

CODE LANGUAGE PREDICTION

FROM GITHUB READMES



AGENDA

01

Acquisition

02

Preparation

03

Features

04

Model

05

Function

Predict a repository language from the README

Predict a project's
language only from
the README

Objectives

Low Quality
READMEs

pip
css

Cache
Frequently

data
android

Big Idea

Findings

Enhancements

13%
Baseline

66%
TF-IDF
Logistic Regression

More READMEs

Tune TF-IDF

Language Specific
Feature Engineering

Predict a project's language only from the README

Objectives

13%
Baseline

Low Quality
READMEs

pip
css

Cache
Frequently

data
android

Big Idea

66%
TF-IDF
Logistic Regression

Findings

More READMEs

Tune TF-IDF

Enhancements

Language Specific
Feature Engineering

Predict a project's language only from the README

Objectives

13%
Baseline

Low Quality
READMEs

pip
css

Cache
Frequently

data
android

Big Idea

66%
TF-IDF
Logistic Regression

Findings

More READMEs

Tune TF-IDF

Enhancements

Language Specific
Feature Engineering

Predict a project's language only from the README

Objectives

Low Quality READMEs

pip
css

Cache Frequently

data
android

Big Idea

Findings

Enhancements

13%
Baseline

66%
TF-IDF
Logistic Regression

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**

Web Scraping for Repository List

100 Repos Per Language

800 READMEs
(Error 429)

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**

Web Scraping for Repository List

100 Repos Per Language

**800 READMEs
(Error 429)**

**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**

Web Scraping for Repository List

100 Repos Per Language

**800 READMEs
(Error 429)**

**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**

Web Scraping for Repository List

100 Repos Per Language

**800 READMEs
(Error 429)**

**GitHub API for
README extraction**

**100 with sleeps
Cached and combined**

Processing Function

What we removed / changed

New Lines

\n

URLS

http\S+

HTML Tags

<.*?>

Hyphens

(replaced with space)

