SCHOOL OF HUMAN SCIENCE & TECHNOLOGY

SOCIAL NETWORK ANALYSIS | IEHST

23F NLP SOCIAL NETWORK ANALYSIS

TWITTER'S API TEXT MINING ANALYSIS
SECTION 2 GROUP 3

Table of Contents

1. Introduction

1. Why tweets related to the 23F?

This study began with an exploration of the ego networks of the different Twitter accounts of youth political leaders in Spain. The goal was to see to what extent their networks were truly meaningful, analysing among other matters, the transitivity of their networks, as well as seeing the communities formed among them. However, due to time and resources constraints, it could not be performed.

The next proposal was to perform a text mining and SNA analysis of the tweets posted during the different protests and demonstrations in favour of Pablo Hasel, a Spanish rapper who was just convicted to prison due to several charges related to insults to the king of Spain and threatening of witnesses. However, the analysis failed due to a late extraction of tweets, which resulted in unrelated words and thus poor networks with very low modularity among their communities.

The final decision and thus, the object of this study was to perform a similar analysis as the one proposed with the tweets related to Pablo Hasel, but with the tweets posted during the 23rd of February of 2021. This day is an important date in Spanish history since it's the 40th anniversary of the attempt of coup d'état by Antonio Tejero, also known as the "Tejerazo". The interest in analysing the different Twitter conversations related to this day arises not only from the importance of the day but also due to the current situation of political polarization and discontent lived in Spain, which is exacerbated by the global pandemic.

2. 23rd of February of 1981

The 23rd of February of 1981 also known as 23-F or the "Tejerazo" was an attempt of coup d'etat performed by Lieutenant-Colonel Antonio Tejero and 200 armed Civil Guard officers who broke into the Congress of Deputies during the vote to elect a President of the Government.

Tejero and his officers held hostage during 18 hours the parliamentarians and minister's hostage for 18 hours. The King Juan Carlos I addressed the nation in a televised speech in which he denounced the coup and called for rule of law and the democracy to continue. Tejero surrendered the next morning and set free the hostages. There were no killings.

During the following years, several conspiracy theories have developed around this event, and to date there is still debate of what really happened on that day. As a result, and due to the current political discontent, it was expected that Spaniards would use social media to express their opinions about such date.

2. Languages used

1. Tweet's extraction – Python

Given the Twitters API (Application Programming Interface) flexibility, our team had the opportunity to use one of python's streaming packages that filters tweets containing a specific track word in the text or a particular hashtag. We noticed that by using python there was no alteration to the text and special characters readability was better than using R. Also, on top of that, python incorporates various packages that structure the information easily from JSON files. This was a major step after compiling the tweets information into three different columns with the aid of "pandas". We could have also used "DPLYR", but we encountered difficulties in the segmentation of different columns with this package.

To extract the tweets from the 23F, our tweets extraction filtered the following words "#reyjuancarlos, #23F, #golpedeestado, #rey, #juancarlos, #españa". These words in the track section from our "Tweets_Ingestion_23F" .py file, track all mentions of a particular domain name. For instance, Twython streamer will ingest those tweets that match the track words, such as "#23F" or "#reyjuancarlos". Once the program reaches the limit set on "self.disconnect", the program dumps all the tweets into a JSON file and waits until it is executed again.

Our EDA has been segmented in different ".py" files. Each ".py" file extracts different fields from our compiled JSON file and executes different data manipulation techniques to produce the final graphical representation in our EDA. It is therefore that for understanding purposes, instead of cramming all the code into one ".py" file, we decided to assign different files with specific purposes. Further enhancing our understanding of the code.

2. Construction of network and identification of metrics – R

To convert the tweets into nodes, edges and assign a weight to each of the relationships we used Dr Juan Camilo Orduz's method. (analysis of Colombia's 2016 Plebiscite Tweets analysis - available on his GitHub.[1])

Here are the steps we followed:

1. Data reading from xlsx file.
2. Text normalization to remove stop words and weird characters
3. Word count
4. Network Analysis
   a. Bigram Analysis: bigram analysis is a frequent text mining method by which a diagram is formed by the sequence of two adjacent elements from a string. Meaning that we find the words that appear next to each other in the text to construct the nodes and edges.
      i. Network definition
      ii. Visualization of the Network
   b. Skip-gram Analysis: Skip-gram analysis is an unsupervised learning technique used to find the most related words for a given word.
      i. Network definition
      ii. Visualization of the Network
5. Node importance –centrality measures
6. Community detection
7. Correlation Analysis
   a. Network definition.
   b. Visualization of the Network

[1] https://juanitorduz.github.io/text-mining-networks-and-visualization-plebiscito-tweets/

The whole script can be accessed on the R Markdown attached to this notebook.

3. Exploratory data analysis

1. Overview

In the Social Network Analysis field, the lack of a contextual analysis before stepping into the plotting stage of a network model can be daunting at first sight. However, developing an exploratory data analysis beforehand, can facilitate the interpretability of the model and reinforce some insights observed in the graphical analysis. In this report, we want to go a step further and focus part of our investigation on additional social media data within the tweets. Although this EDA is barely able to scratch the surface of the data provided by the Twitter API, we will be able to go deeper on diverse topics, such as the computation of the average polarity and subjectivity of our topic, the popularity of different user mentions and words and the most popular hashtags. Gathering and assessing all this data will give us a holistic approach to our tweets ingestion and will give us a better understanding of how we should tackle the network model effectively.

As previously mentioned, we will answer the following questions in this segment of our investigation:

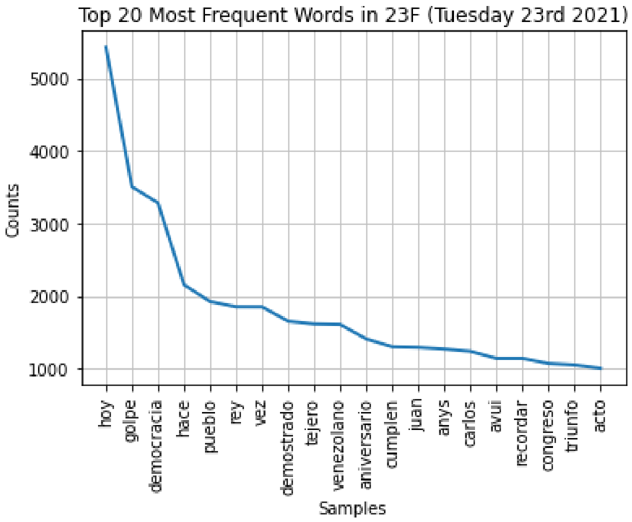
- Which are the top 20 most frequent Words?
- Which are the top 18 most frequent Hashtags?
- Which are the top 10 most Twitter Authors?
- Which are the top 10 User Mentions?
- Which is the most influential tweet?
- What is the average polarity, subjectivity score, and their corresponding distributions?

In each subsegment, we will be explaining the main insights from the graphs and giving a brief description of our coding process. Finally, we will gather all the different insights and suggest some amendments to data before deploying our network model.

2. Exploratory Data Analysis

(1) Most Popular Words

To return this powerful insight, we followed these different steps. As you can see in our file "Diagrams_23F.py", we started by executing a filtering process to extract all the different JSON documents and combine them into a list. Once the extraction was done, we pulled the tweets' text and segmented all the different words using a split function. Consequently, we eliminated stop words in various languages, punctuations, and words related to our filtering tweets extraction. You can see the results of this process on Figure 1.



The following insights can be extracted from figure 1. First, words have a positive connotation towards the 23F. It is a day of celebration which is reminiscent to recall. Words such as "Triunfo", "Aniversario", "Cumplen" and "Democracia" emphasize the importance of this date for the Spanish population. Secondly, some words are in Catalan ("Anys" and "Avui"), we suppose that twitter users from the east of Spain are more active with regards to the 23F. This is interesting, however, there are no fields in the twitter API to identify each tweet's geolocation. This would have allowed us to understand the proportion of tweets in different Spanish regions and determine which one is more active with regards to political social network activity. Finally, we noticed a strange movement with "Venezuela" in the 23F tweets. We will further analyse if there is an actual relation to the Spanish 23F or not in the next subsegments.

(2) Most Popular Hashtags

In this subsegment, we followed a similar approach as in the previous question, but instead of extracting the text of the tweets, we extracted the hashtags from the entities JSON document section. From that point onwards, we followed the same steps as above, until we were surprised by different hashtags related to Venezuela, Japanese topics, and Chinese popular hashtags. It is, therefore, that we decided to remove all these unrelated hashtags before producing the graphical frequency distribution. Figure 2 illustrates our result.

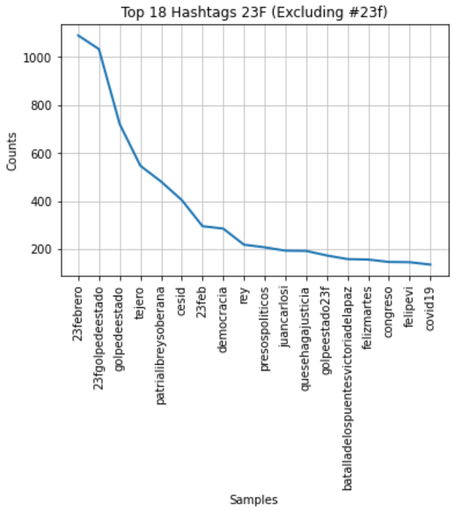


Figure 2 shows us a mix of different hashtags which can be split into two different movements, the 23F in Spain and the 23F in Venezuela. Hashtags such as "Patrialibreysoberana", "Presospoliticos", "Batalladelospuentesvictoriadelapaz" are considered noise in our analysis and do not represent the actual 23F in Spain. Therefore, we will need to remove those Venezuelan tweets which do not add any significance to our analysis. Other than that, the hashtags for the Spanish 23F are focused on remembering the day and stating important figures like "Tejero","Juancarlosi" and "FelipeVI". No different discussions ideologies or discussions are observed in the 23F, it is very pacific, and it is an honourable day for most Spaniards.

(3) Most Frequently Mentioned User Mentions

As it has been observable in the other subsegments above, user mentions are not an exception to gather information from the 23F in Venezuela and Spain. The same procedure has been followed as the one defined in “most popular hashtags”, yet instead of gathering hashtags, we gathered user mentions. Figure 3 highlights the following information.

