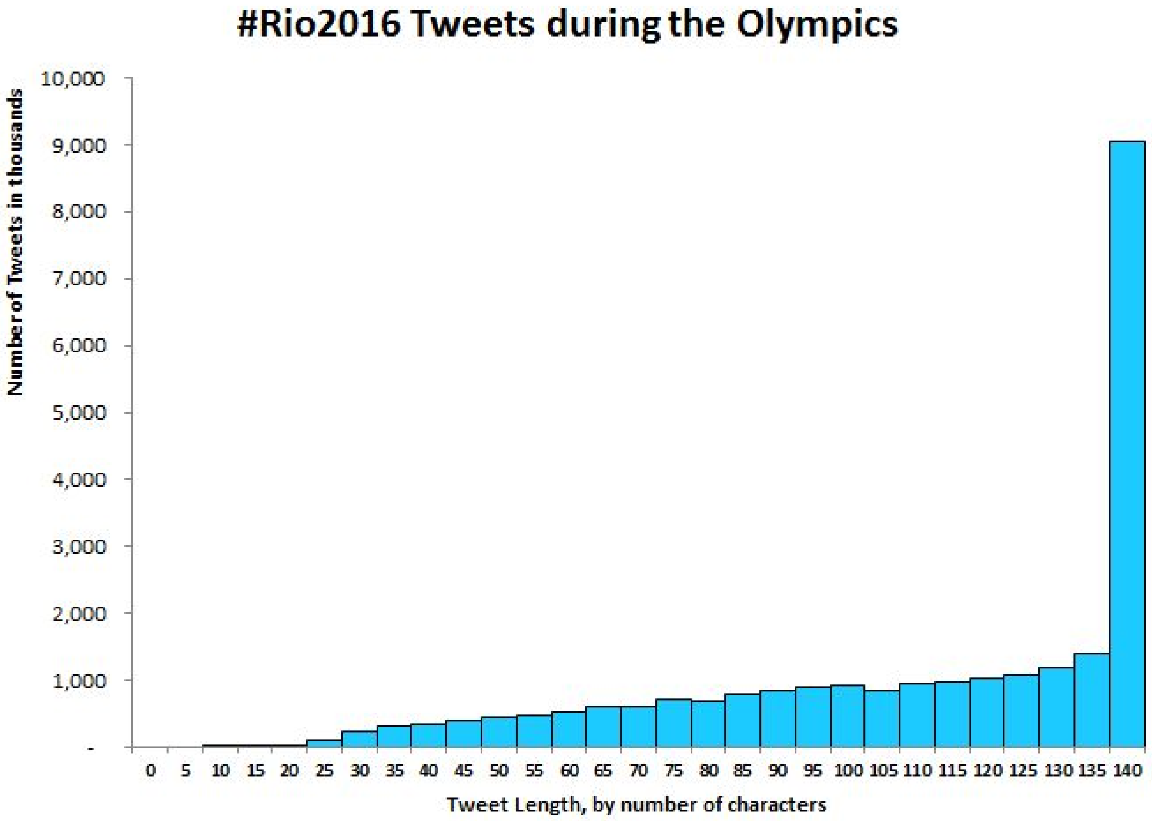
Coursework 1 - Twitter analysis with MapReduce

A. Content Analysis



The histogram above shows the number of tweets in thousands by length, put forth during the 2016 Rio Olympics, using the hashtags #Rio2016 or #rioolympics. The maximum length of a tweet is 140 characters by default.

