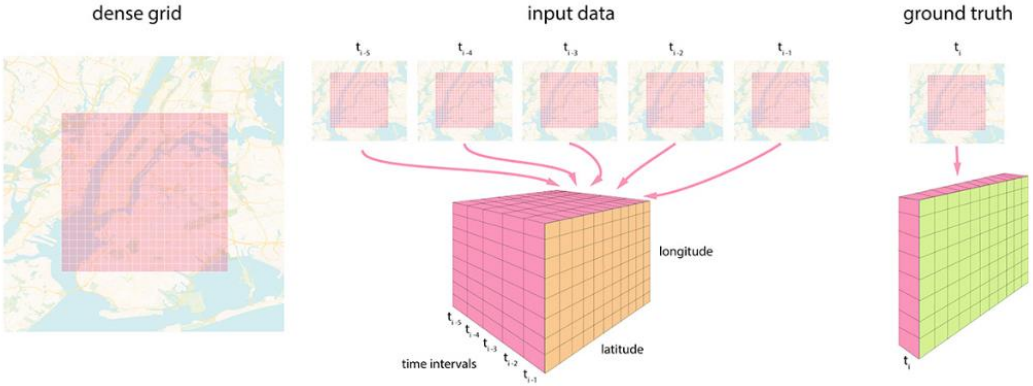
Лабораторное задание №1

Аннотация

В современной жизни человека социальные сети занимают очень важное место и оказывают сильное влияние на повседневную жизнь. Через них передается огромный поток информации, связанной с мыслями, опытом, эмоциями человека и миром, который его окружает. Исследование потоковых данных социальных сетей позволяет наблюдать процессы, происходящие в городе, предсказывать возникающие аномалии и события, позволяя своевременно на них реагировать. В этом задании вам предстоит научиться по историческим данным социальной сети о частоте публикаций в разных городских районах предсказывать ее будущее распределение.

Задание



Вам предоставлен набор данных одной популярной социальной сети, включающий в себя более 8.5 млн записей с метаинформацией публикаций за 13 месяцев (с января 2019 по февраль 2020 года).

Каждая публикация описывается следующей метаинформацией:

- **lon, lat** – координаты геопозиции с округлением до полигона 250x250 метров (географические долгота и широта, соответственно)
- **timestamp** – временная метка публикации с точностью до часа
- **likescount** – количество отметок «лайк» у публикации
- **commentscount** – количество комментариев у публикации
- **symbols_cnt** – общее количество символов в публикации
- **words_cnt** – количество слов (осмысленных, не считая спецсимволов и прочую метаинформацию)
- **hashtags_cnt** – количество хештегов
- **mentions_cnt** – количество упоминаний других пользователей
- **links_cnt** – количество ссылок
- **emoji_cnt** – количество эмодзи.
- **point** – сервисное поле для сопоставления координат из тренировочного, валидационного и тестового датасетов (если point у двух элементов совпадает, то они обладают одними и теме же координатами, сравнение lat и lon может давать погрешность)

Используя эти данные, вам необходимо будет предсказать количество публикаций в каждом полигоне 250x250 метров за каждый час на 4 недели (28 суток) вперед относительно последней публикации в тренировочном наборе.

За границы можно считать квадрат с координатами двух противоположных углов (60.0393322852, 30.5360) и (59.828, 30.142969), гарантируется что в тестовом датасете координаты не выходят за эти границы.

Примечания

- Длина набора данных позволяет учитывать как сезонный фактор изменения активности, так и общий тренд роста/снижения количества публикаций.
- Стоит обратить внимание на то, что сильное влияние на активность могут оказывать временные параметры публикаций.
- Плотность количества публикаций может сильно варьироваться от района к району.

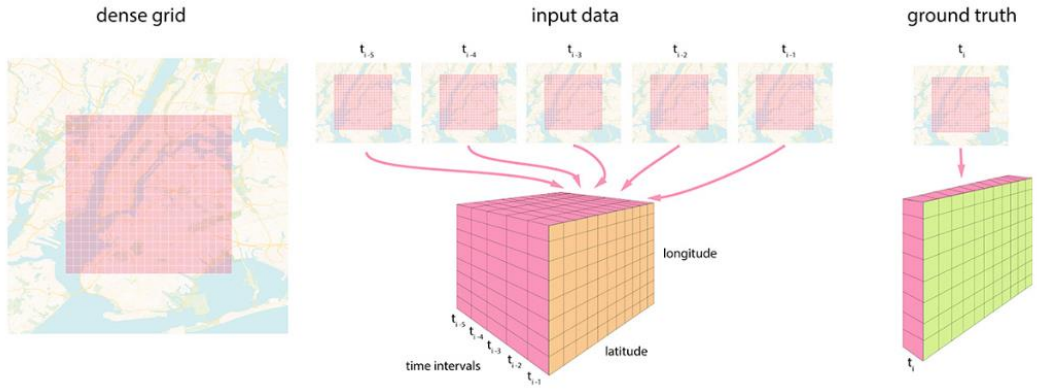
Для выполнения задания вы можете сформировать команды до трех человек (включительно). Данные для выполнения задания и baseline решения будут выложены в начале следующей недели. Обращаем ваше внимание на то, что baseline является ориентировочным и задание будет оцениваться по результатам очной защиты.

Laboratory task 1

Introduction

Social media occupies a very important place in modern human life and has a strong influence on daily life. A huge flow of information related to a person's thoughts, experiences, emotions, and the world that surrounds him/her is transmitted through them. Exploring social media streams allows you to observe processes taking place in the city, to predict emerging anomalies and events, allowing you to react to them in a timely manner. In this assignment you will learn how to use historical social network data about the frequency of publications in different urban areas to predict future distribution.

Assignment



You are presented with a dataset of one popular social network that includes more than 8.5 million records with meta-information of publications over 13 months (January 2019 to February 2020).

Each publication is described by the following meta-information:

- **lon**, **lat** – geoposition coordinates rounded up to a 250x250 meter polygon (geographical longitude and latitude, respectively)
- **timestamp** – timestamp of the publication accurate to one hour
- **likescount** – number of "likes" in the publication
- **commentscount** – number of comments of the publication
- **symbols_cnt** – number of all symbols in the publication
- **words_cnt** – number of words (meaningful, not counting special characters and other meta-information)
- **hashtags_cnt** – number of hashtags
- **mentions_cnt** – the number of mentions of other users
- **links_cnt** – number of links
- **emoji_cnt** – number of emoji.
- **point** – service field for matching coordinates from training, validation and test datasets (if two elements have the same point, they have the same coordinates, comparison of lat and lon may give an error)

Using this data, you will need to predict the number of publications in each 250x250 meter polygon for each hour 4 weeks (28 days) ahead of the last publication in the training set.

The square with coordinates of two opposite corners (60.0393322852, 30.5360) and (59.828, 30.142969) can be considered as boundaries, it is guaranteed that in the test dataset the coordinates do not exceed these boundaries.

Notes

- The length of the data set allows considering both the seasonal factor of activity changes and the general trend of growth/decline in the number of publications.
- It is worth paying attention to the fact that the time parameters of publications can have a strong influence on the activity.
- The density of the number of publications may vary greatly from district to district.

You can form teams of up to three people (inclusive) to complete the assignment. Data for the task and solutions will be posted early next week. Please note that the baseline is indicative, and the assignment will be evaluated based on the results of the in-person defense.