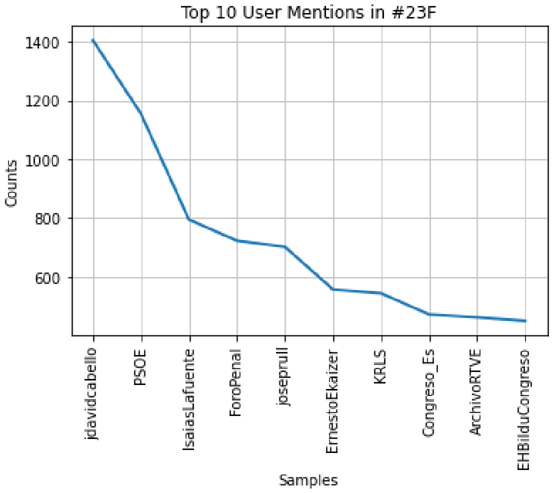


Figure 3 shows a strong user mention connection between Spanish users. Although the top 1 user mentions are related to Venezuela, 8/10 of the top 10 user connections are related to Spanish 23F. This means that even if there is a mix between Spanish users and Venezuelan users, Spanish accounts tend to be more active with regards to social networks as they tend to have higher activity than Venezuelan users. There is higher interconnectivity between accounts, which explains the sparsity in the different user mentions compared to Venezuelan user mentions. Another reason why this could be is the time zone difference. We gathered tweets from 9 am until 10 pm Spanish hour, this means that at the same time we were gathering data from the Venezuelan region from 4 am until 5 pm. Maybe as we weren’t covering the whole-time factor for the two different countries, we had more Spanish user mentions than Venezuelan user mentions.

(4) Most Frequently Mentioned Tweet Authors

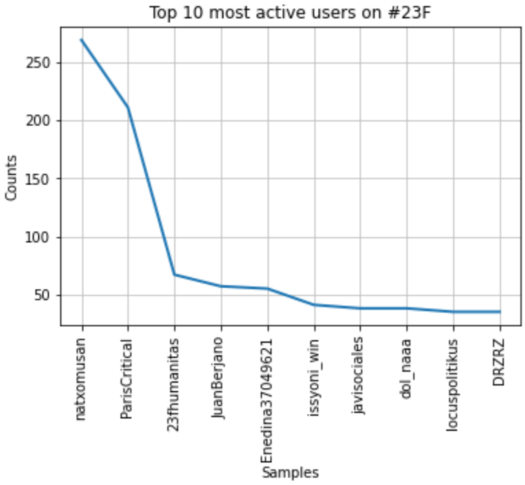
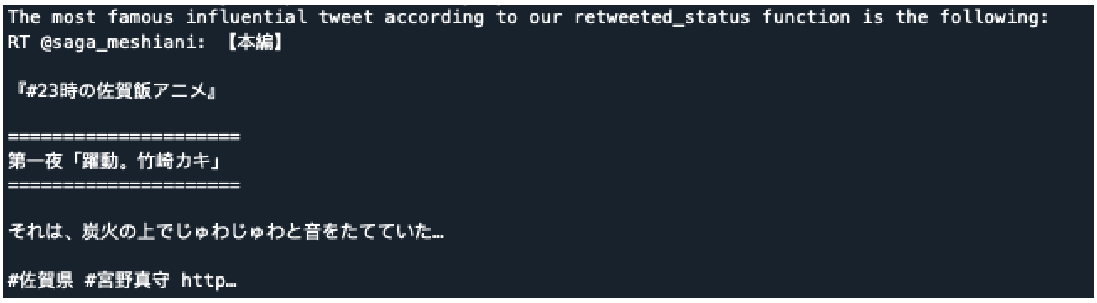


Figure 4 shows a higher activity in Spanish accounts than Venezuelan accounts with regards to the 23F. This completely relates to what has been previously explained in subsection C. Spanish networks as seen as highly reactive and connected in the #23F. An interesting point that hasn't been seen in subsection C, is the different reactions towards the 23F in Spain. Although it was observed in subsection A that this day was highly celebrated by most Spaniards, this isn't completely true. Other profiles such as "natxonmusan" are highly against this celebration movement and propose a different perspective on the network.

(5) Most Influential Tweet from the #23F

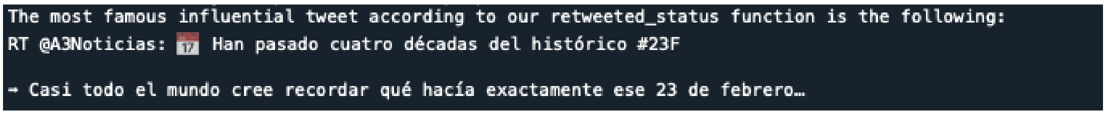
We took a different approach in this subsection. We counted the number of retweets of different posts as well as the ones with the most replies and highly favourited. By doing so and adding all the fields mentioned above, we can compute the "influence score" of every single tweet. Meaning that we can highlight the most influential account in the specific # that we have targeted. The most influential tweet in the 23F is @A3Noticias, a famous broadcasting news account well known in Spain. However, before obtaining that result, we had an unexpected Japanese tweet which was considered the most influential tweet. Since there were not only Venezuelan tweets interfering in our analysis, but we also decided that before doing some graphical analysis with our network models we should filter all those tweets that weren't related to the topic that we are analysing. By doing this we will manage to eliminate the noise from our different network models. Hence, having better interpretability on how network segments the different communities and higher modularity.

Before:



```
The most famous influential tweet according to our retweeted_status function is the following:
RT @saga_meshiani: 【本編】

『#23時の佐賀飯アニメ』

======================
第一夜「躍動。竹崎カキ」
======================

それは、炭火の上でじゅわじゅわと音をたてていた…

#佐賀県 #宮野真守 http…
```

After:

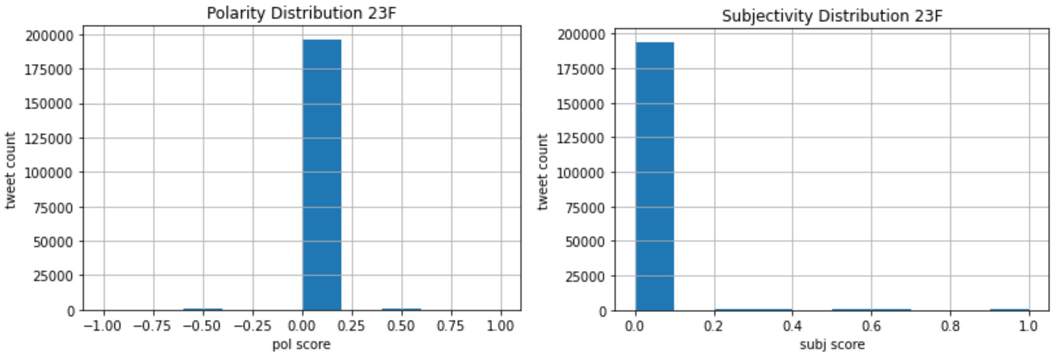


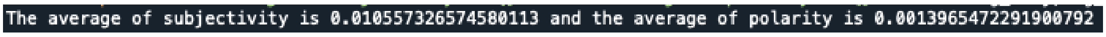
```
The most famous influential tweet according to our retweeted_status function is the following:
RT @A3Noticias: 📅 Han pasado cuatro décadas del histórico #23F

→ Casi todo el mundo cree recordar qué hacía exactamente ese 23 de febrero…
```

(6) Sentiment Analysis

In this subsegment we used the TextBlob library for processing textual data, we analysed the polarity and subjectivity scores for all the words from the tweets that we have extracted. Then we performed an average of all the word scores to get the average polarity and subjectivity. By executing this analysis, we gather insights on the overall reactions to the 23F. For clarification reasons, we will further explain each one of the concepts in this subsegment.

Polarity is assigned as a float and measures the positive and negative phrases between the ranges of [-1, 1]. Neutral sentiments have a polarity score of 0. Subjectivity measures the objectivity of the users' response and ranges from [0,1] where 1 is subjective and 0 is very objective. From the “Wordcloudtext_Sinvenezolano_English.txt”, these were the results obtained:





```
The average of subjectivity is 0.010557326574580113 and the average of polarity is 0.0013965472291900792
```

Following our interpretation from the graphs, tweets were rarely subjective nor polarised. There is a small proportion of words that vary in subjectivity and polarity but aren’t significantly representative of the entire tweet's ingestion. We hoped that there were going to be highly polarised spectrums in this analysis as these kinds of events trigger the emotional component of some users and incite users to share their opinions.

(7) Final Remarks on our EDA

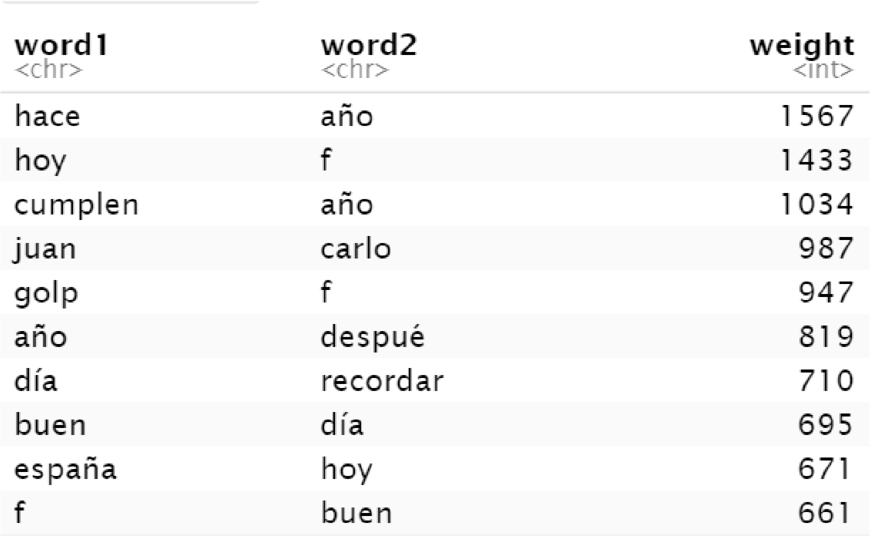
Given all the information given by this investigation, we have a better understanding of the data that we are going to work with.

