Big Data Processing Project Topic Proposal:

Predicting the finest Airbnb hotel:

Finding a top reviewed airbnb hotel in a city

Description:

In this project, we make use of Apache Spark and machine learning algorithms to get the top-rated and reviewed hotel based on certain features from a continuous flow of Airbnb hotel and reviews data. Along with that we also try to contextualist hotel data from different countries and predict the best one. For example, the currency of transaction or valuation is in the countries format where the hotel exists. Here we also intermix the features and data between hotel listings and reviews. In this, we are considering that the new data is being added to the storage continuously even while we predict the data.

Some of the main features of the project are as below:

- Undertaking huge data sets on listed hotels and their reviews and processing them.
- Operating with some diverse set of features like acceptance rate, neighborhood, and property type.
- Operating with a continuous flow of data.

Problem statement:

The main objective of this project is to understand the uninterrupted flow of Airbnb data and make a prediction out of it. Thus with this process, we get a top-reviewed hotel in a particular city we are endeavoring for. In general, the Airbnb data set contains a variety of features which are hard to process with the review data set consisting of mostly text data. Thus with this project, we are trying to process these huge raw data sets into a manageable and predictable data format.

The problems we will be solving for this project are as follows:

- Making the data understandable for general purposes.
- Processing the fast-paced data inputs and analyzing them.
- Making predictions out of the data present.

Predicting the finest Airbnb hotel – 603

Data source:

The data source we are considering consists of 2 data sets. One is listing data with consists of information about Airbnb hotels for 250,000+ listings in 10 major cities. It includes information about hosts, pricing, location, and room type. Next, the review data sets will have over 5 million historical reviews about the hotels we considered before. The total size of data with 2 data sets is 414.32 MB. There needs to be a lot of cleaning to be done on both the data sets as they are in raw format. The link for the data source is as follows:

https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews