PROCOLOMBIA Propensity Investment Model

Team 15 – DS4A/Colombia - Cohort 5

Contents

1. Team 15 members

- Angi Aparicio
- Carlos López
- Daniel Romero
- Dariel Rincones
- Germán Puertas
- Manuel Arias

2. Context

PROCOLOMBIA is a government institution that oversees the commercial promotion of non-traditional exports, international tourism, and foreign investment in Colombia. Their services are geared towards the design and execution of internationalization strategies to generate, develop, and close business opportunities. Foreign investors will find variety of business options to buy products and services in Colombia and will be able to organize business meetings with local companies of their interests.

In addition, PROCOLOMBIA has a national and international office network that offers comprehensive support and advice to clients, through services or instruments aimed at facilitating the design and execution of its internationalization strategy (PROCOLOMBIA, 2021).

3. Project Purpose

As part of its activities, PROCOLOMBIA needs to identify foreign companies with potential investment in Colombia, that may be represented through purchases of product and services, as well as direct foreign investment. Therefore, every year a segmentation exercise is carried out by its commercial area to identify new opportunities for Colombia.

Currently, PROCOLOMBIA is using multiple sources of information such as public data from other government institutions and repositories from private companies dedicated to track and monitor investments across the globe. However, this exercise requires too many resources (man-hours) because it is necessary to match the different databases to guarantee that companies' consolidated information is accurate. Also, the entity does not have the operational capacity to incorporate a more sustainable model for executing this activity.

Taking this into account, the project objective is to improve the way in which foreign companies that have a high investment potential are being selected by PROCOLOMBIA, so the success rate may be increased and leave a more efficient process for the company. It is worth mentioning that in the actual context, obtaining foreign investments for Colombia has a direct impact in economic reactivation, fostering unemployment and poverty reduction, as well as improving the national production efficiency.

4. Dataset Description

Different sources of information were provided by *PROCOLOMBIA*, and below there is a table with a general description:

ID	Source Name	General description	File Size (KB)	Number of rows
1	Gazelle	List of international companies with an associated score indicating the investment potential	1,785	3,311
2	Base Exportaciones	Exports executed by Colombia from 2014 to 2020	412,131	847,703
3	Base Importaciones	Colombia's imports from 2018 to 2019	364,173	1,043,701
4	DIAN	CIF (Cost, Insurance and Freight) and Kg reported by DIAN from 2014 to 2020	11,445	107,310
5	Segmentacion_2020	Shortlist of american companies with investment potential	2	10
6	FdI Markets	Descriptive information about National and International companies performing projects from 2012 to 2017	33,864	87,417
7	Empresas_Extranjeras_USA	American companies based in Colombia	30	75
8	Veritrade	Foreign trade database for different countries (Colombia, Ecuador, Mexico, Panama, Paraguay, Peru)	1,500 (on average per file)	23,942 (on average per file)
9	Orbis	Risk and financial metrics for American Companies	55,243	98,827
10	Base instalados	Foreign companies installed in Colombia	1,526	4,034
11	Empresas Inversionistas en Colombia	Companies that have invested in Colombia since 2010	61	1,227
12	Sociedades extranjeras matriculadas en Colombia	Foreign companies based in Colombia	347	1,919
13	New FdI Markets	Additional information of FdI Markets with most of the information related to projects from 2018 to 2020	6,231	16,306
14	New Orbis	Additional information of Orbis	575,956	1,148,588

Table 1. Datasets summary

Datasets 1 to 9 were provided at the beginning of the project, and the remaining datasets were provided during the project lifetime.

5. Exploratory Analysis

The team performed an exploratory data analysis (EDA) of all datasets that were delivered by *PROCOLOMBIA*. The first approach was to assess all the datasets individually and execute a preliminary data understanding and cleaning, including format fixes, dropping unnecessary columns and name field standardization.

After this initial review, three (3) datasets were considered as the most relevant for further analysis: Gazelle, FdI Markets and Orbis. The main reason is that these datasets contain global information and provided some financial metrics that were considered useful for evaluating foreign investments and/or trade opportunities for Colombia. Due to this reason, only the details of the mentioned datasets are registered in this section.

