**Общие требования**

Файл CVS без заголовка.

Кодировка: UTF-8

Собираем только активные объявления с ценой более 20 млн руб.

Локация: Санкт-Петербург

Цена: всегда делим на 1000. Т.е. если цена 40 000 000, то собираем как 40 000

Собираем данные с Яндекс, Cian, Avito, LifeDeluxe

По итогу сбора получаем 4 файла, которые потом объединяем в 1 общий файл.

Названия файлов csv: проект\_дата\_кол-во\_записей.csv

Например, для cian: cian\_ru\_25.07.2024\_1076.csv

Итоговый файл: дата\_кол-во записей.csv

**Формат файла по полям**

Названия условны, в сам файл они не добавляются (заголовки):

source - источник сбора. Указываем согласно списку ниже

Cian - Ц

Avito - Ф

Yandex - Я

LifeDeluxe - D

**market** - указываем тип жилья, вторичка - "в" или первичка/новостройка - "п"

**district** - район объекта

**rooms** - количество комнат. В основном это цифра, но может быть и "многокомнатные", "Своб. планировка", "Студия" ...

**address** - адрес объекта. В адресе обязательно в начале указывается его название и только затем тип (ул, проспект, переулок ...). Т.е должно быть "Московский проспект, 42/27", а не "проспект Московский, 42/27". Ниже укажем список замен, которые мы использовали.

**floor** - этаж. Указывается в формате какой этаж из всех этажей, например "7/9"

**flatType** - тип дома. Например: Кирпичный, Монолитный, Монолитно-кирпичный ... Есть не везде

**areaTotal** - общая площадь. Цифра. Например 88,2

**areaLive** - жилая площадь

**areaKitchen** - площадь кухни

**wc** - наличие санузна. Раздельный - Р, совместный - С. Ниже будут наши замены.

**price** - цена. Цифра. всегда делим на 1000.

**agency** - продавец/компания/агент

**phone** - номер телефона, даже если номер телефона временный, как на Авито

**additional** - описание объявления

**saleType** - ищем в тексте объявления "Свободная продажа" заменяем на "П", или альтернатива - "В"

**condition** - ищем в тексте объявления "Переуступка", заменяем на "У"

**rendEnd** - срок/год сдачи дома

**url** - url объявления без https://, без www закодированный в base64, например bS5hdml0by5ydS9zYW5rdC1wZXRlcmJ1cmcva3ZhcnRpcnkvMy1rLl9rdmFydGlyYV8xMDdfbV8z

M19ldC5fMzUzNDY5NjUwNA - это m.avito.ru/sankt-peterburg/kvartiry/3-k.\_kvartira\_107\_m\_33\_et.\_3534696504

**Coordinates** – 59.93428; 30.335099 – latitude; longitude через ;

Так как у нас поиск идет по символам адреса, нужно, чтобы в начеле был новзание улиц, а потом уже ул, наб и тд.

Ниже будут приведены наши замены, в формате "что ищем|на что заменяем"

========================

**Особенности сбора (если есть)**

**1. Yandex**

При сборе ссылок, мы собираем № района и зятем добавляем к ссылкам ([PARAM]:#), чтобы при сборе знать район.

Замены № района из ссылки на название:

417962|Курортный

417963|Петроградский

417964|Красногвардейский

417965|Колпинский

417966|Кировский

417967|Фрунзенский

417968|Петродворцовый

417969|Приморский

417970|Василеостровский

417971|Калининский

417972|Пушкинский

417973|Центральный

417974|Невский

417975|Адмиралтейский

417976|Красносельский

417977|Кронштадтский

417978|Московский

417979|Выборгский

=======================

**2. Cian**

**3. Авито**

Авито мы собирали с мобильной версии, чтобы получать телефон.

Тип жилья market мы собираем из хлебных крошек

Замены № района из ссылки на название:

762|Адмиралтейский

763|Василеостровский

764|Выборгский

765|Калининский

766|Кировский

767|Колпинский

768|Красногвардейский

769|Красносельский

772|Московский

773|Невский

774|Петроградский

776|Приморский

778|Фрунзенский

779|Центральный

776|Приморский

**4. LifeDeluxe**

Дополнительно из адресов вырезаем:

