Reddit comments analysis

Carel Kuusk

The code and the dataset that will be used in this project are available at GitHub .

Related work

The aim of this project is to get a better sense of how Reddit evolved in its first years. For this we will be analyzing the network structure of comments from 2005 to 2008. Possible analysis methods are provided in Cordeiro et al. (2018): snapshot analysis to identify more static features of the network structure and aggregate analysis (either via landmark windows or sliding windows) for analysing statistical properties of the graphs within a certain timeframe.

Troy Steinbauer has analyzed the overall properties of the Reddit network in 2011. He identified that for example related subreddit distribution was following the expected power-law distribution.

The behaviour of individual users was analyzed by Thukral et al in 2018. The analysis was partly done on the 2008 and partly on the 2014-2015 period. They discovered, that based on the commenting and posting patterns, the users can be clustered into three types based on when most of their contributions are made (either within a short period from sign-up, stably, or more actively after a long hibernation). Also they discovered a clear separation between commenters and posters.

Dataset

The dataset consists of all of Reddit's comments from 2005 to 2008 (included), that is from the inception of the Reddit. This provides an incredible opportunity to investigate initial stages of an online social network. The data was scraped by Reddit user u/Stuck_In_the_Matrix in 2016 and torrented from the Academic Torrents website. The original dataset includes comments up to at least 2015, but it would not have been feasible to include the whole of Reddit's comment dataset. However, the compressed dataset size is still 932 Mb, and covers the initial growth phase of the site, including the first year during which Reddit let users create their own subreddits.

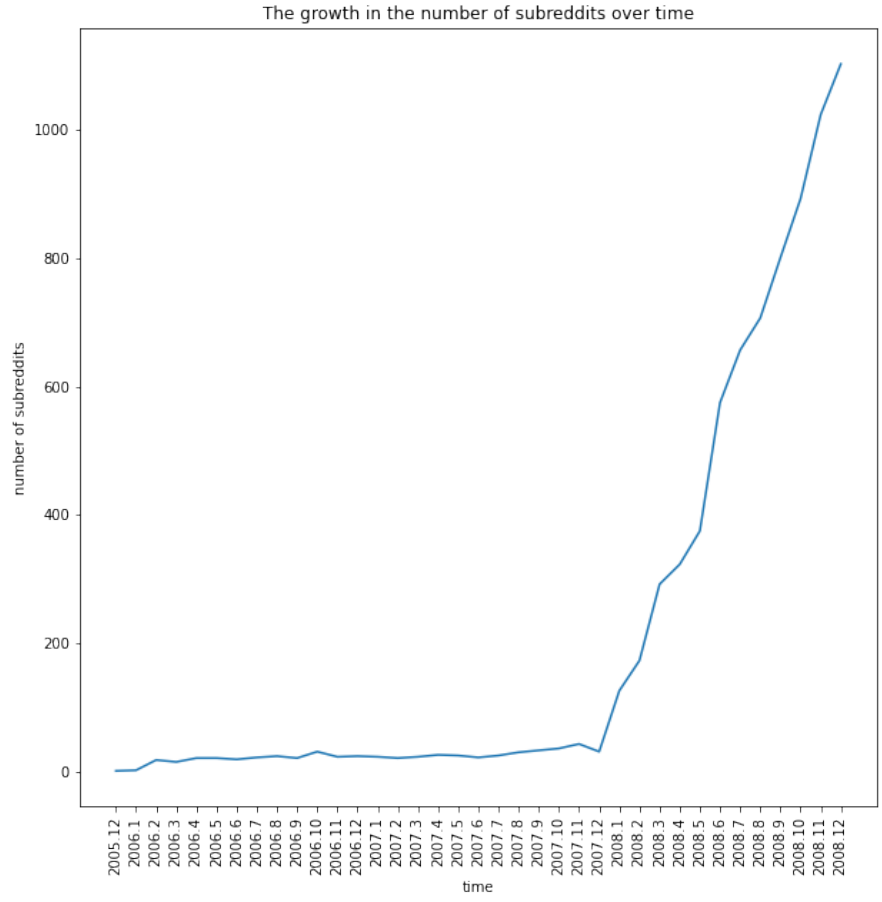
Each line in the decompressed dataset contains data about one comment and is organized as a JSON document. The relevant fields are as follows:

- `subreddit` – the subreddit under which the original link is;
- `subreddit_id` – the ID of the subreddit;
- `author` – the username of the author;
- `body` – the body of the comment;
- `score` – the karma on the comment;
- `link_id` – the ID of the link under which the comment was posted;
- `parent_id` – the ID of the parent, which can be either the link or another comment;

