Zadanie 7

Ivan Filipchuk

Apache Spark

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode, split, lower, regexp_replace,
length
from pyspark.sql.functions import desc
import sys

if __name__ == "__main__":
    if len(sys.argv) != 4:
        print("Usage: spark_wordcount_no_rdd <input_file_path> <word_length>
<min_word_length>")
        sys.exit(1)

    input_file_path = sys.argv[1]
    word_length = int(sys.argv[2])
    min_word_length = int(sys.argv[3])
    spark = SparkSession.builder \
        .appName("WordCountNoRDD") \
        .getOrCreate()
    lines = spark.read.text(input_file_path)
    words = lines.select(explode(split(lines.value, "\s+")).alias("word")) \
                .withColumn("word", lower(regexp_replace("word", "[^a-zA-Z0-
9]", "")))

    words_filtered = words.filter(length("word") == word_length)
    word_counts = words_filtered.groupBy("word").count()
    sorted_word_counts = word_counts.orderBy(desc("count"))
    print(f"Words with length {word_length}, sorted by occurrence:")
    sorted_word_counts.show()
    words_min_length = words.filter(length("word") >= min_word_length)
    word_counts_min_length = words_min_length.groupBy("word").count()
    sorted_word_counts_min_length =
word_counts_min_length.orderBy(desc("count"))
    print(f"\nWords with minimum length {min_word_length}, sorted by
occurrence:")
    sorted_word_counts_min_length.show()
    spark.stop()
```

Dockerfile

```
FROM apache/spark:latest
COPY wordcount.py /
ENTRYPOINT ["/opt/spark/bin/spark-submit", "--master", "local[*]",
"/wordcount.py"]
CMD []
```

docker build -t spark-wordcount-no-rdd .

docker run -it -v /c/Users/filip/Desktop/Ztp-lab6/spark-wordcount-no-rdd/example.txt:/input.txt -v /c/Users/filip/Desktop/Ztp-lab6/spark-wordcount-no-rdd/results:/output spark-wordcount-no-rdd /input.txt 3 5

```
Words with length 3, sorted by occurrence:
+----+-----+
|word|count|
+----+-----+
| the|    2|
| see|    2|
| for|    1|
| how|    1|
| out|    1|
| use|    1|
| and|    1|
+----+-----+


Words with minimum length 5, sorted by occurrence:
```

```
+----------+-----+
|      word|count|
+----------+-----+
|   appears|    1|
|   program|    1|
|     count|    1|
|    should|    1|
|     times|    1|
|   counted|    1|
|   testing|    1|
|separately|    1|
|    sample|    1|
|   results|    1|
+----------+-----+
```