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:

Obraz zawierający tekst, zrzut ekranu, Czcionka, design

Opis wygenerowany automatycznie

Dla długości 7:

Obraz zawierający tekst, zrzut ekranu, Czcionka, design

Opis wygenerowany automatycznie