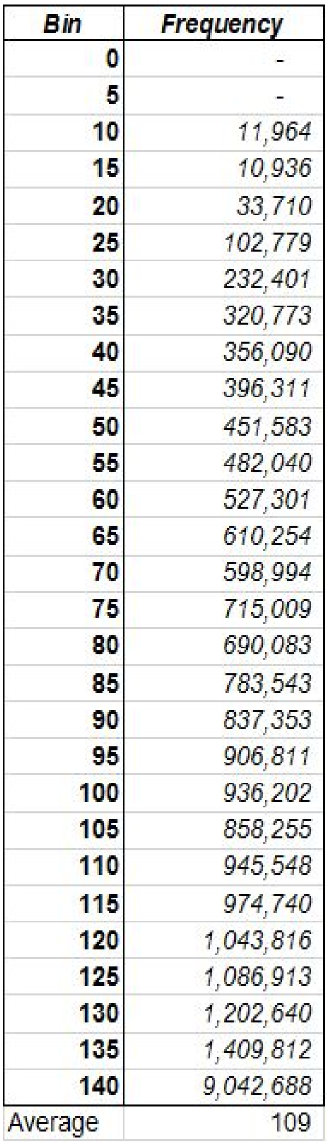
Mapper

The goal of the mapper is to count how many characters were used in a tweet. The mapper input is a Text value with information about a single entry, in the following format:

epoch_time; tweetId; tweet(hashtag contained); device

Example :

1469453965000;757570957502394369; Go Iceland!! #AframIsland #Rio2016; <ahref="http://twitter.com/download/iphone"rel="nofollow">Twitter for iPhone</a>



Bin	Frequency
0	-
5	-
10	11,964
15	10,936
20	33,710
25	102,779
30	232,401
35	320,773
40	356,090
45	396,311
50	451,583
55	482,040
60	527,301
65	610,254
70	598,994
75	715,009
80	690,083
85	783,543
90	837,353
95	906,811
100	936,202
105	858,255
110	945,548
115	974,740
120	1,043,816
125	1,086,913
130	1,202,640
135	1,409,812
140	9,042,688
Average	109

The mapper splits the input by semicolon (;) to separate the fields. The input should only have four fields (time; ID ; tweet; device) so any input that doesn't have exactly exactly four fields is discarded. The mapper extracts the tweet field of the input, calculates the length, categorizes the length into bins of 5 and sends the result to the reducer.

A separate mapper was used to calculate the average length. It executes the same steps as the previous mapper, but does not categorize lengths into bins of 5.

ASCII is a character encoding based on the English alphabet and is commonly used to represent text in computers. The tweet field often contains non-English characters; such as accented characters like "á, ö, ú" or characters from languages like Arabic, Chinese or Japanese to name a few. In some cases these characters are represented in ASCII as being more than one character. This can cause the mapper to overestimate the number of characters in a tweet. In order to prevent that, the mapper replaces all characters not included in the basic ASCII encoding, with a character that is included.[1] Before this implementation the mapper reported tweet lengths of over 140 characters, exceeding the maximum length set by Twitter. After the implementation, there were no reported lengths of over 140 characters.

Reducer

The goal of the reducer is to aggregate data sent from the mapper. The input is in the form of a key and a list of values. Tweet length is the key, followed by a list of values that represent the number of times tweets of the that length were tweeted:

Description: *[Bin Size, count]*

Example: *[140, 1 1 1 1 1 ….]*

The reducer outputs the sums up the list of values, along with the key it's associated to. The results are shown by the histogram and the table above.

A separate reducer was made to calculate the average. It sums up the length of all tweets

