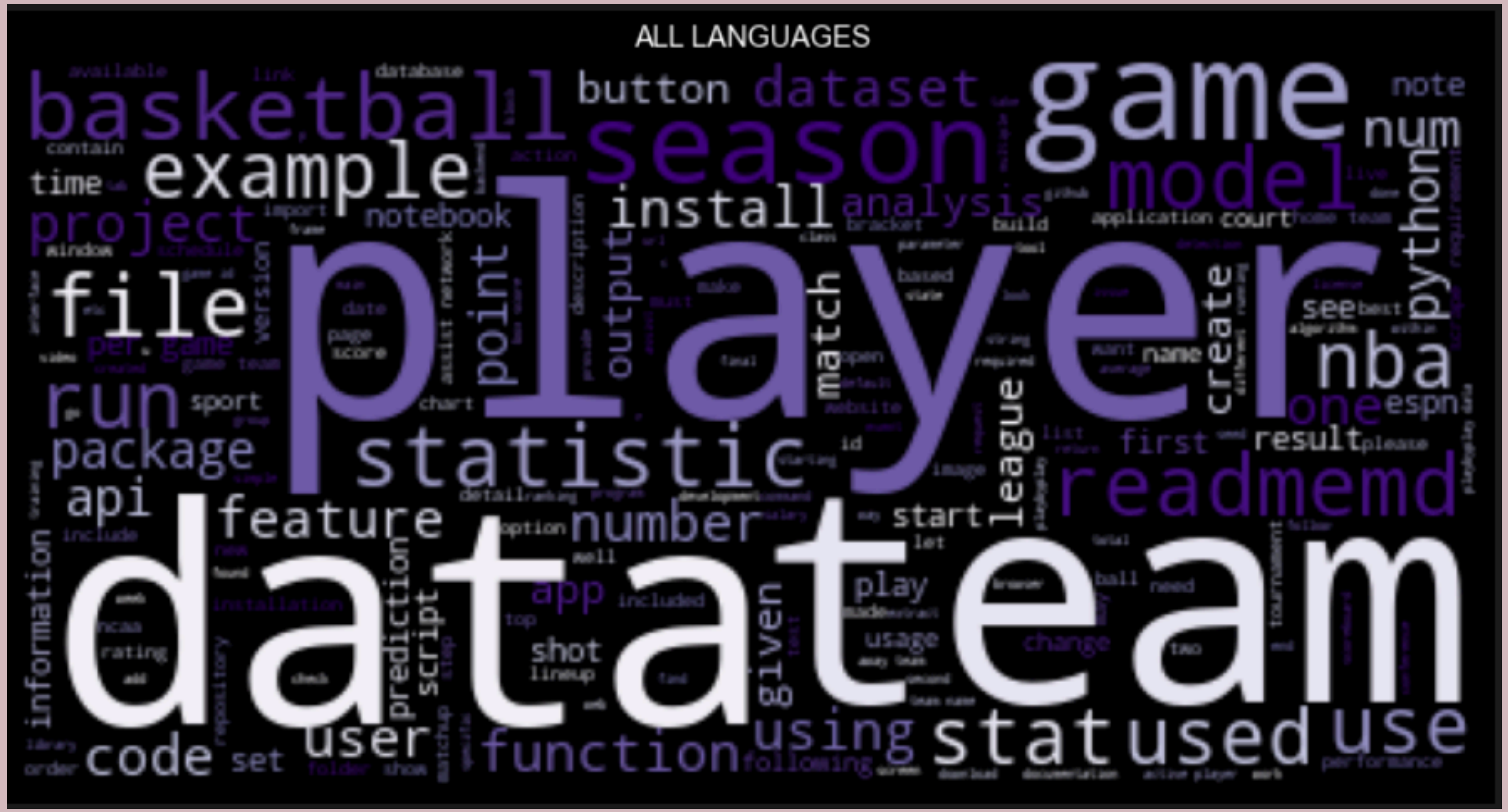
Basketball and NLP

A project by Vincent Banuelos and J. Vincent Shorter

Executive Summary

- Python the most popular language
- README's created in wide varieties of spoken and coding languages
- Most popular words stem from basketball as opposed to coding
- Creation of bigrams was huge help to modeling process

