Queries - Lab Report - 2 (Team - 8)

Shashank Reddy Kandimalla (016799523), Rohith Reddy Vangala (016762109), Sakshi Manish Mukkirwar (016794765), Swati (016702413), Yamini Muthyala (016766165)

1. Users most discussing about it on twitter(tweets):

```
Query:
```

```
db.Tweet.aggregate([ {$match: {Text: /ChatGPT/i}}, {$group: {_id: {$substr: ["$User", 20, -1]}, count: {$sum: 1}}}, {$sort: {count: -1}}, {$limit: 3}]).pretty()
```

Interaction with database: The query uses the MongoDB aggregation pipeline to search for tweets in a collection that contain the string "ChatGPT" in their Text field. It then groups the matching tweets by user_name and counts the number of tweets posted by each user, sorts the results in descending order based on the total number of tweets, and returns only the top 3 users with the highest number of tweets related to "ChatGPT". Finally, the results are formatted in a human-readable format using the pretty function.

2. Most viral tweet (wrt retweets) about the chatgpt(tweet counts, tweets):

Ouerv:

```
db.Tweet.find({Text: /ChatGPT/i}).sort({"RetweetCount": -1}).limit(1).pretty()
```

Interaction with database: The query searches for tweets in a MongoDB collection that contain the string "ChatGPT" in their Text field, sorts them in descending order based on their RetweetCount, returns the top result, and formats it in a human-readable format using the pretty function.

1

3. Hashtags used prominently for the tweets about chatgpt(tweet_hashtags, tweet_hashtag_mappings):

```
Query:
```

Interaction with database: The query searches for tweets in a MongoDB collection that contain the string "ChatGPT" in their Text field, extracts and counts the hashtags used in those tweets, removes any empty hashtags, and sorts the remaining hashtags in descending order based on the count of their occurrences.

Interaction with database: The query uses the MongoDB aggregation pipeline to search for tweets in a collection that contain the string "ChatGPT" in their Text field. It then groups the matching tweets by ConversationId, counts the number of conversations and replies, and selects the Username of the first tweet in each conversation. The results are sorted in descending order based on the conversation count and limited to the top result. Finally, the allowDiskUse option is used to allow MongoDB to use disk space to store temporary data if necessary.

5. Latest tweet(tweet): Query: db.Tweet.find({}, {_id: 1,"Tweet Id":1, Text: 1, Datetime: 1 }).sort({ Datetime: -1 }).limit(1)

Interaction with database: The query searches for tweets in a MongoDB collection and retrieves the _id, Tweet Id, Text, and Datetime fields of each tweet. The results are sorted in descending order based on the Datetime field, and the top result is returned.

```
> db.Tweet.find({},
    { _id: 1,"Tweet Id":1,
    Text: 1, Datetime: 1 }).sort({ Datetime: -1 }).limit(1)

<    {
        id: ObjectId("6445d86638d9db92fc892b9d"),
        Datetime: '2023-01-24 06:58:01+00:00',
        'Tweet Id': 1617778731678044200,
        Text: 'Portland Shop Uses ChatGPT To Tell Family Stories On A Startup Budget https://t.co/rzGvr6yTOc'
}</pre>
```

6. Count of tweets from different devices(tweets):

Query:

```
db.Tweet.aggregate([ \{ surce: \{ surce: \{ surce: \{ surce: \| source", regex: '(?<=\\rangle).*(?=<\\rangle')' \} \} \}, count: \{ sum: 1 \} \} \}, \{ sum: 1 \} \}, \{ surce: \| surce:
```

Interaction with database: The query uses the MongoDB aggregation pipeline to group tweets in a collection by the value of the Source field. It extracts the source name using a regular expression and counts the number of tweets associated with each source. The results are projected to show only the source and count fields and sorted in descending order based on the count.

