Execution Flow for Problem 4: Co-occurring Word Matrix Generation

Step 1: Setup the Hadoop Cluster

1. **Start Hadoop Distributed File System (HDFS)** start-dfs.sh

2. **Start YARN Resource Manager** start-yarn.sh

Step 2: Prepare Data for Processing

1. **Extract Wikipedia Dump File** tar -xvzf Wikipedia-EN-20120601_ARTICLES.tar.gz

2. **Create HDFS Directory for Input Files** hadoop fs -mkdir /10000

3. **Upload Extracted Text Files to HDFS** hadoop fs -put *.txt /10000/

4. **Create Directory for Stopwords** hdfs dfs -mkdir -p /user/abhay/assignment2/stopword/

5. **Upload Stopwords File to HDFS** hdfs dfs -put stopwords.txt /user/abhay/assignment2/stopword/stopwords.txt

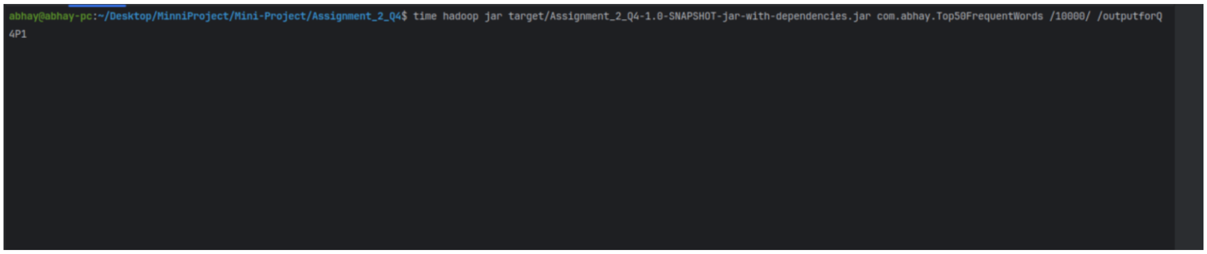
Step 3: Build the Java Program

Compile the Java project using **Maven**:

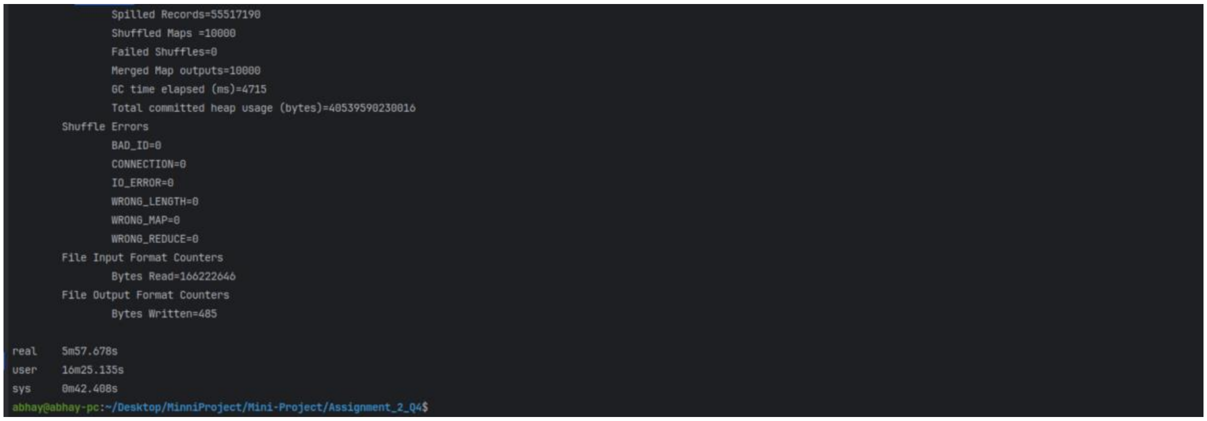
mvn clean package assembly:single

Part A: Identifying Top 50 Most Frequent Words

1. **Run the MapReduce job using the pairs approach** hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \ com.abhay.Top50FrequentWords /10000/ /outputforQ4P1



```
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$ time hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar com.abhay.Top50FrequentWords /10000/ /outputforQ
4P1
```

