BIG DATA PLATFORMS | **ADSP 31013 IP01 | Final Project**

# Will TuringBots replace human software developers?

ARUSHI MAKRARIA | 8th December, 2023

# Executive Summary: Can AI Assistants Replace Humans?

Based on the analysis conducted, AI-assisted tools enhance developers' productivity but are unlikely to replace human software engineers and data scientists.

They augment tasks, improve efficiency, and facilitate learning, but human creativity, problem-solving, and domain expertise remain essential. Collaboration between AI and human experts is crucial for optimal results. If we were given data on codes generated by AI a comparison could be made between the efficiencies of both, but based on the available data AI can only help improve our efficiency not replace us completely.

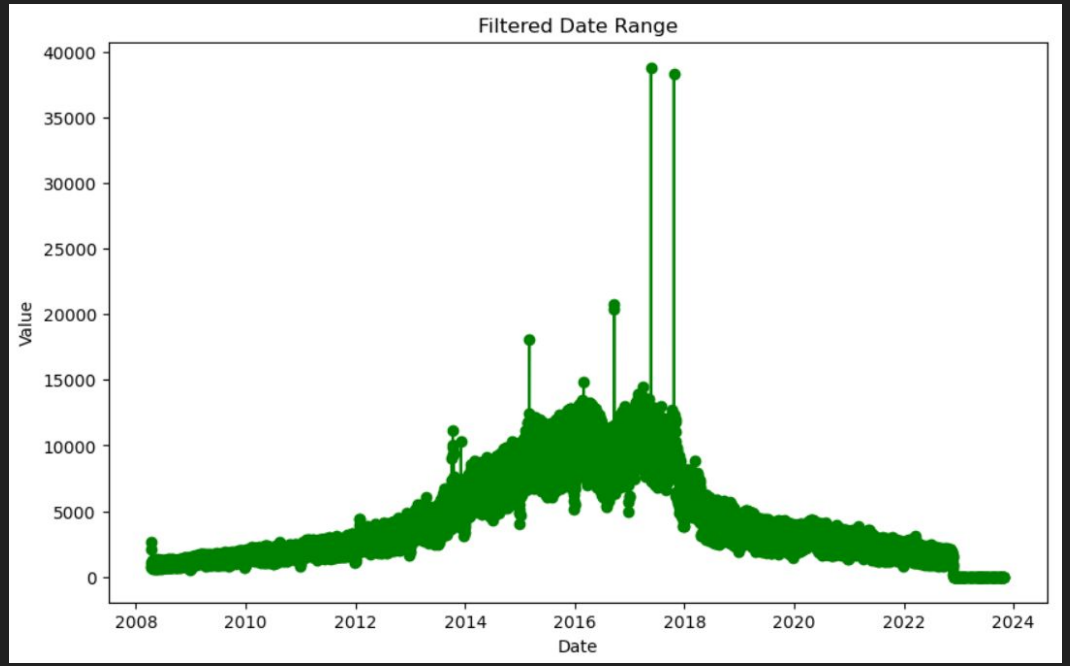
# From Chaos to Clarity: Preparing the Data for Analysis

- We began with a massive dataset of 26,538,161 records, brimming with potential insights.
- Our first step involved identifying and removing irrelevant columns.Columns like "encoding" "trailer" "difference" and "author" were dropped after recognizing that they were not essential for our exploration.
- On examining the dates it was noted that there were no data collection gaps.
- For LHS Analysis stop words were removed. By eliminating these extraneous elements, we created a cleaner dataset, optimized for LSH and ready for our analysis to commence.

# Dataset Timeline Insights: 2008 - 2023

- The dates ranged from before 1980s to 2120 which is factually incorrect so only records between 2008 - 2023 were considered for our analysis.
- The commits peaked in 2017 ignoring the outliers.

- The peak in 2016 could be associated with the launch of ANGULAR which is one of the top most contributed projects on github.

