Subreddit Clustering

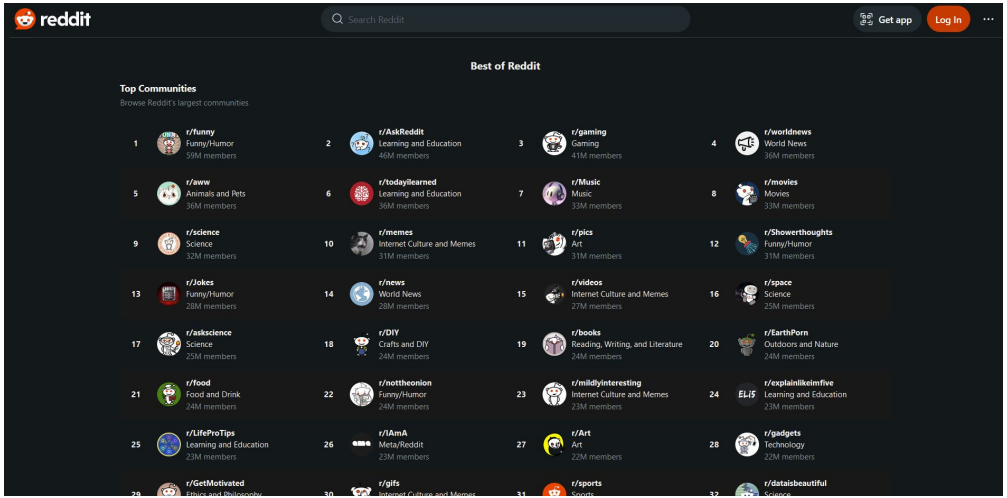
Maddy Powers

Motivation

- Didn’t want to use a pre-existing dataset
- There are a lot of subreddits
- Can they be automatically grouped by topic?
- Does the topic of a subreddit influence user’s language?

Finding a list of subreddits

- Reddit has a list of top subreddits, but it excludes some
  - No r/AmItheAsshole
- Couldn't find any extensive alternatives
- Scraped with 2500 beautifulsoup
  - Wanted a wide variety of subreddit
  - Also wanted to include r/pittsburgh
    - :)

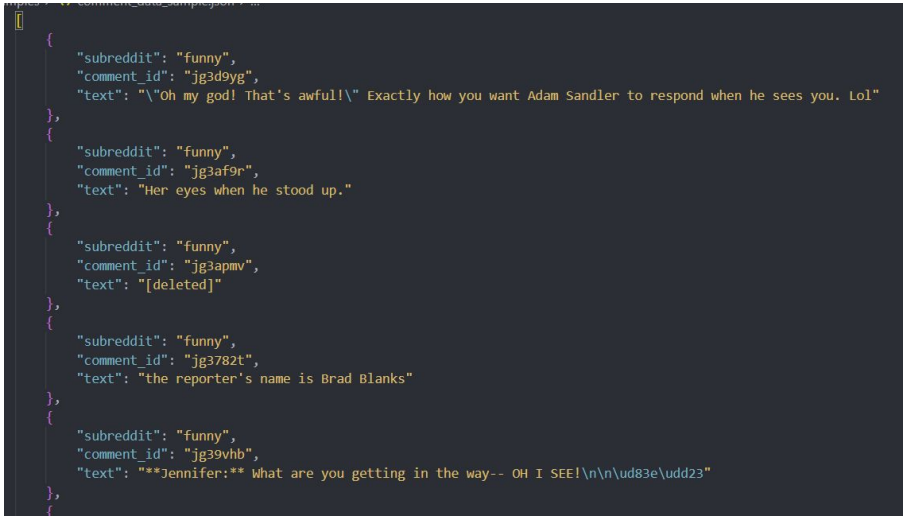


Scraping comments

- Chose comments to avoid dealing with text posts vs image posts
- Used PRAW for scraping
- EXTREMELY slow
  - Rate limits
  - Left computer on overnight twice
  - Made my script resumable
- Ignored AutoModerator

