K-Means Spark

Example of K-Means clustering using PySpark (https://spark.apache.org/docs/latest/api/python/)

Dataset

Open food facts dataset contains data about food products from all over the world. It is available on https://world.openfoodfacts.org/data

Link to csv file: https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv.gz

Clickhouse jar file

Clickhouse jar file is available on https://github.com/ClickHouse/clickhouse-java/releases/download/v0.4.6/clickhouse-jdbc-0.4.6-all.jar and should be placed in jars directory.

Example usage

```
docker-compose up
```

Project structure

- K-Means Spark (src/kmeans.py)
- Preprocessing (src/preprocessing.py)
- Clickhouse (src/clickhouse.py)
- Launching (src/main.py)
- Clickhouse and PySpark integration (docker-compose.yml)