University of Science and Technology in Zewail City
Communications and Information Engineering

Big Data Analytics - CIE 427

Mini Project 1
Analyzing Reddit Comments

Khaled Elbastawisy 201-800-029
Mohamed Ahmed 201-801-898

Data Analysis:

Upon inspection of a 10000 randomly-sampled entries of the reddit comment dataset the following findings were made:

- All comments have zero downvotes.
- About 4.5% of the comments have negative upvotes (which doesn't make sense).
- About 8% of the sampled comments have deleted users. A user could get banned or remove his account but his comments would still be on the platform.
- Some of the most active users are actual bots such as havoc_bot and autowiki bots. They are used for auto-moderating contents.
- Some user entries have nan values
- About 6.5% of the comments have deleted body entries.
- gilded is the amount of reddit golds a comment receives. Almost all the comments have 0 gilded which doesn't make it a useful feature for our analysis.

MapReduce Jobs

1- Top N Subreddits

Mapper: The mapper simply loads the data into a JSON file and then extracts only the value assigned to the label “subreddit”.

Reducer: The reducer takes in “N” as an argument which is supposed to denote the number of top subreddits to show in descending order. The reducer uses a dictionary to count instances of different subreddits appearing in the records. After finishing the count the reducer then sorts the dictionary descendingly and outputs the first N subreddits with their appearance count.

2- Top N Users

Mapper: The mapper simply loads the data into a JSON file and then extracts only the value assigned to the label “author”.

Reducer: The reducer takes in “N” as an argument which is supposed to denote the number of top users to show in descending order. The reducer uses a dictionary to count instances of different users appearing in the records. After finishing the count the reducer then sorts the dictionary descendingly and outputs the first N users with their appearance count.

3- Most Popular Topics in each Subreddit

Mapper 1: The mapper takes an argument of a file name. This is to allow for the processing of most popular topics for only certain subreddits. If not given an argument the mapper will analyze all subreddits. In this instance we used the most popular 20 subreddits only for analysis. The mapper then uses NLTK natural language processing library alongside a regex string to filter out unimportant characters and extract important words or phrases that represent topics for each record. The mapper then sends off two values: the subreddit and the topic.

Reducer 1: The reducer takes in the subreddit-topic duo and uses a dictionary to count each instance for each subreddit-topic tuple. The reducer then passes the subreddit-topic tuple and a third value representing the numbers of its occurrence.

Mapper 2: The job of this mapper is to just pass along the output of Reducer1 without making changes.

Reducer 2: The reducer's job is to analyze the ordered records of each subreddit-topic instance and choose the subreddit-topic tuple with the most occurrences. The reducer then outputs the subreddit, its most popular topic, and the number of times it appeared.

4- Most Popular Topics for each User

Mapper 1: The mapper takes an argument of a file name. This is to allow for the processing of most popular topics for only certain users. If not given an argument the mapper will analyze all users. The mapper ignores deleted users and bots for the rest of the analysis. In this instance we used the most popular 20 users only for analysis. The mapper then uses NLTK natural language processing library alongside a regex string to filter out unimportant characters and extract important words or phrases that represent topics for each record. The mapper then sends off two values: the user and the topic.

Reducer 1: The reducer takes in the user-topic duo and uses a dictionary to count each instance for each user-topic tuple. The reducer then passes the user-topic tuple and a third value representing the numbers of its occurrence.

Mapper 2: The job of this mapper is to just pass along the output of Reducer1 without making changes.

Reducer 2: The reducer's job is to analyze the ordered records of each user-topic instance and choose the user-topic tuple with the most occurrences. The reducer then outputs the user, their most popular topic, and the number of times it appeared.

5- Most Popular Users for each Subreddit

Mapper 1: The mapper takes an argument of a file name. This is to allow for the processing of most popular topics for only certain subreddits. If not given an argument the mapper will analyze all subreddits. In this instance we used the most popular 20 subreddits only for analysis. The mapper goes through sampled records of each of the analyzed subreddits and extracts the user in each record. The mapper then sends off two values: the subreddit and the user.

Reducer 1: The reducer takes in the subreddit-user duo and uses a dictionary to count each instance for each subreddit-user tuple. The reducer then passes the subreddit-user tuple and a third value representing the numbers of its occurrence.

Mapper 2: The job of this mapper is to just pass along the output of Reducer1 without making changes.

