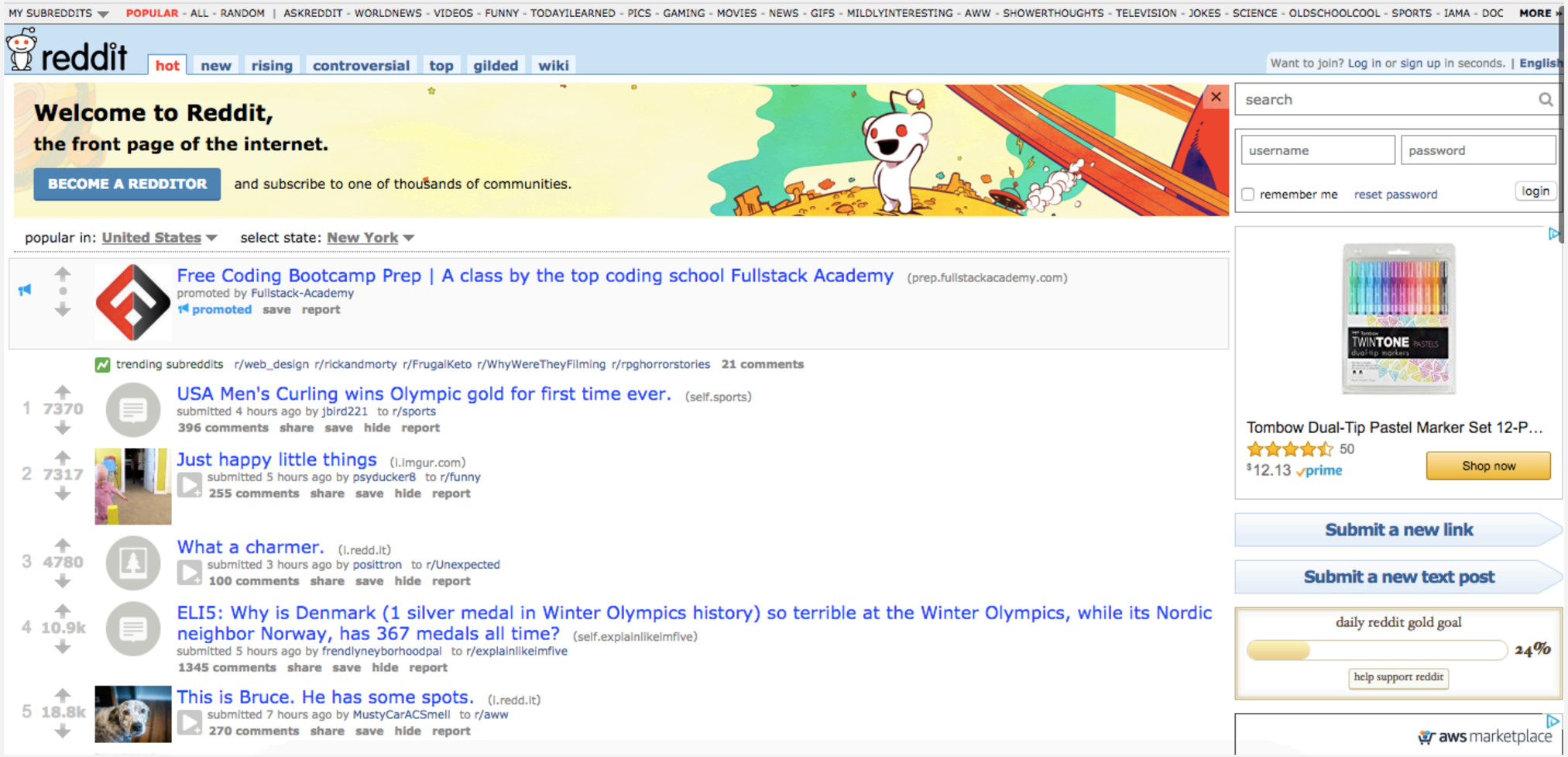
GENERAL ASSEMBLY

Predicting Reddit Comments using Web Scraping and NLP

Marco Tavora

Reddit


MY SUBREDDITS POPULAR - ALL - RANDOM | ASKREDDIT - WORLDNEWS - VIDEOS - FUNNY - TODAYILEARNED - PICS - GAMING - MOVIES - NEWS - GIFS - MILDLYINTERESTING - AWW - SHOWERTHOUGHTS - TELEVISION - JOKES - SCIENCE - OLDSCHOOLCOOL - SPORTS - IAMA - DOC MORE
reddit
hot new rising controversial top gilded wiki
Want to join? Log in or sign up in seconds. | English
Welcome to Reddit,
the front page of the internet.
BECOME A REDDITOR
and subscribe to one of thousands of communities.
search
username
password
remember me
reset password
login
popular in: United States
select state: New York
Free Coding Bootcamp Prep | A class by the top coding school Fullstack Academy (prep.fullstackacademy.com)
promoted by Fullstack-Academy
promoted save report
trending subreddits r/web_design r/rickandmorty r/FrugalKeto r/WhyWereTheyFilming r/rpghorrorstories 21 comments
1 7370
USA Men's Curling wins Olympic gold for first time ever. (self.sports)
submitted 4 hours ago by jbird221 to r/sports
396 comments share save hide report
2 7317
Just happy little things (i.imgur.com)
submitted 5 hours ago by psyducker8 to r/funny
255 comments share save hide report
3 4780
What a charmer. (i.redd.it)
submitted 3 hours ago by posittron to r/Unexpected
100 comments share save hide report
4 10.9k
ELI5: Why is Denmark (1 silver medal in Winter Olympics history) so terrible at the Winter Olympics, while its Nordic neighbor Norway, has 367 medals all time? (self.explainlikeimfive)
submitted 5 hours ago by frendlyneyborhoodpal to r/explainlikeimfive
1345 comments share save hide report
5 18.8k
This is Bruce. He has some spots. (i.redd.it)
submitted 7 hours ago by MustyCarACSmell to r/aww
270 comments share save hide report
Tombow Dual-Tip Pastel Marker Set 12-P...
50
$12.13 prime
Shop now
Submit a new link
Submit a new text post
daily reddit gold goal
24%
help support reddit
aws marketplace

Statement of the problem

Find out what characteristics of a post on Reddit contribute most to the overall interaction as measured by number of comments?

Post characteristics → Number of comments

Statement of the problem

Find out what characteristics of a post on Reddit contribute most to the overall interaction as measured by number of comments?

Post characteristics → Number of comments

A few lines from the scraped dataset

	titles	times	subreddits	nums
0	Remodeling my Grandfather's basement and found...	3 hours ago	r/funny	386 comments
1	me_irl	3 hours ago	r/meirl	197 comments
2	Tree's shadow	2 hours ago	r/oddlysatisfying	112 comments
3	Beautiful...	3 hours ago	r/yesyesyesyesno	24 comments
4	While we're on the subject of Home Owner's Ass...	4 hours ago	r/ProRevenge	369 comments

Preprocessing and EDA

Checking for null values, removing duplicates (in the titles and subreddits columns), removing strings such as "hours ago" and "comments" and converting those cells to integers, and so on.

