Big Data Analytics

Mini Project 1

Mohamed Abdelwahed	201801978
Ibrahim Abdalaal	201800224

Dataset description

The dataset is a bzip2 file (5 GB), we decompress this file and it becomes 31.6 GB. After that, we chose a sample from the file which is 3925 lines and made the sample file format to be a text file.

Data analysis (data problems, patterns, noise, outliers)

We notice from different data samples that downs are “zeros”, in all the data samples that are taken to test our codes. We also notice that in these data there are some noise such as ups being “negative numbers”, from these we can see that there are some data problems with noise.

Challenges faced & how they were solved

Ii. For this task, we had to represent the rate of reply and the controversiality of comments. And as the controversiality is zero, we chose to take “score” to be compared to the rate of replies, and we chose this because we noticed that ups and ‘score’ have the same value.

Optimizations

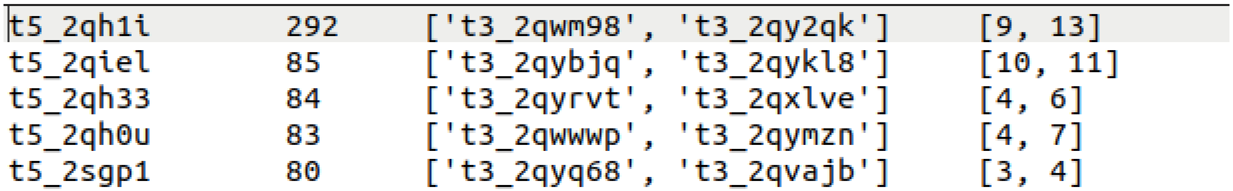
i. For this task, to find the top subreddits with most topics, we choose to make it one mapreduce job instead of two, as the two is only a sample from the large file, so this can be more efficient, and additionally, we add a combiner to reduce the work in the reducer.

Final design of the code detailing each part of the pipeline

1. First task:Most discussed/used topics associated with every subreddit and username with focus on the top subreddits.
    - For this we find the most five subreddits and the most two topics discussed in every subreddit.

- **Mapper:** the output of the mapper of subreddit_id and link_id and ‘1’
- **Combiner:** it takes the output of the mapper and calculates the number of occurrences for each topic from the link_id and the same for subreddit_id.
- **Reducer:**We take the output of the reducer and,we calculate the most five subreddits that are in the dataset, and we try to find the most discussed two topics in these five subreddits.

  Order of the image below(most subreddits,count,most topics,counts)

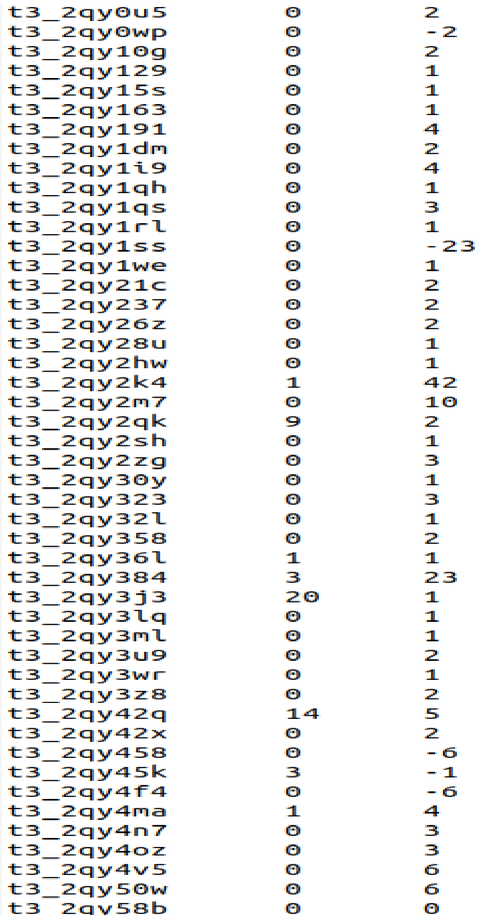
  ```
  t5_2qh1i        292     ['t3_2qwm98', 't3_2qy2qk']     [9, 13]
  t5_2qiel        85      ['t3_2qybjq', 't3_2qykl8']     [10, 11]
  t5_2qh33        84      ['t3_2qyrvt', 't3_2qxlve']     [4, 6]
  t5_2qh0u        83      ['t3_2qwwwp', 't3_2qymzn']     [4, 7]
  t5_2sgp1        80      ['t3_2qyq68', 't3_2qvajb']     [3, 4]
  ```
-
- **Important Note**: for this task,we implemented mapper ,combiner,reducer -files mapper11,combiner11,reducer11 -it works fine on visual studio and gives correct results.However,it gave strange results with hadoop.

  We realized that the combiner is the problem as it might not work in all situations and its execution is not guaranteed.

  So,we implemented another mapper and reducer only for this task(files mapper1,reducer1)

2. Second task:Rate of replies compared to controversiality of comment/post
   - We need to find the number of replies and the controversiality of each comment to show how popular this comment is throughout our datasample.

- So, to Find the main comment we assume that the comment is when the link_id and the parent_id are equal, and we take the score of that comment as the main Controversiality value.
- And to find the number of replies, we count the number of parent_id for every link_id and this represents how many replies this comment has which can be a score for the rate of replay.
- Our first assumption was that parent_id is the the the comment and when the name_id are replies and when name_id is equal to parent_id that can be the main comment, but through that approach we didn’t find any equal values in our sample
- **Mapper:** the output of the mapper is parent_id, contra(which is score value), link_id and ‘1’
- **Reducer:** its takes the output of the mapper, and check the equality of parent_id and link_id to find the main comment, then we calculate the number of parent_id to find the number of number of replies for each link_id and, we show each comment with the the number of replies and the second column is the score.
- From the results, we can see that there is the number of replies for comment, and the score does not correlate with each other.

