Проект 1

Никифоров Саша

23 октября 2024 г.

1 Информация о датасете

1.1 Ссылка на датасет

Здесь.

1.2 Название

Repository dataset

1.3 Описание

Информация о различных репозиториях на гитхабе: название, основной и все языки программирования, кол-во звезд, дата создания и проч.

2 Выводы о результатах

2.1 Гипотезы (Г.) и исследовательские вопросы (И.)

2.1.1 Г. Репозиториев с большим количеством звезд меньше, чем с маленьким

Гипотеза частично верная. Четко прослеживается линейная зависимость между кол-вами звезд и репозиториев, но на отдельных участках разная. Пик кол-ва репозиториев быстро достигается в районе 8 звезд и далее кол-во репозиториев начинает снижатся. Так продолжается до +-10 тыс. звезд, где есть большая нестабильность, но, в целом, скачек вверх, а после на различные большие значения звезд (больше 20тыс.) приходится по одному репозиторию.

2.1.2 И. Япы в репозиториях с +-10тыс. звездами, на которых есть хотя бы 50 проектов

Преобладает JS, Python, TS, Go. Еще, я выделил веб и не-веб языки (довольно грубо, но это не является основной частью исследования, а мне просто захотелось посмотреть). По итогу соотношение веб к не-веб равно 54/46.

2.1.3 Г. У репозиториев с основным языком Rust в среднем больше звезд, чем у C++

У Rust и правда больше, чем у C++ (звезд в среднем). А вообще, больше всего у Go. Выбирались только япы с 5тыс.+ проектами на них, чтобы избежать единичных проектов с огромным кол-вом звезд на редком языке.

2.1.4 Г. Репозитории с linux в названии в среднем имеют больше звезд, чем с win

Да. С linux в среднем 195 звезд, с win - 128.

2.1.5 Г. Чем ближе к нынешнему времени, тем больше соотношение проектов на Rust к проектам на C++

Да, так оно и есть. Хотя проектов на C++ все еще больше (и всегда было больше) и в количестве и в процентном соотношении, но начиная с 2011 г. соотношение Rust/C++, которое тогда было около 0, растет и в 2022 г. оно равно 0.45, что является пиком (не говорю про 2023 г., т.к. там мало данных). Интересно, что пик кол-ва проектов на C++ был в 2018 г., а на Rust - в 2020 г.

В целом, многие гипотезы подтвердились, т.к. были логичными, а исследовательский вопрос и прочие частные вопросы дали логичный результат.

3 Прочее

Так как реализовать некоторые графики не удалось в рамках основных Гипотех и исследовательских вопросов, они были построены, основываясь на тех же данных, но независимо от главной части проекта.

Также, представлены два облака слов (с названиями языков и со словами в названиях репозиториев).

На протяжении всего проекта должен сохраняться кодстайл и старалась сохраняться цветовая палитра (исключения есть в неосновной части: в облаках слов и в последнем графике, что было сделано в целях лучшего разделения слов/частей легенды, т.к. при основной выбранной цветовой палитре не хватает цветов и все сливается).