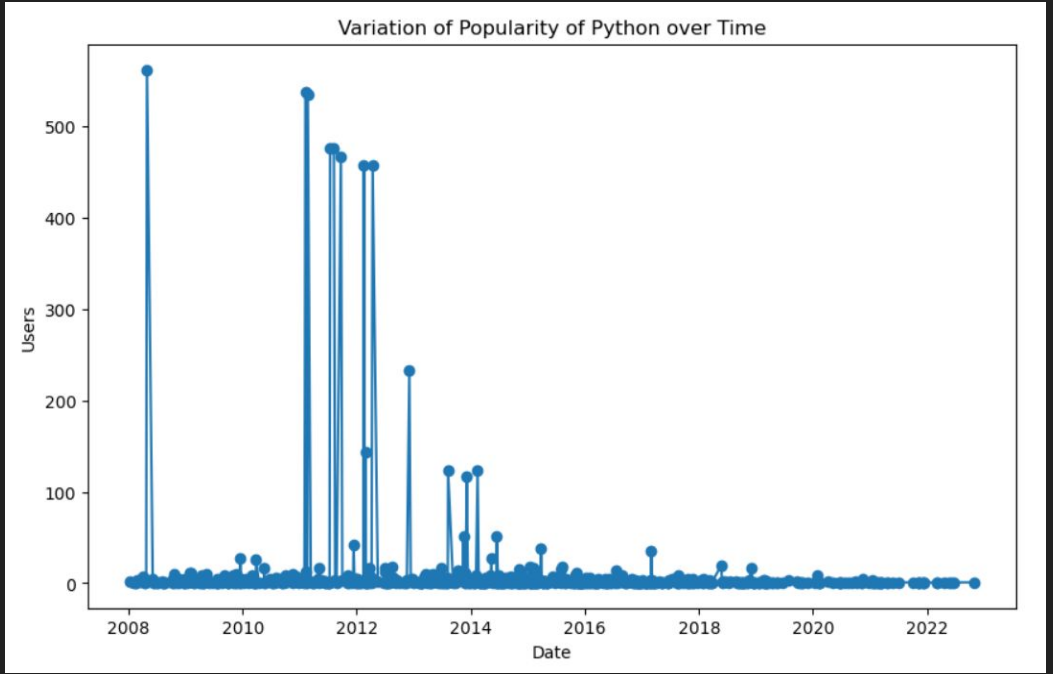


# The Most Popular Programming Languages

On conducting our analysis the five most popular programming languages used on github were:

| Top Languages |
| --- |
| JavaScript |
| CSS |
| Shell |
| HTML |
| Python |

The variation of popularity of Python -
One of the top 5 languages -
over time is shown in the graph.



# GitHub Licensing Landscape: Insights and Trends

- MIT emerges as the most prevalent license across GitHub repositories.
- MIT and JavaScript form the most popular license-language combination.
- Languages like Python and JavaScript show a strong association with the MIT license, while C/C++ embraces a diverse range, including MIT, Apache License 2.0, GPL, and proprietary licenses.
- The table below summarizes the top 5 language-License pairs:

| Language | License |
|---|---|
| JavaScript | MIT |
| Java | Apache2.0 |
| Shell | gpl-2.0 |
| JavaScript | gpl-3.0 |
| Shell | bsd-3-clause |

# Technologies used in Data Science and AI Projects

- On examining the contents of the repositories, the ones linked to data science and Artificial Intelligence used the following Languages, Libraries and Frameworks:

| Language | Libraries | Data Processing | Containerization |
|---|---|---|---|
| Python | TensorFlow | Pandas | Docker |
| R | PyTorch | Numpy | Kubernetes |
| | scikit-learn | Apache Spark | |

