BIG DATA

DESIGN OF A NEW INTERACTIVE DATA ANALYSIS TOOL

David Burrell
Ignacio Martínez
Rodrigo Pueblas

January, 2020

1. Introduction

The objective of this work is to develop a visual analysis tool for a dataset of Tweets. The dataset can be accessed here [1]. The dataset holds a collection of Tweets collected between the 14th and 22nd of April of 2016 in the countries of Spain and Portugal. The visual analysis tool developed will allow users to explore the data in three ways:

1. Space: Where do people Tweet from?
2. Time: When do people Tweet?
3. Content: What do people Tweet about?

2. Data pre-processing

2.1. Data description

The chosen dataset is made of 205,789 Spain and Portugal geolocated Tweets. The dataset has 19 columns, which are shown in Table 1. The Tweets have been captured every minute, starting the 2016-04-14 at 09:18am and finishing the 2016-04-22 at 12:58pm. The Tweets have been gathered from 86,310 accounts and are written in 13 languages.

2.2. Data cleaning

The first step involved in the dataset manipulation is the removal of attributes that will not be used for the visual analysis. Due to the high volume of Tweets, reducing the dimensionality of the data improves the completion time of calculations.

The attributes dropped have been: Tweet Id, User Name, Nickname, Bio, Favs, RTs, Country, Place (as appears on Bio), Profile picture, Followers, Following, Listed, Tweet Url. The dataset has gone from having 19 attributes to being composed of 6.

The final pre-processing step has been to remove those Tweets which null values for the Latitude and Longitude attributes.

Table 1 *Summary of the dataset attributes.*

Attribute	**Description**	**Datatype**
Tweet Id	Tweet’s unique identifier	Numeric
Date	Day the Tweet was posted	String
Hour	Time the Tweet was posted	String
User name	User’s name	String
Nickname	User’ alias	String
Bio	User description	String
Tweet content	Message contained in the Tweet	String
Favs	Number of Favs the Tweet had when it was collected	Numeric
RTs	Number of RTs the Tweet had when it was collected	Numeric
Latitude	Latitude coordinate from where the Tweet was posted	Numeric
Longitude	Longitude coordinate from where the Tweet was posted	Numeric
Country	Country from where the Tweet was posted	String
Place (as appears on Bio)	Geographical information contained in the Bio	String
Profile picture	User’s profile picture	String
Followers	User’s total number of followers	Numeric
Following	Total number of people the user is following	Numeric
Listed	Number of lists the user is present on	Numeric
Tweet language (ISO 639-1)	Language of the Tweet	String
Tweet Url	Tweet’s URL address	String

3. Question 1: Where do people Tweet from?

3.1. Problem characterization in the application domain

The application user will be interested in knowing the locations where people are tweeting from. This allows them to see how the Twitter users are distributed geographically and observe any clusters that appear in certain areas or if they follow a random distribution. Furthermore, the user can also be interested in seeing if the spatial distribution is static or if it changes with time, and may only want to focus on the distribution of certain topics and not the content of all Twitter messages. The geographical attributes (Latitude and Longitude), the temporal ones (Date and Hour) and attributes related to the content (Tweet language (ISO 639-1) and Tweet content) will be the most useful ones for constructing a visualization that answers this question.

3.2. Data and task abstractions

The visualization that answers this question will be used to discover how Tweets are geographically distributed. This will allow the user to verify if his hypothesis about the Twitter users being clustered, which makes sense given that, for example, cities like Madrid and Barcelona are more densely populated than other Spanish areas. The search of elements of interest is therefore a lookup, since the user knows what he is looking for (how Twitter users are geographically distributed) and where to look for it (the geographical distribution of Tweets). Thus, the query is the comparison of the density of Tweets in the geographical area covered by the dataset.

The visual encoding used for this purpose should encode the Latitude and Longitude using the two most important channels. They can be encoded using spatial position combined in a two-dimensional matrix alignment of the marks, acting as key attributes. The marks in this case, could be encoded as areas of a colour representing the number of Tweets in those coordinates. This way, the number of Tweets, the quantitative attribute, is encoded using hue. Since the hue is a channel fully separable from position, both types of information do no conflict.

