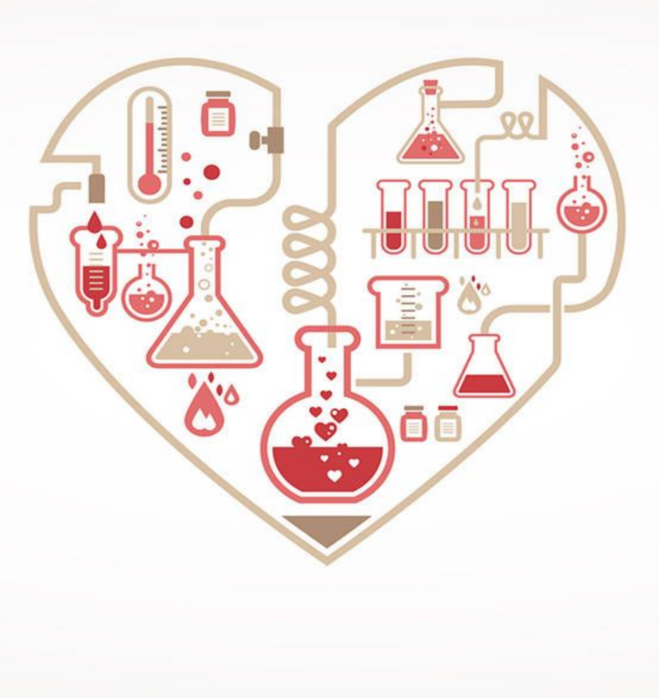
Web scraping & Classification

By Rachel Koenig

Problem Statement



- As a data scientist for the advertising department at reddit, I need to find the most predictive keywords and/or phrases to classify the the dating advice and relationship advice subreddit pages.
- Logistic Regression & Bayes models
- Measure success on accuracy score

r/dating_advice

Community details:

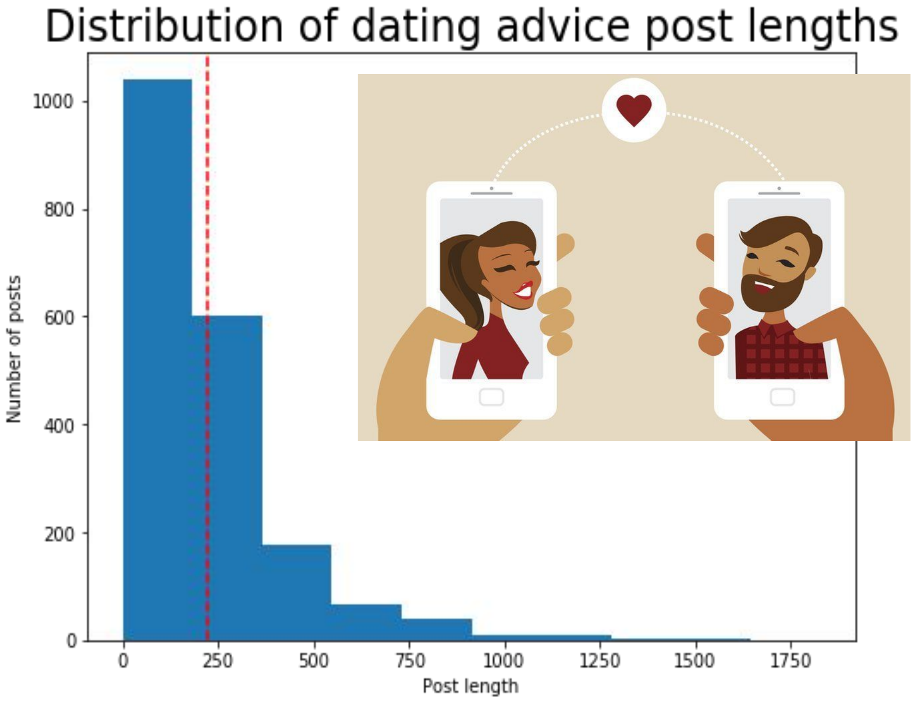
857K subscribers

A page for users to share tips and encouragements about dating or the reddit univers for advice.