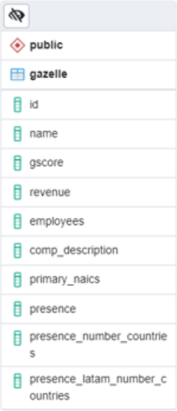


FIGURE 1. GAZELLE ENTITY DIAGRAM

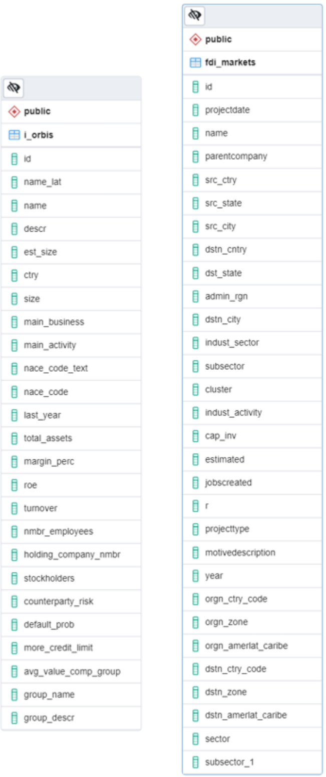


FIGURE 2. ORBIS AND FDI MARKETS ENTITY DIAGRAM

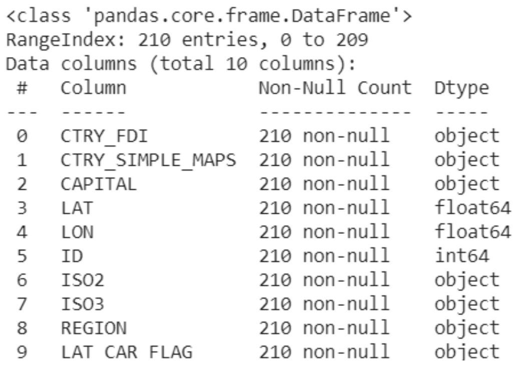
After uploading the selected databases, the approach was to join the related datasets to have a unique data source to perform further analysis. However, it was identified a lack of key fields to perform the join. To overcome this issue, the team took the company name as the main field to execute the join, and after several tries of cleaning and using regex functions to improve the matching, the results were not satisfactory, only maintaining less than 1% of the records in best case scenario for FdI Markets and Orbis:

Total Rows		87,417	2,377	98,827	
	Dataset	**FdI Markets**	**Gazelle**	**Orbis**	**Join Rate**
87,417	FdI Markets	x	656	230	0.8%
2,377	Gazelle		x	77	3.2%
98,827	Orbis			x	
	Join Rate		28%	0.2%	

TABLE 2. JOIN RATES OF GAZELLE, FDI MARKETS AND ORBIS

Due to this situation, the team decided to use only FdI Markets dataset because it was identified as the most meaningful of the first selection to build a descriptive dashboard. In addition to that, it is important to mention that a FdI Markets additional dataset was provided by PROCOLOMBIA, and it was identified that there were new fields and different timeframes comparing with the first FdI Markets dataset.

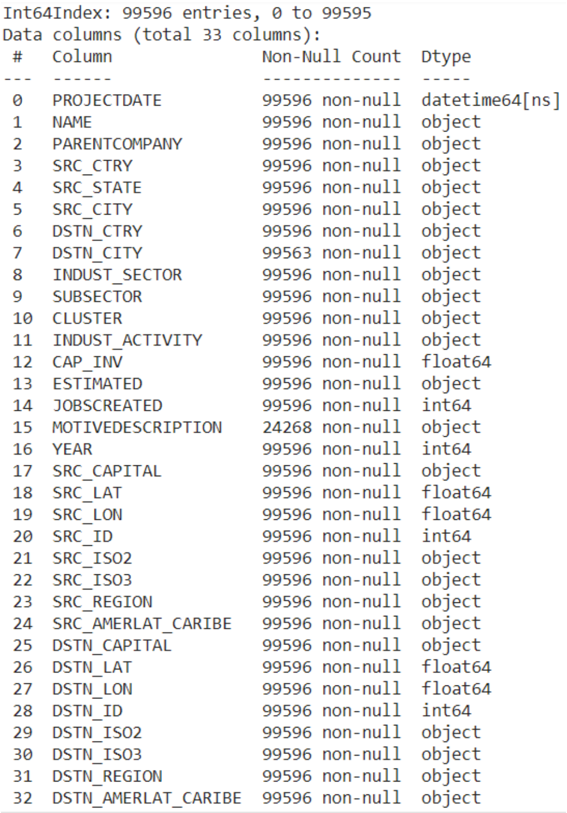
For homologating both datasets, it was required to create a new reference file with information related to georeferenced positions, ISO codes for countries and a flag for determining if the country belongs to Latin America and Caribbean region. It was built with public data from simplemaps.com/world and it allows to have all the FdI Markets countries mapped.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CTRY_FDI         210 non-null    object
 1   CTRY_SIMPLE_MAPS 210 non-null    object
 2   CAPITAL          210 non-null    object
 3   LAT              210 non-null    float64
 4   LON              210 non-null    float64
 5   ID               210 non-null    int64
 6   ISO2             210 non-null    object
 7   ISO3             210 non-null    object
 8   REGION           210 non-null    object
 9   LAT CAR FLAG     210 non-null    object
```

FIGURE 3. INFORMATION OF COUNTRY GEOREF FILE

After the homologation, the FdI Markets final dataset used to build the descriptive dashboard has the following fields and characteristics:



```
Int64Index: 99596 entries, 0 to 99595
Data columns (total 33 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   PROJECTDATE          99596 non-null  datetime64[ns]
 1   NAME                 99596 non-null  object
 2   PARENTCOMPANY        99596 non-null  object
 3   SRC_CTRY             99596 non-null  object
 4   SRC_STATE            99596 non-null  object
 5   SRC_CITY             99596 non-null  object
 6   DSTN_CTRY            99596 non-null  object
 7   DSTN_CITY            99563 non-null  object
 8   INDUST_SECTOR        99596 non-null  object
 9   SUBSECTOR            99596 non-null  object
 10  CLUSTER              99596 non-null  object
 11  INDUST_ACTIVITY      99596 non-null  object
 12  CAP_INV              99596 non-null  float64
 13  ESTIMATED            99596 non-null  object
 14  JOBSCREATED          99596 non-null  int64
 15  MOTIVEDESCRIPTION    24268 non-null  object
 16  YEAR                 99596 non-null  int64
 17  SRC_CAPITAL          99596 non-null  object
 18  SRC_LAT              99596 non-null  float64
 19  SRC_LON              99596 non-null  float64
 20  SRC_ID               99596 non-null  int64
 21  SRC_ISO2             99596 non-null  object
 22  SRC_ISO3             99596 non-null  object
 23  SRC_REGION           99596 non-null  object
 24  SRC_AMERLAT_CARIBE   99596 non-null  object
 25  DSTN_CAPITAL         99596 non-null  object
 26  DSTN_LAT             99596 non-null  float64
 27  DSTN_LON             99596 non-null  float64
 28  DSTN_ID              99596 non-null  int64
 29  DSTN_ISO2            99596 non-null  object
 30  DSTN_ISO3            99596 non-null  object
 31  DSTN_REGION          99596 non-null  object
 32  DSTN_AMERLAT_CARIBE  99596 non-null  object
```

FIGURE 4. INFORMATION OF FINAL FDI MARKETS DATASET

In addition to the descriptive dashboard based on the datasets delivered by PROCOLOMBIA, the team decided to pivot when the join didn't result as expected and work in text analytics from Twitter selecting the most relevant countries for the models. Sample data was extracted to build a word cloud related to keywords as "INVERSIONISTA", "INVESTMENT", "INVESTOR" with "COLOMBIA", obtaining the following result:

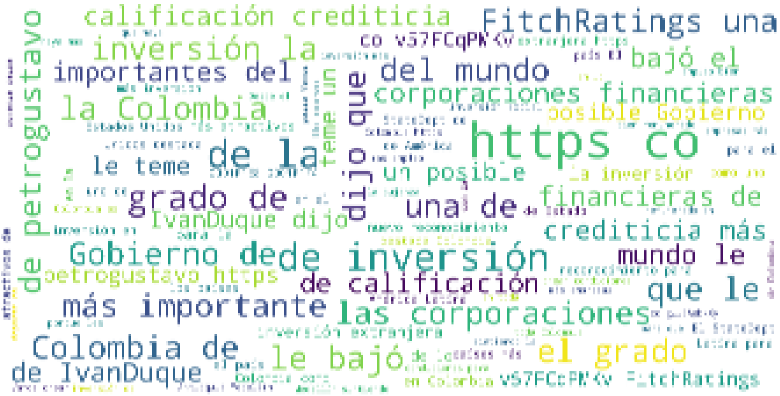


FIGURE 5. INITIAL WORD CLOUD

6. Model

The selected models are unsupervised learning approaches and based on text analytics from Twitter, where the main outputs are descriptive information (word clouds, bigrams, and trigrams), a sentiment analysis and a topic modeling. The selected countries for extracting data were United States, United Kingdom and Canada, as they are the English-spoken most relevant countries for Colombia. Further details can be found below, as well as the previous data processing required.

Data extraction and pre-processing

