Practical - 5

Name: Divya Mahur

Roll No. : 18BCE106

Aim :

⇨ To Apply MapReduce algorithms to find phrase frequency from the given dataset.

Mapper Code :

```
import sys

def clean(word):
    for symbol in ['\n', '\t', ',', '.', ';', '?', '-', '"', "'", '(', ')', "!", ":", "="]:
        word = word.replace(symbol, '')
    return word.lower()

for line in sys.stdin:
    line = line.strip()
    words = line.split()

    for word in words:
        word = clean(word)
        print('%s\t%s' % (word, 1))
```

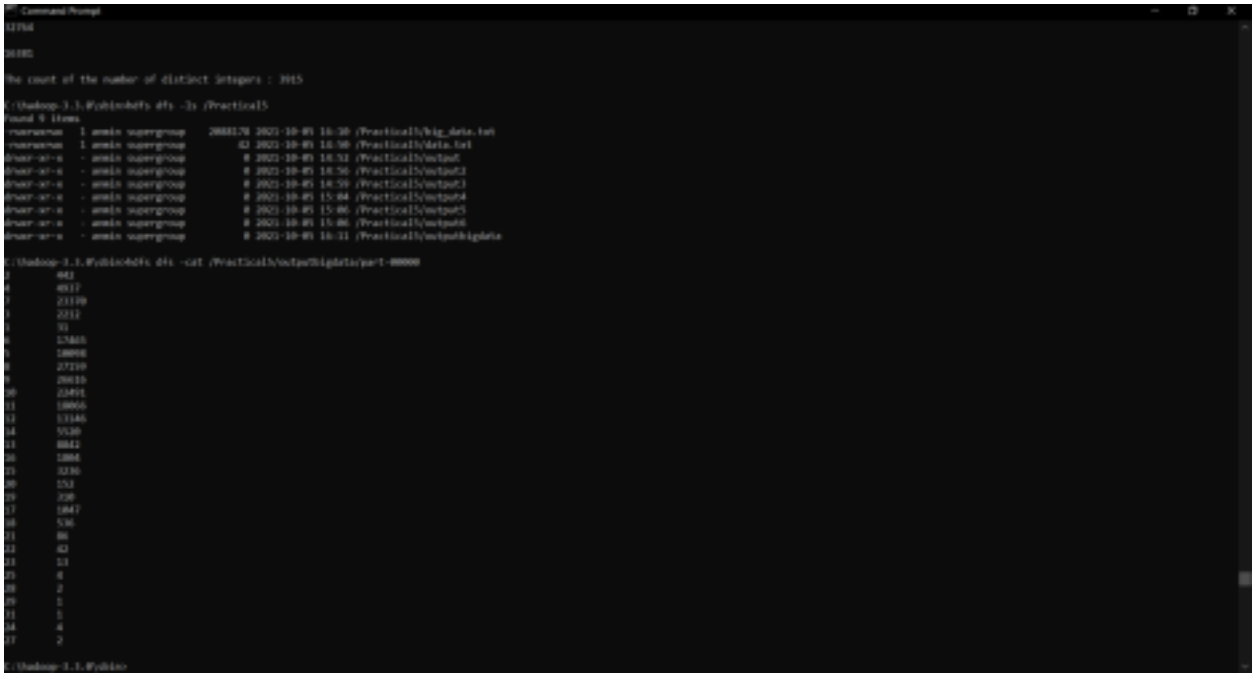
Reducer Code :

```python
import sys

dic = {}

for line in sys.stdin:
    line = line.strip()
    word,count = line.split('\t', 1)

    this_len = len(word)

    if this_len in dic:
        dic[this_len]+=1
    else:
        dic[this_len]=1

for key,val in dic.items():
    print(key,'\t',val)
```

Input :

⇨ Taken 187k+ words for this practical.

Output :



```
The count of the number of distinct integers : 3015

Found 9 items
-rwxrwxrwx   1 amein supergroup    2888178 2021-10-05 16:30 /Practical5/big_data.txt
-rwxrwxrwx   1 amein supergroup         82 2021-10-05 16:50 /Practical5/data.txt
drwxr-xr-x   - amein supergroup          0 2021-10-05 16:52 /Practical5/output
drwxr-xr-x   - amein supergroup          0 2021-10-05 16:56 /Practical5/output2
drwxr-xr-x   - amein supergroup          0 2021-10-05 16:59 /Practical5/output3
drwxr-xr-x   - amein supergroup          0 2021-10-05 15:04 /Practical5/output4
drwxr-xr-x   - amein supergroup          0 2021-10-05 15:06 /Practical5/output5
drwxr-xr-x   - amein supergroup          0 2021-10-05 15:06 /Practical5/output6
drwxr-xr-x   - amein supergroup          0 2021-10-05 16:31 /Practical5/output6bigdata

C:\Hadoop-3.3.0\sbin>hdfs dfs -cat /Practical5/output6bigdata/part-00000
1       461
2       4917
3       23170
4       2211
5       51
6       17465
7       18098
8       27219
9       20015
10      23491
11      18066
12      13146
13      9520
14      6842
15      3886
16      3236
17      153
18      238
19      1847
20      536
21      86
22      43
23      13
25      4
28      2
29      1
31      1
34      4
37      2
```

Conclusion

⇨ I learned how Map/Reduce can be used to find phrase frequency from the given dataset.