Dedeepya Palakurthi
XZ74495

Assignment 2 - MapReduce

Data Extraction :

Select the book which corresponds to your birth month. For birth month 8-12, divide by 2 and round up.

Once you selected the book, go to page number that corresponds to your birth date (1-31) and extract next 10 pages of the book to a text file (file1.txt).

Next, go to page number that corresponds to your birth year (last 2 digits). For year 2000 onwards, use 1 infront of the year number to find the page number (so year 2000 becomes 100, 2001 - 101 and so on). Extract next 10 pages into another text file (file2.txt).

data_extraction.py

```
import PyPDF2

# Open the PDF file
with open("C:\\Users\\hp\\OneDrive\\Desktop\\UMBC\\UMBC-DS\\Semester2 -
Spring 2024 (29 Jan - 27 May)\\603 - Platforms for Big Data
Processing\\Assignments\\Assignment 2 (14-03-2024,
Thu)\\Harry_Potter_(www.ztcprep.com).pdf", 'rb') as file:
    pdf_reader = PyPDF2.PdfReader(file)

    # Get the total number of pages in the PDF
    num_pages = len(pdf_reader.pages)

    # Store the data of birth
    # My DOB: 09-August-2001
    birth_month = 8
    birth_date = 9
    birth_year = 2001

    # Calculate the book number based on birth month
    selected_book = (birth_month + 1) // 2

    # Calculate the page number for birth date
    birth_date_page = birth_date - 1

    # Calculate the page number for birth year
    birth_year_page = int(str(birth_year % 100).zfill(2))

    # Extract text from the selected pages and write to new files
    with open('file.txt', 'w') as file1, open('file2.txt', 'w') as file2:
        # Extract pages for birth date
        for page_num in range(birth_date_page, min(num_pages,
birth_date_page + 10)):
            page = pdf_reader.pages[page_num]
            file1.write(page.extract_text())

        # Extract pages for birth year
        for page_num in range(birth_year_page, min(num_pages,
birth_year_page + 10)):
            page = pdf_reader.pages[page_num]
            file2.write(page.extract_text())
```

Dedeepya Palakurthi
XZ74495

Output files :

file1.txt file2.txt

1. Write Python code and use MapReduce to count occurrences of each word in the first text file (file.txt). How many times each word is repeated?

mapper1.py

```
import sys

# Read entire line from STDIN (standard input)
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()
    # Split the line into words
    words = line.split()
    # Assign count one to each word
    for word in words:
        word = word.replace('\"', "")
        word = word.replace(",", "")
        word = word.replace(";", "")
        word = word.replace("!", "")
        word = word.replace("-", "")
        word = word.replace("_", "")
        word = word.replace("?", "")
        word = word.replace("|", "")
        print(f'{word}\t{1}')
```

reducer1.py

```
import sys

# Initialize variables to store previous and current word counts
previous_word = None
previous_count = 0
current_word = None

# Read input from standard input (STDIN)
for line in sys.stdin:
    # Strip whitespace and split the line into word and count
    line = line.strip()
    current_word = line.split('\t')[0]
    count = 1

    # If the current word is the same as the previous word, update the
count
    if previous_word == current_word:
        previous_count += int(count)
    else:
        # If the current word is different from the previous word,
        # print the previous word and its count
        if previous_word:
```

```
            print(f'{previous_word}\t{previous_count}')
        # Reset the count and update the previous word to the current word
        previous_count = count
        previous_word = current_word

# Print the last word and its count
if previous_word == current_word:
    print(f'{previous_word}\t{previous_count}')
```

Commands to run MapReduce using Python :

```
type file1.txt | python mapper1.py | sort | python reducer1.py >
file1_output.txt
```

Commands to run MapReduce using Hadoop :

```
C:\Windows\System32\cmd.e

Microsoft Windows [Version 10.0.22631.3155]
(c) Microsoft Corporation. All rights reserved.

C:\DP\hadoop>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
starting yarn daemons
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.

C:\DP\hadoop>jps
22912 RunJar
22600
26776 NodeManager
12892 DataNode
16716 Jps
22780 ResourceManager
27980 NameNode

C:\DP\hadoop>hdfs dfs -ls
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
ls: `.': No such file or directory

C:\DP\hadoop>hdfs dfs -ls \
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 4 items
drwxr-xr-x   - hp supergroup          0 2024-03-13 12:43 /non_english_word_count_in_python
drwxr-xr-x   - hp supergroup          0 2024-03-13 12:00 /tmp
drwxr-xr-x   - hp supergroup          0 2024-03-07 11:52 /training
drwxr-xr-x   - hp supergroup          0 2024-03-13 11:58 /word_count_in_python

C:\DP\hadoop>hdfs dfs -put "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\fil
e1.txt" /word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.

C:\DP\hadoop>hdfs dfs -put "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\fil
e2.txt" /non_english_word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
put: '/non_english_word_count_in_python/file2.txt': File exists

C:\DP\hadoop>hdfs dfs -rm -r /non_english_word_count_in_python/file2.txt
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Deleted /non_english_word_count_in_python/file2.txt

C:\DP\hadoop>hdfs dfs -put "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\fil
e2.txt" /non_english_word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
```

