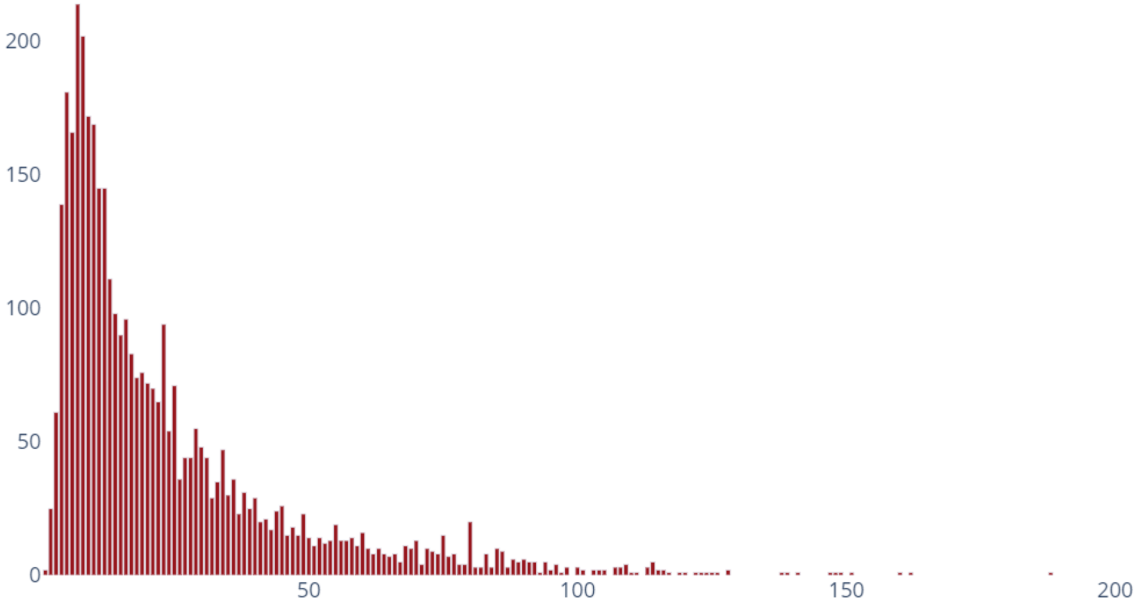


7. Results

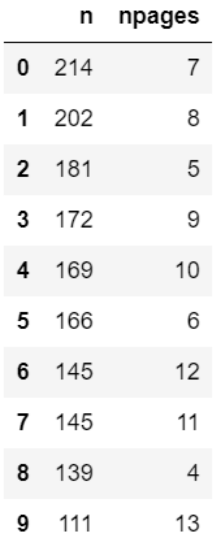
From the exploratory analysis, several findings were encountered:

There is a great disparity among Conpes documents. For example, the number of pages is very heterogeneous. The next figure shows the histogram of number of pages.

Figure 17 - Distribution of Conpes documents by number of pages



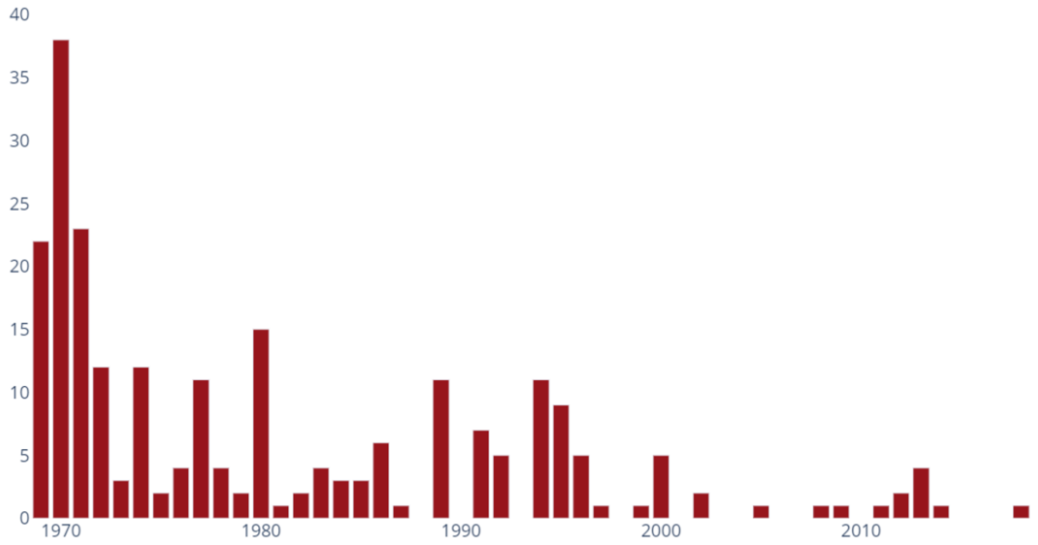
The most frequent, in terms of number of pages, are documents of 7 pages long, followed by 8 and 5. In the top 10 most frequent size of documents by number of pages



	n	npages
0	214	7
1	202	8
2	181	5
3	172	9
4	169	10
5	166	6
6	145	12
7	145	11
8	139	4
9	111	13

