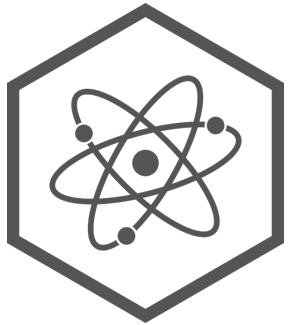
Reddit NLP Project

Jennifer Boyles
DSIR-11302020

Data Science Question



- Can we classify r/biochemistry and r/biology related questions based on scientific terminology?

Data Acquisition & Cleaning

- Subreddit data scraped using the PushShift's API

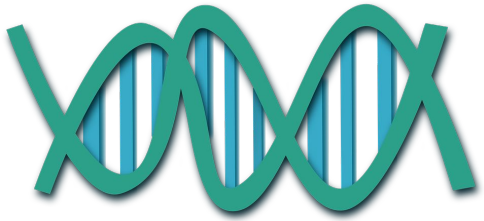
- Dropped [removed] & [deleted] posts

- Removal of extraneous items

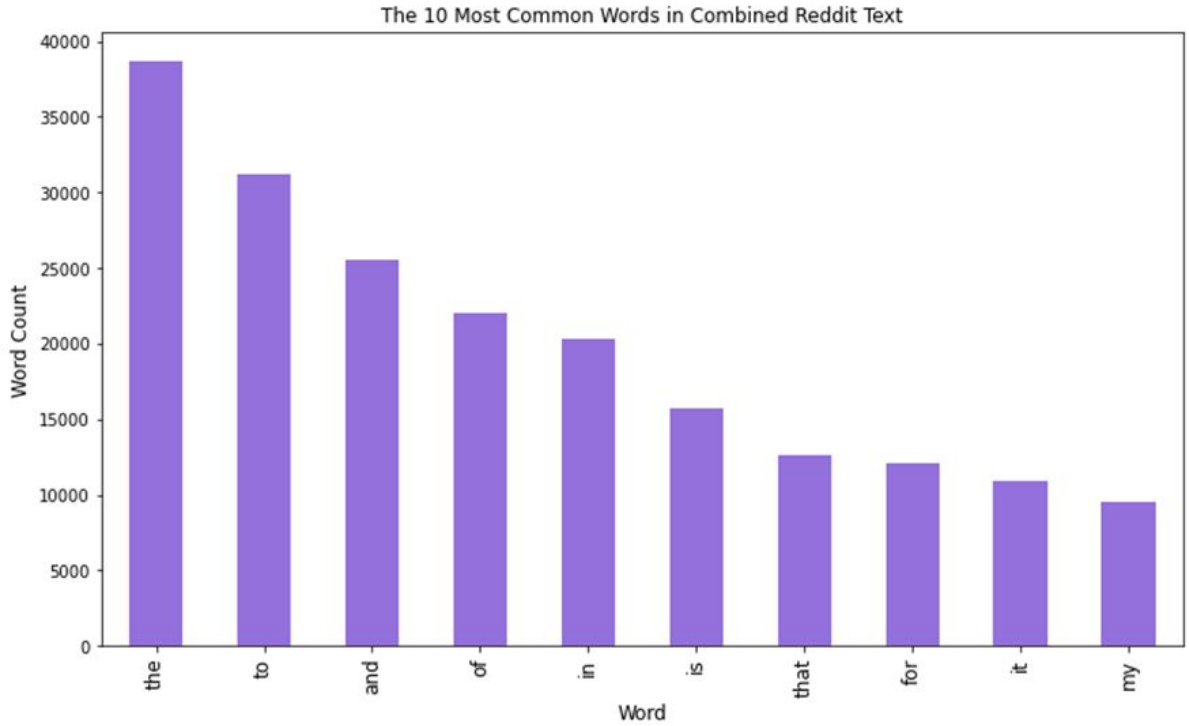
https://github.com/pushshift/api

High Impact Removal

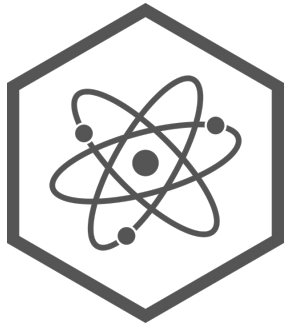
- Removal of subreddit name variations:
  - **biology, bio, biochemistry, biochem**



