NLP Project

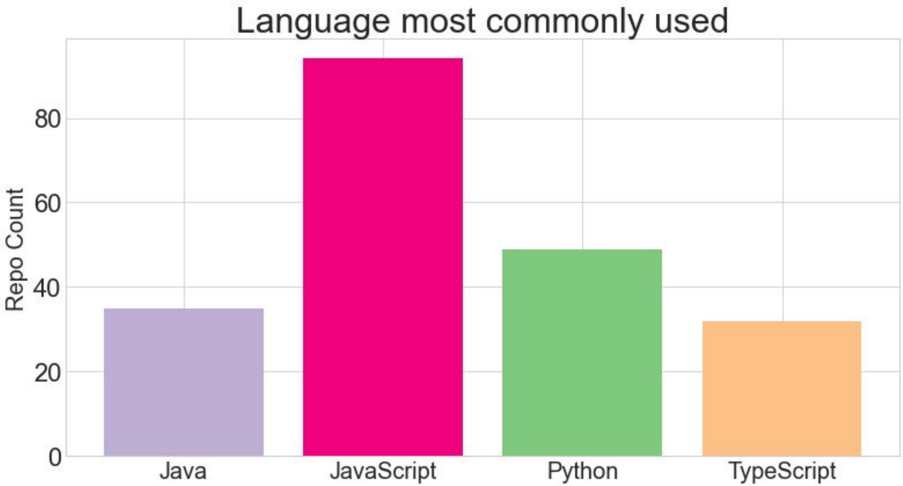
-Most Starred Github Repositories

By: Burton Barnes, Yuvia Cardenas, Yvette Ibarra

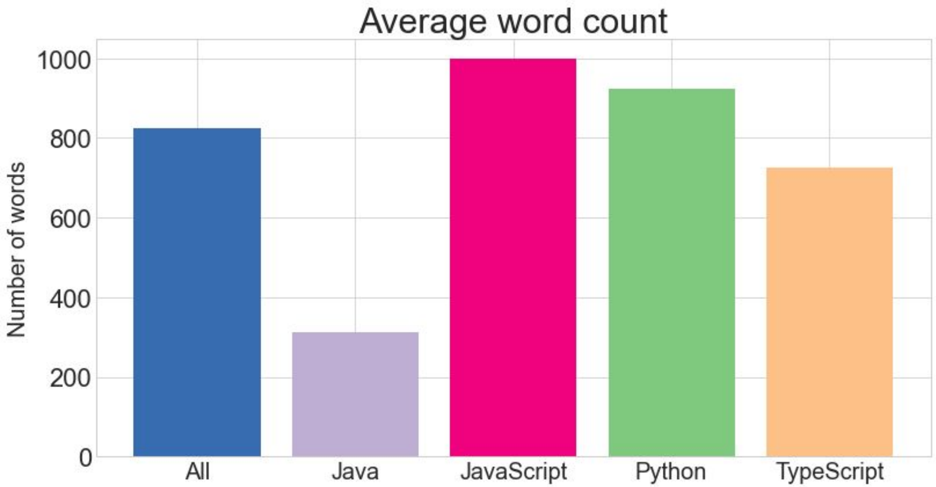
Executive Summary

- **The goal** is to first identify key words of the programming language and create a machine learning model that can effectively predict the primary language use of github repositories trending as "most starred".
- **Key Takeaways:**
  - **Acquire** data using BeautifulSoup to create a usable data frame from extracted text and primary language used within each ReadMe.
  - **Exploration** revealed JavaScript is most common language, Java word count is much lower, JavaScript has best sentiment score, most common word is yes.
  - **Modeling** revealed Decision Tree is best model with a 67% accuracy score.
- **Next steps and recommendations** include getting larger “text” datasets, hyperparameter tuning, and gradient boosting algorithms.

Which Language is the most used?