All Words



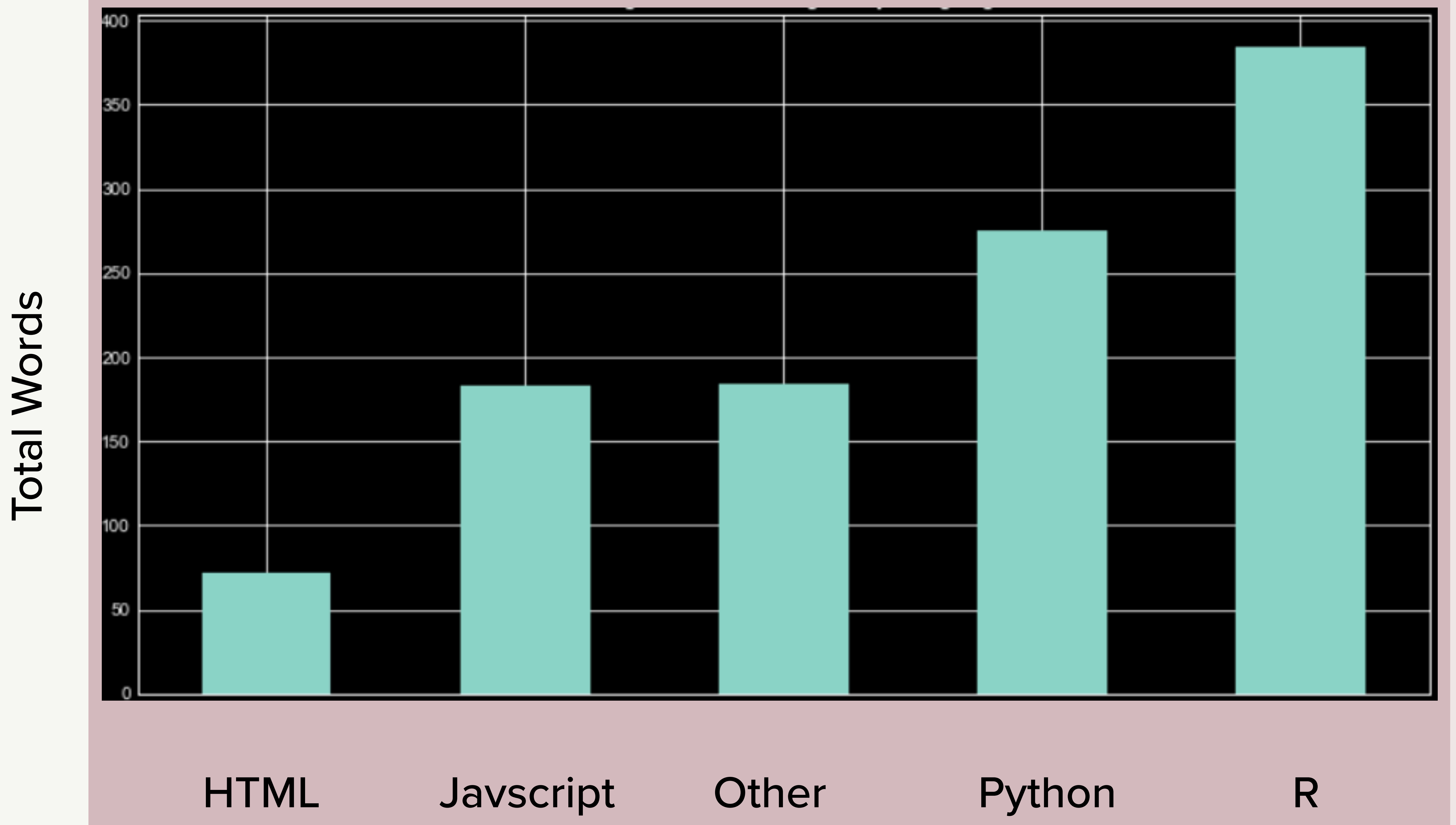
Total Word Count: 49395 (8932 Unique)

