kmeans-spark

This project shows example of training K-means model using spark. ClickHouse was used as a data source. Predictions were stored there either.

Dataset

OpenFoodFacts dataset consists of the descriptions of different food products. More info could be found here (https://world.openfoodfacts.org/data)

Data preparation

Data was preprocessed with removing of unimportant features and null columns filling.

Project structure

1. Research notebook (notebooks/)
2. Preprocessor (src/preprocessing/)
3. Model trainer (src/model.py)
4. Training with ClickHouse using docker-compose (docker-compose.yml)

Consider put *clickhouse-jdbc-0.4.6-all.jar* in jars (jars) folder (used for clickhouse connection).