The data

- JSON file with a list of comment objects
- Each comment has
  - Subreddit
  - Comment id
  - Comment Text



Sharing the data

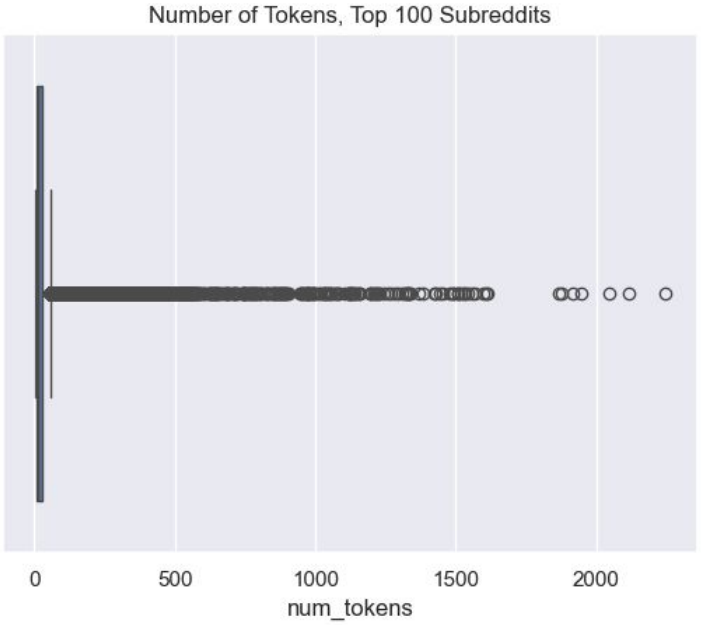
- Inspired by GUM corpus
- Uploaded to github with text field removed
- Script to fetch exact comments used
- PRAW lets you batch requests
  - Faster than initial scraping
  - Still will take a few hours for biggest file
- Needed git LFS for the large files

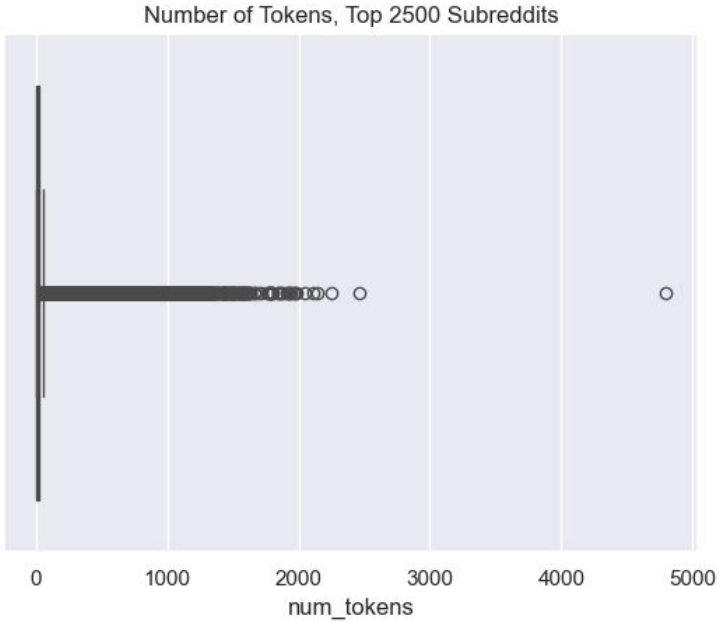

```
[{"subreddit": "funny", "comment_id": "jg3d9yg"}, {"subreddit": "funny", "comment_id": "jg3af9r"}, {"subreddit": "funny", "comment_id":
"jg3apmv"}, {"subreddit": "funny", "comment_id": "jg3782t"}, {"subreddit": "funny", "comment_id": "jg39vhb"}, {"subreddit": "funny", "comment_id":
"jg36sqa"}, {"subreddit": "funny", "comment_id": "jg35kg8"}, {"subreddit": "funny", "comment_id": "jg35mu5"}, {"subreddit": "funny", "comment_id":
"jg39chw"}, {"subreddit": "funny", "comment_id": "jg3brtc"}, {"subreddit": "funny", "comment_id": "jg3czp8"}, {"subreddit": "funny", "comment_id":
"jg3dgxk"}, {"subreddit": "funny", "comment_id": "jg3bw93"}, {"subreddit": "funny", "comment_id": "jg3pgwz"}, {"subreddit": "funny", "comment_id":
"jg3csm1"}, {"subreddit": "funny", "comment_id": "jg3g507"}, {"subreddit": "funny", "comment_id": "jg3g81v"}, {"subreddit": "funny", "comment_id":
"jg3he4u"}, {"subreddit": "funny", "comment_id": "jg388km"}, {"subreddit": "funny", "comment_id": "jg3a4u3"}, {"subreddit": "funny", "comment_id":
"jg3cnjx"}, {"subreddit": "funny", "comment_id": "jg3xdc5"}, {"subreddit": "funny", "comment_id": "jg3f8ub"}, {"subreddit": "funny", "comment_id":
"jg3h5ps"}, {"subreddit": "funny", "comment_id": "jg3ksvj"}, {"subreddit": "funny", "comment_id": "jg39mkf"}, {"subreddit": "funny", "comment_id":
"jg3di2c"}, {"subreddit": "funny", "comment_id": "jg3i6uj"}, {"subreddit": "funny", "comment_id": "jg399h0"}, {"subreddit": "funny", "comment_id":
"jg3cinq"}, {"subreddit": "funny", "comment_id": "jg3cy8c"}, {"subreddit": "funny", "comment_id": "jg3f6z5"}, {"subreddit": "funny", "comment_id":
"jg3jm9x"}, {"subreddit": "funny", "comment_id": "jg3f2k2"}, {"subreddit": "funny", "comment_id": "jg55hjt"}, {"subreddit": "funny", "comment_id":
"jg3gsuf"}, {"subreddit": "funny", "comment_id": "jg3y5ii"}, {"subreddit": "funny", "comment_id": "jg3c9f0"}, {"subreddit": "funny", "comment_id":
```