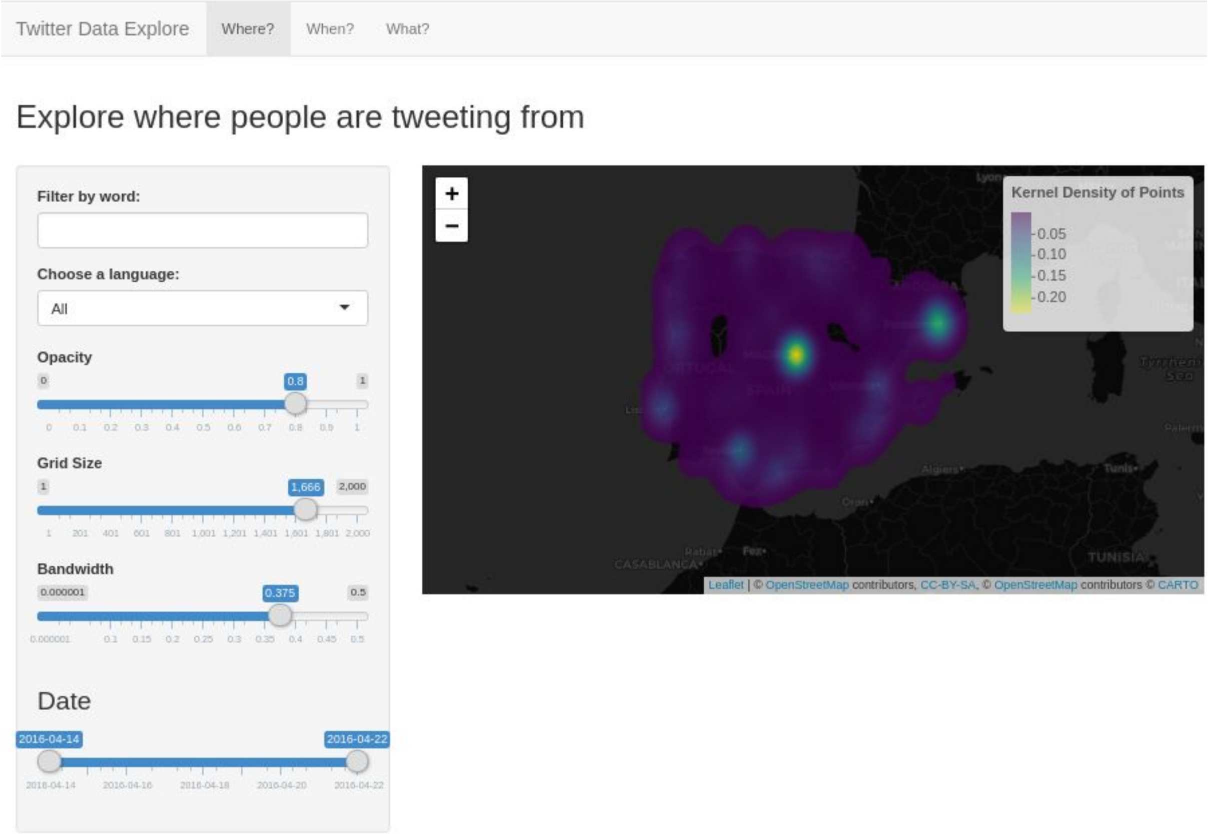
3.3. Interaction and visual encoding

The visual idiom chosen to encode the visualization is a heatmap. The Latitude and Longitude are encoded using spatial position as a two-dimensional matrix of area marks coloured according to the number of Tweets. This idiom allows the user to find clusters of points that are highlighted against the colour of the less dense zones. The heatmap is superimposed on a map of the Iberian Peninsula, the area where the Tweets have been collected, so that the user can match the clusters with real locations.

Concerning interaction, it could be useful if the user could adjust the visualization by filtering the Tweets by word in the case he is interested in a particular topic, as well as choosing the language of the Tweets that are processed. The user may also want to filter the dates that are taken into account, making it possible to analyse all the time span available or focus on a shorter interval.

3.4. Algorithmic implementation

As there were some outliers in the data that are outside the longitude and latitude of the Iberian Peninsula the dataset for displaying on the map is filtered to contain those within the desired coordinates. Then using the bkde2D function available in the KernSmooth R package a 2D binned kernel density estimate is computed across all the coordinates. It is the results of this that are imposed on to the real world map to visualise to the user where the highest concentrations of Tweets originated from. Scale used to encode the number of Tweets is encoded using a sequential colour-blind friendly palette [2]. As this pallet increases with lightness as the number of Tweets increases it quickly draws the users attention to "hot spots" of activity. To increase this effect the "real world" map is generated using the leaflet package in R. Not only does this generate the map using the same longitude and latitude coordinates to ensure maximum accuracy, it allows for the darkening of the background colour to enhance the brightness of the kernel density scale as seen in Figure 1.