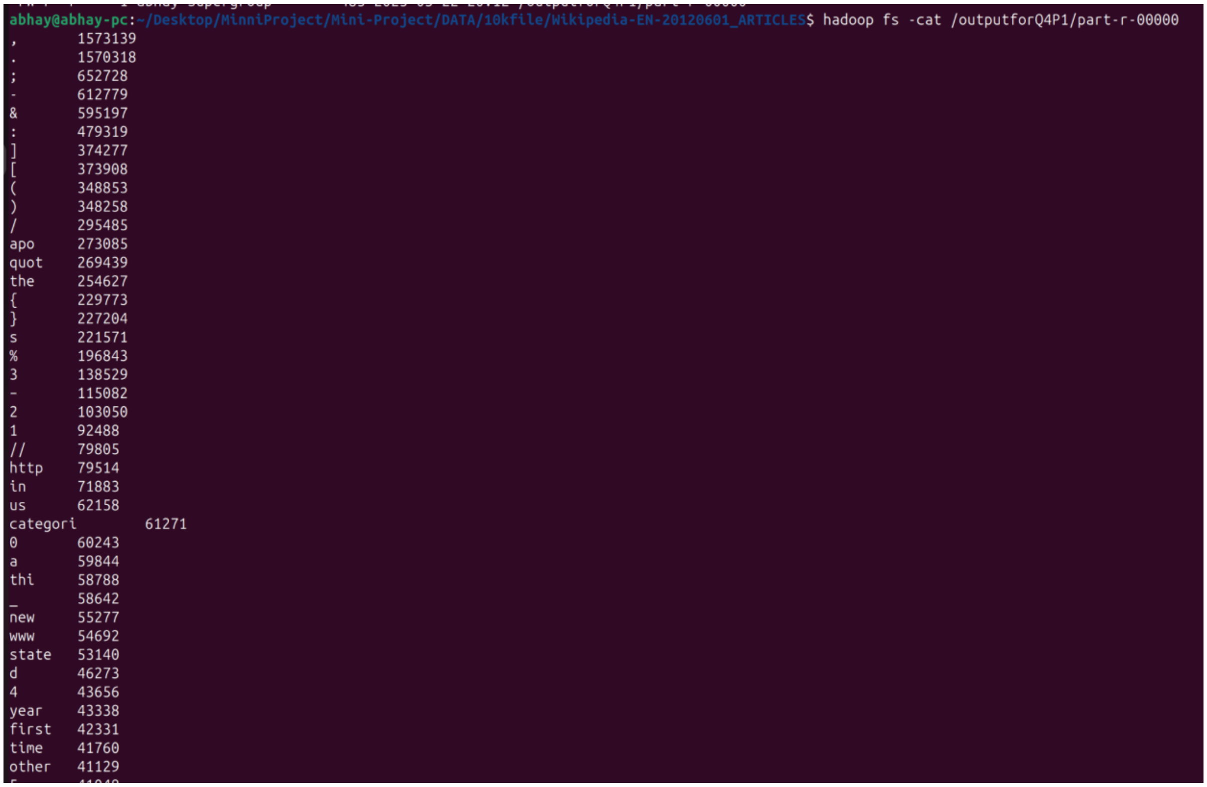


```
                Spilled Records=55517190
                Shuffled Maps =10000
                Failed Shuffles=0
                Merged Map outputs=10000
                GC time elapsed (ms)=4715
                Total committed heap usage (bytes)=40539590230016
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=166222646
        File Output Format Counters
                Bytes Written=485

real    5m57.678s
user    16m25.135s
sys     0m42.408s
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$
```

Completion Time

2. **Check the output** hadoop fs -cat /outputforQ4P1/part-r-00000



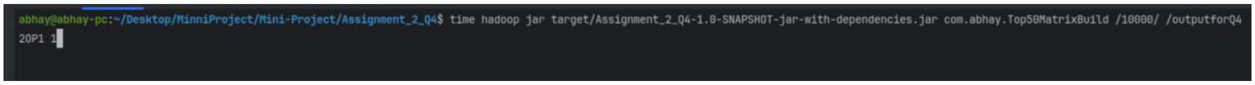
```
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/DATA/10kfile/Wikipedia-EN-20120601_ARTICLES$ hadoop fs -cat /outputforQ4P1/part-r-00000
,       1573139
.       1570318
;       652728
-       612779
&       595197
:       479319
]       374277
[       373908
(       348853
)       348258
/       295485
apo     273085
quot    269439
the     254627
{       229773
}       227204
s       221571
%       196843
3       138529
–       115082
2       103050
1       92488
//      79805
http    79514
in      71883
us      62158
categori        61271
0       60243
a       59844
thi     58788
_       58642
new     55277
www     54692
state   53140
d       46273
4       43656
year    43338
first   42331
time    41760
other   41129
```

