Referendum in Catalonia

26 OCTOBER 2017 on politics

Exploratory Data Analysis

Ville Saarinen, Emmi Lahtisalo, Neli Noykova

On 10th of October the president of the Catalonian Government ***Carles Puigdemont*** came close to announcing independence. Two days before that, on October 8th, approximately 350,000 people marched on the streets of Barcelona against the independence (crowd estimate by local police). The on-going situation in Spain is an interesting addition to the list of recent political developments in the West.

Both of these events were discussed actively on Twitter, which emphasize the growing influence of social media over our lives. Tools such as Twitter may offers us a new and interesting lens through which we may gain better understanding of such important political and sociological developments. Our aim is to show that using publicly available tools (open source, open data) and exploratory data analysis methods, we can gain better insight about these developments.

Process

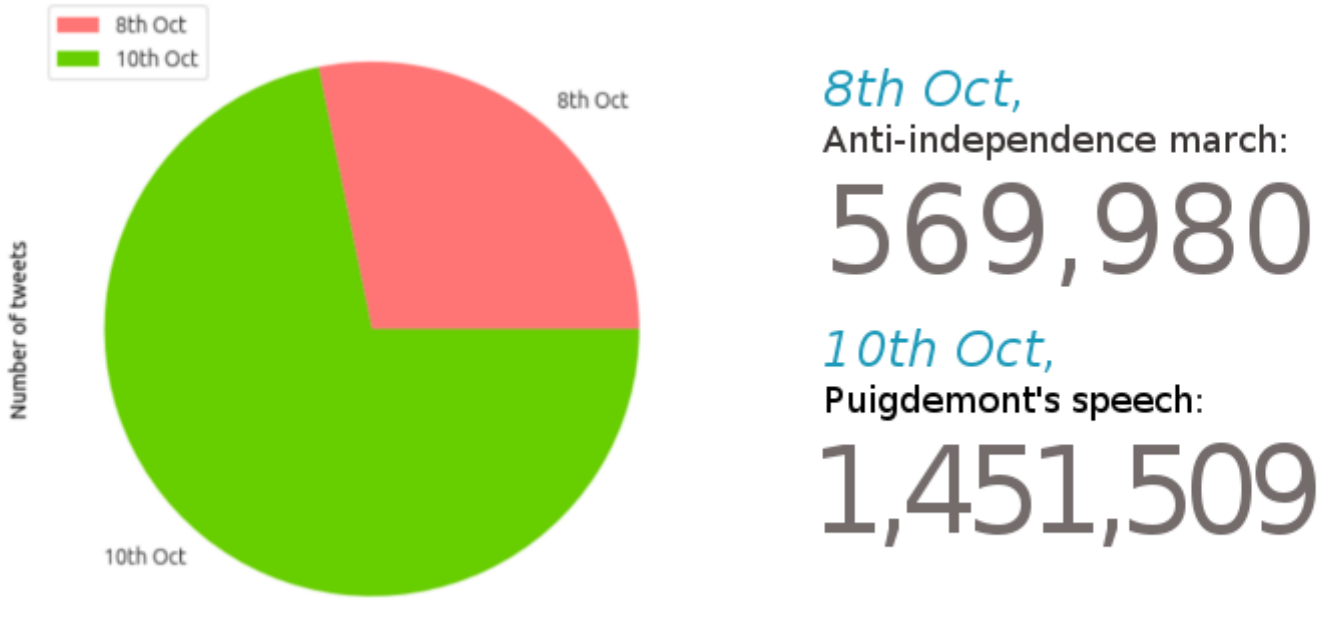
We retrieved the data through Twitter API. The criteria we set to get a relevant set of tweets was the specific timing, and a mix of appropriate hashtags and keywords. The data was then multilingual, the most commonly used languages being Spanish and English.

Keywords were:

- catalonia, cataluna, *#referendumCAT*, independence, independencia, *#referendumcatalonia*, *#recuperemelseny*, *#cataloniareferendum*, cataluña, *#catalanreferendum*, *#recuperem*, *#parlemhablemos*, francoland, democracia, separatistas, rajoy, puigdemont, corrupcion, justicia, constituciõn

General look at the data

As a result we obtained two datasets with a combined size of over ***2 million tweets.***



8th Oct,
Anti-independence march:
569,980

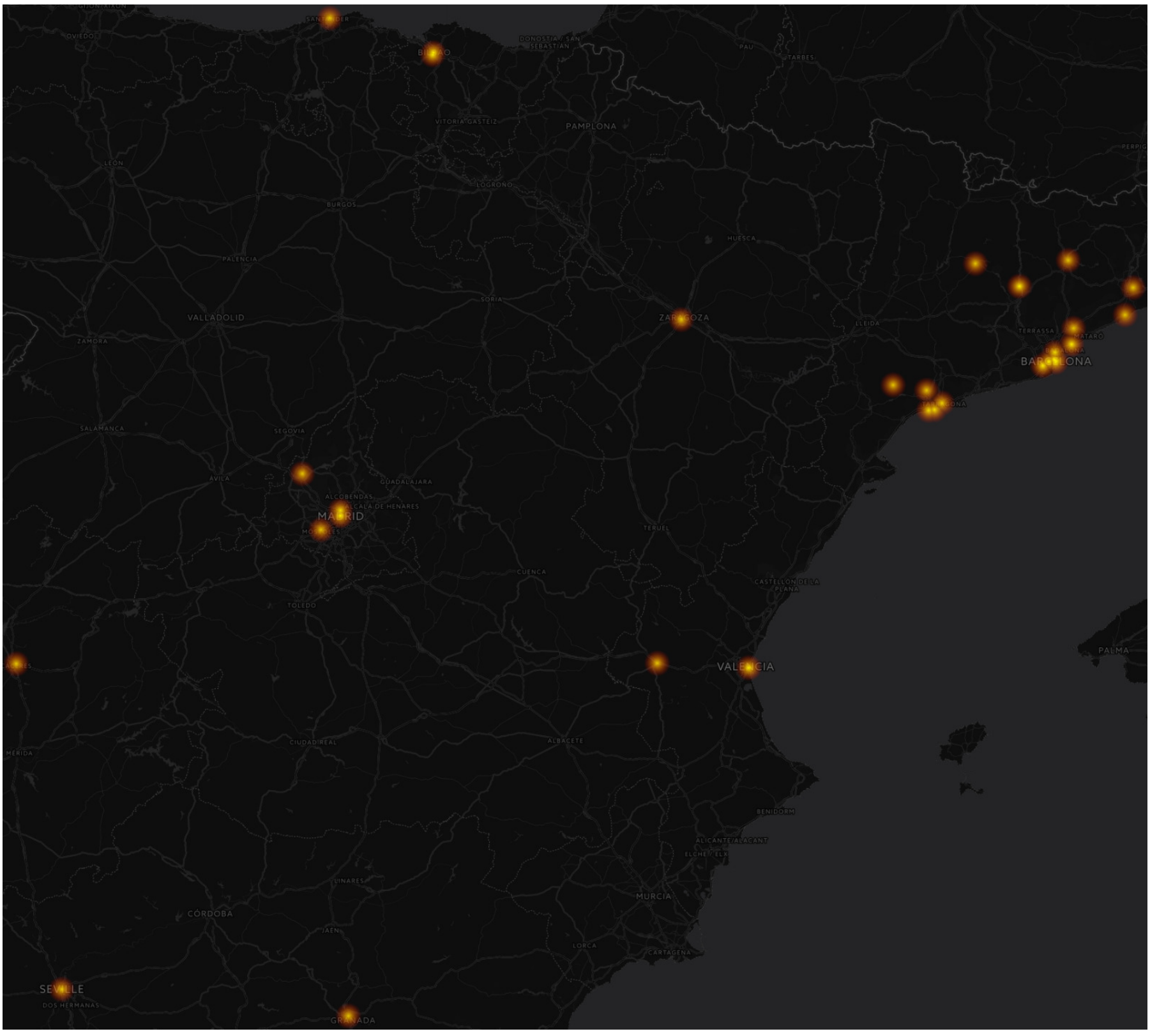
10th Oct,
Puigdemont's speech:
1,451,509

Location data

We generated a heatmap showing the locations of all of the around 200 spanish tweets which included coordinates. Note that only **less than one percent** of the data included accurate geolocation data.

