Subreddit Differentiation

Nick Calow

Overview

- Two subreddits, one machine.
- Can a model be trained to adequately differentiate the two?

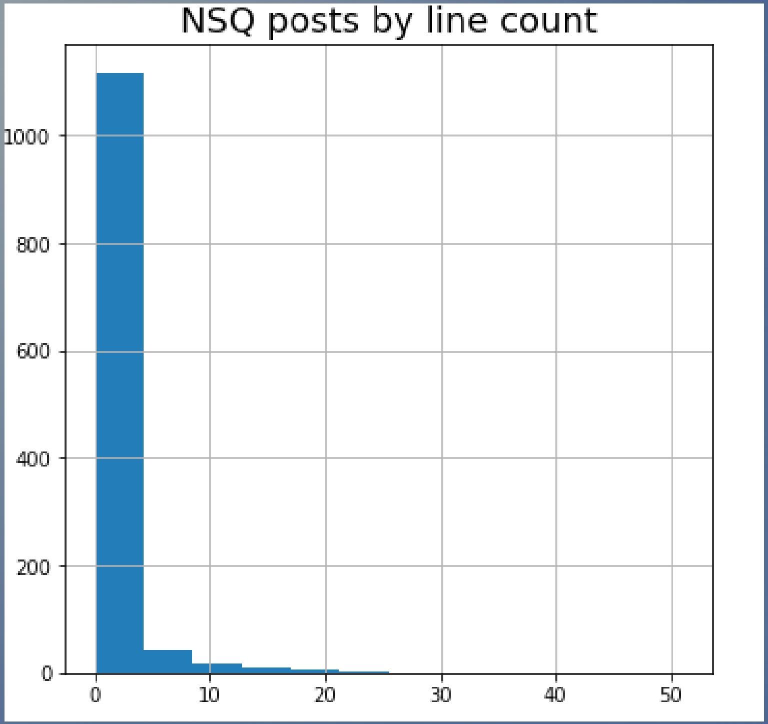
TIFU

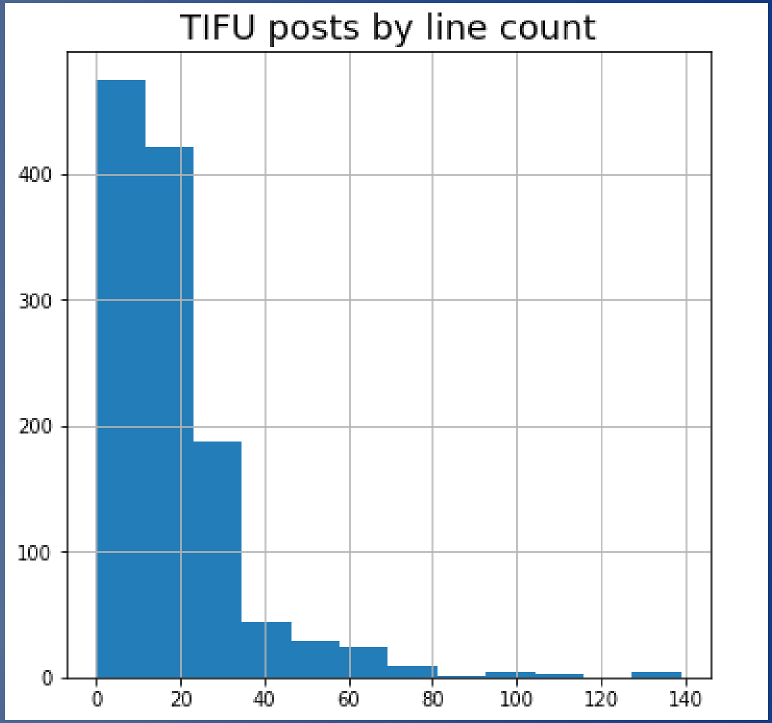
- Today I Fucked up.
- Every post must start with TIFU
- Every post must have at least 750 characters of content
- Must be a story that contains a "fuckup", not a death/injury

