# WEB SCRAPING

# OBJECTIVE

Extract and analyze trending repositories from GitHub to identify popular projects and programming language distributions

## PLATFORM:

**PYTHON ENVIRONMENT**

## TOOLS:

**PYTHON, REQUESTS, BEAUTIFULSOUP, PANDAS, MATPLOTLIB, SEABORN**

# OVERVIEW // INTRODUCTION

THIS PROJECT FOCUSES ON EXTRACTING AND ANALYZING DATA FROM GITHUB'S TRENDING REPOSITORIES PAGE, WHICH SHOWCASES POPULAR REPOSITORIES BASED ON STARS, FORKS, AND RECENT ACTIVITY.

USING A COMBINATION OF REQUESTS AND BEAUTIFULSOUP, WE AUTOMATED THE PROCESS OF VISITING THE GITHUB TRENDING PAGE, EXTRACTING REPOSITORY DATA, AND VISUALIZING KEY METRICS INCLUDING STAR COUNTS AND PROGRAMMING LANGUAGE DISTRIBUTION.

THE GOAL WAS TO GAIN INSIGHTS INTO CURRENT DEVELOPMENT TRENDS, IDENTIFY POPULAR OPEN-SOURCE PROJECTS, AND APPLY WEB SCRAPING TECHNIQUES IN A REAL-WORLD CONTEXT.

# TOOLS & ENVIRONMENT

- REQUESTS – USED TO FETCH THE HTML CONTENT FROM THE GITHUB TRENDING PAGE
- BEAUTIFULSOUP – PARSED THE HTML CONTENT AND EXTRACTED REPOSITORY DETAILS
- RE (REGULAR EXPRESSIONS) – HELPED CLEAN AND EXTRACT NUMERIC VALUES FROM TEXT
- PANDAS – STORED THE SCRAPED DATA IN A STRUCTURED DATAFRAME AND EXPORTED IT AS A CSV FILE
- MATPLOTLIB – PROVIDED A BASE FOR CREATING PLOTS AND CUSTOMIZING VISUALS
- SEABORN – USED TO CREATE CLEAN, AESTHETICALLY PLEASING VISUALIZATIONS

# WEB SCRAPING PROCESS

THIS PROJECT INVOLVES WEB SCRAPING GITHUB'S TRENDING REPOSITORIES PAGE TO EXTRACT DATA ABOUT THE MOST POPULAR PROJECTS.