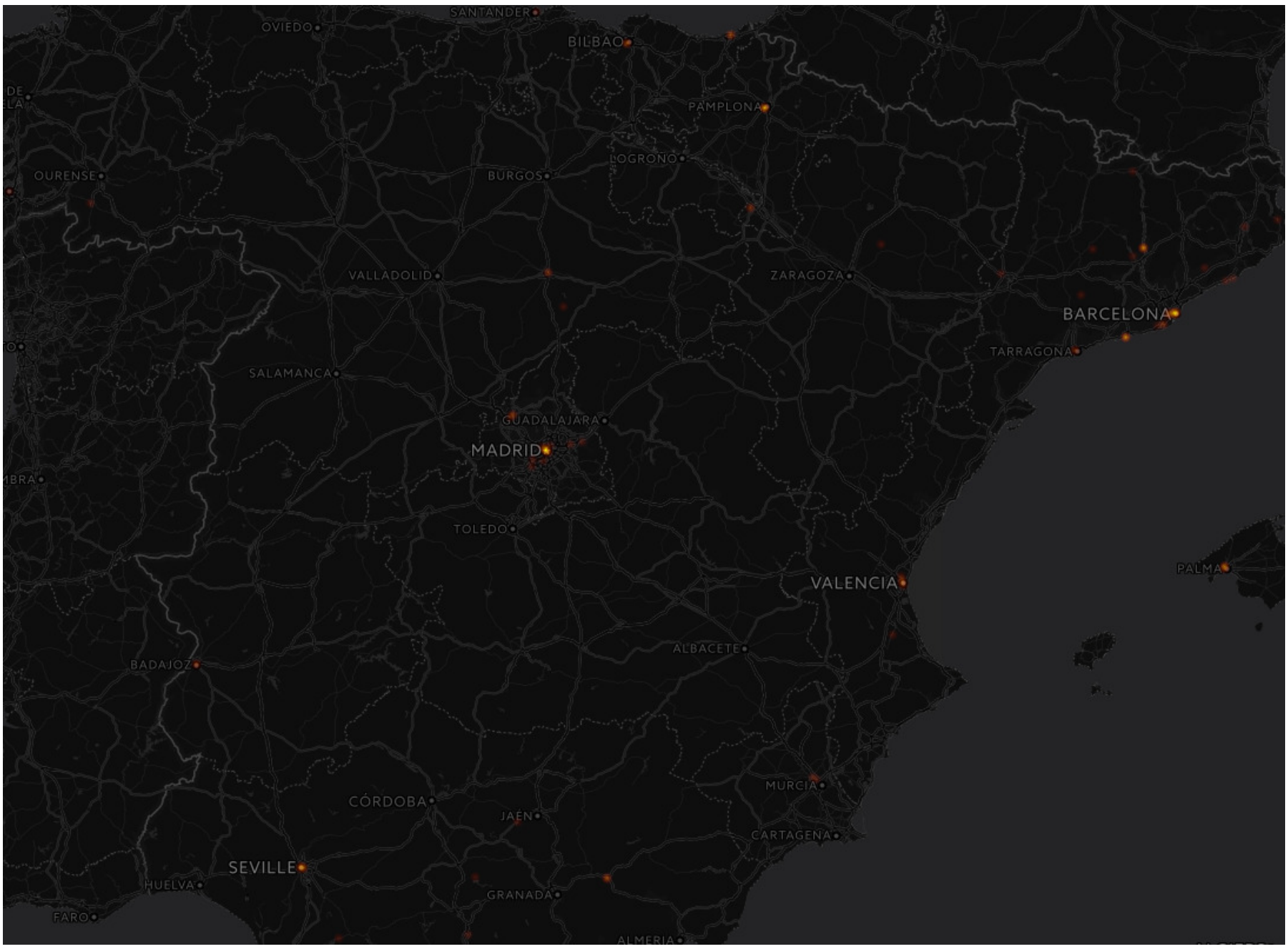
Can we draw some conclusions about this?
Maybe Twitter users are well aware of the ramifications of losing their privacy?

8th October



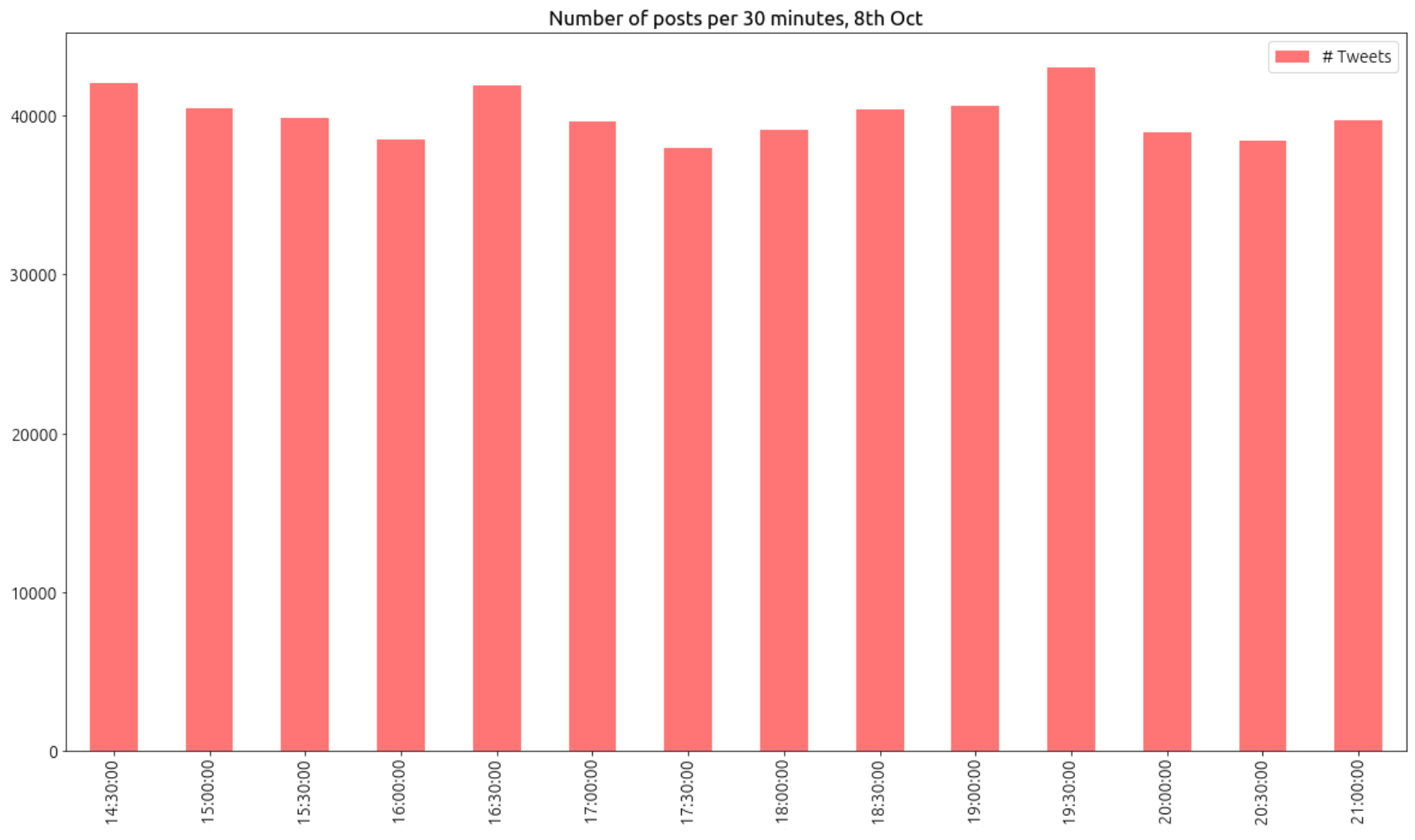
On the day of the protest march, there was notably more activity around Barcelona.

