HW4

CS 5665

Tutorial referred : http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/ , also discussed with Bhagyashree about the approaches to follow for questions 2-c and 2-d.

Overview

In this homework, you will write Map and Reduce functions to perform following two tasks:

Task 1: Word Count

A) Given the provided file (Tolstoy's War and Peace), create a complete count of each word that appears in the text. Which word appears the most?

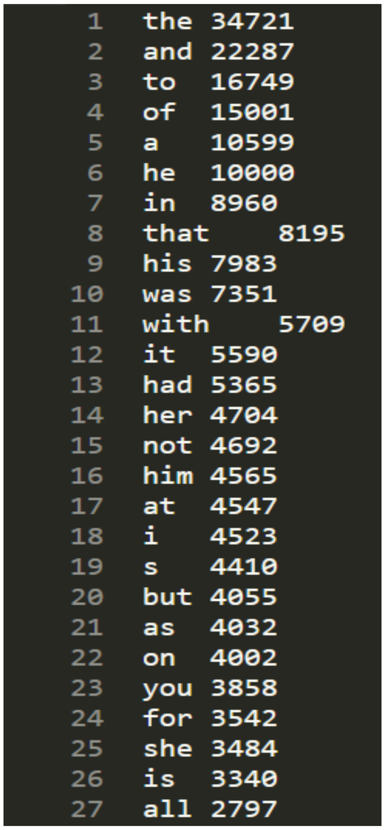
Ans: - Word "the" appears the most in the document warandpeace.txt:

Command used to copy the file into hdfs: hdfs dfs -copyFromLocal warandpeace.txt

Command used to run the mapper and reducer on Hadoop:

```
hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-2.6.0-mr1-cdh5.8.0.jar -Dmapred.reduce.task=1 -file /home/cloudera/Desktop/mapper.py /home/cloudera/Desktop/reducer.py -mapper "python mapper.py" -reducer "python reducer.py" -input /user/cloudera/warandpeace.txt -output /user/cloudera/output1.a
```

Converted the output file generated(part-00000) to output1.a.txt for better visualization and readability.Output1.a:

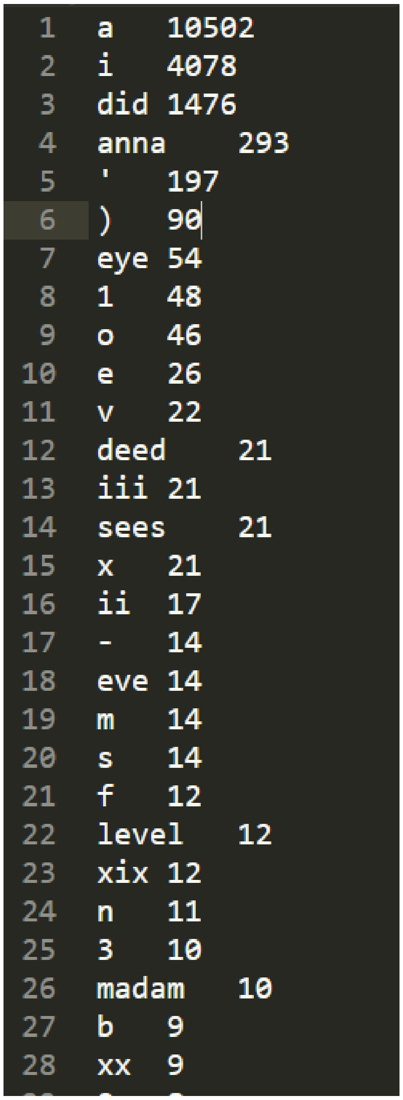


```
the	34721
and	22287
to	16749
of	15001
a	10599
he	10000
in	8960
that	8195
his	7983
was	7351
with	5709
it	5590
had	5365
her	4704
not	4692
him	4565
at	4547
i	4523
s	4410
but	4055
as	4032
on	4002
you	3858
for	3542
she	3484
is	3340
all	2797
```

B) Create a count of all the palindromes that occur in the text. Which palindrome occurs most often?

