Subreddit Classification Using Natural Language Processing

Caroline Clark

October 9th, 2020

Which Subreddit?

Build a model to classify a subreddit post as belonging to either r/artificial or r/datascience

Project Pipeline

Data Acquisition

Data Pre-Processing

Data Visualization

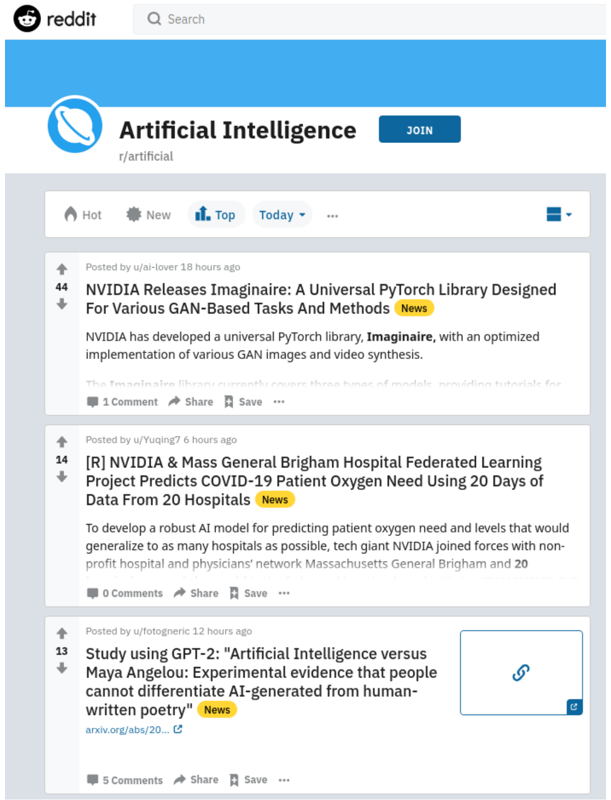
Baseline Naïve Bayes

Optimal Model

Conclusions

Data Collected with Pushshift API

r/artificial



Posted by u/ai-lover 18 hours ago
44
NVIDIA Releases Imaginaire: A Universal PyTorch Library Designed For Various GAN-Based Tasks And Methods News

NVIDIA has developed a universal PyTorch library, **Imaginaire,** with an optimized implementation of various GAN images and video synthesis.

The **Imaginaire** library currently covers three types of models, providing tutorials for

1 Comment Share Save ...

Posted by u/Yuqing7 6 hours ago
14
[R] NVIDIA & Mass General Brigham Hospital Federated Learning Project Predicts COVID-19 Patient Oxygen Need Using 20 Days of Data From 20 Hospitals News

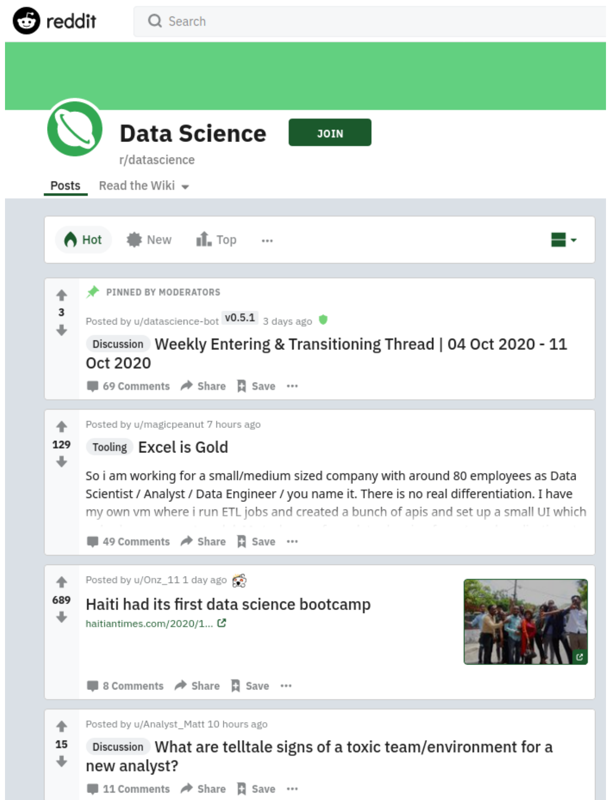
To develop a robust AI model for predicting patient oxygen need and levels that would generalize to as many hospitals as possible, tech giant NVIDIA joined forces with non-profit hospital and physicians' network Massachusetts General Brigham and **20**

0 Comments Share Save ...

Posted by u/fotogneric 12 hours ago
13
Study using GPT-2: "Artificial Intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry" News
arxiv.org/abs/20...