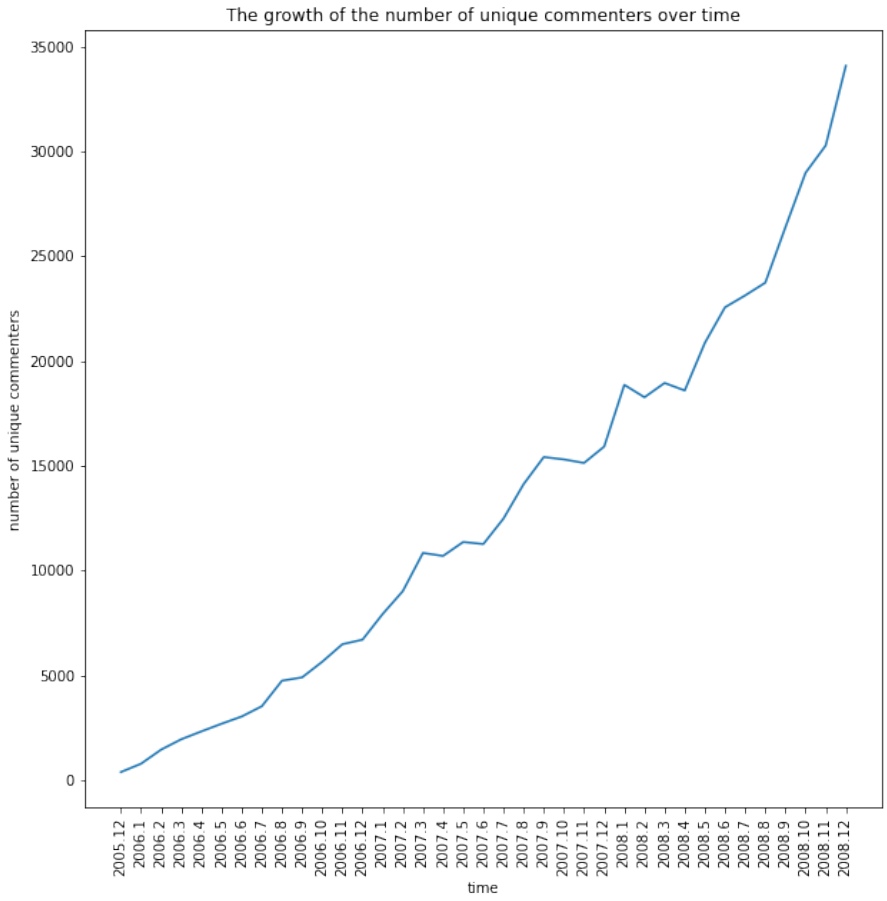
- `id` – the ID of the comment itself;
- `created_utc` – the UNIX timestamp of the comment creation.

There are more metadata, e.g. a field `controversiality`, which measures whether a comment has received a similar number of upvotes and downvotes, etc. Also, during the first years there were less fields (downvotes for example did not appear before 2008), so the selection was partly motivated by ensuring that all the fields would be present during the whole period under analysis.

Some basic cleaning was applied to the data. First, the fields `link_id`, `parent_id` needed some preprocessing to be compatible and comparable with the field `id`. Secondly, due to a relatively large number of comments authored by now-deleted users, some preliminary analysis steps required eliminating comments that were authored by the deleted users. This is because all the deleted users were renamed to `"[deleted]"`, i.e. the statistics about user comment distributions, max number of posts, etc were skewed due to the there being many more comments authored by a deleted user.



From the figure above we can see that the number of subreddits grew rapidly and then exploded at the start of 2008 due to Reddit allowing users to create their own subreddits.



The steady but fast growth of users is also clear from the data.

Methodology

The methodology in this project follows the steps below.

1. Finding and downloading data.
2. Preliminary preprocessing, ensuring the high quality of data.
3. Preliminary data analysis, investigating potentially interesting further research paths.
4. Add a sentiment feature to comments (by applying a pre-trained model on the comments).
5. Create graph(s).
6. Link prediction.

(For the second checkpoint, the first three are done). For the graphs, there are many options, which I will probably explore. First, creating a bipartite graph of users and subreddits and predicting from past behaviour, if and which people would join which subreddits. It is also possible to create a graph of co-commentators – a graph of people who have posted under the same post. This enables to explore communities and their formation more in detail.

A good thing about this dataset is that this is naturally temporally ordered. The dataset naturally comes in months, but each comment also has a timestamp on it, so it is relatively natural to analyze community formations.

There will be two types of analysis that will be carried out.