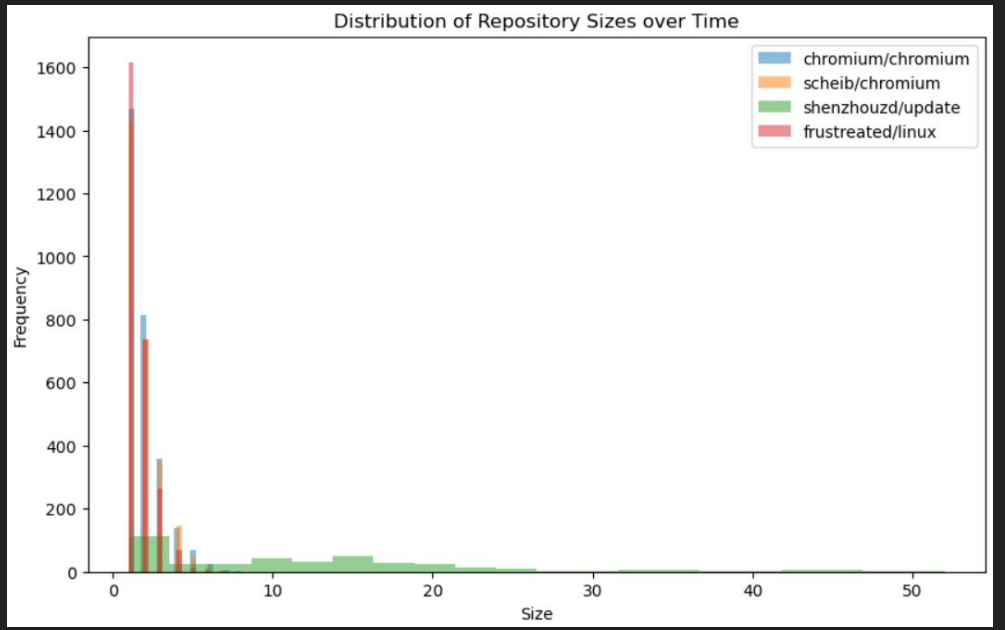
# Technologies on GitHub and their BigTech Associations

- Some of the most popular technologies on GitHub are React-Native,VS Code, TensorFlow, Docker and Kubernetes.
- VS Code is open sourced by Microsoft and React-Native is backed by Meta which are both big tech companies.
- The most popular languages were JavaScript, HTML & CSS implying that web development is a very popular domain on GitHub.
- Technologies associated with emerging fields like Artificial Intelligence, Machine Learning, and Web Development are gaining traction and contributing to GitHub's dynamic ecosystem.

# Repository Analysis

The Most popular repositories are listed below:

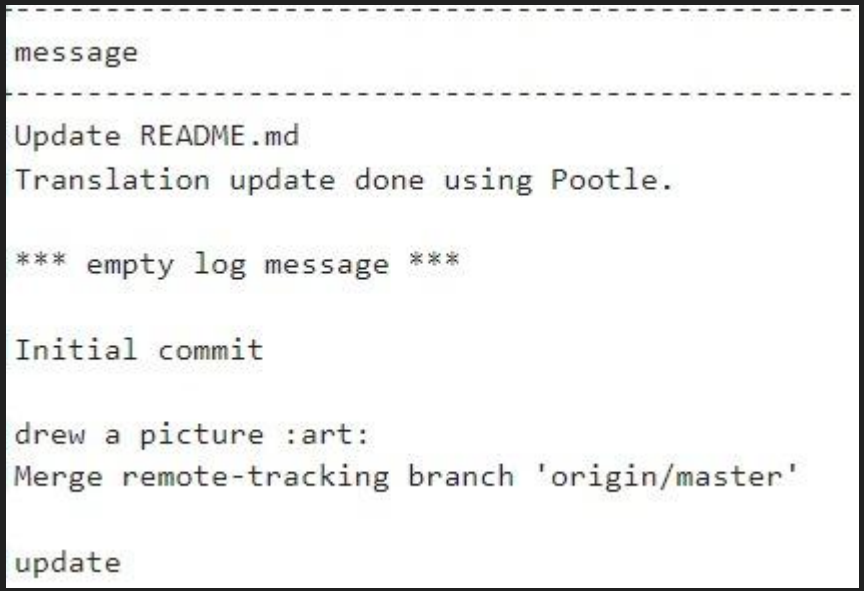
| repo_name | count |
|---|---|
| chromium/chromium | 119717 |
| shenzhouzd/update | 118597 |
| scheib/chromium | 110513 |
| cminyard/linux-li... | 108625 |
| frustreated/linux | 107459 |




The Graph summarizes the distribution of their size over time. While shenzhouzd/update has been consistent in is count, frustrated had a spike of popularity then died down.