Ans :- The most occurring palindrome in the file is : “a”.

Converted the output file generated(part-00000) to output1.b.txt for better visualization and readability. Output1.b:



```
a   10502
i   4078
did 1476
anna    293
'   197
)   90
eye 54
1   48
o   46
e   26
v   22
deed    21
iii 21
sees    21
x   21
ii  17
-   14
eve 14
m   14
s   14
f   12
level   12
xix 12
n   11
3   10
madam   10
b   9
xx  9
```

Task 2: Election Fraud

In this task your job is to investigate whether there was election fraud in 2008. You have 2006 and 2008

election data files: (i) 2006 data file; and (ii) 2008 data file. The files are of the format where each line is a vote

in the election.

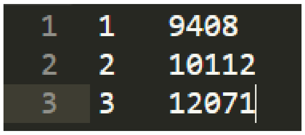
The format of the text file is:

VoterID \t CountyID \t PartyID

A) Which party won the election in 2008?

Ans:- **Party 3** won the election in 2008.

Converted the output file generated(part-00000) to output2.a.txt for better visualization and readability. Output2.a:



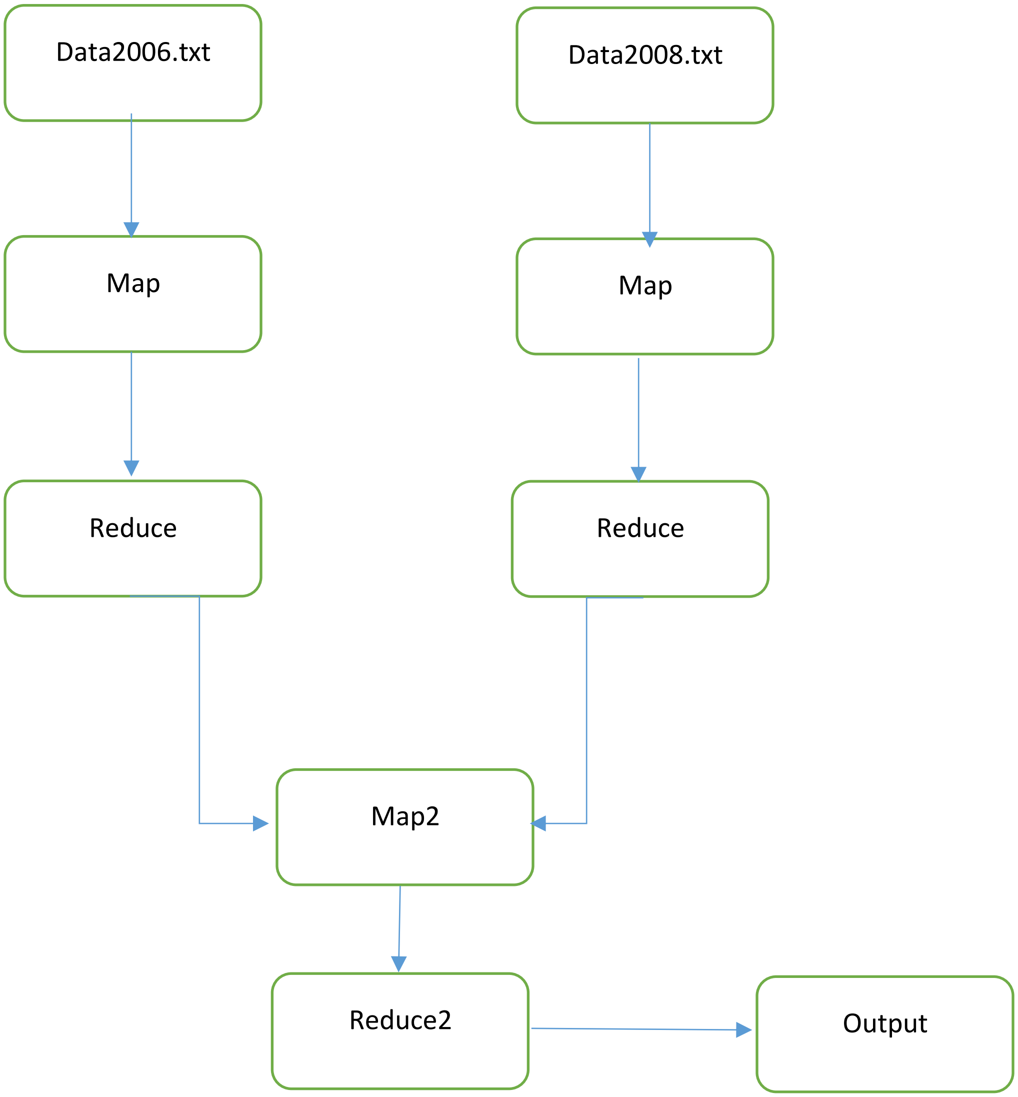
```
1	9408
2	10112
3	12071
```

B) In 2006, which county was the most monolithic in the manner in which they voted? (i.e. which county came

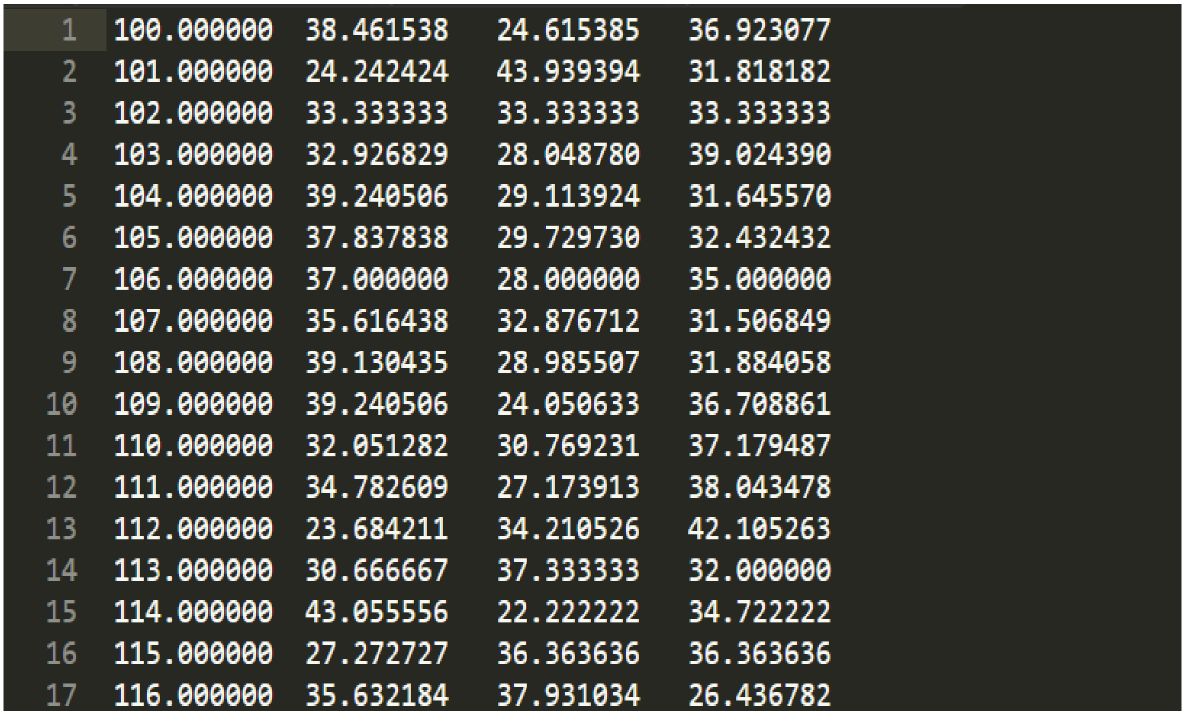
closest to voting 100% for a single party).

Ans :- County # 277 was most monolithic in the manner they voted with party 3 getting 51.6% votes.