Part B: Constructing the Co-occurring Word Matrix (Pairs Approach)

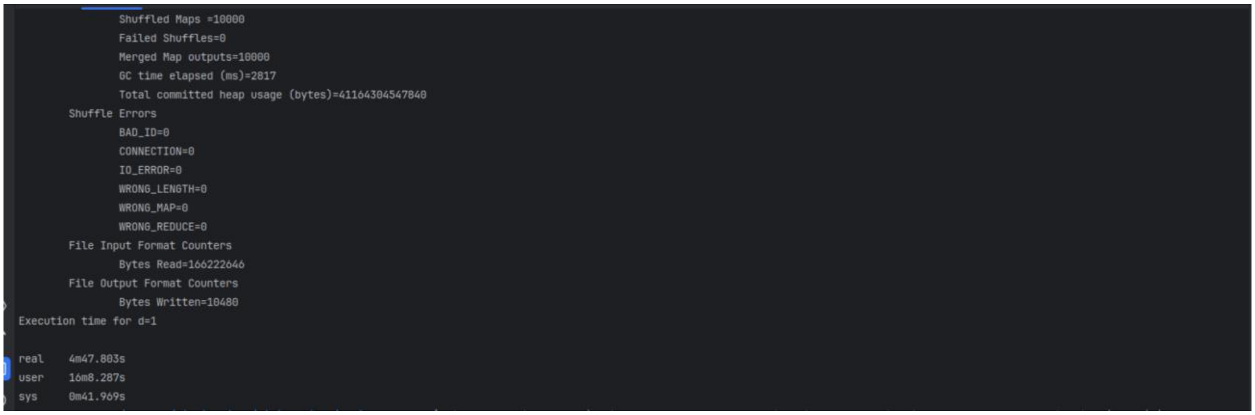
For different values of D (word distance), execute the following:

For D = 1

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50MatrixBuild /10000/ /outputforQ42OP1 1



```
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$ time hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar com.abhay.Top50MatrixBuild /10000/ /outputforQ4
2OP1 1
```

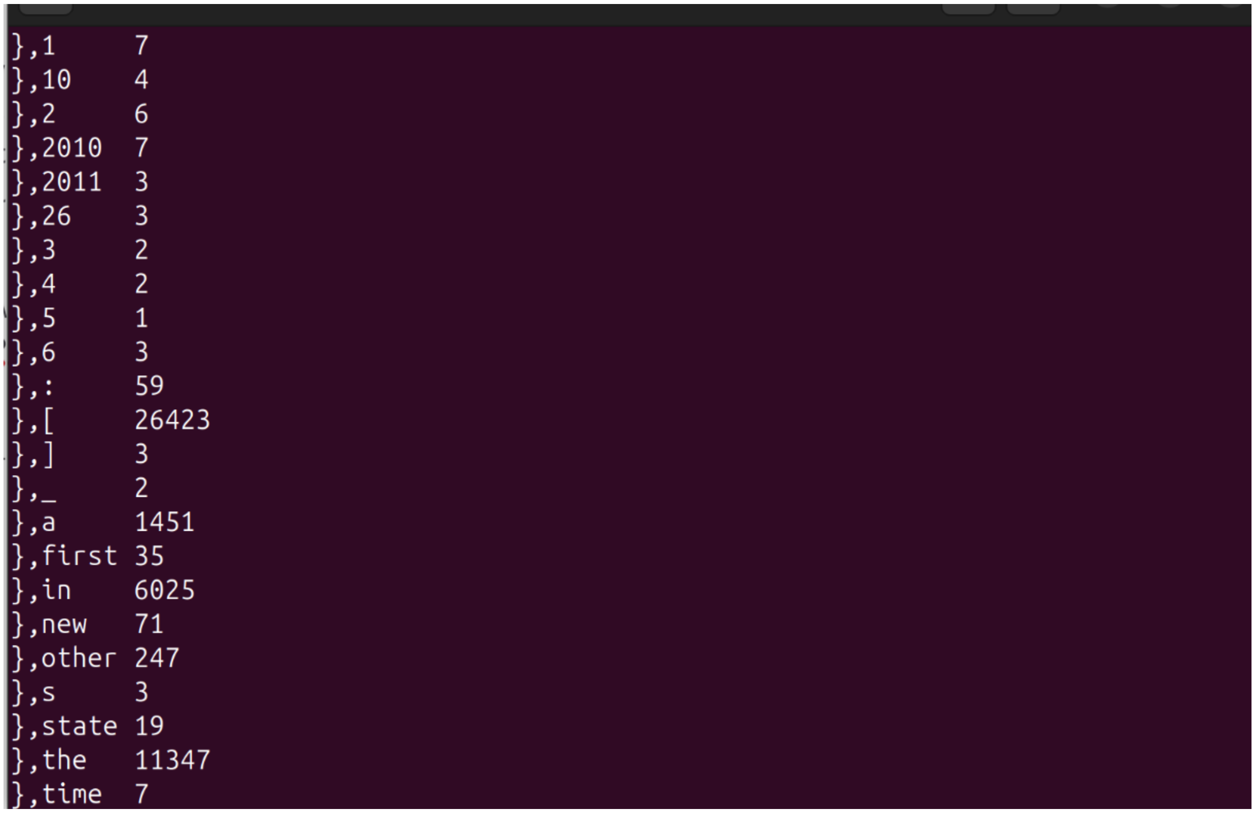


```
                Shuffled Maps =10000
                Failed Shuffles=0
                Merged Map outputs=10000
                GC time elapsed (ms)=2817
                Total committed heap usage (bytes)=41164304547840
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=166222646
        File Output Format Counters
                Bytes Written=10480
Execution time for d=1

real    4m47.803s
user    16m8.287s
sys     0m41.969s
```