Most Common Words

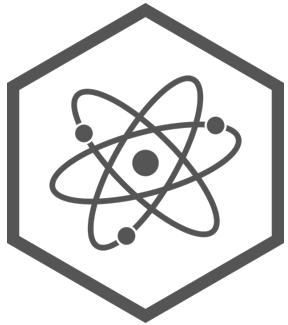


Null Model



- Null model for comparison to modeled classifiers
- Based on the most frequent value
- Our null model: 56.6% accuracy

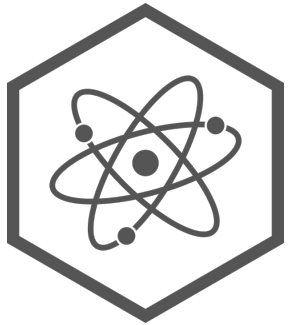
Evaluation Metrics



- Optimize Accuracy: users can be banned for off-topic posts

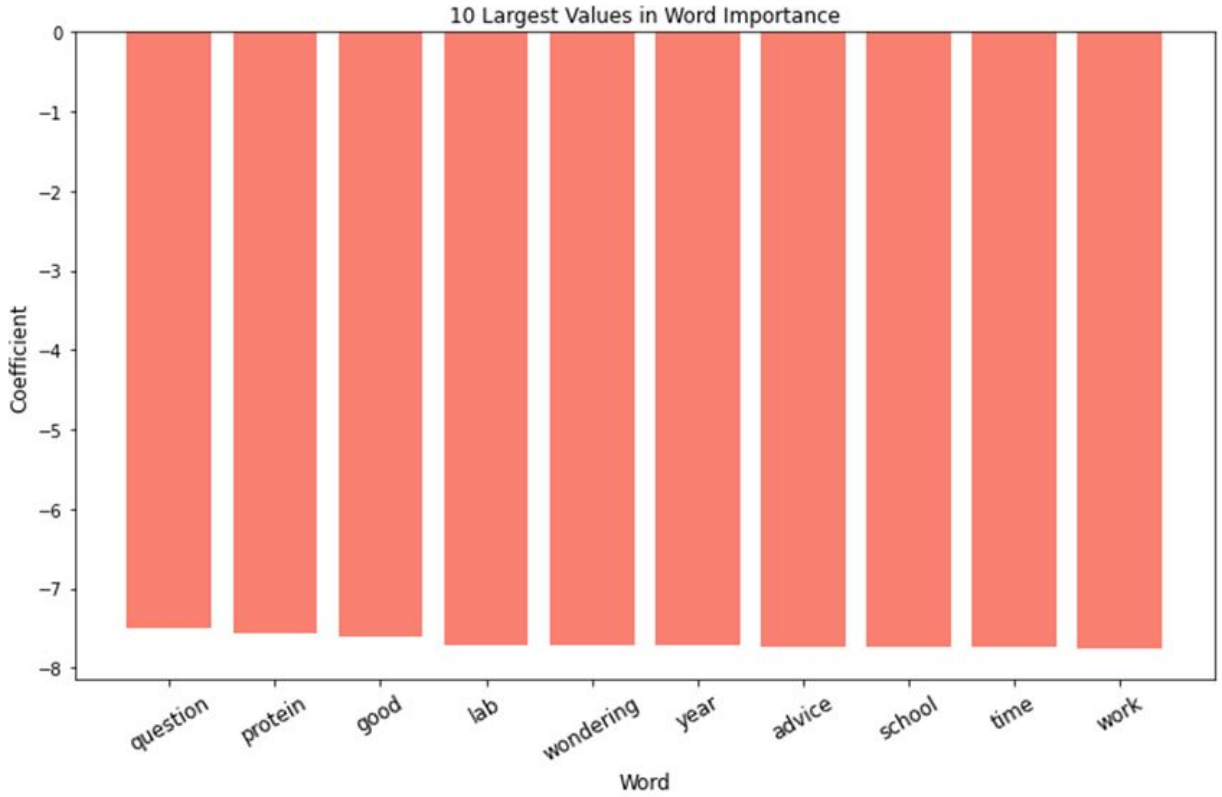
- Optimize Recall: get specific posts where they belong!