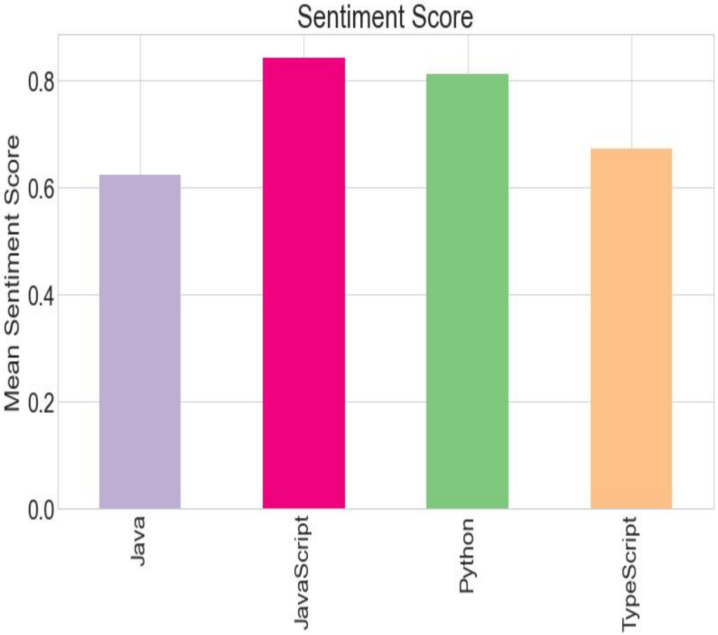


What is the average word count ?

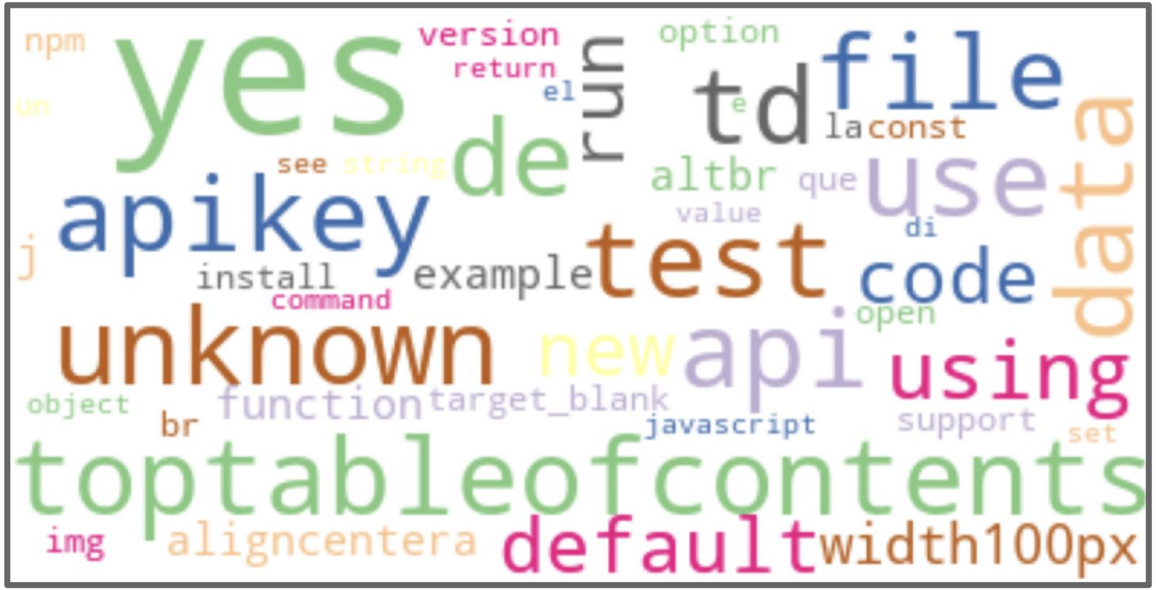


Explore

Is there a difference in sentiment by language?

