FORKING SOLUTIONS

MOST FORKED REPOSITORIES

PRESENTED BY
Derek Bixby
Alfred W. S. Pirovits Jr
Miatta Sinayoko

FORK YOUR REPO

Agenda

EXECUTIVE SUMMARY

FINDINGS

MODELING

CONCLUSION

EXECUTIVE SUMMARY

Goal: Use 100 "most forked" web scraped Github READMEs to create a categorization machine learning model to predict the programming language of a Git Repository based upon the language in that repository's README.

Common Words

- Title
- Desk
- Height
- Blob

Word Frequencies

- Dictionary
- All languages
- Vary in accordance

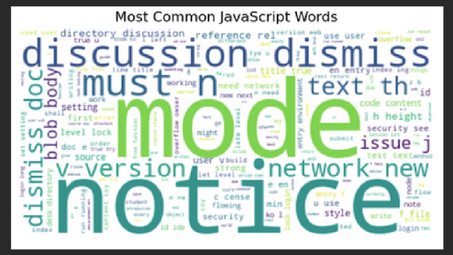
Python Words

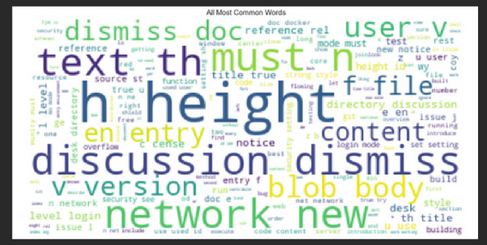
- Varied uniformly
- Distribution not random