222 - average words per post

18 - average upvotes per post

9 - average # of comments


Distribution of dating advice post lengths

r/relationship_advice

Community details:

1.7 million members

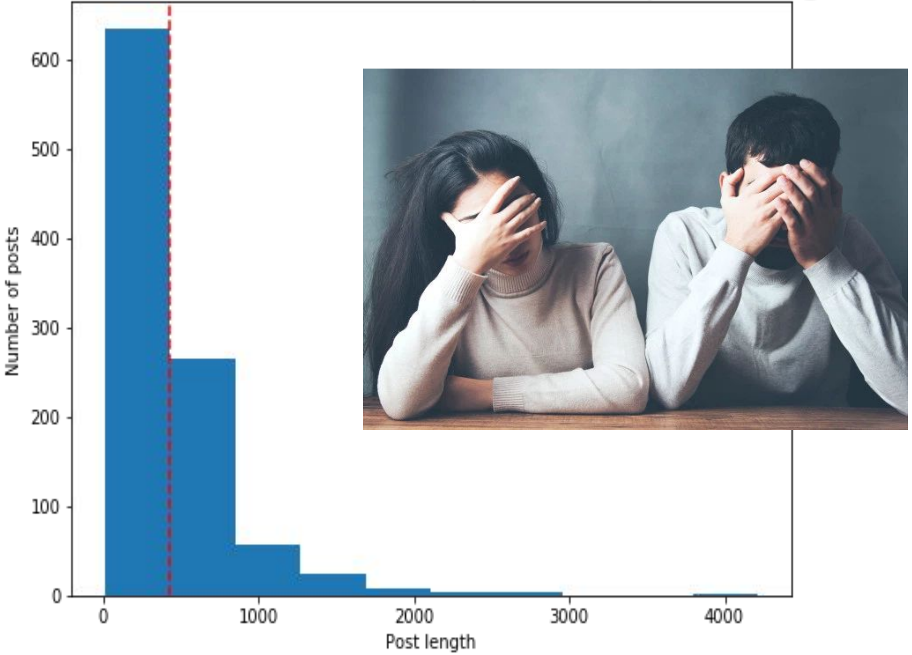
A page for users to get help with your relationships from romantic to friendship, family, co-workers,