Analysis of the data

Start simple

We build a model using only subreddits as predictors as shown on the right

But before doing anything, what is our baseline accuracy?

For any classification model, the baseline accuracy must be calculated. The baseline accuracy in this case is the accuracy we would get if we always predict that the number of comments is larger than the median:

	subreddits	binary
25	funny	1
28	shittyrobots	0
29	malelivingspace	1
30	todayilearned	1
31	worldnews	1

$$\text{baseline} = \frac{n_{\text{above median}}}{n_{\text{total}}} \approx 0.517$$

Our model must perform better!

Dummies

We cannot build a mathematical model using the subreddits column as it is. We must convert this column into dummy columns which is straighforward to do in Python.

```
df = pd.concat([df,pd.get_dummies(df['subreddits'])], axis
    1).drop('subreddits',axis=1).drop('4PanelCringe',axis=1)
```

	binary	ATBGE	AbandonedPorn	AccidentalRacism	AccidentalRenaissance	Android	Art	AskReddit	BeAmazed
25	1	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
29	1	0	0	0	0	0	0	0	0
30	1	0	0	0	0	0	0	0	0
31	1	0	0	0	0	0	0	0	0

Models

Random Forest Classifier using only subreddits as predictors

- We used a Random Forest Classifier to model our data. The random forest score obtained is approximately 0.47 which quite poor, **lower than the baseline**.

- Adding a few extra features consisting of names appearing in the threads (am "dummifying" them as before) almost did not change the score chosen (the score value didn't change, only the variance decreased slighly), probably because:

  - The words could have been better chosen
  - The dummy columns corresponding to the chosen words were heavily unbalanced. Balancing them properly could increase the score (to be done yet).

Models (continued)

Random Forest Classifier using subreddits and titles as predictors

- To convert the titles column into predictors that allow for algebraic manipulations we preprocess the column data using Countvectorizer for the SciKitLearn library. What this tool does is to count the ocurrences of each word in the text. After countvectorizing, the titles column is converted into the following:

Android	Art	AskReddit	BeAmazed	BetterEveryLoop	...	yesterday	york	yosemite	younger	yvette	zedog	zelda	z
0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	...	0	0	0	0	0	0	0	0
0	0	0	0	0	...	0	0	0	0	0	0	0	0

Models (continued)

Random Forest Classifier using subreddits and titles as predictors

- The score for the RFC was very close to 1.0
- ScikitLearn has a very convenient functionality which tells us the most important features after the RFC fitted the data and tested it

	Features	Importance Score
933	times	0.055592
400	kidding	0.022022
142	choice	0.020371
322	happened	0.017418
822	subreddits_FellowKids	0.016534

Using now a logistic regression again using subreddits and titles as predictors:

- The score is roughly 0.85 which is slighly worse