Predicting GitHub repo language using NLP

A quest to predict programming language from a single aspect of an entire body of work

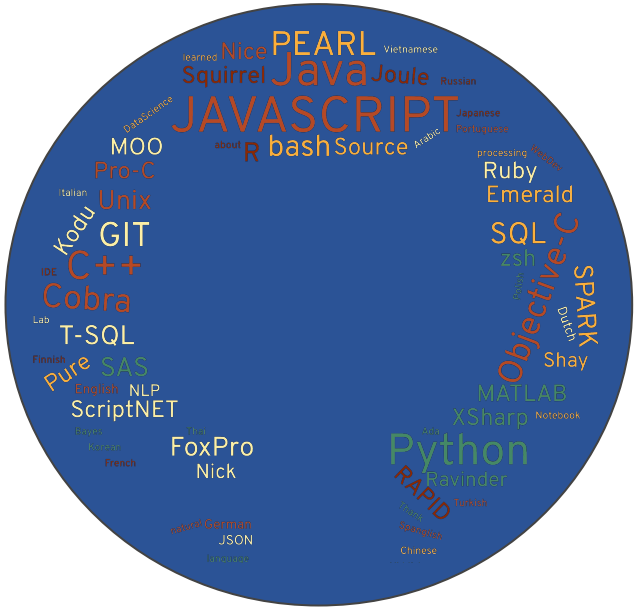
By Shay Altshue, Ravinder Singh, and Nick Joseph

“You shall know a word by the company it keeps.”
– *J.R. Firth, Professor of General Linguistics, 1957*

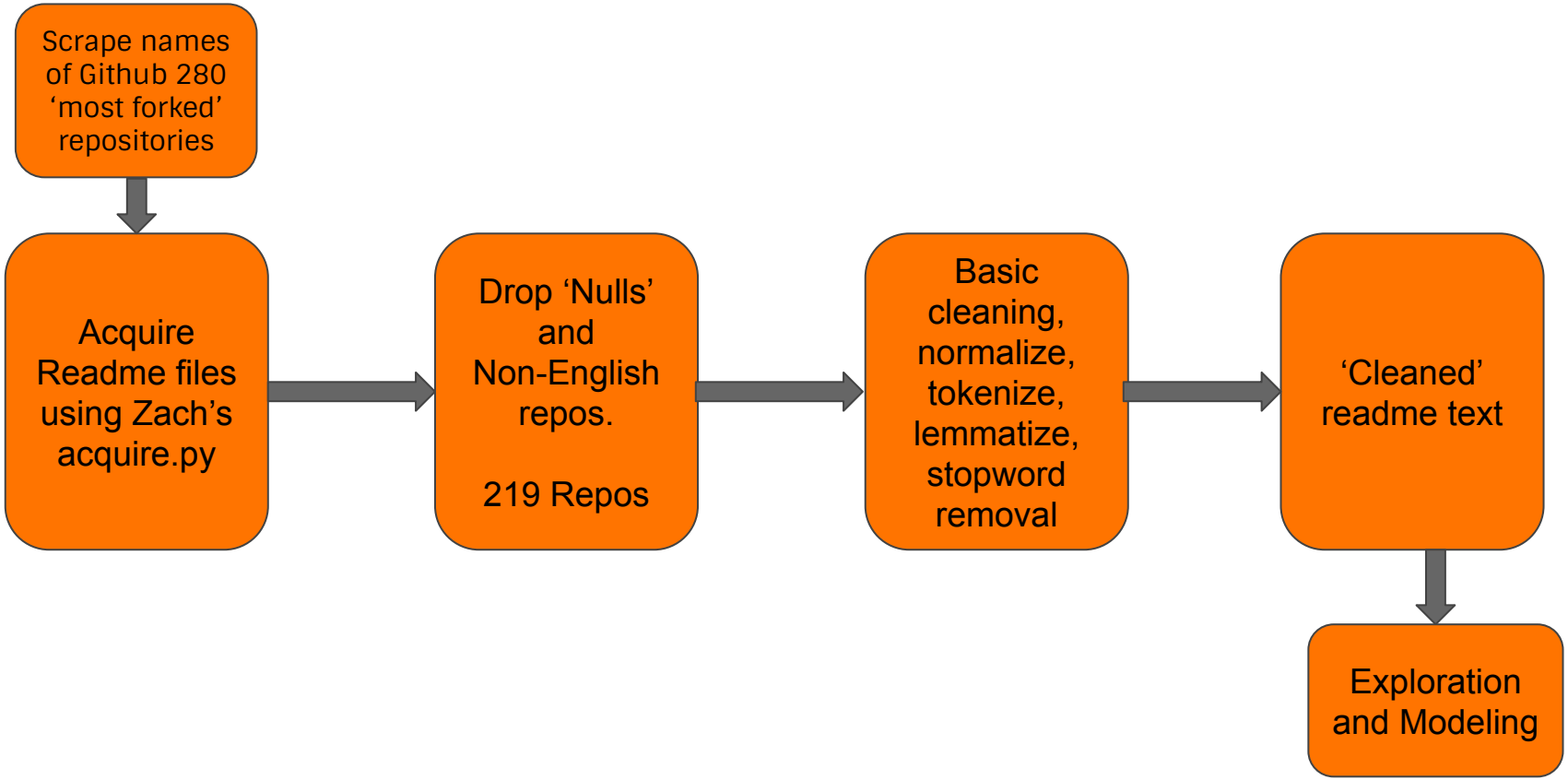
”Hope it ain’t this company.”
- *Ravinder Singh, Co-Founder, The Tree Musketeers, 2020*

