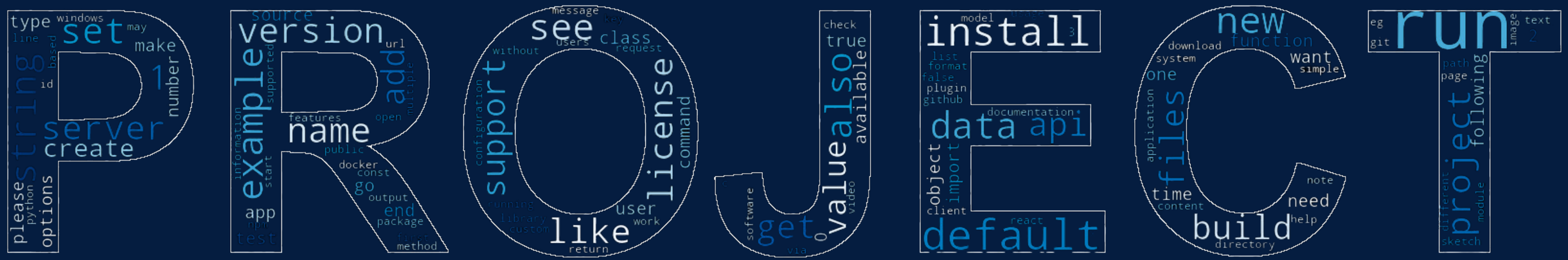
CODE LANGUAGE PREDICTION

FROM GITHUB READMES

ZACHARIA SCHMITZ
ASWATHY RADHA
JOSHUA CLICK

24 October, 2023



AGENDA

01 Acquisition

02 Preparation

03 Features

04 Model

05 Function

Predict a repository language from the README

Objectives

Predict a project's language only from the README

Big Idea

13%
Baseline

66%
TF-IDF
Logistic Regression

Findings

Low Quality READMES

Cache Frequently

pip
css

data
android

Enhancements

More READMEs

Tune TF-IDF

Language Specific Feature Engineering

Objectives

Predict a project's language only from the README

Big Idea

13%
Baseline

66%
TF-IDF
Logistic Regression

Findings

Low Quality READMES

Cache Frequently

pip
css

data
android

Enhancements

More READMEs

Tune TF-IDF

Language Specific Feature Engineering

Objectives

Predict a project's language only from the README

Big Idea

13%
Baseline

66%
TF-IDF
Logistic Regression

Findings

Low Quality READMES

pip
css

Cache Frequently

data
android

Enhancements

More READMEs

Tune TF-IDF

Language Specific Feature Engineering

Objectives

Predict a project's language only from the README

Big Idea

13%
Baseline

66%
TF-IDF
Logistic Regression

Findings

Low Quality READMES

Cache Frequently

pip
css

data
android

Enhancements

More READMEs

Tune TF-IDF

Language Specific Feature Engineering

PLAN & ACQUIRE

PREPARE
EXPLORE
MODEL
DELIVER

Chosen Languages

typescript, go, ruby, c++,
html, java, python, javascript

100 Repos Per Language

800 READMEs

(Error 429)

Web Scraping for
Repository List

GitHub API for
README extraction

100 with sleeps

Cached and combined

PLAN & ACQUIRE

PREPARE
EXPLORE
MODEL
DELIVER

Chosen Languages

typescript, go, ruby, c++,
html, java, python, javascript

100 Repos Per Language

800 READMEs

(Error 429)

Web Scraping for Repository List

GitHub API for README extraction

100 with sleeps

Cached and combined

PLAN & ACQUIRE

PREPARE
EXPLORE
MODEL
DELIVER

Chosen Languages

typescript, go, ruby, c++,
html, java, python, javascript

100 Repos Per Language

800 READMEs

(Error 429)

Web Scraping for Repository List

GitHub API for
README extraction

100 with sleeps

Cached and combined

PLAN & ACQUIRE

PREPARE
EXPLORE
MODEL
DELIVER

Chosen Languages

typescript, go, ruby, c++,
html, java, python, javascript

100 Repos Per Language

800 READMEs

(Error 429)

Web Scraping for
Repository List

GitHub API for README extraction

100 with sleeps
Cached and combined

PLAN & ACQUIRE

PREPARE

EXPLORE
MODEL
DELIVER

Processing Function

What we removed / changed

New Lines
\n

Hyphens
(replaced with space)

Punctuation

URLS
http\S+

Extra White Space

Upper Casing

HTML Tags
<.*?>

Stop Words
use, using, used, code, codes, file

Tokenized

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Initial Exploration:

- Some READMEs <10 Words
- Removed <50
- 86 - 93 Samples

Javascript
92

Typescript
92

Python
93

Go
92

Java
90

Ruby
91

C++
86

HTML
67

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Initial Exploration:

