Тестовое задание

Цель

Вам необходимо обработать сырые данные, поступившие от парсера.

Описание

В файле файл с данными две вкладки platform1, platform2. Это данные о гостиницах от двух разных платформ в одном регионе.

Поля

Поле	Описание
id	Идентификатор в БД
create_time	Дата создания
title	Название
hotel_type_original	Тип гостиницы
city	Город
address	Адрес
rating	Рейтинг
rating_5	Рейтинг по 5-ти бальной шкале
review_count	Количество отзывов
star_rating	Звездность
rooms_count	Количество номеров
contact_social	Контакты соц. сетей
description	Описание
email	email строкой, несколько значений через запятую
phone	телефон строкой, несколько значений через запятую
website	сайты строкой, несколько значений через запятую
uid	Уникальный идентификатор гостиницы на платформе, не может быть разным у одной гостиницы, и не может повторятся у разных гостиниц, но в рамках одной платформы
parsing_time	Время сбора
lat	Широта

Поле	Описание
lon	Долгота

Состав данных

- Записи могут повторятся, и самые актуальные данные идут в конце.
- Если какое-то поле по конкретной гостинице пустое в последней записи, но встречалось ранее, то необходимо его взять из более ранних записей.
- Поля телефона, email, сайтов могу содержать лишние символы, несколько записей и прочее.
- Формат записи названия, адреса, типа гостиницы отличается в разных платформах.

Задача

1. Собрать от каждой платформы финальный список гостиниц, в котором по каждой гостинице внутри платформы будет только одна запись с самыми актуальными и полными данными.
2. Почистить данные.
3. Поля телефона, email, сайтов распарсить и сохранить как списки в одинаковом формате (address@domen.org, 79234553322, domen.ru)
4. Вывести топ 10 по каждой платформе, по параметрам:
    a. больше всего телефонов,
    b. больше всего отзывов.
5. Вывести квадрат координат размером 1км на 1км, где больше всего гостиниц.
6. Задача со *, объединить данные от двух платформ, по критерию, который вы придумаете
    a. Вывести все гостиницы, которые есть в платформе 1 и нет в платформе 2
    b. Вывести топ 10 гостиниц которые есть в обоих платформах, по суммарному количеству отзывов

Требования

1. Результат должен быть представлен Jupyter notebook
2. Результат должен воспроизводится автоматически и нуля при повторном запуске, ручные правки должны быть учтены в коде
3. Комментарии приветствуются