5 Comments Share Save ...

r/datascience



PINNED BY MODERATORS
3
Posted by u/datascience-bot v0.5.1 3 days ago
Discussion Weekly Entering & Transitioning Thread | 04 Oct 2020 - 11 Oct 2020
69 Comments Share Save ...

Posted by u/magicpeanut 7 hours ago
129
Tooling Excel is Gold

So i am working for a small/medium sized company with around 80 employees as Data Scientist / Analyst / Data Engineer / you name it. There is no real differentiation. I have my own vm where i run ETL jobs and created a bunch of apis and set up a small UI which

49 Comments Share Save ...

Posted by u/Onz_11 1 day ago
689
Haiti had its first data science bootcamp
haitiantimes.com/2020/1...
8 Comments Share Save ...

Posted by u/Analyst_Matt 10 hours ago
15
Discussion What are telltale signs of a toxic team/environment for a new analyst?
11 Comments Share Save ...

Data Cleaning Reduced Noise

- Dropped [removed] and [deleted] posts
- Dropped entries with missing post text
- Created 'all_text' feature
- Regex

Regex Used to Clean Text

"That's my goal for the next few years!" says Yann LeCun, the charismatic leader of Facebook AI. Hot off the Press! Here are the links to the November 2018 issue of **Computer Vision News**, the magazine of the algorithm community published by **RSIP Vision**: exclusive interview with **Yann LeCun**, many more articles about computer vision and **free subscription at page 32**.

[HTML5 version (recommended)](https://www.rsipvision.com/ComputerVisionNews-2018November/)

[PDF version](https://www.rsipvision.com/computer-vision-news-2018-november-pdf/)

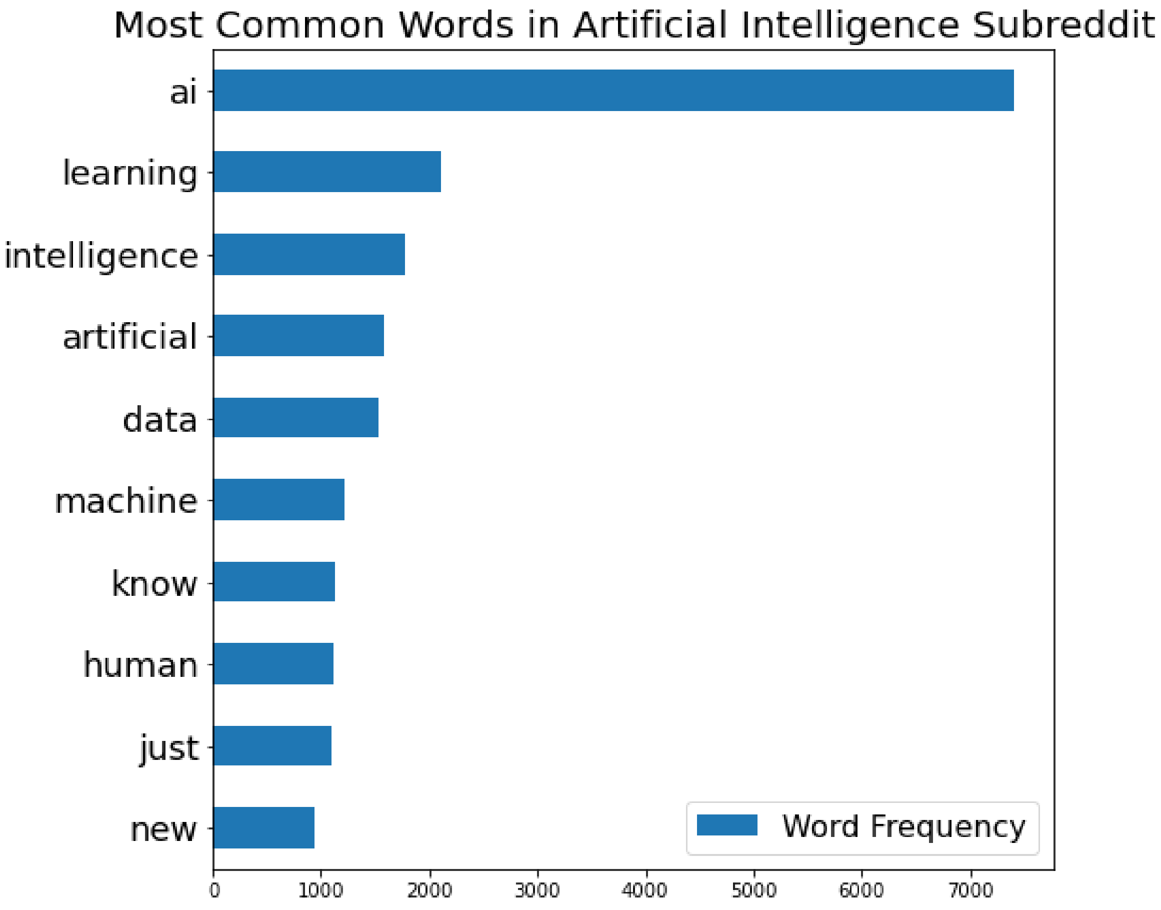
Enjoy!