[1] https://en.wikipedia.org/wiki/ASCII

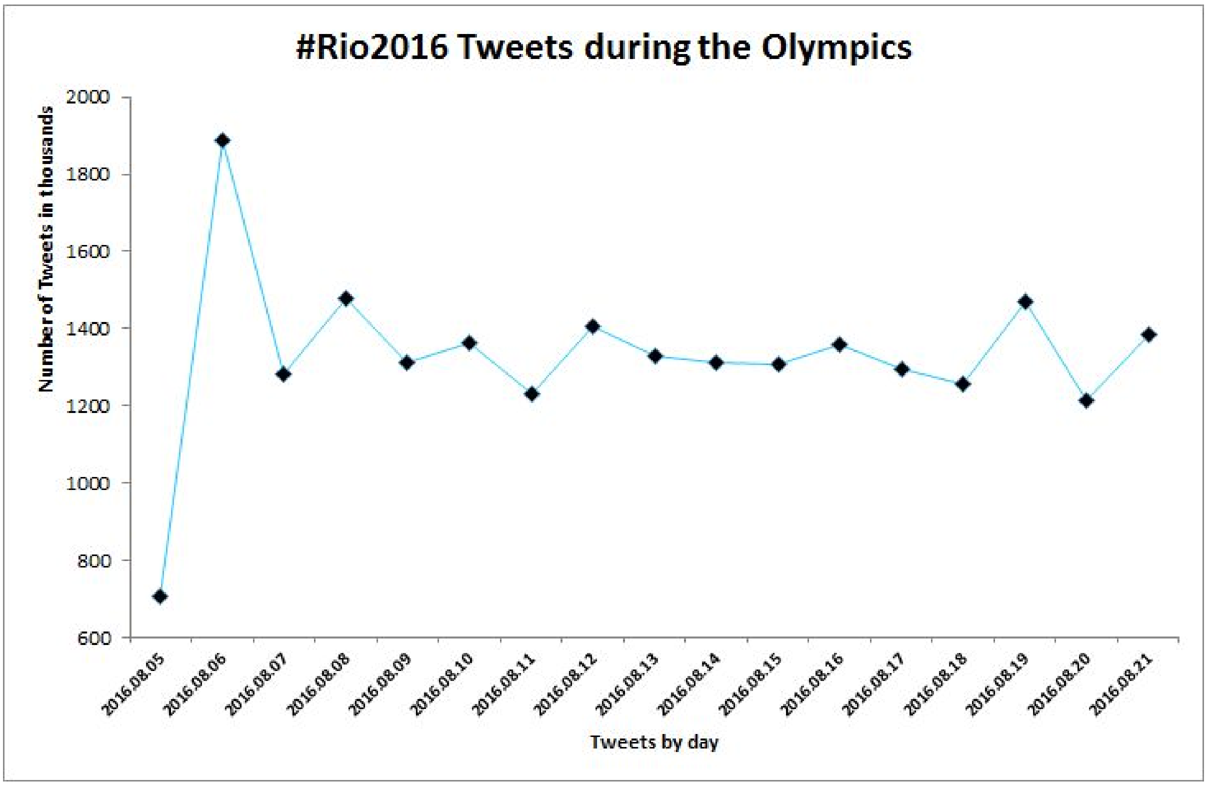
and divides the sum by the number of tweets.

Considerations and results

The majority of tweets are of length 140, accounting for roughly 40% of all tweets. When tweets are written they are often at first too long and then edited down to fit the maximum length. In other cases they are simply split into multiple tweets.[2] Links, images and re-tweets are also handled differently and do not count the same way towards the 140 character limit.[3] The average length of a tweet was 109.

There are no tweets tweets in bins 0 and 5. This is because the hashtags that were used to gather the data, #Rio2016 and #rioolympics, are included in the tweet lengths, making the minimum length of an individual tweet 8.

B. Time Analysis



The time series shows the number of tweets in thousands, per day of the 2016 Rio Olympics. The opening day was 5th of August and the games ended on 21st of August.

[2] http://thenextweb.com/twitter/2012/01/07/interesting-fact-most-tweets-posted-are-approximately-30-characters-long/

[3] https://blog.twitter.com/2016/doing-more-with-140-characters

Mapper

The mapper finds the day a tweet was posted. Like in the previous mapper the input is a single line of text, which the mapper splits by semicolon (;) to separate the fields of the tweet. Inputs that don't have exactly exactly four fields are discarded. The mapper then extracts the timestamp from the the input. The timestamp is stored by default as an epoch time, or unix time, which the mapper converts to a date in the form of yyyy.MM.DD and sends it to the reducer.

Reducer

The reducer aggregates the results of the mapper. The input is a key and a list of values. The date of the tweet is the key and a list of values is the number of times a tweet of was posted on that day:

Description:	*[Date, count]*
Example:	*[2016.08.20, 1 1 1 1 1 ….]*

The reducer outputs the sum of the the list of values, along with the key it's associated to. The results are shown by the time series above.

