Это многоразовый парсер. Мы будем использовать его каждый день.

Заходим на сайт https://uk.indeed.com/

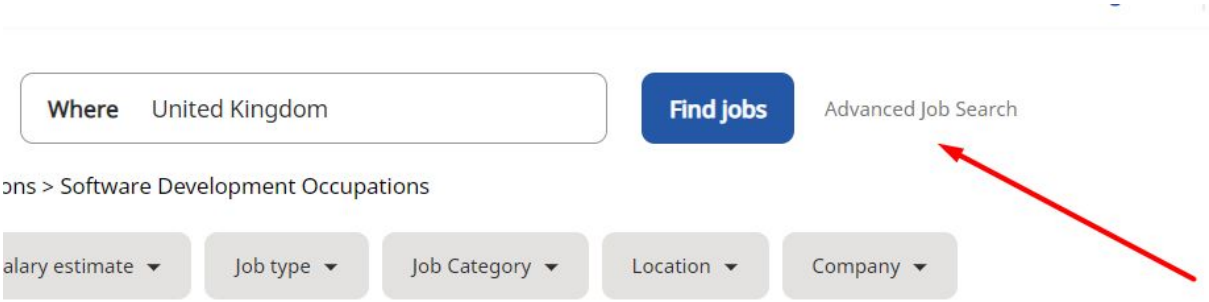
Важно: Indeed это очень популярный сайт. ТОП-50 в мире. Его скрапили уже тысячи раз разные люди, поэтому в интернете достаточно много различных примеров кода, возможно вам это поможет.

Первый парсинг будет самый большой. Потом ежедневные будут короче.

В разделе What вбиваем слово **developer** в разделе Where вбиваем **United Kingdom**

Нажимаем кнопку Find jobs.

Нажимаем Advanced Job Search



В первом поле вводим developer

Здесь оставляем всё как есть (Все работы, Все сайты, Исключить агентства):

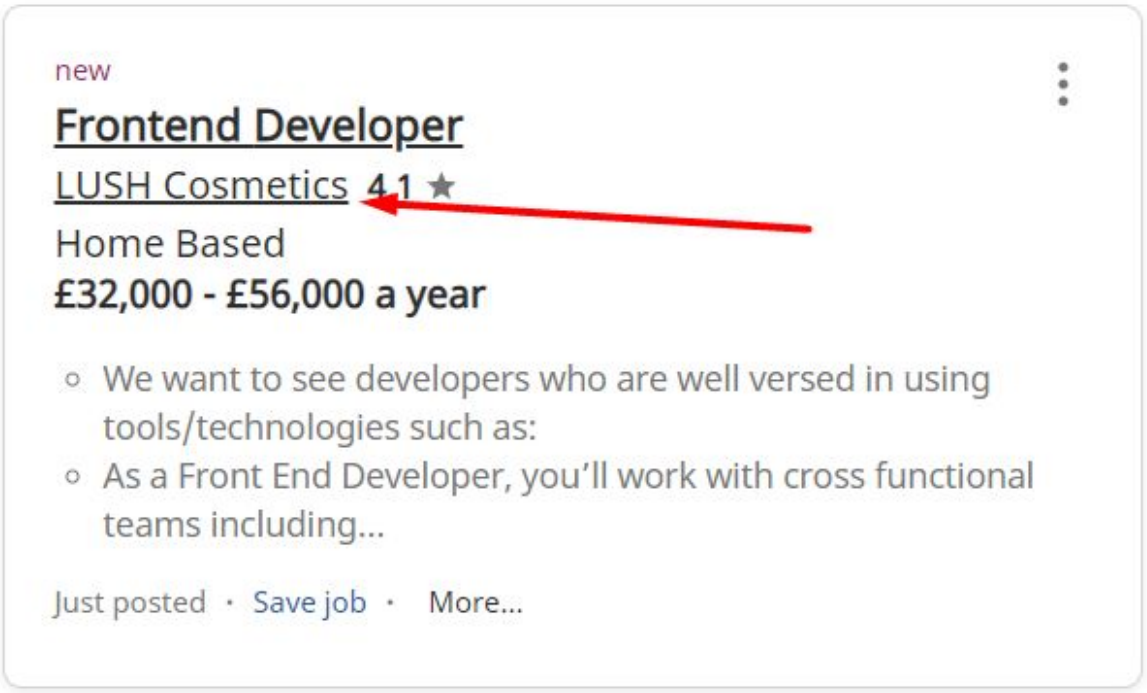
Выбираем only in

Выбираем within 15 days, отображать по 50 и сортировать по дате:

Нажимаем кнопку, получаем примерно 2800 страниц по 50 карточек (140,000 карточек):

Ссылка выглядит вот так:
https://uk.indeed.com/jobs?as_and=developer&as_phr=&as_any=&as_not=&as_ttl=&as_cmp=&jt=all&st=&sr=directhire&salary=&radius=0&l=United+Kingdom&fromage=15&limit=50&sort=date&psf=advsrch&from=advancedsearch

Для начала со всех карточек (которых 140.000) парсим только название компании и ссылку на профиль компании:



Ссылки на профиль выглядят вот так: https://uk.indeed.com/cmp/Lush-Cosmetics

На некоторых названиях компании нет ссылки на профиль, не знаю почему, видимо у таких компаний нет профиля. В таком случае мы парсим название компании и ссылку на вакансию (это та что при нажатии открывается справа). Ссылка на вакансию имеет свой идентификатор и выглядит вот так: https://uk.indeed.com/jobs?q=developer&l=United%20Kingdom&radius=0&sort=date&limit=50&sr=directhire&fromage=15&vjk=fa982a5f899be079

Все названия компаний и ссылки на профиль/вакансию нужно добавить в нашу БД. Для БД мы хотим использовать что-то что имеет визуальный интерфейс и интеграцию с Zapier. Например **Firebase** или **Airtable**.

Далее нам нужно взять все профили компаний и удалить дубли по названию (либо изначально перед каждым добавлением в БД смотреть есть ли там уже такая компания, если есть - не добавлять). Дублей будет очень много, т.к. большинство вакансий публикуют одни и те же компании.

Те компании у которых нет ссылки на профиль (только название и ссылка на вакансию) мы объединяем. Т.е. если у нас было вот так:

Application Solutions LTD	Link 1
Application Solutions LTD	Link 2
Application Solutions LTD	Link 3

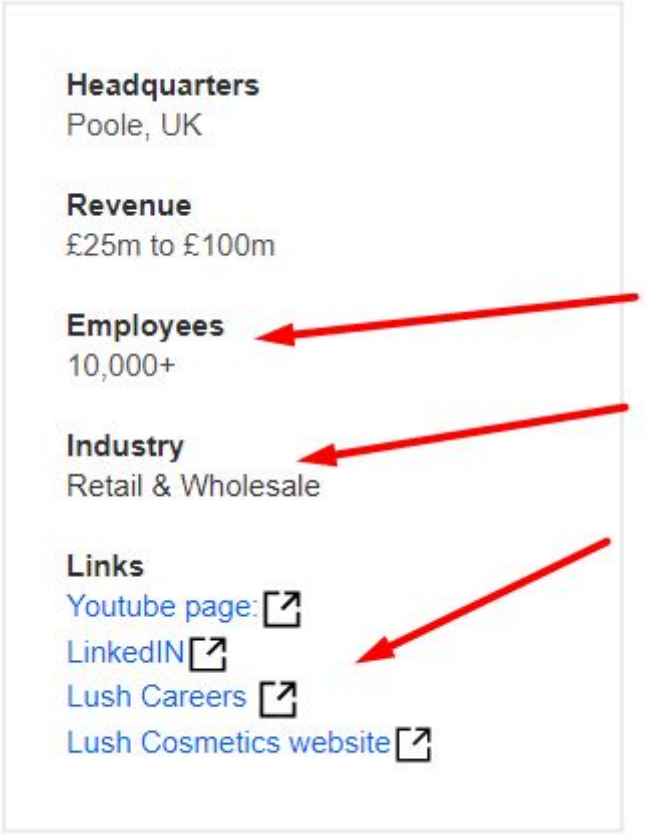
То теперь будет одна строка:

Application Solutions LTD	Link 1 на вакансию Link 2 на вакансию Link 3 на вакансию