&amp;#x200B;

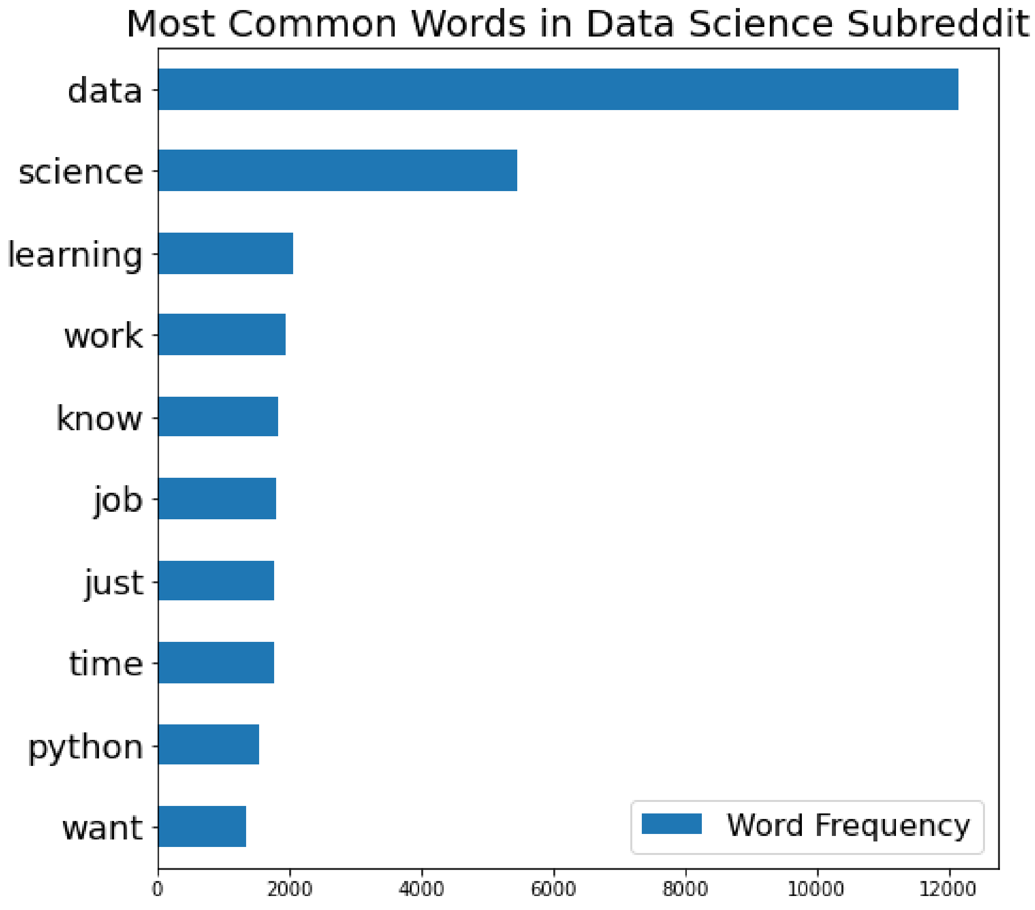
https://i.redd.it/79khdghqwpw11.jpg

That s my goal for the next few years! says Yann LeCun the charismatic leader of Facebook AI Hot off the Press! Here are the links to the November issue of Computer Vision News the magazine of the algorithm community published by RSIP Vision exclusive interview with Yann LeCun many more articles about computer vision and free subscription at page HTML version recommended PDF version Enjoy! &amp

How Much Overlap in Top Words?

