

Coding Language Predictions in Google Github README's

NLP-Classification Project
Amanda Gomez
11 APRIL 2020



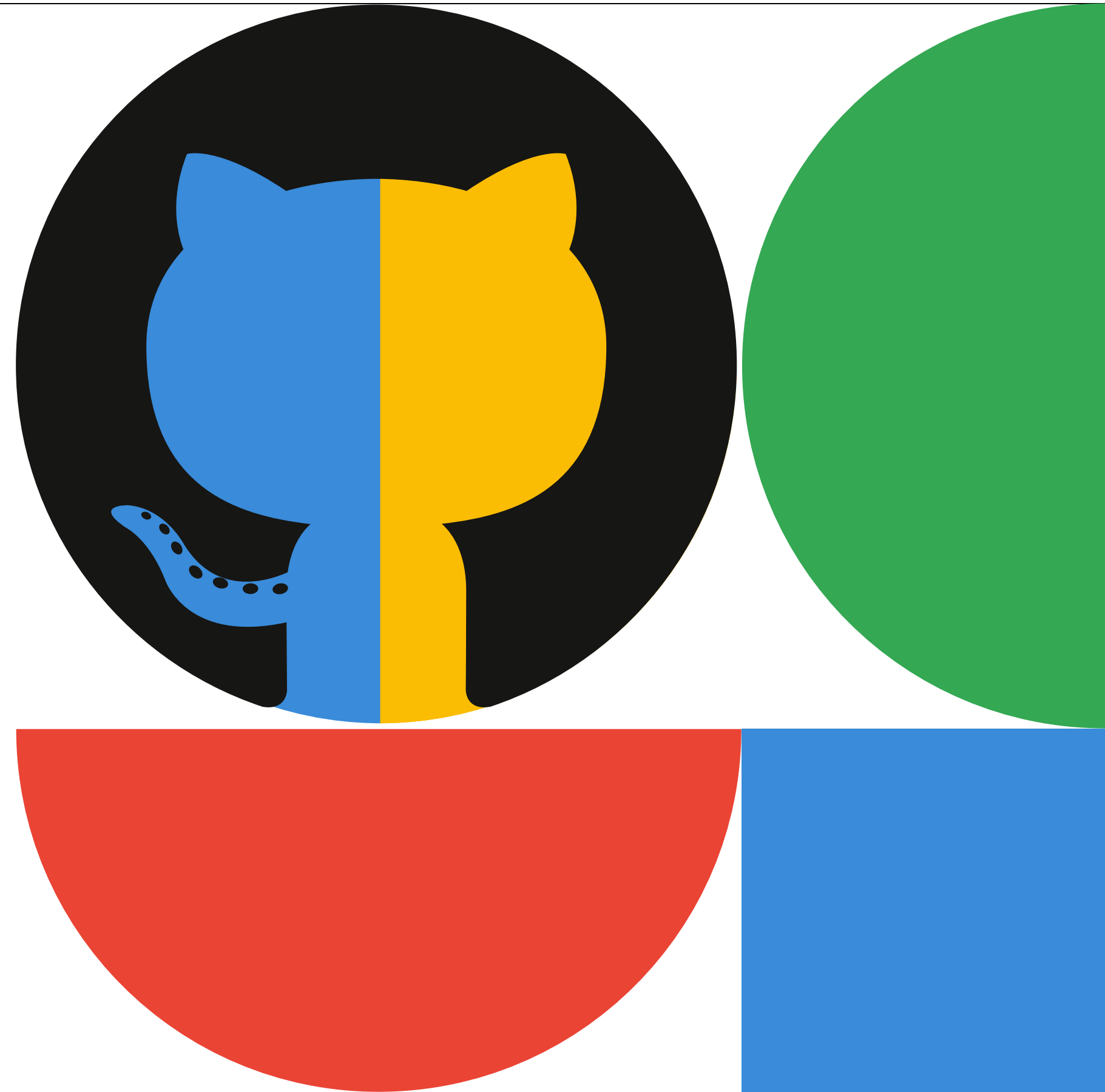
Today's Agenda

Executive Summary

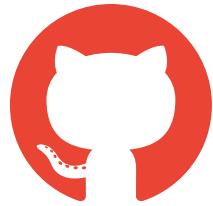
Data Analysis

Conclusion

Appendix

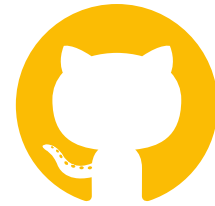


Executive Summary



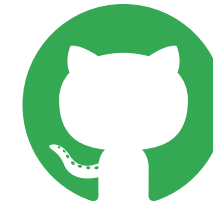
Goal

The goal for this project is to create a model that will accurately predict the primary coding language of a Github Repository given text from a README.