Executive Summary

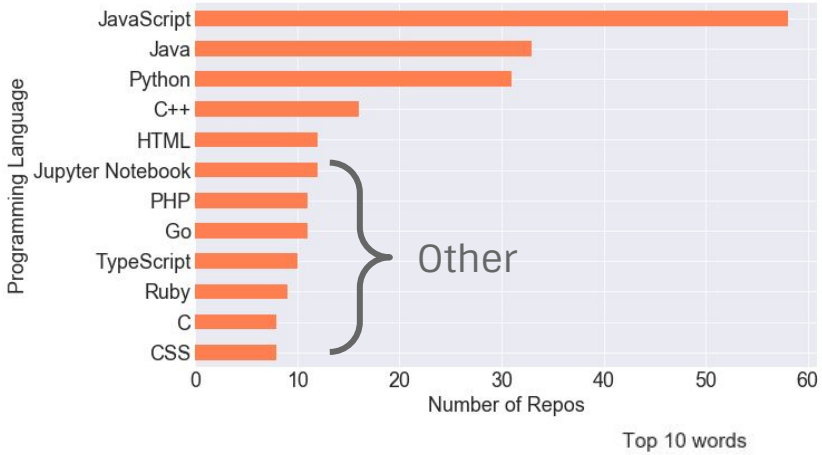
- Scraped the list of 280 of the “most forked” repositories on Github.
- After cleanup, we had 219 readme files, which we used to train the model
- Created a baseline model, which only predicts a repository is using JavaScript (the most recurring language). Predicted accurately 26% of the time
- Our model was able to accurately predict the programming language used in a given repository 68% of the time



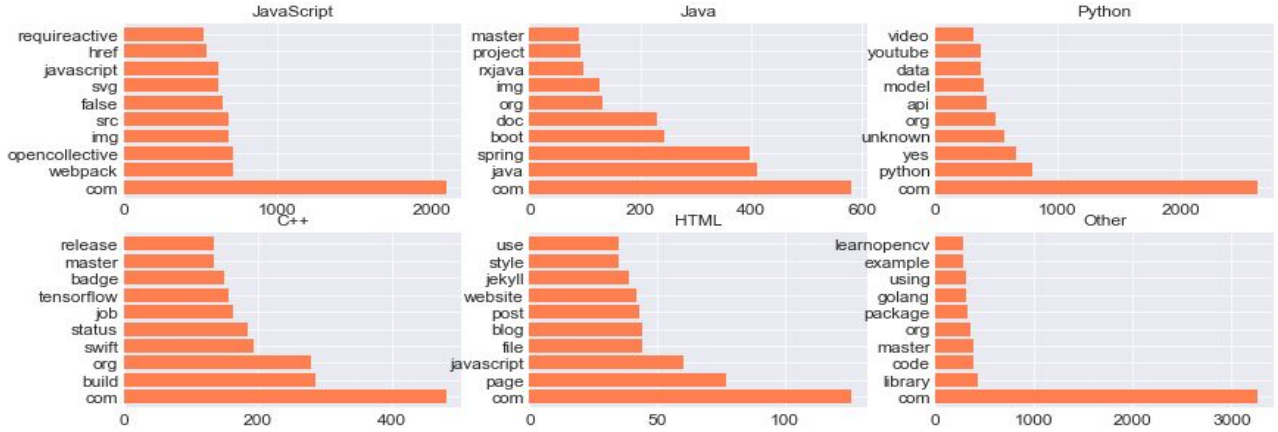
Data Acquisition and Preparation Pipeline