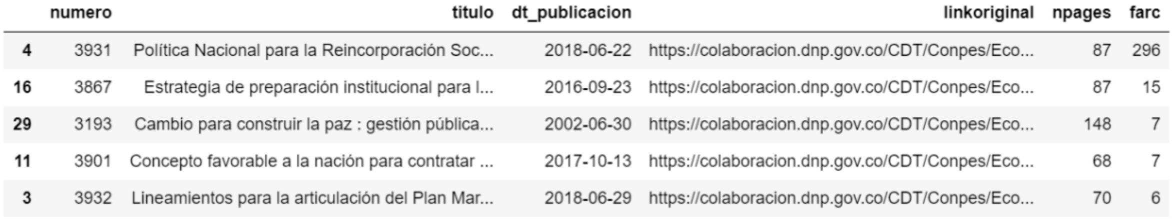
Another interesting finding, from the exploratory analysis and the functionalities developed, is that some words that are supposed to be of current or recent interest in public policy, were addressed with more interest in the past. Sometimes, **counterintuitive** results can be achieved. For example, the word tecnologia (technology), is supposed to be a novel concept, but it was addressed with higher frequency in the 70´s than in recent years, as shown below.

Figure 18 - Frequency over time of the word Technology



Also, depending on the social, economic, and political juncture, the long term public policy is adjusted. One example of this is the peace process. The term Farc had an unusual high frequency in the context of the peace process. In the Conpes document regarding the incorporation of the Farc members, the frequency of the word exceeds almost 20 times the next Conpes document.

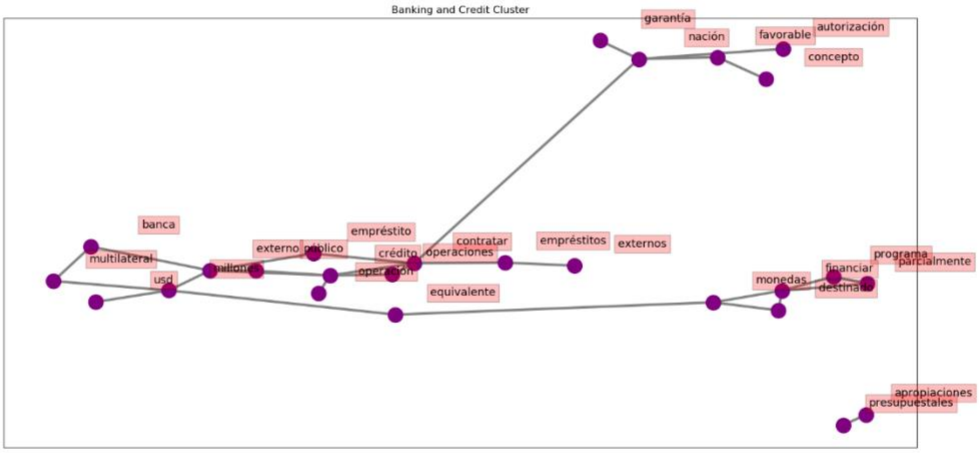
Figure 19 - Frequency of Farc within Conpes documents



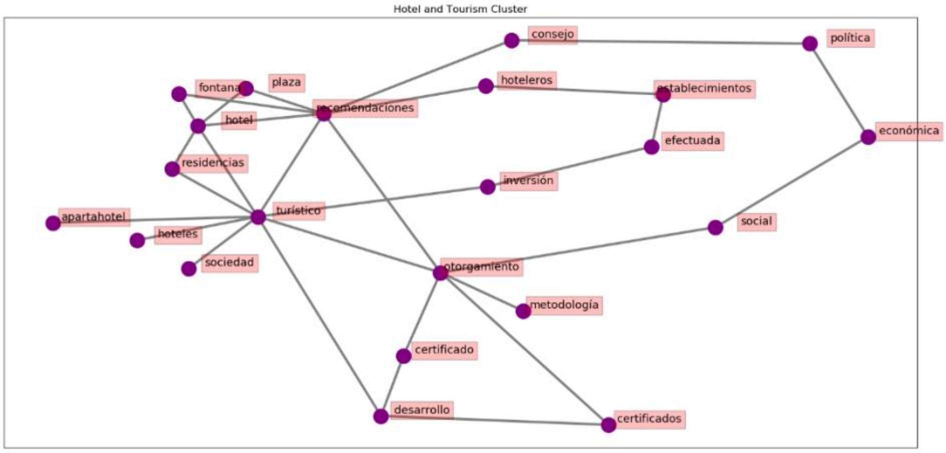
	numero	titulo	dt_publicacion	linkoriginal	npages	farc
4	3931	Política Nacional para la Reincorporación Soc...	2018-06-22	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	87	296
16	3867	Estrategia de preparación institucional para l...	2016-09-23	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	87	15
29	3193	Cambio para construir la paz : gestión pública...	2002-06-30	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	148	7
11	3901	Concepto favorable a la nación para contratar ...	2017-10-13	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	68	7
3	3932	Lineamientos para la articulación del Plan Mar...	2018-06-29	https://colaboracion.dnp.gov.co/CDT/Conpes/Eco...	70	6