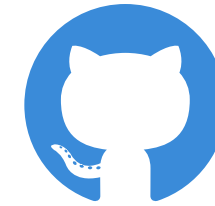
Data Set

This data set was scraped from Google's Github site. Not all pages were obtained due to empty repositories, although the csv used contains 1020 observations.



Findings

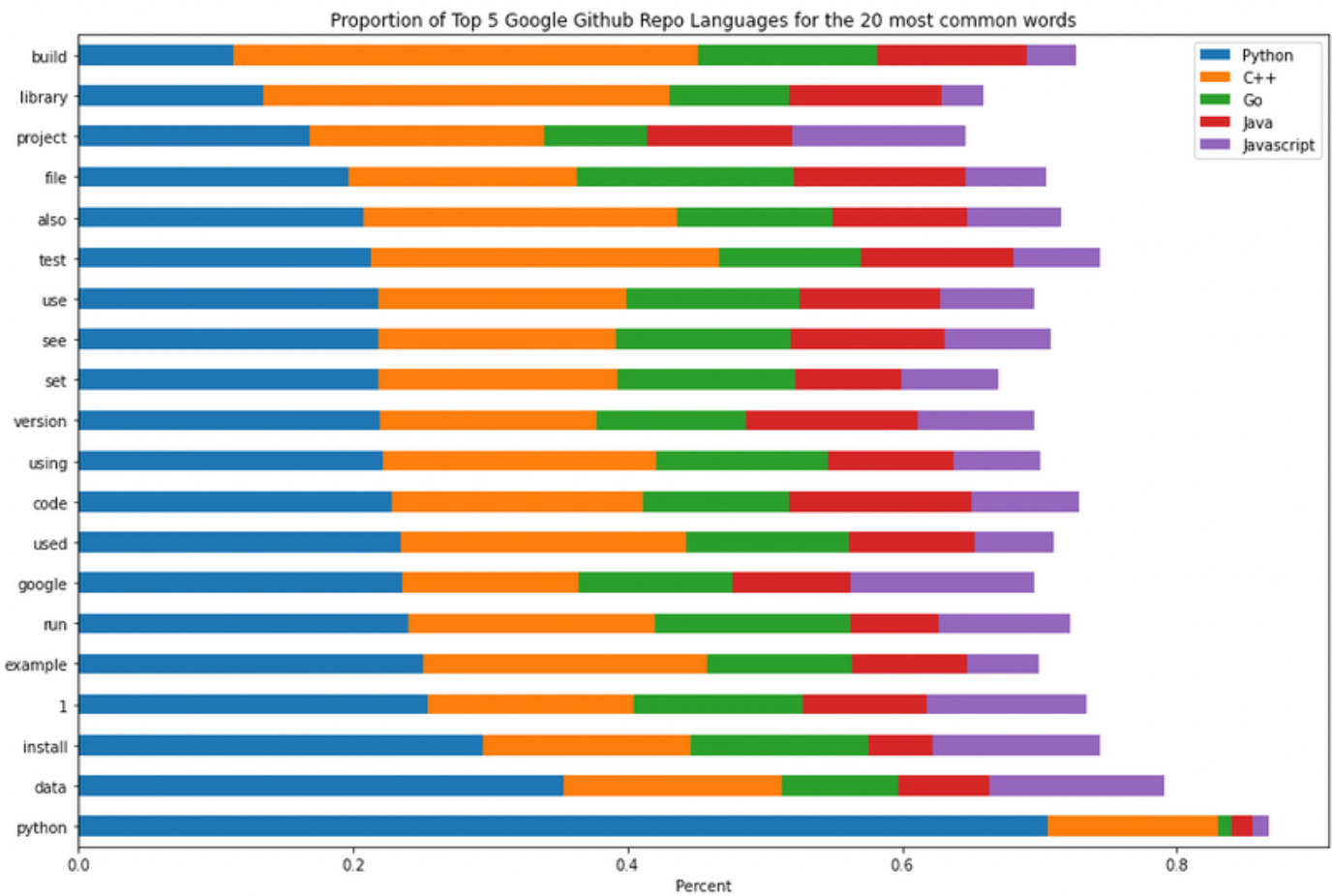
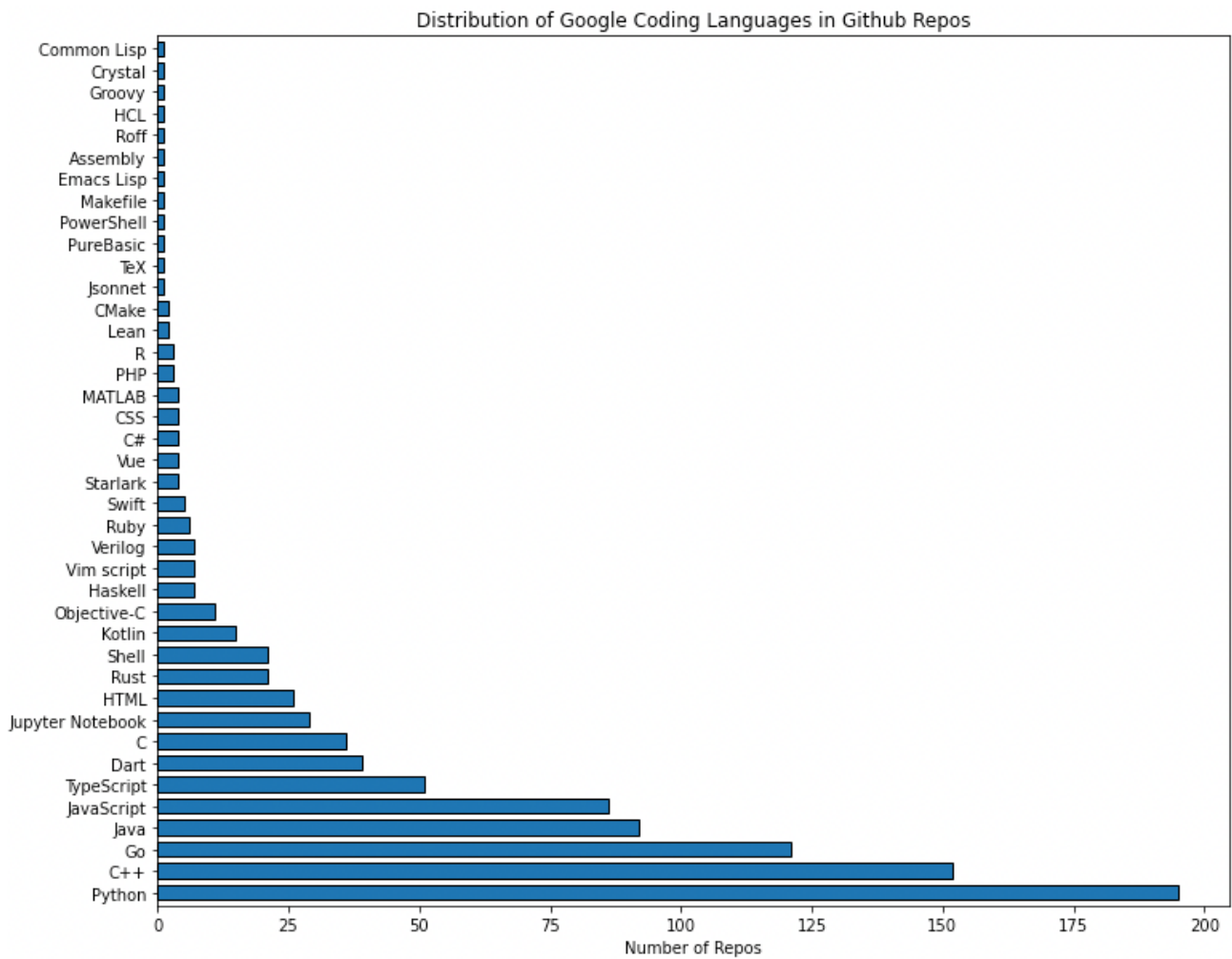
- Google uses over 30 different languages in their repos, which drove down the accuracy.
- The top common words used across the languages were verbs.