Looking at the data

- 120,856 comments from top 100 subreddits
- 2,320,731 from top 2500
- Some comments are “[deleted]” or “[removed]”
  - Need to be filtered for text based analysis
  - Could also be interesting to leave in for other types of analysis
- A few subreddits that were in the top 2500 had no posts in the past year

Distribution of comment length





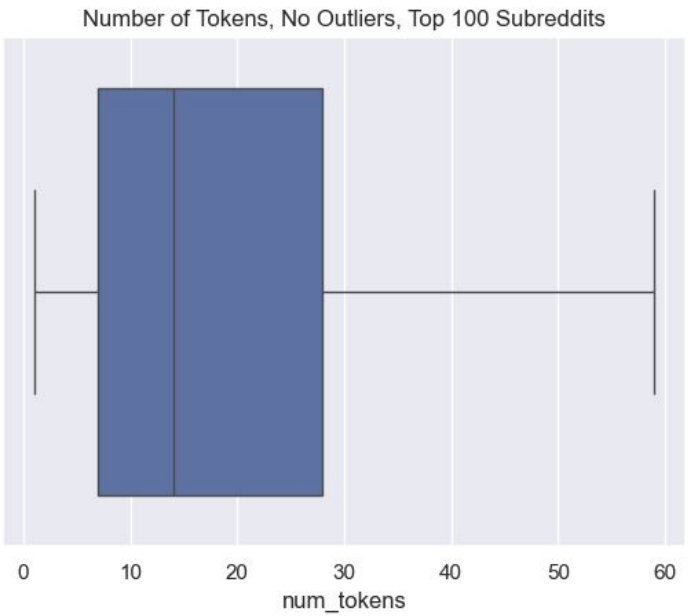
A strange outlier

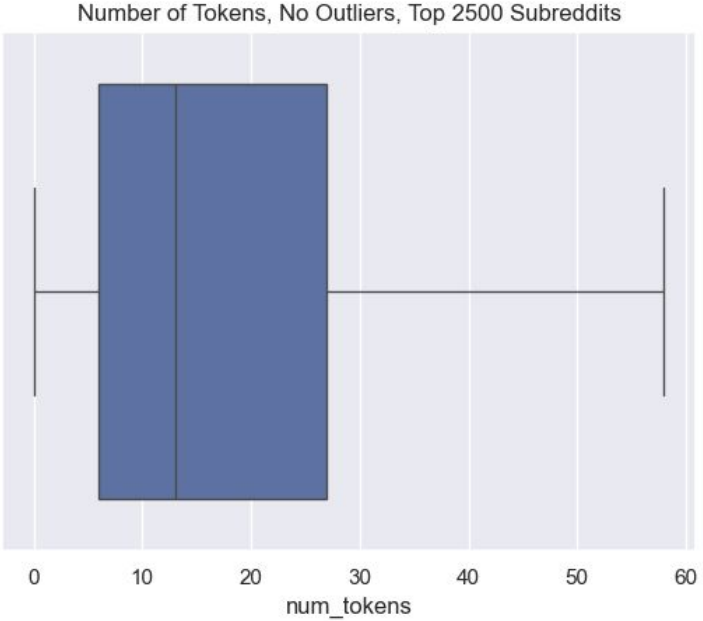
2^3 + 3^3 + 4^3 + 5^3 + 6^3 + 7^3 + 8^3 + 9^3 = 2^10 + 10^3

10^3 - 9^3 - 8^3 - 7^3 - 6^3 - 5^3 - 4^3 - 3^3 - 2^3 = -2^10