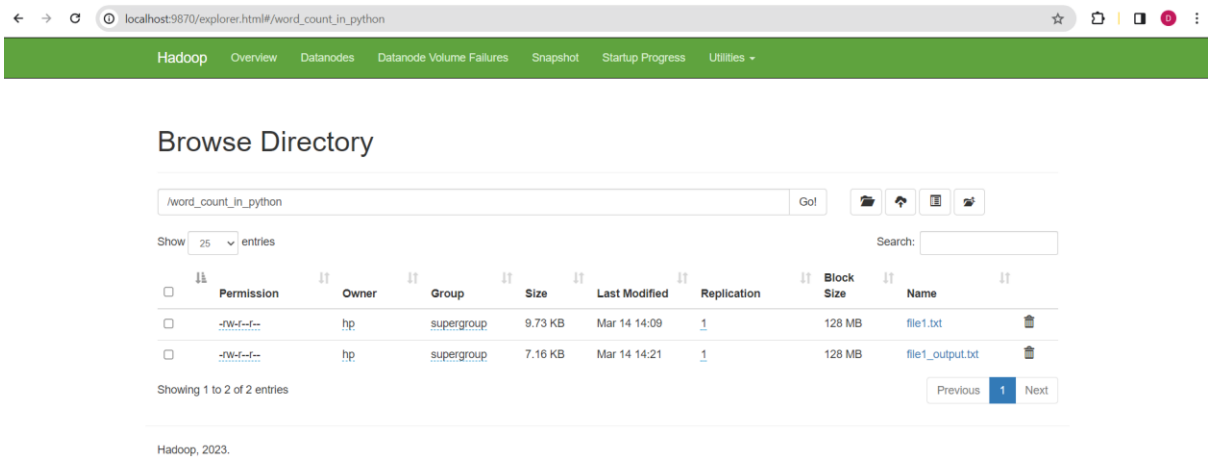
Dedeepya Palakurthi
XZ74495

```
C:\Windows\System32\cmd.e

operable program or batch file.
Deleted /word_count_in_python/file.txt

C:\DP\hadoop>hdfs dfs -ls \word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 1 items
-rw-r--r--   1 hp supergroup       9965 2024-03-14 14:09 /word_count_in_python/file1.txt

C:\DP\hadoop>hdfs dfs -ls \non_english_word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 1 items
-rw-r--r--   1 hp supergroup       8082 2024-03-14 14:11 /non_english_word_count_in_python/file2.txt

C:\DP\hadoop>icacls "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\mapper1.py
" /grant Everyone : F
Invalid parameter "Everyone"

C:\DP\hadoop>icacls "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\mapper1.py
" /grant Everyone:F
processed file: C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\mapper1.py
Successfully processed 1 files; Failed processing 0 files

C:\DP\hadoop>icacls "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\reducer1.p
y" /grant Everyone:F
processed file: C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\reducer1.py
Successfully processed 1 files; Failed processing 0 files

C:\DP\hadoop>icacls "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\mapper2.py
" /grant Everyone:F
processed file: C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\mapper2.py
Successfully processed 1 files; Failed processing 0 files

C:\DP\hadoop>icacls "C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\reducer2.p
y" /grant Everyone:F
processed file: C:\Users\hp\OneDrive\Desktop\UMBC\UMBC-DS\Semester2 - Spring 2024 (29 Jan - 27 May)\603 - Platforms for Big Data Processing\Assignments\Assignment 2 (14-03-2024, Thu)\reducer2.py
Successfully processed 1 files; Failed processing 0 files

C:\DP\hadoop>hadoop jar "C:/DP/hadoop-streaming-2.7.3.jar" -input /word_count_in_python/file1.txt -output /word_count_in_python/file1_output/ -mapper "C:/Users/hp/OneDrive/Desktop/UMBC/UMBC-DS/Semes
ter2 - Spring 2024 (29 Jan - 27 May)/603 - Platforms for Big Data Processing/Assignments/Assignment 2 (14-03-2024, Thu)/mapper1.py" -reducer "C:/Users/hp/OneDrive/Desktop/UMBC/UMBC-DS/Semester2 - Sp
ring 2024 (29 Jan - 27 May)/603 - Platforms for Big Data Processing/Assignments/Assignment 2 (14-03-2024, Thu)/reducer1.py"
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
packageJobJar: [/C:/Users/hp/AppData/Local/Temp/hadoop-unjar4886256257762258814/] [] C:\Users\hp\AppData\Local\Temp\streamjob964259201439859293.jar tmpDir=null
2024-03-14 14:16:22,495 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-14 14:16:22,844 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-03-14 14:16:23,607 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hp/.staging/job_1710435331578_0002
2024-03-14 14:16:23,948 INFO mapred.FileInputFormat: Total input files to process : 1
2024-03-14 14:16:24,031 INFO mapreduce.JobSubmitter: number of splits:2
2024-03-14 14:16:24,182 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1710435331578_0002
2024-03-14 14:16:24,182 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-03-14 14:16:24,356 INFO conf.Configuration: resource-types.xml not found
2024-03-14 14:16:24,356 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-03-14 14:16:24,609 INFO impl.YarnClientImpl: Submitted application application_1710435331578_0002
2024-03-14 14:16:24,656 INFO mapreduce.Job: The url to track the job: http://DESKTOP-697REVG:8088/proxy/application_1710435331578_0002/
2024-03-14 14:16:24,656 INFO mapreduce.Job: Running job: job_1710435331578_0002
```

