Reddit Review Analysis

STAT 605 project in CHTC

Lingfeng ZHU; Yijie LIU; Zhao LI; Kunning WANG

December 9, 2019

UW Madison

Introduction

The main purpose of this project is to analyze the popularity and user actvations of subreddits based on the May 2015 Reddit Comments dataset.

- Filtered out 100 subreddits with the most amount of comments during May 2015;
- Analyzed the word cloud plots of these subreddits;
- Found some most active subreddits by sorting them by the average amount of comments per user;
- Analyzed the tendency of popularity over time of popular subreddits.

Dataset

Our dataset is May 2015 Reddit Comments from Kaggle. The dataset's size is 30GB. It contains 1.7 billion comments from Reddit in May 2015 and related information such as auther, subreddit, etc.

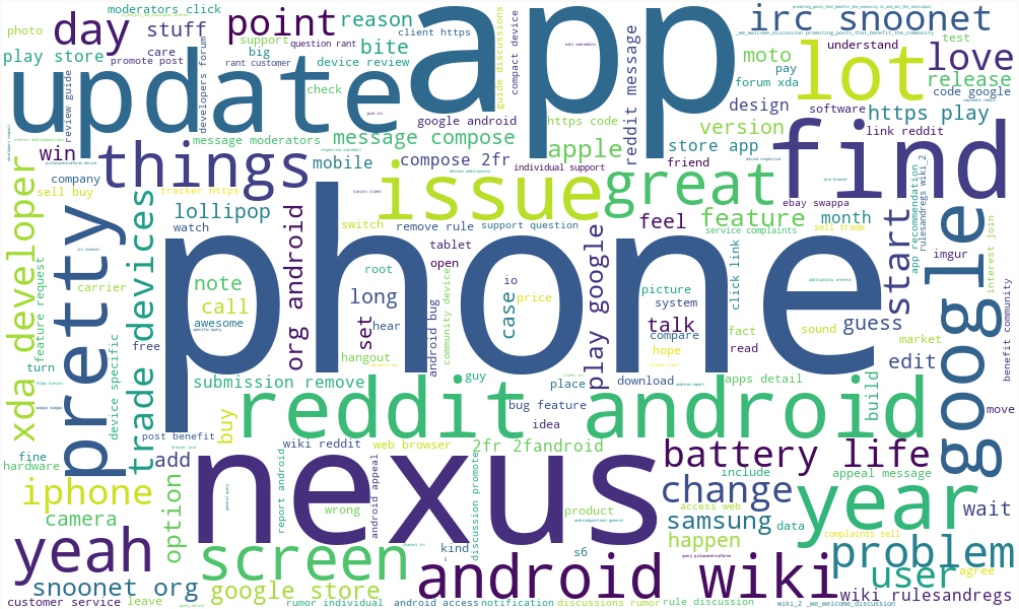
We used following variables:

- **subreddit**: The subreddit in which the comment appears
- **author**: The author of the comment
- **body**: The body text of the comment
- **id**: ID of the comment

Wordcloud of Subreddits

We ran 100 parallel jobs for each subreddits. And in each job we:

- combine the comments into a single text
- transfer all the letters into lower case
- transfer all the words into prototype
- set stopwords to delete meaningless words
- plot wordcloud



Android subreddit

Most active subreddits

We used the average amount of comments per user to evaluate the active level of each subreddit in May 2015. For a subreddit, its active level is given by:

$$activeLevel(subreddit) = \frac{\text{\# of comments in subreddit}}{\text{\# of related users}}$$

It was reasonable to use this metrics to measure the active level because the subreddit with a higher average user active level could be considered to be more active.

