Project 3 - Subreddit Identification

Matthew DeVay, ATX-DSI-11

Problem:

1. Build a classification model to determine which of two subreddits a given post had been submitted.
2. My friend who works in machine learning doesn’t think of machine learning as data science.
3. I think of them as inextricably linked.
4. Prove it!

Process:

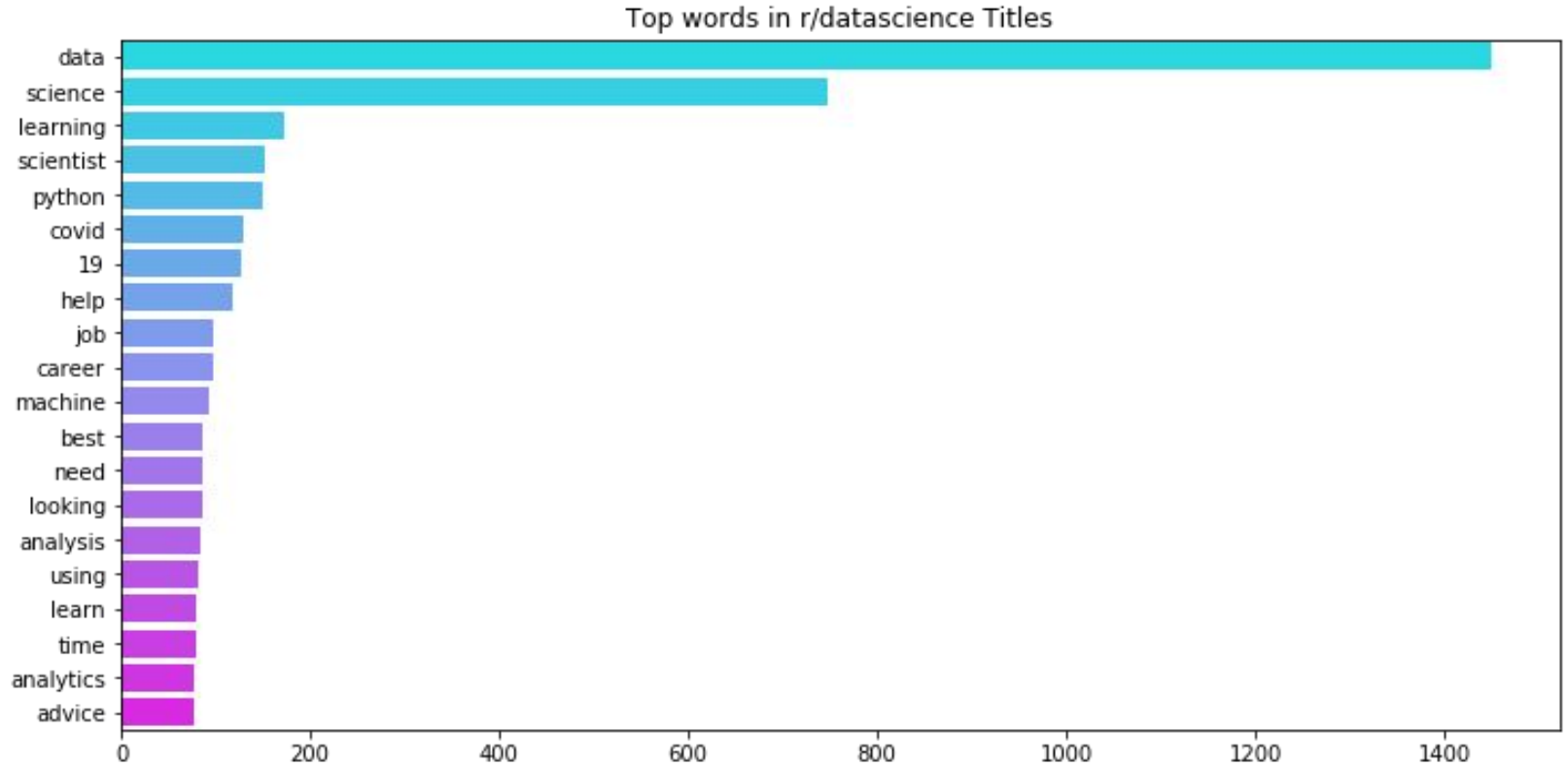
1. Collect Data from Reddit subs r/learnmachinelearning and r/datascience
2. Build a Naive Bayes model to classify a given submission as either r/machinelearning or r/datascience
3. Build a second classification model for the same purpose, other than Naive Bayes
4. Refine the models if possible
5. Prove friend wrong