- Some READMEs <10 Words
- Removed <50
- 86 - 93 Samples

Javascript
92

Typescript
92

Python
93

Go
92

Java
90

Ruby
91

C++
86

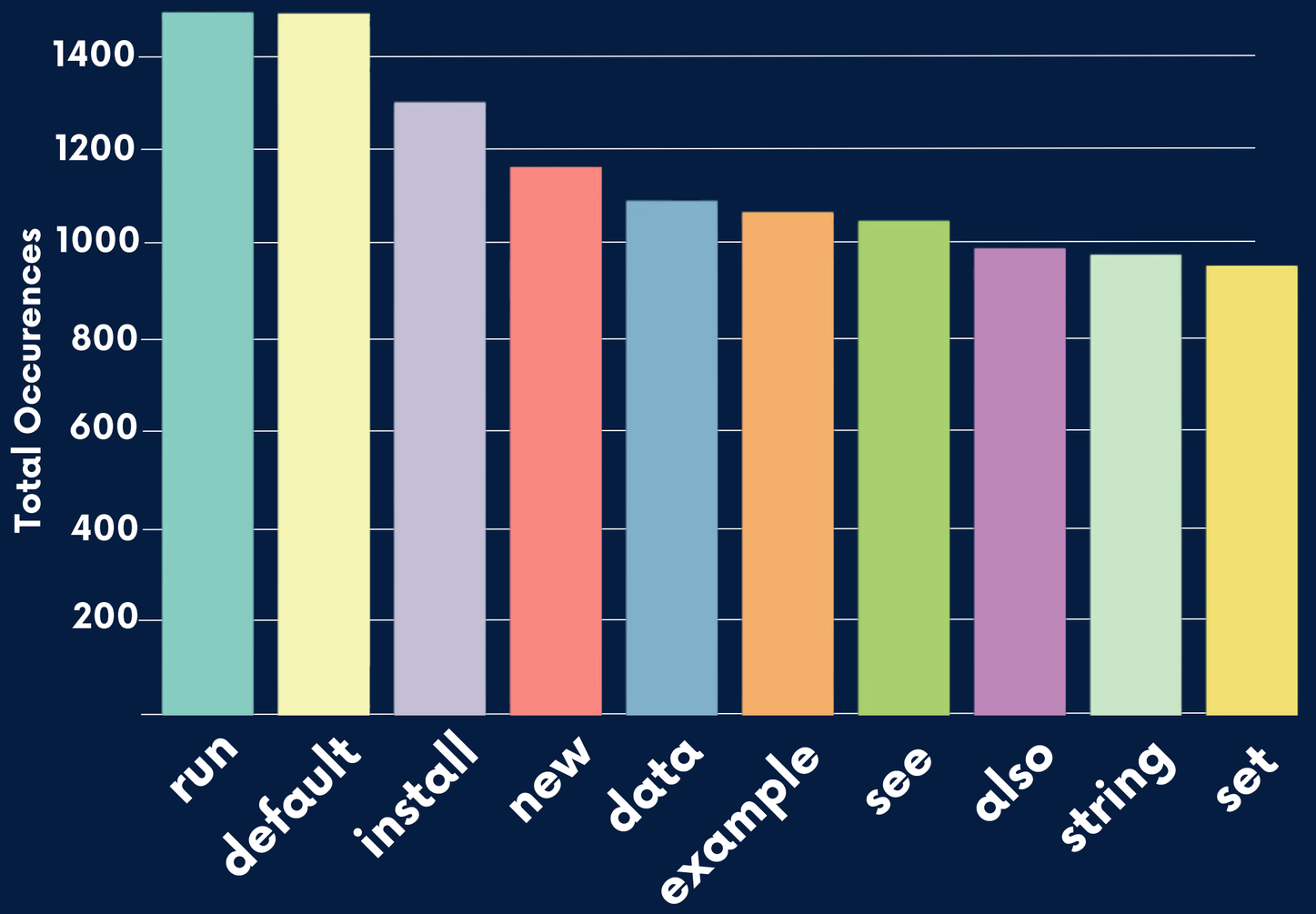
HTML
67

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

10 Most Common Words

(Entire Corpus)

Pinpointing Languages?