Check output:

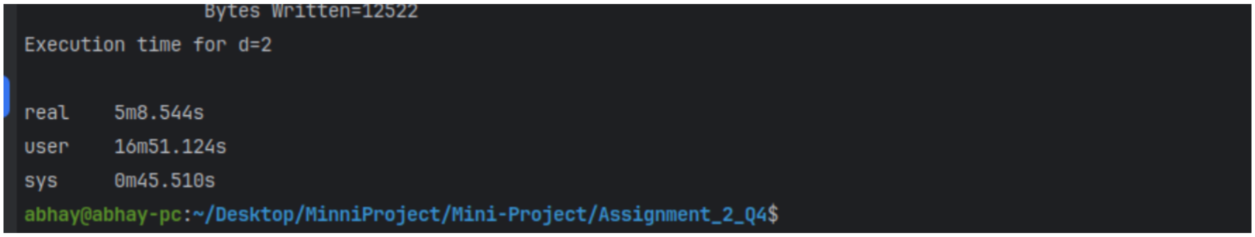
hadoop fs -cat /outputforQ42OP1/part-r-00000



```
},1     7
},10    4
},2     6
},2010  7
},2011  3
},26    3
},3     2
},4     2
},5     1
},6     3
},:     59
},[     26423
},]     3
},_     2
},a     1451
},first 35
},in    6025
},new   71
},other 247
},s     3
},state 19
},the   11347
},time  7
```

For D = 2

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50MatrixBuild /10000/ /outputforQ42OP2 2

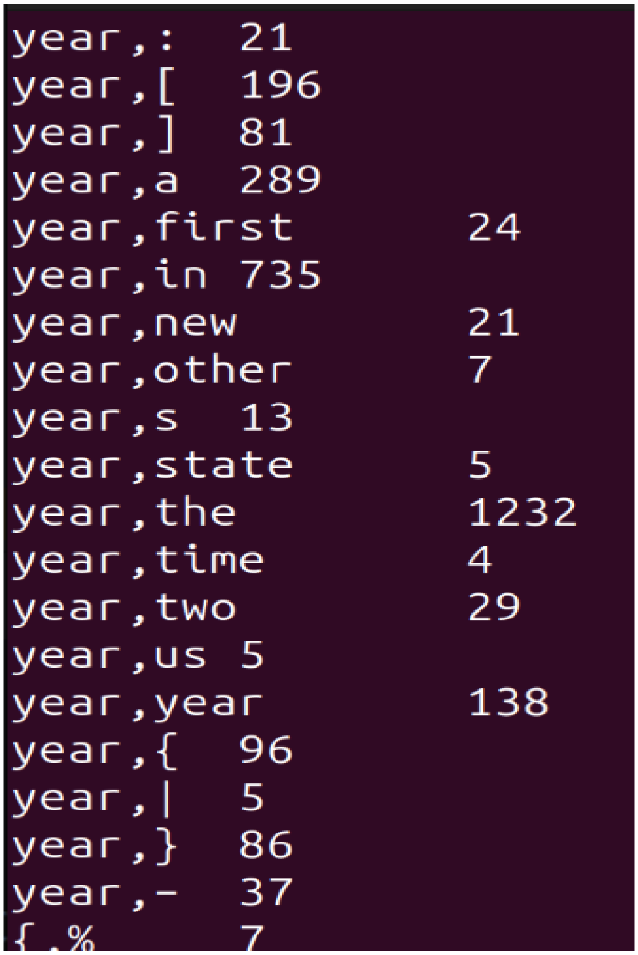


```
                Bytes Written=12522
Execution time for d=2

real    5m8.544s
user    16m51.124s
sys     0m45.510s
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$
```