# Reasons for Commiting to GitHub Repositories

- The highest commits were for Update README which is used for documentation purposes.
- The users were also committing their log files.
- The third most popular reason was the initial commit which is done at the start of a project.
- The top 5 reasons are summarized in the table.

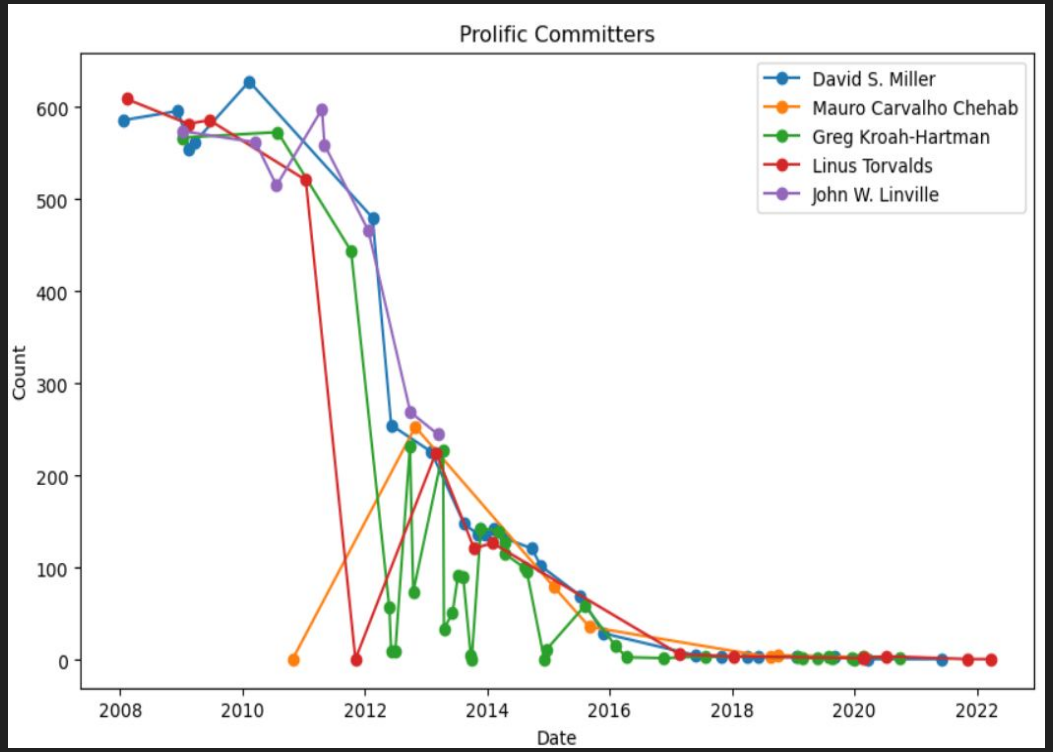


# Unveiling GitHub's Most Influential Contributors

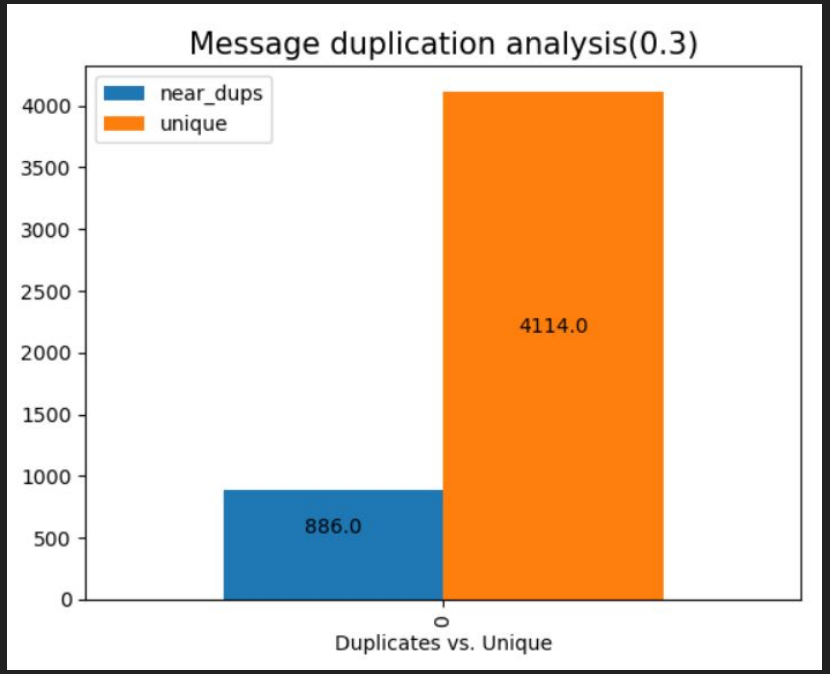
>The trends of the most prolific committers are depicted on the right-hand graph, spanning the period 2008 to 2023.