Extra White Space

Punctuation

Upper Casing

Stop Words

use, using, used, code, codes, file

Tokenized

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

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

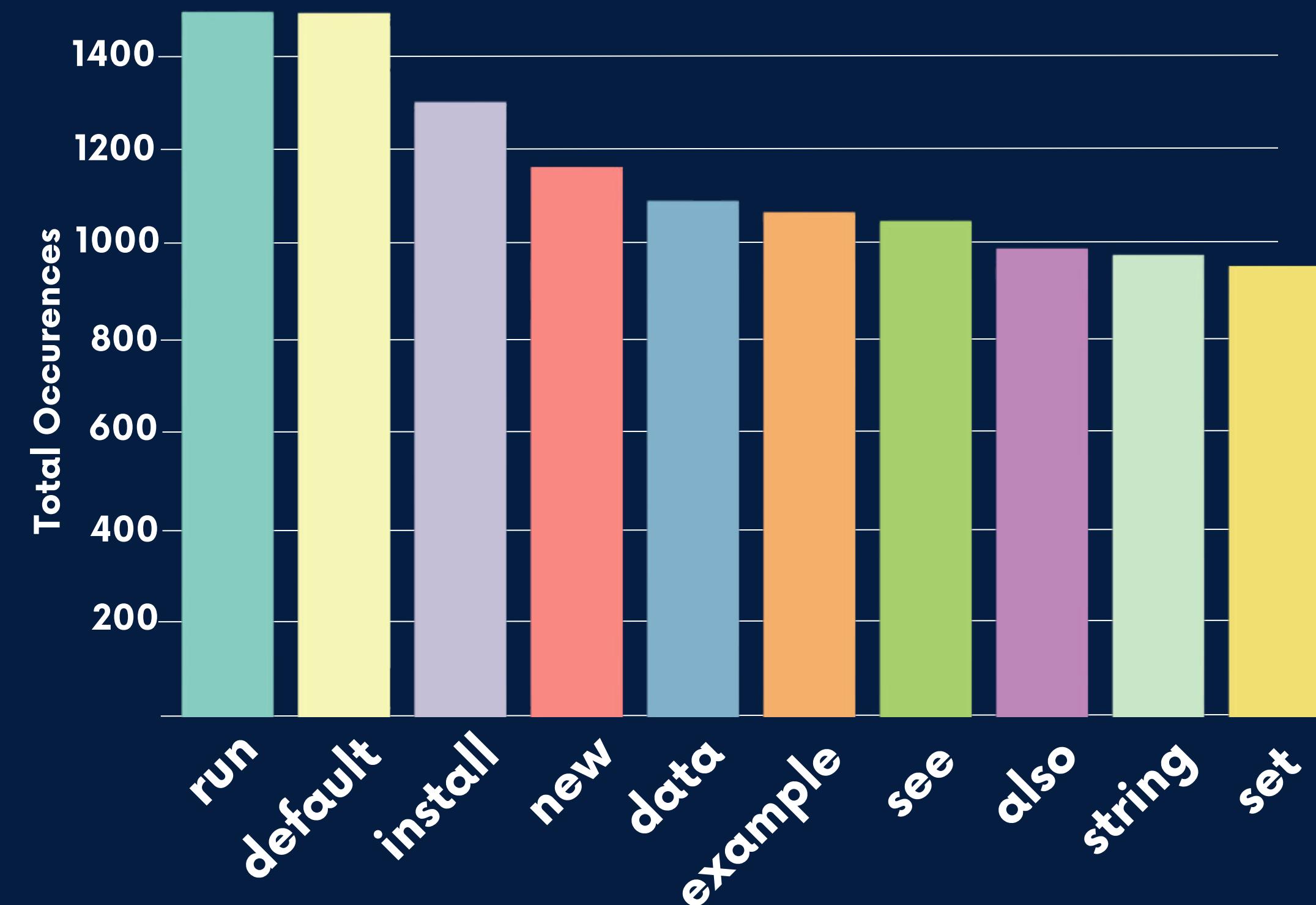
HTML

67

Pinpointing Languages?

10 Most Common Words

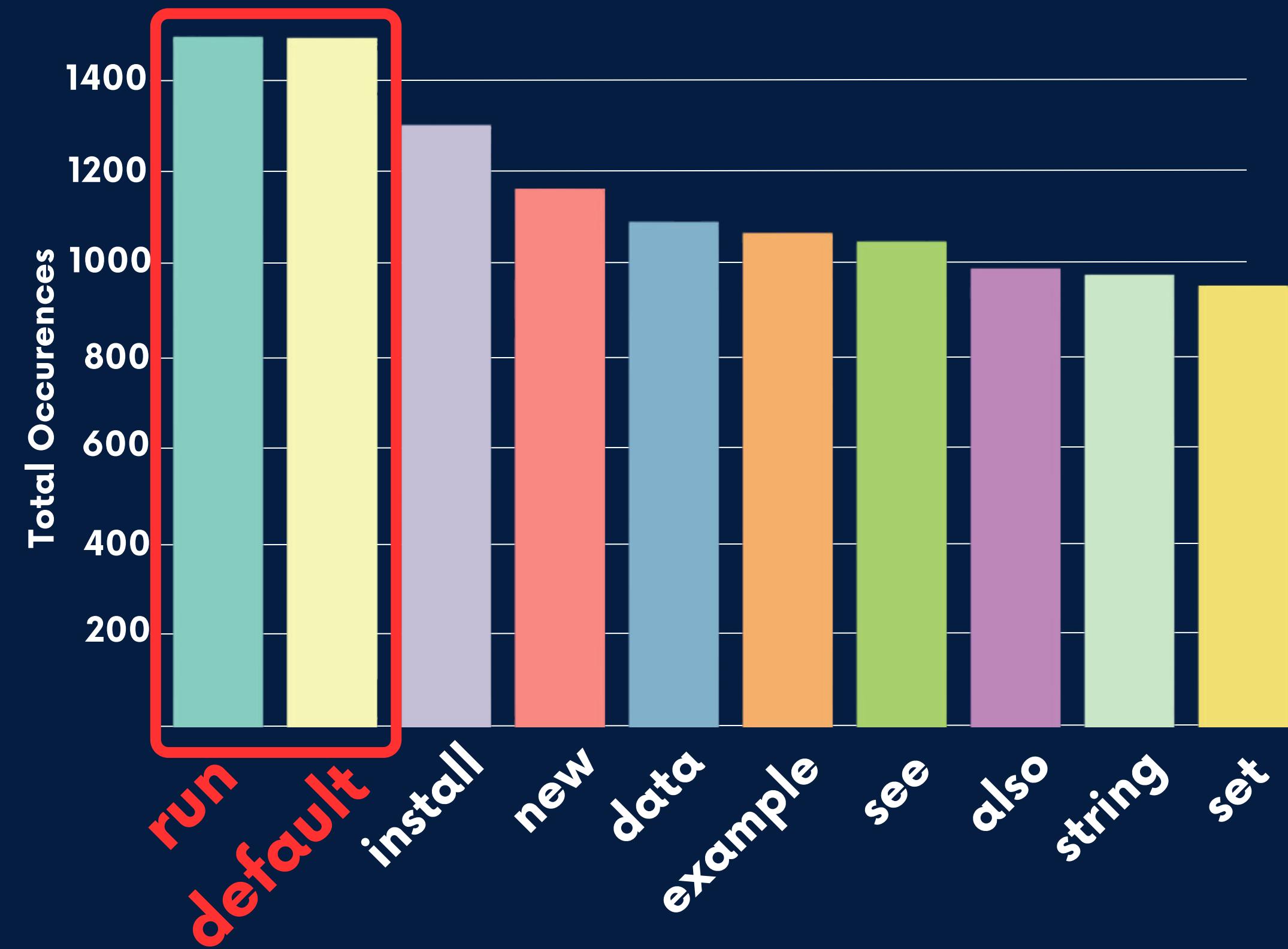
(Entire Corpus)