Exploration



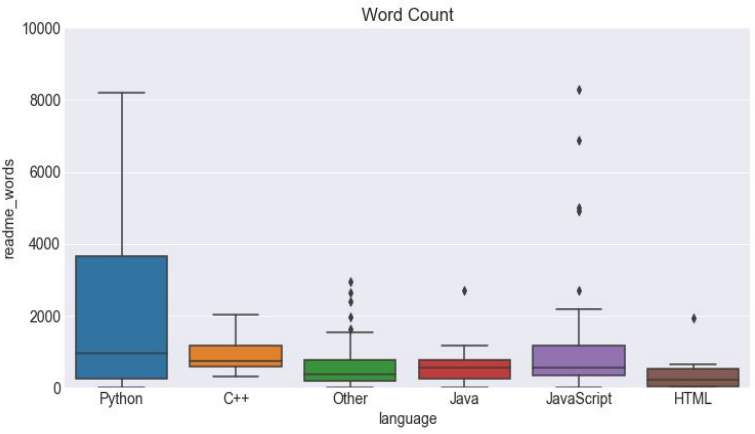
Top 10 words

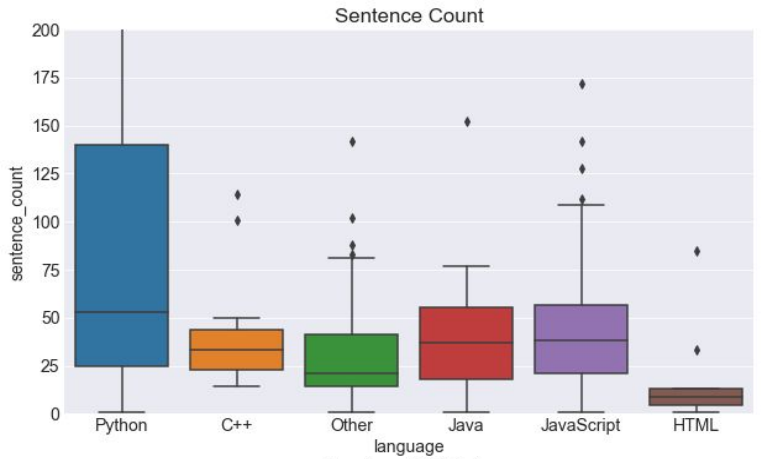


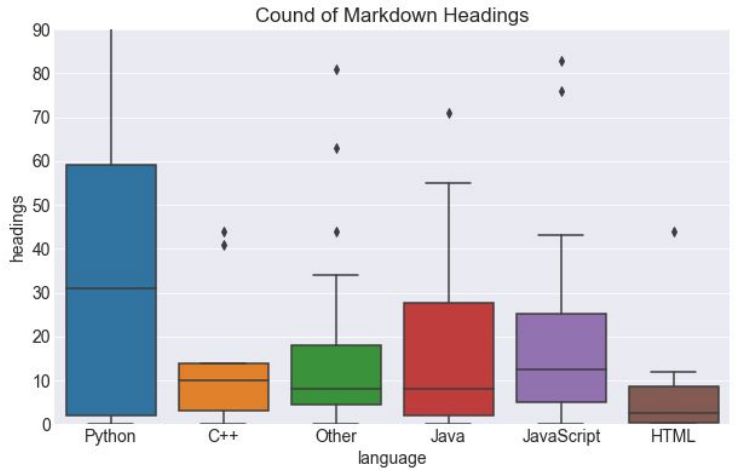
Takeaways:

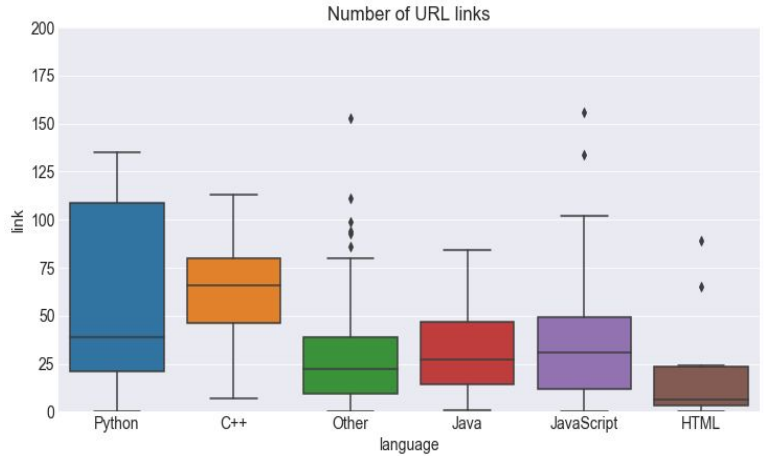
- JavaScript is most often used language in the repos analyzed.
- ‘Python’, ‘yes’ and ‘unknown’ are most frequently used words for Python
- Top 10 words for each language is quite distinct.

Feature Exploration









Of all the features explored:

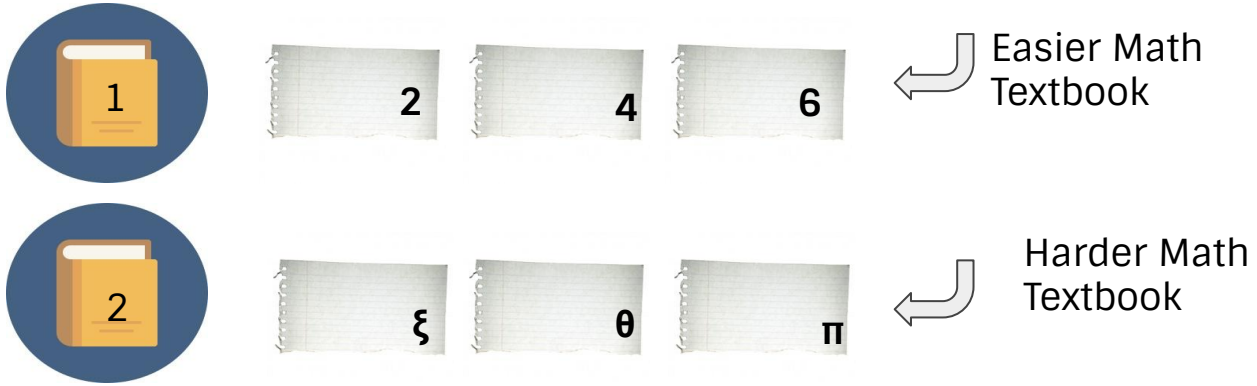
Python has highest ‘spread’ (IQR).

HTML has lowest

Modeling

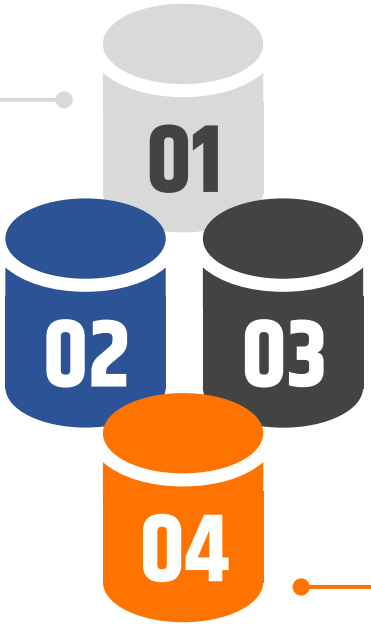
Model	Accuracy of Predictions
Naive Bayes	**68%**
Logistic Regression	61%
Decision Tree	56%
Random Forest	56%
K Nearest Neighbor	54%
Baseline (JavaScript)	26%

Naive Bayes: How it Works



Conclusion

Getting a good sample of Github Repos is difficult



Finding new ways to differentiate the languages

Readmes are chaotic and challenging to fine tune predictions from

Scraping the files in a repo to predict language, not the readmes