10th October



Overall there was around three times more activity on the 10th of November compared to the 8th October. This represents the same ratio as the ratio between the total sizes of the datasets.

Tweeting Activity

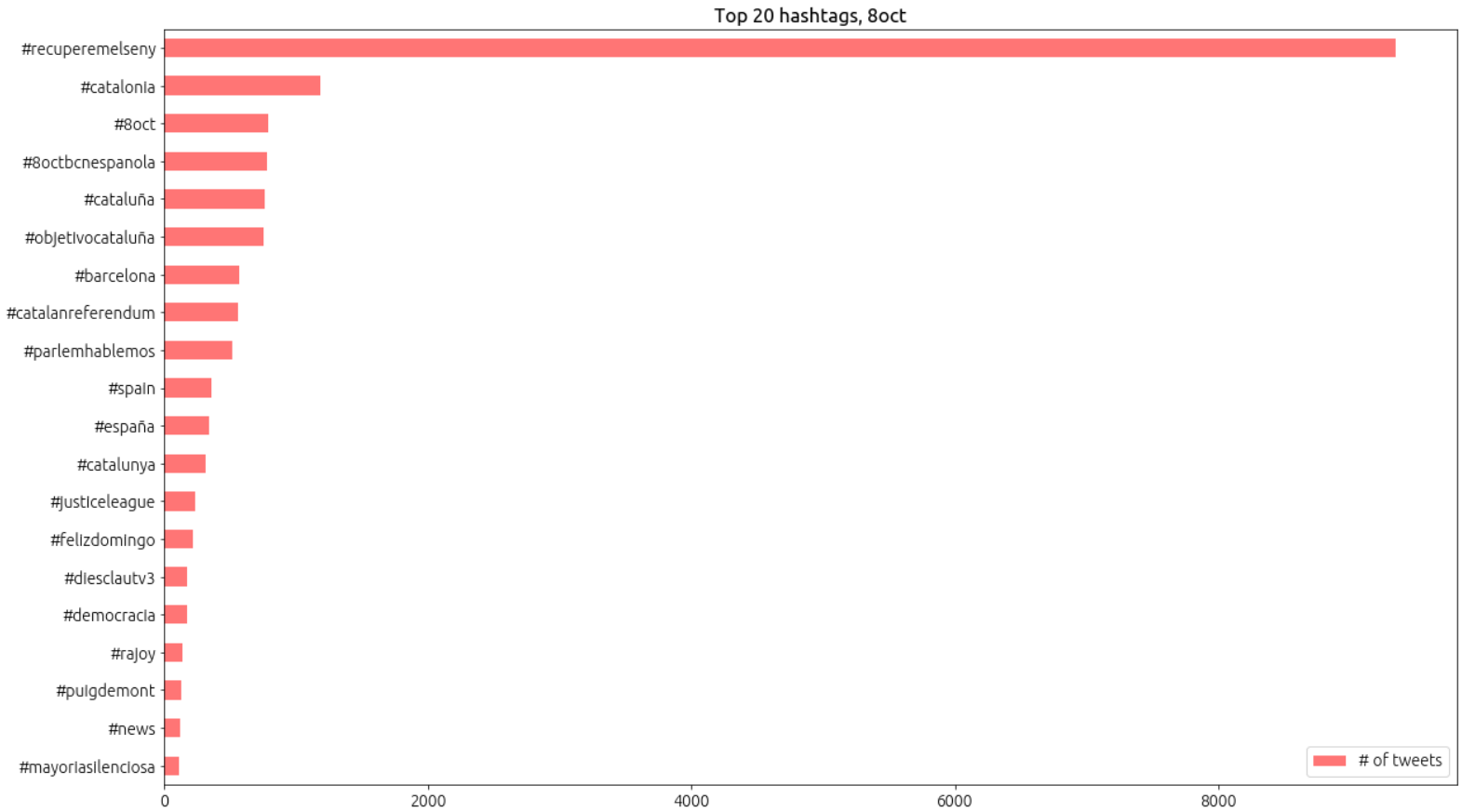


The variation in tweet activity and the difference between the two days could be explained with different nature of the events. The March on October 8th 2017 continued through the whole day, so it is understandable that the discussion about the march continued evenly throughout the day as well. Puigdemont's speech on the other hand was given in the evening. The speech should have started at 17:00 GMT, but was postponed by an hour. There is a peak in tweeting activity around 17:30 GMT. The activity seemed to gradually calm down after it.

