Brendan Herger
Data Visualization
April 23, 2015

Project Dataset Submission

Reddit for the win!

Original location

This dataset contains information from http://www.reddit.com/r/askgaybros/, which is a safe for work subreddit. The code for compiling this data from Reddit's API, and an example dataset are available at https://github.com/bjherger/reddit_stats.

License information

Licensing for Reddit content is available here: http://www.reddit.com/wiki/licensing

Licensing for Reddit API is available here: https://github.com/reddit/reddit/wiki/API

Basically, as this is a non-commercial project, I just have to abide by the API rate limits.

Data types

Table 1

Variable	Type	Comment
author	String	
created	float (epoch)	
created_utc	float	
distinguished	string	
domain	string	only looking at self.askgaybros
downs	int	number of down votes
edited	bool	whether the post has been edited
gilded	int	number of reddit gold post has received
id	str	unique identifier
name	str	unique identifier, based on id
num_comments	int	number of comments
over_18	bool	whether post is age restricted (nsfw)

Table 1-1

permalink	str	link to post
report_reasons	str	
score	str	ups - downs
selftext	str	body text of post
stickied	bool	
subreddit	str	only looking at self.askgaybros
subreddit_id	str	
title	str	title for post
ups	int	number of upvotes
url	str	
user_reports	str	
visited	bool	

Why I like this dataset

As a member of the askgaybros community, I'm interested in exploring the community, and seeing what it is like. I'd like to see if I can discover sub-topics using topic modeling, and display how these subtopics differ. I would also like to see the number of posts over time, and what factors effect up votes and down votes.