Filter by word:

Choose a language: All

Opacity: 0.8 (0 – 1)

Grid Size: 1,666 (1 – 2,000)

Bandwidth: 0.375 (0.000001 – 0.5)

Date: 2016-04-14 – 2016-04-22

Figure 1 Screenshot of application showing heatmap of where Tweets were sent from.

To allow for better visualisation depending on the user's needs they have access to modify the inputs to the kernel density function. By increasing the grid size they are able to have a more granular view as more grid sections (buckets) are generated. While increasing the bandwidth will smooth the approximation of the distribution, increasing the size of each of the grid sections. The user can also zoom in and out and reposition the map to focus on a specific area within the map.

To allow for the filtering of the data several controls have been added. First, a text input box that allows the user to search for a specific word, filtering out all which do not have the given string in their content. If no Tweets are found to contain the user is informed that this is the case. Additionally the user can filter on the language of the person tweeting using a dropdown with the different languages available in the dataset. By filtering this way they can see how people from different regions are distributed throughout the region. And the final filter is a date slider. Using this the user can select the date range of the tweeting to investigate if over time the “hot spot” location change.

4. Question 2: When do people Tweet?

4.1. Problem characterization in the application domain

The user will be interested in knowing when people are tweeting. That is, how the frequency of Tweets varies through time. This allows the user to identify trends in the Tweets time distribution and to see if there are times of the day when Twitter users are more or less active. The temporal information present in the dataset, the Date and Hour attributes, will be the most useful for constructing a visualization that answers this question.

4.2. Data and task abstractions

The visualization that answers this question will be used to discover how Tweets are distributed in time. The task in question is a lookup search, since the user knows what to look for (how Tweets are distributed in time) and where to look for it (the Tweets time distribution). Therefore, the query is the identification of the trend in the Tweets time distribution.

The visual encoding used for this purpose should express the time information and the posting frequency in the two most important channels. Both are numeric attributes (time information is ordered and number of Tweets is quantitative) so the use of spatial position to encode them will be the most effective. Also, the possibility of filtering the data would help in exploring the data: the user should be able to have an overview of the whole time distribution or focus on a specific interval.

4.3. Interaction and visual encoding

The visual idiom chosen to encode the visualization has been a line chart. The ordered attribute, time, is encoded using the spatial position in a horizontal axis. The values are ordered in horizontal regions using the value as key. The quantitative attribute, the number of Tweets, is encoded using the aligned spatial position in a vertical axis. The marks used are points, which are connected between horizontal positions with lines. This linkage helps to identify trends in the data.

Concerning the user interaction with the visualization, a way of filtering the time span that allows the user to filter the time interval he wants to focus on would be useful. This way, he can zoom in and out, obtaining an overview of the time trend across multiple days, or see how the Tweets are distributed in single days. Also, since the line chart is composed of a high number of points, a way of displaying information about each one if the user hovers over the line chart would help to increase the understanding of the data.

4.4. Algorithmic implementation

The first step was to keep only the attributes necessary for the task. In order to do so, a new attribute that combines the Date and Hour is created, called times. Then, a new variable called time.series is created that aggregates the Tweets based on the times and Date attributes. This way, the time.series variable becomes a data frame with three columns: times (date and time that the Tweet was posted), Date and Tweets (the number of Tweets posted at a given date and time).

The next step is to generate the line chart. First, the minimum and maximum date of the time interval under study are obtained. Then, the time.series is filtered to include only the rows between those two dates. Finally, the time.series data frame is used to make a line chart with the ggplot2 package. The x-axis contains the time information (times attribute of the time.series data frame) representing one minute intervals (starting the 2016-04-14 at 09:18am and finishing the 2016-04-22 at 12:58pm by default). The y-axis encodes the number of Tweets that were posted during each one minute interval (Tweets attribute). Finally, the x-axis label is composed of the Date attribute. The final result can be seen in Figure 2.

