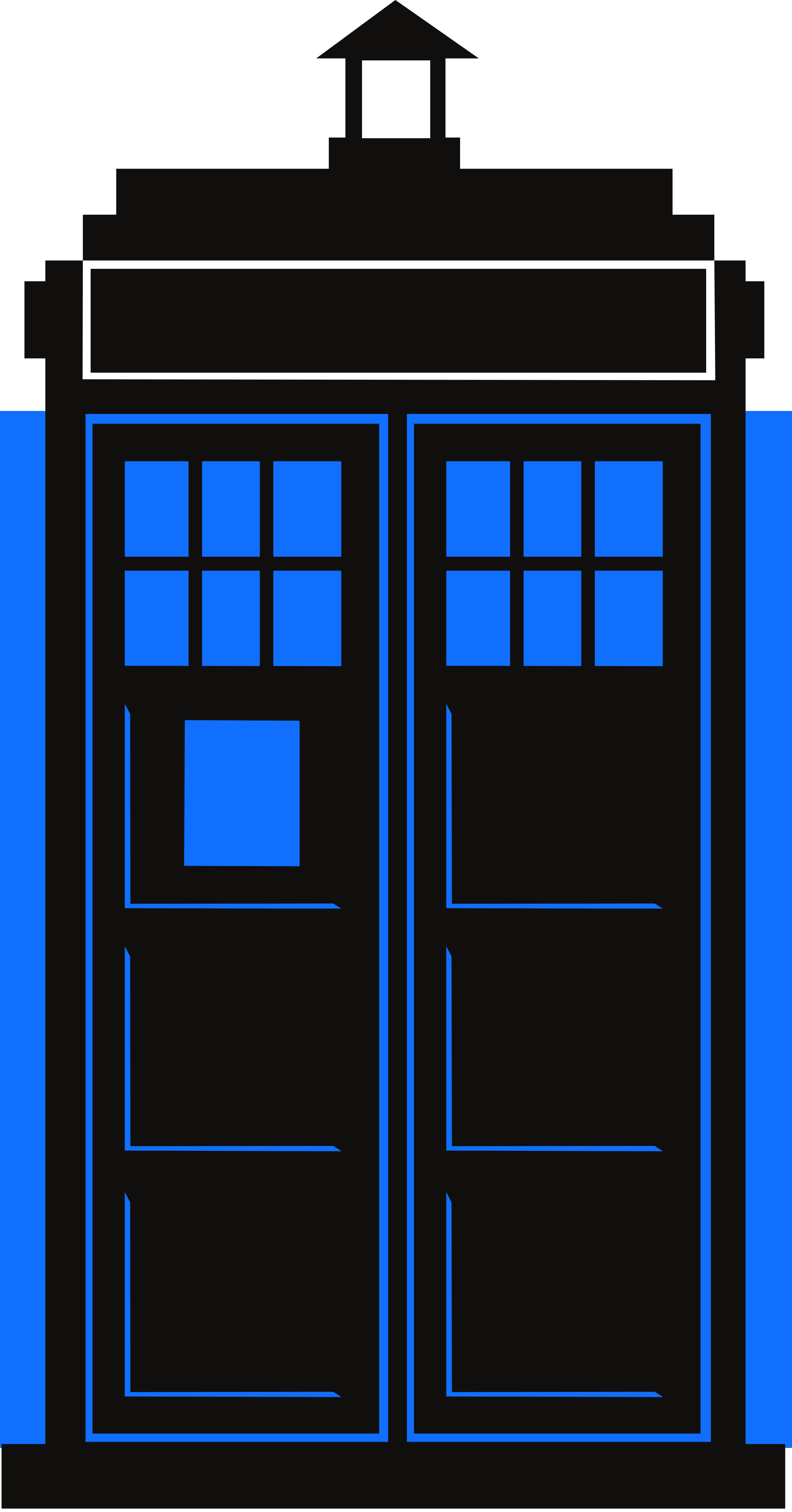


Predicting Coding Languages in Doctor Who Readme Files

By Caitlyn Carney

Executive Summary

What should you expect going into this?

Explore

What was learned in exploration?

Model

What models were ran and which one was the best.

Conclusion

What was learned overall.

Questions

Are there specific words used in different programming languages that can they help us predict the language being used?

Goal

Create a NLP model to predict the programming language used in a github repository based on the words and word combinations.

The Process

Acquire through web-scrapping, prepare the content, explore the data, create models, and pick the most accurate one to use.

Key Findings

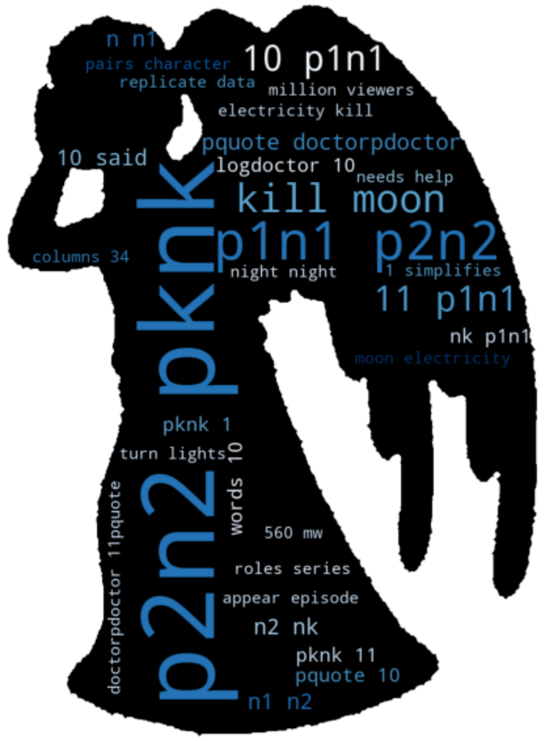
Each language has different frequent words from each other, Java Script is the most commonly used, and the top 10 most frequent words of all are all Java Script terms.

Conclusion

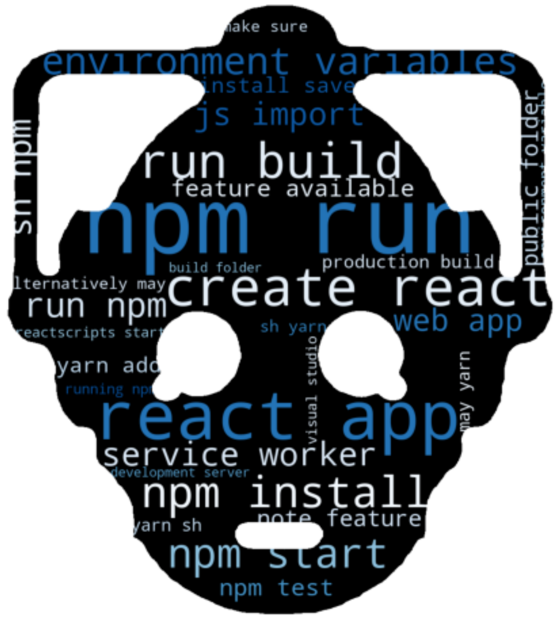
There are specific words used in readme files that key into what language is being used. The SGD Classifier model performed the best, and beat the base line by 14.3%.

Exploration

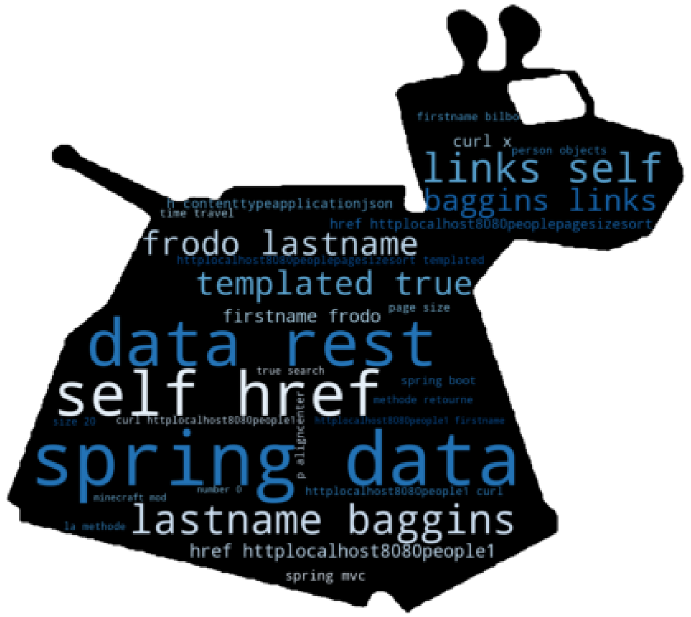
Python Word Combos



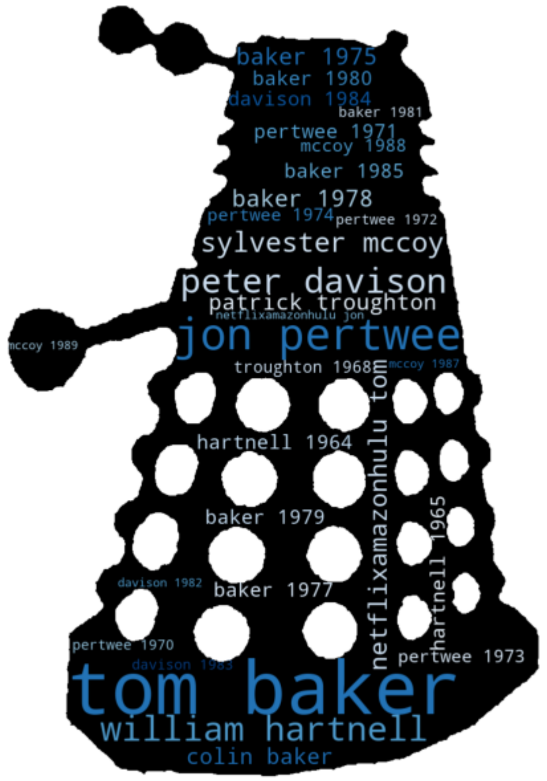
Java Script Word Combos



Java Script Word Combos



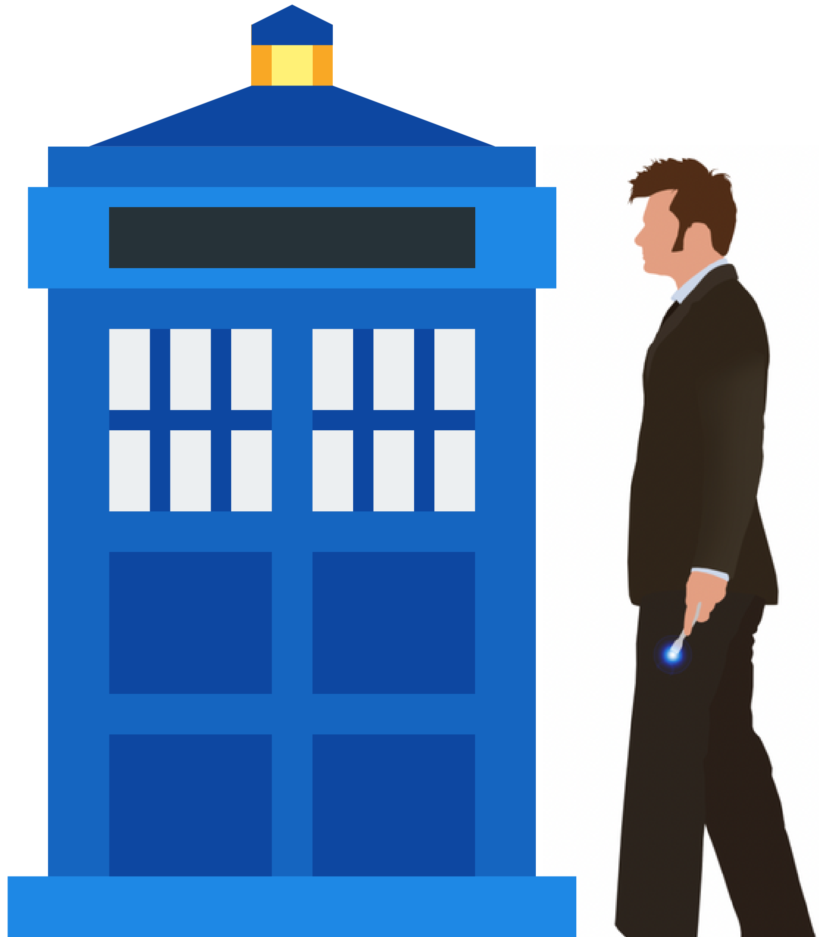
HTML Word Combos



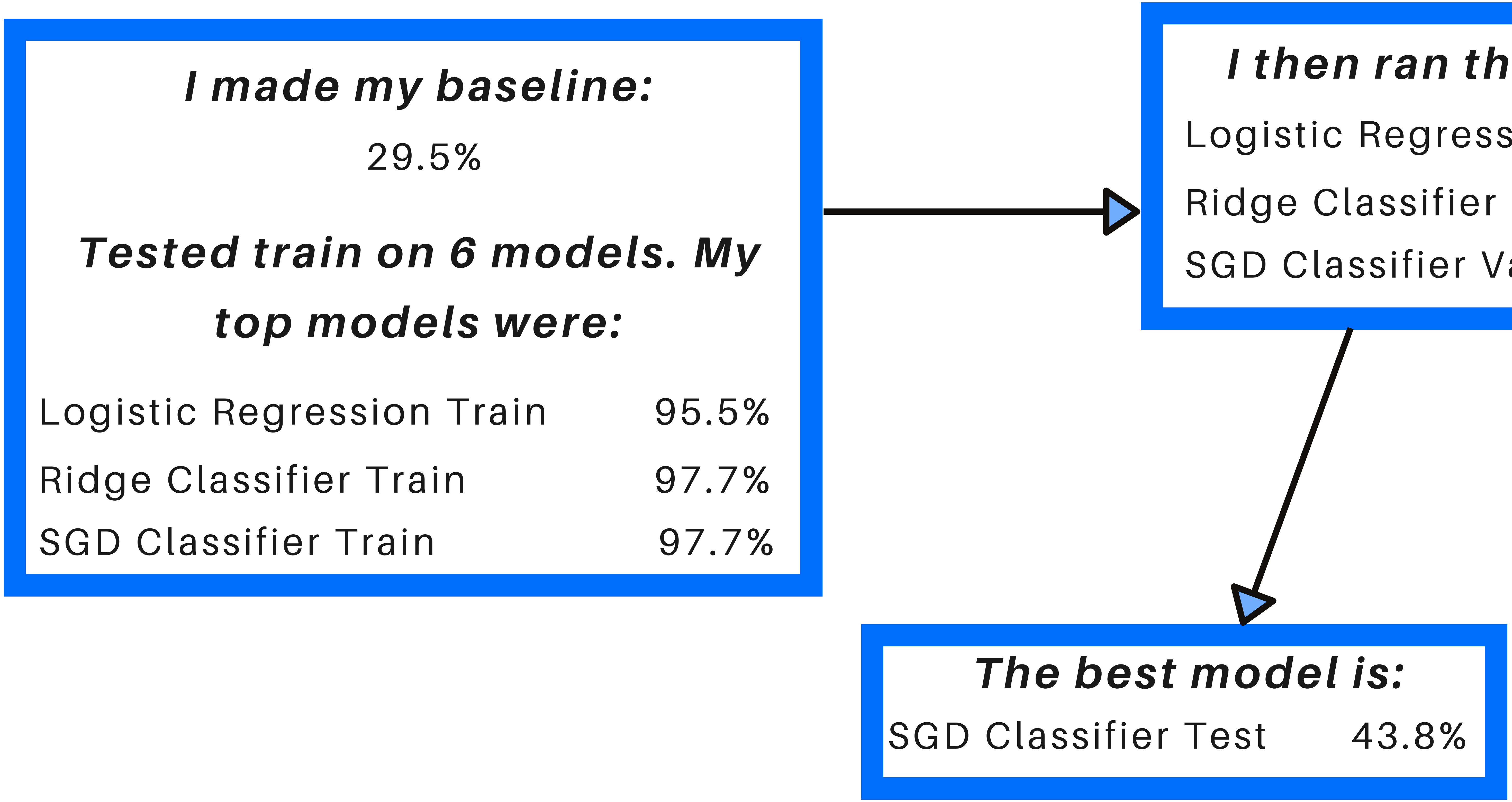
I learned that there are specific phrases/words that are used more often in each coding language.

I also learned that there were no commonalities between the languages and their top 20 words/phrases.

Java script is the most commonly used coding language and makes up the top 10 most frequent words of all.



Modeling



I made my baseline:

29.5%

Tested train on 6 models. My top models were:

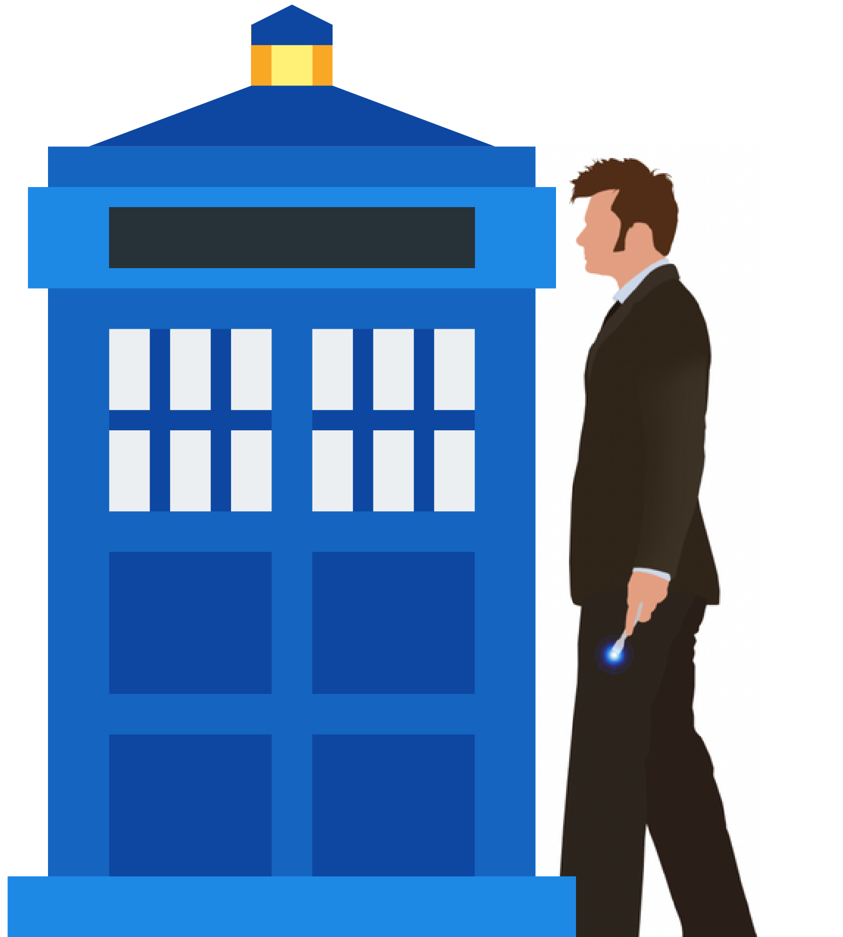
Logistic Regression Train	95.5%
Ridge Classifier Train	97.7%
SGD Classifier Train	97.7%

I then ran them on validate:

Logistic Regression Validate	42.1%
Ridge Classifier Validate	36.8%
SGD Classifier Validate	57.9%

The best model is:

SGD Classifier Test	43.8%



Conclusion

Exploration

There are specific words and phrases that feed into each coding language individually, with no overlap within the top 20 of each language.

Java Script is the most prominently used coding language out of them all. This coding language makes up for the top 10 most frequently used words across all languages.

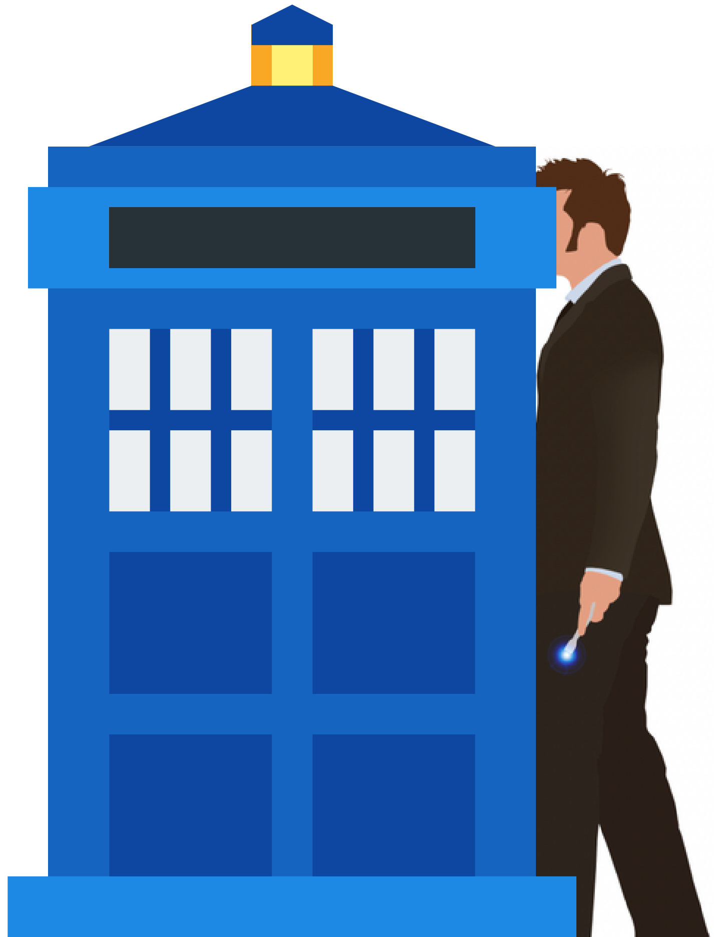
Modeling

Each model beat the baseline in train, and validate.

However, the best model overall was the SGD Classification model which beat my the baseline of 29.5% by 14.3%. Putting it at a 43.8% accuracy reading.

With Further Time

I would like to add in more repositories to see it's affect on the models accuracy



Questions?

https://github.com/CaitlynCarney/coding_language_prediction

https://www.linkedin.com/in/caitlyn-carney/