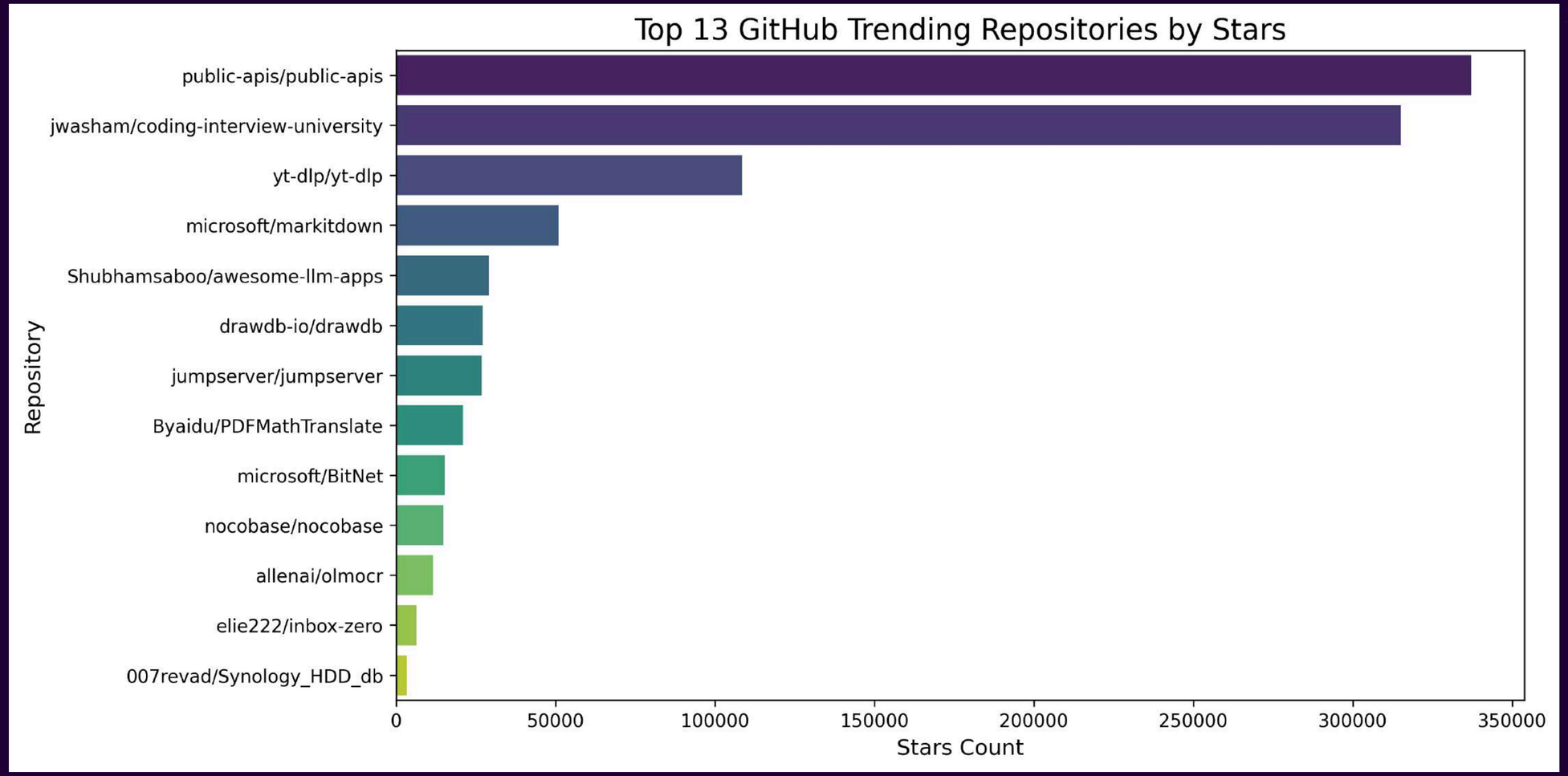
WE USED THE REQUESTS LIBRARY TO FETCH THE PAGE CONTENT AND BEAUTIFULSOUP TO PARSE AND EXTRACT STRUCTURED DATA FROM THE HTML. SPECIFICALLY, WE TARGETED THE REPOSITORY ARTICLES ON THE PAGE, WHICH CONTAIN THE REPOSITORY NAMES, PROGRAMMING LANGUAGES, STAR COUNTS, AND FORK COUNTS.