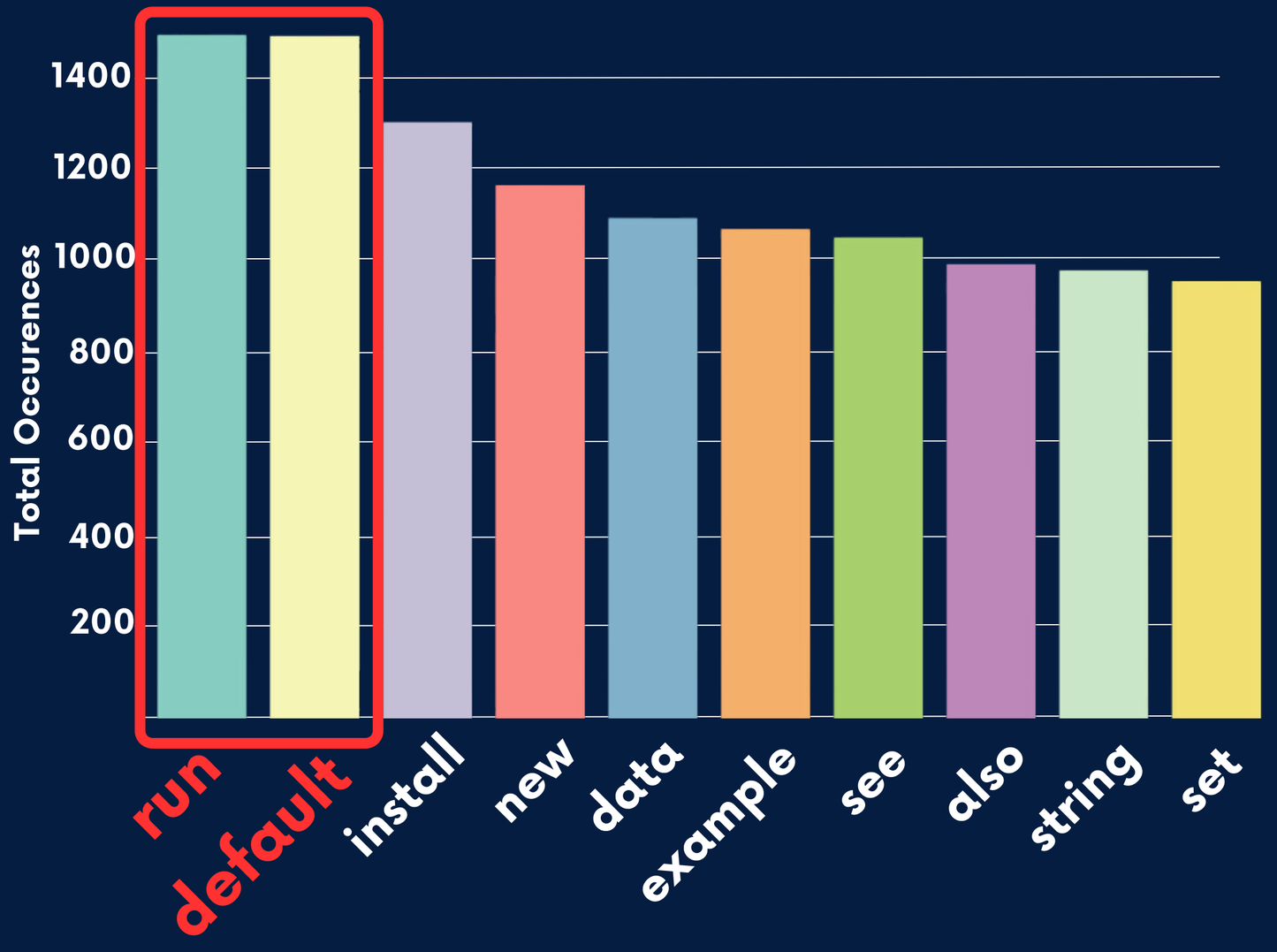


PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

10 Most Common Words

(Entire Corpus)

Pinpointing Languages?