We have noticed that we need to filter those tweets that do not belong to the Spanish 23F (For instance: Venezuelan, Japanese and Chinese tweets) and that there are other perspectives on the successful 23F coup d’état. Unfortunately, we were unable to see apparent different perspectives on our subject matter in the sentiment analysis. Nevertheless, we remain hopeful to see this in our graphical network representation once we get rid of all the bias in our ingested tweets.

4. Analysis of the Bigram network

1. Description of network

As mentioned before, each node represents a word, connected to the word to which it appears the most often together. Due to the considerable number of words extracted from the tweets, we decided to set up a threshold of at least appearing 300 times in the data set to be considered as a node. The Bigram network consists of 160 nodes and 188 edges. The weighted vertex degree was calculated with the function strength(), which sums up the edge weights of the adjacent edges for each vertex.

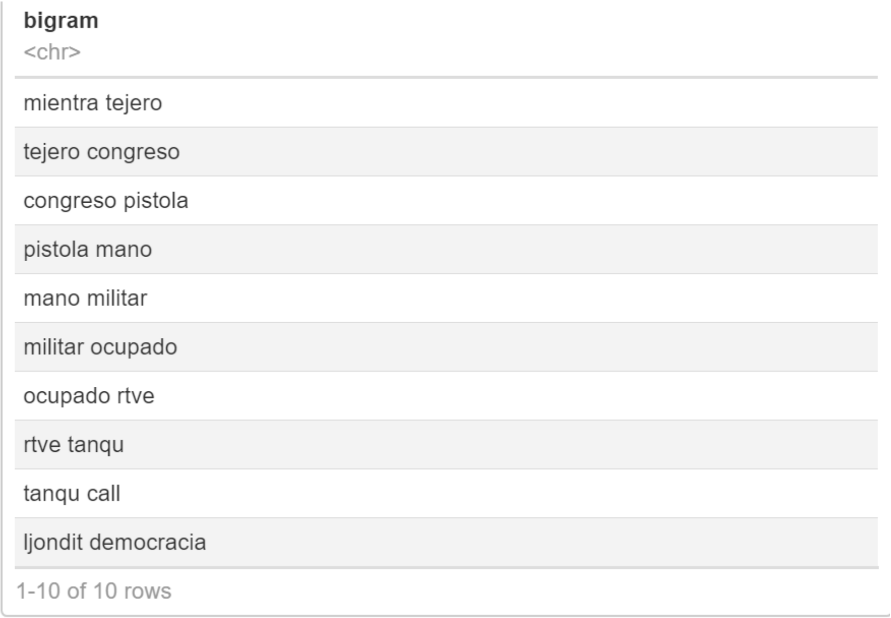


word1	word2	weight
<chr>	<chr>	<int>
hace	año	1567
hoy	f	1433
cumplen	año	1034
juan	carlo	987
golp	f	947
año	despué	819
día	recordar	710
buen	día	695
españa	hoy	671
f	buen	661

2. Network visualization and analysis

1. Setting up the network:

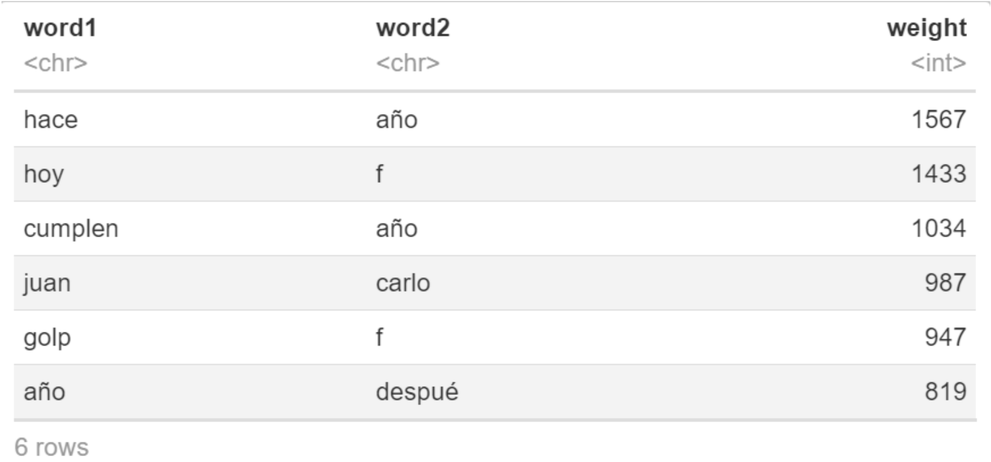
To create a bigram network, it is necessary to do a previous step in which we divide the sentences into couples of words that appear together often. In this table, we can see a small sample of the different partitions that are created. Since this is a random sample, the conclusions that we can draw would not be representative of the network. Nonetheless, but we can recognise some expected terms as we are talking about a coup d'état: "militar", "pistola", "congreso"…



bigram <chr>
mientra tejero
tejero congreso
congreso pistola
pistola mano
mano militar
militar ocupado
ocupado rtve
rtve tanqu
tanqu call
ljondit democracia

1-10 of 10 rows

But this is not enough to build our network. We need to divide each pair of words into “from” and “to” columns and assign them a weight (their absolute frequency). We can see the first 6 rows in the following table:

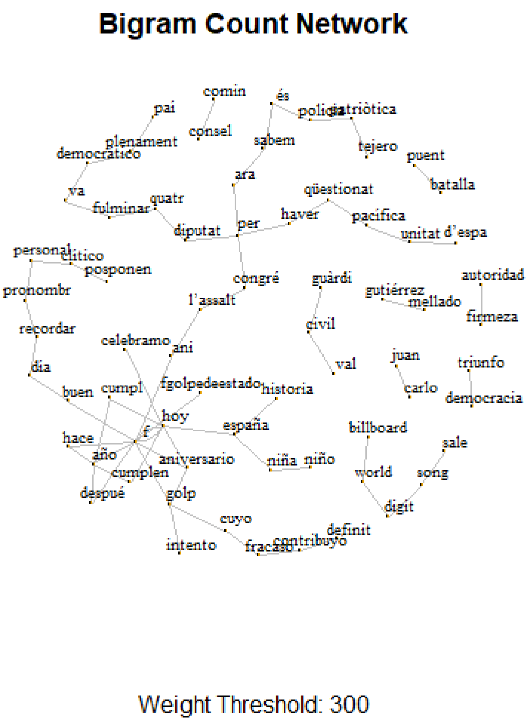


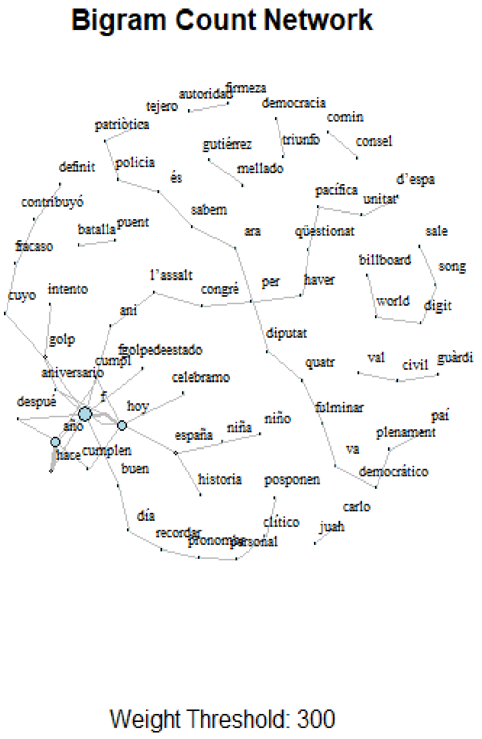
word1 <chr>	word2 <chr>	weight <int>
hace	año	1567
hoy	f	1433
cumplen	año	1034
juan	carlo	987
golp	f	947
año	despué	819

6 rows

Once this has been done, we can proceed to the plotting of our network.

2. Plotting the Bigram network





From these two graphs (unweighted and weighted network) we can draw some quick conclusions:

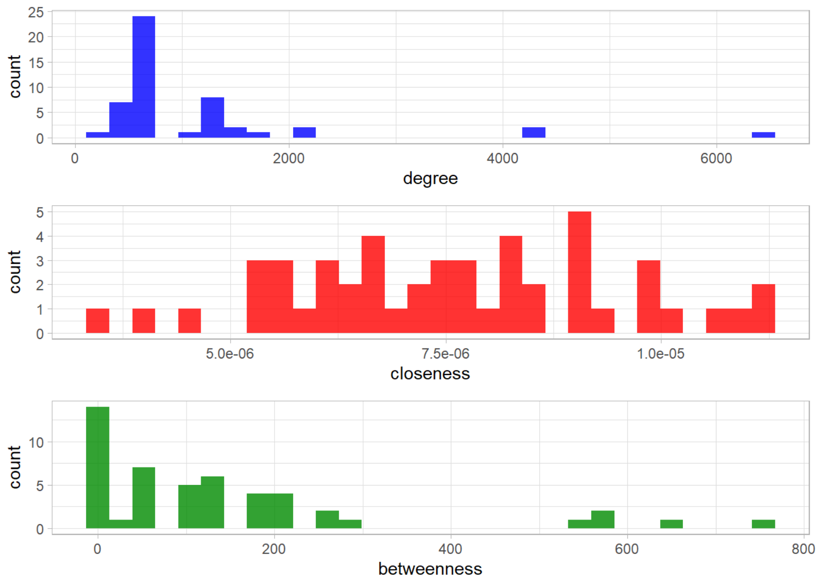
- The letter "f", as expected is an especially important node since it the point of departure of many edges. It is expected to be a key node since on twitter the main hashtag used was #23febrero
- The words "hoy" and "año" also are key nodes of the network.
- It is interesting to see how "democracia" was in the top 20 most used words, but not among the important nodes. It is present, but only related to the word "triunfo". This makes sense since it represents the fact that the coup failed, and democracy "survived" it could be expected that it was not an isolated community.
- There are communities formed around words in different languages (English, Catalan and Spanish).
- Lastly, no words from the most relevant tweet have a significant role in the network.