Reducer 2: The reducer's job is to analyze the ordered records of each subreddit-user instance and choose the subreddit-user tuple with the most occurrences. The reducer then outputs the subreddit, its most popular user, and the number of times they engaged.

6- Most Upvoted Topics

Mapper 1: The mapper takes an argument that should be 0 or 1. The argument is supposed to denote whether to work on “most upvoted” mode or “most downvoted” mode. The “most downvoted” mode was not used in this analysis ue to all downvotes in the sampled million instances are zero. The mapper then uses NLTK natural language processing library alongside a regex string to filter out unimportant characters and extract important words or phrases that represent topics for each record. The mapper then sends off two values: the topic and the upvote count.

Reducer: The reducer takes in “N” as an argument which is supposed to denote the number of top-voted topics to show in descending order. The reducer uses a dictionary to count instances of different topics appearing in the records. After finishing the count the reducer then sorts the dictionary descendingly and outputs the first N topics with their appearance count.

7- Attitude of user based on their comments (sentiment analysis)

This job analyzes the comments of users and outputs a ratio of positivity to negativity of their overall activity on all subreddits. A ratio greater than one means an overall positive attitude and the larger the number the merrier. A ratio between zero and one implies a negative attitude being more negative as it gets closer to zero.

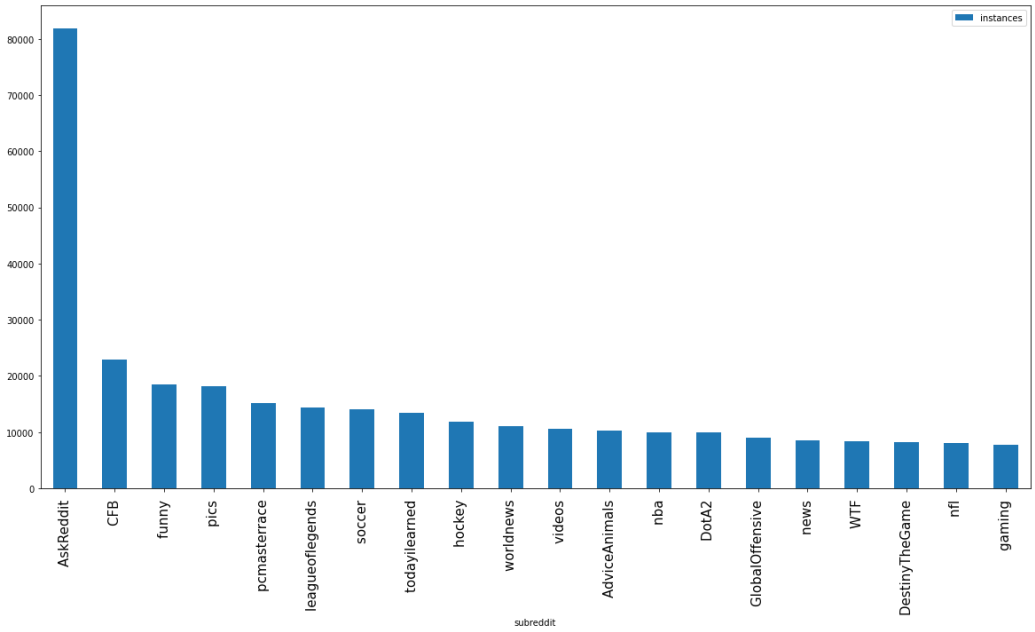
Mapper 1: The mapper takes an argument of a file name. This is to allow for the processing of most popular topics for only certain users. If not given an argument the mapper will analyze all users. In this instance we used the most popular 20 subreddits only for analysis. The mapper goes through sampled records of each of the analyzed users and extracts two scores of negative or positive sentiment for each comment. The mapper sends three values: the user an the positive and negative scores

Reducer: The reducer sums the positive and negative values for each user. After that, a ratio of positive : negative is calculated and assigned to each user. If the negative score is zero the positive score is divided by 0.01. The reducer then outputs the user and the ratio values.

