
How to go viral in Reddit

- Your perfect guide
- powered by Natural Language Processing with Machine Learning

General Assembly

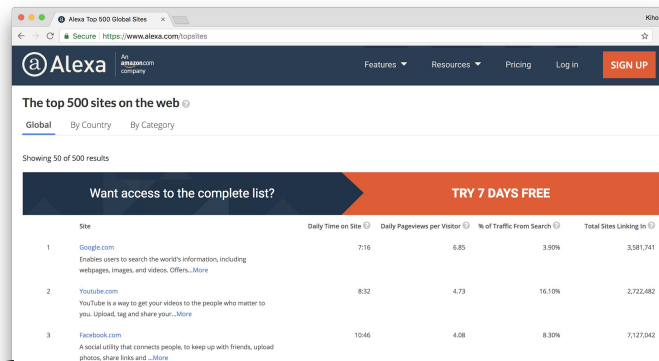
| Kihoon Sohn, Data Scientist

June 4, 2018

Before we dived in...

Reddit is the 6th most visited website in the world.

Source: [Alexa global topsites](#), [Wikipedia](#) (as of May 16, 2018)



Site	Daily Time on Site	Daily Pageviews per Visitor	% of Traffic From Search	Total Sites Linking In
1 Google.com Enables users to search the world's information, including webpages, images, and videos. Offers... More	7:16	6.85	3.90%	3,581,741
2 Youtube.com YouTube is a way to get your videos to the people who matter to you. Upload, tag and share your... More	8:32	4.73	16.10%	2,722,482
3 Facebook.com A social utility that connects people, to keep up with friends, upload photos, share links and... More	10:46	4.08	8.30%	7,127,042

Site	Daily Time on Site	Daily Pageviews per Visitor	% of Traffic From Search	Total Sites Linking In
6 Reddit.com User-generated news links. Votes promote stories to the front page.	15:08	9.71	16.10%	478,691

Reddit is powerful medium to spread words!



—

What kinds of aspects make a reddit post popular?

Are you in the right keywords/subreddit?

51k popular Reddit posts fetched, analyzed

- In 7 day period, total 105k realtime reddit posts fetched from `reddit.com/hot.json`, the most popular one, by using Reddit's public API. After cleaning the data, **51k unique posts remained to be analyzed.**
(fetched between May-25-2018 to Jun-1-2018)
- Out of 80+ features in the dataset, I selectively choose the following for my model in Data Science.

Machine Learning Model

Most used models/libraries

- CountVectorizer
- TfidfVectorizer
- Random Forest
- Logistic Regression

Features

4 core features selected

- Title
- Subreddit
- # of comments
- Post Age (minutes)
= Fetched time - Created time

Target

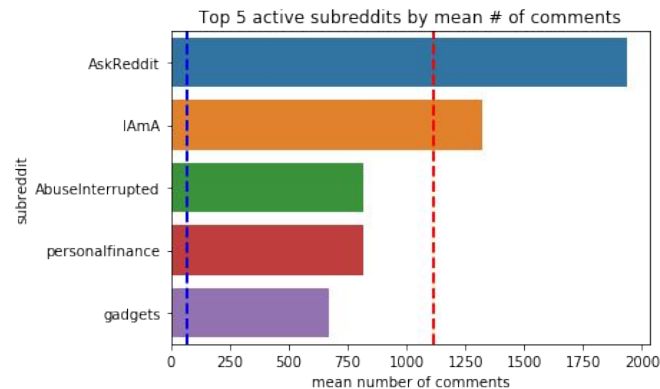
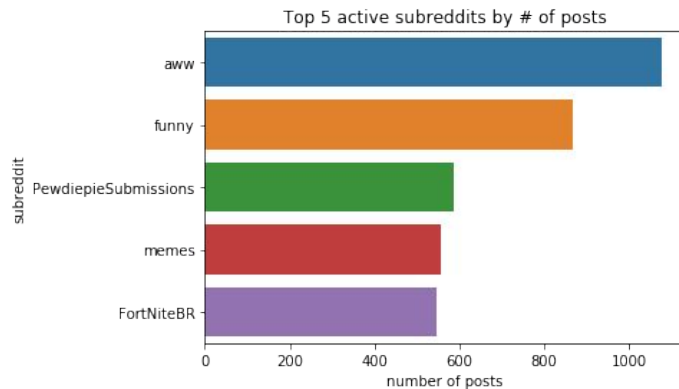
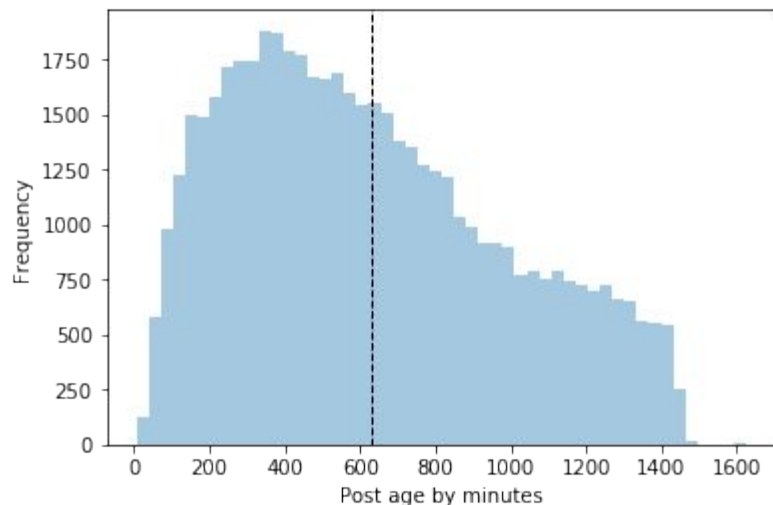
Binary classification in the number of comments.

- High / Low
Separator - 75th percentile

(50th - 16, 75th - 43, max-28,236)

Post age distribution & Active subreddits

- Hot posts stay in the list: 627 mins in average.
- # of comments in each post: 69 comments in avg.
- Top 5 subreddits by mean # of comments(right-bottom) gets 1,112 comments in average.



Final model scored .83

- With features `title`, `subreddit`, `post age`, by TfidfVectorizer, Logistic Regression.
- Predictive subreddits to get higher comments
 - `r/AskReddit`, `r/cars/`, `r/CFB`, `r/MemeEconomy`, `r/Drama`
- Top 10 appeared words in title
 - new, one, first, now, time, day, will, made, game, today

Conclusion

- Post your story at ***`r/AskReddit`, `r/cars`***
- Include, **`new`, `one`, `first`** words in your title.

→ **These will make your post more attractive.**

Next steps

for the catchy
fivethirtyeight article

- Build model with different features
 - e.g. ups, is_video, is_url, etc.
 - Compare subreddits to predict
 - `r/music` vs. `r/kpop`
 - `r/republican` vs `r/democrats`
 - Data Science tuning
 - Word Cloud
 - Lemmatize, Tokenize
 - ngram_range
-

—

Thank you