Final Model



- Model: Multinomial Naive Bayes, CountVectorizer
- Accuracy: 75%
- Recall: 86%

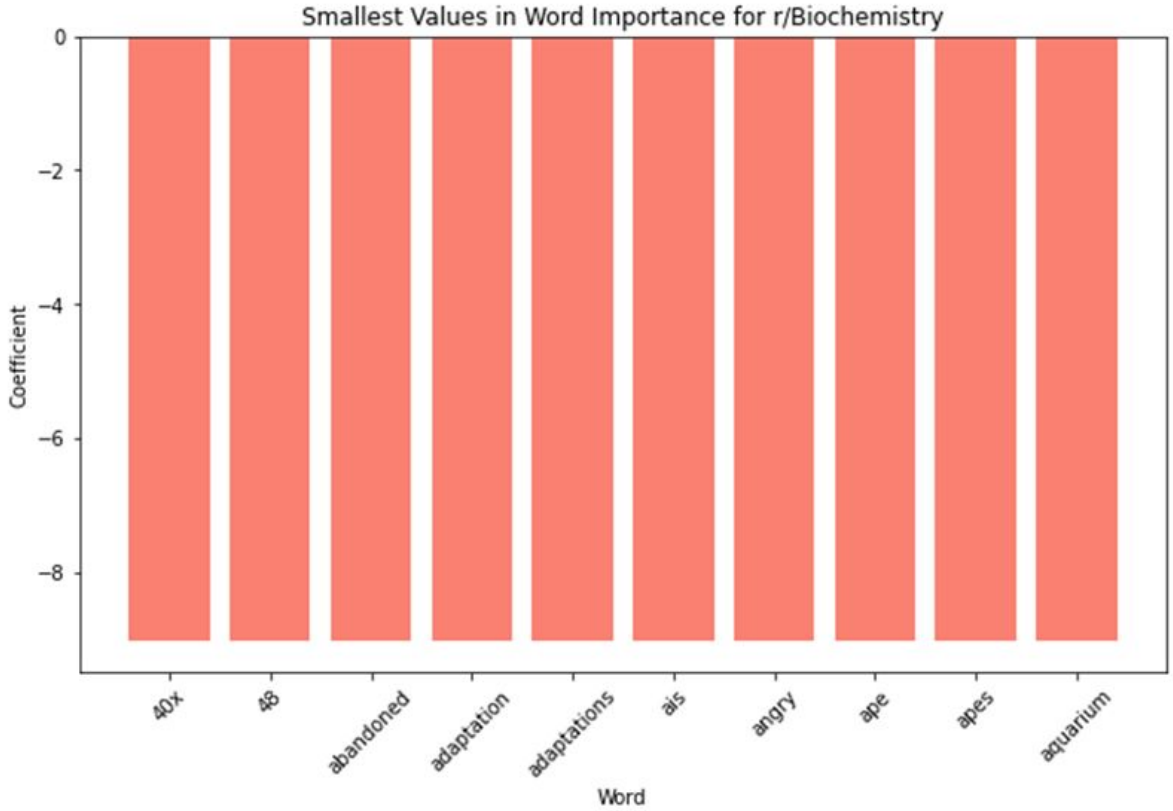
Most Important Words



Least Important Words

Coefficient Value:

-9.05

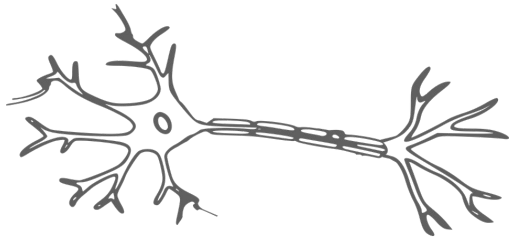


Other Least Important Words

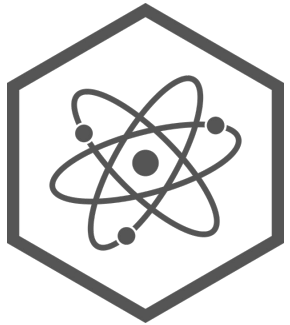
~250 words with the smallest coefficient (-9.05):

Many were biology specific terms:

Gametes, Phylogenetic Tree, Meiosis, Oocytes, Fitness, Niches



Conclusion

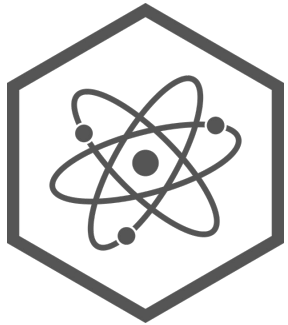


- Better performance than null
- Successful modeling of scientific terminology

  r/biochemistry: ‘cargo transport in our cells dynein’

  r/biology: ‘what could happen if the cells could survive a day longer in the body'

Next Steps/Improvements



More data -- larger training set of posts with more scientific terms.

Dealing with noise.

Thank you!

Questions?