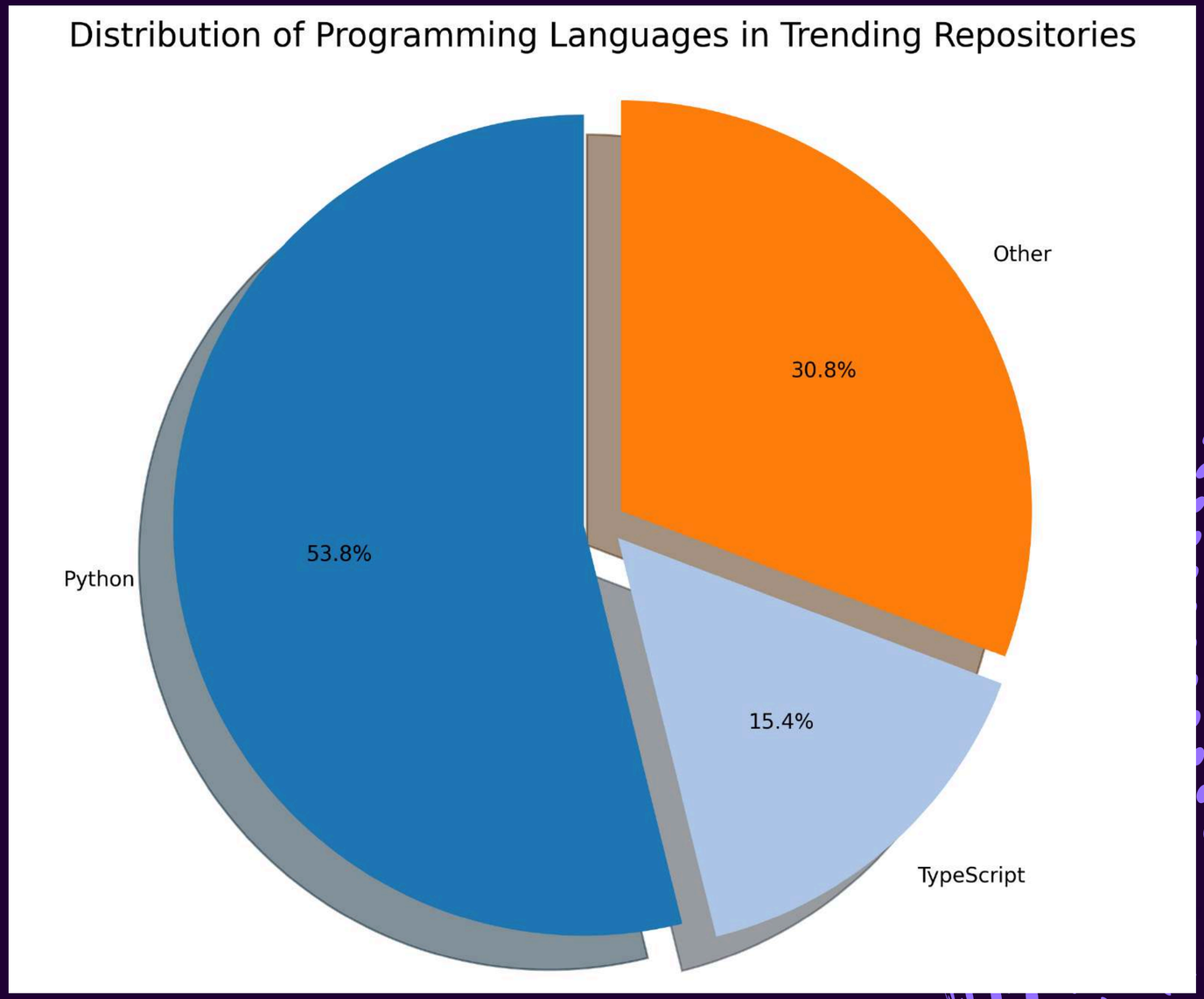
BY IDENTIFYING AND ISOLATING THESE ELEMENTS USING CSS SELECTORS, WE ACCURATELY CAPTURED THE REPOSITORY DATA. THE COLLECTED INFORMATION WAS THEN STORED IN A PANDAS DATAFRAME AND EXPORTED TO A CSV FILE (GITHUB_TRENDING_REPOS.CSV) FOR FURTHER ANALYSIS OR VISUALIZATION.

# DATA EXTRACTION

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Repository | Language | Stars | Forks |
| 2 | public-apis/public-apis | Python | 337058 | 35619 |
| 3 | jwasham/coding-interview-university | Unknown | 315074 | 78492 |
| 4 | yt-dlp/yt-dlp | Python | 108553 | 8529 |
| 5 | microsoft/markitdown | Python | 51061 | 2486 |
| 6 | Shubhamsaboo/awesome-llm-apps | Python | 29052 | 3257 |
| 7 | drawdb-io/drawdb | JavaScript | 27217 | 1932 |
| 8 | jumpserver/jumpserver | Python | 26894 | 5447 |
| 9 | Byaidu/PDFMathTranslate | Python | 21068 | 1783 |
| 10 | microsoft/BitNet | C++ | 15320 | 1085 |
| 11 | nocobase/nocobase | TypeScript | 14936 | 1643 |
| 12 | allenai/olmocr | Python | 11571 | 781 |
| 13 | elie222/inbox-zero | TypeScript | 6506 | 672 |
| 14 | 007revad/Synology_HDD_db | Shell | 3507 | 225 |
| 15 | | | | |

# DATA VISUALIZATION




Distribution of Programming Languages in Trending Repositories
Other
30.8%
53.8%
Python
15.4%
TypeScript

# CONCLUSION

- Successfully scraped GitHub's trending page to extract repository data
- Used a combination of requests, BeautifulSoup, and pandas to automate data collection
- Visualized the data using Seaborn, making trends easier to interpret
- Gained hands-on experience in web scraping, data handling, and visualization using Python
- Created a reusable framework for monitoring GitHub trends over time

# THANK YOU

Presented by :- Ya.sh Raj
(03920803122)