The Art & Science of Reddit

An Exploration of Words and Phrases

by Teng Mao, Data Science Fellow at General Assembly

What is the name of Reddit's mascot?

Snoo!

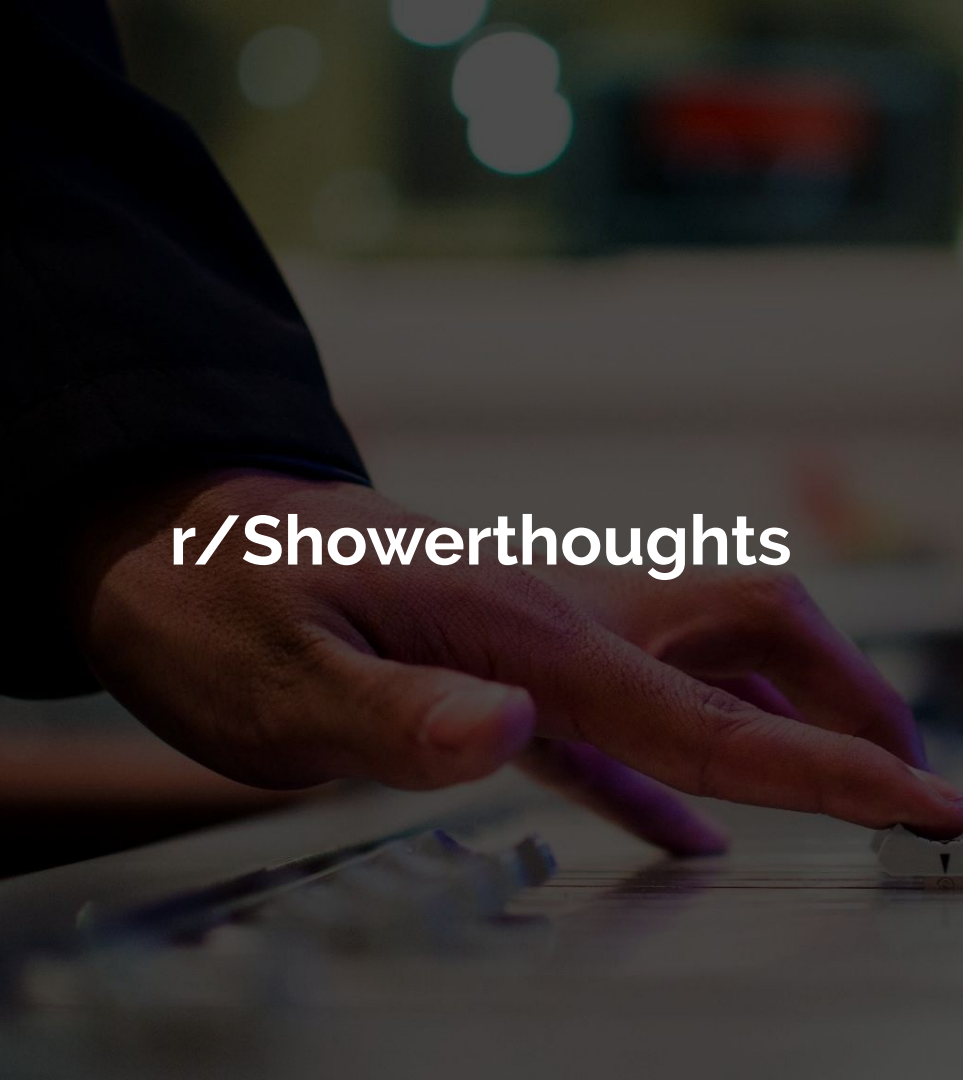
Reddit: the front page of the internet - Agenda

- State the Problem
- Goal of the Project
- Steps to Build a Model
- Exploratory Data Analysis
- Optimal Pipeline Used
- Top Words
- Conclusion
- Questions

r/AskReddit

What is the problem Reddit is trying to solve?

Given a random post, to which subreddit should the post belong?