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

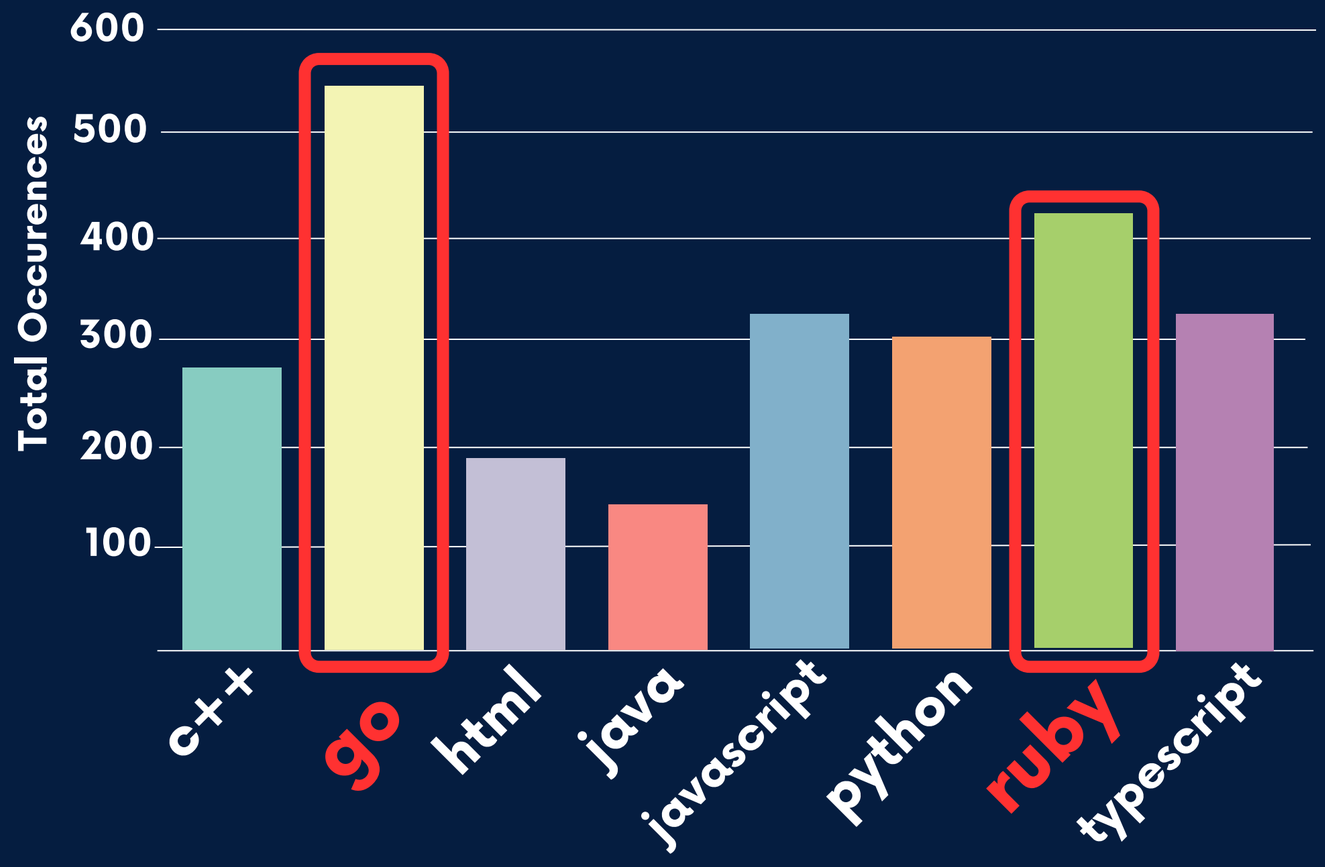
‘run’ Occurence
(per language)

Pinpointing Languages?

Run
Scripting

New
Object Oriented, common in Ruby

Data
(Python)



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

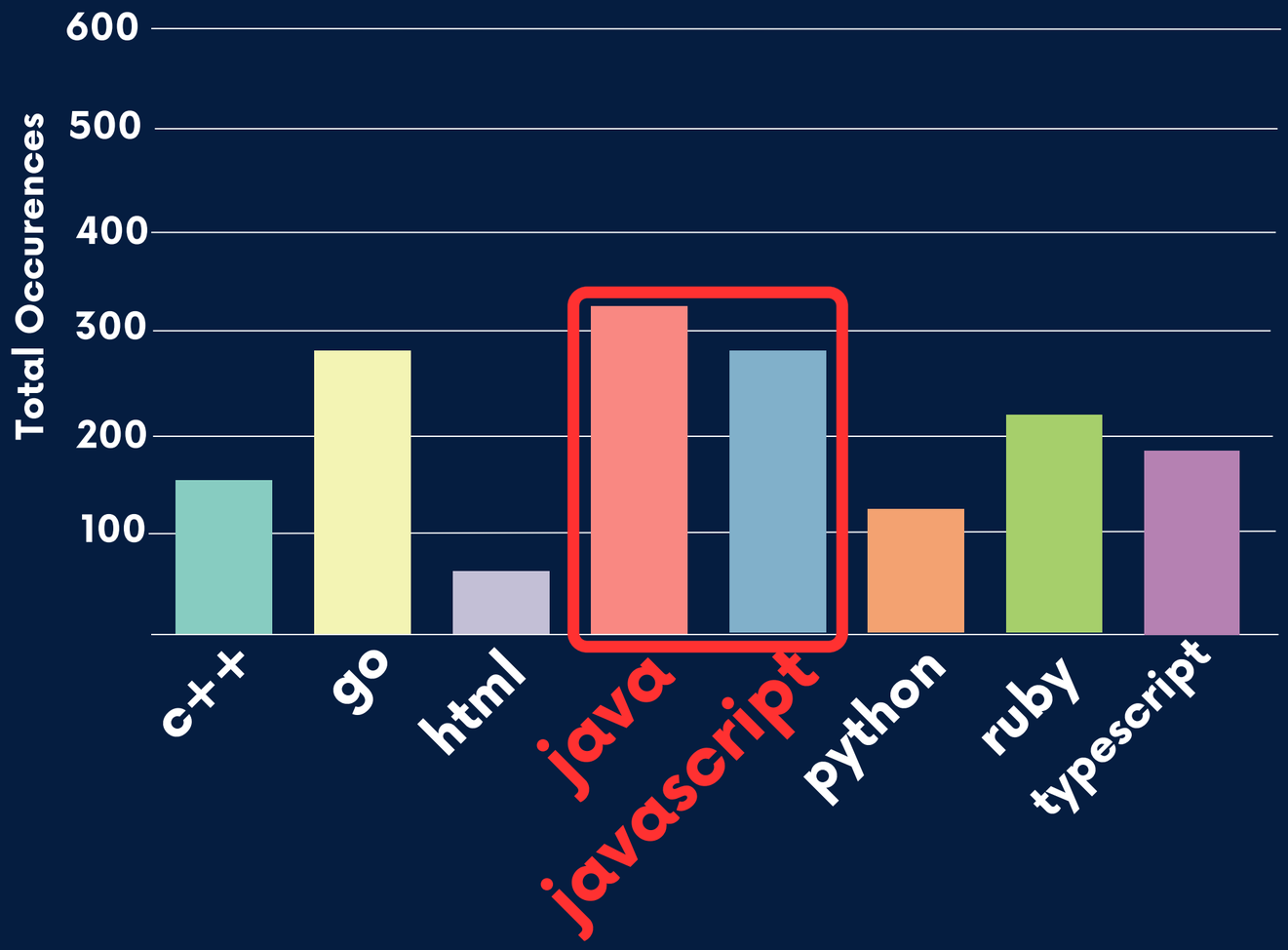
Pinpointing Languages?