425 - average words per post

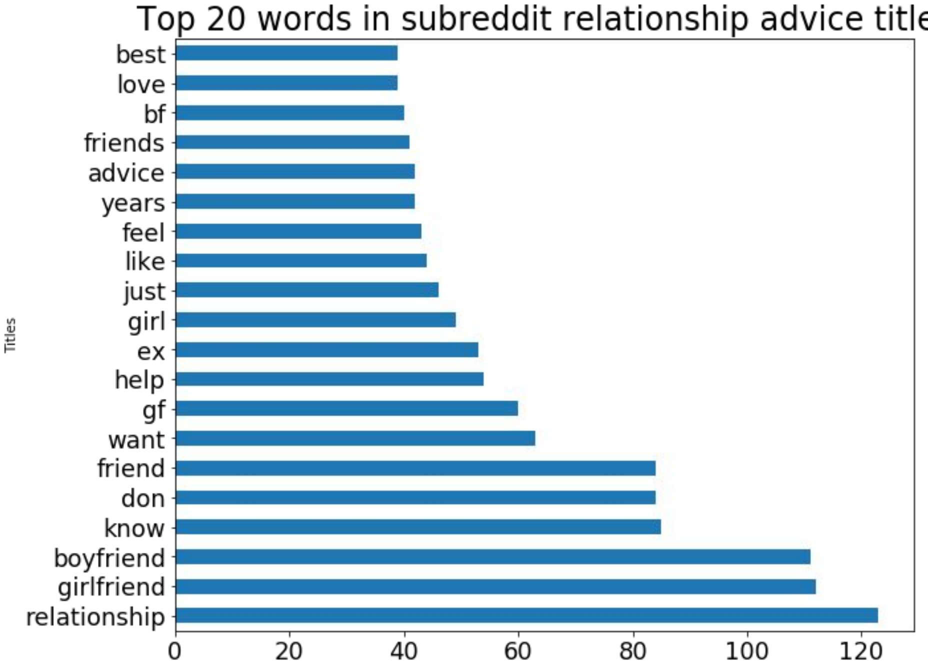
15 - average upvotes per post

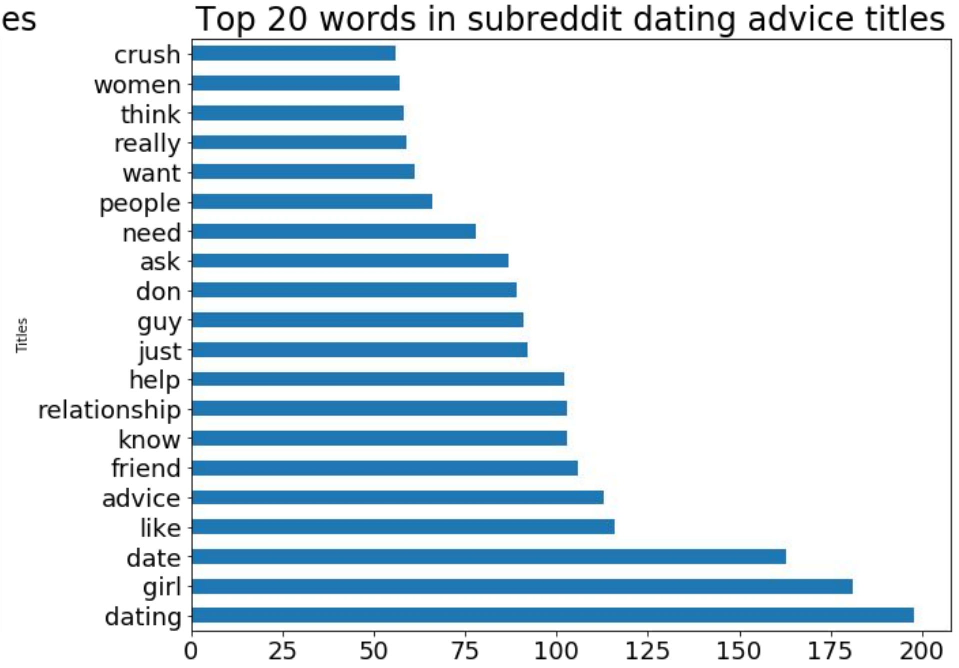
11 - average # of comments

Distribution of relationship advice post lengths



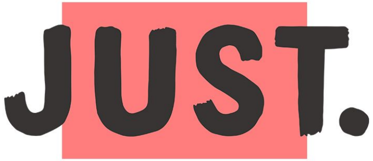
Pre-NLP Comparisons

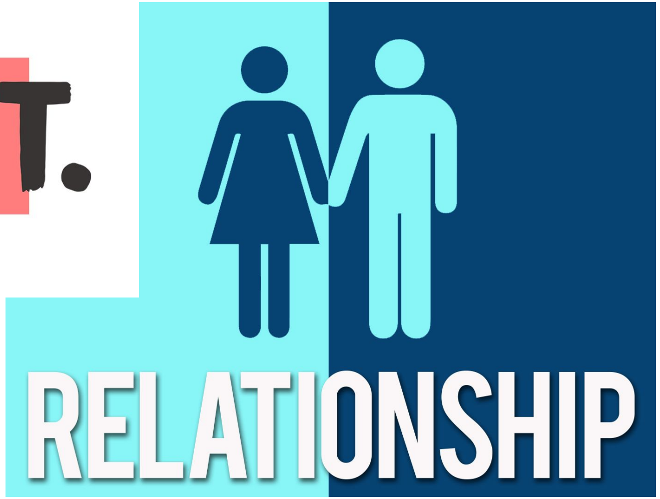

Top 20 words in subreddit relationship advice titles

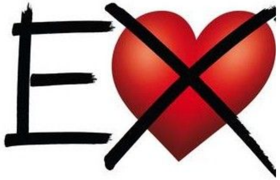

Top 20 words in subreddit dating advice titles

NLP Steps

- Concatenated dataframes
- title + selftext = new alltext col
- Binarized “subreddit” column
- Set X = alltext & y = subreddit
- Used a function to
  a. Clean html
  b. Remove non-letters
  c. Find & remove stop words
  d. Join everything back together as a giant string.