Output file of MapReduce :

file1_output.txt

Hadoop Cluster in Web UI :



Dedeepya Palakurthi
XZ74495

2. From the second text file (file2.txt), write Python code and use MapReduct to count how many times non-English words (names, places, spells etc.) were used. List those words and how many times each was repeated.
There are multiple ways of doing this. You can use
pyenchant (https://pypi.org/project/pyenchant/),
pyspellchecker (https://pyspellchecker.readthedocs.io/en/latest/) or just download a list of words (http://www.gwicks.net/dictionaries.htm) and search through them.

mapper2.py

```
import sys
import re
import enchant

#Initialize the English dictionary
english_dict = enchant.Dict("en_US")

# Read entire line from STDIN (standard input)
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()
    # Split the line into words
    words = line.lower().split(" ")
    # Assign count one to each word
    for word in words:
        word = word.replace('\"', "")
        word = word.replace(",", "")
        word = word.replace(";", "")
        word = word.replace("!", "")
        word = word.replace("-", "")
        word = word.replace("_", "")
        word = word.replace("?", "")
        word = word.replace("|", "")
        if word != "":
            if not english_dict.check(word):
                print(f"{word}\t1")
```

reducer2.py

```
import sys

# Initialize variables to store previous word and count
previous_word = None
previous_count = 0

# Read input from standard input (STDIN) line by line
for line in sys.stdin:
    # Remove leading and trailing whitespace from the line
    line = line.strip()

    # Split the line into word and count based on tab delimiter
    current_word = line.split('\t')[0]
    count = 1  # Since each line represents one word, count is always 1

    # Check if the current word is the same as the previous word
```

```
    if previous_word == current_word:
        # If the current word is the same, increment the count
        previous_count += int(count)
    else:
        # If the current word is different, print the previous word and its
count
        if previous_word:
            print(f"{previous_word}\t{previous_count}")

        # Update previous_word to the current word and reset count
        previous_word = current_word
        previous_count = count

# Print the last word and its count if it exists
if previous_word:
    print(f"{previous_word}\t{previous_count}")
```

Commands to run MapReduce using Python :

```
type file2.txt | python mapper2.py | sort | python reducer2.py >
file2_output.txt
```

Dedeepya Palakurthi
XZ74495

Commands to run MapReduce using Hadoop :

```
C:\Windows\System32\cmd.e

C:\DP\hadoop>hdfs dfs -ls /word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 2 items
-rw-r--r--   1 hp supergroup       9965 2024-03-14 14:09 /word_count_in_python/file1.txt
-rw-r--r--   1 hp supergroup       7335 2024-03-14 14:21 /word_count_in_python/file1_output.txt

C:\DP\hadoop>hdfs dfs -ls /non_english_word_count_in_python
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
Found 2 items
-rw-r--r--   1 hp supergroup       8082 2024-03-14 14:11 /non_english_word_count_in_python/file2.txt
-rw-r--r--   1 hp supergroup        436 2024-03-14 14:21 /non_english_word_count_in_python/file2_output.txt

C:\DP\hadoop>hdfs dfs -cat /non_english_word_count_in_python/file2_output.txt
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
à       4
Æ       2
couldnÆ 4
dayö    1
didnÆ   10
dif     1
dudley  5
dursley 25
dursleyÆ        1
dursleys        4
emeraldgreen    1
ferent  1
fic     2
fice    2
ge      4
getups  1
goodfor 1
grunnings       2
hadnÆ   2
harold. 1
harvey  1
j.k.    4
knowwho 1
mr      20
mrs.    9
ö       1
ôdonÆ   1
ôlittle 1
openmouthed     1
ôsorry  1
ôthe    1
```

Output file of MapReduce :

file2_output.txt

Dedeepya Palakurthi
XZ74495

Hadoop Cluster in Web UI :