This text is in protest against reddit forcing its new user interface on mobile users regardless of whether they're opted out or not. They know it sucks and know users hate it and now they're forcing it on those users. If reddit wants to play silly games then so can us users. Each comment can be upto 10,000 characters in length and data costs money to store and serve. So, this is me doing my bit making reddit pay for its action. If we all adopt this measure, costs may start to add up for them. This signature uses "-" to pad out the comment to 10,000 chars which show up as a horizontal line.

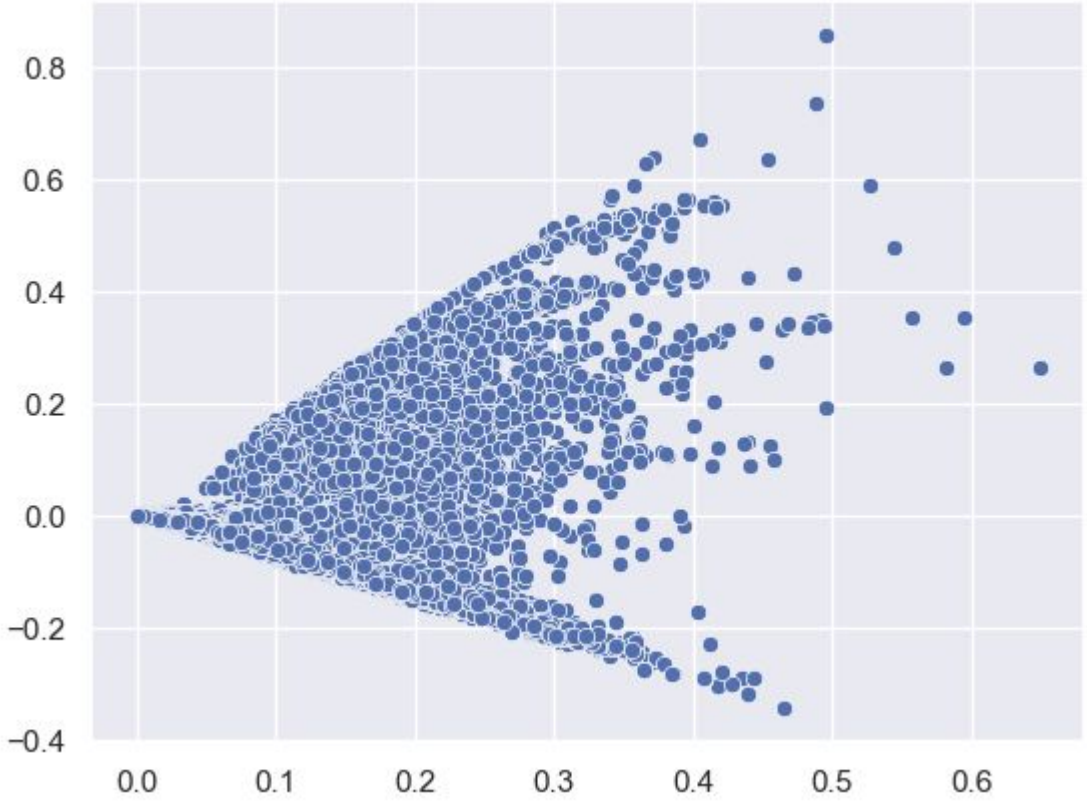
Distribution of comment length, outliers hidden