Hadoop architecture used to solve this problem:



Converted the output file generated(part-00000) to output2.a.txt for better visualization and readability. Output2.a:



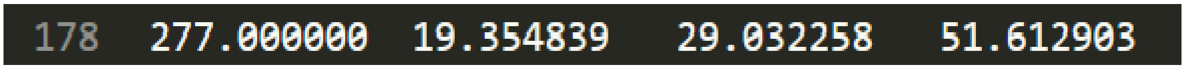
```
100.000000  38.461538   24.615385   36.923077
101.000000  24.242424   43.939394   31.818182
102.000000  33.333333   33.333333   33.333333
103.000000  32.926829   28.048780   39.024390
104.000000  39.240506   29.113924   31.645570
105.000000  37.837838   29.729730   32.432432
106.000000  37.000000   28.000000   35.000000
107.000000  35.616438   32.876712   31.506849
108.000000  39.130435   28.985507   31.884058
109.000000  39.240506   24.050633   36.708861
110.000000  32.051282   30.769231   37.179487
111.000000  34.782609   27.173913   38.043478
112.000000  23.684211   34.210526   42.105263
113.000000  30.666667   37.333333   32.000000
114.000000  43.055556   22.222222   34.722222
115.000000  27.272727   36.363636   36.363636
116.000000  35.632184   37.931034   26.436782
```

.

.

.

.



```
277.000000  19.354839   29.032258   51.612903
```

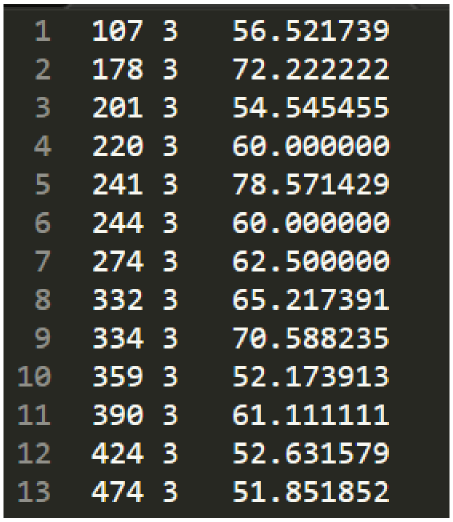
C) Studies have shown if a political party gains more than 50% in voting percentage from one election cycle to

the next, then most likely fraud has occurred. (Example, if party A received 100 votes in 2006 in county B, then

received 200 votes in 2008, fraud may have occurred). In which counties in 2008 did voter fraud likely occur?