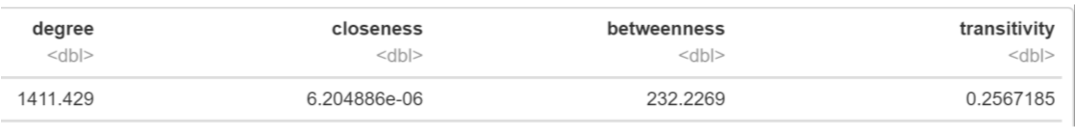
Let's zoom out a little bit to see better the whole bigram network. The dynamic graph created with NetworkD3 can be visualized in the Rmarkdown attached to this report. Since it's dynamic and we have zoomed out (by setting the threshold to 150) we can see better the network that we have created. Even if it's an interesting graph in which we can see all the connections among words, we will not analyse it. We'll proceed to an in-depth analysis of the network, but once we perform a skip-gram analysis.

3. Centrality measures

Since we have created two types of networks (bigram and skip gram) but only the latter was used for the creation of communities and clusters, we will be analysing that one: the skip gram network.

Visually we can presume that there are certain nodes which will have a higher degree centrality (the node "f" for example) than others. From this table and graph, we can have a further insight into the distribution of the centrality measures: degree, closeness and betweenness.





degree <dbl>	closeness <dbl>	betweenness <dbl>	transitivity <dbl>
1411.429	6.204886e-06	232.2269	0.2567185

4. Degree centrality:

Since what we have created is a non-directed graph, we define the degree of a node as the number of direct connections such node has with the other nodes. In our case, the amount of direct "one hop" connections each word has to the others in the network. We assign a score based on the number of links each word has. So, the higher the degree of a word, the more important it is.

In terms of degree, we can see how there are around 35 nodes that have a low degree (around 5000 links), and in second place we find around 10 nodes that have more than 1000 links. On average, we have a 1107 degree. Meaning that even if in the graph these chunks of nodes seem to be in the lower bound, they are not. The "matter" is that we have two groups of outliers, among which we find the words "f" and "hoy", which even if lack content, we already noted that they are important nodes since they tend to be at the beginning of the tweets.

5. Closeness centrality

Closeness centrality measures the shortest paths between all nodes and then assigns each node a score based on its sum of shortest paths. Here, we find an extremely low average value of 0.00000758692. In this case, we see how the values are better distributed in comparison with the previous metric. For our network of words, this value implies that words do not transmit information well among each other. Even if this seems odd for a network of words, since they are not people, it is important since in such a politically polarized context, if we had a very high closeness centrality measure it would mean that the discourse could change quickly and thus become rapidly polarised. In addition, we could understand that it would be difficult of spread fake news.

6. Betweenness centrality

Betweenness centrality measures the number of times a node is found on the shortest path between other nodes. In other words, it tells us the degree to which nodes will stand between each other. In our case, we find an average value of 144 and a left-skewed distribution with a high number of nodes with 0 betweenness centrality. Again, we could expect this distribution and values. The reason is that since we are looking at connected words, it is difficult that this value is high since conversations are different. We of course have some outliers as always, which correspond to the key-words of the discourse of the day.

7. Transitivity

Transitivity refers to the probability that the adjacent vertices of a vertex are connected. We find a transitivity value of 0.1081081. This value is very low, but it was expected since what we have created is a network of related words, and thus the fact that they appear together does not mean that they will be connected in triads. Furthermore, since what we have created is a bigram network, it is expected that we do not have high transitivity since there are no clusters of words connected, contrasting to what we will see in the skip gram analysis.

5. Skip gram Analysis

1. Description of the network

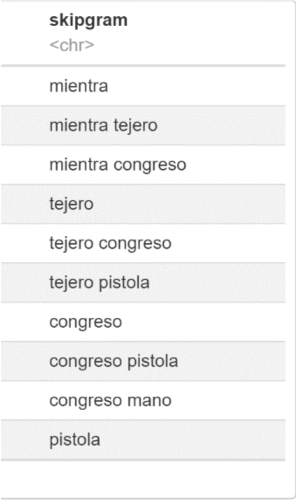
In the skip gram network, every single node represents a word, but in this case, the word will be connected to a word to which it is related. Before we were looking at words that appear ***together***, here we are looking at words that are ***related***. We will use this second network to define the communities and clusters.

Here, the threshold was set at 200 meaning that we consider words to be related if they appear together at least 200 times. The Skip gram network consists of 145 nodes and 302 edges. The weighted vertex degree was calculated with the function strength(), which sums up the edge weights of the adjacent edges for each vertex.

2. Visualization of the network

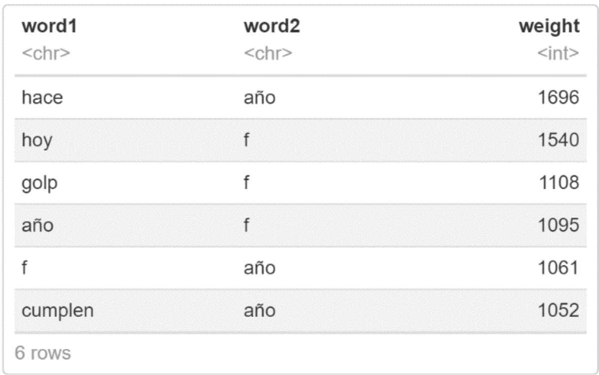
a. Setting up the network:

We are going to induce word embeddings by exploiting the signal from the word-context co-occurrence.[2] To do so we also must create pairs of words but now they will not necessarily appear next to each other, but they will be related.



skipgram <chr>
mientra
mientra tejero
mientra congreso
tejero
tejero congreso
tejero pistola
congreso
congreso pistola
congreso mano
pistola

As we see in this table, words that appear have some sort of relationship with each other but do not necessarily appear together. Take "tejero", "tejero congreso" and "tejero pistola". Tejero was the head of the coup d'état and was the one who threatened the deputies that day at the congress. Thus, the words "congreso" and "pistola" appear together. As it was done before, we separate the words and give each a weight. We can see the first 6, bellow.



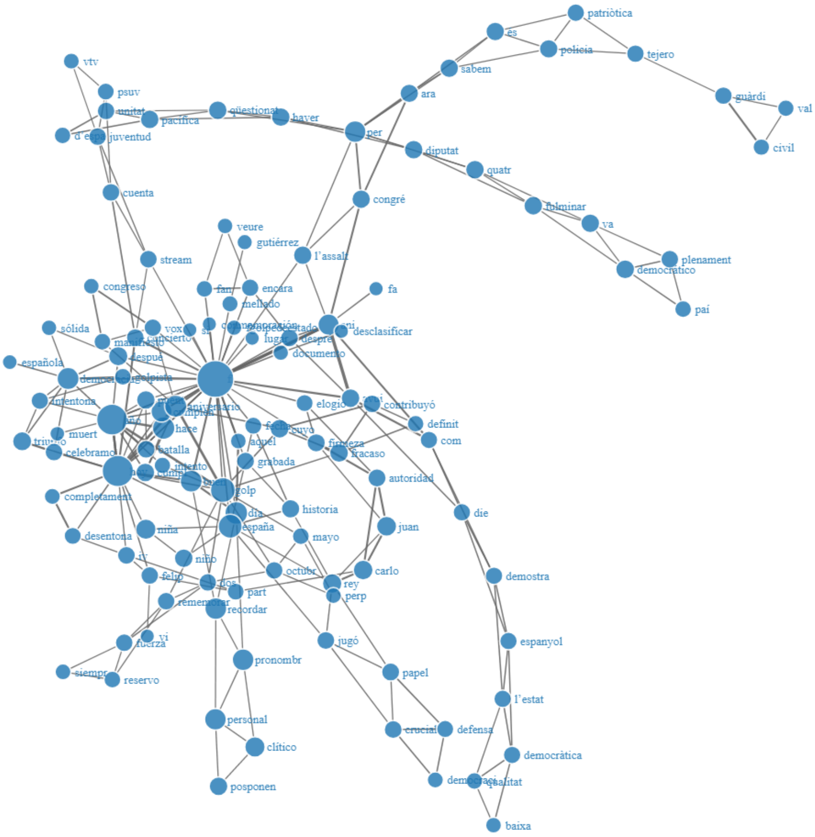
word1 <chr>	word2 <chr>	weight <int>
hace	año	1696
hoy	f	1540
golp	f	1108
año	f	1095
f	año	1061
cumplen	año	1052

6 rows

[2] (Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis, 2021)

a. Plotting the Skip-gram network

While this is the static version of our graph, you can access the dynamic version created with NetworkD3 in the Rmarkdown attached to this report.

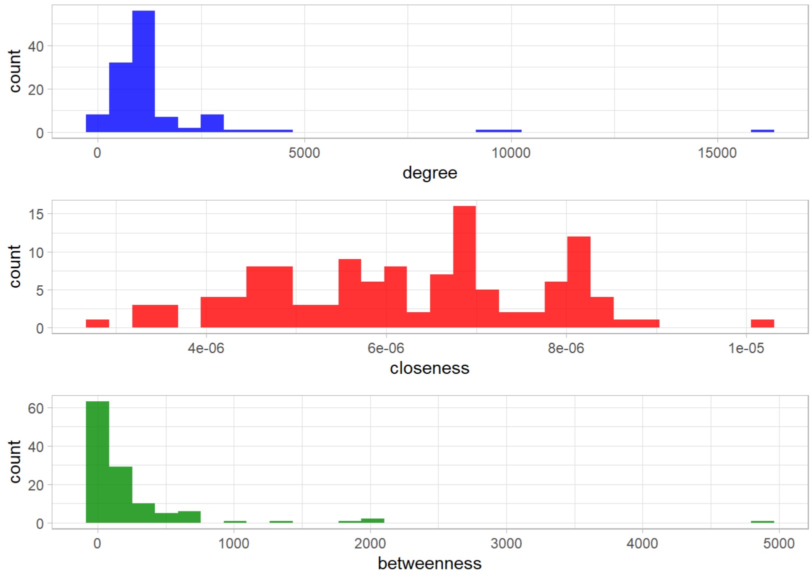


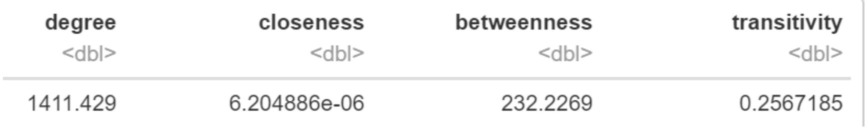
From this weighted graph we can also extract some quick insights:

- Overall, we see all the connections that we saw before in the bigram, but now with better connections among nodes.
- Again, the letter "f" is the most important node, and the central one.
- In comparison with the bigram, we see how in the word "hoy" we see that It's related to the prominent figures of the day, such as "Felipe iv".
- Along these lines we also see how "Juan Carlos" is no longer an isolated node, but it's connected directly to the word "autoridad". Juan Carlos was the king of Spain at the time, and the person who was able to put an end the coup d'état.

