Reddit comments, May 2015

A reddit user has released a giant data (https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/) set with reddit comments. There are already a wealth of visualizations being done on this dataset, but better ones could be made and interactivity could uncover even more info. Each data entry contains the following features:

- created_utc
- ups
- subreddit_id
- link_id
- name
- score_hidden
- authorflaircss_class
- authorflairtext
- subreddit
- id
- removal_reason
- gilded
- downs
- archived
- author
- score
- retrieved_on
- body
- distinguished
- edited
- controversiality
- parent_id

The following is an example of a single entry:

```
{
   "archived": false,
   "author": "TheDukeofEtown",
   "author_flair_css_class": "male",
   "author_flair_text": "Male",
   "body": "I cant agree with passing the blame, but Im glad to hear its 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 its gonna be alright. I dont have to be shackled to my mistakes and I dont have to be afraid of making them. ",
   "controversiality": 0,
   "created_utc": "1420070668",
   "distinguished": null,
   "downs": 0,
   "edited": false,
   "gilded": 0,
   "id": "cnasd6x",
   "link_id": "t3_2qyhmp",
   "name": "t1_cnasd6x",
   "parent_id": "t1_cnapn0k",
   "retrieved_on": 1425124228,
   "score": 3,
   "score_hidden": false,
   "subreddit": "AskMen",
   "subreddit_id": "t5_2s30g",
   "ups": 3
}
```

Because of the size of the set, we suggest only looking at the May 2015 comments (https://www.kaggle.com/reddit/reddit-comments-may-2015/downloads/reddit-comments-may-2015.7z) (use 7zip to decompress). Examples of “data stories” that could be investigated:

- How often can we infer gender from flair and is voting response (ups, downs, controversiality) dependent on this?
- Which subreddits are nicer to participants?
- Are gilded users popular in general?