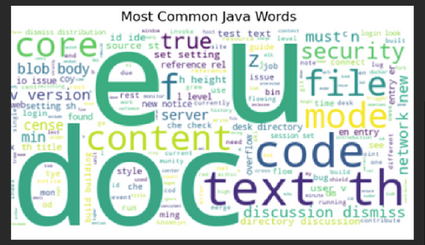
Least Frequent Words

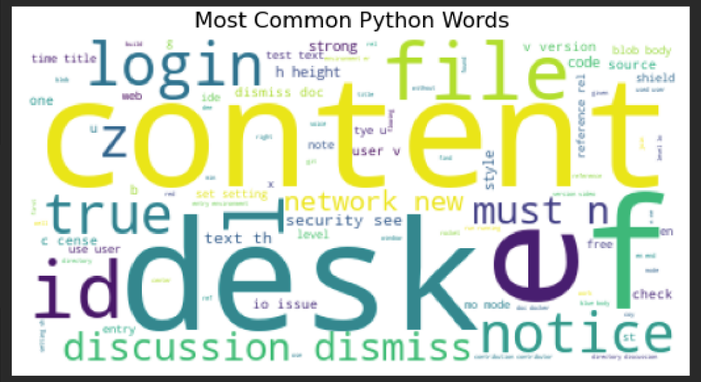
- Often partial
- Not unique

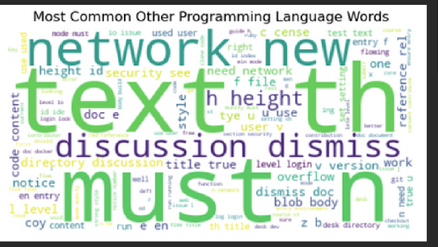
Common Words