Modeling Process

- CountVectorizer
- Logistic Regression
- Pipeline & Gridsearch

= test scores increased each time but were still overfit

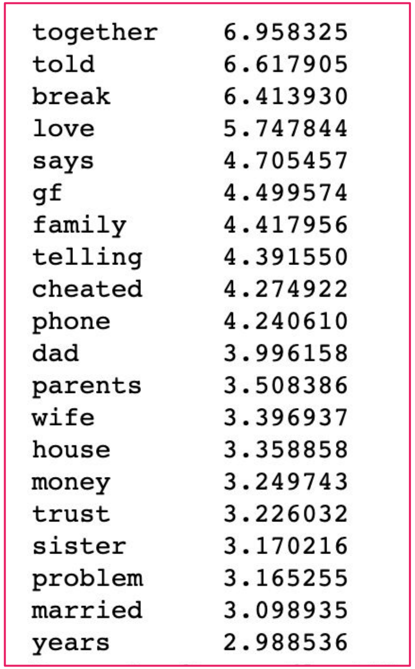
- K Nearest Neighbors = performed worse than baseline model
- Bayes Multinomial = lower variance

- TfidfVectorizer = lowered train score and increased test score for lower variance

Evaluation & Conclusions

Baseline score: 63.3%

Final train score: 87.6% | test: 76% | cross val: 79%

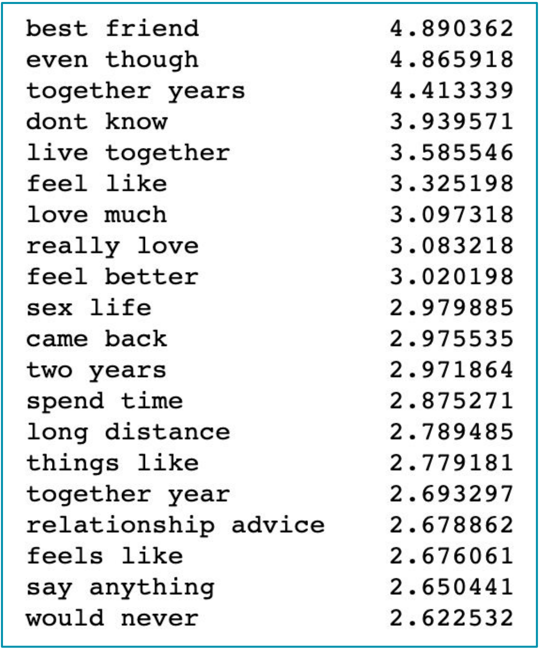


```
together     6.958325
told         6.617905
break        6.413930
love         5.747844
says         4.705457
gf           4.499574
family       4.417956
telling      4.391550
cheated      4.274922
phone        4.240610
dad          3.996158
parents      3.508386
wife         3.396937
house        3.358858
money        3.249743
trust        3.226032
sister       3.170216
problem      3.165255
married      3.098935
years        2.988536
```

Together is 6.9 times as likely to predict the relationship advice page.

Best friend is 4.9 times as likely

Even though I would like to have a higher test score, I was able to successfully lower the variance so I think the model is ready to launch a test. If advertising engagement increases, the same key words could be used to find other potentially lucrative pages.



```
best friend                 4.890362
even though                 4.865918
together years              4.413339
dont know                   3.939571
live together               3.585546
feel like                   3.325198
love much                   3.097318
really love                 3.083218
feel better                 3.020198
sex life                    2.979885
came back                   2.975535
two years                   2.971864
spend time                  2.875271
long distance               2.789485
things like                 2.779181
together year               2.693297
relationship advice         2.678862
feels like                  2.676061
say anything                2.650441
would never                 2.622532
```

Sources

Page 2: image http://metropolismanagement.com/the-science-of-love/

Page 3: image http://www.studentprintz.com/online-dating-is-ruining-romance-heres-why/

Page 4: image https://nypost.com/2018/11/19/heres-whats-keeping-you-in-your-miserable-relationship/

Page 6: images https://puropeople.com.au/just-when-such-a-small-word-speaks-volumes/, https://www.stitcher.com/podcast/i-do-podcast, https://www.earlyyearscount.earlychildhood.qld.gov.au/, https://www.theatlantic.com/entertainment/archive/2016/11/the-evolution-of-like/507614/, https://pandagossips.com/posts/925