Clustering, attempt 1

SVD decomposition

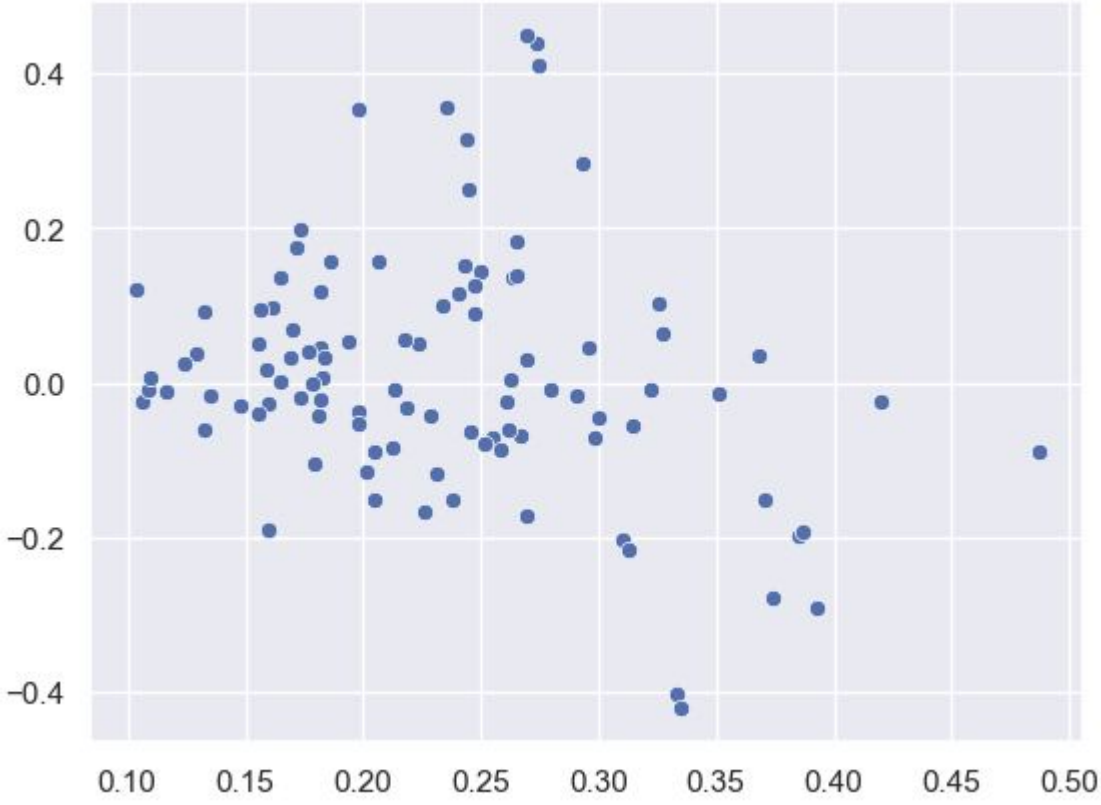


Clusters?

(see notebook)

Clustering, attempt 2

SVD decomposition



Clusters!

(see notebook)

Future steps

- Further analysis of clusters now that I have good ones
- Notebook clean up
- Try to reduce the size of the shared data a bit
- Biggest problem has been time management

Questions?