NoStupidQuestions

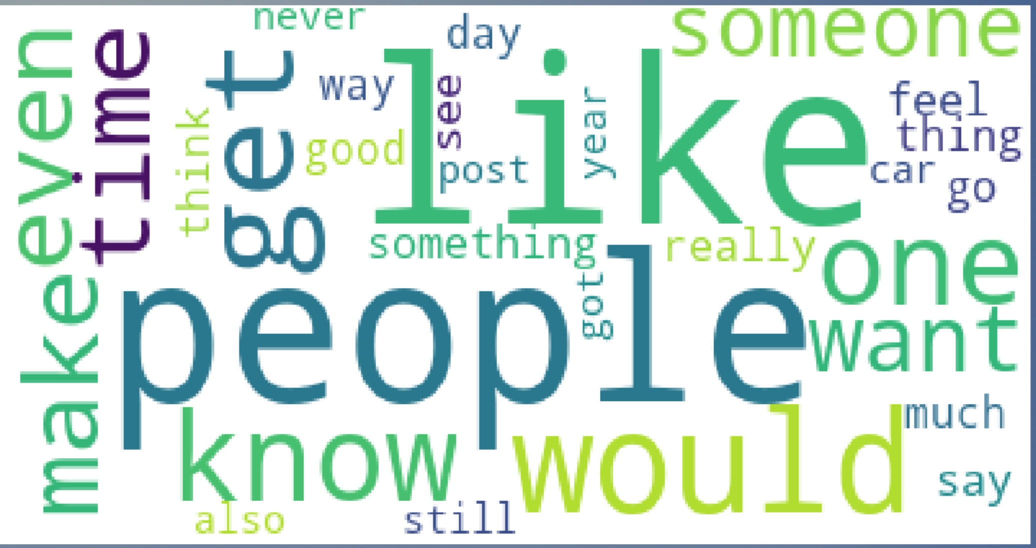
- No Stupid Questions.
- No repeat questions
- No joke questions
- No questions relating to suicide or sexuality(was I raped/harassed? etc.)
- No self-promotion
- No illegal/unethical/disturbing material