Check output:

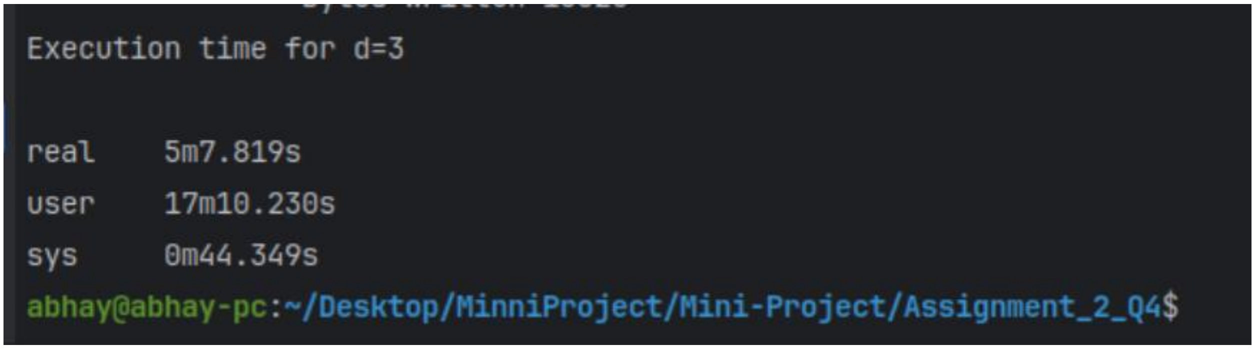
hadoop fs -cat /outputforQ42OP2/part-r-00000



```
year,:  21
year,[  196
year,]  81
year,a  289
year,first      24
year,in 735
year,new        21
year,other      7
year,s  13
year,state      5
year,the        1232
year,time       4
year,two        29
year,us 5
year,year       138
year,{  96
year,|  5
year,}  86
year,-  37
{,%     7
```

For D = 3

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50MatrixBuild /10000/ /outputforQ42OP3 3

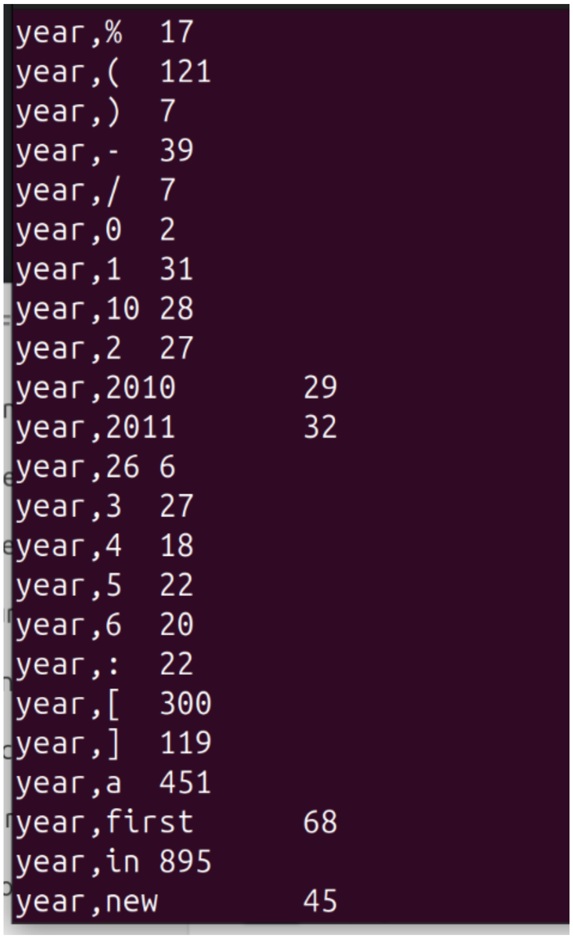


```
Execution time for d=3


real    5m7.819s
user    17m10.230s
sys     0m44.349s
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$
```