- However, now we can see the connections among the words, we see many triangles, meaning that we can expect a higher transitivity.
- There is a big node centred and from there we see two stems that go up and which are joined in one node, both in Catalan. There is also another stem in Catalan, but in this case going down.

3. Centrality measures





degree	closeness	betweenness	transitivity
<dbl>	<dbl>	<dbl>	<dbl>
1411.429	6.204886e-06	232.2269	0.2567185

a) Degree centrality

Unsurprisingly, we see a similar distribution to the previous network. However, we have a higher average degree. In this case we have a mean of 1411. Why is this the case? As we mentioned, bigram only creates a network of words that appear together often, while in this case we have words that are related. It follows then that the degree centrality will be higher, since the connections are more (we no longer need to have words appearing next to each other) and better. The node with the higher degree is, as expected, the letter “f”.

b) Closeness centrality

In terms of closeness, the average value is 0.000006204886. In this case we have a worse metric than before, it is more difficult to spread information among related nodes than just nodes that appear often together. This is quite surprising, since we could expect that if words are related to each other, it would become less difficult to transmit information along those connections. However, both values are extremely low, so the impact is not meaningful. In the distribution we see how it is homogenously distributed, but we find two important spikes around two standard deviations away from the mean, in which around 15 and 10 nodes have much higher closeness. If we look at which are these nodes, we see again the word "f" but also, the word "vox" with a closeness value of 0.000008410429. Such value showcase something that we have been noticing in the past years: populist rhetoric is transmitted much rapidly.

c) Betweenness centrality

The betweenness centrality is higher than before. We find an average value of 232, but the distribution is fairly the same as in the bigram. It's a left skewed distribution in which most of the nodes lack betweenness centrality.

d) Transitivity

In this network we see how the value of transitivity has almost tripled from the previous one (from 0.1081081 to 0.2567185). Just like we mentioned above, this was expected since the skip gram method looks at the relationships between words and thus creates interrelated pairs and triples of nodes.

4. Community detection

i. Louvain

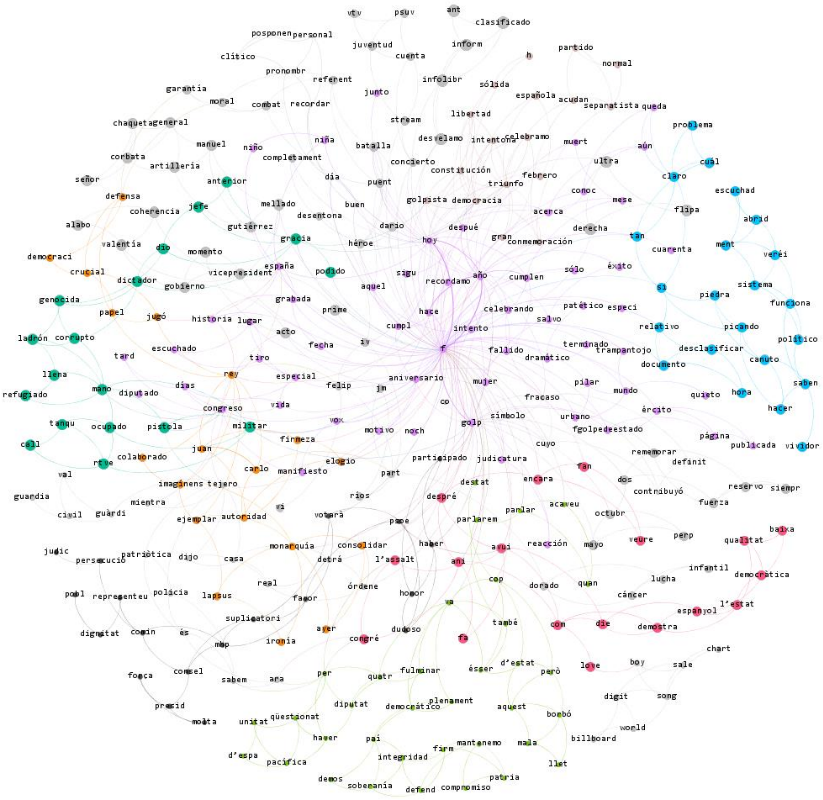
The method used to find communities was the Louvain method. Using this method allows to optimize modularity. Since our network is quite special because it does not connect persons, flights… but words, we need a method that can optimize this parameter as much as possible.

ii. Modularity and number of communities

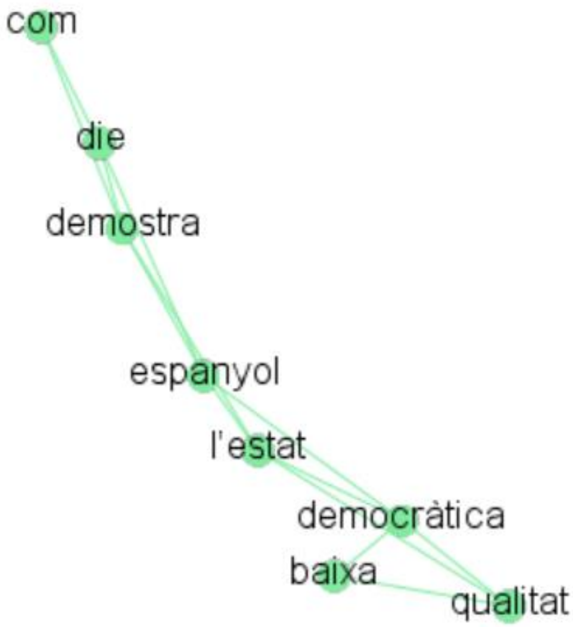
The Louvain method in R can find 13 groups with a modularity of 0.64 (1 being the maximum value and thus having a full modular clustering). A modularity close to 0.7 means that we have a quite high density of edges inside the communities in respect to the edges outside the communities. The communities are really interconnected among themselves.

iii. Visualization of communities

We have visualized the network both in R and Gephi. In the R version we have a dynamic graph, but for the purposes of the report we will focus on the graph created in Gephi. We can see the overall distribution in the different 13 communities but breaking each one down will allow us a better understanding. The dynamic version of the graph in the Rmarkdown is attached to this report.



iv. Community one

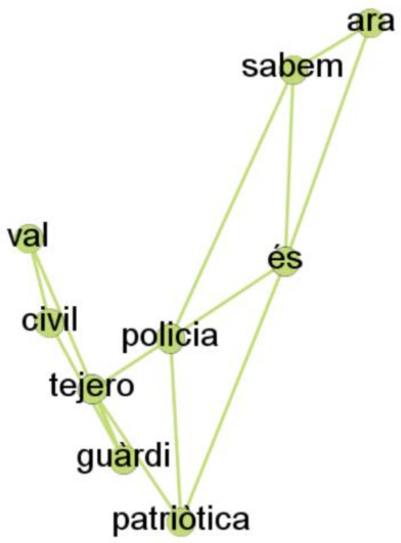


In this first community we see words in Catalan. We understand the current division of the Spanish society regarding the independence (or not) of Cataluña. Elections for the presidency of the Generatlitat (the main institution by which Cataluña organises its autonomous competencies).

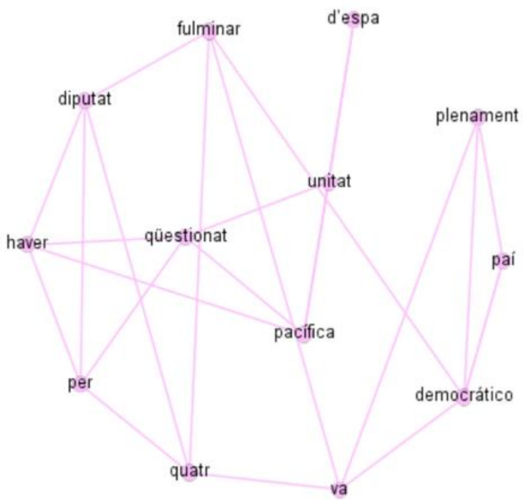
What we see here is that the catalán community on twitter does not really hold dearly the idea of the Spanish state. What they tweeted in that day, instead of the flattery words of the rest of the network (“triunfo”, “sólida”…), were words discrediting it. We see that they have taken this opportunity of the 23F to “remind” society that this is a clear example of the low quality of the Spanish democratic state.

v. Community two

In contrast to the previous Catalan community that was analysed, we can see how in this one there are no political connotations, they are just stating the facts. They are talking about Tejero, the police, the civil guards… they are not making comments about the day. We could say that “patriòtica” could have some connotations but in this context of words, we can extract much.



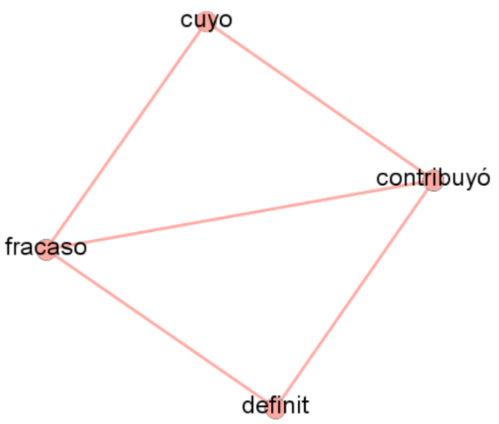
vi. Community three

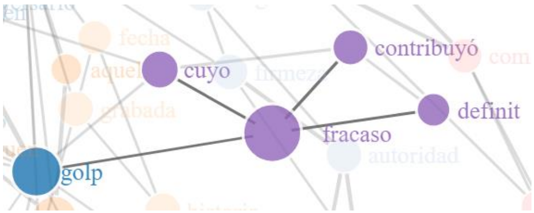


This community is also in Catalan, and it's also a politically weighted one. The thing that makes it different from the previous one is that instead of talking about the Spanish democracy, they also talk here about the unity of the nation. At the end of the day, they are asking to re-think the unity of the country and thus to put forward their independence messages. What is concerning is that the word "pacifica" is directly linked to "questionat" and "unitat". We saw that in general there was not high closeness centrality, therefore there should not be any concerns regarding polarizing the debate.

vii. Community four

These words seem to be clustered into one community because they all talk about the end of the coup d'état. If we zoom out and see to which other words, they are related, we see that they are talking as expected, about the coup.





