Luke Tibbott
Project Fletcher
November 4, 2018

What makes a shower thought good?

A shower thought is a witty, profound aphorism that one might come up with while in the shower. On reddit.com/r/showerthoughts, users submit their own shower thoughts and other users upvote or downvote based on their like or dislike of the shower thought. An example of a highly upvoted shower thought currently on the front page of /r/showerthoughts is “People who walk up two stairs at a time are both active and lazy.” I like this topic because it’s hard to quantify how good a shower thought is, but a good shower thought might be identified only through a “I know it when I see/hear it” methodology from a human’s point of view. Good shower thoughts have that *je ne sais quo* that makes you think “oh, wow! I never thought about it that way!” My goal is to distill this indefinable quality from shower thoughts.

My plan is to use natural language processing (NLP) to determine what makes a post to reddit.com/r/showerthoughts successful or unsuccessful. A metric for success has yet to be defined, but I have two thoughts: use linear regression and make success a continuous variable (the number of upvotes a post will receive), or define posts above the median number of upvotes as successful and posts below the median number of upvotes as unsuccessful and turn this in to a classification problem. I’m leaning towards the classification interpretation of this problem at the moment. My belief is that the number of upvotes a post receives on reddit is not deterministic, and rather random. There are numerous examples of identical posts getting wildly different amounts of upvotes. I think linear regression would therefore be a poor predictor of upvotes.

I think it’s also important to consider another determinant of a post’s success: the time of day it’s posted. I’ll be collecting this variable and including it in my analysis, along with features created via unsupervised learning (most likely clustering) in conjunction with NLP.

The data

I don’t know *exactly* where I’m getting my data, but I have several options.

The reddit API only allows one to query the most recent 1000 posts in any given subreddit. I discovered this fact when I tried to use the Reddit API Python wrapper (PRAW) to collect my data and was unable to get more than 1000 posts. I would prefer more than 1000 posts -- ideally 10,000, and hopefully closer to 50,000 (if showerthoughts even has 50,000 posts in it). The reddit API will likely not be of much use to me.

I’m currently investigating using pushshift.io, a more robust version of reddit’s API. I’ll likely use Google’s BigQuery tool to extract only the features I’m interested in.

Another option is to use a 269 GB data dump of all reddit posts from 2006-2015 that’s available on Kaggle. I could use Google BigQuery in conjunction with Amazon Web Services to query this giant dataset to select only the subreddit I want. I’d probably have to change to a different subreddit if I go this route, as /r/showerthoughts was only created in 2012 and would have only three years worth of posts. Alternative subreddits include askreddit, askscience, and legaladvice. Showerthoughts is my first choice, but any popular text-based subreddit (rather than image-based) will do.

Technical challenges

Getting the data will be my first major challenge. Aside from that, it’s very possible that the phrasing of a submission only has a meager impact on a post’s success. It’s possible that my model won’t have strong predictive power, but that wouldn’t be the end of the world.

Projects similar to this one have been done before -- predicting the success of social media posts is an extremely important business question -- but they often use Keras and Tensorflow. I am currently unfamiliar with these tools. If I discover early on in this project that neural networks are necessary tools for this problem, I might use this project idea for project 5, but I don’t anticipate needing to entirely scrap this idea.