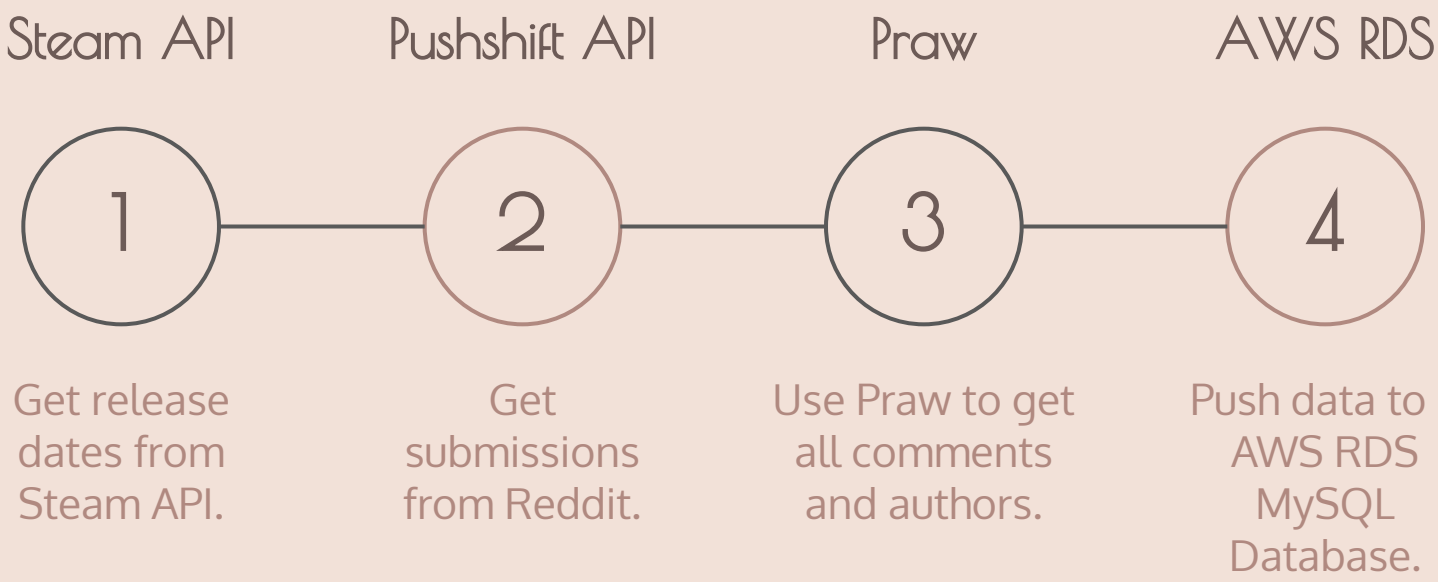
Reddit User Segmentation

GA DSIR Capstone
Samuel Park

Project Goal

Cluster redditors, specifically active authors of comments and submissions around the release date of video games. By clustering users, the end goal, is to gain insight into the structure and features of reddit users, their comments, and submissions for future modeling.

Data Collection



Data Cleaning & EDA

- Drop columns with too many null values
- Cleanup data types, especially booleans
- Aggregate comment and submission data by author
- Merge data into a final dataset
- Look at distribution of data

Model Selection

Agglomerative Clustering

Data is very hierarchical especially due to boolean features and very positively skewed distributions.

DBSCAN

Deals with outliers and odd cluster shapes.