Regarding the clustering results for all the titles, knowledge graphs were made in order to visualize the most relevant clusters, since the word cloud sometimes is not that good of an option.

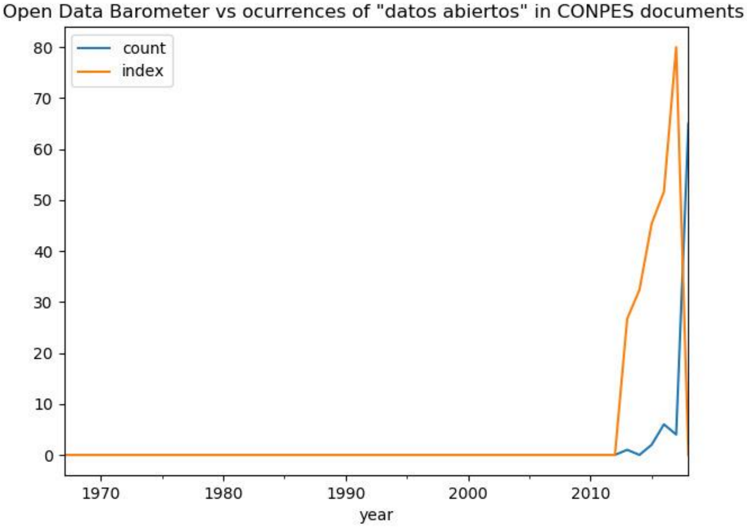
There was a predominant cluster containing words related to banking, credit operations and foreign investment. This makes sense since 'credit' is a type of Conpes document, which has become more popular in the later years. The credit Conpes allow public entities to borrow money from the government in order to make all kinds of projects, from bridges and schools to social programs.



The second predominant cluster contains words related to tourism, which is a positive indicator after the peace process.



Lastly, having all the words at our disposition from the documents, one could plot the occurrences of words against indicators to see if the documents share any real life relations to what is going with the country. The next plot shows the Open Data Barometer produced by the World Wide Web Foundation versus the number of times "datos abiertos" is mentioned in the corpus.



We can observe a clear relation between these numbers, which demonstrates that Conpes documents are indeed important for the country public policy.

8. Feasible next steps

In order to upgrade the functionalities included into the application, the most immediate enhancements can be performed:

- Better search functionality: include not only one or multiple consecutive words (i.e, big data) but multiple words to search within the documents. Include search operators such as AND or OR, words near other words (using n-grams), subquery search results (be able to search within search results), export results to Excel or other platforms.

- Geospatial connection: develop maps in order to identify topics within the Conpes documents geospatially related, i.e., municipalities named within Conpes documents.
- External documents matching: Develop a model (forecast, cluster or another technique) in order to inject unstructured data (i.e. news, research articles, laws) and match the most similar Conpes documents related with the documents provided.
- Provide automatic reports and conclusions.

9. Applications and extensions

The applications and extensions of the platform are multiple. They were categorized in five main topics

RegTech[1]: Identify tendencies in new regulation and laws, fines, global enforcement cases and sanctions. Patterns in contracts. Patterns in mutual evaluations from organizations such as FATF, OCDE in order to identify regulatory gaps. Further applications, with broader use of AI can be performed, such as automatic generation of regulations or laws based on documents similarities.

SupTech[2] – Government: Analyze supervision in-situ reports. Establish customer complaints tendencies. Adjust supervisory activities and analyze legal claims against government agencies. Spot atypical patterns in government contracts. Support activities in government license applications.

Customer service: Support churn models in order to attract and retain customers based on analyzing tendencies in unstructured data (tweets, customer services). Identify patterns of service quality and frauds.

Academic Research: Pre-analysis of state of the art. Main topics and unattended fields of research; Tendencies in topics during bibliographic research. Group literature according to main topics.

Risk Management: Identify patterns within fraudulent claims, establish groups within medical patients, spot patterns in fraudulent credit applications, Analysis of Suspicious activities Reports (SARs), analyze cases by GPOs in order to design macro strategies to tackle organized crime.

[1] https://www.investopedia.com/terms/r/regtech.asp

[2] https://www.centralbanking.com/central-banks/economics/data/3650941/suptech-more-than-just-a-new-name-for-solving-an-old-problem