Лисий Нос,|

Левашово,|

Стрельна,|

Парголово,|

Осиновая Роща,|

Реки|

=======================

=======================

**Замены для адреса** (используем регулярные выражения). Замены отрабатываю по порядку, сверху вниз. В общем, убирается в начале все лишнее (если есть), город, поселок, территория, район ..., дальше тип адреса переставляется после названия. Например "Римского-Корсакова проспект 8/18", вместо "проспект Римского-Корсакова 8/18"

re:Санкт-Петербург[, ]|

re:[, ]+Санкт-Петербург|

re:^([^<>,]\*?) пр-т Васильевского острова([^<>,]\*?),|$1 Васильевского острова проспект,

re:^([^<>,]\*?) проспект Васильевского острова([^<>,]\*?),|$1 Васильевского острова проспект,

re:^([^<>,]\*?) пр-т Петроградской стороны([^<>,]\*?),|$1 Петроградской стороны проспект,

re:^([^<>,]\*?) проспект Петроградской стороны([^<>,]\*?),|$1 Петроградской стороны проспект,

re:^[^<>,]\*?проспект ([^<>,]\*?),|$1 проспект,

re:^[^<>,]\*?проезд ([^<>,]\*?),|$1 проезд,

re:^[^<>,]\*?аллея ([^<>,]\*?),|$1 аллея,

re:^[^<>,]\*?бульвар ([^<>,]\*?),|$1 бульвар,

re:^[^<>,]\*?набережная ([^<>,]\*?),|$1 набережная,

re:^[^<>,]\*?переулок ([^<>,]\*?),|$1 переулок,

re:^[^<>,]\*?улица ([^<>,]\*?),|$1 улица,

re:^[^<>,]\*?дорога ([^<>,]\*?),|$1 дорога,

re:^[^<>,]\*?площадь ([^<>,]\*?),|$1 площадь,

re:^[^<>,]\*?пр\. ([^<>,]\*?)$|$1 проспект

re:^[^<>,]\*?ул\. ([^<>,]\*?),|$1 улица

re:^[^<>,]\*?б-р ([^<>,]\*?),|$1 бульвар,

re:^[^<>,]\*?дорога ([^<>,]\*?),|$1 дорога,

re:, улица ([^<>,]\*?),|, $1 улица,

re:, ул. ([^<>,]\*?),|, $1 улица,

re:^ул. ([^<>,/]\*?)$|$1 улица

re:^ул. ([^<>,/]\*?) / ул. ([^<>,]\*?)$|$1 улица / $2 улица

re:^улица ([^<>,]\*?)$|$1 улица

re:^[^<>,]\*?[кК]анала ([^<>, ]\*?) ([^<>,]\*?),|$1 канала $2,

re:^[^<>,]\*?площадь ([^<>,]\*?),|$1 площадь,

re:^[^<>,]\*?пл. ([^<>,]\*?),|$1 площадь,

re:^[^<>,]\*?пр. ([^<>,]\*?),|$1 проспект,

re:^[^<>,]\*?пр-кт ([^<>,]\*?),|$1 проспект,

re:^[^<>,]\*?[рР]еки ([^<>, ]\*?) ([^<>,]\*?),|$1 реки $2,

re:^[^<>,]\*?территория ([^<>,]\*?),|$1 территория,

re:^[^<>,]\*?шоссе ([^<>,]\*?),|$1 шоссе,

re:^([\d]{1,3})-я|$1

re:^([\d]{1,3}) ([^<>,]\*?) линия|$1 линия $2

Василеостровский район, жилой комплекс|

re:Васильевского острова ([\d])|$1

re:^[^<>,]\*?жилой комплекс|

re:^[^<>]\*?жилой комплекс|

re:^[^<>,]\*?Курортный р-н, [^<>,]\*?,|

re:^[^<>,]\*?Лахта территория,|

re:^[^<>,]\*?Петродворцовый р-н, Петергоф,|

re:^[^<>,]\*?Пушкинский р-н, Павловск,|

re:^[^<>,]\*?Пушкинский р-н, пос\. [^<>,]\*?,|

re:^[^<>,]\*?Пушкинский р-н, Пушкин,|

re:^[^<>,]\*?пос. [^<>,]\*?,|

re:^[^<>,]\*?посёлок[^<>,]\*?,|

re:^[^<>,]\*?Пушкин,|

re:^[^<>,]\*?Сестрорецк,|

re:^[^<>,]\*?Центральный район,|

re:^[^<>,]\*?Шувалово территория,|

Замены для санузла:

раздельный|Р

Санузел Р|Р

совместный|С

Санузел совмещённый|С

санузлов|

**Стартовые URL для сбора ссылок объявлений приложил. Для каждого проекта отдельно.**

**Приложил отдельно пример сбора по каждому проекту отдельно и итоговый файл.**