Run
All Scripting

New
Object Oriented, common in Ruby

Data
(Python)

‘new’ Occurence
(per language)



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

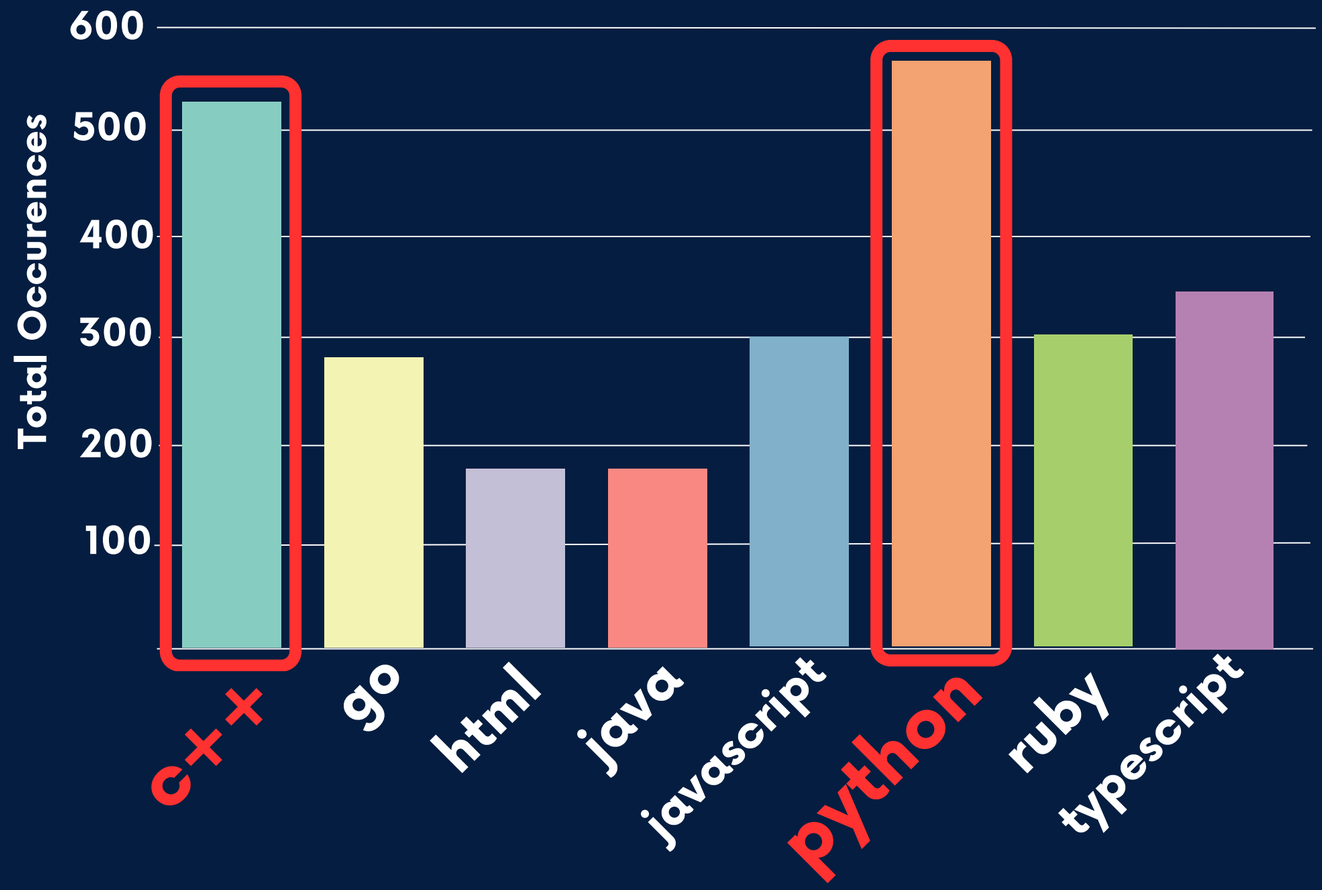
‘data’ Occurence
(per language)