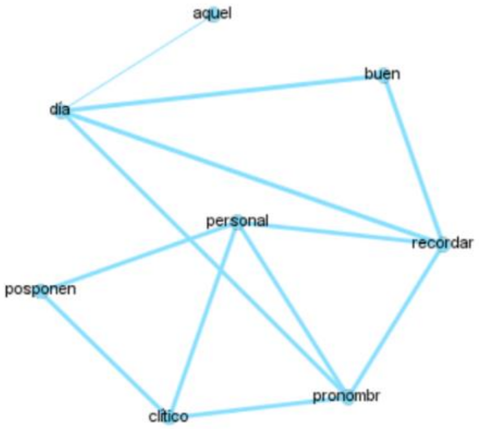
viii. Community five

mellado
gutiérrez

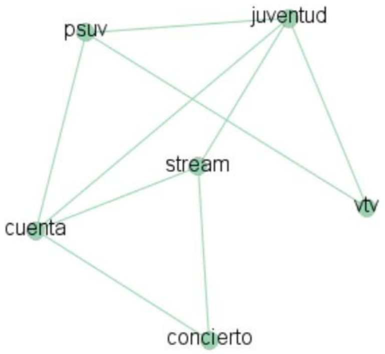
This community was formed because it links the words "Mellado" and "Gutiérrez" which correspond to the name Manuel Gutiérrez Mellado[3], a spy of the Franco dictatorial regime, but that in the day of the 23rd of February was the and vice president of the Government of Adolfo Suárez. Without any doubt he decided to jump from his seat when the coup plotters entered the building asking them to leave and hand in their weapons.

ix. Community six

This community is uninteresting, there is no message that we can extract from it other than that people are talking about their personal experiences of the given day.



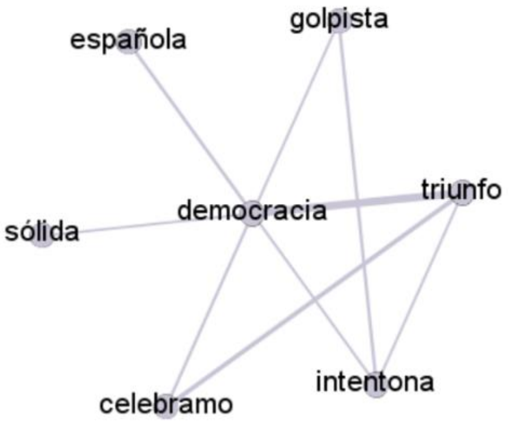
x. Community seven



This community seems random if we put it into the context of the 23rd of February. We can consider that there must have been some concert in streaming and that is why these words were clustered into a community.

[3] (Manuel Gutiérrez Mellado, de espiar para Franco a batirse por la democracia el 23-F, 2021)

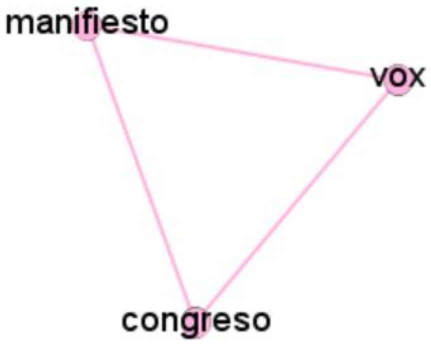
xi. Community eight

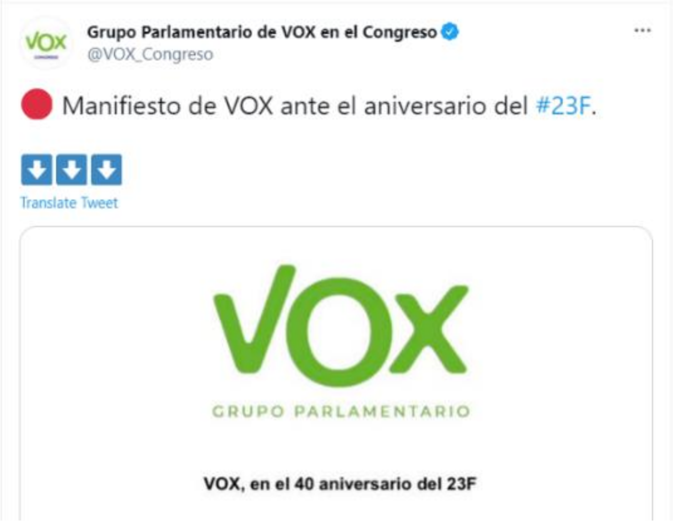


Here we can see some of the most popular words in the dataset, the words "democracia", "solida", "triunfo"…In addition, we can see from the weights of the edges how the words "triunfo", "democracia" and "celebramos" appear more often in the tweets. All in all, this small community represents the overall sentiment that we have seen across the analysis: Spaniards see this day as a day of celebrating how solid and good their democracy is.

xii. Community nine

This triad is fascinating. The only political party whose name has appeared in the whole network is VOX, the extreme right party of Spain. Not only they have been able to make an entrance in the network, meaning that Spaniards have talked about it quite a lot since we set up several filtering thresholds, but also, they are connected to an especially important word: "congreso".




Grupo Parlamentario de VOX en el Congreso
@VOX_Congreso

🔴 Manifiesto de VOX ante el aniversario del #23F.

⬇⬇⬇

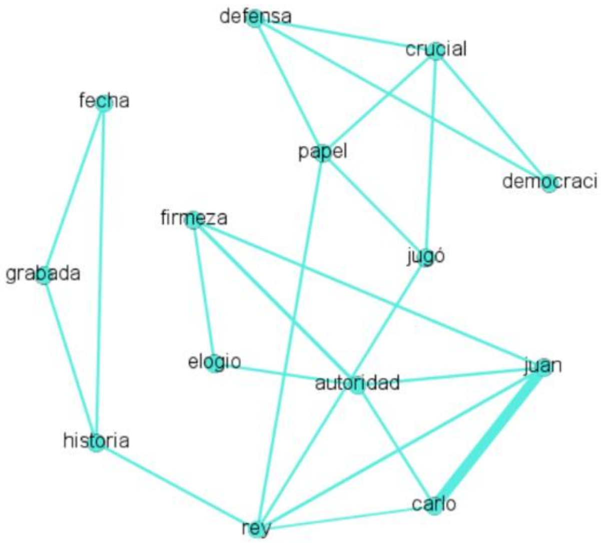
Translate Tweet

VOX
GRUPO PARLAMENTARIO

VOX, en el 40 aniversario del 23F

But why is this the case? A quick search on twitter gives us the result straight away: Vox released a manifesto on that day. In the given manifesto Vox tries to change the focus of the attention to the failures of the current coalition government. They present themselves as the only alternative for a unified Spain who recognises the role and importance of the king of Spain.

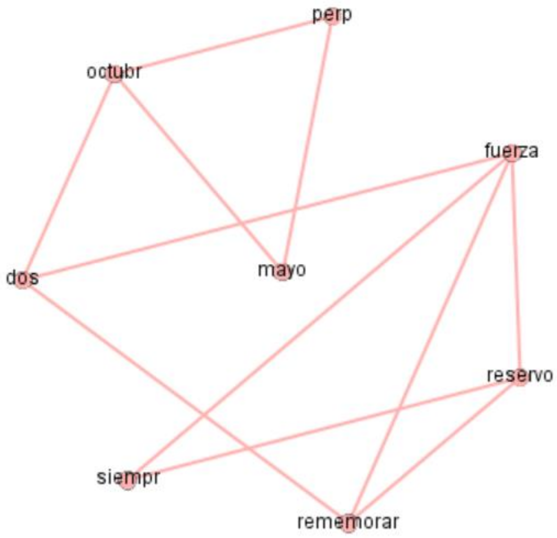
xiii. Community ten



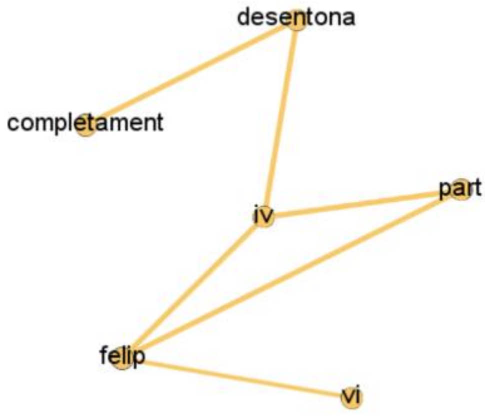
Here we see another community in which they describe the events of the day, but with a preference towards the king. They talk about how important his role during this day was since even if he does not have the powers, he used to have before, he was the only authority that could break down the coup. If we were to delve deeper into the political analysis, we could presume that this community would belong to the right-wing, since the tone of the words around the figure of the king is not confrontational but praising.

xiv. Community eleven

In this case, we see how they have mentioned another important date of Spanish history, the 2nd of May. That day the people of Madrid rebelled against the French occupation, leading to higher repression. It's interesting to see how users connect different historical events. We also see the word "octubre" which may refer to the 1st of October when there was an attempt of independence by Cataluña. As we can see all of them are events related to the social uprising, in one way or another.

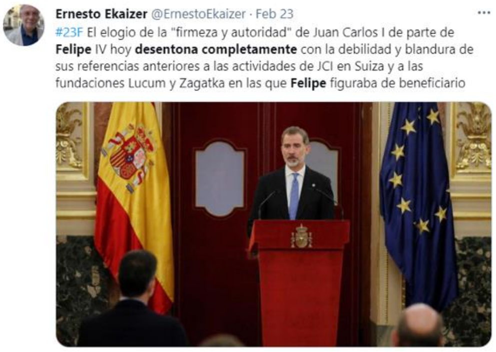


xv. Community twelve



As we explained, the king has a delicate reputation nowadays. Both the emeritus and the current king are always being criticised, especially by the Spanish left which has a strong republican ideology. In the present community we see how the current king, Felipe VI is being discredited by saying that their presence today is "out of tune".

