Тестовое задание на вакансию Python3 backend developer

Ссылка на вакансию: https://hh.ru/vacancy/85784343

Компания: https://scan-factory.ru

Предисловие: Одна из функций нашего продукта - поиск публично-доступных ресурсов, принадлежащих Заказчику. Заказчик (для примера, это будет компания "Яндекс") вводит в нашу платформу свой домен второго уровня (например, yandex.ru), а наша платформа ищет доменные имена 3, 4, 5 уровней в рамках этого домена (например, мы найдём subdomain.yandex.ru, admin.subdomain.yandex.ru, и тд).

Иногда возникает проблема, связанная с тем, что DNS-сервер Заказчика настроен таким образом, что на запрос любого домена 3, 4, 5 уровня будет считаться, как реально существующий (если зайти по ссылке https://non-existent-random-subdomain.yandex.ru, откроется некий сайт Заказчика). Такая ситуация приводит к тому, что наша база данных заполняется "мусорными" доменами, которые нужно отсеивать от реальных.

Мы сымитировали ситуацию, описанную выше: создали базу данных sqlite с набором тестовых доменов. Скачать тестовую базу данных: https://disk.yandex.ru/d/EPzAQWYhvxoUAA

База данных имеет следующую структуру:

Таблица domains - содержит колонки name, project_id. Таблица заполнена набором доменов.

Таблица rules содержит колонки regexp, project_id. Таблица пустая.

Задача:

1. Скачать файл с базой данных sqlite: https://disk.yandex.ru/d/EPzAQWYhvxoUAA

2. Изучить данные, которыми заполнена база данных

3. Написать скрипт на Python, который для каждого проекта (project_id) создала регулярное выражение в таблице regexp, которое будет отсеивать "мусорные" домены.

4. По итогу работы программы таблица rules должна содержать регулярные выражения, вместе с project_id проекта, которому они соответствуют. **То есть, на выходе в БД должны оказаться регулярки**, **отсеивающие домены вида `*.sub.yyy.com` и `*static.developer.xxx.com`**

Важно: не нужно хардкодить регулярные выражения на основе предоставленных данных (например, не нужно хардкодить правило "[a-z].sub.yyy.com"). Мы ждём от вас обобщенное решение, которое работало бы на любом наборе входных данных -- т.е. алгоритм, который вычленит "мусорные" паттерны из набора доменов.

При оценке работы учитывается как правильность выполнения задачи, так и качество самого кода, как если бы этот код деплоился в production.

Код заливайте в репозиторий на гитхабе, и ссылку присылайте на почту info@sf-cloud.ru вместе с Вашим резюме.

Ваши решения принимаем до конца дня 14.09.2023.

Каждому, кто пришлёт нам рабочее решение, мы переведём 500 рублей (укажите, пожалуйста номер телефона/номер карты в емейле, куда перевести бонус) в качестве "спасибо" за участие в решении тестового задания.

Спасибо за ваше участие!