Be Coder

Задача на финал

«Кто виноват и что делать?»

История коммитов в репозитории может дать немало полезной информации, особенно если разработчики старательно пишут сообщения. Например, во многих публичных проектах есть соглашения об именовании коммитов, которые являются баг-фиксами. Репозитории для примера: https://github.com/angular/angular, https://github.com/usememos/memos, https://github.com/knockout/knockout. Видно, что сообщение коммита, который исправляет дефект, содержит какую-либо форму слова fix. Для решения этой задачи рекомендуется выбрать один из публичных git-репозиториев, в котором число авторов не менее 5.

Если существует фикс, значит, была ошибка. Для простоты будем считать, что ошибку внес тот, кто последним правил код, который позже исправили в фиксе.

Гипотеза 1. Один QA-инженер выдвинул гипотезу, что один и тот же разработчик в одних и тех же файлах ошибается чаще, чем в других. То есть у него есть определенные пробелы в знании отдельных частей системы, которые отражаются на качестве кода. Правда, инженер не придумал, как учесть тот факт, что разработчик со временем обучается.

Гипотеза 2. Еще инженер выдвинул вторую гипотезу о том, что разработчик с наибольшей вероятностью ошибется в коде (в файле), с которым он еще не работал. Потом пришел Senior QA-инженер и переформулировал эту гипотезу так: разработчик чаще ошибается в коде, с которым он реже работает. Правда, четких критериев «чаще» и «реже» он не сформулировал.

Вам предлагается выполнить 2 задания. Первое нам кажется сложнее. Выполнение будет оцениваться независимо, то есть допустимо решать вторую задачу, не решив первую.

Задание 1. Проверить одну из гипотез (или обе).

Можно использовать метод проверки статистических гипотез.

Можно подбирать параметры модели (и даже применять машинное обучение) для предсказания вероятности ошибки и сверять прогноз с фактическими ошибками. Для оценки модели можно использовать критерии полноты (recall) и точности (precision) или другие, которые вы считаете подходящими.

В процессе решения у вас могут возникнуть другие гипотезы на основе полученных данных. Будет плюсом, если вы опишете их и подкрепите данными.

Проверка гипотезы на нескольких репозиториях тоже будет плюсом.

Предполагается, что получение необходимых данных из репозитория и проверка гипотез будут легко повторяемыми, то есть полностью или частично автоматизированными.

Задание 2. Предположим, что первая гипотеза верна (независимо от результатов, полученных в задаче 1). Тогда мы могли бы определять:

1. Коммиты, в которых весьма вероятна ошибка (по сочетанию автора и измененных файлов). Такие коммиты надо тщательнее тестировать.
2. Разработчика, которому лучше доверить ревью таких коммитов или исправление ошибок в них. Критерии выбора такого разработчика вам предстоит придумать.

Ожидается, что решение будет содержать код для получения данных из репозитория, определения проблемных коммитов и выбора разработчика для ревью и исправления. Однако,

при оценке выполнения задания основное внимание будет уделяться презентации идеи решения и подробностей расчета, уточнению важных деталей, проработке особых случаев.

Общее для двух заданий. Важен не столько финальный результат решения задач, сколько выбранные пути решения, принятые допущения, четкость описания, оптимальный план реализации в коде. Любые наработки и промежуточные результаты будут оценены.