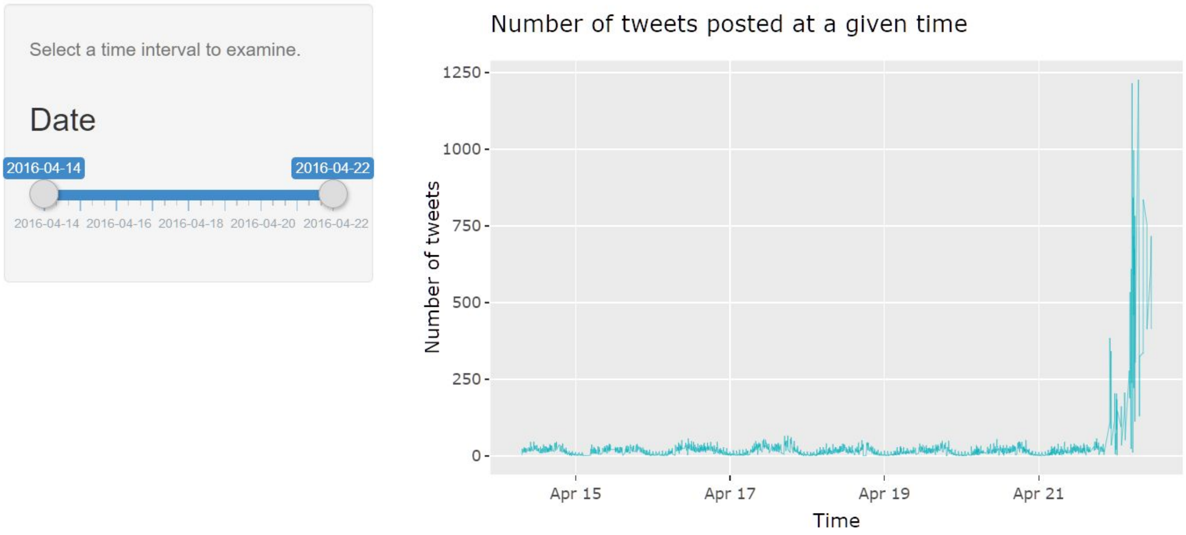


Figure 2 *Time chart displaying the time distribution of Tweets across eight days.*

The user can interact with the plot in two ways. The first one consists in adjusting a slider that controls the value of the minimum and maximum dates of the time interval under study. This allows the user to explore the whole time distribution or zoom in and analyse a shorter time interval in depth (see Figure 3). The whole visualization is redone every time the user adjusts the slider. The second way of interaction, is by obtaining detailed information of each point of the line chart by hovering the mouse over it. A popup appears showing the date, time and number of Tweets corresponding to the mark in question. This has been achieved by using the plotly library to display the ggplot2 plot.

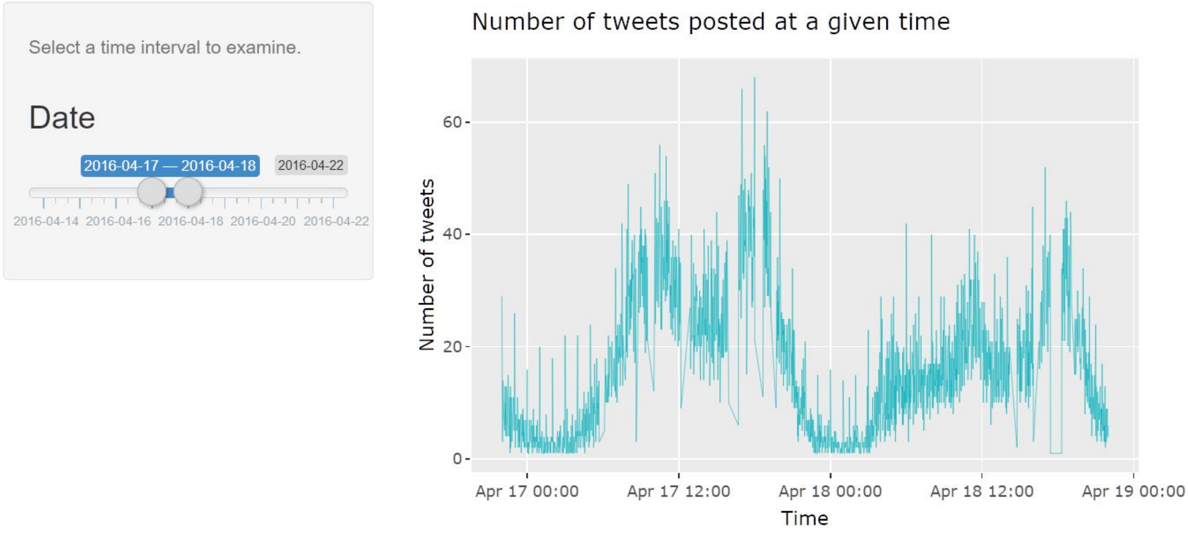


Figure 3 *Time chart displaying the time distribution of Tweets across two days. Notice that the slider has been used to adjust the date interval and that the plot offers a more detailed view than the one shown in Figure 2.*

5. Question 3: What do people Tweet about?

5.1. Problem characterization in the application domain

A user will be interested in what was the content of the Tweets to understand what were popular topics at the time the data was captured. As there is a large amount of text across all the Tweets this information needs to be distilled into a format that can be easily digested by the application user. So rather than using entire Tweet content only hashtag words will be used. In Tweets words that have a # character at the beginning are used to describe the topic or theme of Tweet. This allows for them to be categorized on the Twitter platform. The application will use these to show the user what topics are popular within the dataset.