Pinpointing Languages?

Run
All Scripting

New
Object Oriented, common in Ruby

Data
(Python)



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Unique Identifier Words

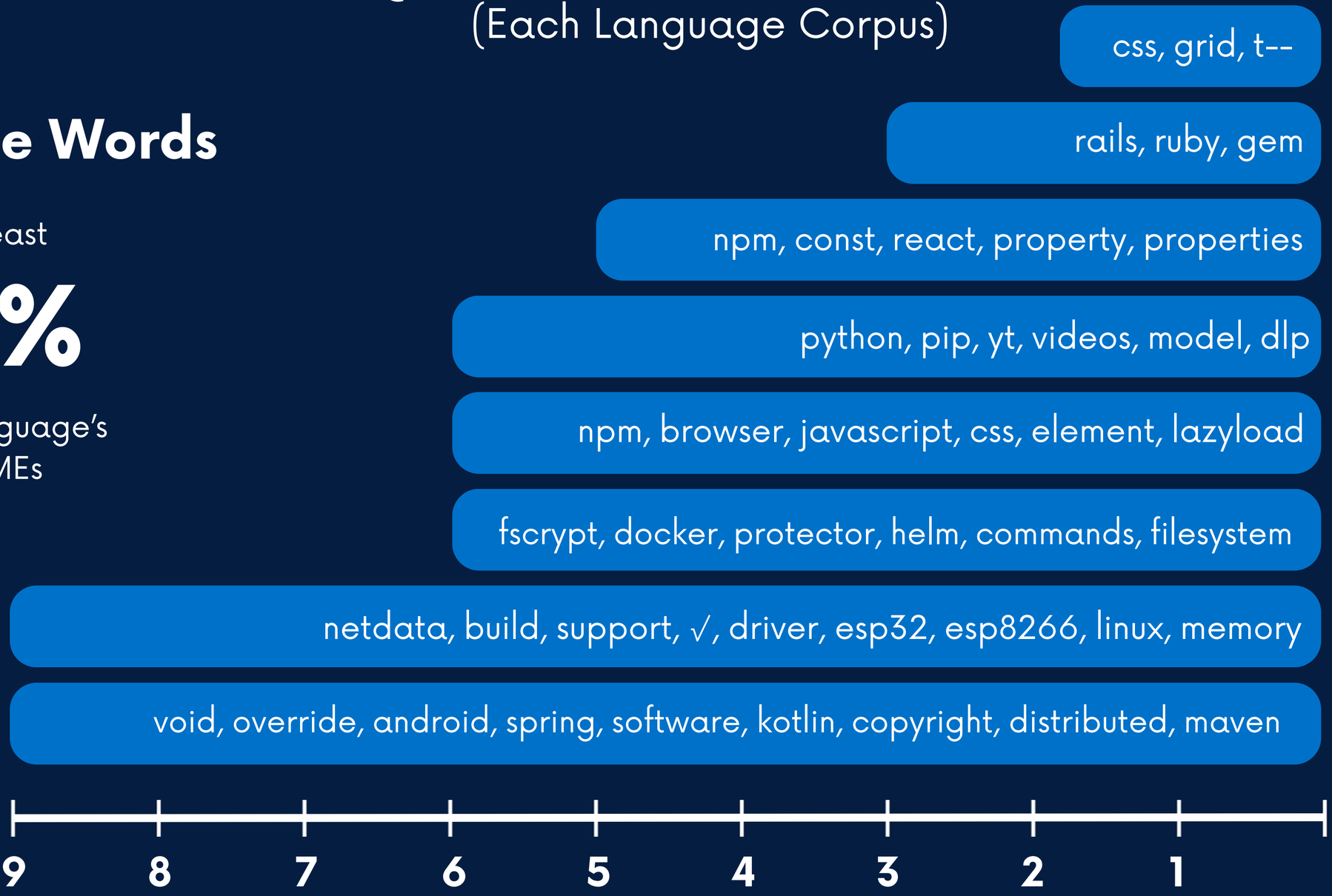
(Each Language Corpus)