7. Analyzing which websites or pages are most commonly shared on Twitter(outlinks):

Query:

```
db.Tweet.aggregate([{$unwind: "$Outlinks"},
   {$project: {outlink: {$trim: {input: "$Outlinks", chars: "[]""}}}},
   {$group: {_id: "$outlink", count: {$sum: 1}}},
   {$sort: {count: -1}},{$limit: 3}])
```

Interaction with database: The query extracts outlinks from tweets in a MongoDB collection and counts the number of occurrences of each unique outlink. It uses the \$unwind operator to create a separate document for each outlink, \$project to trim the outlink, and \$group to group by the outlink and count its occurrences. The results are sorted in descending order based on the count.

8. Websites which are not secure in outlinks:

Query:

```
 db.Tweet.aggregate([ { $match: { Outlinks: { $regex: /^(?!https:\\\/).+/ } } }, { $group: { _id: "$Outlinks", count: { $sum: 1 } }, { $sort: { count: -1 } }, { $project: { _id: 0, "Tweet Id":1, Outlinks: "$_id", count: 1 } }, { $limit: 3 } ])
```

Interaction with database: The query uses the MongoDB aggregation pipeline to extract outlinks from tweets in a collection that do not start with "https://". It then counts the number of occurrences of each unique outlink, sorts the results in descending order based on the count, and limits the output to the top 3 results. The \$project operator is used to shape the output to include only the tweet ID, outlink, and count of occurrences.

9. tweets by language:

Query:

```
db.Tweet.aggregate([ {$group: {_id: "$Language", count: { $sum: 1 }}}, {$sort: { count: -1} },{$limit: 3}])
```

Interaction with database: The query uses the MongoDB aggregation pipeline to group tweets by language and count the number of tweets in each language. It then sorts the results in descending order based on the count. This query can be used to identify the number of tweets in each language present in the collection.

10. Time where there are more tweets:

Query:

```
db.Tweet.aggregate([ { $addFields: { $toDate: "$Datetime" }} }, {$group: {_id: { $hour: "$parsedDate" },count: {$sum: 1} }}, {$sort: { count: -1 }}, {$limit: 3}])
```

Interaction with database: This query adds a new field to the documents in the Tweet collection that represents the parsed date from the Datetime field. It then groups the tweets by the hour of the parsed date and counts the number of tweets in each group. Finally, the results are sorted in descending order based on the tweet count.

11. Average tweets at particular time:

Query:

```
db.Tweet.aggregate([ {$addFields: { parsedDate: {$toDate: "$Datetime" }} },{ $group: {_id: { $hour: "$parsedDate" },avg_likes: { $avg: "$LikeCount" } }}, {$sort: {_id: 1 }} ,{$limit: 3}])
```

Interaction with database: The given MongoDB query calculates the average number of likes for tweets posted at each hour of the day. It first converts the Datetime field to a date type. It then groups the tweets based on the hour of the day and calculates the average number of likes for each group. Finally, it sorts the results by the hour of the day in ascending order.

12. Top 3 media based on like count:

Query:

```
db.Tweet.find({ "Media": { $exists: true } }, {"_id": 0, "TweetId": 1, "Text": 1, "LikeCount": 1, "Media": 1} ).sort({ "LikeCount": -1 }).limit(3)
```

Interaction with database: This MongoDB query retrieves the top 3 media objects (tweets that contain media) based on their LikeCount in descending order. The query first filters for tweets that have the "Media" field and selects the relevant fields (TweetId, Text, LikeCount, and Media) for those tweets. The results are sorted in descending order of LikeCount and limited to 3.

13. Finding the top 3 users with the highest engagement (total sum of reply count, retweet count, like count, and quote count):

Query:

```
db.Tweet.aggregate([ {$group: {
    _id: "$Username", totalEngagement: {$sum: { $add: [ "$ReplyCount", "$RetweetCount", "$LikeCount", "$QuoteCount" ]}
}} },{$sort: {"totalEngagement": -1 }}, {$limit: 3 }])
```

Interaction with database: This MongoDB query aggregates tweets by the Username field and calculates the total engagement for each user by summing up the ReplyCount, RetweetCount, LikeCount, and QuoteCount. The resulting documents are then sorted by the totalEngagement field in descending order and only the top 3 are returned.

