STREAM PROCESSING

"FIFA WORLD CUP TWEETS' ANALYSIS"

COURSE: ITC6001- INTRODUCTION TO BIG DATA

SEMESTER: FALL 2020

INSTRUCTOR: Sofoklis Efremidis

PARTICIPANT: Joana Veizi

ID: 240571

Student: Joana Veizi
Id: 240571

Table of Contents

Student: Joana Veizi
Id: 240571

ABSTRACT

With FIFA World Cup tournament taking place every fourth year, more and more tweet generation is triggered. Massive posts and retweets exchanged among users, challenge data analysts to collect, clean and analyze the corresponding data streams. This assignment aims to use 3 different analyzing methods – traditional approach, Count-Min Sketch and HyperLogLog algorithm – in order to determine the Heavy Hitters showed in the tweets. The performance of the two aforementioned sophisticated algorithms will not only be evaluated but also compared and represented.

1. INTRODUCTION

FIFA World Cup known as well as World Cup, is one of the most popular global football competitions. In 2018**,** 3.572 billion people watched FIFA World Cup Russia (fifa.com, 2018), which is an indication of viewers' attraction. These remarkable records provoke more and more analysts, to investigate data associated with the tournament and acquire essential insights in various aspects. The corresponding calibrated statistics may help organizations and companies learn more about people's preferences and interests, so they could potentially adjust properly and increase profits.

But where could be data found and collected? The data sources could be various. However we will focus on streaming data from Twitter. To achieve this goal we should define what data streaming is and how could be such kind of data manipulated.

Data Streaming is a continuous flow of data generated by a variety of sources (Rouse, 2019). Some indicative examples are, data generated from social networks, ecommerce purchases, bank transactions and so on. What makes them popular, is the fact that they are produced and analyzed real-time, keeping companies updated and enabling them react immediately. This kind of data flows in different volumes and different speeds.

The continuous characteristic and the huge volume of the data streams require customized techniques for their filtering and analysis. In our assignment we will briefly use a three-dimensional methodology - traditional approach, Count-Min Sketch, HyperLogLog algorithm - in order to analyze the Word Cup tweets and detect the most frequent mentioned tags and users. In other words we will search for the Heavy Hitters. The analysis was enriched with further statistics and plots so that data could be deeply understood.

Student: Joana Veizi
Id: 240571

2. SIMPLE APPROACH

The dataset used for our assignment was relied on 46 json files, containg tweets. It is assumed that the data flow as a stream and the volume does not allow their storage.

The first step we took was to open the json files and parse the data they contain, line by line. This step is essential for the analyst, in order to obtain access to the data.

Having reached the accessibility, a python program that is able to distinguish tweets and collect all mentioned tags and user-names among the tweets, was built. As a result the total number of tweets found within the json files was 23881.

More in detail, within the tweets, 28606 hashtags and 48434 mentioned users were detected. The research for tags was relied on words that begin with the symbol "#" whereas the mentioned users were pointed out by their "@" symbol (How to use advanced search). It should be underlined that the aforementioned numbers do not refer to the total number of unique tags or user names. The following Statistics' table clearly shows the difference between the two measurements.

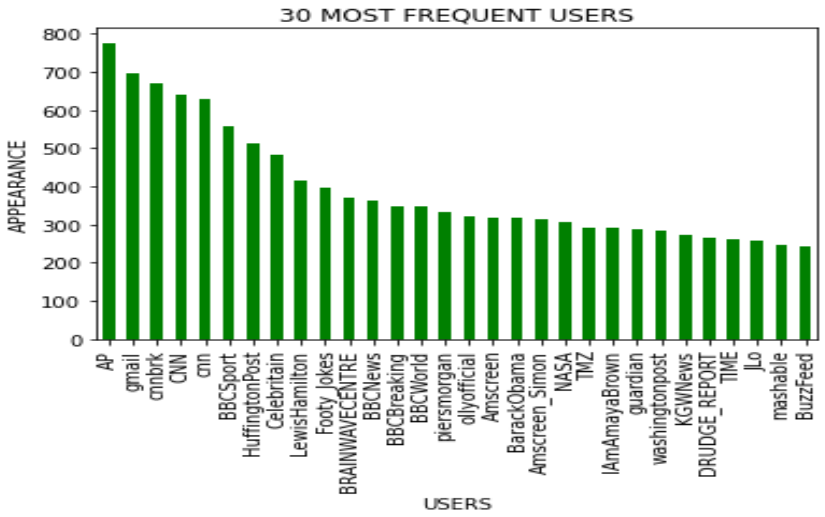
Statistics' Table

STATISTICS	
USERS	**TAGS**
Total number of mentioned users: 48434	**Total number of found tags: 28606**
Total number of unique users: 7279	**Total number of unique tags: 6299**