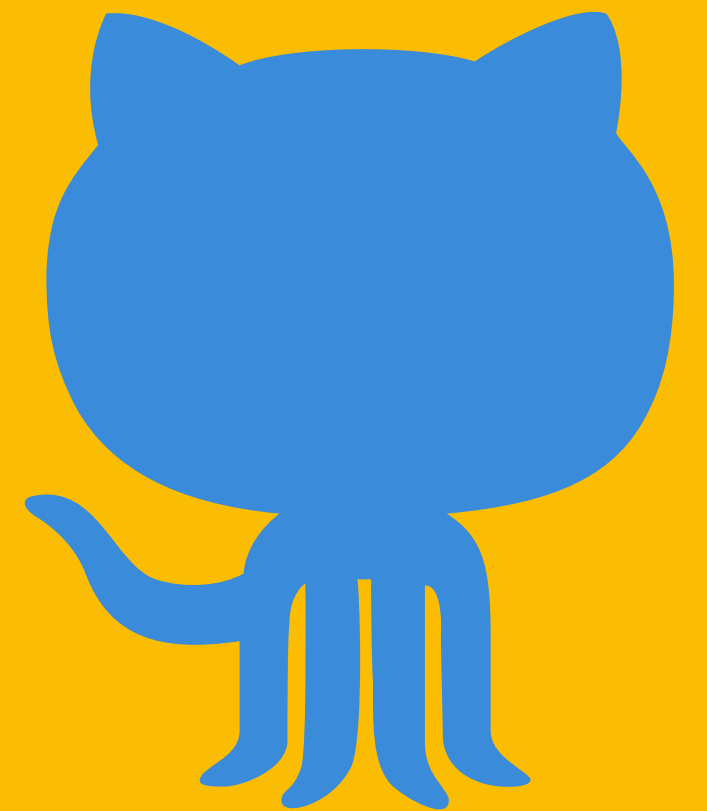
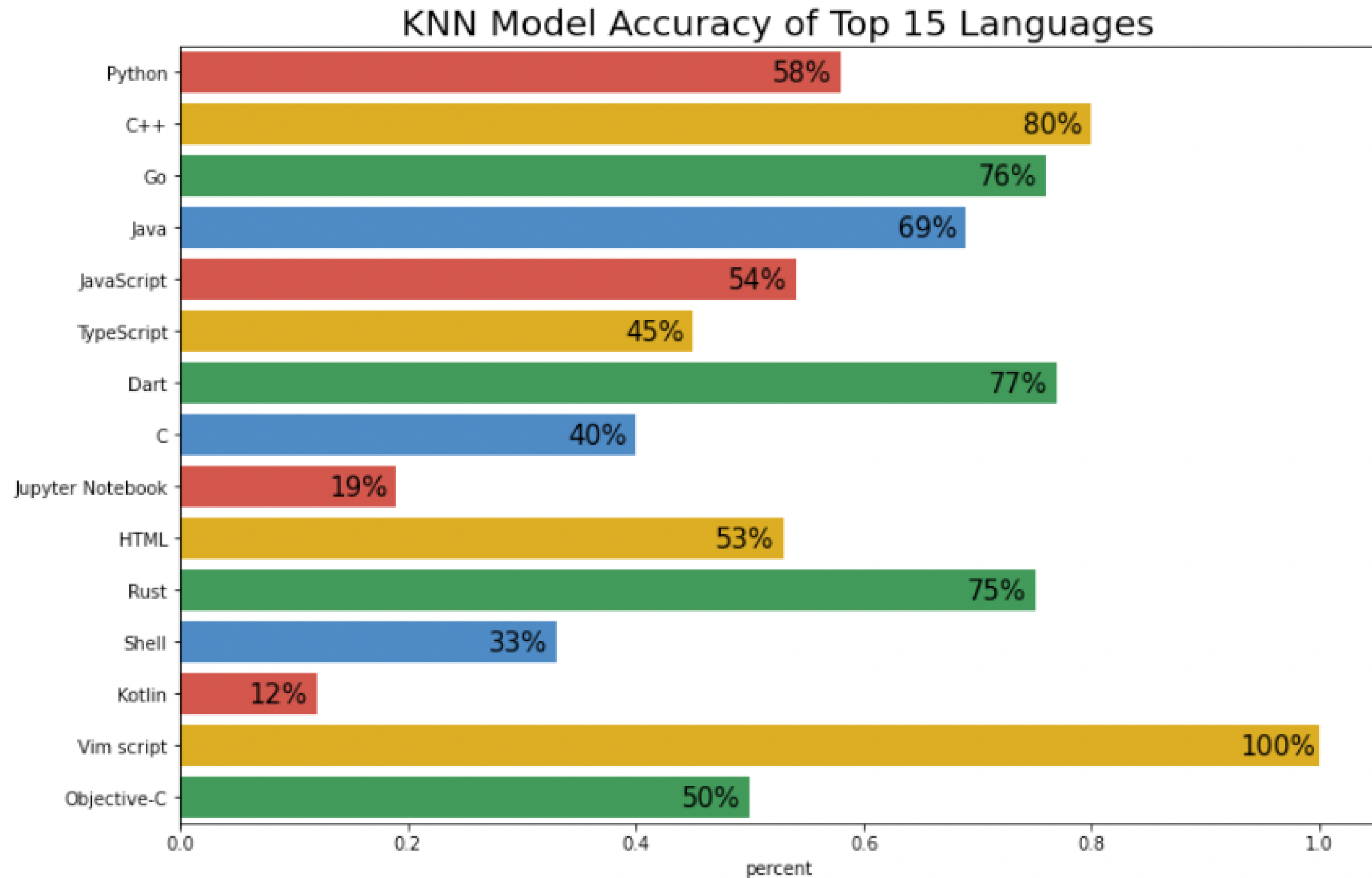
Model Performance