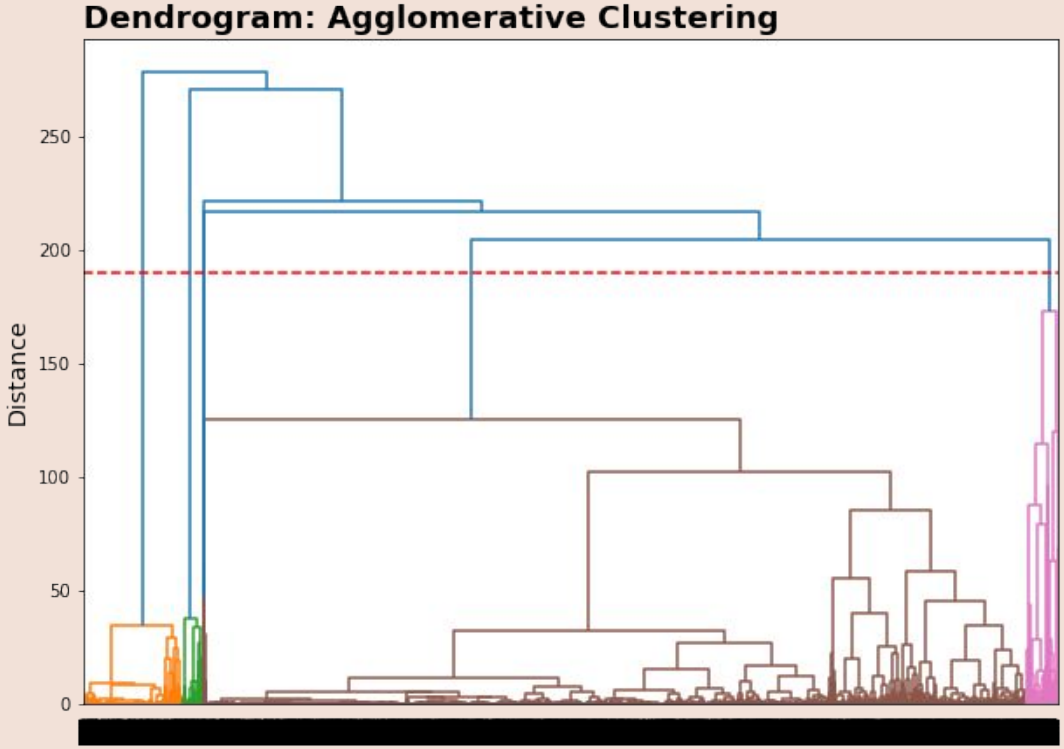
Silhouette Scores

Agglomerative Clustering

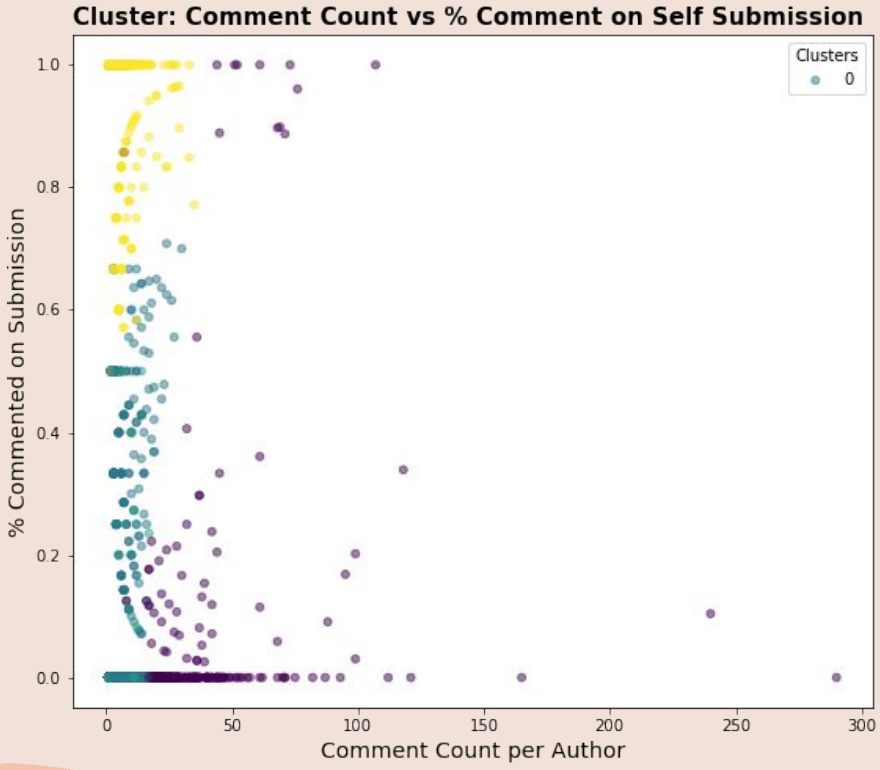
Hierarchical Clustering

DBSCAN

Density-Based Clustering


Dendrogram: Agglomerative Clustering
Distance
250
200
150
100
50
0

Clusters



- Number of comments posted by Redditor.
- Percentage of comments posted on self posted submission.

Clusters and Features

cluster	0	1	2	3	4	5
CommentKarma	317865.52	1314544.25	25660.37	19023.37	203750.60	16880.77
CreatedUTC	1336813520.37	1370720540.25	1349456156.14	1359887518.56	1312777316.30	1355127240.04
VerifiedEmail	0.96	1.00	1.00	0.00	0.90	0.91
IsMod	0.58	1.00	0.30	0.14	0.40	0.28
LinkKarma	66036.48	4343624.50	5622.83	1954.57	5517.00	10451.91
created_hour	12.18	13.25	11.99	11.75	8.80	11.79
name_length	10.49	9.75	10.08	10.74	9.80	10.17
comment_count	11.59	4.50	1.94	1.76	1.30	4.06
avg_comment_length	162.32	699.50	102.34	102.40	158.80	141.55
avg_edited_comments	3.17	0.00	2.47	2.57	12.70	2.72
avg_is_submitter	0.02	0.00	0.00	0.00	0.00	0.96
avg_comment_score	8.30	25.79	4.64	3.87	1960.75	3.06

Conclusion & Limitations

Conclusions

1. Using limited features silhouette score of 0.741
2. `LinkKarma`, `comment_count`, `avg_is_submitter`, and `avg_comment_score` features were most distinct per cluster
3. For more descriptive clusters, NLP is necessary

Limitations

- Model can only be used on Reddit data
- Limited analysis of clusters based on numeric features
- Selection bias for data

Next Steps

More Data

- Different Games
- Author posts and submissions

Too Much Data

- Move process to AWS

Subset Analysis

- Pre and post release date
- By game

NLP

- Self Focus Analysis
- Parts of Speech
- Sentiment Analysis

THANKS

Questions?
contact@tungstenandsons.com
tungstenandsons.com