MapReduce Results Analysis and Visualization

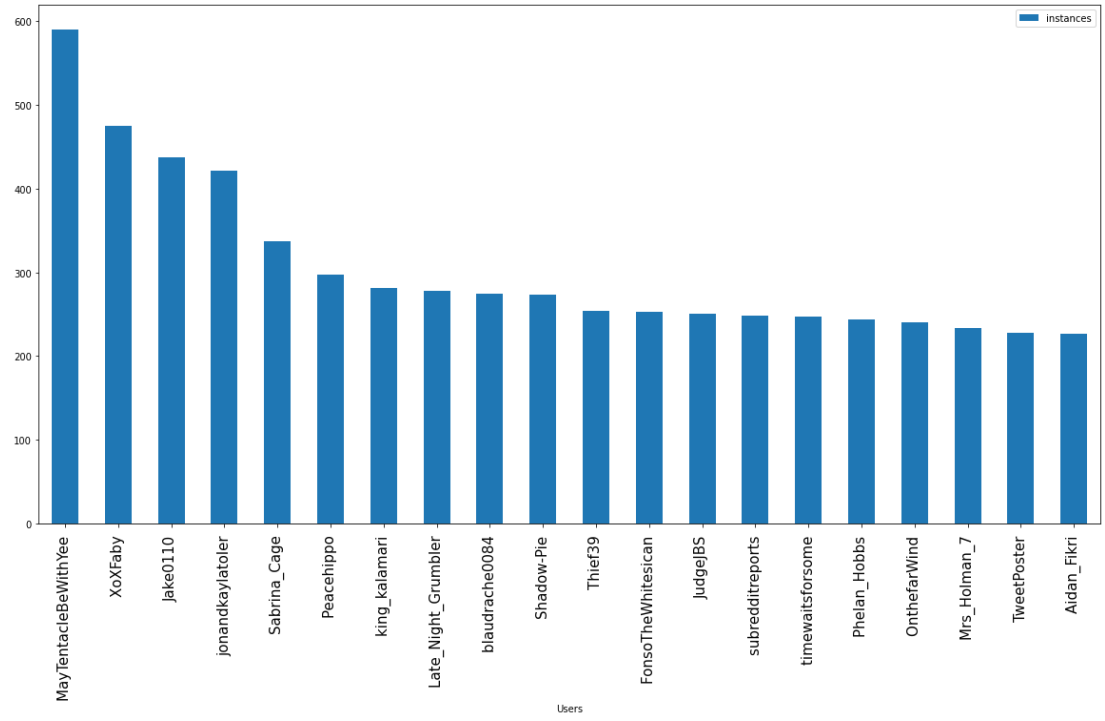
1- Top N Subreddits

The top 20 subreddits are shown in the bar graph below.