5.2. Data and task abstractions

The task the user wishes to perform is a discovery task, where they find out what hashtags are being used within the dataset. Then they require the ability to browse through the data, looking for words that have high usage. This means they have a “one attribute” target, the individual count value of each word. The data is categorical, of attribute type string and attribute semantic of being a real world topic. It should be noted though that the meaning of a topic is given by the person who created the Tweet, and may not be used in a way that is obvious to the application user.

5.3. Interaction and visual encoding

The data showing the frequency of different words is shown via two idioms: a word cloud and a bar chart. The word cloud uses the size and position of each word to indicate its frequency, with larger more central words being used more often. By using this technique an application user can quickly see what words are most common. Although this technique very quickly allows a user to determine the most popular words it can be hard for them to determine the actual number of uses of each word. To make up for this shortcoming in the word cloud is accompanied by a bar chart showing the counts of each word. This idiom uses ordering (from most to least frequent) to allow the user to discover the most frequent word(s) quickly. Then, by aligning the bars they can see, approximately, proportionally much frequent different words are to each other.

To allow users to be able to explore the data more efficiently some filtering is also required. Based on the features available in the dataset the available features are the user's language, the date of Tweet and the maximum number of words to be displayed. To keep the information as useful as possible, only the top 10 hashtagged words are displayed on the barchart. If a user is interested in the count of other words the library used will display the frequency count when a user hovers over the word.

5.4. Algorithmic implementation

To make the logic calculating the word frequencies as efficient as possible a new feature "HashTags" was added to the dataset. This holds all hashtags within a Tweet's content. By preprocessing this data it means the application will be more efficient at run time.

Then when generating a word cloud there are two steps. Firstly the frequency of each word needs to be calculated. To do this the hashtags feature column is converted into a document corpus using the tm package in R. Then transformations are performed to remove undseriable tokens such as blank space and special characters. The # is included in characters that are removed as the user already knows they are in use so including them in the displayed words would be redundant. Then special words strings "https, "tco", "..." and "com" are removed as they were found to have high counts but did not have any meaning. Once completed the tm library is used again to generate a frequency matrix with all the unique words found and how many times they were used.

Then R library wordcloud2 is uses the generated frequency matrix to create the word cloud, using the required shiny methods for it to be displayed (Figure 4). To implement the barchart plot the ggplot library was used taking the first 10 values, or all if there is less than 10, from the terms matrix and displaying them in ascending order of frequency (Figure 5). Both the words in the word cloud and the bars in barchart use the same colour. This is so that user does not get confused if the word cloud words are in one colour and the bars in another.