Pinpointing Languages?

10 Most Common Words

(Entire Corpus)



Pinpointing Languages?

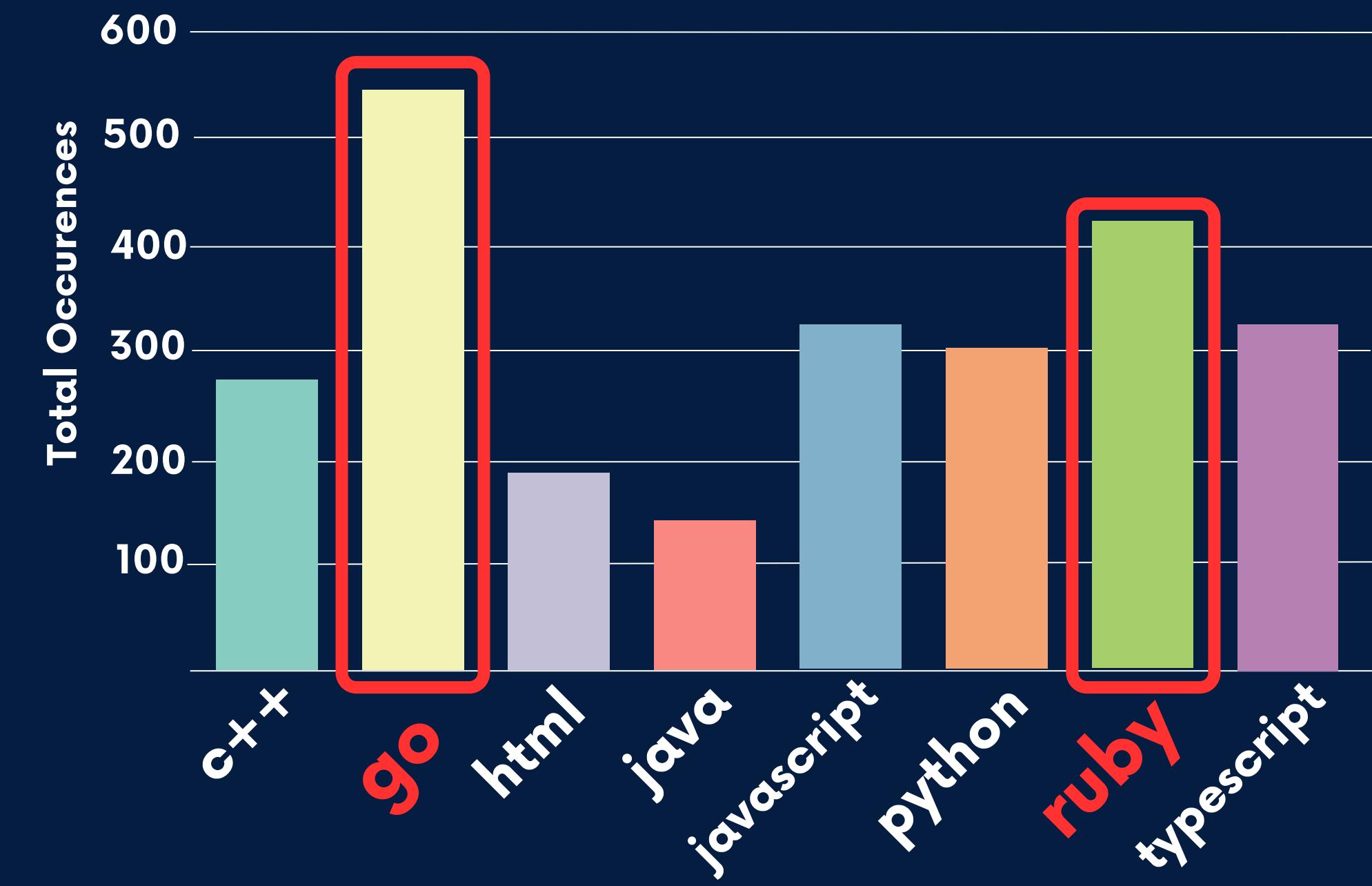
Run

Scripting

New
Object Oriented,
common in Ruby

Data
(Python)