The forenamed Statistics provide an overview of the data but there is need for further investigation. Questions such as, how many times each Tag was showed and who are the most frequently mentioned users should be answered as well.

The next graph shows in detail the 30 most frequent stated users. We can infer from the bar graph that the American television channel, CNN and the British public service broadcaster, BBC are great examples of Heavy Hitters.

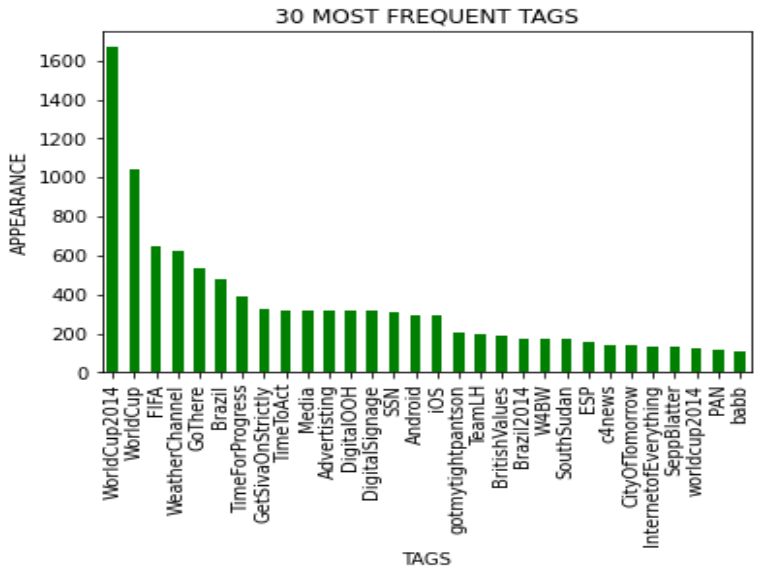
Most frequent appeared users.



Both users are quite expected, since they are news' resources.

Similar non-surprising results are found through the tags research. The protagonists of the tags are tags that relate to the Wold Cup. For instance, WorldCup2014, WorldCup and FIFA tags can be individuated. Since the WorldCup2014 hashtag is the leader of tags, we could assume that the tweets we are analyzing were generated in 2014.

Most frequent appeared tags



Memory Usage

The users' and tags' frequencies were briefly recorded by two data frames and the size consumed by the two data frames is projected by the "Data frame size in bytes" table. Data frame constructed for users' consumes more memory than the tags' data frame but this can be explained due to the greater total and unique number of users found in the tweets.

Data frame size in bytes

CONSUMED SIZE BY DATA FRAME

Users'	Tags'
549324	467206

The results of the unique users and tags associated with their frequencies where not only calculated but also saved into csv files named as "user-counter" and "tags-counter" correspondingly. In this way more details can be found in the relative files.

Student: Joana Veizi
Id: 240571

3. COUNT-MIN SKETCH APPROACH

The analysis of massive data sets using traditional tools accompanies two important drawbacks. Time consume and Terabytes of memory usage. Such problems can be overcome with approaches such as Count-Min Sketch and HyperLogLog. Count-Min Sketch is a probabilistic data structure that builds a table containing the frequency of events taking place in data streams (Count-Min Sketch, 2019).

Functionality

Initially all the cells of the table have a 0 value. When an event is faced it passes through a hash function. The output of the hash function indicates a row value, and the counter at the corresponding row column is incremented by 1. The procedure is repeated until the event parses all defined hash functions. To determine the frequency of the event, we select the minimum count over all of the hash functions (Shukla, 2018).

CMS

	VALUE 1	VALUE 2	VALUE 3
HASH FUNCTION 1	0	0	0
HASH FUNCTION 2	0	0	0

Does this method show any disadvantages?