“It’s more of an art than science…”

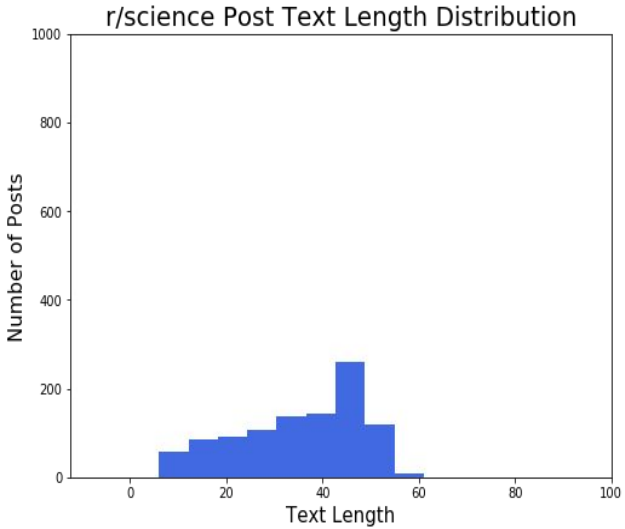
What are the keywords that distinguish between the two subreddits - r/science and r/Art?

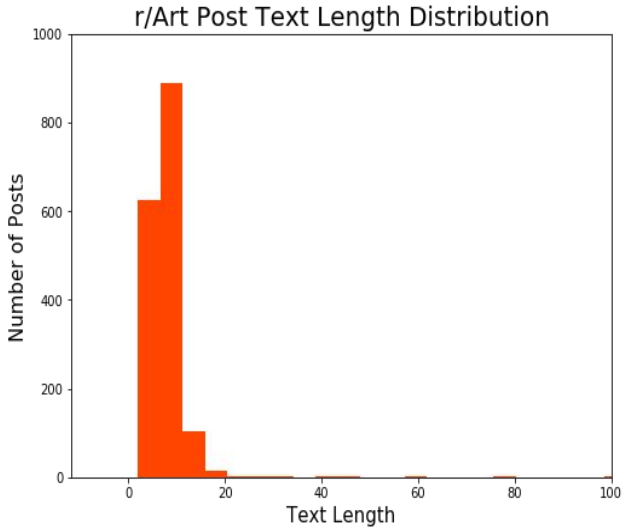
r/DIY - Steps to Build a Model

Below is a summary of steps to create a model that categorizes posts to a subreddit:

1. Gather posts from target subreddits and combine into one dataset
2. Clean the data and do preliminary EDA
3. Utilize stemming, lemmatization, regex, and/or stop words to create clean tokens
4. Choose transformers and models to use in a pipeline
5. Gridsearch for optimal parameters to fit the pipeline
6. Check the accuracy scores of the models
7. Explore the features, create a Confusion Matrix, calculate error metrics and make adjustments to clean tokens if necessary

r/mildlyinteresting - EDA





- 62% of the data coming from r/Art (1,649 posts)
- 38% coming from r/science (1,017 posts)
- r/science averaged of 36 words per post
- r/Art averaged of 8 words per post

r/therewasanattempt - Optimal Pipeline Used

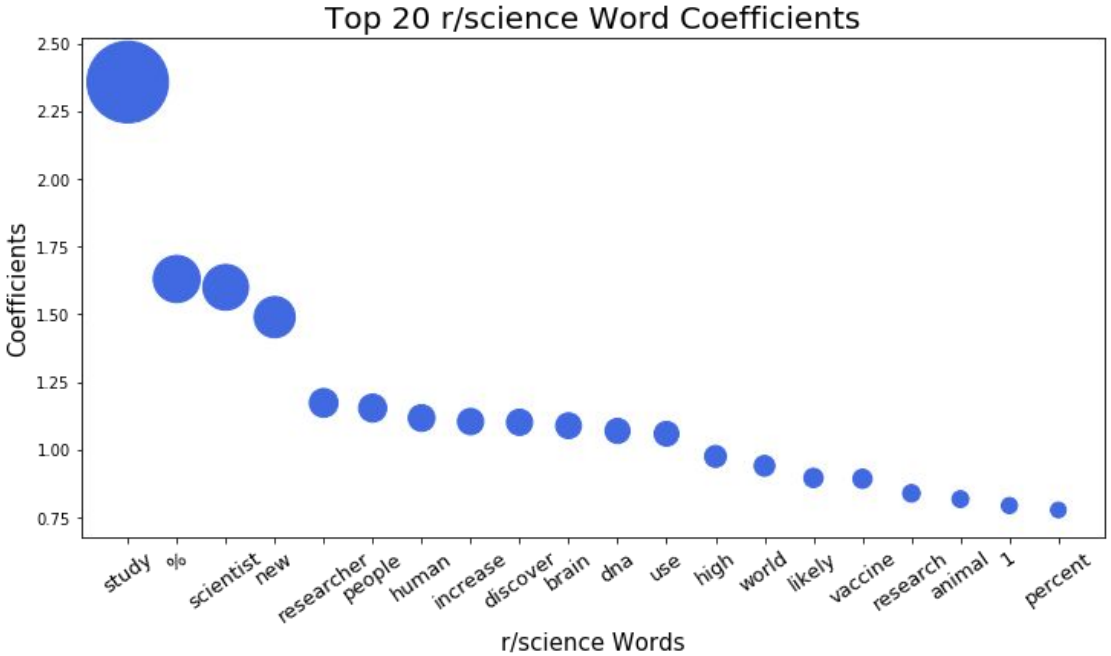
Optimal Pipeline:

- CountVectorizer transformer
- Multinomial Naive Bayes model
- Accuracy score of 99.6% on the train data and 99.4% on the test data
- This exceeds the baseline accuracy of 61.9%.

Confusion Matrix	Predicted r/Art	Predicted r/Science
Actual r/Art	409	4
Actual r/science	0	254

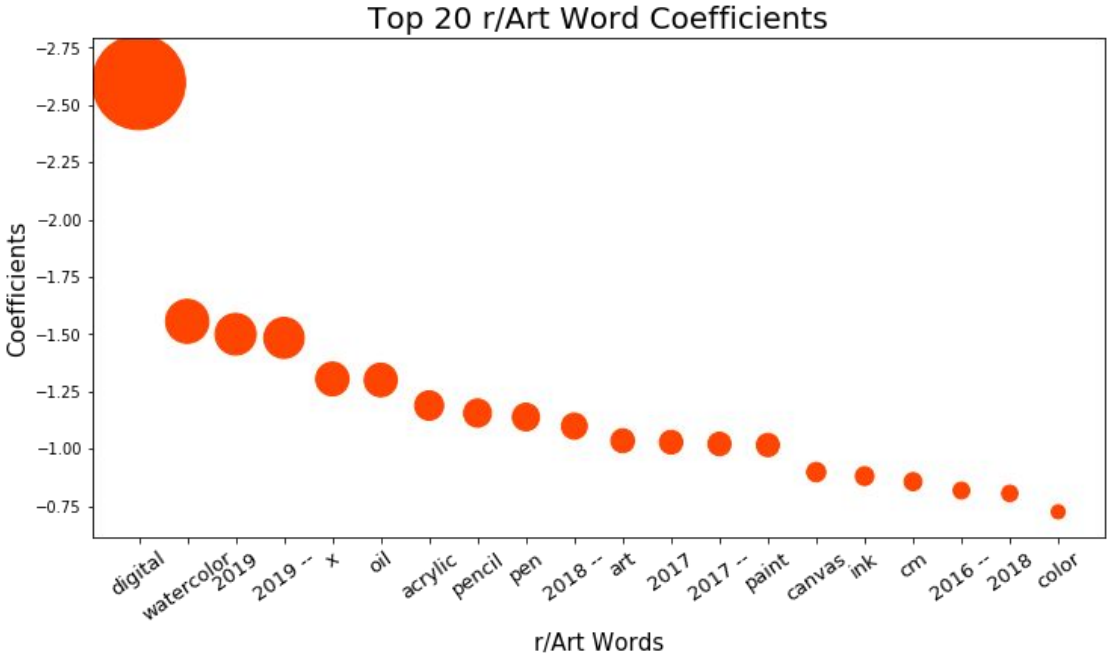
- No type II errors (false negatives)
- 4 type I errors (false positives)
- Sensitivity/Recall score was 100%
- 4 of type I errors generated a Precision score of 98%.

r/science - Top Words



Based on the coefficient of 2.4 for "study", the appearance of that word in a post would indicate an 91.4% chance that the post is from r/science.

r/Art - Top Words



Based on the coefficient of -2.6 for "digital", the appearance of that word in a post would indicate a 93.1% chance that the post is from r/Art.

r/todayilearned - Conclusion

Follow the 7 steps to build a model for post classification to a subreddit

Next Steps:

- Utilize regex to clean the tokens more thoroughly
- Investigate other models such as knn, decision trees, random forests, SVM, etc.
- Perform the same analysis on the comments section of the subreddit
- Explore the 4 posts that caused type I errors

Final thought: if someone says "it's more of an art than science", paint them a picture!

r/NoStupidQuestions

Any questions?