As was mentioned before, Twitter is the data source for all the analysis. The extraction was performed using a python library called Tweepy, that allows to access the Twitter API requesting an authentication process and afterwards, it is possible to download tweets according to certain parameters like language, location, keywords, among others. It is important to mention that Tweepy has some limitations, and it only allows to retrieve tweets up to 7 days ago and extract a maximum of 18,000 tweets per a fifteen-minute window.

The keyword that was selected as criteria for tweet extraction was “Colombia”, and it is worth mentioning that code is easily adjustable to replace or include more keywords to customize the extraction. Once the selected countries tweets have been extracted, the next step was to build a corpus, that is a text collection (in this case a tweet collection) ready for executing the analysis and modeling. Thus, it was required to perform a text standardization and cleaning process. This involved the following actions over the text from the tweets:

- Stop words (words commonly used in the language and not adding meaning to the text)
- Special characters (e.g. “!”,“#”)
- Links and URLs (tweets may contain links that are not part of the text)
- Remove retweets (duplicated data)
- Convert all text to lower case characters

Text descriptive analysis

The first analysis after having the corpus was to have an initial descriptive analysis of the combination of words that were more common in the data that was extracted. For this, there were functions to generate n-grams, that are the sequence of n words in a text (tweet) where n goes from 1 to 3, to have initial insights of the most common words that are being used filtering all tweets that mention “Colombia”.

Sentiment Analysis

A sentiment analysis is a process of determining if the data, in this case text (tweets), are positive, neutral, or negative. The idea is to obtain insights from this classification to help PROCOLOMBIA monitor the sentiment about Colombia in the selected countries (United States, United Kingdom, and Canada), and have additional information to identify investment opportunities.

VADER (Valence Aware Dictionary and sEntiment Reasoner) was the selected tool to perform the analysis, and it is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto, 2021) built in python language. It is worth mentioning that the tool is sensitive both the polarity and the intensity of sentiments expressed in social media contexts.

For this purpose, VADER uses a normalized, weighted composite score (between -1 and 1) to determine a measure of sentiment for a given text, and the typical threshold values are:

1. Positive sentiment: compound score >= 0.05
2. Neutral sentiment: (compound score > -0.05) and (compound score < 0.05)
3. Negative sentiment: compound score <= -0.05

Topic Modeling