Top hashtags

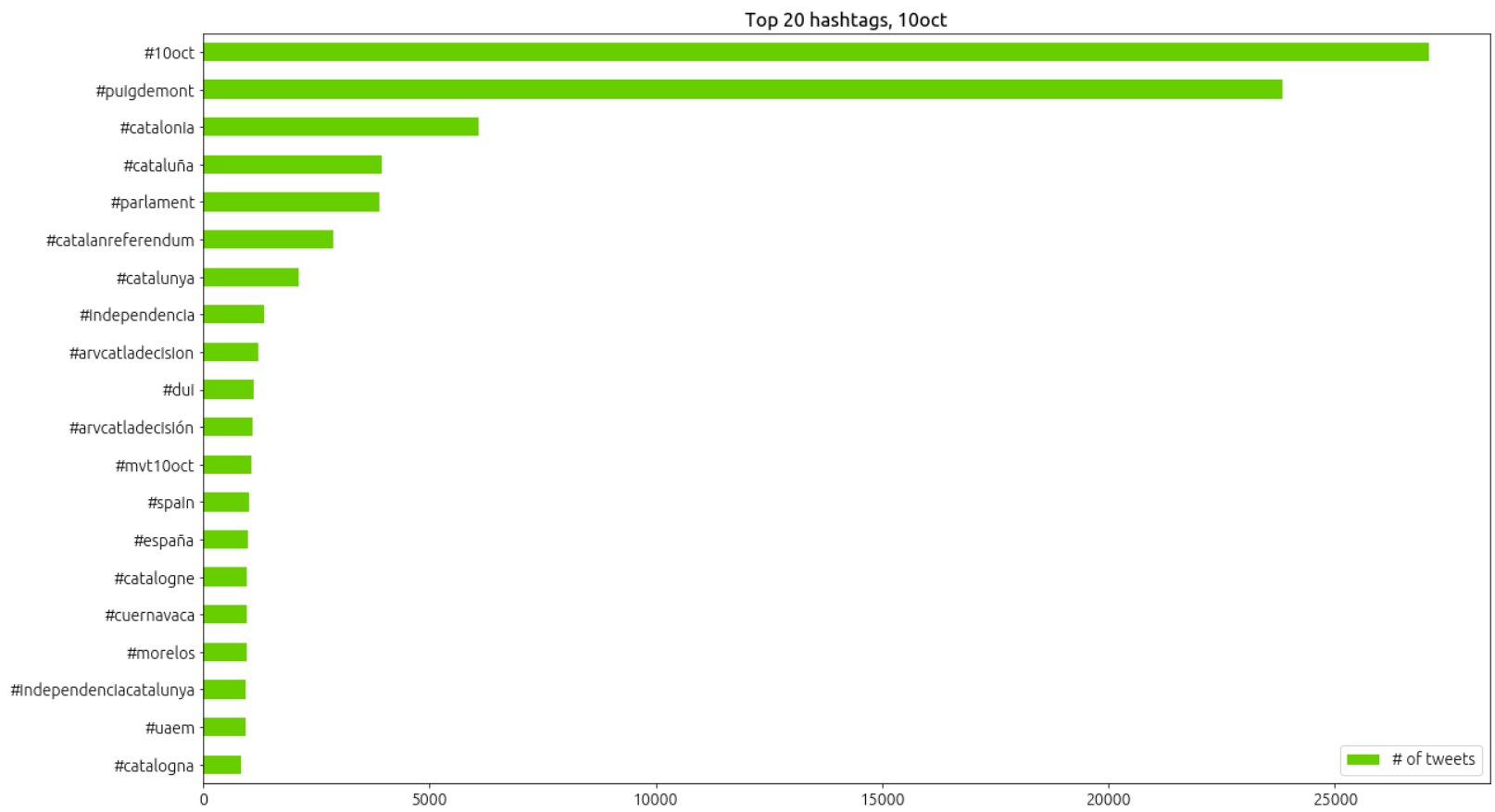
Next we wanted to take a look and see which ***hashtags*** were trending in both of these two days and what kind of differences could we find.

8th October



At the **8th of October** we can see that the slogan of the day, "*recuperem el seny*" which translates to "let's reclaim common sense", was also most often mentioned in the Twitter.