Considerations and results

The reducer reported dates ranging from 25th of July to the 9th of September. Only tweets posted during the days of the event (5 - 21 August) are reported in the timeseries.

The date in the dataset is stored as unix time, which is based on UTC time.[4] The opening ceremony of the games started at 23:00 UTC, 5th of August and ended at 03:00 UTC, 6th of August. Many tweets were posted during the opening ceremony which is why we see a spike on the 6th of August.[5] We can also see a smaller spike on the 19th of August as that day many group sports such as hockey and badminton concluded, as well as Usain Bolt, a world famous runner from Jamaica, earned his 3rd gold during course of the games.[6]

[4] https://en.wikipedia.org/wiki/Unix_time

[5] https://en.wikipedia.org/wiki/2016_Summer_Olympics_opening_ceremony

[6] https://www.rio2016.com/en/news/ten-highlights-of-friday-19-august-at-the-rio-2016-olympic-games

C. Hashtag Analysis

This part of the report covers which countries got the most support during the course of the games.

Mapper

The goal of the mapper is to find out which country or countries (if any) a particular tweet is supporting. As in the previous mappers the input is a single line of text, which the mapper splits by semicolon (;) to separate the tweet into fields. Inputs that don't have exactly exactly four fields are discarded. The mapper takes the tweet field and splits the text into individual words. It then removes non-ASCII characters, with a few exceptions.[7] This is done so that similar hashtags are grouped together.

Example: #USA , #USA!, #USA? ,#U.S.A are all converted into #USA

The only drawback is that this step removes words from non-latin based languages such as in Hindi, Mandarin or Japanese. As adding language support for all known languages would have been too complex for a project this size, the drawback was deemed acceptable and latin based alphabets were made the focus of this analysis. For this reason there is a heavy bias towards English speaking countries[8] although an attempt was made to reduce this bias as much as possible.[9]

The mapper decides that a word is a hashtag if the first character of the word is a '#'. In some cases a word contains more than one hashtag, such as #USA#SPAIN. If that happens the mapper separates them into two hashtags. The mapper then compares the hashtag to a list of country names and variants of their names, such as {"An ISO Alpha-3 country code"[10], "Name of country in local language"}.

Example : {"Spain", "ESP", "España"}

Some exceptions were made to this comparison however. For some countries the local name was not always available. In other cases the 3 digit country codes were mistakenly matched with hashtags that were not related to those countries. In those cases the 3 digit

[7] The following characters were left intact, as they form part of the local variant of some non-English speaking country names: çÉñÍÖé'

[8] http://qmplus.qmul.ac.uk/mod/assign/view.php?id=354303

[9] See footnote 7

[10] http://www.nationsonline.org/oneworld/country_code_list.htm

code was removed from that country's list.[11]

If a hashtag contained one of those variants of their name, the hashtag was associated with that country. A single hashtag could be associated with more than one country.