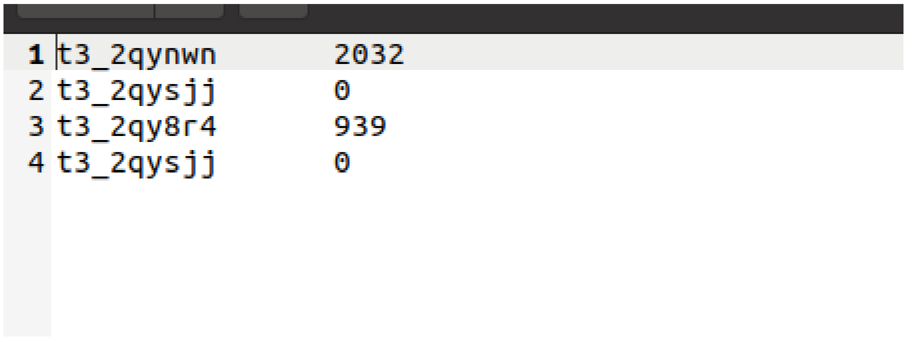


```
t3_2qy0u5       0       2
t3_2qy0wp       0       -2
t3_2qy10g       0       2
t3_2qy129       0       1
t3_2qy15s       0       1
t3_2qy163       0       1
t3_2qy191       0       4
t3_2qy1dm       0       2
t3_2qy1i9       0       4
t3_2qy1qh       0       1
t3_2qy1qs       0       3
t3_2qy1rl       0       1
t3_2qy1ss       0       -23
t3_2qy1we       0       1
t3_2qy21c       0       2
t3_2qy237       0       2
t3_2qy26z       0       2
t3_2qy28u       0       1
t3_2qy2hw       0       1
t3_2qy2k4       1       42
t3_2qy2m7       0       10
t3_2qy2qk       9       2
t3_2qy2sh       0       1
t3_2qy2zg       0       3
t3_2qy30y       0       1
t3_2qy323       0       3
t3_2qy32l       0       1
t3_2qy358       0       2
t3_2qy36l       1       1
t3_2qy384       3       23
t3_2qy3j3       20      1
t3_2qy3lq       0       1
t3_2qy3ml       0       1
t3_2qy3u9       0       2
t3_2qy3wr       0       1
t3_2qy3z8       0       2
t3_2qy42q       14      5
t3_2qy42x       0       2
t3_2qy458       0       -6
t3_2qy45k       3       -1
t3_2qy4f4       0       -6
t3_2qy4ma       1       4
t3_2qy4n7       0       3
t3_2qy4oz       0       3
t3_2qy4v5       0       6
t3_2qy50w       0       6
t3_2qy58b       0       0
```

3. Third task:Topics that yield the highest number of upvotes and/or lowest of downvotes

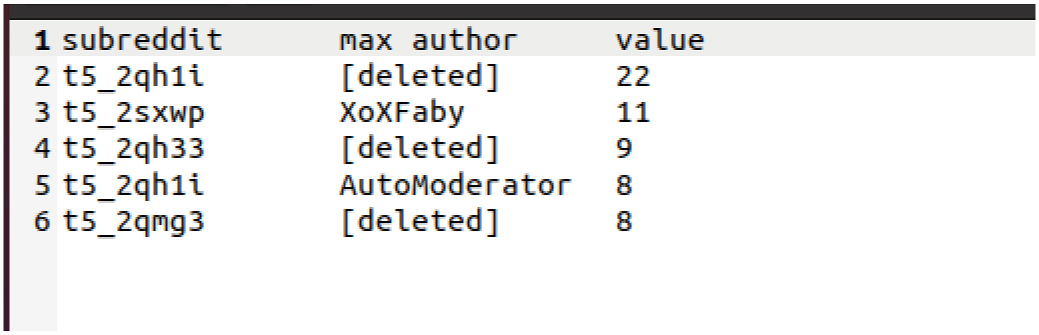
    - The task is trying to find the highest number of votes for some topics and the lowest number of votes for the other topic.
    - **Mapper:** The output of the mapper is the link_id, its corresponding ups and its corresponding downs.
    - **Reducer:** We make a summation for all upvotes and downvotes for each link_id, and we choose to show the highest two upvotes and lowest two downvotes along with link_id

- We notice that the min is the same as it is zero and from our data analysis all downvotes are zero.



```
t3_2qynwn       2032
t3_2qysjj       0
t3_2qy8r4       939
t3_2qysjj       0
```

4. Fourth task(Creative/Innovative Requirements to get more insights, information and/or suggestions): we choose the highest author that has been active on a subreddit.
    - For this task we can know who are most active members, so we try to find number of the different authors through the subreddit.
    - **Mapper:** the output is subreddit_id and the author
    - **Reducer:** we count the number of all authors in each subreddit, and we find the most active author from its number of occurrences for each subreddit



```
subreddit       max author      value
t5_2qh1i        [deleted]       22
t5_2sxwp        XoXFaby         11
t5_2qh33        [deleted]       9
t5_2qh1i        AutoModerator   8
t5_2qmg3        [deleted]       8
```