10th October

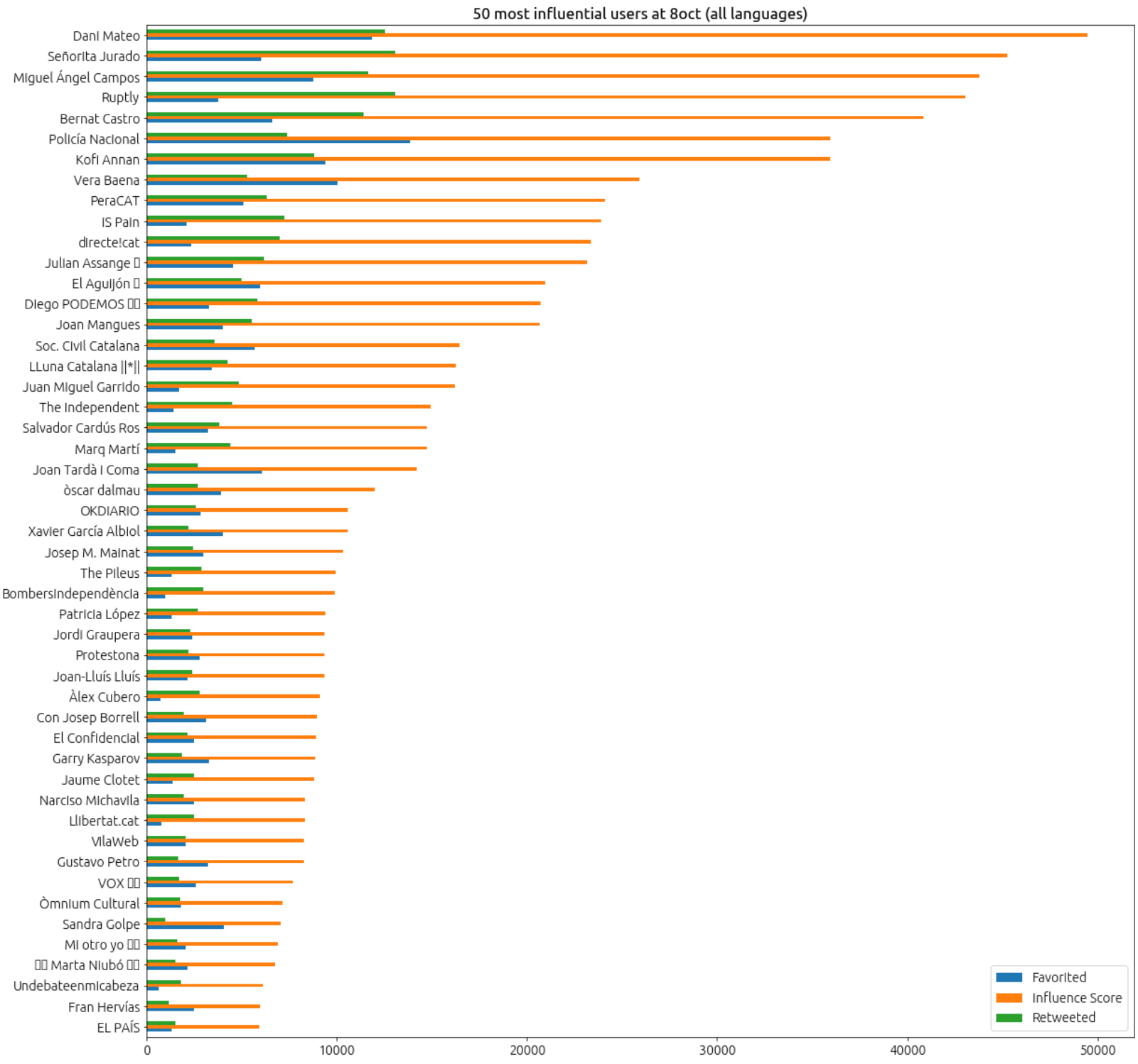


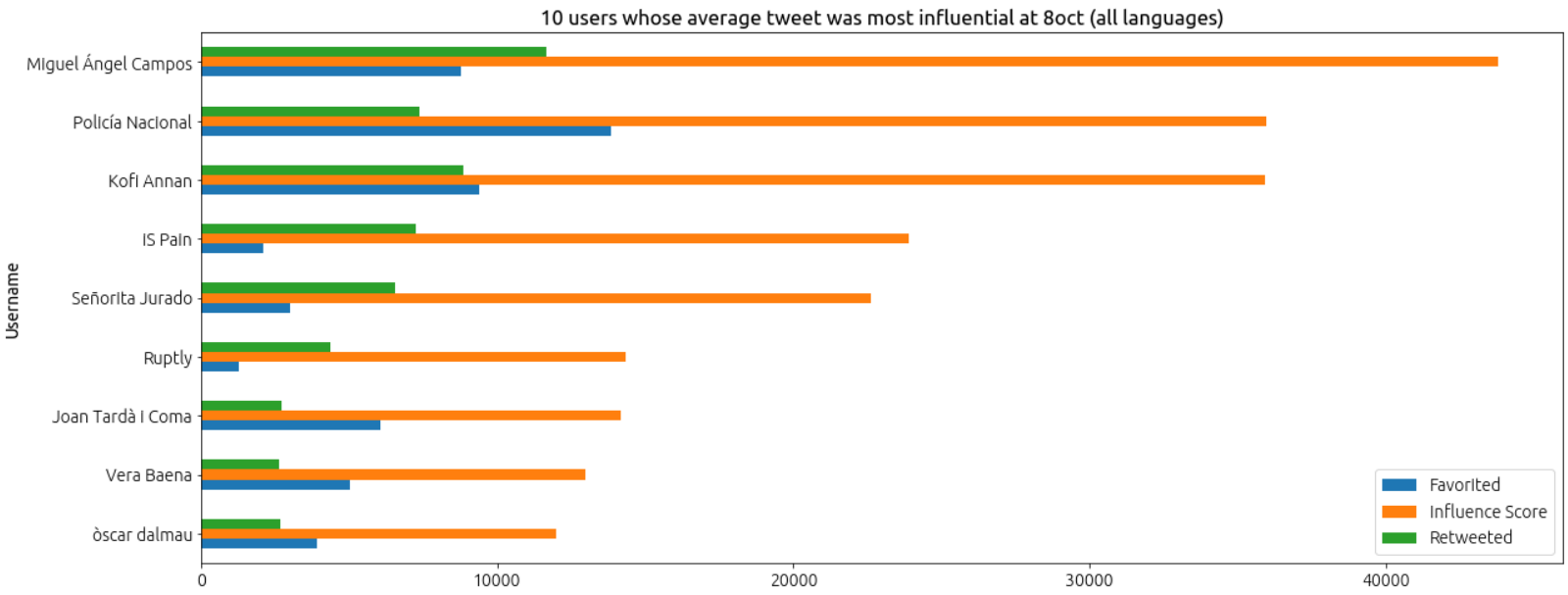
From the day of Puigdemont's speech the most popular hashtag in the data set was #10oct, #puigdemont coming second. The results were not too surprising and give a good overall image of the type of data we retrieved from this day. Many of the most frequently mentioned hashtags were also in our search query which of course affects the outcome.

Quick look at the users

We created a simple algorithm that calculated an influence score for tweets. This was based on likes and retweets that a tweet had received by the time we collected the data. By calculating averages and combining scores of tweets from individual users, we were able to find which users posted the "most influential" tweets during this time frame. This way we could also calculate which user gained the highest influence score throughout the whole day. It is however good to keep in mind that this algorithm is quite crude and does not take into account things such as numbers of followers for the users.

8th of October





Here we've provided two different graphs from the day of the pro-unity march, 8th October. For the first one we have summed up the influence scores of all tweets from a user, and for the second one we have calculated an average influence score. The graphs show different names because of this.

One hypothesis was that the most influential users would be mostly public figures or institutions. It was interesting to see that there was a good mixture of both private users and public figures among the most liked and retweeted users. The first list is of course easy to climb with a big number of posts with smaller influence scores. The most popular posts of the day on the other hand have most likely been posted by users on the second list. These figures were based on multilingual data.

Notable medias:

- Ruptly
- The Independent
- El Confidencial
- El País

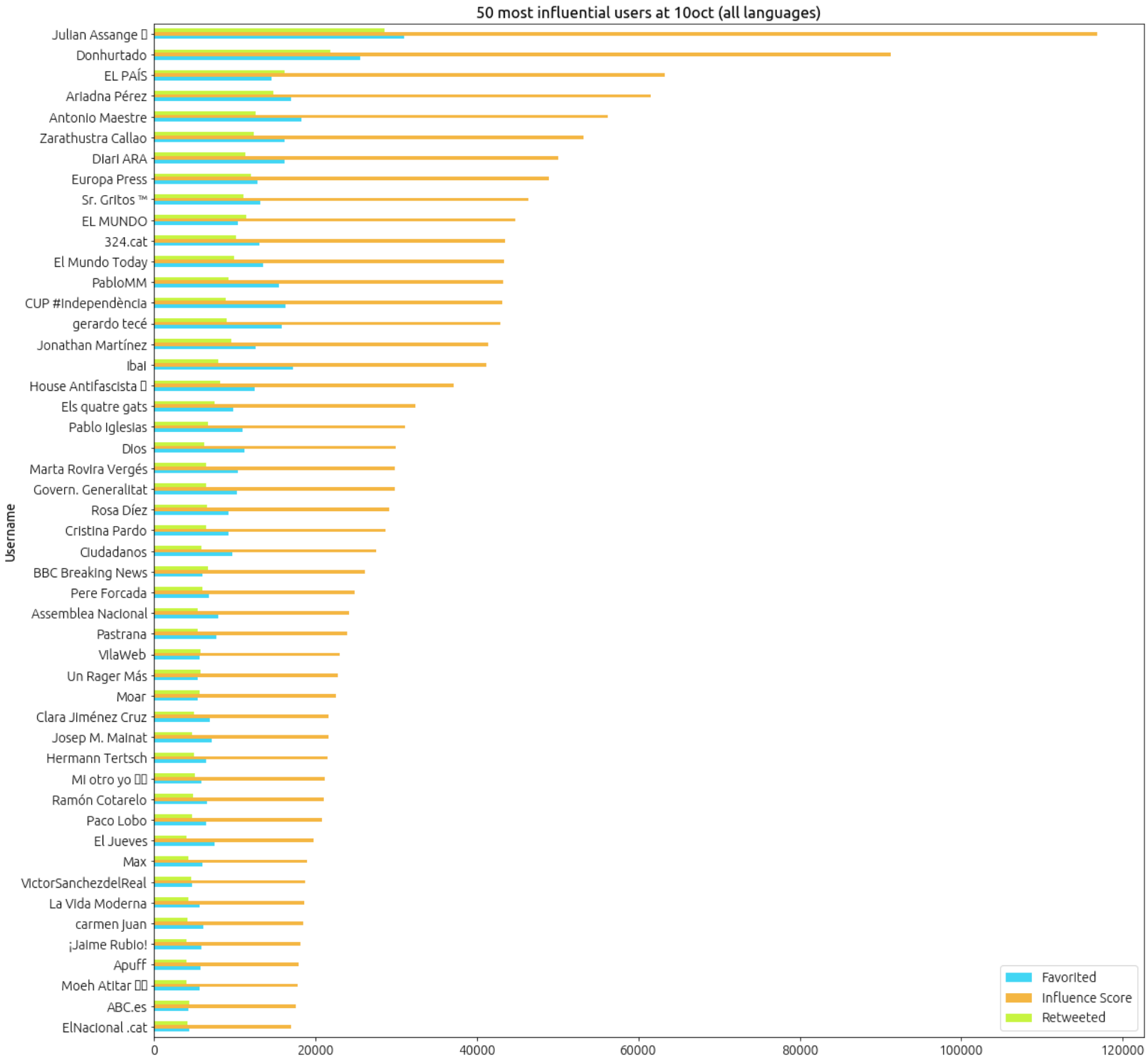
Notable political/other public figures:

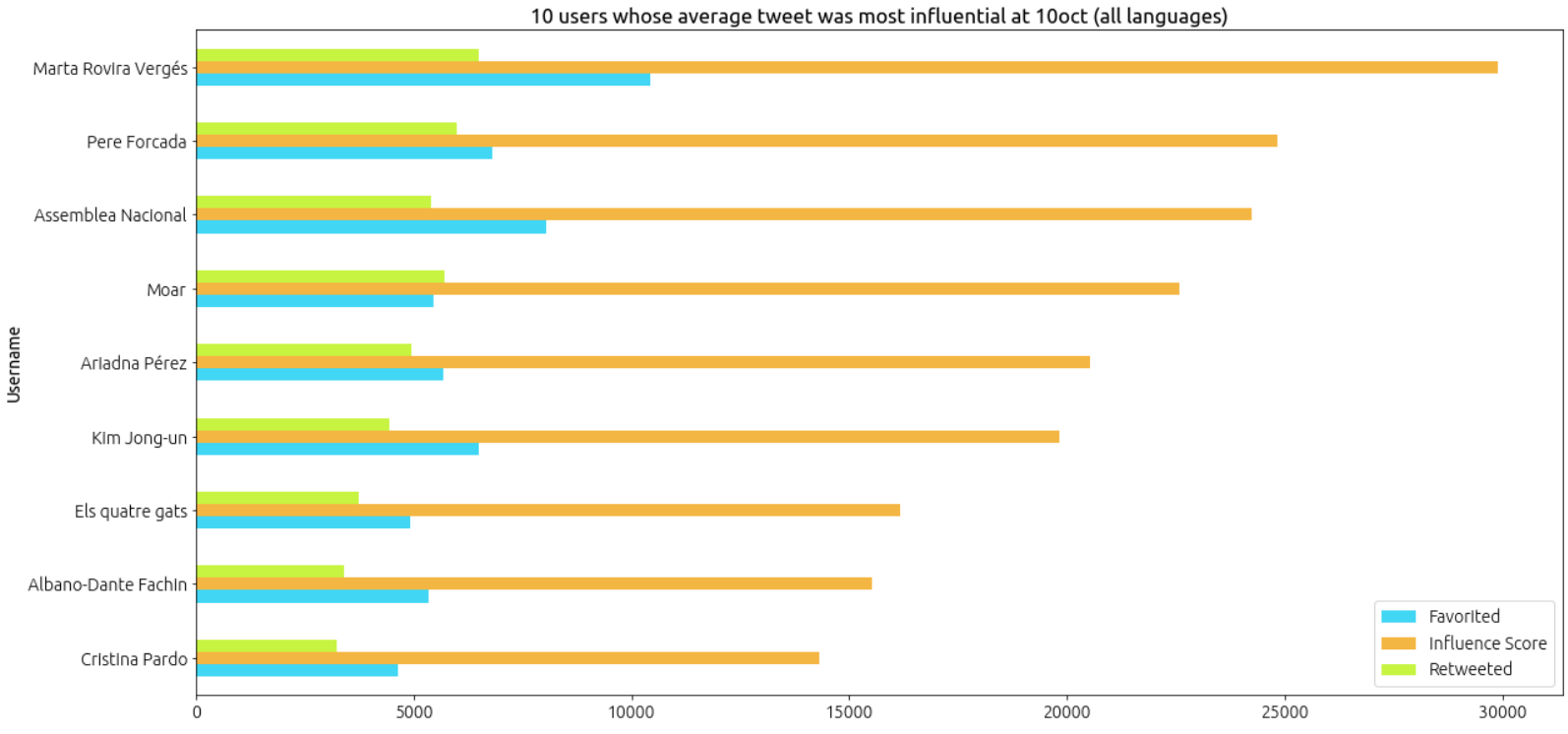
- DiegoPodemos
- JulianAssange
- ***Kofi Annan***
- SPAINonymous

Notable institutions:

- Policia
- Ciudananos (Political party)

10th of October





For 10th October the lists were slightly different, but again there was an interesting mixture of private users and public figures. What is interesting is that there seem to be a larger number of well-known Catalan independence activists among the top influencers, compared to Oct 8th. Julian Assange was hugely influential on this day.

From the 6th place of the top 10 influencers one can find and interesting name, Kim Jong-un. We would like to point out that this is a satire account. The supreme leader of North Korea is no doubt a highly influential figure, but does not use Twitter, as far as we are aware. :)

Another interesting point is that all three major Spanish newspapers, El Mundo, El País, and ABC, show on these lists. On 8th October only El País made it to the Top 50 influencers.

Notable medias:

- El País
- El Mundo
- ABC
- BBC Breaking News

Notable political/other public figures:

- JulianAssange
- Marta Rovira Vergés (Catalan pro-independence politician)
- Carme Forcada (Catalan pro-independence politician)
- Chumel Torres (Mexican comedian/satirist)

Notable institutions:

- Assemblea Nacional Catalana (Catalan National Assembly)

Top 50 tweeters with no original tweets

After providing some simple grouping and excluding original posts, we were able to list the top most active users, who only retweet other users tweets. As the results suggest **('CatalanRobot')** some of these could be bots, which only automatically retweet original tweets.

Possible bias in the data

Bots that retweet certain tweets in large quantities can affect not only the data itself, but also what hashtags are found popular by different popularity measures. Retweeting bots also affect the influence score so that a post seems to be influential even though in reality the retweets aren't done by "real people".

We noticed that some of the users repeat the same tweet, but post it as their original post, not as a retweet. This was mostly done by different news agencies (BBC, BBCNews, Reuters, AP, Guardian, Nytimes, BILD). It is common for agencies to share a piece of news they have all retrieved from the same source. This can however create a bias in the data, if it happens in large numbers. In our data this was not a huge problem because of the small amount of these repeated tweets, but it is something to keep in mind.

Sentiment Analysis

Sentiment analysis is the term used for analyzing the general sentiment from a piece of text. This can be done manually, but also by using computational means. At the most basic level Sentiment Analysis gives a polar value of Positive, Negative, or Neutral to a piece of text.

Manual classification

Initially we were interested in the idea of computational classificator for political sentiment. We attempted to do this by first classifying some of the English language posts manually in to two opposing categories: pro-independence and anti-independence.

For a significant portion of the posts we had difficulties with the analysis. Reasons for this ranged from simple ambiguity to cryptic symbolic meanings in the text. Some referred to the Bible, others required extensive knowledge of Spain's historical and political context.

We came to the conclusion that given the time we had available, the classified sample we could attain by manually inspecting the tweets wouldn't meet the size required for training a reliable machine learning model.

Computational methods

After the manual classification of political sentiment turned out to be a bit too time consuming for this project, we turned to computational methods.