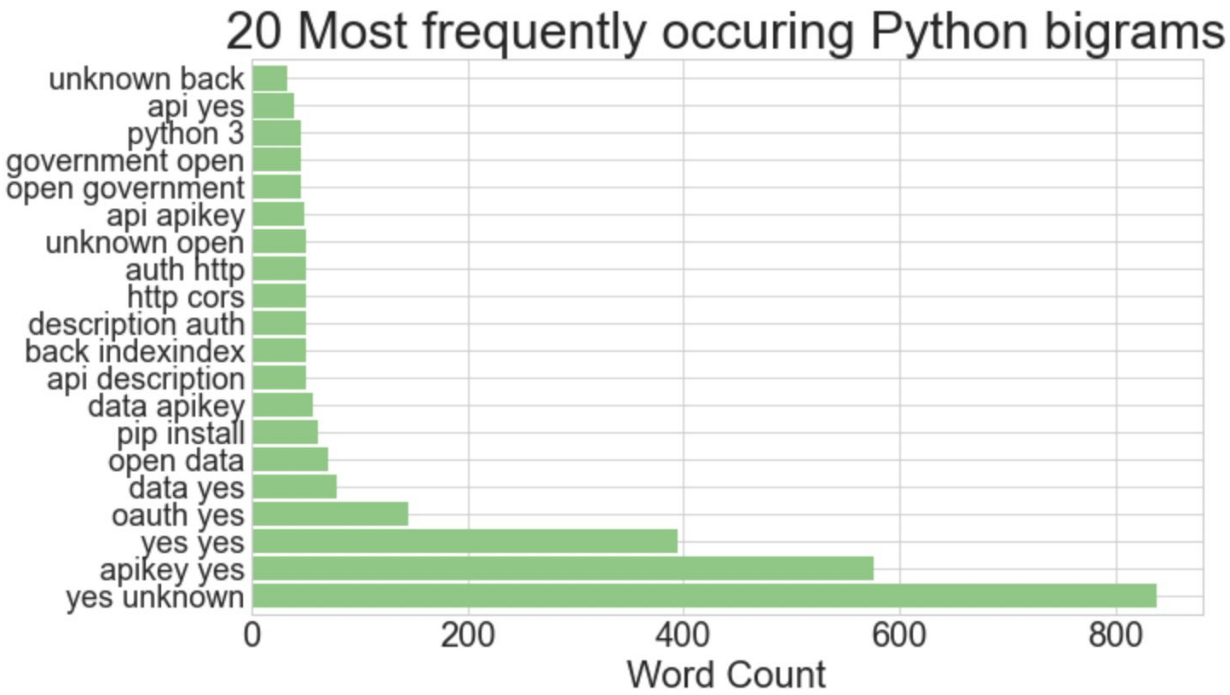


Top All Languages Words

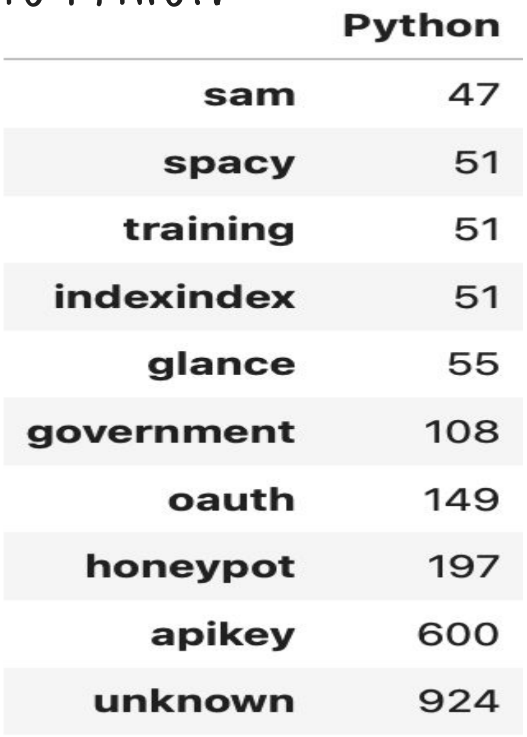


Explore

Top Python BiGrams



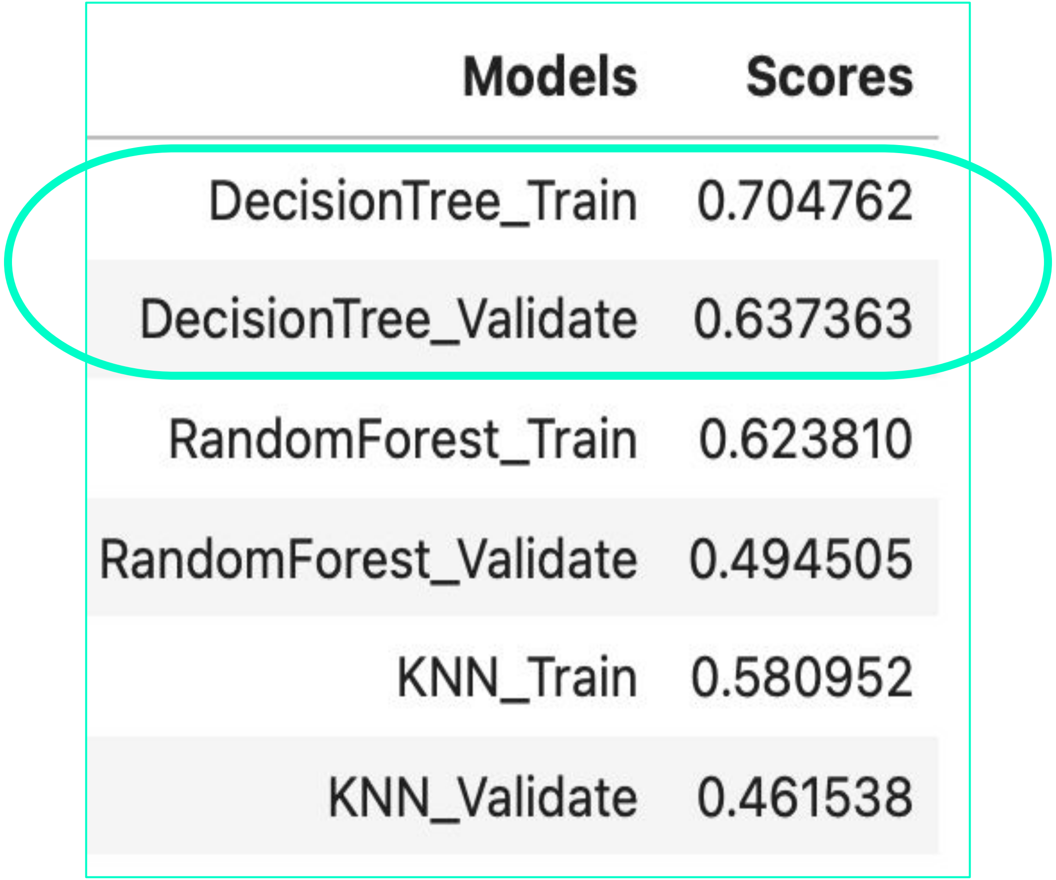
Top 10 Words Unique to Python



	Python
sam	47
spacy	51
training	51
indexindex	51
glance	55
government	108
oauth	149
honeypot	197
apikey	600
unknown	924

Explore

Decision Tree, Random Forest, & KNN



Models	Scores
DecisionTree_Train	0.704762
DecisionTree_Validate	0.637363
RandomForest_Train	0.623810
RandomForest_Validate	0.494505
KNN_Train	0.580952
KNN_Validate	0.461538

Baseline will be 45 % accuracy

- Since the biggest language in our data set is JavaScript which makes up 45% of the data

The accuracy of the Decision Tree model is above the baseline in both train and validate.