14. Most mentioned users:

Query:

```
db.Tweet.aggregate([{ $unwind: "$MentionedUsers"}, {$group: {_id: "$MentionedUsers", count: { $sum: 1 }} },{$sort: {count: -1 }}, {$limit: 3 }])
```

Interaction with database: This MongoDB query starts by unwinding the MentionedUsers array, then grouping the tweets by the mentioned user and counting the number of occurrences. The result is then sorted in descending order by the count and limited to the top 3 mentioned users.

```
db.Tweet.aggregate([{ Sunwind: "SMentionedUsers"}, {Sgroup: { _id: "SMentionedUsers", count: { Ssum: 1 }} }, {Ssort: {count: -1 }}, {Slimit: 3 }])

< {
    __id: "[User(username='GRDecter', id=1281457267582177280, displayname='Genevieve Roch-Decter, CFA', description=None, rawDescription=None, descriptionUrls=None, verified=None, count: 332
}

{
    __id: "[User(username='OpenAI', id=4398626122, displayname='OpenAI', description=None, rawDescription=None, descriptionUrls=None, verified=None, created=None, followersCount=No count: 387
}

{
    __id: "[User(username='YouTube', id=10228272, displayname='YouTube', description=None, rawDescription=None, descriptionUrls=None, verified=None, created=None, followersCount=No count: 283
}</pre>
```

15. Most occurred word in the text field:

Query:

```
db.Tweet.aggregate([{"$match": {"Text": {"$regex": "^(?!.*ChatGPT).*", "$options": "i"}}}, {"$project": {"words": {"$split": ["$Text", " "]}}}, {"$unwind": "$words"}, {"$group": {"_id": "$words", "count": {"$sum": 1}}},{"$sort": {"count": -1}},{"$limit": 3} ])
```

Interaction with database: This query matches tweets that don't contain the phrase "ChatGPT" in the text field, splits the text into words, unwinds the resulting array, groups by each unique word and counts their occurrences, sorts the result in descending order of count, and returns the top 10 most occurred words.

16. Calculate the average engagement per like, retweet, and reply across the entire collection:

Query:

Interaction with database: This MongoDB query calculates the average engagement per like, retweet, and reply across the entire collection of tweets. It first matches tweets with at least one engagement (like, retweet, or reply), then groups all tweets to calculate the total number of likes, retweets, replies, and total engagement. Finally, it projects the average engagement per like, retweet, and reply by dividing the total engagement by the total number of likes, retweets, and replies, respectively. The result does not have an " id" field and contains only the average engagement per like, retweet, and reply.

17. Finding the tweets with the highest ratio of retweets to likes:

Query:

```
db.Tweet.aggregate([{$match: {RetweetCount: {$gt: 0}}, LikeCount: {$gt: 0}}}, {$project: {_id: 0, "Tweet Id": 1, RetweetToLikeRatio: {$divide: ["$RetweetCount", "$LikeCount"]}}}, {$sort: {RetweetToLikeRatio: -1}}, {$limit: 3} ])
```

Interaction with database: This MongoDB query finds the tweets with the highest ratio of retweets to likes in the collection of tweets. It first filters for tweets with non-zero RetweetCount and LikeCount, then calculates the RetweetToLikeRatio by dividing the RetweetCount by the LikeCount for each tweet. The results are sorted in descending order by RetweetToLikeRatio and limited to the top 10 tweets. The output includes the Tweet Id and RetweetToLikeRatio fields for each of the 10 tweets.

18. Updating the Languages name to their Full Name::

```
Query:
```

```
db.Tweet.updateMany(
    { Language: "en" },
    { $set: { Language: "English" } }
);
```

Interaction with database: This MongoDB query updates all documents in the "Tweet" collection where the "Language" field equals "en". It sets the "Language" field to "English".

19. Deleting a particular Tweet:

Query:

db.Tweet.deleteOne({ "Tweet Id": 1617156308926349312 })

Interaction with database: This MongoDB query deletes a single document from the "Tweet" collection where the value of the "Tweet Id" field is equal to 1617156308926349312.

THE END