15826 – Multimedia Databases and Data Mining
Fall 2013, C. Faloutsos
Homework 3
Q1 – Hadoop

Emma R. Zhang
runyunz@andrew.cmu.edu

November 14, 2013

1 Code

```
1
2  echo "Q0: Clean up directory"
3  bin/hadoop fs -rmr /reddit_titles/
4  bin/hadoop fs -rmr /reddit_output/
5
6  echo "Q1: Create directory and upload dataset"
7  bin/hadoop fs -mkdir /reddit_titles/
8  bin/hadoop fs -put reddit_titles/reddit_titles.csv /
       reddit_titles/
9
10 echo "Q2: Compile WordCount"
11 cd ngram
12 ./compile.sh
13 cd ..
14
15 echo "Q3: Run WordCount and report the number of unique words"
16 bin/hadoop jar ngram/WordCount.jar WordCount /reddit_titles/ /
       reddit_output/word_count
17 bin/hadoop fs -cat /reddit_output/word_count/part-00000 | wc -l
18
19 echo "Q4: Run NGram and Output"
20 cd ngram
21 ./compile_ngram.sh
22 cd ..
23
```

```
echo "Q5: Report number of bigrams(n=2) occur at least 100 times"
bin/hadoop jar ngram/NGram.jar NGram /reddit_titles/ /reddit_output/ngram
echo "Number of trigrams(n=3) occur at least 100 times"
bin/hadoop fs -cat /reddit_output/ngram/part-00000 | wc -l

echo "Q6: List the top 20 most common bigrams and the number of times they occur"
bin/hadoop fs -cat /reddit_output/ngram/part-00000 | sort -nrk 2 | head -n20

echo "Q7: Report number of trigrams(n=3) occur at least 100 times"
cd ngram
./compile_tringram.sh
cd ..

bin/hadoop jar ngram/TriNGram.jar TriNGram /reddit_titles/ /reddit_output/tringram
echo "Number of trigrams(n=3) occur at least 100 times"
bin/hadoop fs -cat /reddit_output/tringram/part-00000 | wc -l

echo "Q8: List the top 20 most common trigrams and the number of times they occur"
bin/hadoop fs -cat /reddit_output/tringram/part-00000 | sort -nrk 2 | head -n20
```

2 Result

```
Q3: Run WordCount and report the number of unique words
38064

Q5: Report number of bigrams(n=2) occur at least 100 times
599

Q6: ist the top 20 most common bigrams and the number of times they occur
how+i 5606
i+feel 4230
when+i 3692
this+is 3405
in+the 2149
```

```
of+the 1981
xpost+from 1614
on+the 1614
i+see 1369
feel+when 1307
in+a 1214
i+dont 1192
to+the 1147
every+time 1138
i+think 1080
on+my 1066
when+my 990
for+the 984
i+have 964
what+i 939

Q7: Report number of trigrams(n=3) occur at least 100 times
116

Q8: ist the top 20 most common trigrams and the number of times they occur
how+i+feel 3904
i+feel+when 1259
this+is+how 790
every+time+i 679
the+front+page 616
how+i+felt 595
feel+when+i 581
is+how+i 548
my+cake+day 482
this+is+what 473
when+i+see 404
i+feel+after 395
the+first+time 390
i+feel+about 361
my+reaction+when 315
i+see+a 312
one+of+my 300
for+the+first 299
xpost+from+rfunny 298
this+is+the 285
```