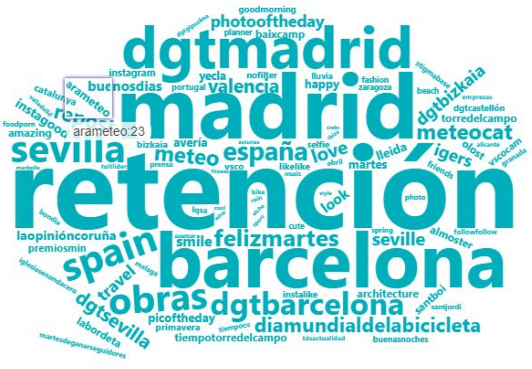


Figure 4 *Output of word cloud.*

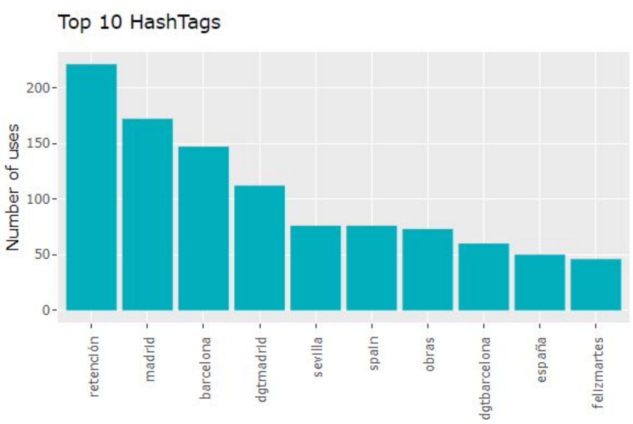
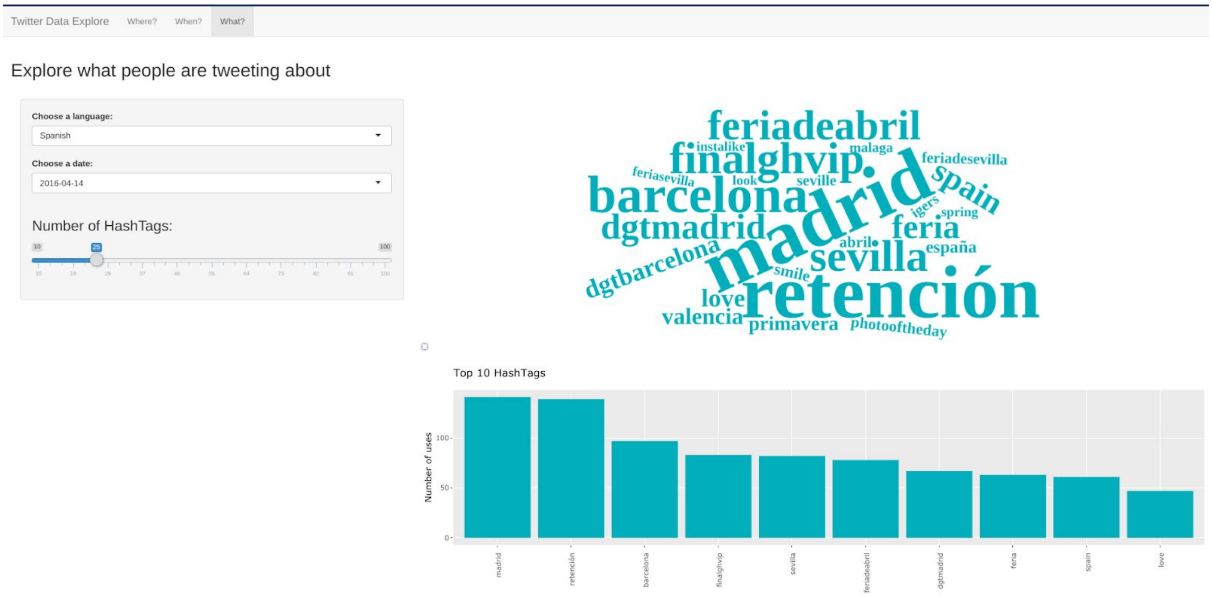


Figure 5 *Output of barchart.*

To allow for the browsing experience the user has access to three filtering controls. The first is language, this is a drop down populated with all the languages of the Tweets. The next is date which is a drop down with each section corresponding to one of the days available in the dataset. The final is the number of hashtags. This is a slider that allows the select the number of words displayed in the word cloud, for example if the slider is set to 33 then the the 33 most frequent words are shown in the cloud. The slider ranges from 10 to 100, where any number can be selected in between. 100 was selected as the maximum as any higher and words that have very low counts (including 1) are shown. These values do not meet the requirements of the application to show the top Tweeted hashtags. As the implemented logic recalculates the word frequencies each time the language or date filters are changed the results are cached using the R memoise package. This reduces load times if a user want to revisit a specific filter combination. Changing the number of hashtags does not require a recalculating of the terms matrix, it just returns the same data but changing the number of words returned.