Most active subreddits

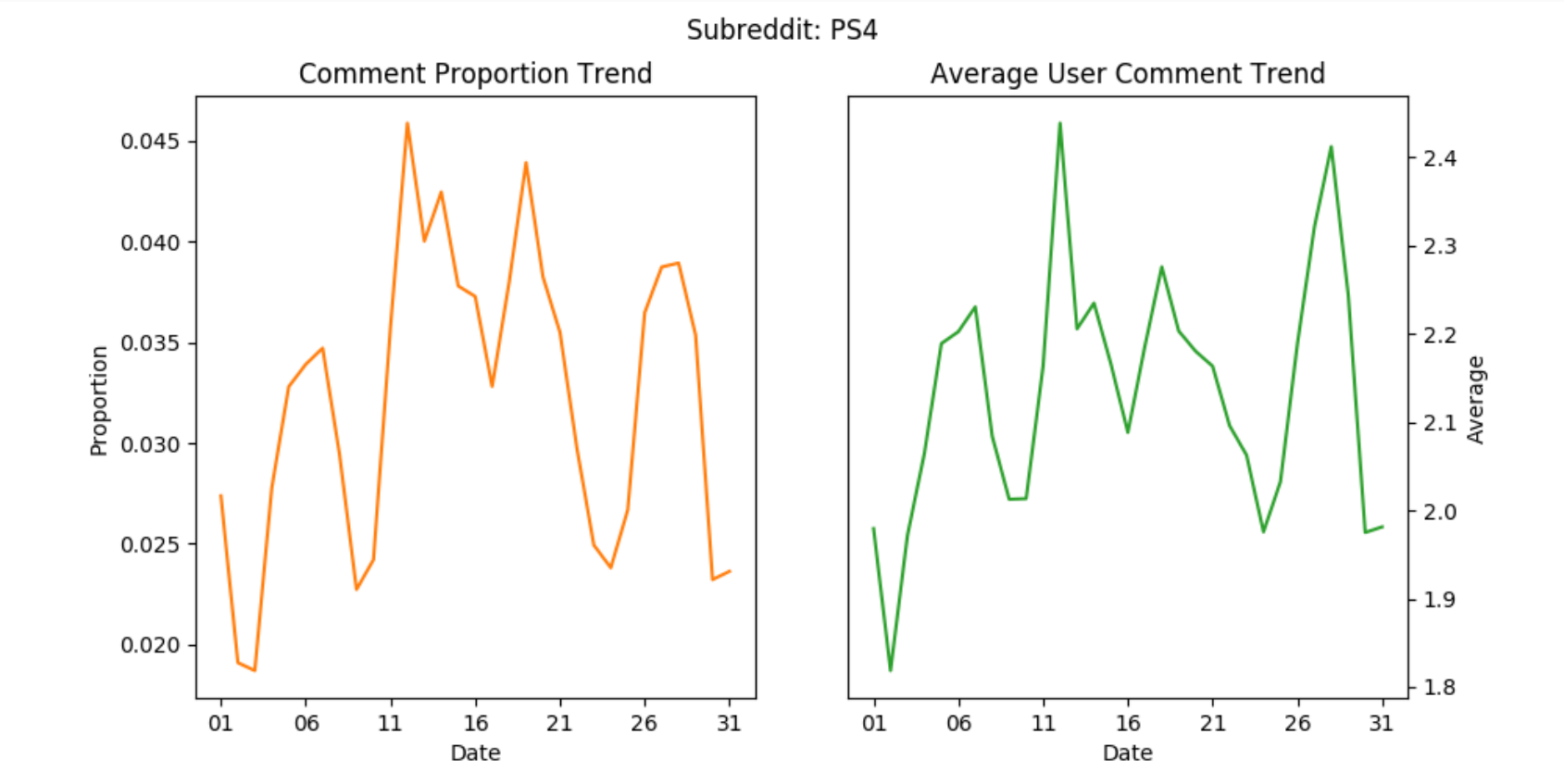
We can compute the value of this metrics for every subreddit in parallel using CHTC. Then, we can sort the subreddits by the active level, and the top 5 most active subreddits can be seen bellow:

Table 1: Top 5 Active Subreddits

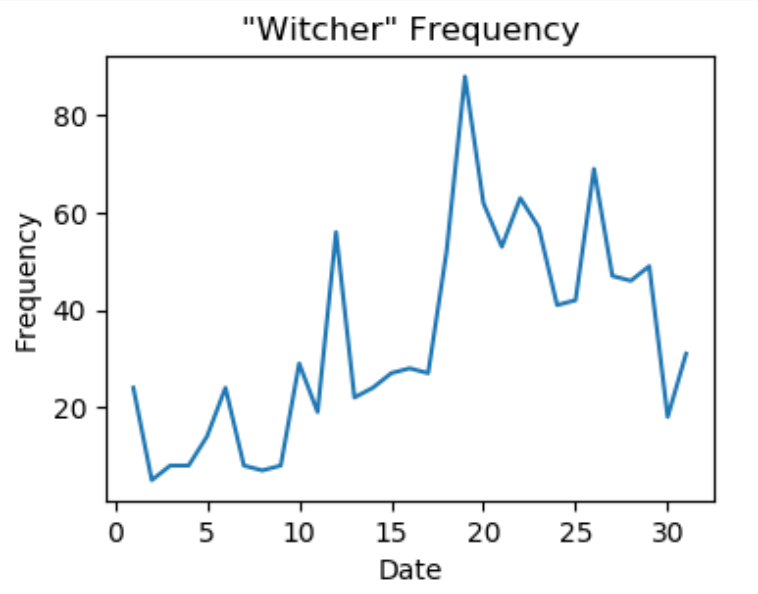
Top	subreddit	Active Level
1	Random Acts Of Amazon	54.53407683
2	newsokur	52.71981321
3	Pokemon Trades	43.17164179
4	Global Offensive Trade	29.00911372
5	India	20.83660602

Daily Trend

With the data from every day in May, we can also analysis the daily trend of comments and average comments per user to have a more detailed insight. We also use parallel computing in each subreddit to make plots.



Further Insight



Top 5 Active Subreddits

Game	Launch Date
Final Fantasy IV	May 12
Project CARS	May 12
The Witcher 3: Wild Hunt	May 19
Splatoon	May 29

The launch date of these popular games all correspond to the peaks of the plots of Subreddit Trend.

Wordcloud of r/PS4

People mentions a lot about 'Witcher' and like to share their feeling about this game and other games. This month players mostly have a positive attitude towards new games.

Weakness

We only did preliminary work. We only roughly show hot topics in these subreddit and how the subreddit activity trend goes in May 2015. If someone is interested in a particular subreddit or topic, he/she needs to analyse him/herself with our method.

Results

We found out active communities such as Random Acts Of Amazon, Pokemon Trades. We also plotted user activity trend and word activity trend. Based on our work, Reddit can:

- Monitor active communities and users.
- Relate comment-activity trend with big events such as game lunch and Black Friday together, then allocate sources more efficiently.
- Monitor the daily trend of a word appearance such as ”witcher” to see what topics are popular.

Questions?