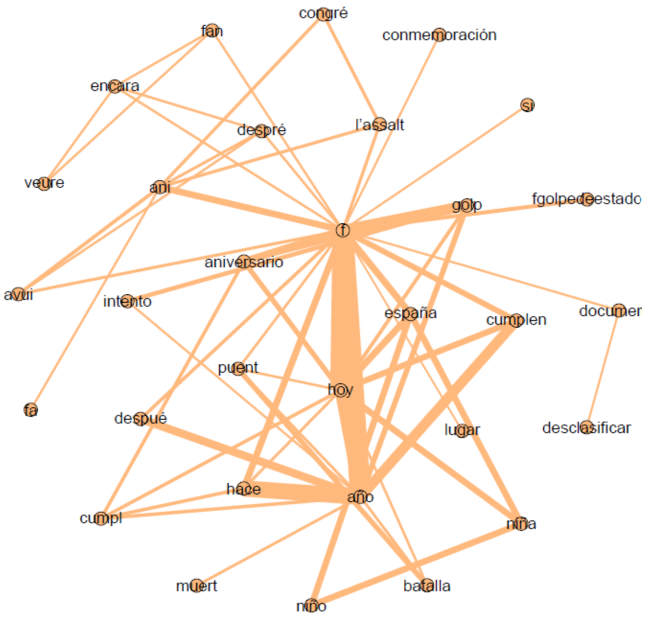
The reason behind it is that there was a viral tweet criticizing the speech he gave, in which he was congratulating his dad (the emeritus king) for this firmness and authority held on the 23rd of February.


Ernesto Ekaizer @ErnestoEkaizer · Feb 23 ···

#23F El elogio de la "firmeza y autoridad" de Juan Carlos I de parte de **Felipe** IV hoy **desentona completamente** con la debilidad y blandura de sus referencias anteriores a las actividades de JCI en Suiza y a las fundaciones Lucum y Zagatka en las que **Felipe** figuraba de beneficiario

xvi. Community thirteen

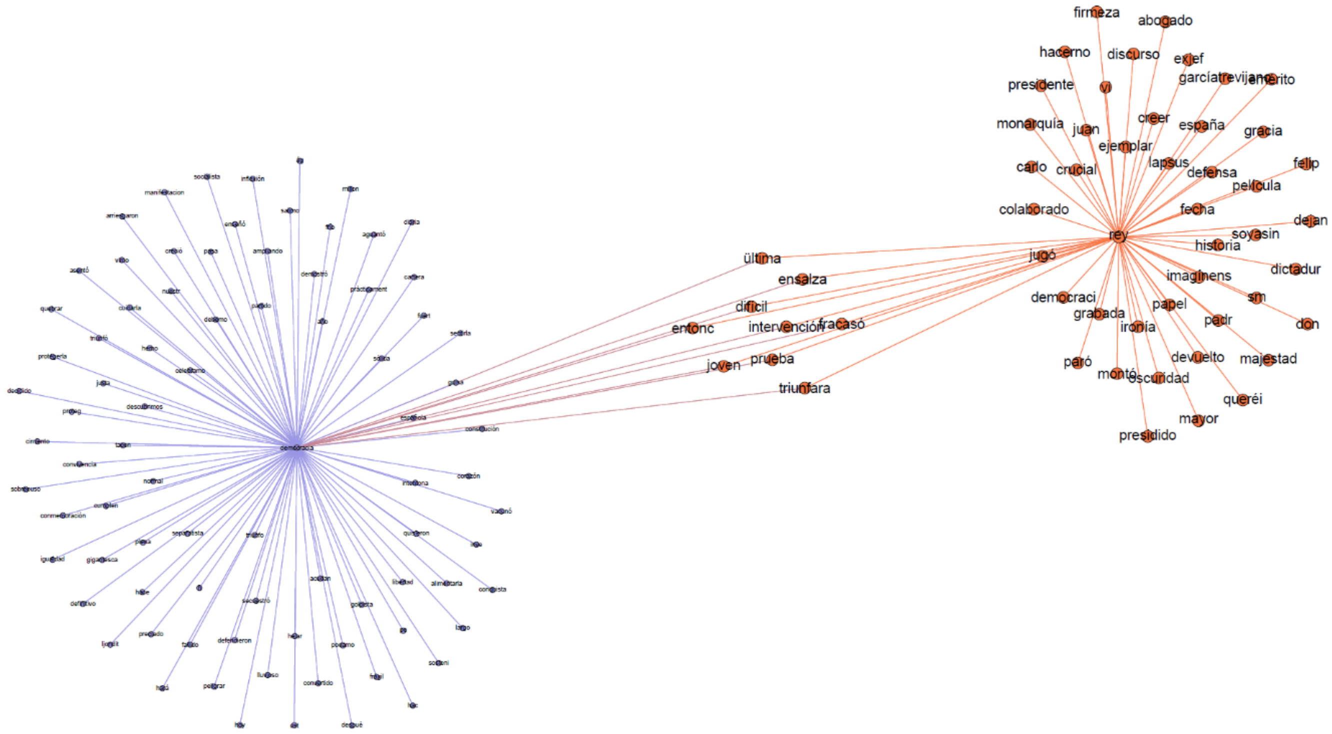
In this community, we see words both in Catalan and Spanish. It is the most general one out of the thirteen communities. If we look at the general graph, we see how this community is at the centre (this can be intuited since the word "f" is in this community and we already know that it is key to our analysis). It does not contain any political connotations even if we see some words in Catalan. The only important nodes that we should highlight are the "desclasificar" and "documentos" ones. They refer to the fact that most of the documents from the coup d'état trial are still confidential, and there is high interest in Spanish society to gain access to them.



5. Phi Correlation coefficient

Lastly, we will divide the skip gram into three clusters of “our choice”. We will use the phi correlation coefficient, a measure used often in text mining analysis. It measures how much more likely it is that either both word X and Y appear, or neither do, than that one appears without the other.[4] It’s important to note that the relationships that we are seeing here are symmetrical, not directional.

Quick reminder: the top 3 words were “hoy”, “golpe” and “democracia”. Since “hoy” is a word without any political connotations, we have decided to perform the correlation coefficient around the words: “golpe”, “democracia” and “rey” ("golpe“ is the 6th most popular word). It will calculate the correlation between each pair of words and then filter by one of the three clusters that we have chosen as topic words. Furthermore, after a lot of trial and error, we decided to set up a 0.1 correlation threshold. The result is the following graph, which can be visualized here:



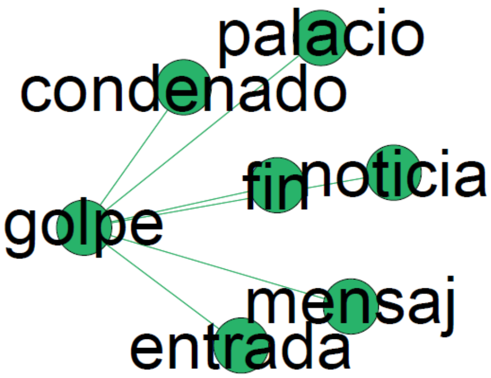
[4] (Robinson, 2021)

The top 5 pairs of words that have the highest correlation among them are: “democracia-triunfo” with a value of 0.4125428, “jugó-rey” with a value of 0.3930265, “rey-grabada” with a value of 0.3913712, “rey-crucial” with value of 0.3833579 and lastly, “rey-democracia” with a value of 0.3765380.

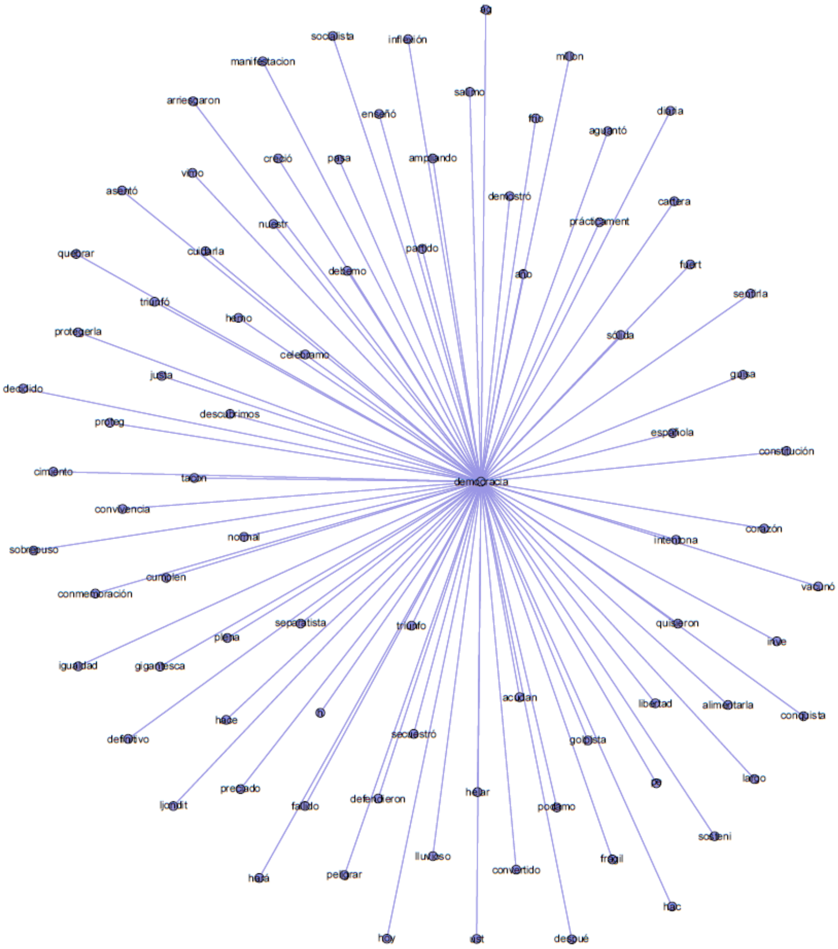
The words that we see in the middle are those that are both correlated with the word “democracia” and “rey”. Let us explore them. They are “última”, “enlaza”, “difícil”, “intervención”, “fracasó”, “prueba”, “entonces”, “joven”. Except for “entonces” all refer to the same thing: the difficult task the king of Spain had that day, in which democracy became so fragile for some hours. We also find the word “jóven”. In this case, it does not refer to “when I was young…” But to the fact that Spain’s democracy was very young at that time and thus there was a higher risk of falling back to a dictatorship.

