README.md

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

Data preparation

Data was preprocessed with removing of unimportant features, null columns filling and scaling. Preprocessing was done on Scala side. Preprocessed data was obtained by model service from data mart and predictions are saved by data mart either.

Project structure

1. Research notebook
2. Preprocessor
3. Model trainer
4. Training with ClickHouse using docker-compose

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