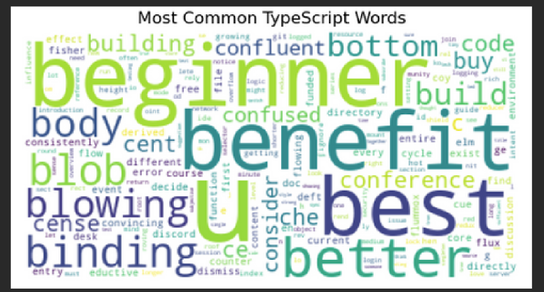






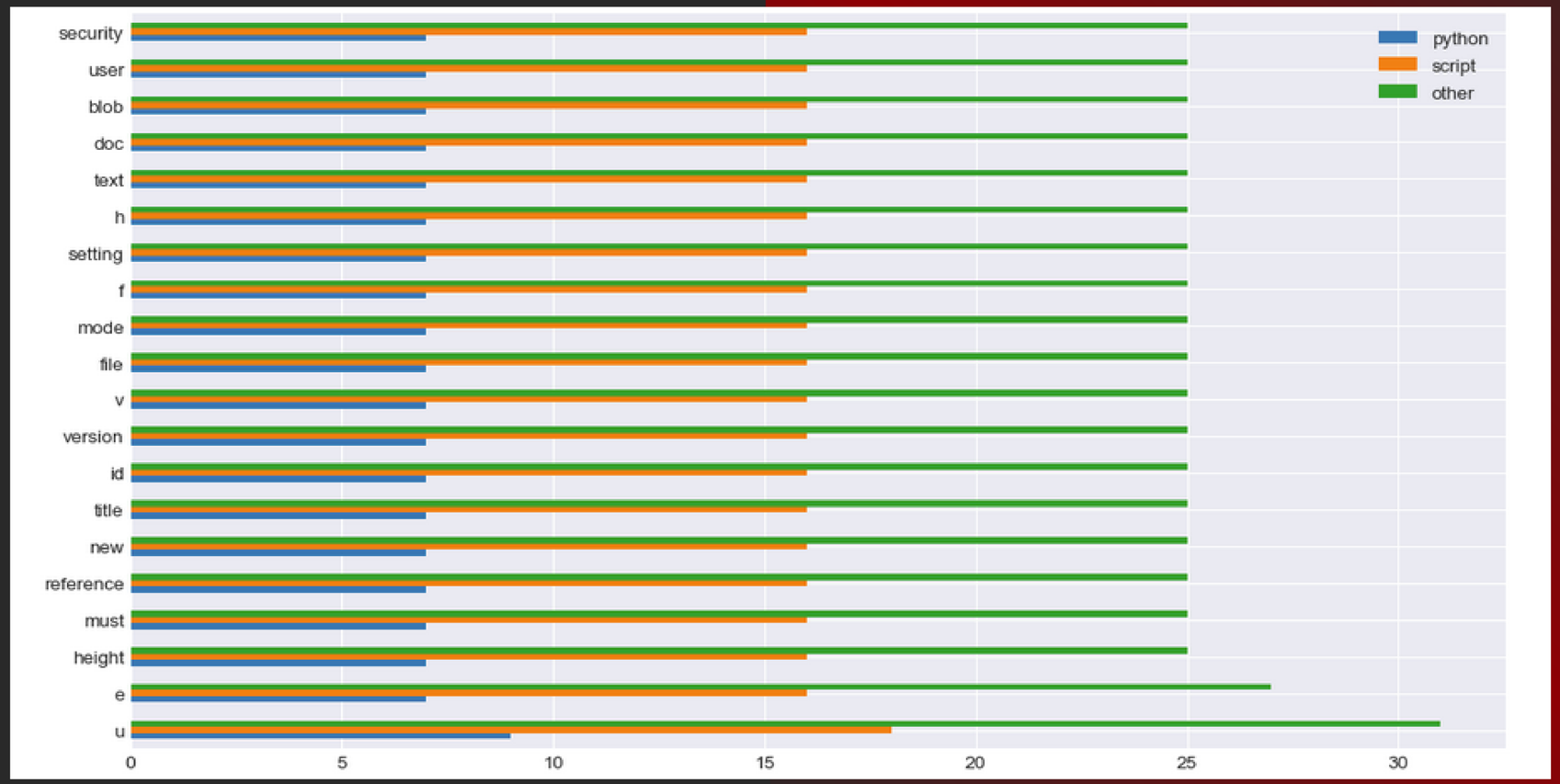






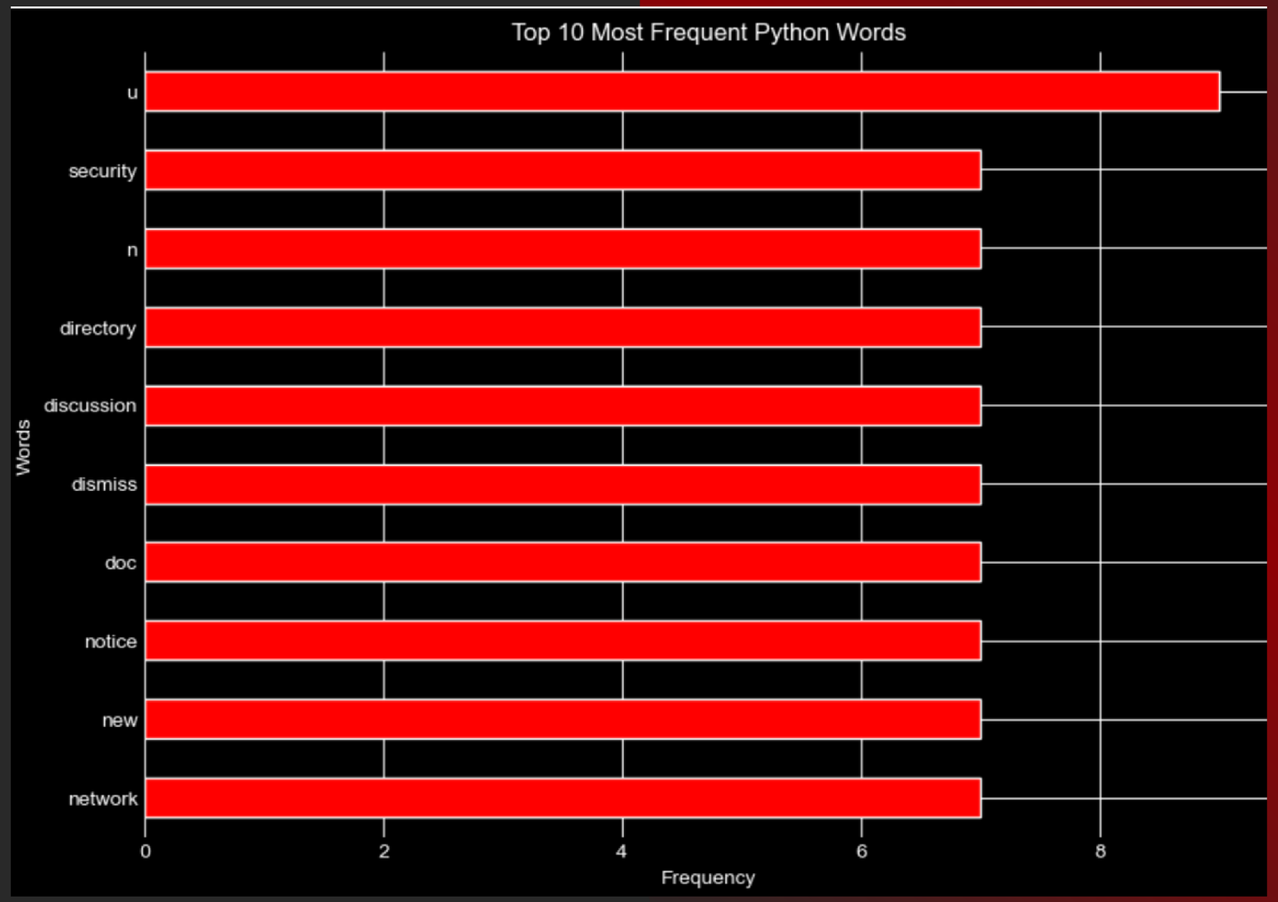
- Most common words
- Height, must, reference, desk
- Common across all languages