Explore what people are tweeting about

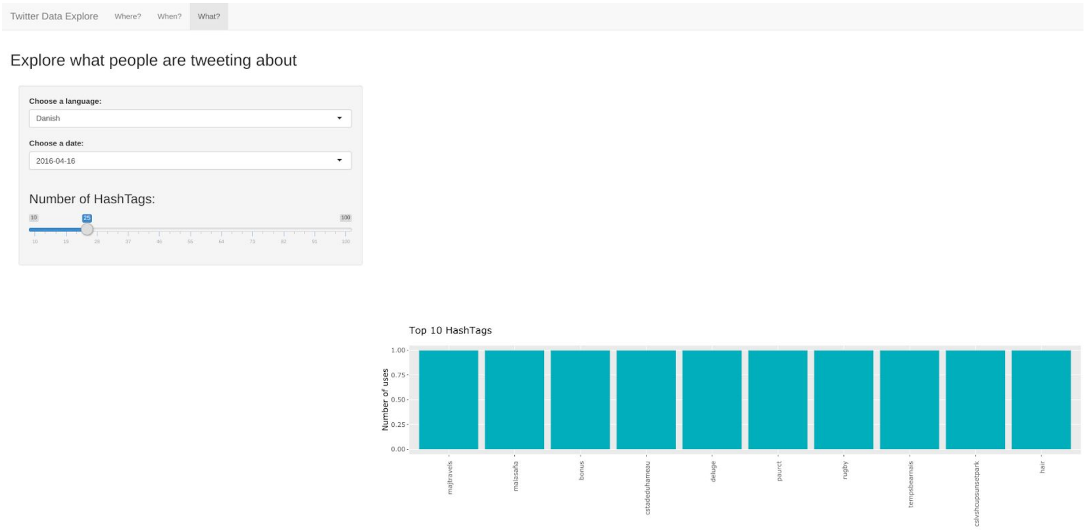
Choose a language: Spanish

Choose a date: 2016-04-14

Number of HashTags: 25

Figure 6 *Screenshot of application showing both plots and associated filters.*

It was found during development that there are some limitations to this implementation.
Due to memory issues the cloud could not be generated when given all the data at once. Thus the filtering options are used to only select data for one language and date each time the cloud is generated. Please note that although this has stopped memory errors occurring on local machines, the memory restriction applied on shinyapps.io means the error can still occur for specific language and date combinations, e.g. “Spanish” and “2016-04-16”. It was also seen that there was not always hashtags available to be displayed, e.g. “Russian” and “2016-04-14”. When this occurs the application returns a message to the user that there is no data to display so they do not think an error has occurred. Finally there is a limitation in the wordcloud2 library where sometimes it fails to display words that are to large in the word cloud, especially if they are within the higher frequency ones. This issue can even lead to issues where no words are displayed (Figure 7). This is where the advantage of the barchart comes in as it has no limitation caused by the word size, and clearly displayed the most frequent words.



Explore what people are tweeting about

Choose a language: Danish

Choose a date: 2016-04-16

Number of HashTags: 25

***Figure** 7 Showing issue where word cloud is not displayed.*

6. Insights

6.1. Question 1: Where do people Tweet from?

The most noticeable insight the heatmap shows us is the more populated an area the more Tweets are sent from there. Madrid is the most brightly coloured area of the map followed by Barcelona. This is actually shows that the populations in Spanish cities are higher than those in Portugal, as even smaller ones in Spain match the density by the larger cities in Portugal (presuming number of Tweets shows the number of people living in a location). Although adjusting the date we see no real difference in the distribution of Tweets sent showing that different regions do not differ on a day to day basis.

When looking at the distributions of what language people Tweet in we see no surprise that Spanish and Portuguese people are more likely to Tweet from their home countries. While Tweets in English, Dainish, Italian generally follow the same pattern as all the languages put together. But the other languages follow different patterns. Those in German and French are more densely found in the north of Spain than any other language. Conversely Tweets in Hungarian, Turkish and Dutch are focused on the south. Tweets in Russian have a higher density in the south than any other non-Spanish languages. Finnish Tweets are the only ones with a low density in central Spain, i.e. Madrid, being found instead in the north and south. While Swedish Tweets are almost exclusively found in Barcelona. Although these distributions may not be a true reflection of where people who speak certain languages congregate when on the Iberian Peninsula, due to the time frame not being at the height of the summer holidays, it does produce some interesting results.

6.2. Question 2: When do people Tweet?

As it can be seen in Figure 2, the trend of the Tweets across time is regular, except on the last day of the record. In the rest of the days, we can observe some cyclical components that repeat periodically every twenty-four hours. Figure 3 shows a time distribution across two days, both showing the same trend although the exact tweeting frequency varies.

At the start of the day there is a progressive rise in the frequency of Tweets being posted, which culminates with a peak around 11:00 am. Then, the frequency values drop down until around 16:00 to rise again and peak at around 17:30. Finally, the frequency goes progressively down, reaching a valley around 22:00. This trend could reflect the daily schedule of the users. As they wake up, more users are active, reaching a peak at 11:00am. Then, at 16:00 there is another peak which may coincide with people getting out of work. The low frequencies may represent the times of the day users are working and sleeping.

6.3. Question 3: What do people Tweet about?

The main topic contained across all days and languages was locations. Not surprisingly, either the words "Spain" and "Portugal" are included in the top ten tags for most combinations of languages and dates. Though cities were more popular with "Madrid" being a very popular term followed by "Barcelona" then smaller Spanish cities it can be presumed this was the order due to population sizes and popularity with tourists. Cities in Portugal also featured but were more popular with those tweeting in Portuguese. When filtering by language it can be seen that the pattern of most Tweeted about cities followed the patterns described when investigating where people Tweeted about. Out of all the "foreign" languages in the dataset (not Spanish or Portogease) people tweeting in German were the only ones to mention cities from from their home country. The next most common theme after cities is positive words such as "love" and "happy". These can be found across the languages. Presumably these are people describing pictures of their lives or tourists showing their feelings towards the cities they are visiting. In fact there were no negative words found in the top words for any language or date.