A topic modeling is a technique that allows to detect word and phrase patterns to determine clusters composed of words that have similarities and can be associated to a same topic. In the project's framework, this allows to identify topics that are related to investment opportunities in Colombia, and then deep dive into the most common words per topic for assessing if there are additional insights (e.g. sector, specific locations, economic situation).

For performing this analysis, the Latent Dirichlet Allocation ("LDA") model was selected, and it is widely used in Natural Language Processing ("NLP"). The idea behind LDA is that the model assumes that each document is a mix of topics, and each topic is a mix of words. An advantage is that one does not have to know in advance what the topics will look like. By tuning the LDA parameters to fit different dataset shapes, one can explore topic formation and resulting document clusters. (Zhao & Cen, 2014).

The parameters that can be modified in LDA model are:

- 'α': document-topic density factor – number of topics expected in the document (tweet)
- 'β': topic-word density factor – distribution of words per topic
- 'K': number of topics to be considered

Additionally, to have more clarity of the steps followed by an LDA model, a diagram is shown below:

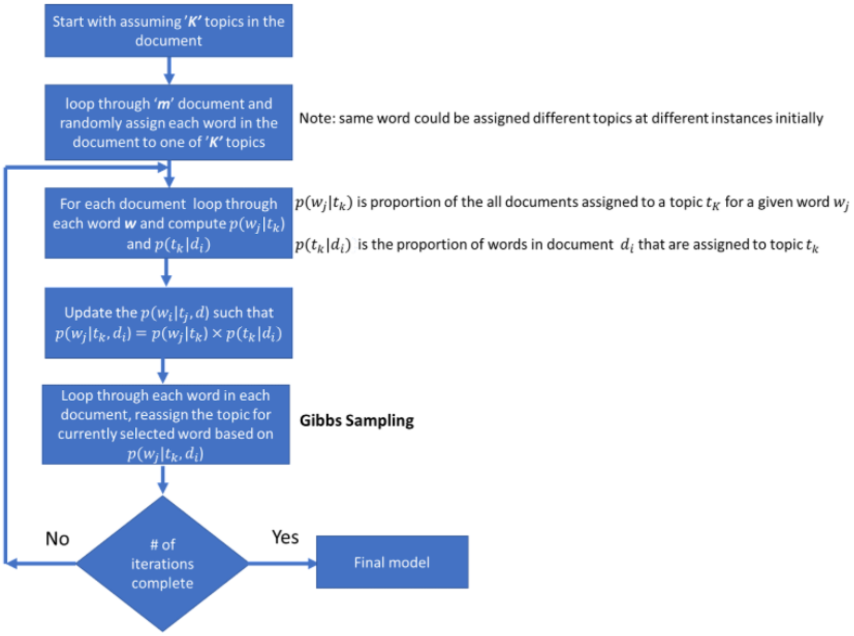

Start with assuming '*K*' topics in the document
loop through '*m*' document and randomly assign each word in the document to one of '*K*' topics
Note: same word could be assigned different topics at different instances initially
For each document loop through each word *w* and compute $p(w_j|t_k)$ and $p(t_k|d_i)$
$p(w_j|t_k)$ is proportion of the all documents assigned to a topic t_K for a given word w_j
$p(t_k|d_i)$ is the proportion of words in document d_i that are assigned to topic t_k
Update the $p(w_i|t_j,d)$ such that $p(w_j|t_k,d_i) = p(w_j|t_k) \times p(t_k|d_i)$
Loop through each word in each document, reassign the topic for currently selected word based on $p(w_j|t_k,d_i)$
Gibbs Sampling
No
# of iterations complete
Yes
Final model

FIGURE 6. GENERAL STEPS IN LDA (Great Learning Team, 2020)

7. Backend and dashboard Implementation

Backend

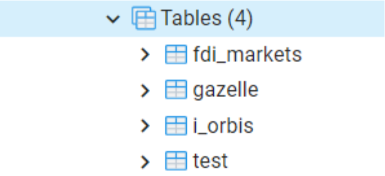
The backend was built using the free tier of Amazon Web Services (AWS). There were two services that were configured during the project and are explained below:

RDS

RDS is the relational database service offered by AWS. The team decided to create a Postgres database engine to upload the information related to the most relevant datasets delivered by PROCOLOMBIA (Gazelle, Orbis, fDi Markets). For this purpose, the configuration that was selected in AWS is as follows:

- DB instance class: Burstable classes with 1 vCPU and 1GiB RAM
- Storage: General purpose (SSD) with 20GiB of allocated storage

After creating the instance, pgAdmin was used for stablishing a connection with the database and to manage, create the tables and upload the datasets:



- Tables (4)
  - fdi_markets
  - gazelle
  - i_orbis
  - test

FIGURE 7. TABLES CREATED IN RDS INSTANCE

It is important to mention that this implementation was performed in parallel with EDA activities, and as it was mentioned in Exploratory Analysis part, it was thought for stablishing the join between datasets and work in further analysis with the merged information. Since the results of the join between datasets was not satisfactory, RDS instance is functional but does not take part of the final backend implementation.

EC2

EC2 stands for Elastic Compute Cloud and is a virtual server instance that can be used to run applications in AWS infrastructure. Configuration details can be found below:

- t2.micro
- Ubuntu Server 20.04 LTS (HVM), SSD Volume Type – 64 bits (x86)
- 1 vCPU, 2.5 GHz and 1GiB RAM
- Elastic IP address: 3.14.187.26 (maintain the same address even when server instance is down and restarted)

The idea of having this server instance is to run all the processes and models associated to the dashboard and maintain the dashboard online. In addition to that, MobaXTerm was used for connecting with EC2 and manage all the files supporting the descriptive visualizations, model execution and results, and code related to Dash application (dashboard).

For performing these tasks, it was required to install the following Python libraries to run the processes:

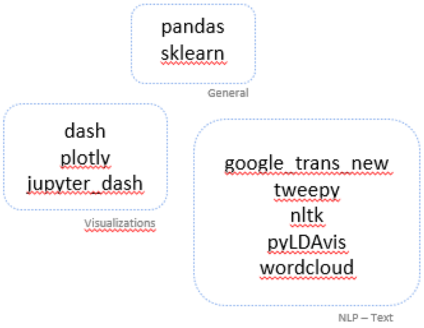


FIGURE 8. PYTHON LIBRARIES INSTALLED IN EC2

Moreover, there were some relevant decisions that were taken in the backend, and that directly affect the reflected results in the dashboard:

- Installing the screen command, which allows to use multiple shell windows from one SSH session and keep a shell session active for long times (even when there are network disturbances). This allows to have the dashboard running and work after a disconnection without losing the progress.
- Programming a daily process in which the Twitter data extraction occurs at a determined time. This allows to refresh the data and run the models to obtain updated information.
- Generate the model results as .html files, to further uploading in Dash and maintain the interactive features in the dashboard.

Dashboard

The dashboard was built in Dash, and it was created to show the project results in an online application. From the technical view, Cascading Style Sheets (CSS) were used to design the layout of the content that is been displayed in the dashboard, as well as selecting colors and fonts. Additionally, callbacks were used to create filters and allow the possibility of having interactive features.

The dashboard has a main menu on the left side where the user can find the following tabs:

Statistical analysis

As was mentioned before, the descriptive visualizations are based on FdI Markets merged dataset and is composed of:

- Global heatmap representing the capital investment (in million USD) from the source country. A year slicer allows to select a specific timeframe.
- Scatter chart with sector as y-axis and capital investment in x-axis (in million USD) to identify the most relevant sectors. It can be filtered by source country.
- Scatter chart with sector as y-axis and jobs created in x-axis. A capital investment (in million USD) legend that changes color according to the value is also included to relate the two metrics mentioned before. It can be filtered by source country.
- Bar chart where countries are in x-axis and capital investment (in million USD) in y-axis. Each country has two bars, one for outward and one for inward, in order to have a deeper understanding of the relevance of each country and its foreign direct investment balance. A destination region filter was included for selecting specific countries.

Model

This tab shows all the model results using Twitter data and it has additional tabs for selecting each of the countries that were chosen for the analysis (United States, United Kingdom, Canada). Once a country is selected, the dashboard shows the following options:

- Descriptive analysis: first graph is a word cloud with the most common words of the tweets for giving a first glance of the information. Below, there are bar charts with top unigrams, bigrams and trigrams for each relevant country where the user can identify different trends based on 1-word, 2-word and 3-word combinations.
- Sentiment analysis: there are two bar charts. On the left side there is a histogram showing the frequency of score assigned to each tweet. This gives a measure of the intensity of the sentiments. On the right side, a bar chart shows a positive, neutral, and negative sentiment tweet count.
- Topic modeling: results are shown in an interactive bubble graph where tweets are grouped by topic and selecting one of the bubbles will show the top-n relevant words per topic on the right side.

About us

- Team-15 members with LinkedIn url.

For summarizing, an architecture diagram can be found below:

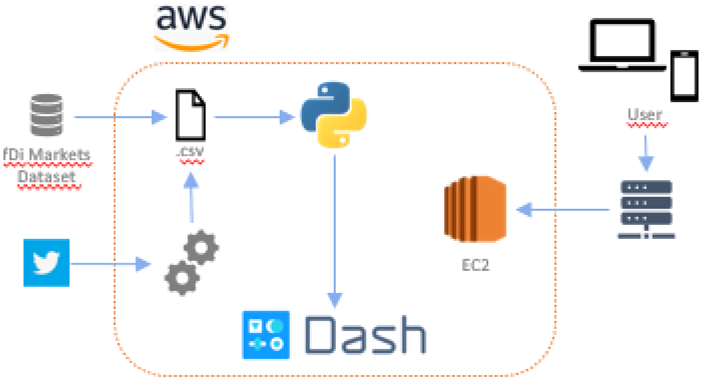


FIGURE 9. ARCHITECTURE DIAGRAM

8. Results

The results were registered in the application, and can be divided in two streams according to its data source:

fDi Markets

fDi markets worked as a data source for building visualizations related to global investments and get insights of which countries are the most relevant around the globe. In addition, it is possible to select a time window with the slicer to identify trends and see the distribution in recent years:

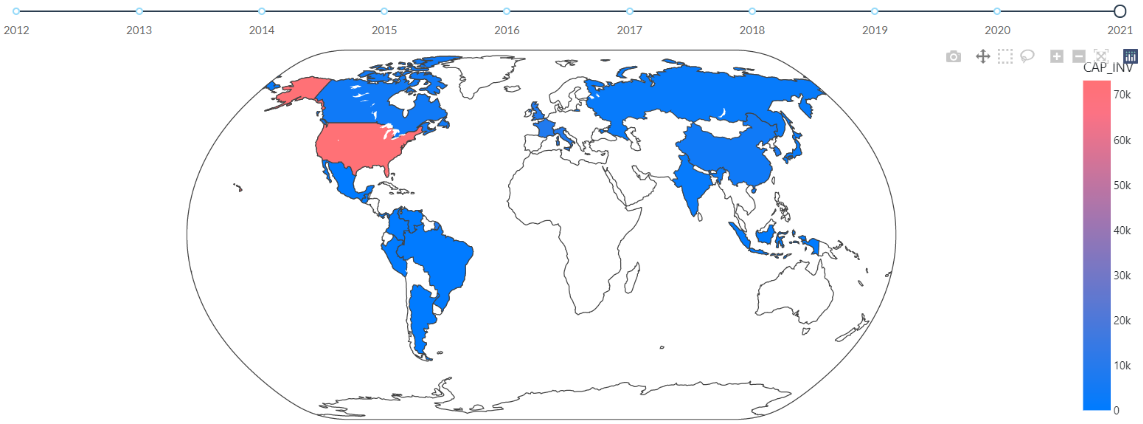
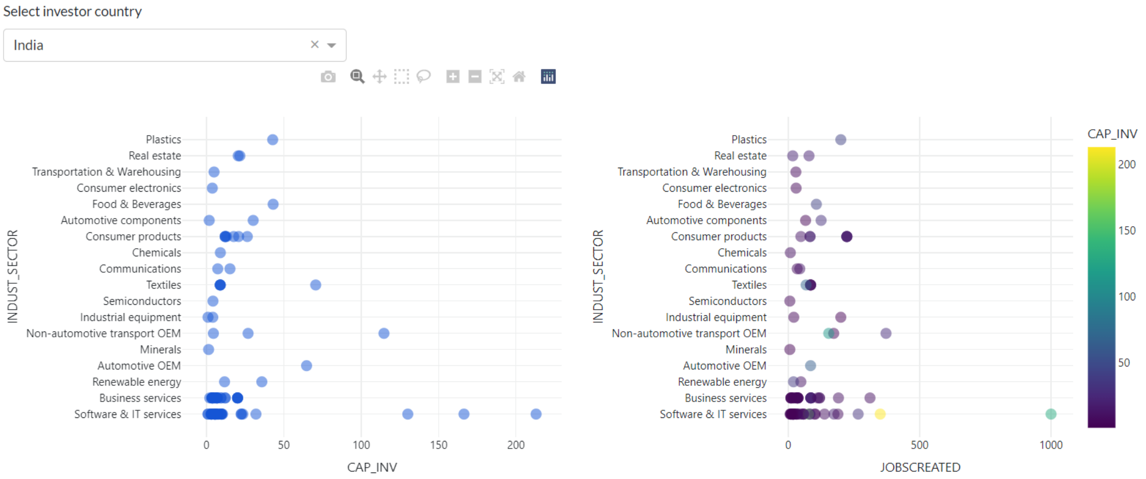


FIGURE 10. GLOBAL HEATMAP WITH CAPITAL INVESTMENTS (IN M USD)