The answer to the stated question is, positive. The most common drawback of the estimator is that is considered a biased estimator. In other words, the output may be overestimated. However this issue may be minimized by increasing the width of the table. The higher the width of the table, the less the probability of collisions is. This happens simply because we increase the counters we store in memory.

Count-Min Sketch implementation for the FiFa World Cup tweets analysis

The Count-Min Sketch data structure is a two dimensional matrix with w columns and d rows. The parameters w, d are determined when the sketch is created and they are chosen based on two quotations, $w = [e/\varepsilon]$ and $d = [\ln 1/\delta]$ where ε is the error factor and δ the error probability.

We implemented the CMS algorithm in order to check the number of occurrences of each user and tag. The algorithm was evaluated twice when two different error rates where applied. For the evaluation part, Accuracy was computed. Accuracy is the fraction of the correct predictions the algorithm made (Classification: Accuracy). As a result, we were reassured that with lower error rate the better accuracy scores the algorithm produces. More in detail, as the bellow table describes with a 0.0001 error rate, our algorithm achieves a 99% Accuracy.

Student: Joana Veizi
Id: 240571

Count-Min Sketch Accuracy

Accuracy (confidence= 0.99, error= 0.001)		Accuracy (confidence= 0.99, error= 0.0001)	
USERS	**TAGS**	**USERS**	**TAGS**
17%	27%	99%	99%

Memory usage

The choice of the parameters may be influenced by two factors. The time and the memory. Initially most users would choose a lower error rate to accomplish better results, but the choice is not brief since the memory usage should be taken into consideration.

Having defined a 0.99 confidence and a 0.0001 error as Count-Min Sketch parameters, the CMS memory usage, in bytes was calculated. The next table provides basic size information for both users' and tags. Clearly, the CMS constructed for users consumes more size than the CMS constructed for tags. The output was anticipated since the total and unique number of users was higher than the tags.

CMS size in bytes

CONSUMED SIZE BY CMS

Users' CMS	Tags' CMS
1132800	1127776

Student: Joana Veizi
Id: 240571

4. HYPERLOGLOG APPROACH

The HyperLogLog is another highly used algorithm for streaming data analysis. This approach computes approximately the unique events (Cardinality) in a data set. The output is a number that refers to high or low Cardinality. The higher Cardinality the more distinct elements in a data set.

Functionality

An event is hashed and for each of those hashes we count the number of leading zeros. The estimation is based on the longest sequence of zeroes found while parsing the data set. An estimation for the unique values in a dataset is calculated by the formula 2^x , where x is the total number of zeroes found in the longest sequence of zeroes.

The next table includes a representative example. We assume that an event is hashed. After the hashing, 3 different hash values arise. The longest sequence of zeros that appears in the hash values is 3. Thus the estimator's result is 2^3=8

Example
00**110110 =>2**
0**1010111 =>1**
000**10110 =>3**
Max= 3
Estimator: 2^3=8

Does the method accompany any disadvantages?

The main disadvantage of the algorithm is the high variance. The problem can be reduced by splitting the dataset into chunks. When the dataset is divided into subsets, the maximum number of leading zeros in each chunk is computed. The mean value of the results is calculated providing an estimation of the cardinality of the entire set (HyperLogLog, 2020).

HyperLogLog implementation for the FiFa World Cup tweets analysis

In the last part of our analysis we implemented the HyperLogLog algorithm with a given 5% counting error, in order to detect the tags' and users' distinctiveness.

The following table illustrates the Cardinality results' showed after implementing at first a simple estimator and then the HyperLogLog algorithm. It is noticeable that the two approaches exhibit different results when they have to compute the distinct values. On the one hand the estimated users' cardinality is underestimated by the HyperLogLog algorithm, whereas on the other hand the tags' cardinality is overestimated. Overall, the results may differ based on the used approach; however it is important to be mentioned, that the deviations are not too large to reject the algorithm.

Student: Joana Veizi
Id: 240571

Traditional approach VS HLL algorithm (5% error)

USERS' RESULTS		
	Traditional Approach	HyperLogLog Algorithm
Total Number	48434	48434
Total Unique Number	7279	6825

TAGS' RESULTS		
	Traditional Approach	HyperLogLog Algorithm
Total Number	28606	28606
Total Unique Number	6299	6789

Recommendations for improvement

It is evident, that our algorithm needs further improvements. Returning some steps back, we face the primary assumption. The counting error is defined to 5%. If we decrease the counting error to 2%, then the algorithm will produce less deviation from the true results.