I. Cluster 1: “Golpe”

This first cluster does not have a lot of relevance, and we cannot even see it in the above image since it was too small. Even if the word “golpe” was a very important node in our network and it was one of the most repeated words, it has not resulted in a relevant cluster. The words attached to it are very few and aren't very meaningful. We only see “fin” and “condenado” which are directly related to the failed attempt and the further prosecution of the instigators.



II. Cluster 2: "Democracia"

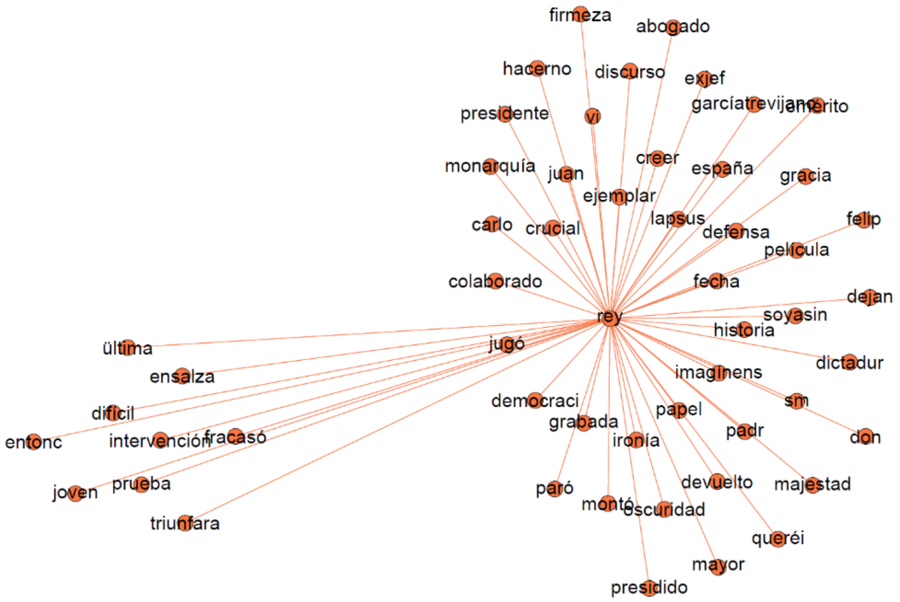


The "democracia" cluster is the one that has the biggest number of words related to it. As we saw before, "democracia" is included among the pairs of words with the highest correlation among them. IN this cluster we find the following top five pairs of correlated words: "democracia-triunfo" (0.4125428), "democracia-española" (0.2773625), "democracia-sólida" (0.2653900), "democracia-intentona" (0.2619193) and lastly "democracia-año" (0.2405064).

These sets of words show how Spaniards consider this day a victory a democracy in which even if there was an attempt to bring it down, it showed how solid it was (and still is). The very high correlation between "democracia" and "triunfo" and "democracia" and "sólida" portray the positive relationship Spaniards have with this concept.

Other words that are not as correlated (0.2115423) but that still made the cut to our threshold and are interesting to look at are, for example, the pairing "democracia-separatista". It shows how even on a national holiday, polarised sentiments of independence are still very much the talk of the town.

III. Cluster 3: Rey



Our final cluster is the "rey" one. We can presume that tweets talk about the emeritus king, since the current one did not intervene in the coup d'état. The emeritus king has been a contentious figure for the past years due to the different corruption scandals there have been revealed.

As a result, one could expect the conversation to be a bit contaminated by those topics. So, let us delve into the top 5 correlated words: "rey-jugó" (0.3930265), "rey-grabada" (0.3913712), "rey-crucial" (0.3833579), "rey-democracia" (0.3765380), and "rey-fecha" (0.3519099). Surprisingly we do not see many interesting relationships, other than the word "democracia" that we already analysed, and "crucial". Indeed, the king was a key figure of the day.

Let us look at some other correlated words since the top 5 did not reveal much. Another pair that is quite interesting and that relates more to the current situation in Spain is the following: "rey-ironía" (0.3151374). But what is even more interesting is to see that the next pair in the list is "rey-ejemplar" (0.3131045). This dichotomy perfectly portrays the political discourse in Spain, where the figure of the king is either idolized and protected (right-wing parties) or it's criticised (left-wing parties).

6. Conclusions

In this report, we have explored the networks that resulted from the different Twitter conversations of the 23rd of February.

We chose this date given the historical importance of the day for Spaniards. 40 years ago, Spain witnessed an attempt of coup d'état. This coup was performed by Lieutenant colonel Tejero and 200 civil guards. Given the current political tension present in Spain as a result of the COVID-19 pandemic, it became relevant to delve into the conversions Twitter users would have on that day.

Tweets were extracted using Python, especially the Twython library. Several hashtags were chosen, to better tune the search. After filtering duplicates, we reached a count of around 30,000 tweets.

Before jumping to the social network analysis, an EDA was performed in Python to have a wider and deeper understanding of the contents of the tweets. We calculated the most frequent words, most relevant users... During the EDA process, iterative filtering took place as we detected several outliers. We found two types of outliers: tweets in other languages that are not spoken locally in Spain (mainly English, French and Japanese) and tweets related to Venezuela. On the 23rd of February, several demonstrations took place in which the hashtag #23F was used, and thus contaminated our research. Once this process was done, we proceeded to the social networks' analysis in R.

To analyse the networks, we first needed to build them. We chose two methods for that task: bigrams and skip-grams. Bigrams were used to have a first image of how the network would look like, as well as ordering the data for the skip-gram analysis.

The bigram network paired up the words that appear together. While it did provide a network in which we could differentiate important nodes based on their weight, it was not able to link nodes among each other into meaningful connections. Since we wanted to detect important communities, we proceeded to the network analysing through a Skip-gram analysis. The skip-gram unlike the bigram allows to find relations among words and couples them up.

In both networks, centrality measures as well as transitivity were low. Nevertheless, we see a subtle increase in the skip-gram network. The reason why lies in its ability to pair up related words.

We set up a threshold of 200 meaning that a node will only be plotted if it appears in the skip-gram pairing at least 200 times. Since the Louvain method is the one that optimizes modularity the most, it was the one chosen to build the communities. We find a modularity of 0,64 and 13 communities.

The communities found can be divided into four categories:

1. General communities in which users talk about their experiences that day, and in which we see repeatedly the words "democracia", "triunfo" and "sólida". Overall, these communities have a positive image of the day since it showed the stregth of the Spanish democracy, even if there might be attempts to bring it down, it will prevail. We can also identify the figure of the king, who is praised for his role during the day, putting an end to the coup.
2. Catalan speaking communities: in this case, we have two subtypes of communities: critics and general comments. The critical users have taken this date to push forward the independentist agenda, stating how this day showcases how fragile and failed the Spanish state is.
3. Political communities: in this case we find VOX, who used this day to publish a manifesto against the current government administration, a user who criticized the speech king Felipe VI gave and lastly users stalking about the 2nd of May, another important day in Spanish history, in which they tried to free themselves from the French invasion. It was unusual not to see more political parties' names in the network, or the name of the president of the government, Pedro Sánchez.
4. Lastly, we found random communities in which they talked about unrelated events but that we were not able to locate in the EDA.

To finalize our analysis, we used the Phi-coefficient in order to build three communities around some of the most important nodes of the network. The nodes chosen were "rey", "democracia" and "golpe". Although "golpe" was an important word (top 3), it did not have many words correlated to it based on the phi-coefficient so we chose to ignore it. The cluster "democracy" was the most important one, since it had the biggest number of words correlated to it. The most important pairs of words were "democracia- triunfo", "democracia-española" and "democracia-sólida", which are a clear summary of the sentiments we have seen through the analysis. Lastly, we have the cluster "rey" which had also meaningful connections, such as "rey-ironía" or "rey-ejemplar", which portray the dichotomy of the figure of the king on such an important day for Spanish democracy.

7. References

- Aclweb.org. 2021. *Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis*. [online] Available at: <https://www.aclweb.org/anthology/E17-2028.pdf> [Accessed 27 February 2021].

- Ebi.ac.uk. 2021. *Properties of PPINs: transitivity | Network analysis of protein interaction data*. [online] Available at: <https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/protein-protein-interaction-networks/properties-of-ppins-transitivity/> [Accessed 27 February 2021].

- Elconfidencial.com. 2021. *Manuel Gutiérrez Mellado, de espiar para Franco a batirse por la democracia el 23-F*. [online] Available at: <https://www.elconfidencial.com/cultura/2020-02-23/manuel-gutierrez-mellado-23-f-biografia_2465447/> [Accessed 27 February 2021].

- Faculty.ucr.edu. 2021. *Introduction to Social Network Methods: Chapter 8: More Properties of Networks and Actors*. [online] Available at: <https://faculty.ucr.edu/~hanneman/nettext/C8_Embedding.html> [Accessed 27 February 2021].

- Juanitorduz.github.io. 2021. *Text Mining, Networks and Visualization: Plebiscito Tweets - Dr. Juan Camilo Orduz*. [online] Available at: <https://juanitorduz.github.io/text-mining-networks-and-visualization-plebiscito-tweets/> [Accessed 27 February 2021].

- Medium. 2021. *Measuring Network Centrality*. [online] Available at: <https://medium.com/cantors-paradise/measuring-network-centrality-2a76b0045410> [Accessed 27 February 2021].

- Medium. 2021. *Skip-Gram: NLP context words prediction algorithm*. [online] Available at: <https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c> [Accessed 27 February 2021].

- Robinson, J., 2021. *4 Relationships between words: n-grams and correlations | Text Mining with R*. [online] Tidytextmining.com. Available at: <https://www.tidytextmining.com/ngrams.html> [Accessed 27 February 2021].

- Uc-r.github.io. 2021. *Text Mining: Word Relationships · UC Business Analytics R Programming Guide*. [online] Available at: <https://uc-r.github.io/word_relationships> [Accessed 27 February 2021].