Seekho Bigdata
Data is the New Oil

PySpark Coding Practice Day 1

Seekho Bigdata Institute

Kal ki Soch, Aaj ki Shiksha

Data is the New Oil

JOIN NOW

https://www.seekhobigdata.com/
9988454737

Find the top N most frequent words in a large text file

Sample Data

```
Hello world
Hello from PySpark
PySpark is awesome
Hello PySpark world
```

1. **Load the Data:** Read the text file into a DataFrame or RDD.
2. **Tokenize the Text:** Split the text into words.
3. **Count Word Frequencies:** Count the occurrences of each word.
4. **Sort and Extract Top N Words:** Sort the words by frequency and extract the top N.

PySpark

```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode, split, col
from pyspark.sql.types import StringType

# Initialize Spark session
spark = SparkSession.builder \
    .appName("Top N Frequent Words") \
    .getOrCreate()

# Load the data
file_path = "path/to/sample.txt"
df = spark.read.text(file_path)

# Tokenize the text into words
words_df = df.select(explode(split(col("value"), " ")).alias("word"))

# Convert to lower case for case insensitivity
words_df = words_df.withColumn("word", col("word").lower())

# Count word frequencies
word_counts_df = words_df.groupBy("word").count()

# Sort by frequency in descending order
sorted_word_counts_df = word_counts_df.orderBy(col("count").desc())

# Extract top N words (e.g., top 3)
top_n = 3
top_words_df = sorted_word_counts_df.limit(top_n)

# Show the results
top_words_df.show()

# Stop the Spark session
spark.stop()
```

Explanation

Initialization: Create a Spark session.

Load Data: Read the text file into a DataFrame.

Tokenize Text: Use split to break text into words and explode to flatten the array into rows.

Normalize Case: Convert all words to lowercase to ensure case-insensitive counting.

Count Frequencies: Group by word and count occurrences.

Sort and Limit: Sort by frequency and limit the results to the top N words.

Show Results: Display the top words.
Replace "path/to/sample.txt" with the actual path to your text file.

SB

Seekho Bigdata

Data is the New Oil

https://www.seekhobigdata.com/