Use of my model **out performs the baseline by 28%** when predicting a coding language based on a repository's README file.

Data Analysis



Data Analysis



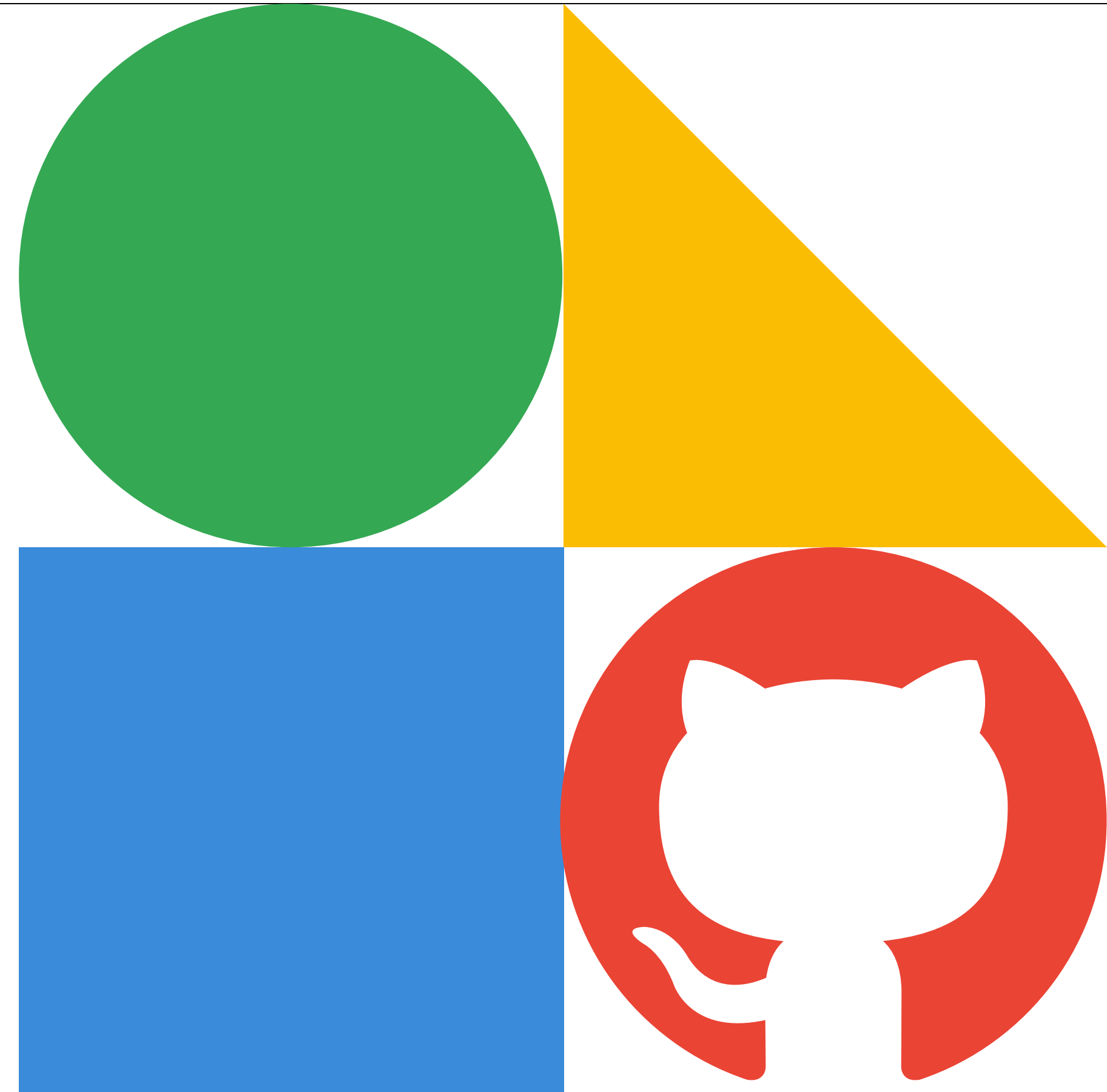
Conclusion

KNN predictive model out performed the baseline model by 28%

- This is likely due to the fact that Google Github repositories use 40 different programming languages.
- Most of which are verbs

Next Steps

Given more time, I'd like to filter my scraping for a limited amount of languages to identify keywords for each language.



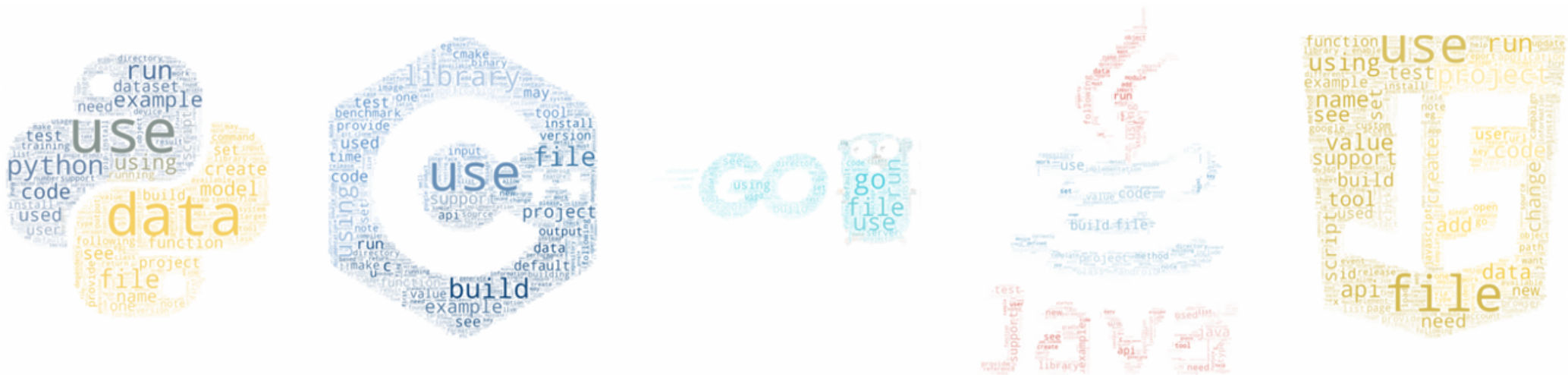
Appendix

DATA DICTIONARY

column_name	description	key	dtype
repo	Link suffix in <code>username / repo_name</code> formatting.		object
language	Primary coding language used in the repo.		object
readme_contents	String of text scraped from repo's README file.		object
readme_length	Length of <code>README</code> text		int64
clean_content	String of <code>README</code> text that has been cleaned by <code>clean()</code> function		object
cleaned_length	Length of <code>clean_content</code> text		int64

GITHUB

https://github.com/o0amandagomez0o/nlp_project-readme_prediction



The GitHub Octocat logo is a red silhouette of an octopus-like creature with a large head, small body, and a long, curved tail with a dotted line. It is centered within a large yellow circle. The background of the slide features a white area on the right and a colorful geometric pattern on the left consisting of blue and green triangles and a yellow circle.

Thank you!

Any questions or comments?