>Based on the commits volume the top contributors are summarized in the table below.

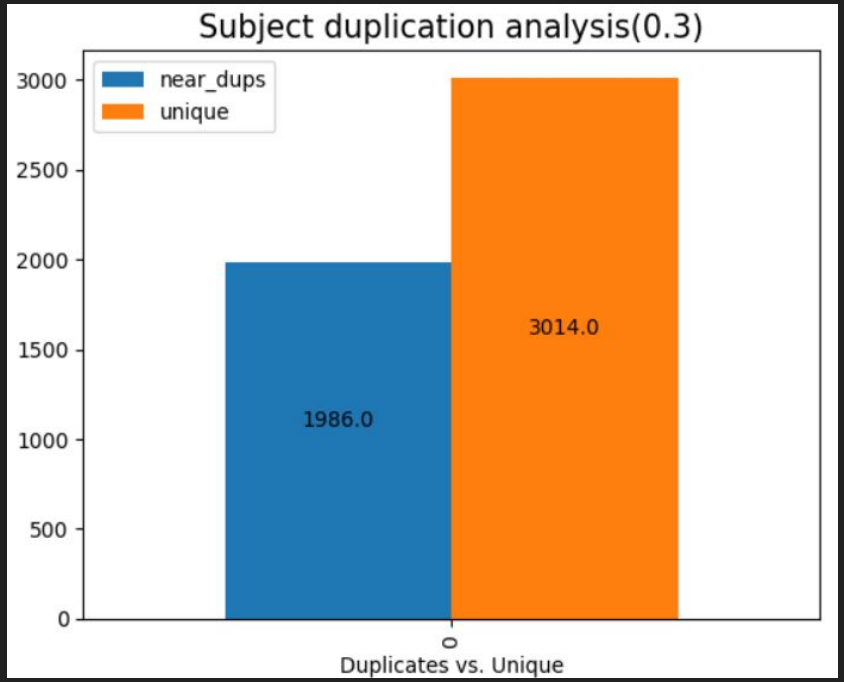
| Top Contributors | Commits |
|---|---|
| Linus Torvalds | 289,534,682 |
| David S. Miller | 197,194,819 |
| Greg Kroah-Hartman | 170,286,941 |
| Mauro Carvalho Chehab | 89,475,155 |
| John W. Linville | 84,503,430 |



