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

Example usage

```
python src/main.py \
    --data_path=<path_to_data> \
    --save_path=<path_to_model> \
    --columns_json_path=config/columns.json \
    --k=2 \
    --max_iter=10 \
    --distance_measure="euclidian" \
    --tol=1e-4 \
    --seed=42 \
    --driver_cores=2 \
    --driver_memory="4g" \
    --executor_memory="10g"
```

Project structure

- K-Means Spark (src/kmeans.py)
- Preprocessing (src/preprocessing.py)
- Main (src/main.py)