The most common words Tweeted from the "English" language users across all the dates are "Endomondo" and "Endorphins". This seems strange at first but upon inspecting these Tweets they are automatically generated ones from people using the Endomondo personal trainer app. As these words are the highest on weekdays it shows people are more likely to exercise during the week rather than on weekends.

It can also that some musical events happened during the time frame the data was captured. When filtering the language "English" on the dates "2016-04-17" and "2016-04-18" the term "florenceandthemachine" appears in the top 10 results. This was due to them having a concert in Madrid at this time. The word "prince" has a large number of occurrences when filtering for "English" on the date "2016-04-22". Some research shows this was the day after his death and explains the sudden use of this hashtag.

There are some apparently strangely popular hashtags. When filtering by the language "German" the word "quiz" is the most used hashtag. Attempting to figure out why this was the case was not successful. With "retención" being the second highest word on the 2016-04-14, just two behind the highest for that day, and the highest on the 2016-04-21, by over a 100 mentions, traffic jams in Spain are the worst on Thursday - or at least that is when people complain about them. On 2016-04-22 there was a huge spike in the use of the word "tuitutil" a reference to a no longer available application to find out who has unfollowed you on twitter.

7. Conclusions

Using the application developed above it is possible for a user to successfully investigate the given Twitter data set to answer the questions proposed at the beginning of this report:

Where do people Tweet from?
In the given dataset mainly from populated urban areas, although they can be found throughout the region investigated. A deciding factor on what area they are most likely to Tweet from depends on the language they choose to Tweet in.

When do people Tweet?
People Tweet as they go about their daily lives. Focusing more on when they have "free" time to discuss events and activities they do outside of their work. The lowest posting frequency appears during the night hours, when most users sleep.

What do people Tweet about?
People mainly Tweet about the location they are in, to highlight where a topic they are talking about is taking place. The topics they post about they are mainly positive about, presumably to share happy moments with those that follow them.

The objective of the work has been achieved. A visual tool has been developed that answers the three addressed questions. Furthermore, the tool offers the user the opportunity to interact with the visualizations, adjusting the amount of data that is displayed or filtering the data based on some condition. This helps in managing the amount of information the user needs to focus on and allows them to drive their attention on a more specific portion of the data.

If this work was to be taken forward the main goal would be to implement an application that would be able to integrate the three application streams into one. That would be an application that could track the usage of a hashtag over geographical space and usage over time. This would give real insight into how a topic grows over time to be utilized by people. To achieve this more research would need to be done into managing the large amount of data required as currently this is the bottleneck to an efficient response time.

8. Running the application

The online application can be found at https://dburrell.shinyapps.io/BD_E2_TD/ but as previously mentioned due to the memory limitations with the shiny apps server issues may occur and it is advised to run the application locally for the best results.

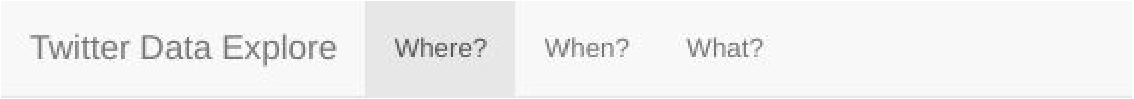
To run the application the following R libraries are required:

- shiny
- shinydahboard
- dplyr

- RColorBrewer
- leaflet.extras
- KernSmooth
- data.table
- rgdal
- tm
- wordcloud
- wordcloud2
- stopwords
- memoise
- raster
- ggplot2
- plotly

To run the application unzip the attached VisualisationProject.zip file and open the app.R file in Rstudio ensuring the clean-tweitter-data.csv and dataGeneration.R files are within the same folder. Then select the “Run App” command. It may take some minutes as the new features are generated and added to the dataset.

To see each section of the application use the navigation bar along the top of the application:



Each tab will display the visualisation for a specification question:

- Where? : The heatmap that allows users to investigate where Tweets were sent from
- When? : Time series plot that allows users to investigate when Tweets were sent
- What?: Wordcloud and barchart that allow users to investigate what has been Tweet topics

9. References

1. 200,000 Spain and Portugal geolocated Tweets. Free Twitter Dataset - (Iberia) - Followthehashtag // Free twitter search analytics and business intelligence tool. Retrieved 26 January 2020, from http://www.followthehashtag.com/datasets/spain_portugal/

2. Rudis, B., Ross, N., & Garnier, S. The viridis color palettes. Retrieved 27 January 2020, from https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html