Besides, there are two scatter charts that allows to identify the investments the capital investment and jobs created by sector, and allows to filter by the source country (investor):

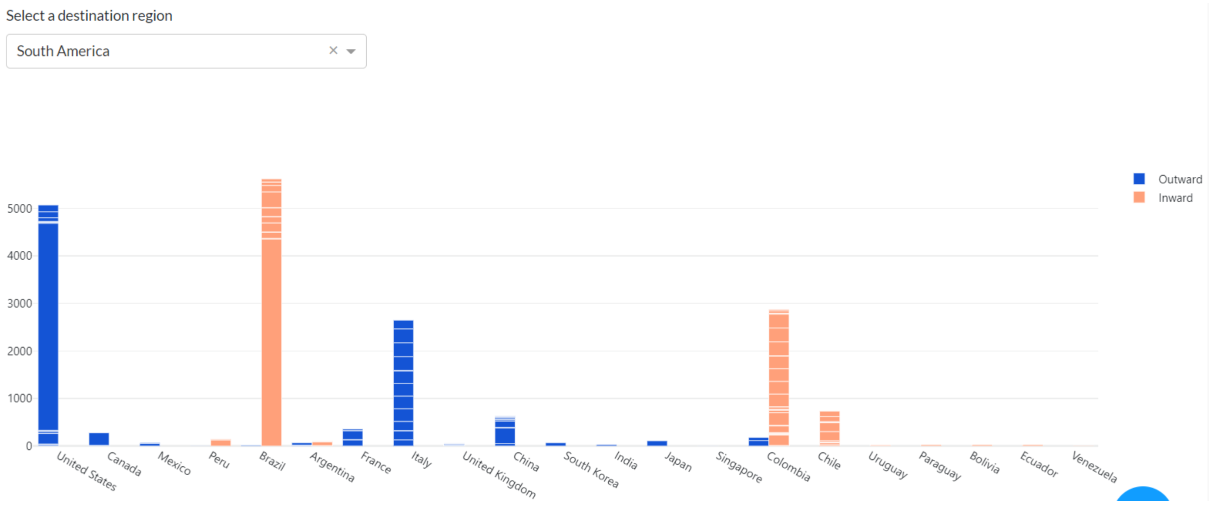


Select investor country

India

FIGURE 11. SCATTER CHARTS WITH CAPITAL INVESTMENT (IN M USD) AND JOBS CREATED BY SECTOR

The last results related to fDi Markets is related to the capital inward and outward flows by country:



Select a destination region

South America

FIGURE 12. BAR CHART WITH CAPITAL FLOWS BY COUNTRY

With all the visualizations described above, United States is the country that has the highest foreign direct investment in the world. Also, it can be confirmed that the English-spoken most relevant countries following United States are Canada and United Kingdom. However, it also shows that Italy is getting relevant for the region, being the second with most outward flows with South America as destination region, and China is also showing signs of increasing foreign direct investment.

Twitter based

There are results per each selected country. Below, Canada's results can be observed:

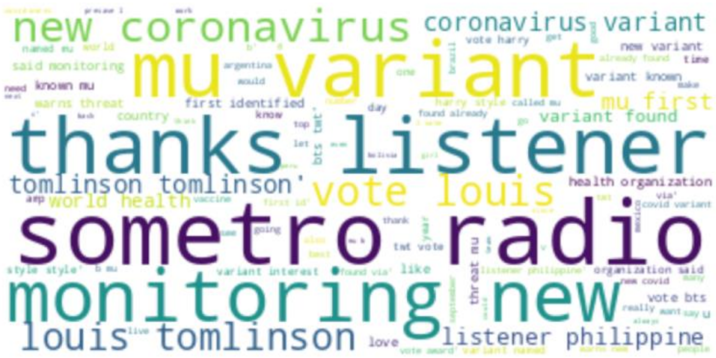


FIGURE 13. CANADA WORD CLOUD - TWEETS FROM SEPTEMBER 1ST

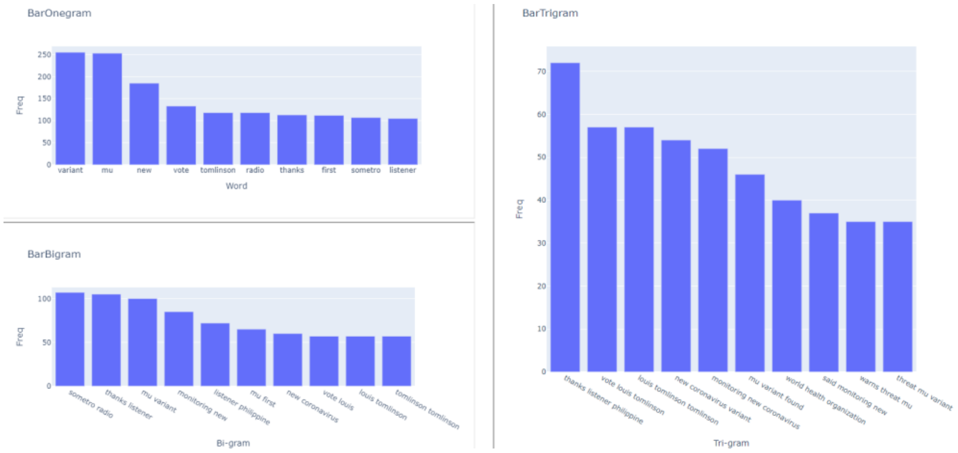


FIGURE 14. CANADA'S BAR CHARTS WITH N-GRAMS (N FROM 1 TO 3) – TWEETS FROM SEPTEMBER 1ST

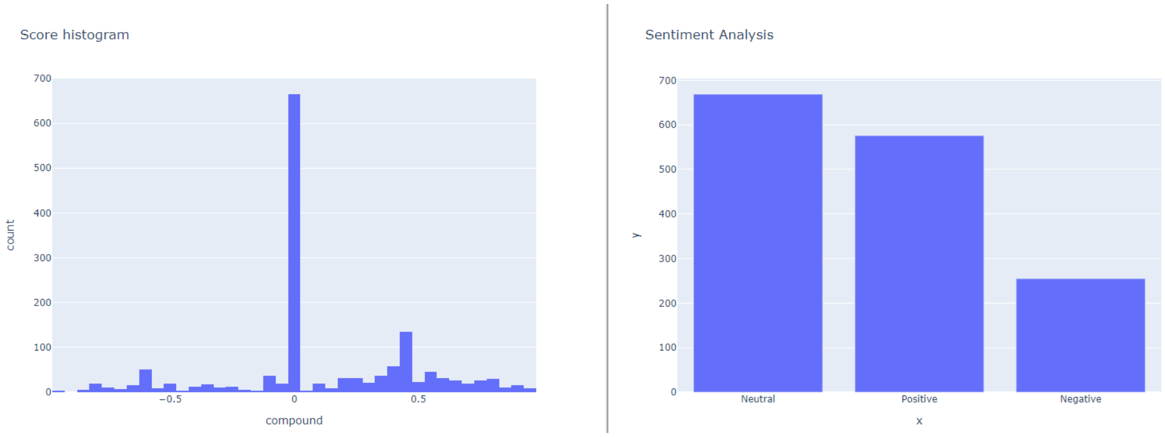
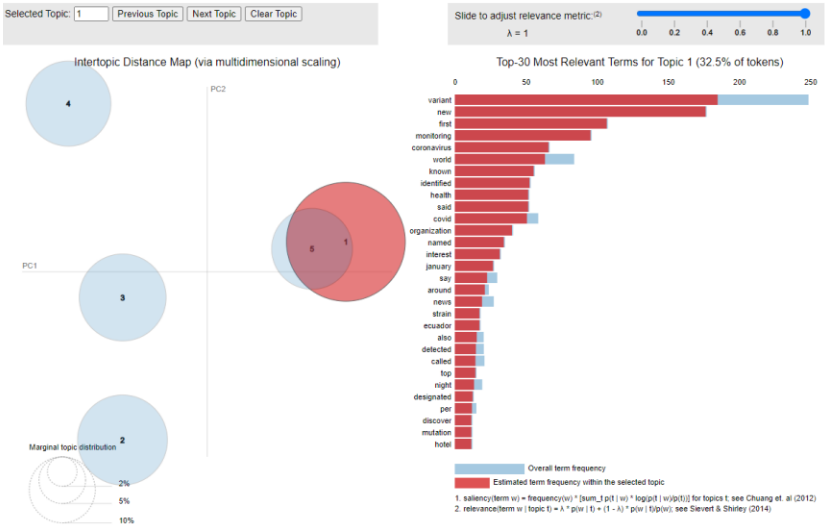


FIGURE 15. CANADA'S SENTIMENT ANALYSIS WITH COMPOUND SCORE – TWEETS FROM SEPTEMBER 1ST



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

FIGURE 16. CANADA'S TOPIC MODELING – TWEETS FROM SEPTEMBER 1ST

The same visualizations can be obtained for United States and United Kingdom, and regarding the modeling, different conclusions may arise depending on Twitter extraction. For the case of Canada with tweets from September 1st, it was observed that the mu variant, related to a new Coronavirus case identified in Colombia, was the main topic in this social media with "Colombia" as keyword, and that the sentiment about Colombia was neutral, followed by positive tweets, that may be influenced by music and sports related topics.

To conclude, the model was defined to give a general overview of the sentiment and topics that are been discussed in other countries about Colombia. However, the user may tune and select different keywords to refine the tweet selection and obtain results focused on specific activities.

The dashboard is available online in the following link:

http://3.13.197.145:8010

In addition, all the relevant information can be found in the following GitHub repository:

https://github.com/darielr/PROCOLOMBIA_TEAM15

References

- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- https://docs.tweepy.org/en/stable/
- https://developer.twitter.com/en/docs/twitter-api/v1/rules-and-filtering/search-operators
- https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/