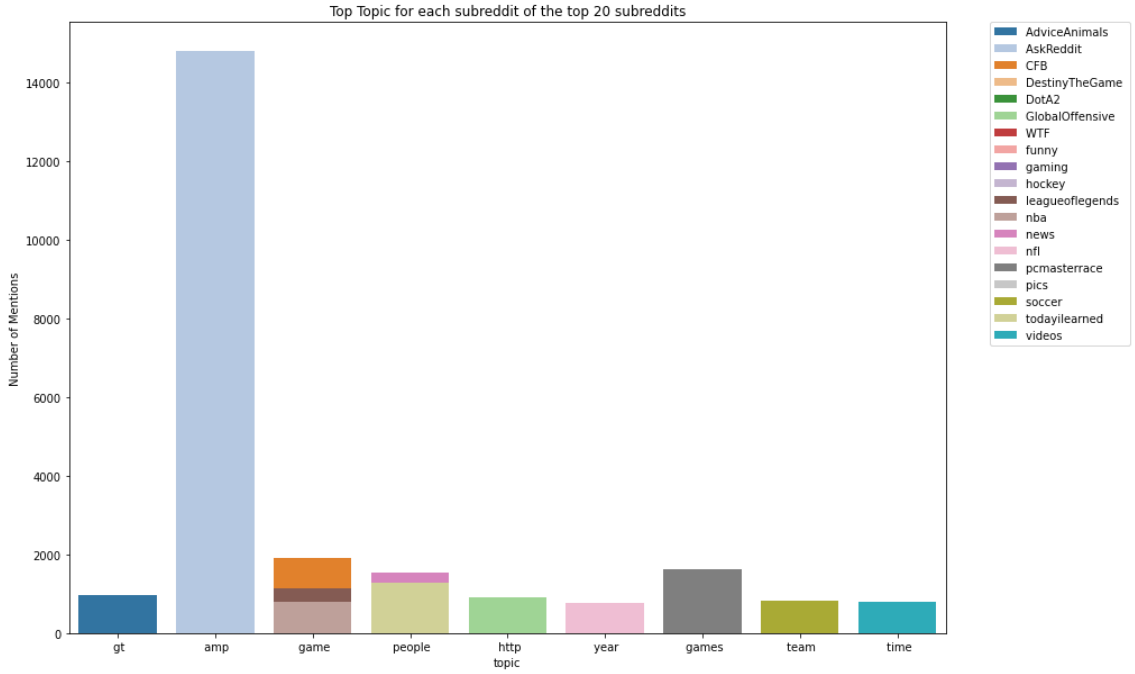
2- Top N Users

The top 20 active users are shown in the bar graph below.



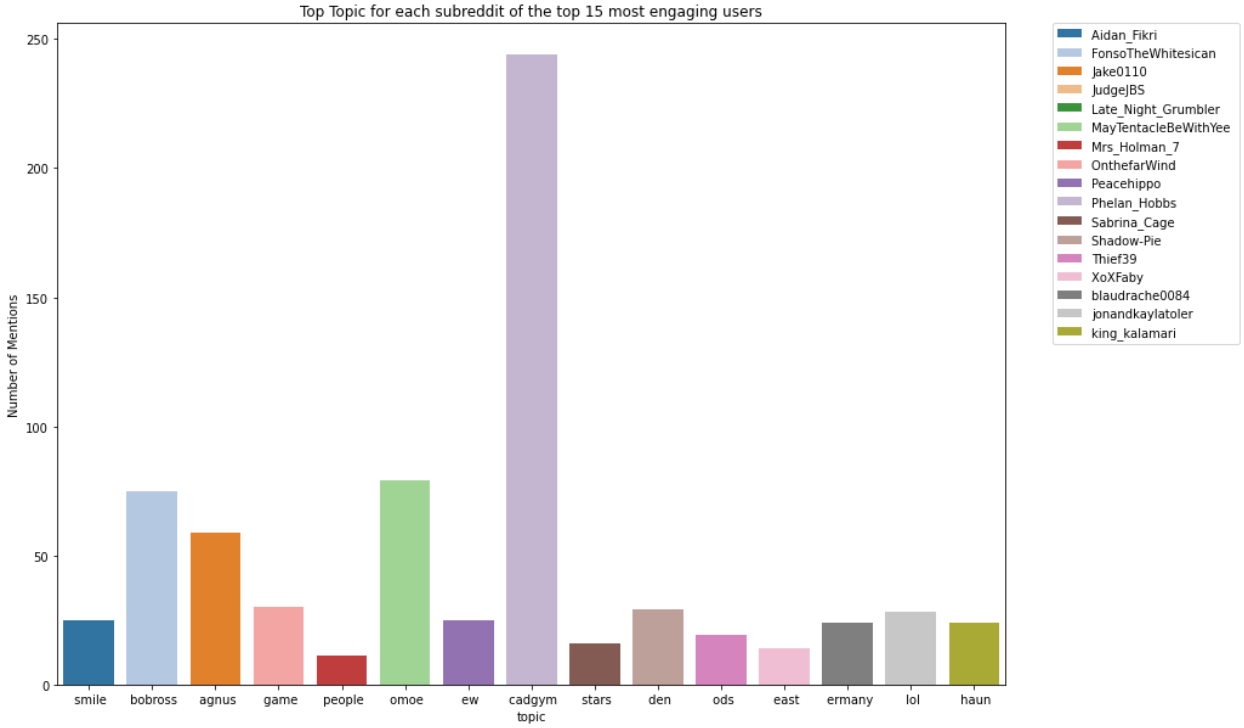
3- Top Topics per Subreddit

The most discussed topics in the most active subreddits are summarized in the graph below