Average README Word Count: 276

Most Used Word: Team

Number of Repositories: 179

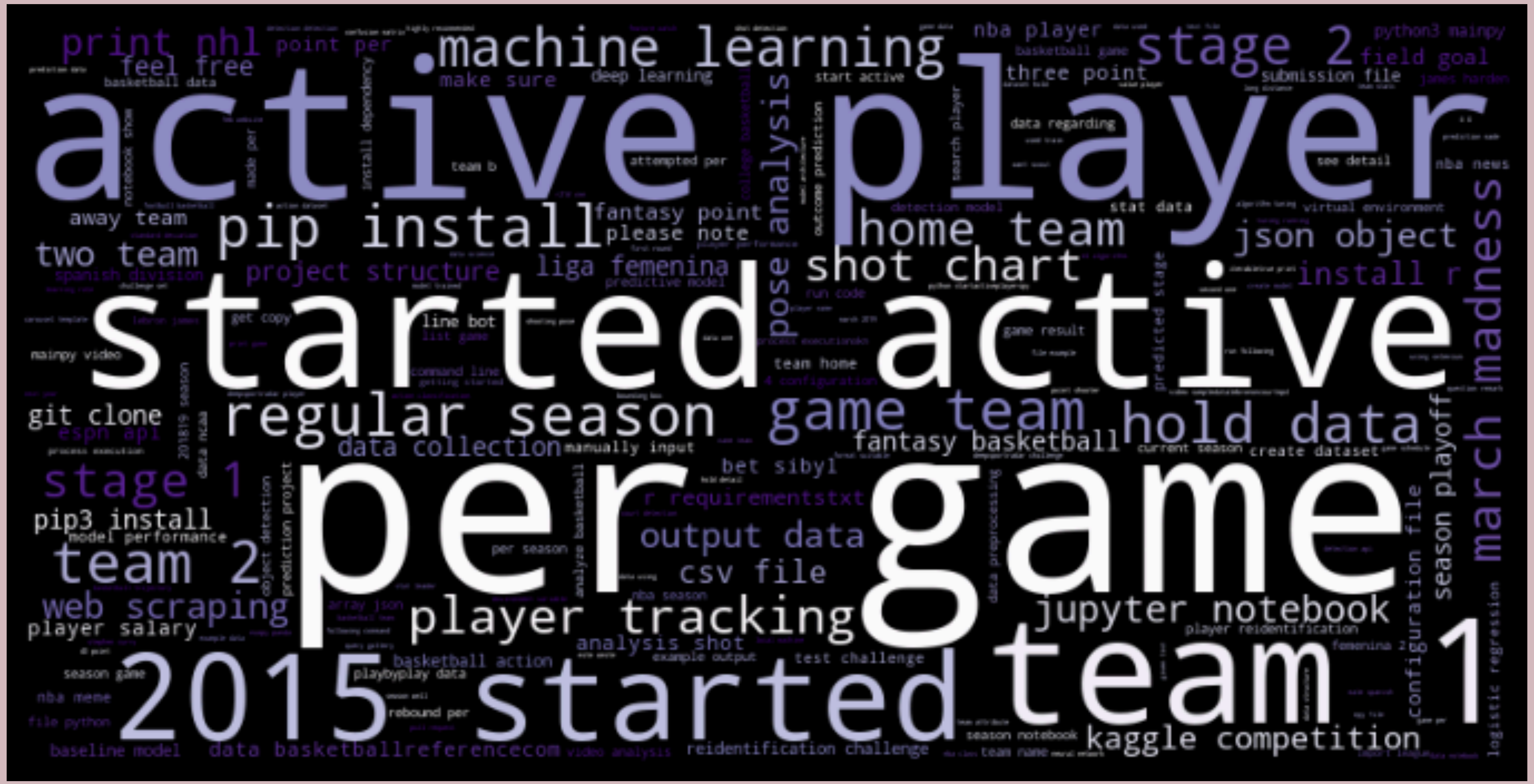
Average README Length By Words



- READMEs in R are the most verbose
  - Assist/Network most popular bigram in that R
- Top 5 R Words:
  - Team: 208
  - Data: 150
  - Game: 137
  - Season: 107
  - Player: 93

Python is most popular language

40% of Repositories tested



Unique Bigrams Across Data: 14824

Most Used Unique Bigram: Per/Game

Avg README Word Count: 249

Number of Repositories: 73

Top Performing Model

DTC_0

Type:

Decision Tree Classifier

Features:

Lemmatized Words

Word Bigrams

Word Count

Parameters:

Depth - 5

Accuracy(Score):

78%

Conclusion

DTC Models were consistent performers

Python by far the most popular language

Most README’s used large amounts of basketball related language

An affirmative next step would be to expand the number of README’s pulled in and target less popular languages

With more time we would like to work on sentiment analysis with +/- in the direction of basketball or coding

Thank You

This has been a Double Vincent Production