Top Unique Words

In at least

25%

of that language's READMEs



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Unique Identifier Words

(Each Language Corpus)

Top Unique Words

In at least

50%

of that language's READMEs

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Unique Identifier Words

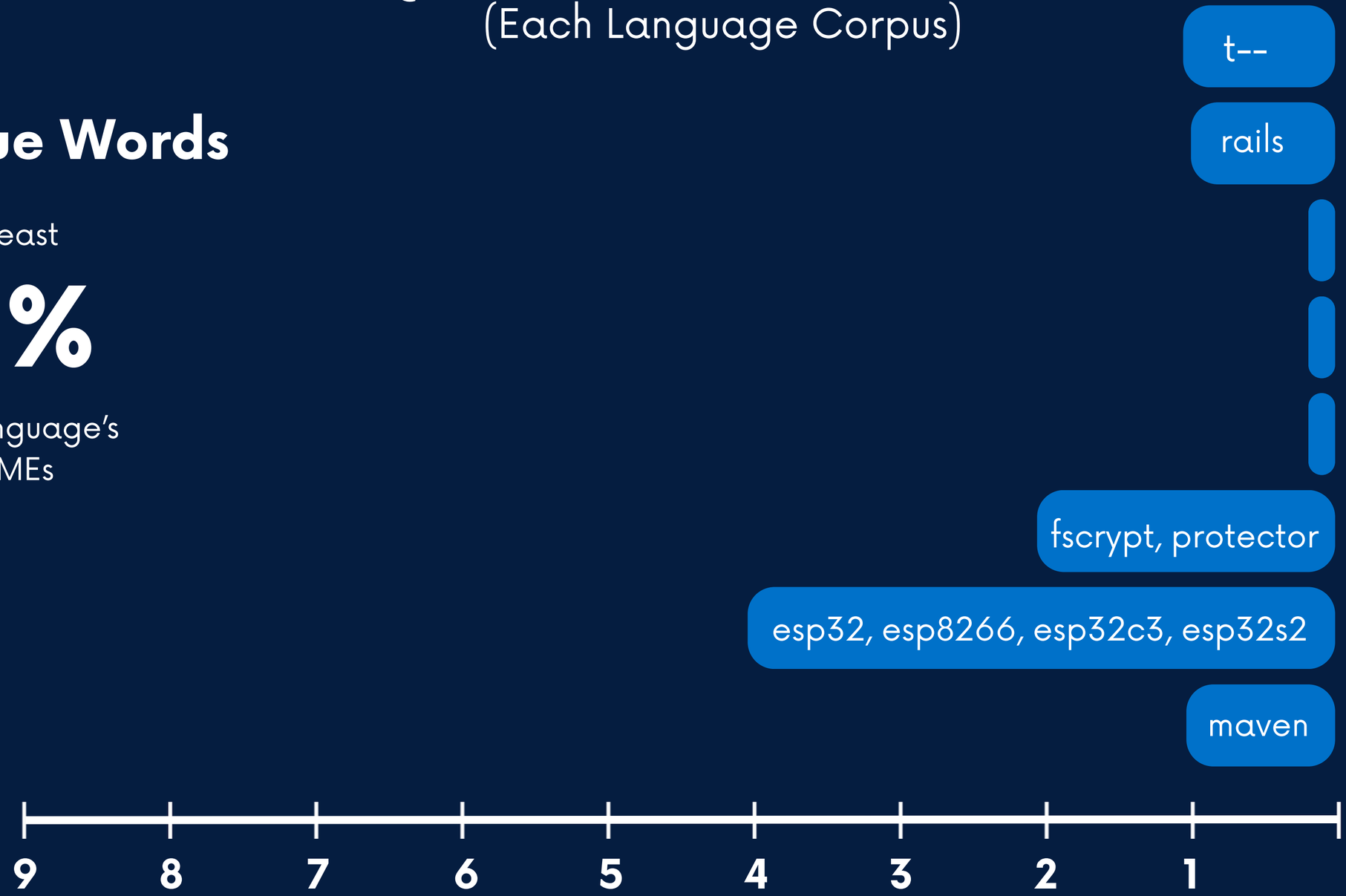
(Each Language Corpus)

