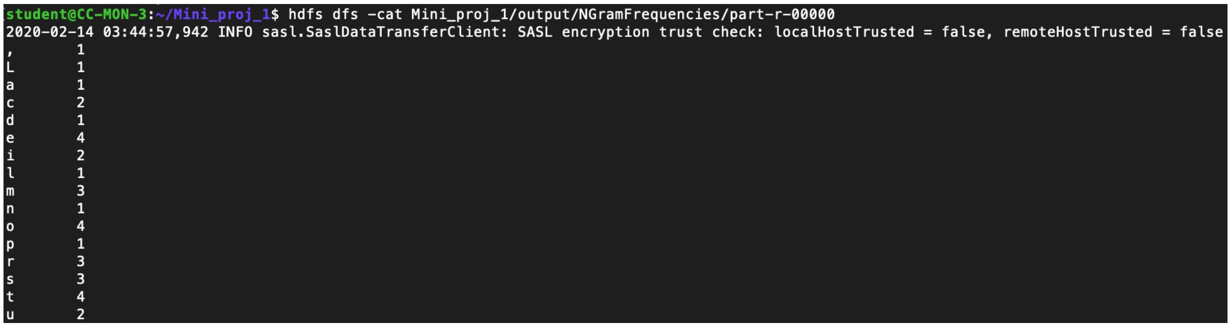
INFSCI 2750 - Miniproject 1: Outputs Part 3 and Part 4

Aishwarya Jakka, aij12@pitt.edu
Kwesi Aguillera, kra40@pitt.edu
Shreyank Ranganath, shr74@pitt.edu

==
Part 3: Hadoop program that produces the n-gram frequencies of the text
==

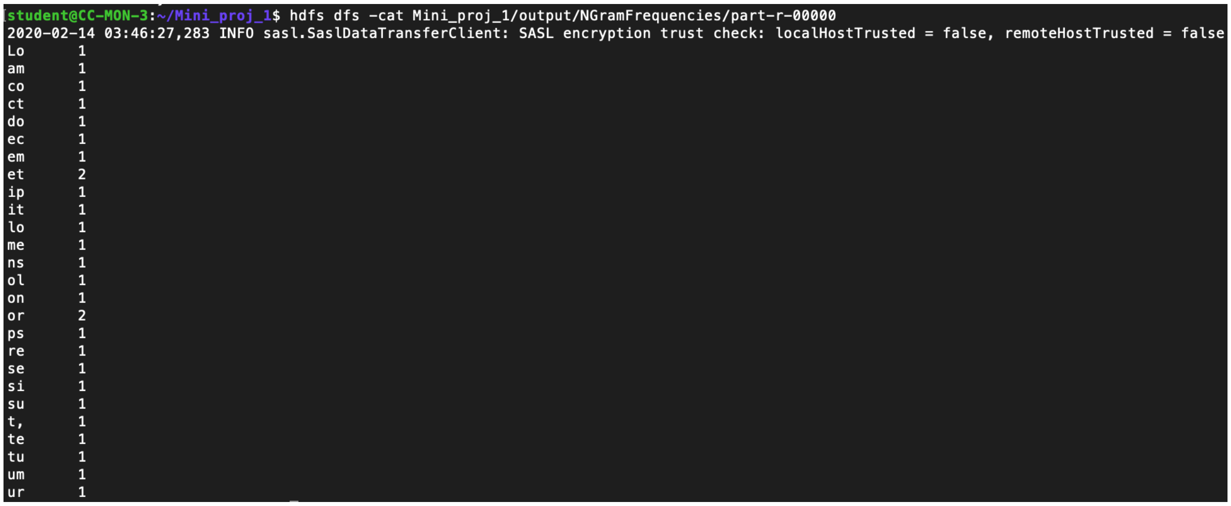
Input txt file: ngram.txt (Attached with submission): “Lorem ipsum dolor sit amet, consectetur”
N-gram frequencies output:

N=1



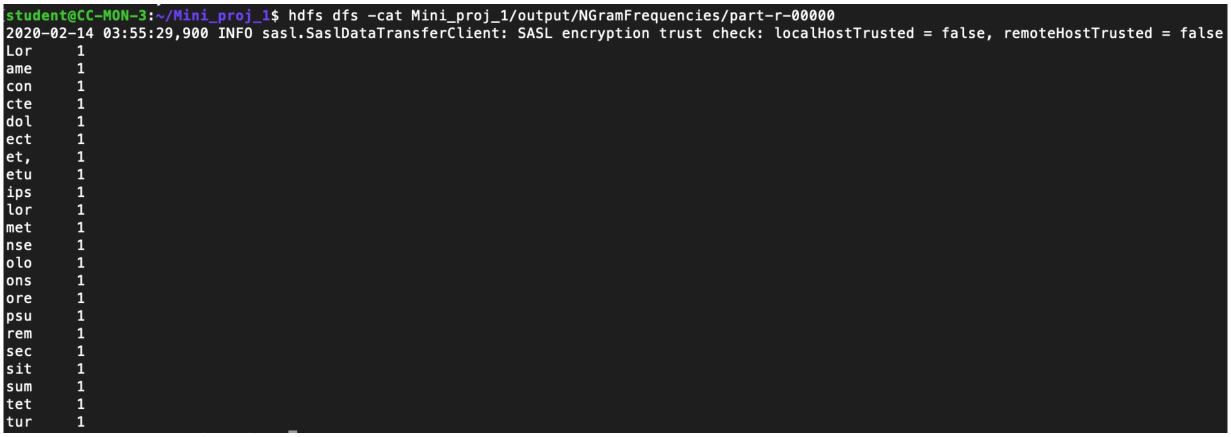
```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/NGramFrequencies/part-r-00000
2020-02-14 03:44:57,942 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
,       1
L       1
a       1
c       2
d       1
e       4
i       2
l       1
m       3
n       1
o       4
p       1
r       3
s       3
t       4
u       2
```

N=2



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/NGramFrequencies/part-r-00000
2020-02-14 03:46:27,283 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Lo      1
am      1
co      1
ct      1
do      1
ec      1
em      1
et      2
ip      1
it      1
lo      1
me      1
ns      1
ol      1
on      1
or      2
ps      1
re      1
se      1
si      1
su      1
t,      1
te      1
tu      1
um      1
ur      1
```

N=3



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/NGramFrequencies/part-r-00000
2020-02-14 03:55:29,900 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Lor	1
ame	1
con	1
cte	1
dol	1
ect	1
et,	1
etu	1
ips	1
lor	1
met	1
nse	1
olo	1
ons	1
ore	1
psu	1
rem	1
sec	1
sit	1
sum	1
tet	1
tur	1
```

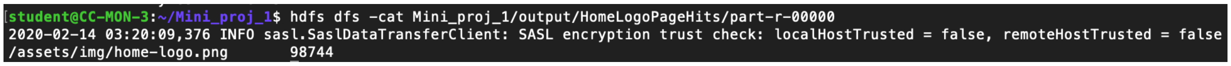
=====================================

Part 4: Hadoop program to analyze real logs

=====================================

Q1. How many hits were made to the website item “/assets/img/home-logo.png”?

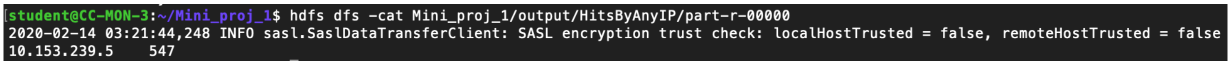
Hits: 98744



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/HomeLogoPageHits/part-r-00000
2020-02-14 03:20:09,376 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
/assets/img/home-logo.png	98744
```

Q2. How many hits were made from the IP: 10.153.239.5

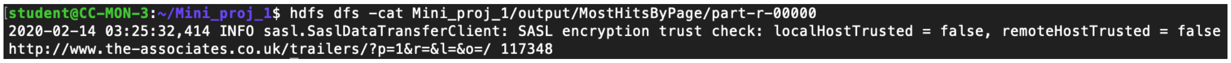
Hits: 547



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/HitsByAnyIP/part-r-00000
2020-02-14 03:21:44,248 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
10.153.239.5	547
```

Q3. Which path in the website has been hit most? How many hits were made to the path?

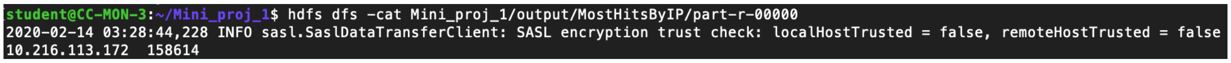
Path: http://www.the-associates.co.uk/trailers/?p=1&r=&l=&o=/
Number of hits: 117348



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/MostHitsByPage/part-r-00000
2020-02-14 03:25:32,414 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
http://www.the-associates.co.uk/trailers/?p=1&r=&l=&o=/ 117348
```

Q4. Which IP accesses the website most? How many accesses were made by it?

IP: 10.216.113.172
Number of Accesses: 158614



```
student@CC-MON-3:~/Mini_proj_1$ hdfs dfs -cat Mini_proj_1/output/MostHitsByIP/part-r-00000
2020-02-14 03:28:44,228 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
10.216.113.172	158614
```