Traditional approach VS HLL algorithm (2% error)

USERS' RESULTS		
	Traditional Approach	HyperLogLog Algorithm
Total Number	48434	48434
Total Unique Number	7279	7255

TAGS' RESULTS		
	Traditional Approach	HyperLogLog Algorithm
Total Number	28606	28606
Total Unique Number	6299	6433

The two tables show clearly the effect of the counting error decrease. As we can observe, the change we made had a harmonic correction both to users' and tags' results.

Therefore, we conclude that there is enough evidence that we can gain better results by simply changing the counting error. The corresponding error of HLL can be easily computed by the formula $1.04/\sqrt{m}$, where m is the number of registers.

Memory Usage

Having defined a 0.02 error as the parameter of HLL algorithm the size consumed by the HLL was computed. The next table allows comparison between users and tags and shows the memory usage of HLL when was applied for the research of users and tags cardinality. What seems interesting at this point is the fact that HLL consumes less memory than CMS.

HLL size in bytes

CONSUMED SIZE BY HLL	
Users' HLL	Tags' HLL
34512	34448

5. CONCLUSION

Overall Data streams need dedicated algorithms to be filtered and analyzed. Our assignment has engaged different approaches for the tweets data streaming manipulation. The three analyzed approaches imply both advantages and disadvantages. Nevertheless, the algorithms may improve performance with the implementation of different solutions. Our recommendation for each individual user is to define a goal and a cost range (time consume, memory storage, complexity) and act correspondingly. The whole process should be by any means customized since different users have different needs.

6. REFERENCES

fifa.com. (2018, December 21). Retrieved from https://www.fifa.com/worldcup/news/more-than-half-the-world-watched-record-breaking-2018-world-cup

Count-Min Sketch. (2019, July). Retrieved from https://florian.github.io/count-min-sketch/

HyperLogLog. (2020, August 10). Retrieved from wikipedia: https://en.wikipedia.org/wiki/HyperLogLog

Classification: Accuracy. (n.d.). Retrieved from https://developers.google.com/machine-learning/crash-course/classification/accuracy

Crahen, E. (n.d.). *Count-Min Sketching, Configuration & Point-Queries*. Retrieved from https://crahen.github.io/algorithm/stream/count-min-sketch-point-query.html

How to use advanced search. (n.d.). Retrieved from https://help.twitter.com/en/using-twitter/twitter-advanced-search

poweredtemplate.com. (n.d.). Retrieved from https://poweredtemplate.com/word-templates/abstract-textures/12211/0/index.html#

Rouse, M. (2019, March). *Definition data streaming*. Retrieved from SearchNetworking: https://searchnetworking.techtarget.com/definition/data-streaming

Shukla, K. (2018, July). *Big Data with Sketchy Structures, Part 1 — the Count-Min Sketch*. Retrieved from towards data science: https://towardsdatascience.com/big-data-with-sketchy-structures-part-1-the-count-min-sketch-b73fb3a33e2a

7. APPENDIX

In various points of our analysis it is mentioned that both algorithms Count-Min Sketch and HyperLogLog, with given parameters consume a certain memory. In other words if the parameters of the applied algorithms change the memory usage will change as well.

For experiment purposes, we altered the parameters of the algorithms and recorded the influence the parameter's alteration had in memory usage.

In CMS algorithm we can infer that when it was used a higher error the memory usage was decreased. The bellow table illustrates the memory behavior when a 0.001 error was defined as CMS parameter.

CMS size in bytes (error= 0.001)

CONSUMED SIZE BY CMS

Users' CMS	Tags' CMS
126752	121312

In HLL algorithm it was used a 0.01 error parameter (less than the 0.02 error used in the analysis) and the size used by the HLL increased.

HLL size in bytes (error=0.01)

CONSUMED SIZE BY HLL

Users' HLL	Tags' HLL
141168	141104

To sum up, in both algorithms Count-Min Sketch and HyperLogLog the error increase leads to memory size growth.