# Similarity Analysis for Messages and Subject

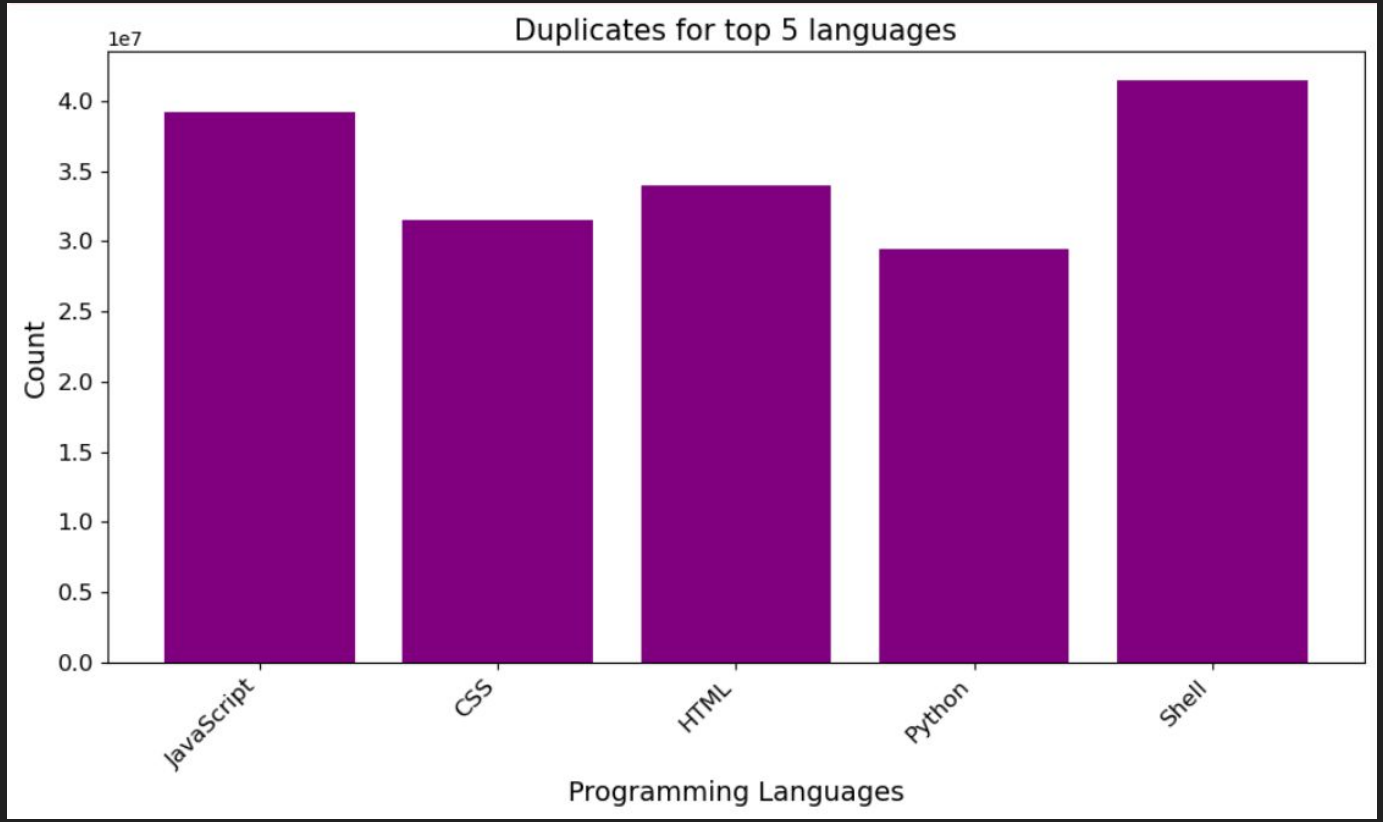



The message column seems mostly unique.




The Subject column seems a little more balanced implying that people might be copying and pasting the subject.

# “subject” and “message” duplication for each of the top 5 programming languages



# Conclusions

- Some things we can conclude after our analysis are that there was a trend where github was immensely popular but its use has been declining lately.There are other platforms like GitLab and Bitbucket that offer similar services. Depending on specific needs or preferences, developers might choose one platform over another.
- Some of the most popular languages were JavaScript, CSS & HTML indicating that Web Development is very popular on GitHub.
- For AI and Data Science projects people preferred to use Python Apache Spark and TensorFlow.

# Recommendations

- While the boom of various AI assistants are trending it is rational to use them only as assistants with our work and not be completely dependent on them to complete tasks.
- AI systems, including assistants, can inadvertently perpetuate biases present in their training data. Depending heavily on AI in software development raises ethical concerns related to biased decision-making and unintentional propagation of discriminatory practices. It is crucial to maintain human oversight to ensure fair, ethical, and unbiased coding practices.
- Relying solely on AI could lead to code that lacks the necessary nuanced understanding of the project's objectives. It is important to be able to understand and eliminate redundancy.