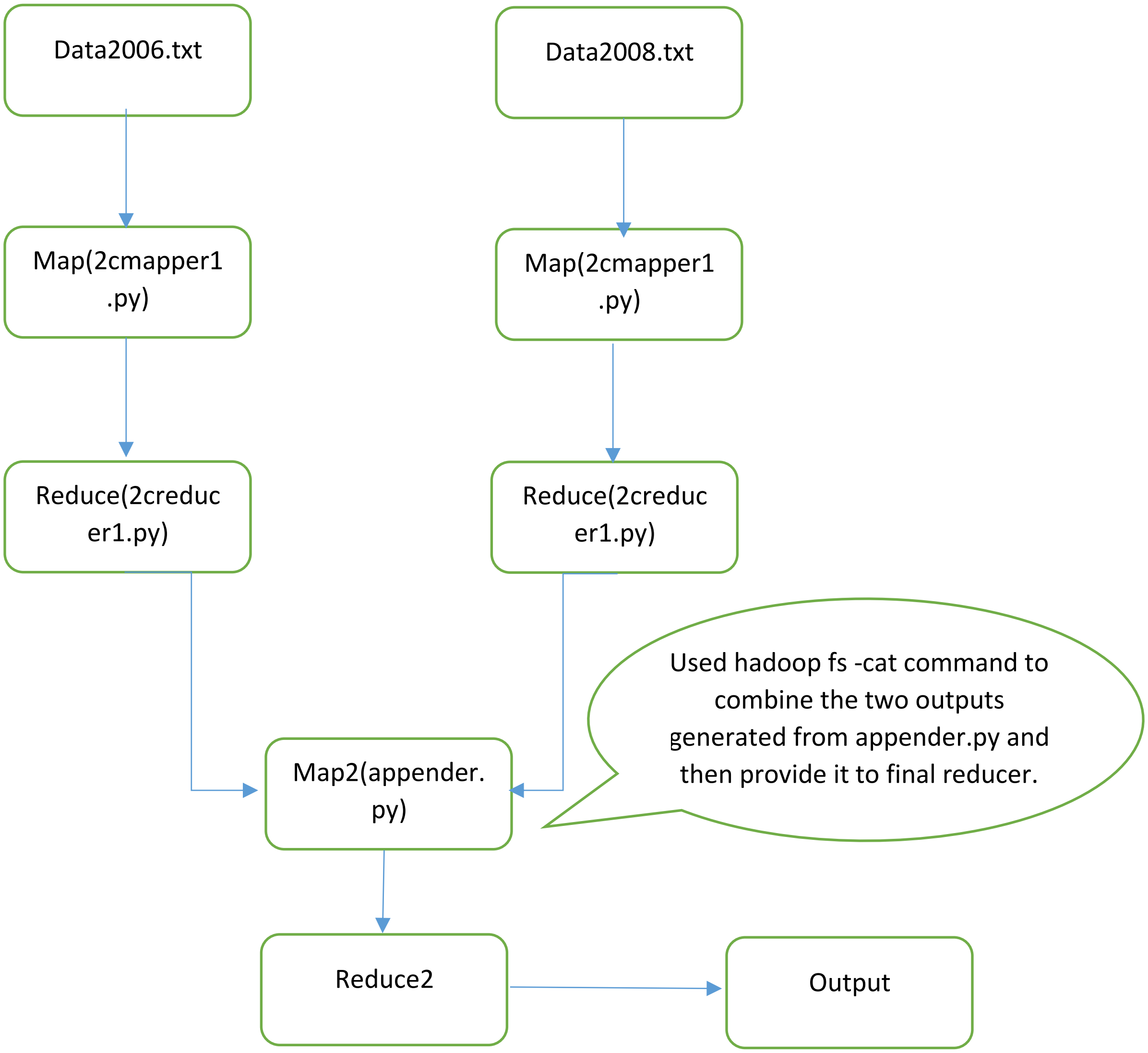
Ans :- Counties in 2008 in which fraud occurred:

Converted the output file generated(part-00000) to output2.c.txt for better visualization and readability. Output2.c:



```
107 3	56.521739
178 3	72.222222
201 3	54.545455
220 3	60.000000
241 3	78.571429
244 3	60.000000
274 3	62.500000
332 3	65.217391
334 3	70.588235
359 3	52.173913
390 3	61.111111
424 3	52.631579
474 3	51.851852
```

Hadoop architecture used to solve this problem:

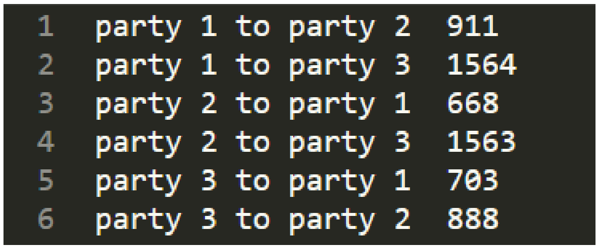


Used hadoop fs -cat command to combine the two outputs generated from appender.py and then provide it to final reducer.

D) From 2006 to 2008 how many voters changed which party they voted for? What is the most common type of change?

Ans :- 6297 voters changed the party they voted for from 2006 to 2008, most common type of change was from party 1 to party 3. Following is the output:

Converted the output file generated(part-00000) to output2.d.txt for better visualization and readability. Output2.d:



```
party 1 to party 2  911
party 1 to party 3  1564
party 2 to party 1  668
party 2 to party 3  1563
party 3 to party 1  703
party 3 to party 2  888
```

Combined the two input files into one and then processed the file as a whole using single mapper and reducer.