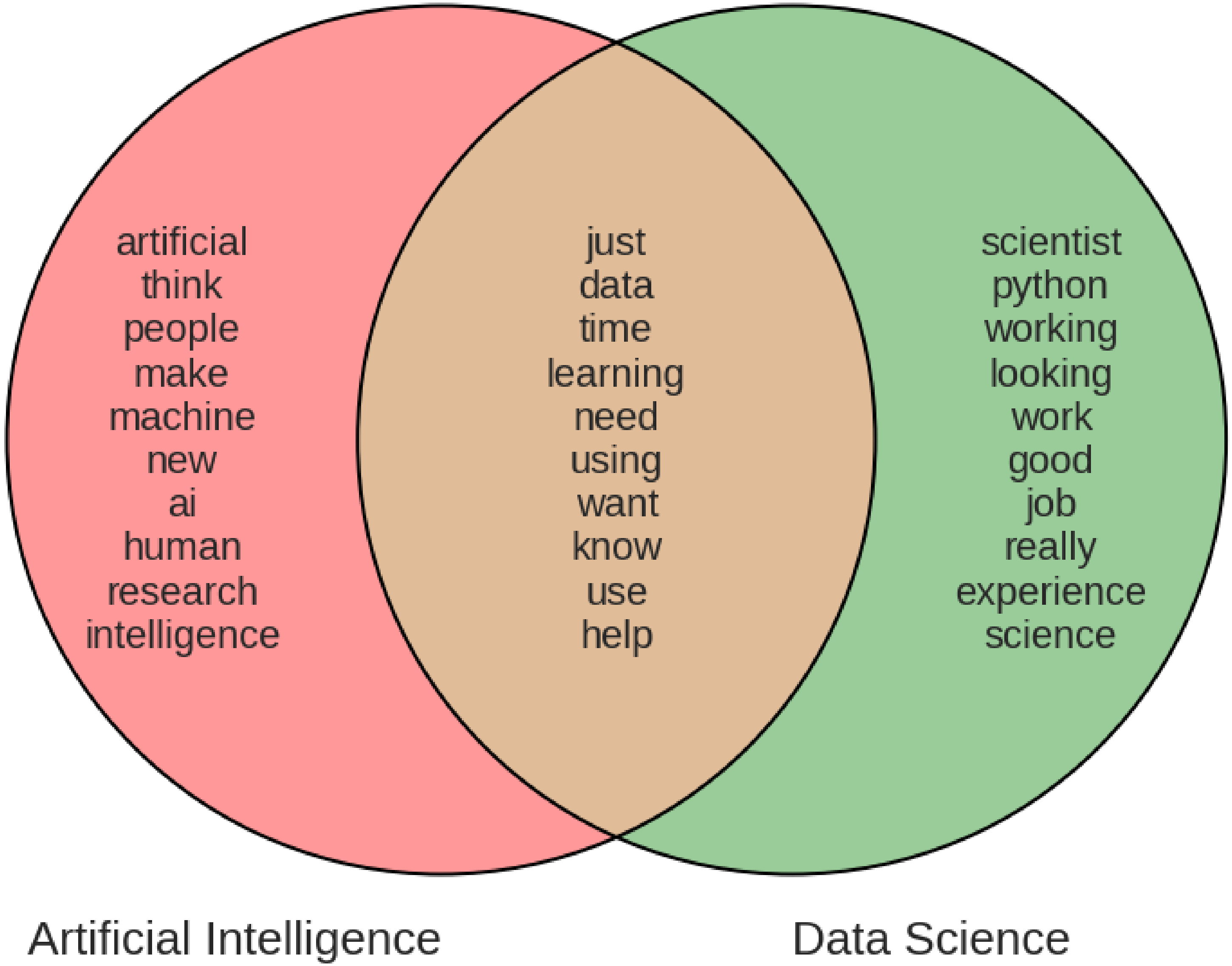


Most Common Words in Artificial Intelligence Subreddit



Most Common Words in Data Science Subreddit

Top Words Overlap ~70%



Data Pre-Processing and Vectorization

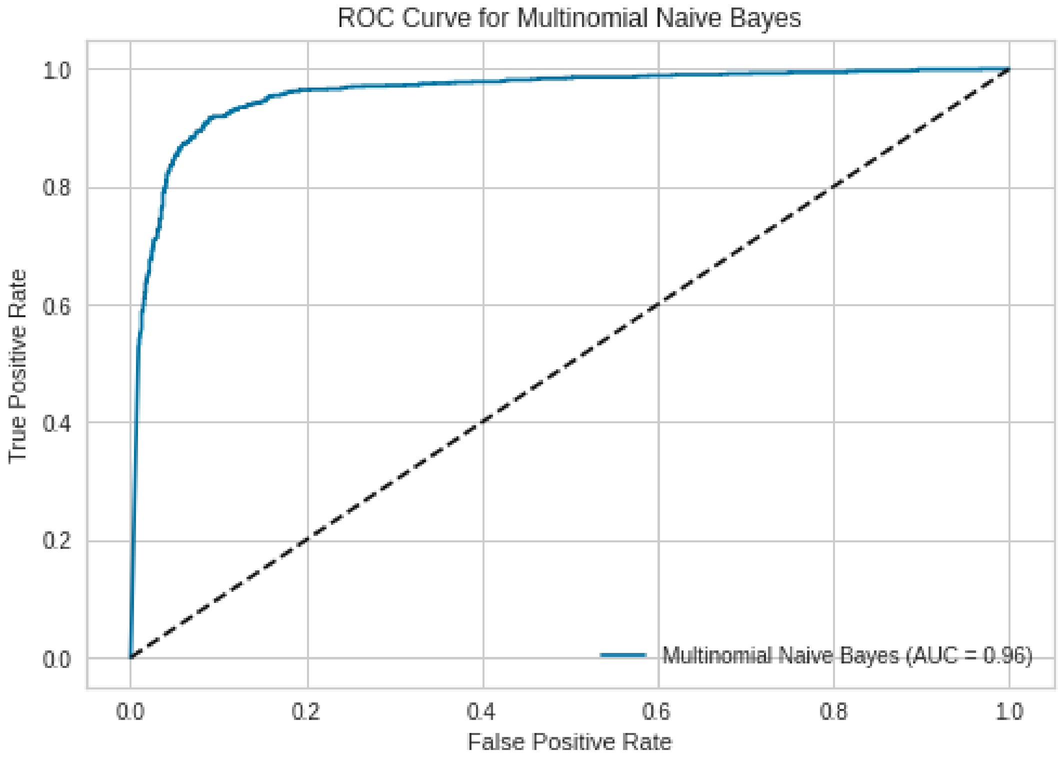
Pre-Processing

- Tokenizing
- Stemming
- Lemmatizing

Vectorization

- CountVectorizer