EDA

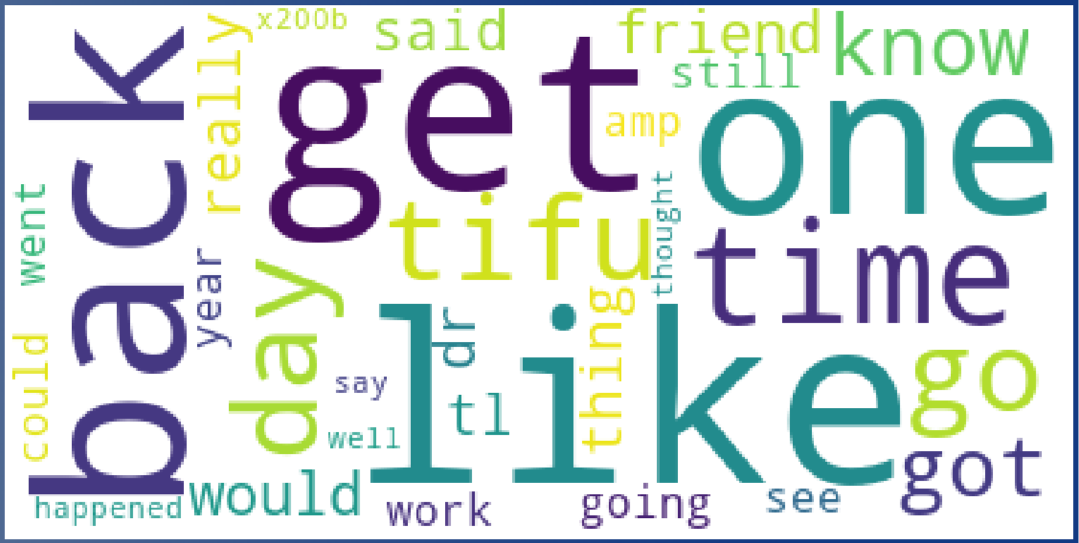

NSQ posts by line count


TIFU posts by line count

- Line Count
- Post Length
- Content



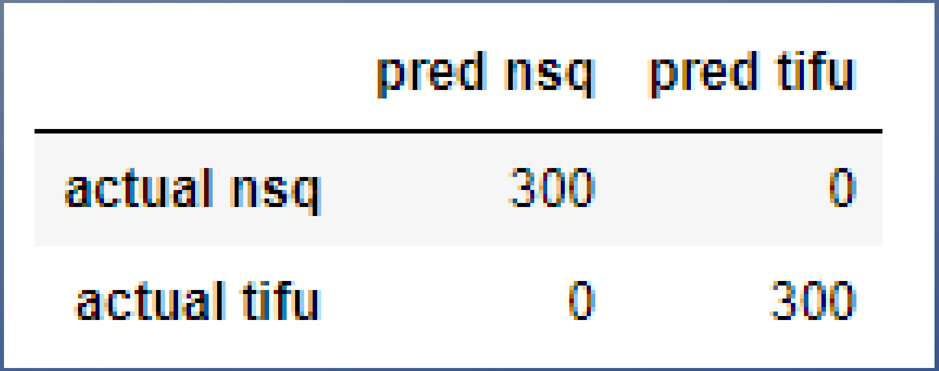
NoStupidQuestions



TIFU

- Wordclouds are great
- Limitation: Like
  - TIFU: 1989
  - NSQ: 390

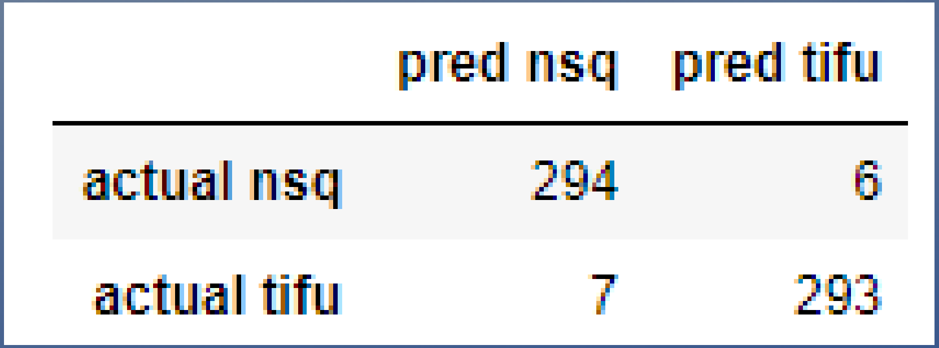
Model 1: DecisionTreeClassifier



	pred nsq	pred tifu
actual nsq	300	0
actual tifu	0	300

- Turns out, having TIFU at the beginning of every post makes it too easy
- 100% Accuracy

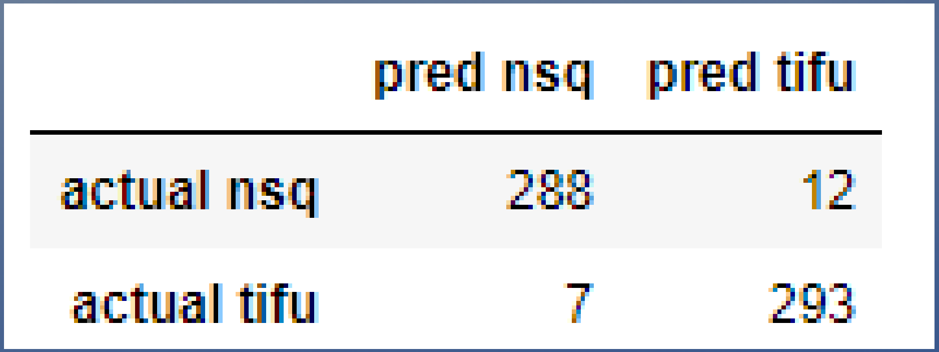
Model 2: DecisionTreeClassifier(2)



	pred nsq	pred tifu
actual nsq	294	6
actual tifu	7	293

- I took out all version of TIFU and ran it again
- Tuned my parameters more closely, still not bad
- 97.8% Accuracy

Model 3: Bernoulli Bayesian



	pred nsq	pred tifu
actual nsq	288	12
actual tifu	7	293

- Decided to try another model to compare
- Designed for discrete features
- Probabilistic instead of tree-like
- 96.8% Accuracy

Conclusions

- The indicator really helped. Too much
- Even without it, model was still strong. Very different kinds of subreddits
- Future rooms for improvement:
  - More involved models
  - More posts
  - Comments