Homework (2 weeks)

Goal to build a simple application packaged into a docker.
The application is processing some public news data.

Application Logic

1) Download public data from **(test** dataset only); AG News
   https://huggingface.co/datasets/sh0416/ag_news

2) Write code to generate a table with those columns:
   word : w
   word_count: Total frequency of <word> which appears in the column "description" of AG News dataset

   Example:
   word : "make",
   count: 457 "make" appears 457 times in the News/description.
   Each appearance of "make" must be counted (case sensitive)

   Column "word" have only 3 rows with values: ["president", "the", "Asia"]

   Save the table on disk as parquet file with this format
   "word_count_{YYYMMDD). parquet"

   Command line to generate should be:
   python src/run.py process_data -cfg config/cfg.yaml -dataset news -dirout "ztmp/data/"

3) Write code to generate another similar table with those columns:
   word : w
   count: Total frequency of <word> which appears in the column "description" of AG News

   Column "word" have all the unique word in News/ Description column
   (example: "Today this is raining" word : ['today', 'this', 'is', 'raining']

   Save the table on disk as parquet file with this format
   "word_count_all_{YYYMMDD). parquet"

   Command line to generate should be:
   python src/run.py process_data_all -cfg config/cfg.yaml -dataset news -dirout "ztmp/data/"

4) Package the application into a docker with requirements
   Run the application with command line

Code implementation constraints

Code must use **pyspark** to process the data and save on disk.
In addition to pyspark, other packages can be used/installed.
Docstring, TypeHints, Logging must be added.
Basic tests should be added (we do not ask for complex testing, just basic)
Attention to the code quality/structure is required.

Docker Requirements

+ Base OS should linux debian
+ Python environment must contain conda
+ Environment should contain those packages:

pyspark, pytorch, numpy, pandas, scipy, scikit-learn.polars, orjson, pyarrow.awswrangler, transformers, accelerate, duckdb, neo4j, s3fs, umap-learn, smart-open,onnxruntime.spacy, seqeval.gensim.numba, sqlalchemy.pytest

+ python should be 3.11
+ Additional packages can be added if needed.
+ Docker must be built using a Github Action script.

Submission

One single Zip file: {name)_YYYYMMDD.zip
Zip file should contain folders with this organization

code/

github_build_action.yml
Dockerfile.Dockerfile
script/run.sh : Bash Script to start to generate the 2 files.

Config files in code/config:
config.yaml : config file in yaml format in config/ sub-folder

Source code in code/src/

screenshots/

docker_build.png
pip_freeze.png
dataprocessed.png
data_processed_all.png

logs/

Docker_build.txt : Docker build log
pip_list.txt : pip list inside the docker
Data_processed.txt: Pipeline logs
Data_processed all.txt: Pipeline logs|

outputs/

word_count_{YYYMMDD). parquet
word_count_all_{YYYMMDD}.parquet

More screenshots and logs can be added (this is advised to add more screenshots/logs).
More files can be added if required.