- TfidfVectorizer

Multinomial Naïve Bayes with CountVectorizer Achieved 90.67% Accuracy



Testing Accuracy	Training Accuracy
90.67%	95.53%

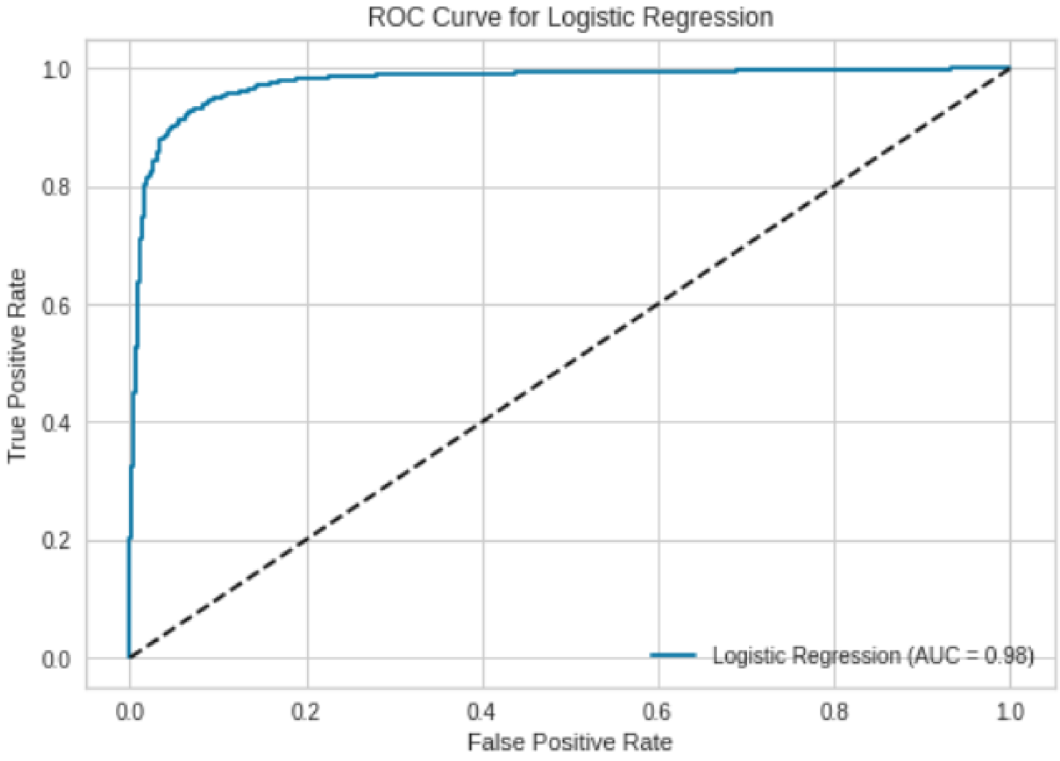
Multinomial Naïve Bayes

`alpha`	2

CountVectorizer

`ngram_range`	(1, 3)
`max_df`	0.5
`min_df`	2

Logistic Regression with Stemming and TfidfVectorizer Achieved 92.66% Accuracy