Example #USAvsTeamGB Associated with both USA and UK

Reducer

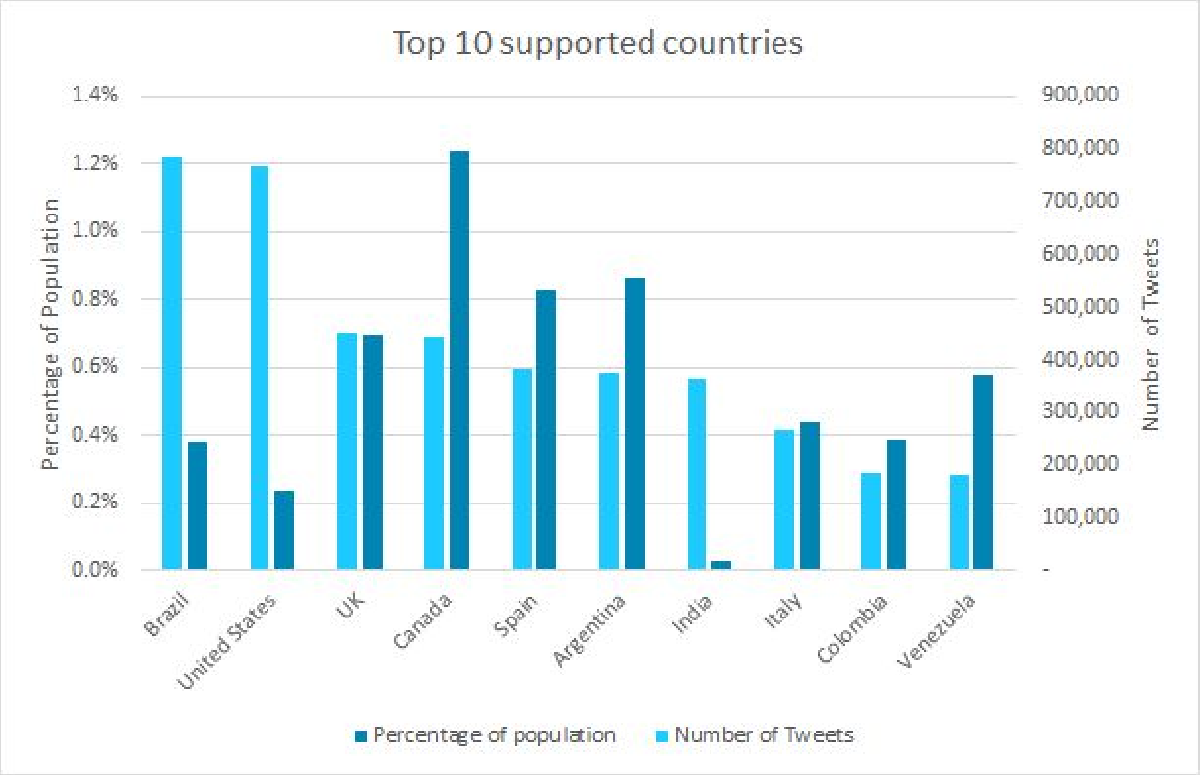
The reducer aggregates the results of the mapper. The input is a key and a list of values. The country name is the key and a list of values is the number of tweets in support of that country:

Description: [Country Name, count]

Example: [UK, 1 1 1 1 1 ….]

The reducer outputs the sums up the list of values, along with the key it's associated to.

Results



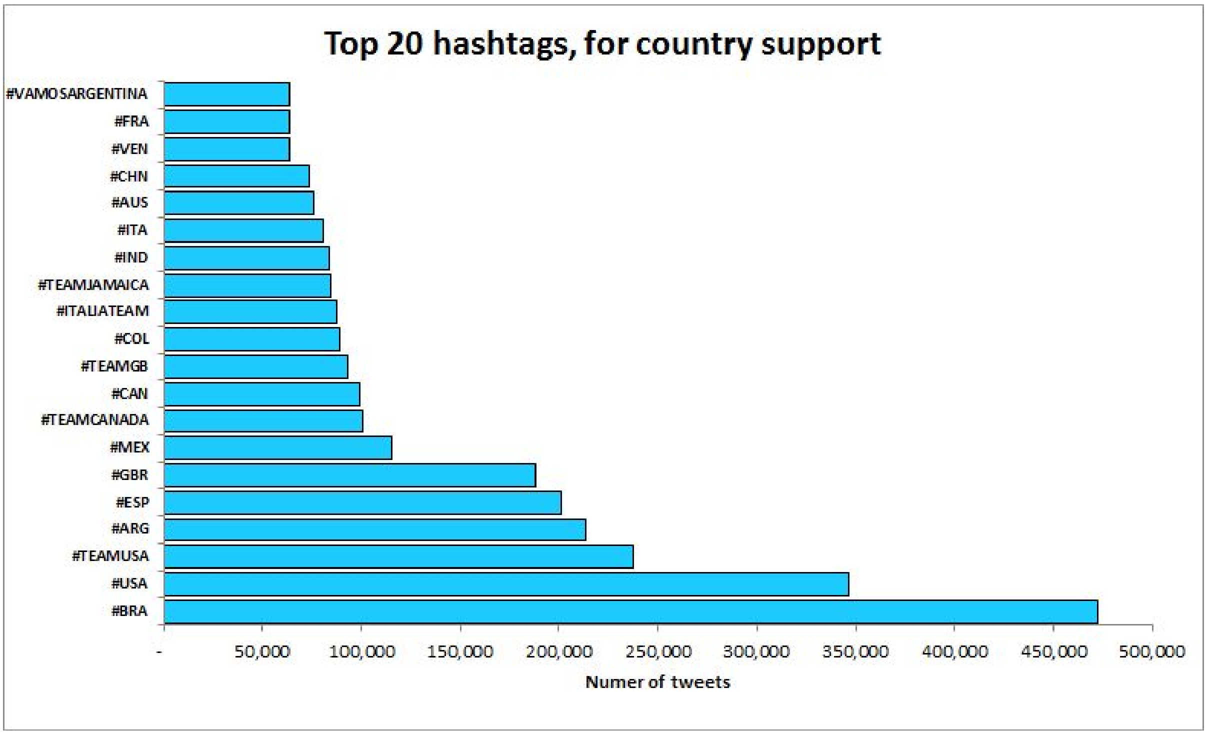
The two most supported countries are Brazil and USA respectively, which is consistent with results reported by others.[12] [13] This was to be expected as the games were held in Brazil as