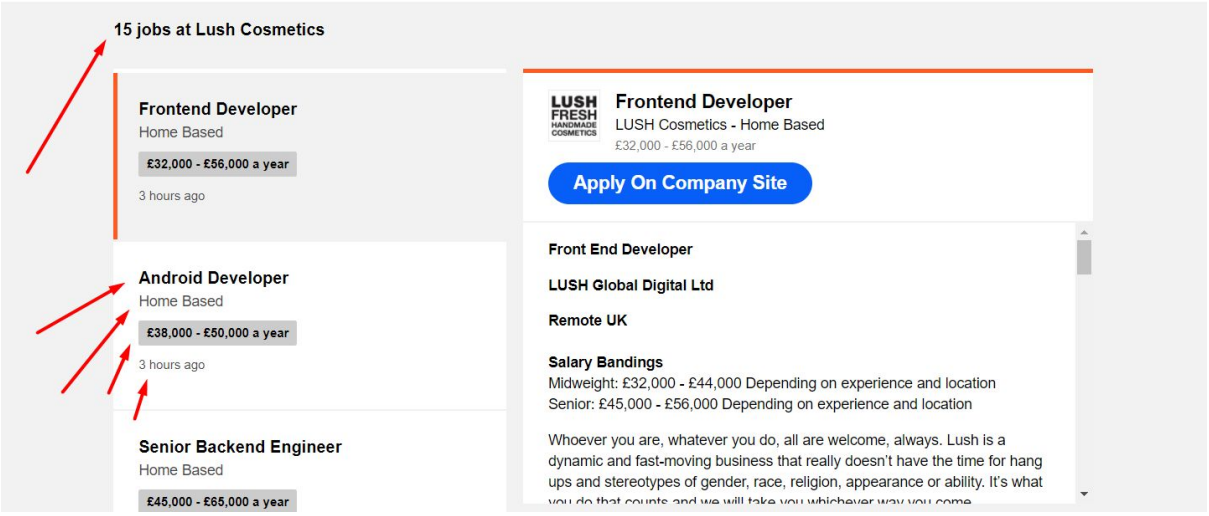
Итого у нас появляется что-то вроде карточки компании. Теперь нам эти карточки нужно наполнить, как это сделать:

1 - Те у кого есть ссылка на профиль компании, например https://uk.indeed.com/cmp/Lush-Cosmetics

Открываем раздел **Why Join Us** https://uk.indeed.com/cmp/Lush-Cosmetics/about и справа парсим количество сотрудников, индсустрию и весь раздел Links. Кол-во сотрудников и индустрию мы будем использовать в фильтрах.



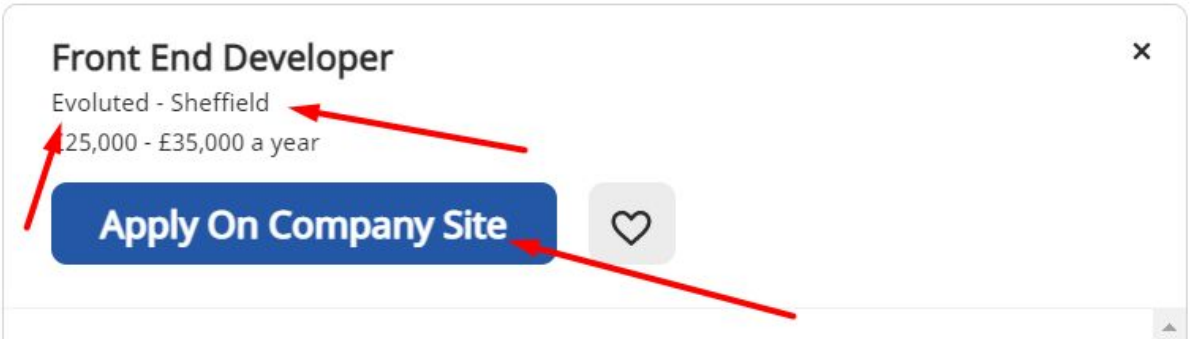
Затем переходим в раздел **Jobs** https://uk.indeed.com/cmp/Lush-Cosmetics/jobs и парсим общее количество вакансий, и список открытых вакансий в формате Название, Город (на скрине Home Based), зарплата, дата публикации:

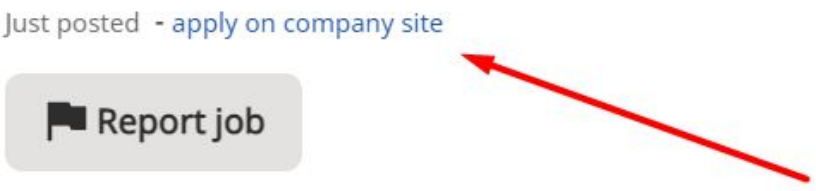


2 - Те у кого только название компании и ссылки на вакансии:

Открываем ссылку на любую вакансию (например самую свежую по дате) и парсим из нее следующую инфу (и добавляем в нашу “карточку компании”):

- Название компании
- Город
- Ссылка с большой кнопки (эта кнопка может называться по разному, не только Apply On Company Site, но нам в любом случае нужна ссылка с этой кнопки)
- Ссылка внизу описания, которая ведет на сайт компании





Итого у нас есть несколько сотен или даже тысяч **уникальных** “карточек компаний”. В которых у нас есть различная инфа об этих компаниях:

- Название компании и ссылка на профиль Indeed
- Кол-во сотрудников
- Индустрия
- Общее количество вакансий
- Ссылка на сайт
- Прочие ссылки
- Дата публикации самой свежей вакансии (нужно определять скриптом)
- Отдельным блоком список всех открытых вакансий:
    - Название
    - Город
    - Зарплата
    - Дата публикации

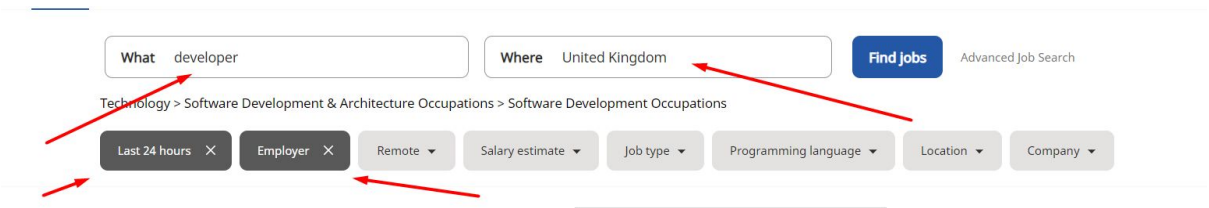
Надо обернуть это в простенький и удобный веб-интерфейс, который будет удобно развивать дальше и наращивать мясо. Например список всех компаний таблицей и при клика пусть открывается список вакансий. Хотя зачем нам его открывать у себя на сайте, если можно сразу открывать Indeed в разделе Jobs.
Для веб-интерфейса можно взять шаблон отсюда
https://www.creative-tim.com/templates/free

Повторный парсинг.
Вакансии добавляются каждый день (около 6.000 в сутки), поэтому парсер должен быть многоразовый. Запускаться **автоматически раз в сутки**.

2 основные задачи повторного парсинга:

1. Добавить в БД новые компании, которых у нас ещё нет
2. Обновить столбец “дата последней вакансии” у тех компаний, которые уже есть в нашей БД (чтобы видеть тех, кто допустим месяц никого не искал, а сейчас вновь ищет)

Спустя 24 часа после первого большого парсинга (который within 15 days), нужно делать всё тоже самое, только выбирать в фильтре “Date posted: Last 24 hours” и “Posted by: Employer”.



Единственное отличие в том, что теперь он каждое спарсенное название компании не сразу добавляет в БД, а смотрит есть ли у нас уже такая компания. Если компании с

таким именем нет - создает новую “карточку”. Если такая компания есть - то он обновляет столбец “дата последней вакансии” на сегодняшний день.

ВАЖНО: **не надо сразу парсить все 140.000 карточек, давайте для начала обкатаем всё на 100 карточек, и когда будет работать как нам нужно - спарсим всё за 15 дней**.

Это первая часть ТЗ. После того как сделаем - будет сразу следующая, т.к. помимо парсинга надо нужен будет дополнительный функционал работы в веб-интерфейсе с тем что спарсили.