Laboratorium 7

Maisuradze Yurii

Spark

Kod aplikacji:

Example.txt:

```
test test
ex-librist ex-librist ex-librist
tesasdasdas asd sadsad sad
test exocannibalism
exobiological
asdsadas
tt
sad sad adsfsadfdsaf
exocannibalism exocannibalism exocannibalism exocannibalism exocannibalism
exobiological
adsfsadfdsaf
untightening untightening untightening untightening untightening sd
untightening untightening untightening
```

Main.py:

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

def count_words(file_path, word_length):
    spark = SparkSession.builder \
        .appName("WordCount") \
        .getOrCreate()
    df = spark.read.text(file_path)
    words = df.select(explode(split(regexp_replace(lower(df.value), "[^a-
zA-Z0-9\\s]", ""), "\\s+")).alias("slowo")) \
              .filter(length("slowo") >= word_length)
    word_counts = words.groupBy("slowo").count()
    sorted_word_counts = word_counts.orderBy("count", ascending=False)
    sorted_word_counts.show(sorted_word_counts.count(), False)
    spark.stop()

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("Usage: main.py <file_path> <min_word_length>")
        sys.exit(1)

    file_path = sys.argv[1]
    word_length = int(sys.argv[2])

    count_words(file_path, word_length)
```

Konfiguracja dockera

Dockerfile:

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

Lista komend dockera używanych do uruchomienia zadania

main.py, Dockerfile oraz example.txt są w jednym folderze

docker build -t spark_words .

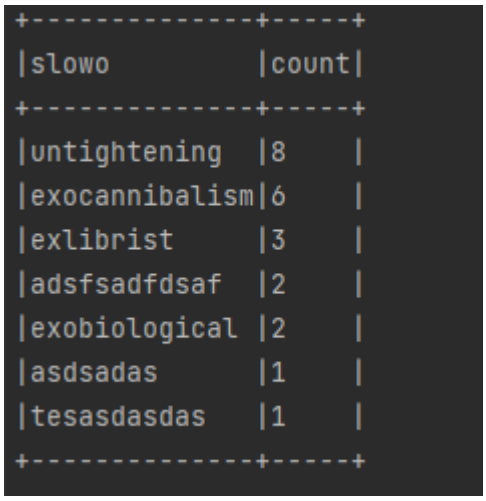
docker run --rm -v "C:/Users/maiso/Desktop/Studia 2 stopien/I sem/ZTP/Lab7/spark:/data" spark_words "/data/example.txt" 3

Zrzuty ekranu pokazujące działanie programu

Dla długości 3:

```
+--------------+-----+
|slowo         |count|
+--------------+-----+
|untightening  |8    |
|exocannibalism|6    |
|sad           |3    |
|exlibrist     |3    |
|test          |3    |
|adsfsadfdsaf  |2    |
|exobiological |2    |
|asd           |1    |
|sadsad        |1    |
|asdsadas      |1    |
|tesasdasdas   |1    |
+--------------+-----+
```

Dla długości 7:



slowo	count
untightening	8
exocannibalism	6
exlibrist	3
adsfsadfdsaf	2
exobiological	2
asdsadas	1
tesasdasdas	1