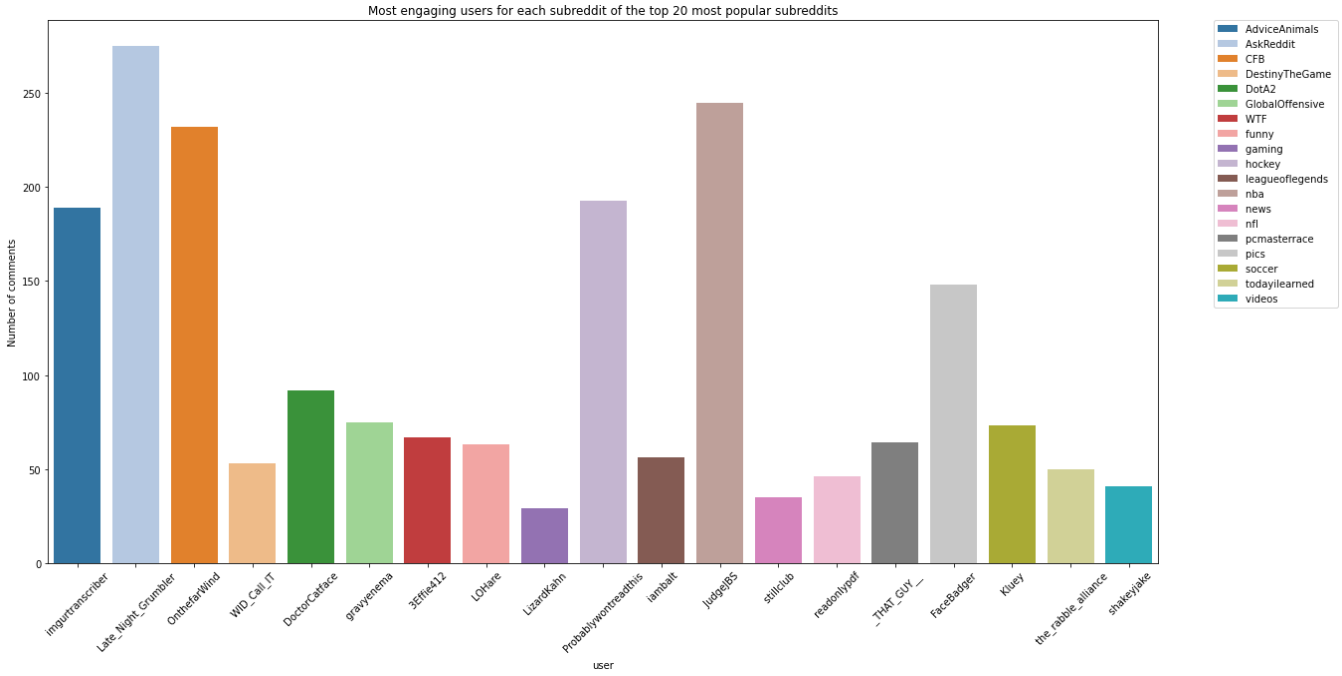
4- Top Topics per User

The most discussed topics by the most engaging reddit users are summarized in the graph below.



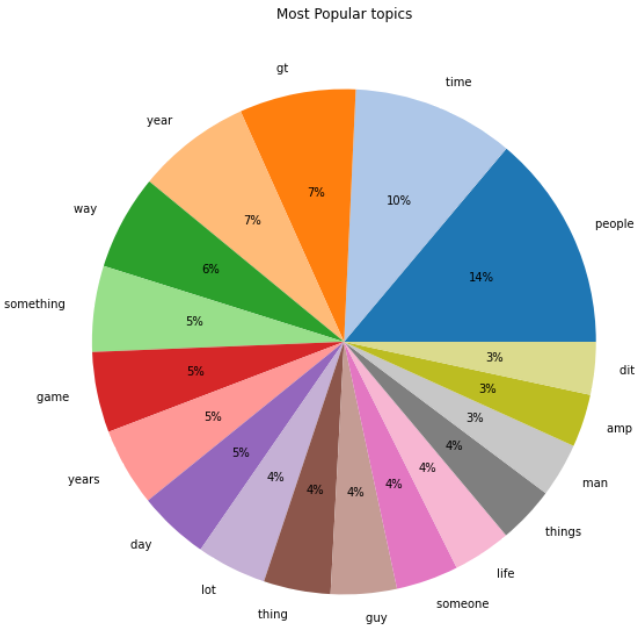
5- Top Users per Subreddit

The below graph shows the most active reddit user in each of the top 20 subreddit.



5- Top Upvoted Topics

The most upvoted topics in the dataset comments are summarized in the pie chart below.



Notes and References

- We used google colab as an interactive shell and coding environment to run our map-reduce tasks. We used the following tutorial to install Hadoop dfs on colab. https://github.com/anjalysam/Hadoop
- For the map-reduce tasks we run, one million instances of the dataset were used as we did not have immediate access to a high performance cluster to submit jobs on the whole dataset efficiently.