Data:

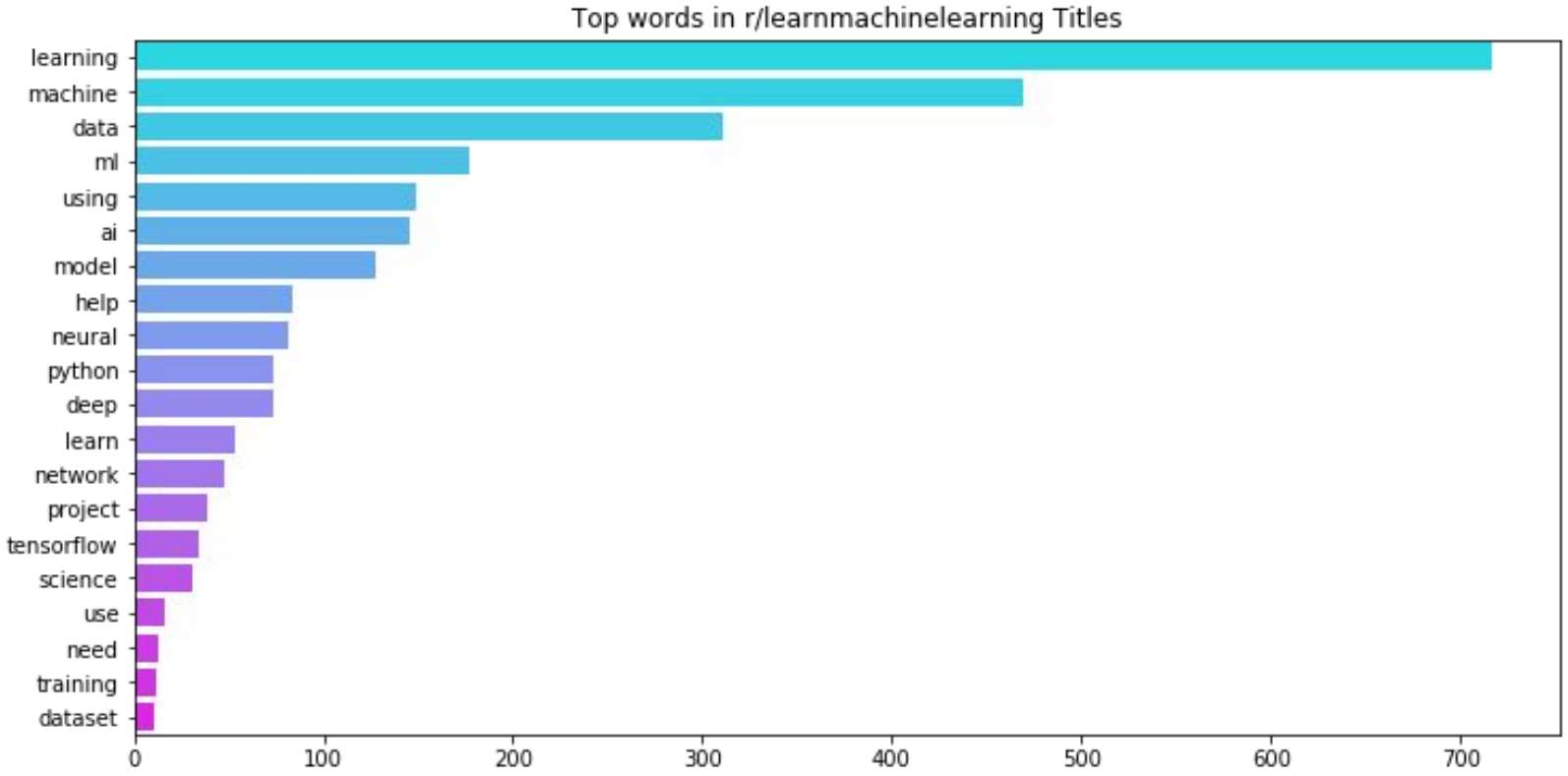
Reddit submissions to r/learnmachinelearning and r/datascience:

- 3,000 from each
- mod/user removed posts rejected
- Crossposts rejected
- Only selfposts
- Data very clean due to control of import

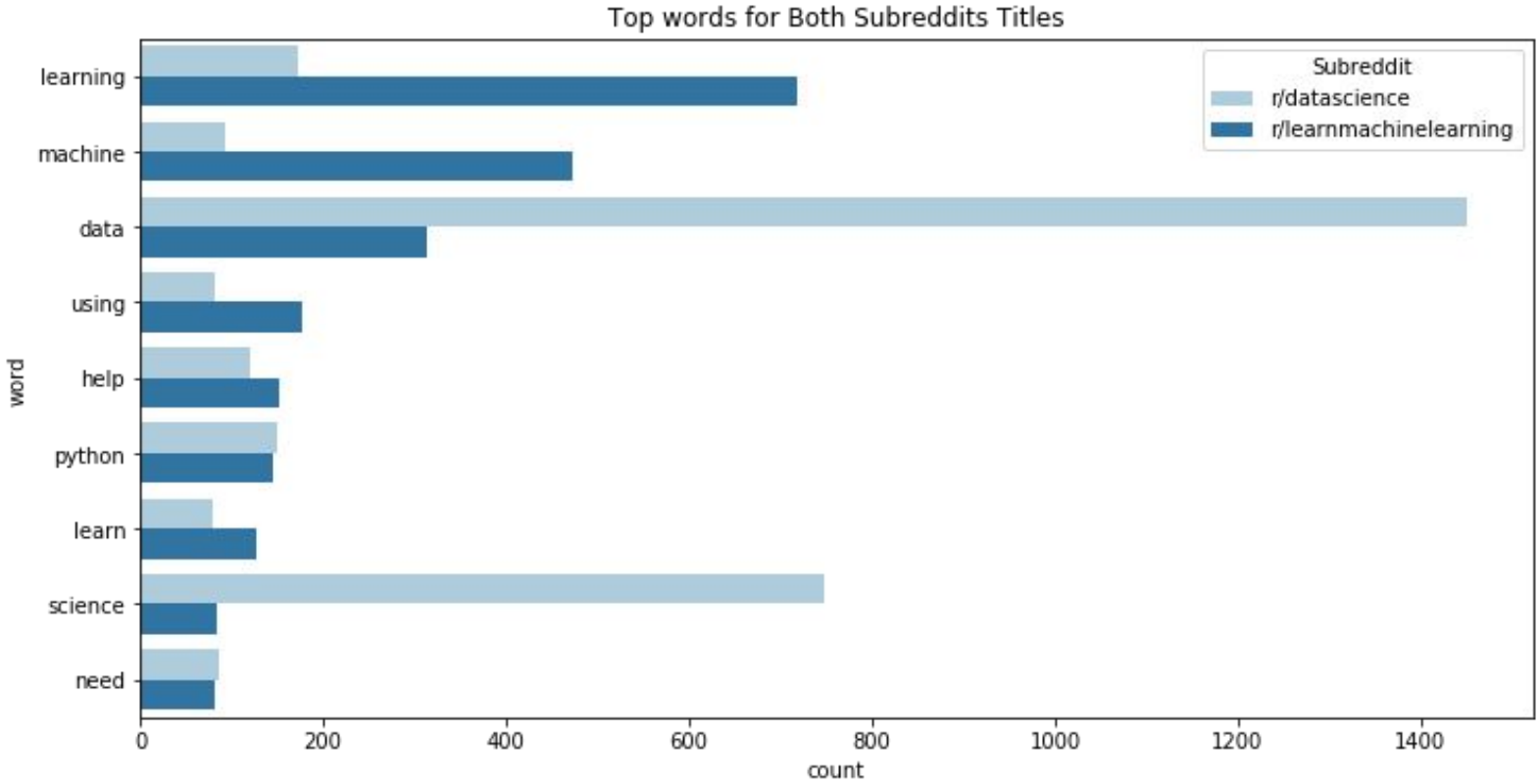
Top words for Data Science



Top words for Learn Machine Learning



Top words present in both subreddits:



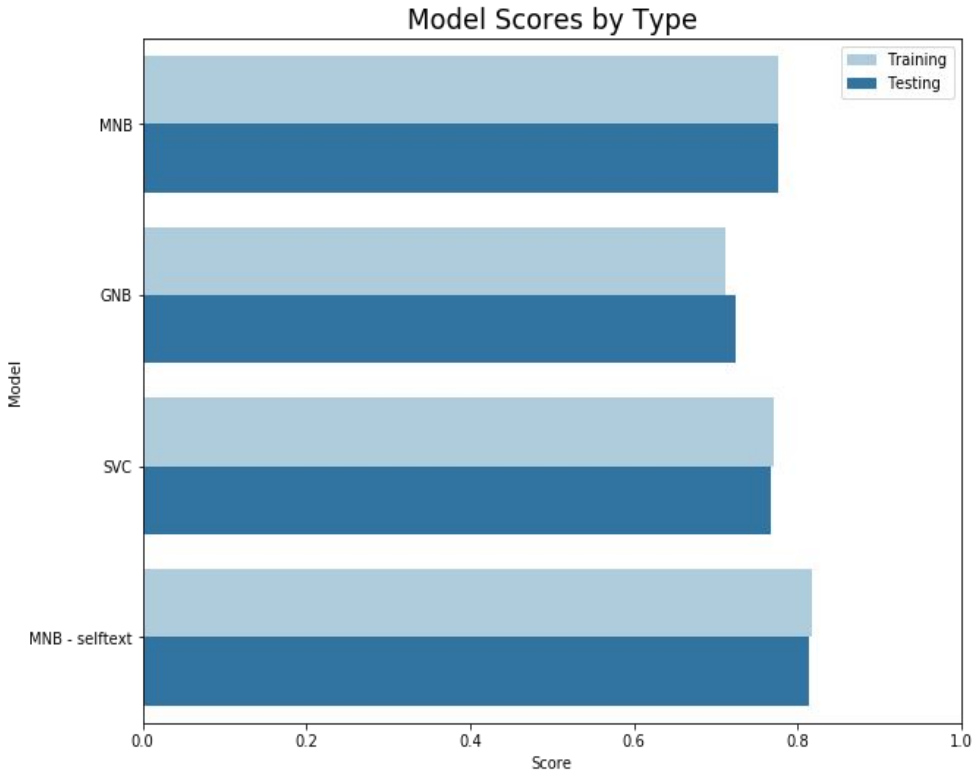
Modeling

- Gaussian Naive Bayes with TF-IDF Vectorizer, titles only
- Multinomial Naive Bayes with Count Vectorizer, titles only
- Support Vector Machine with TF-IDF Vectorizer, titles only
- Multinomial Naive Bayes with Count Vectorizer, selfpost only

- Each was performed as a grid search covering

Model Performance

- MNB on self-text was the best scorer
- Also computationally expensive
- Could only be performed on a subset



Conclusion:

- It’s hard to consistently classify which subreddit a given post came from
- My friend should acknowledge that machine learning and data science are inextricably linked!