Testing Accuracy	Training Accuracy
92.66%	97.85%

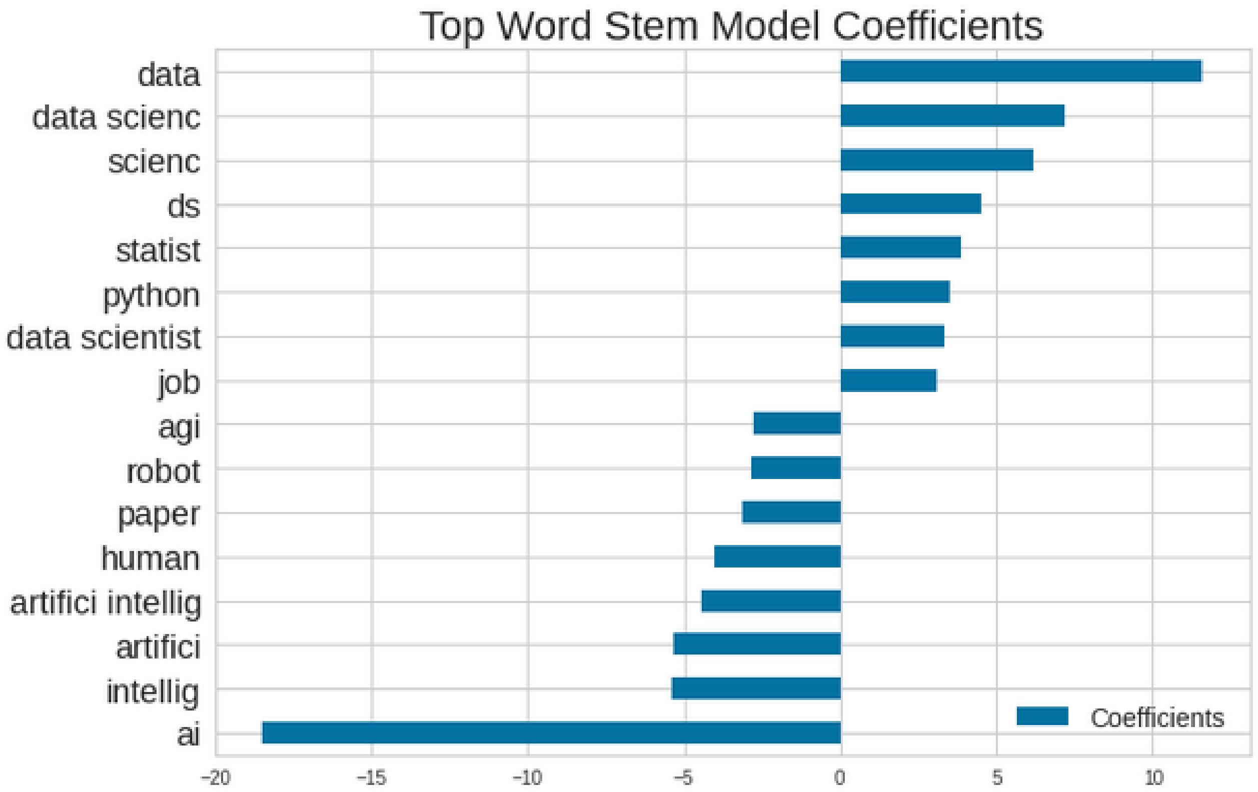
Logistic Regression

`C`	2

TfidfVectorizer

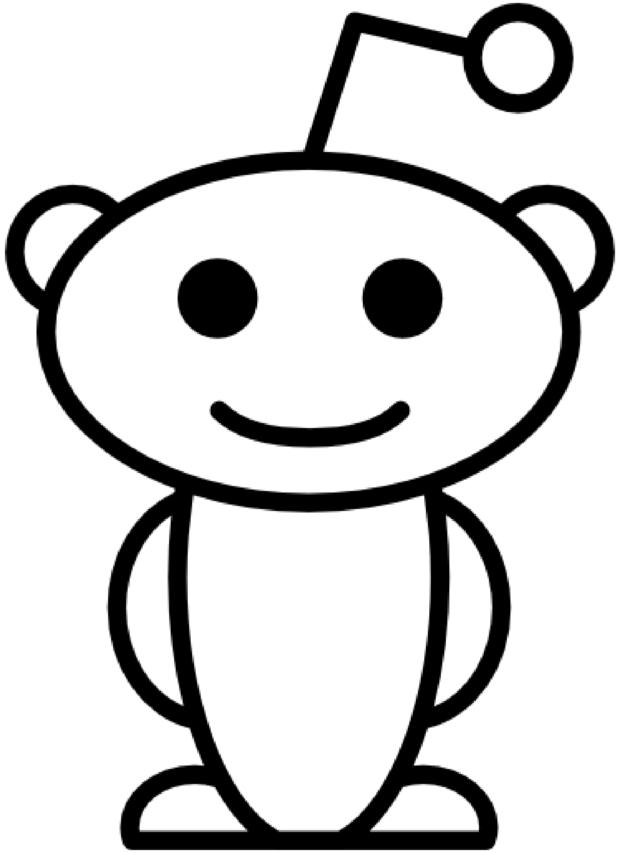
`preprocessor`	stemming
`ngram_range`	(1, 3)
`max_df`	0.5
`min_df`	2

Word Stems that Most Influenced the Classifier



Target Labels

r/datascience	*1*
r/artificial	*0*



Conclusions

- Logistic regression with tuned hyperparameters
  - 92.66% accuracy
  - Coefficients
- More text = better model
  - More relevant?
- Next Steps
  - Bias/variance tradeoff
  - New posts in r/datascience, r/artificial
  - r/deeplearning, r/learnmachinelearning

Questions?

Thank you!