'run' Occurrence (per language)



Pinpointing Languages?

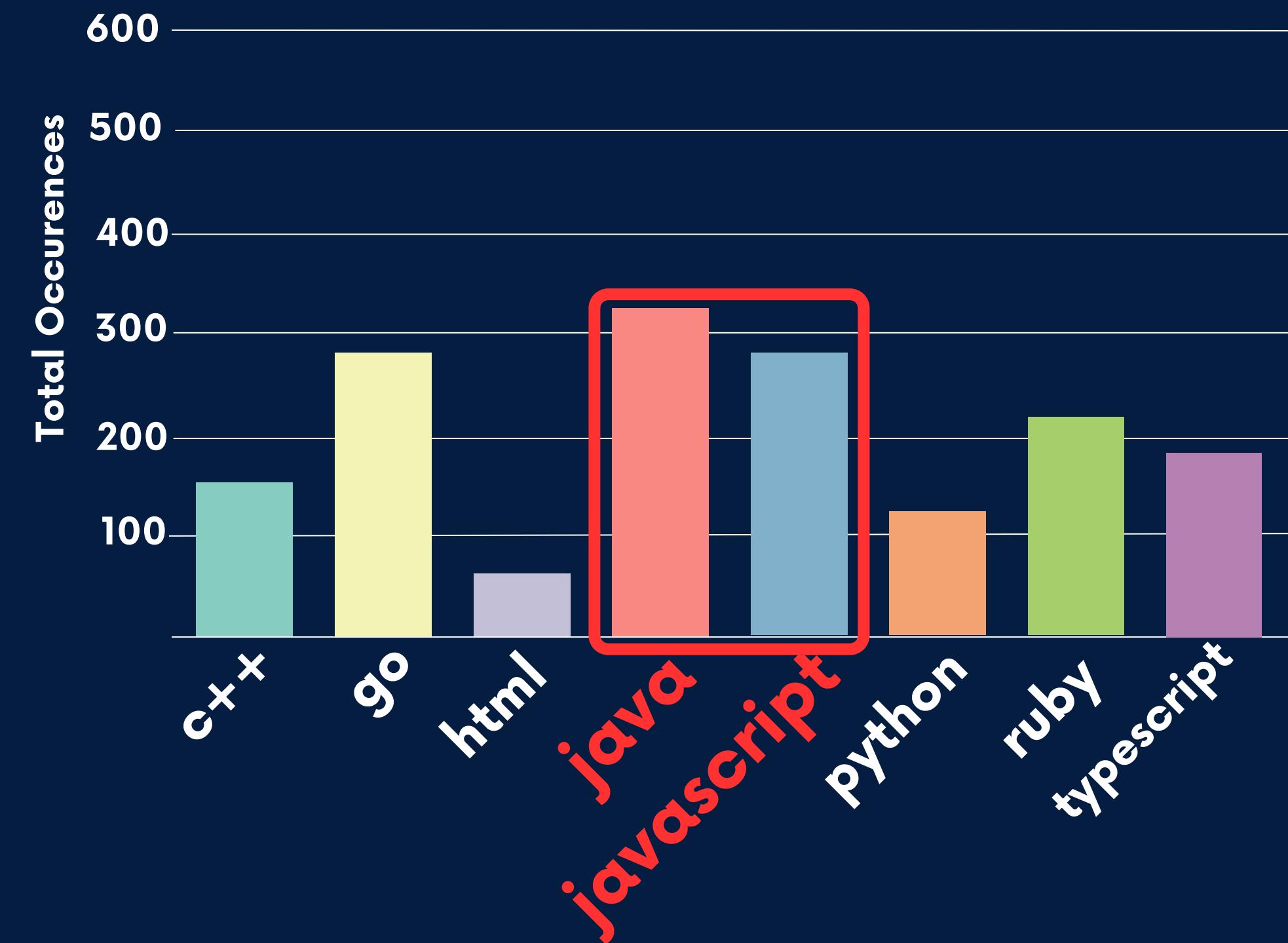
Run
All Scripting

New
Object Oriented,
common in Ruby

Data
(Python)

'new' Occurrence

(per language)



Pinpointing Languages?