Sentiment Analysis in Spanish

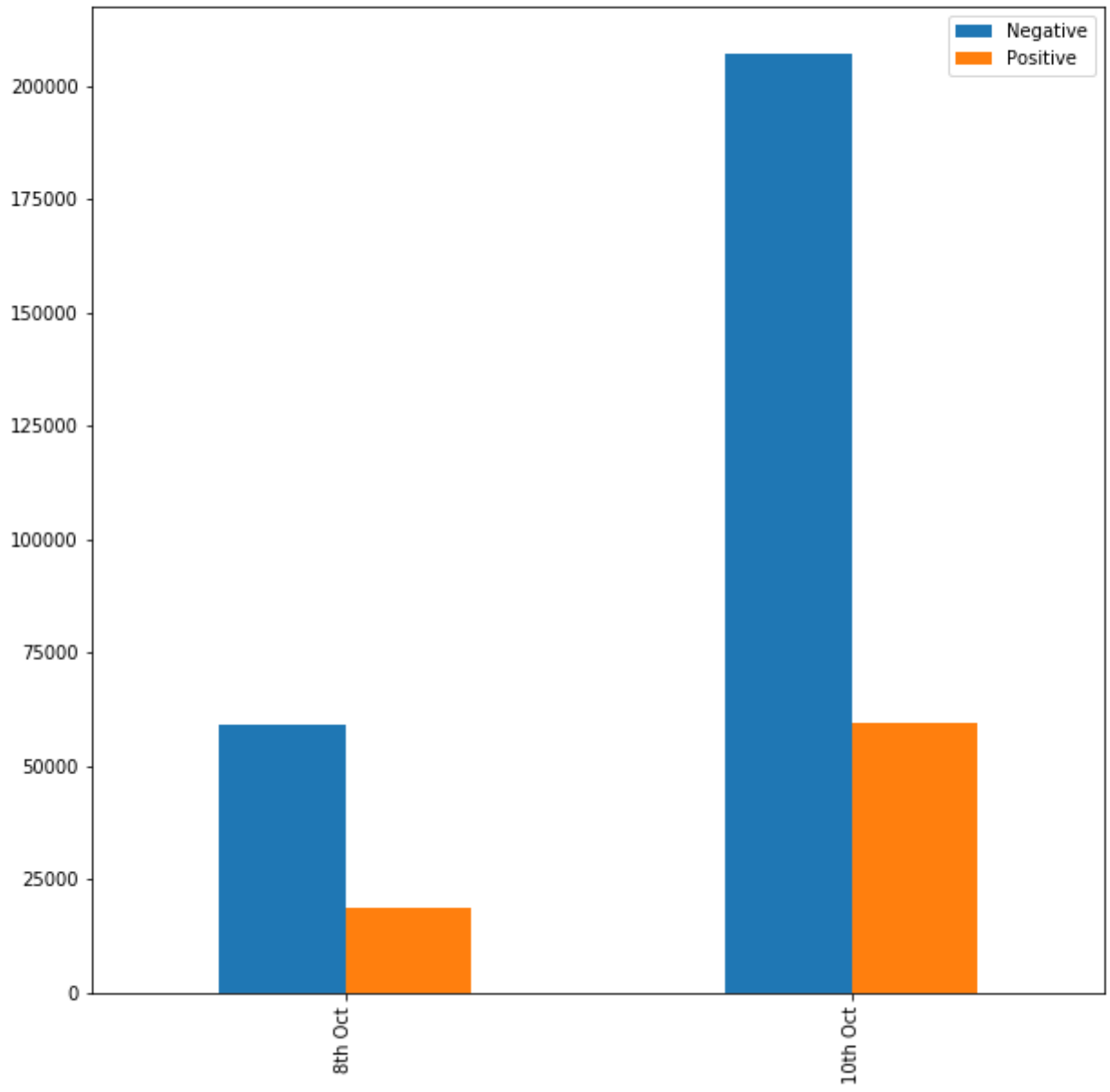
We were interested to find out about the possibilities of machine learning in political sentiment analysis. However as the most of the data was in Spanish we needed a Spanish sentiment analysis tool. For that purpose we obtained a data set from SEPLN which included labeled sentiment data from different areas of life. Using that data we trained our own machine learning model.

SEPLN offers a training data set labelled as "political" data. We first trained our model using this data, but unfortunately the data set turned out to be very small for this kind of a task. The accuracy score of the test run was only 0.64. We decided to combine all of the sets offered by SEPLN, from many different fields. After training the model with this much larger set, we were able to get a test score of 0.82, so the larger data set improved the accuracy quite remarkably.

One major problem with this training data was that the sources weren't clear about how the data was labelled and on which basis. This created quite a bit of ambiguity and unfortunately made the results quite difficult to interpret. Much of the inner workings were copied from this great blog post from Manuel Garrido. Sample of results can be seen here:

	text	polarity
8	Síntesis de la corrupción en Pilar https://t.co/s4tx54OLXY	0
13	Los catalanes no les importa perder su poblacion en un proceso de independencia que puede llevar a ser mas minoria etnica	0
14	#undiavasaganarvos \nEl domingo 22 apoyemos la boleta de @1PaisUnido, se necesita justicia social e igualdad de oportunidades!!	1
15	@Deisagamarra Y bueno...por lo visto te aferras a tapar el mal de los clubes y de la corrupcion reinante	0
23	Me gustó un video de @YouTube https://t.co/W5sogrFkTG Liga de la Justicia - Trailer Final - Subtitulado	0
24	Enoooormes valientes!!Gracias por acompañarnos @LuisSalvador, @veroprial, @amarlos71 💪😘😘 #Mejorunidos… https://t.co/7buF5gDHEn	0
29	Pues eso... el mundo no está ciego\n\nPro-Spain supporters make fascist salutes ahead of demonstration https://t.co/7QPekb8DjF via @MailOnline	0
36	¿Que en Cataluña se ha revestido todo de pacifismo y democracia? Sí. ¿Que eso oculta algunas bastantes actitudes fascistas? Las oculta.	0
40	@InesArrimades Se han tomado tu mensaje al pie de la letra. Es lo que pasa por pedir a catetos que integren la mayo… https://t.co/tKdmR1GKVt	0
51	Rajoy es el Bartomeu de España	0
55	Que pasaría si Cat se independiza y en las 1ras elecciones gana Junts x la reconeccio? 😕😕😕\n #ObjetivoCataluña @_anapastor_ #RecuperemElSeny	0
56	@Baxayaun El fascismo en España «murió» en la cama, y a sus simpatizantes se les amnistió en la Transición. A Españ… https://t.co/GZy48S0gMZ	1
75	Mientras, los medios españoles ni mencionan que había nazis en una manifestación. Todo era alegría y democrácia.	1
84	Abertis y Colonial deciden este lunes si trasladan su sede fuera de Cataluña https://t.co/vKjYDNE5cg vía @el_pais	0
86	LIBEREN A PATRICIO FONTANET \nHIJOS DE MIL PUTAS\nCON PRUEBAS FALSAS NO VA\nLUEGO HABLAN DE CORRUPCION Y EL GOBIERNO HACE CORRUPCION	0
95	@mirasanlucar Fascistas estos. Que vergüenza y después dicen que quieren que Cataluña se quede.\n😡	0
96	'Fin de la hegemonía' (#Cataluña despierta!!!) https://t.co/UmobNbqAas vía @el_pais #RecuperemElSeny #EstadoDeDerecho #EstamosPorTi	0
103	@jagolo82 @Fran_Iliones @Agurvenurs @davmiranda @policia Pero no legitime q un gobierno regional se salte las leyes… https://t.co/niQ0x6yDZB	0
130	Una quincena de grandes empresas se llevan la sede social de Cataluña en una semana https://t.co/rGeTTaeu4C	0
137	Siguiendo #ObjetivoCataluña. Qué bueno sería que debates como este tuvieran cabida en las cadenas públicas @rtve y @tv3cat #ParlemHablemos	1
147	@prosperoventura Pero hay que invertir otros millones en campaña pro corrupcion	0
150	Vaya país, esos que hoy buscan revertir el avance lucha contra la impunidad y la corrupción pueden hoy manifestarse por esa misma lucha.	0
158	#ObjetivoCataluña #Puigdemont y lo está consiguiendo!!!!!!!!!!! https://t.co/C3W8ypadgf	1
168	El artículo 8 de la constitución, ósea serán detenidos por golpistas, ósea q si Tejero se pasó 25 años preso q vaya… https://t.co/RbreUO9kdT	0
169	#ObjetivoCataluña "El Rey es una persona joven, viajada...". Si y muy formada, ya. Pero es de derechas. como Rajoy.	0
172	CUATRO CASOS AISLADOS‼!#ObjetiboCataluña #8Oct #CatalanReferendum https://t.co/YyHRKzNNx3	0
174	@elferranet @sanchomdv Sin DUI hay creo posibilidades que Rajoy acepte un referéndum.	0
180	@bbcmundo Si...lo llevará tan lejos\nque tendrán que recoger los pedacitos de Cataluña hasta en la Luna.	1
183	A la salida del Palacio de Justicia y con sus respectivas boletas de libertad en mano se retiraron los periodistas https://t.co/LaYKHlTLrZ	0
193	@momouruguay Y por justicia divina Perú se quedará sin mundial. Y desgraciadamente Chile a repechaje.	0
194	@adnradiochile @alvaroramis La tasa delictual, los más casos de corrupción, mirar la educación como bien de consumo… https://t.co/o7XmffAeY6	0
198	El rey quedó como parte del problema y no de la solución. Un proyecto con Cataluña pasa por poner fin al régimen del 78. #ObjetivoCataluña	0
203	Rajoy después de demonizar a los catalanes, se está preparando para después de la DUI culparlos de la crisis terrible que sufre España.	0
206	@Pevelasco No creo q pueda soportar una semana mas ni de calor ni de Puigdemont...q vengael otoño las lluvias y q g… https://t.co/fmgQSRU4AS	0
248	ESTOS ACTOS Y MAYORES SI ES POSIBLE..SIN VIOLENCIA PEERO MASIVOS Y EN TODO EL PAIS Y TODA CATALUÑA ..HASTa LA REFOR… https://t.co/W3sss0ODbc	0
272	Y dale en la Sexta con el referéndum pactado. Que el Gobierno no puede pactar la soberanía, pesados. Y UK no tiene Constitución escrita	1
287	#objetivoCataluña ESTA CLARO ESTA ESTABA EN CADAQUES TOCANDO L GUITARRA CON TRAPERO RAHOLA PUIGDEMONT LAPORTA K OBJETIVIDAD D UNA PERIODISTA	0
290	Claves referéndum Cataluña: El día en que TV3 demostró que es la televisión de solo una parte de Cataluña\nhttps://t.co/oVfLv7clc3	0

