Mini Project 1 - Hadoop Ecosystem

Overview

Reddit is a website comprising user-generated content—including photos, videos, links, and text-based posts—and discussions of this content in what is essentially a bulletin board system. The name "Reddit" is a play-on-words with the phrase "read it", i.e., "I read it on Reddit.". As of 2018, there are approximately 330 million Reddit users, called "redditors". The site's content is divided into categories or communities known on-site as "subreddits", of which there are more than 138,000 active communities.

As a network of communities, Reddit's core content consists of posts from its users. Users can comment on others' posts to continue the conversation. A key feature to Reddit is that users can cast positive or negative votes, called upvotes and downvotes respectively, for each post and comment on the site. The number of upvotes or downvotes determines the posts' visibility on the site, so the most popular content is displayed to the most people.

Dataset

The Dataset consists of a list of comments taken for the Reddit public API from various subreddits.

- The dataset is one file with each comment in JSON format in one line (check appendix section for format)
- Total Comments: 53,851,542
- Compression Type: bzip2 (5,452,413,560 bytes compressed | 31,648,374,104 bytes uncompressed)
- Dataset can be found here (Google Drive)
- Do NOT open the dataset file on an editor as it may cause a crash or a freeze of your application or your system
- You can explore/generate your own samples using these commands
  Windows (Powershell):

```
gc <file_name> | select -first <line_number> >> sample.out
```

  Linux:

```
head -n <line_number> <file_name> >> sample.out
```

Requirements

The dataset is useful for a wide range of experiments/analyses because it's a large collection of time stamped events with interesting features (username, body text, post location). There are three categories of requirements.

1. Basic Requirements:
    a. Write "CLEAN", well-documented code & produce indicative visualisations to:
        i. Most discussed/used topics associated with every subreddit and username with focus on the top subreddits
        ii. Rate of replies compared to controversiality of comment/post
        iii. Topics that yield the highest number of upvotes and/or lowest of downvotes
    b. Write a "CLEAN", organized document detailing the process needed to reach the results with focus on these points
        i. Data analysis (data problems, patterns, noise, outliers)
        ii. Challenges faced & how they were solved
        iii. Optimizations
        iv. Final design of the code detailing each part of the pipeline
2. Discussions & Executions with the following process:
    a. Do experiments, trial runs, and/or section runs
    b. Discuss findings
    c. Running on the cluster in the HPC lab
3. Creative/Innovative Requirements to get more insights, information and/or suggestions (This is completely up to the students, the following are just suggestions):
    a. % of negative / positive attitude of comments per subreddits/users
    b. Predict posts/subreddits a user will next engage with (i.e. recommender systems)
    c. Community detection with ground truth (subreddits)
    d. Track memes

Rubric

- 60% - Code to get the 3 basic requirements (Frequent topics, rate of replies, topics impact on upvotes/downvotes)
- 10% - Process document
- 10% - Discussions & executions (discussion, experiment runs, and final cluster run)
- 20% - Creative/innovative requirements
- - 5% per instruction broken, for any submission/delivery criteria not followed
- AN INSTANT ZERO (if you are lucky) for plagiarism/cheating/copying for all students involved

Restrictions

1. You have to use Hadoop Ecosystem
2. You have to use the supplied dataset

3. Your code has to run OFFLINE, i.e. no internet connection available
4. NO restriction on language
5. NO restriction on utility libraries (but you have to discuss with me first for cluster setup)
6. NO restriction on number of MapReduce jobs (even though less means more efficient which means higher marks)

Delivering Process

Deliverables:

1. MapReduce job code
2. Script to run/setup the job requirements to be run on the cluster namenode/master
3. Document with visualisations

Process milestones:

1. Team assignment (teams of 2)
2. Mid-project discussion (online, per student request)
3. Cluster run (details will be shared later)

Appendix

Dataset Comment Format:


```
{
    "gilded": 0,
    "author_flair_text": "Male",
    "author_flair_css_class": "male",
    "retrieved_on": 1425124228,
    "ups": 3,
    "subreddit_id": "t5_2s30g",
    "edited": false,
    "controversiality": 0,
    "parent_id": "t1_cnapn0k",
    "subreddit": "AskMen",
    "body": "I can't agree with passing the blame, but I'm glad to hear it's at least helping you with the anxiety. I went the other direction and started taking responsibility for everything. I had to realize that people make mistakes including myself and it's gonna be alright. I don't have to be shackled to my mistakes and I don't have to be afraid of making them. ",
    "created_utc": "1420070668",
    "downs": 0,
    "score": 3,
    "author": "TheDukeofEtown",
    "archived": false,
    "distinguished": null,
    "id": "cnasd6x",
    "score_hidden": false,
    "name": "t1_cnasd6x",
    "link_id": "t3_2qyhmp"
}
```