[11] See "Considerations and Results" for more in-depth

[12] https://blog.twitter.com/2016/the-rio2016-twitter-data-recap

[13] https://icgx.asia/blog/the-olympics-most-tweeted/

well as Twitter being a very popular social media platform in the USA.[14] Canada received the most support per capita.



As the most supported countries are Brazil and USA, it's logical that the top hashtags are related to them, with #BRA, #USA and #TEAMUSA being the most used hashtags related with country support. The most common forms of country support hashtags, are hashtags consisting of only the 3 digit country codes or combination the word "TEAM" and the three digit country codes. Exceptions being #VAMOSARGENTINA and #TEAMGB, which were hashtags marketed in their respective countries to support their teams in the Olympics.

Considerations

At first all hashtags were compared to an unfiltered list of countries and their 3 digit country code. This lead to some hashtags being mistakenly matched with countries that had common forms of country codes, such as Andorra (AND), Antarctica (ANT) or The Vatican (VAT). To filter out these errors the tweets per country were compared to the population of each country. Top 50 most supported countries that had an abnormal amount of support (higher than 20 tweets per capita) had their 3 digit code removed.

Second time around the tweets and the country they supported were arranged from most frequent to least frequent. This lead to discoveries such as #WaterPolo was being counted as support for Poland (POL) and #CeremoniaDeApertura (#OpeningCeremony in English)

[14] http://bit.ly/2eqU3fM

was being matched with Turkey (TUR). In those cases the countries also had their 3 digit code removed from the list. This increases bias towards countries that have had their country codes removed. See below table for more examples:

Code	*Country*	*Reason for removal // Number of tweets per capita*
AND	*Andorra*	*Matched with any country that ends with -land*
POL	*Poland*	*#Matched with #WaterPolo*
TUR	*Turkey*	*Matched with #CerimoniaDeAbertura, a common hashtag in Spanish*
ARM	*Armenia*	*Code is common in English*
ANT	*Antarctica*	*Code is common in English // Over 10.000 tweets per capita*
BEN	*Benin*	*Common name in english // Over 50 tweets per capita*
AIA	*Anguilla*	*Over 20 tweets per capita*
ABW	*Aruba*	*Over 20 tweets per pecapita*
ANT	*Antarctica*	*Over 20 tweets per capita*
EST	*Estonia*	*Code is common in English*
ERI	*Eritrea*	*Code is common in English*
FIN	*Finland*	*Code very common // Over 20 tweets per capita*

An alternative approach was considered; only to match the country if the hashtag contained nothing more than the country code, such as #ESP or #USA. This was thought to be a inferior implementation as it would overlook hashtags that correctly used the country code in combination with sport’s names or other words used to cheer on teams.

Examples #GoUSA #FutolBRA #VamosESP #AframÍsland