Word Frequencies



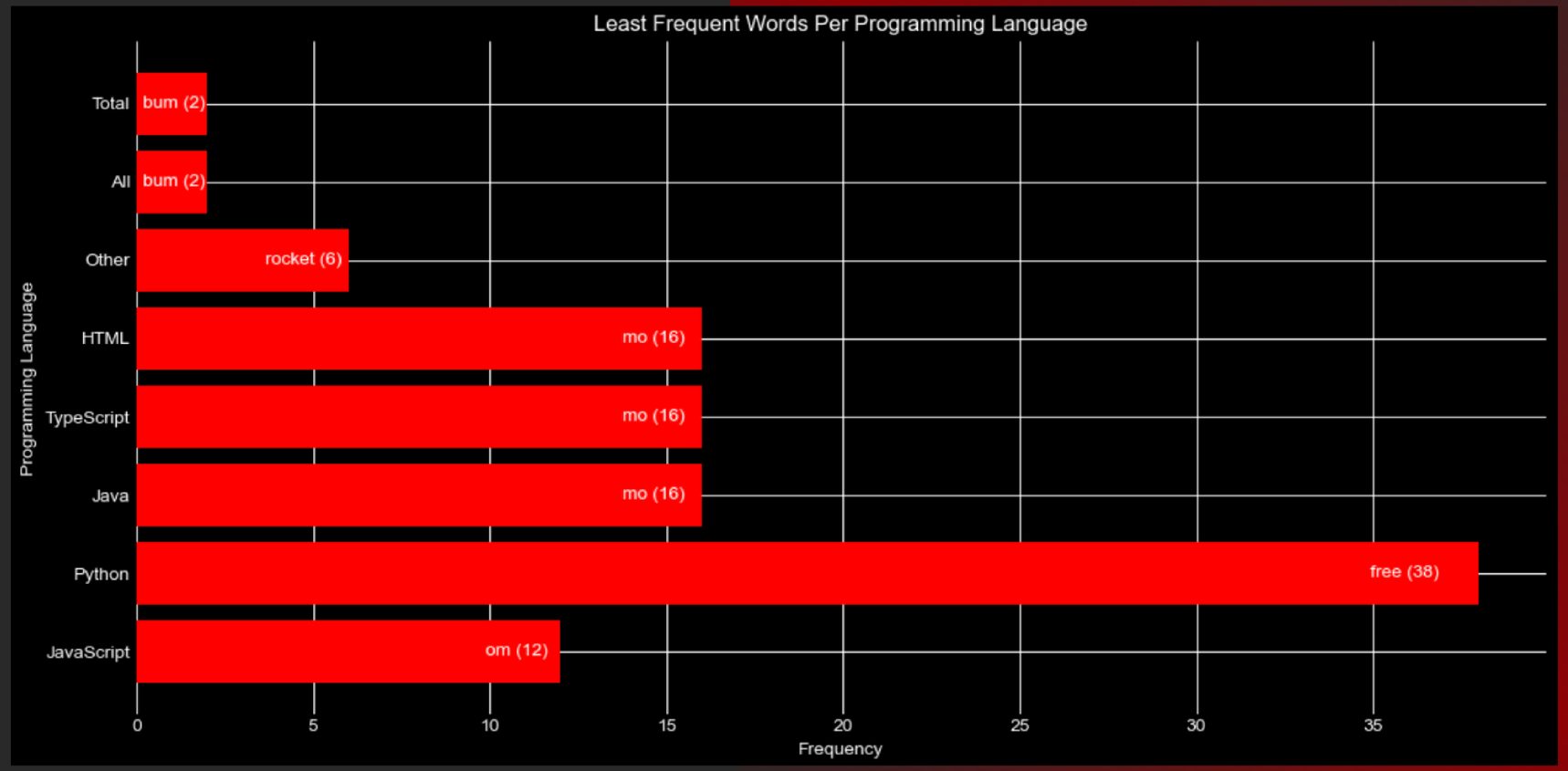
Word frequencies are an artifact of the cleaning process

Python Words



Words within each language are also far too evenly distributed.

Least Frequent Words



The least frequent words were often partial words and were not a useful measure for most languages

Modeling

Decision Tree Model

Baseline

Baseline was established based upon guessing "Other" programming language and had accuracy score of .42

Tested KNN and Decision Tree

Decision Tree performed best between the 2 models

Final Model

Decision Tree with a depth of 1 performed best, but did not beat baseline with a .4 accuracy test score

Performance

Models performed best when differentiating between "Python" and "Other" languages

CONCLUSION

Need a larger sample size

Too many unique words and languages had a small or single sample size

Need a different cleaning function

Cleaning resulted in too neat of a dataset and not enough unique words

Python

Focus on differentiating "Python" from "Other" languages

Contact

https://github.com/fork-your-repository/nlp-project.git

Links

Derek Bixby

Alfred W. S. Pirovits Jr

Miatta Sinayoko

Thank you

HAPPY FORKING