Here we can see the distribution of sentiment in the Spanish tweets:



Sentiment analysis in English

For English there is a great variety of different ready-made tools available for natural language processing, and indeed sentiment analysis as well. We decided to compare two different commonly used sentiment analyzing tools with this data, called TextBlob and Vader.

- TextBlob is a text processing tool that has simple sentiment analysis as one of its features.
- Vader is a sentiment analyzing tools specifically created for social media texts. It's advantage is that it analyses not only words, but also punctuation marks and for example simple emoticons.

Both of the analyzers gave quite high numbers of "neutral" labels, Vader less so. Of course it is natural to have tweets with neutral sentiment, as not all tweets really have a positive or negative attitude. The amount of neutral labels was so however so great, that we decided to take a closer look at the data. After applying both classifiers to one of the data sets we observed that Vader provided more accurate estimates that TextBlob.

	text	polarity_blob	polarity_vader
2	Why is the _Commission turning a blind eye to this?	-0.500000	-0.4019
6	Britain needs to lead, why isn't May speaking up for Catalonia? We should encourage others to LEAVE and j...	0.000000	0.3328
10	Images on fascists for the union in Barcelona. Sad to call for union of Spain against independence of Catalonia, by...	-0.500000	-0.5994
13	Catalonia's Independence Referendum, in Photographs	0.000000	0.0000
28	Wishing to see the independent circus Cataluña broke & misery This is what bastards deserve!!	0.000000	-0.8802
37	And the material conditions of Armenian proletariat has dropped. "but at least there is independence /s" 😑	-0.300000	0.0000
57	Catalan President Puigdemont has 'independence at his core' via JusSwaggTV	0.000000	0.0000
65	Catalonia yes, Catalonia not, but, meanwhile, the central gang of thieves is still stealing.	0.000000	-0.8641
70	Pro-Spain supporters make fascist salutes ahead of demonstration via	0.000000	-0.0772
72	Now they chose to make photos and become bystanders while one week before they chose to fire bullets at people will...	0.000000	-0.3400
74	Spanish unionists find their voice in huge Barcelona rally - ABC News	0.200000	0.3182
86	Hold on, here's a better video of those intrepid human tower builders from Catalonia in Reus (and...	0.250000	0.4404
98	Thousands protest in Barcelona against Catalonia's independence from Spain	0.000000	-0.2500
112	there were a couple sets of fireworks that went off and it was CRAZY. right before they announce independence	-0.157143	-0.4824
113	Spanish violent people against the intependence go to Barna.Catalans want th independence from Spain.See why(this i...	-0.400000	-0.5574
122	First Catalunya declares independence, they, if they overcome repulsion, will talk	0.250000	0.0000
131	Franco seems to be alive and well... EU turning a blind eye to it all ..	-0.200000	-0.0258
139	Just heard John Swinny state that the SNP will now concentrate on a new independence campaign . Well as a dissalusioned yes voter .never	0.136364	0.5859
142	Hundreds of thousands rally against Catalonia secession vía	0.000000	0.0000
149	is not Spain.	0.000000	0.0000
158	Spain's Colonial calls board meeting for Monday to discuss moving head office from Cata...	0.000000	0.0000
161	Catalonia: hundreds of thousands join anti-independence rally in Barcelona	0.000000	0.2960
168	Catalonia will apply referendum law calling for independence declaration: leader	0.000000	0.0000
178	Google Top stories: Spanish unionist...	0.250000	0.2023
184	Catalan President Puigdemont has 'independence at his core' - He's either a freedom fighter, the defender of th...	0.000000	0.7351
196	You were fooled about Catalonia by mainstream media.This is Barcelona today. This is SPAIN 🇪🇸 We love our country❤ \n	0.500000	0.3818
199	Catalan President Puigdemont has 'independence at his core' - He's either a freedom fighter, the defender of th...	0.000000	0.7351
204	Catalan President Puigdemont has 'independence at his core' - He's either a freedom fighter, the defender of th...	0.000000	0.7351
207	NYT: 'I Am Spanish': Thousands in Barcelona Protest a Push for Independence	0.000000	-0.2500
208	NYT: Asia and Australia Edition: North Korea, Catalonia, Hurricane Nate: Your Monday Briefing	0.000000	0.0000

Conclusions

After applying some quite simple data wrangling and visualization techniques on two data sets from Twitter we were able to show and analyze tweeters activity in time, most active users, the users with highest influence score, presence of bots, which only retweet, medias presence and activity, as well as to detect the activity of some public persons and institutions. Using more complicated classifiers we were also able to provide sentiment analyses in two languages – Spanish and English. We have applied one classifier for posts in Spanish, and 2 for English text. The results showed different, but still useful accuracy.
These results confirmed our expectations that it is possible to obtain insight from fast paced social media such as Twitter, concerning political and sociological developments.

Possible directions for future work

One challenge that was revealed during manual analysis of the tweets: how to use computational methods to decipher a tweet's political stance? This would be an interesting challenge for a future project.

Soon after beginning this project we realized how incredible the possibilities for a data analysis are with data like this. Here we show only a small collection of

examples of ways to analyze Twitter data, but it could be stretched to much more.

> *"We just need two more weeks..."*
>
> *-Anonymous Team Member*

More detailed information and also the scripts concerning data scraping, wrangling, and other techniques used in our exploratory data analyses are available on https://github.com/DeepIntuition/DataScienceProject2017 .

Data Science Team

Read more posts by this author.

Share this post