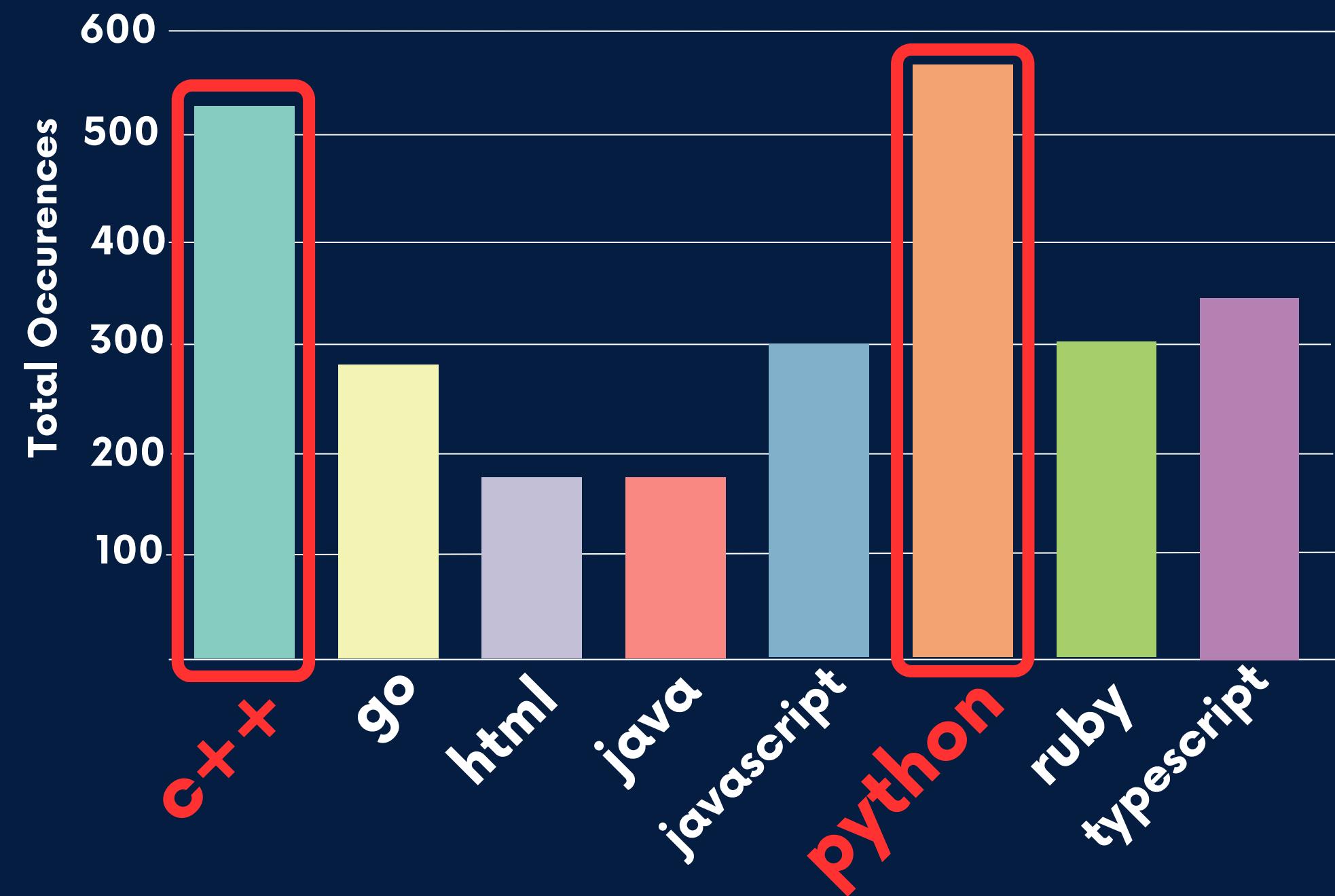
Run
All Scripting

New
Object Oriented,
common in Ruby

Data
(Python)