The accuracy of the KNN is slightly above the baseline in both train and validate.

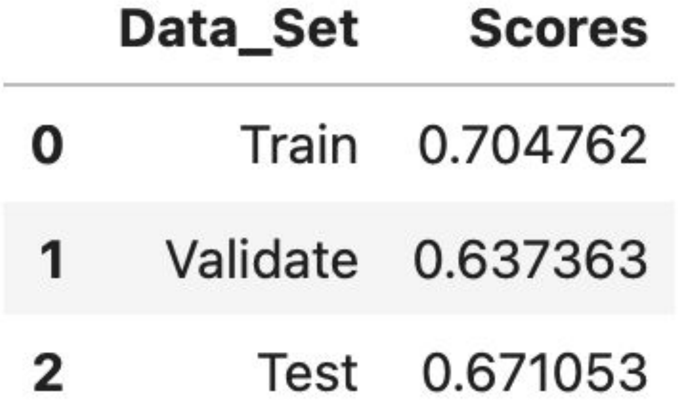
The accuracy of the Random Forest model is above the baseline in both train and slightly above in validate.

Model

Conclusions

- **Expl. Conclusions:** JavaScript is most common language, Java word count is much lower, JavaScript has best sentiment score, most common word is yes
- **Model Conclusions:** Final model beat the baseline of 45% by 22% in terms of accuracy score.
- **Next Steps:** Consider getting larger “text” datasets, hyperparameter tuning, gradient boosting algorithms

Decision Tree on TEST



	Data_Set	Scores
0	Train	0.704762
1	Validate	0.637363
2	Test	0.671053

Conclusion