Top Unique Words

In at least

75%

of that language's READMEs



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Unique Identifier Words

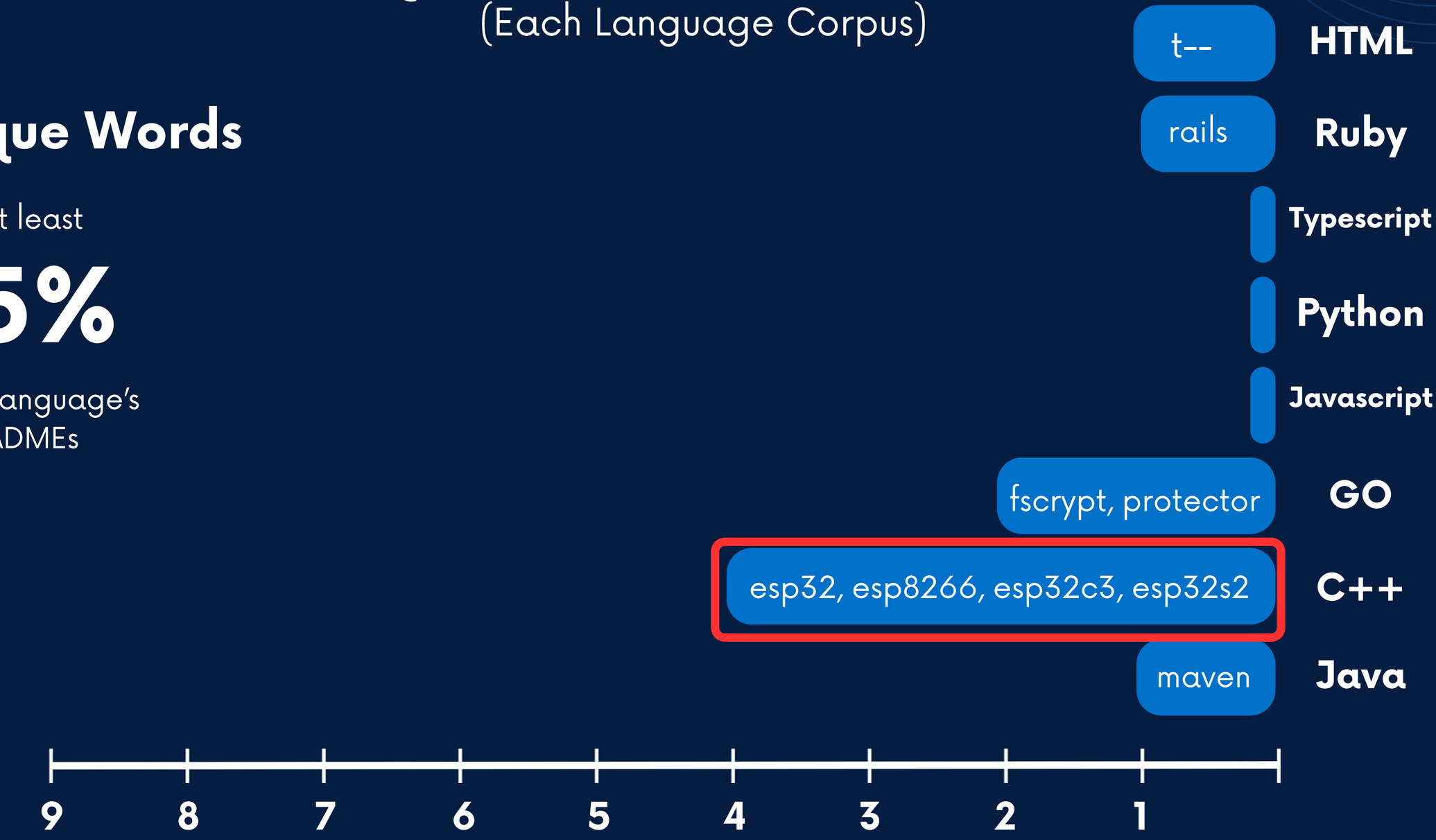
(Each Language Corpus)

Top Unique Words

In at least

75%

of that language's READMEs



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

Utilizing GridSearch

13%
(baseline)

```
TfidVectorizer(
max_df=250,
max_features=500,
min_df=25,
ngram_range=(1, 2))
```

```
LogisticRegression(
C=1,
penalty='l2')
```

87%
(train)

66%
(test)

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

FUTURE MODEL IMPROVEMENTS

Feature Engineering

!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)

Ensemble Methods

Voting Classifiers

More Samples

Quality Samples

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

FUTURE MODEL IMPROVEMENTS

Feature Engineering

!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)

Ensemble Methods

Voting Classifiers

More Samples

Quality Samples

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

FUTURE MODEL IMPROVEMENTS

Feature Engineering

!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)

Ensemble Methods

Voting Classifiers

More Samples

Quality Samples

PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL
DELIVER

LESSONS LEARNED

Implement Sleeps

Break It Up

API If Possible

robots.txt

Github: /AswathyRadha100
LinkedIn: /aswathy-radha83

Github: /Zacharia-Schmitz
LinkedIn: /zschmitz

Github: /Joshua-Click
LinkedIn: /joshua-r-click