'data' Occurrence

(per language)



Unique Identifier Words

(Each Language Corpus)

css, grid, t--

HTML

rails, ruby, gem

Ruby

npm, const, react, property, properties

TypeScript

python, pip, yt, videos, model, dlp

Python

npm, browser, javascript, css, element, lazyload

Javascript

fscrypt, docker, protector, helm, commands, filesystem

GO

netdata, build, support, √, driver, esp32, esp8266, linux, memory

C++

void, override, android, spring, software, kotlin, copyright, distributed, maven

Java

Top Unique Words

In at least

25%of that language's
READMEs

Unique Identifier Words

(Each Language Corpus)

Top Unique Words

In at least

50%of that language's
READMEs

t--

HTML

rails, ruby

Ruby

TypeScript

python, dlp

Python

Javascript

fscrypt, protector

GO

esp32, esp8266, esp32c3, esp32s2

C++

android, kotlin, maven

Java

Unique Identifier Words

(Each Language Corpus)

Top Unique Words

In at least

75%of that language's
READMEs

t--

HTML

rails

Ruby**TypeScript****Python****Javascript**

fscrypt, protector

GO

esp32, esp8266, esp32c3, esp32s2

C++

maven

Java

Unique Identifier Words

(Each Language Corpus)

Top Unique Words

In at least

75%of that language's
READMEs

t--

HTML

rails

Ruby**TypeScript****Python****Javascript**

fscrypt, protector

GO

esp32, esp8266, esp32c3, esp32s2

C++

maven

Java

Utilizing GridSearch

13%
(baseline)

```
TfidfVectorizer(  
    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

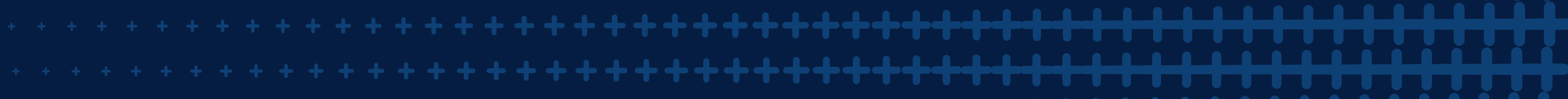
```
!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)
```

Ensemble Methods

Voting Classifiers

More Samples

Quality Samples



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL

DELIVER

FUTURE MODEL IMPROVEMENTS



Feature Engineering

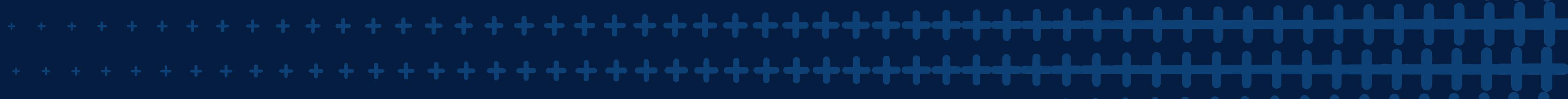
```
!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)
```

More Samples

Ensemble Methods

Voting Classifiers

Quality Samples



PLAN & ACQUIRE
PREPARE
EXPLORE
MODEL

DELIVER

FUTURE MODEL IMPROVEMENTS



Feature Engineering

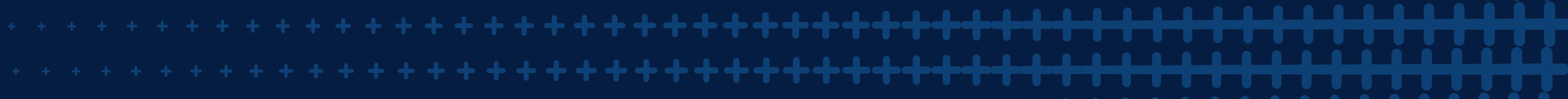
```
!pip install (Python)
<> (HTML)
&& (Ruby)
>> (C++)
```

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

run files help
yes npm
start users features
time information
version number

available options
default see
command configuration
get one based

data object 0
library class project
script path pay
option output node
key page documentation
true create value
access eg software
windows feature run

install native end
values need user
post size github
required sketch
message file
running go
text update browser
single main examples

example release 10 changes
name function via different requests
list many json error android
function tests note model c video
type instead

also license set
app module rails
server web client
import git uses
new build supported
source instead

following like automatically
environment