Check output:

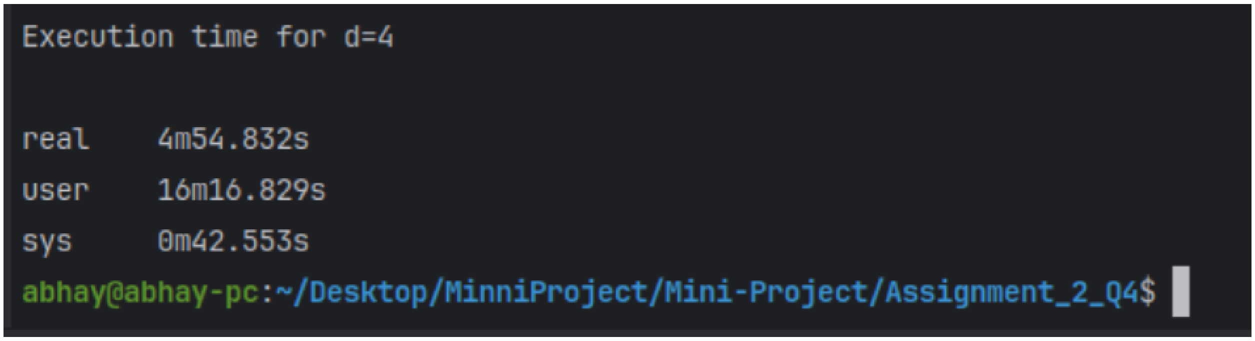
hadoop fs -cat /outputforQ42OP3/part-r-00000



```
year,%  17
year,(  121
year,)  7
year,-  39
year,/  7
year,0  2
year,1  31
year,10 28
year,2  27
year,2010       29
year,2011       32
year,26 6
year,3  27
year,4  18
year,5  22
year,6  20
year,:  22
year,[  300
year,]  119
year,a  451
year,first      68
year,in 895
year,new        45
```

For D = 4

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50MatrixBuild /10000/ /outputforQ42OP4 4

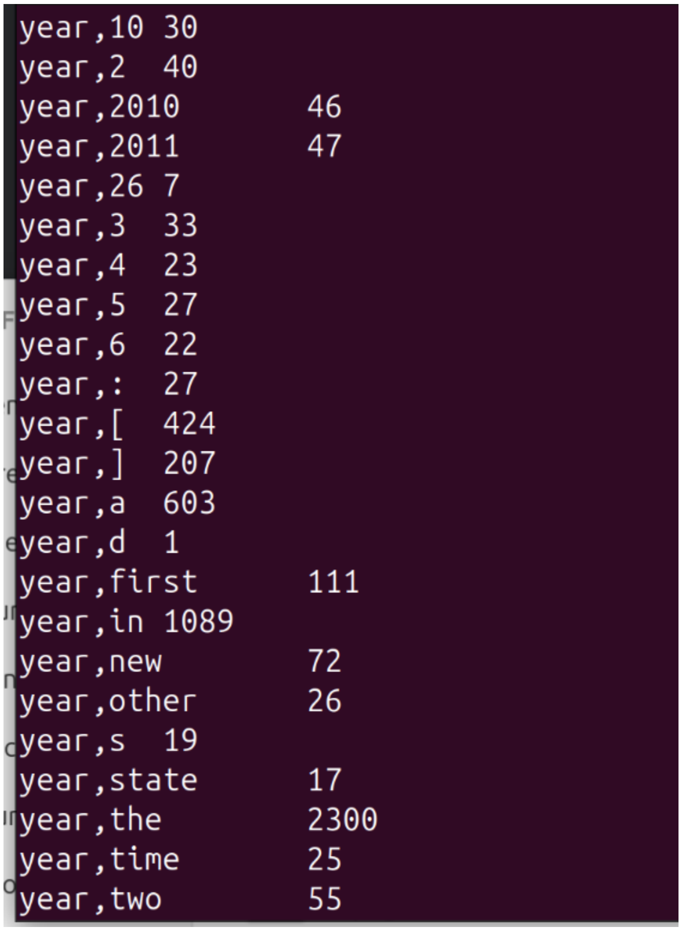


```
Execution time for d=4

real    4m54.832s
user    16m16.829s
sys     0m42.553s
abhay@abhay-pc:~/Desktop/MinniProject/Mini-Project/Assignment_2_Q4$
```

Check output:

hadoop fs -cat /outputforQ42OP4/part-r-00000



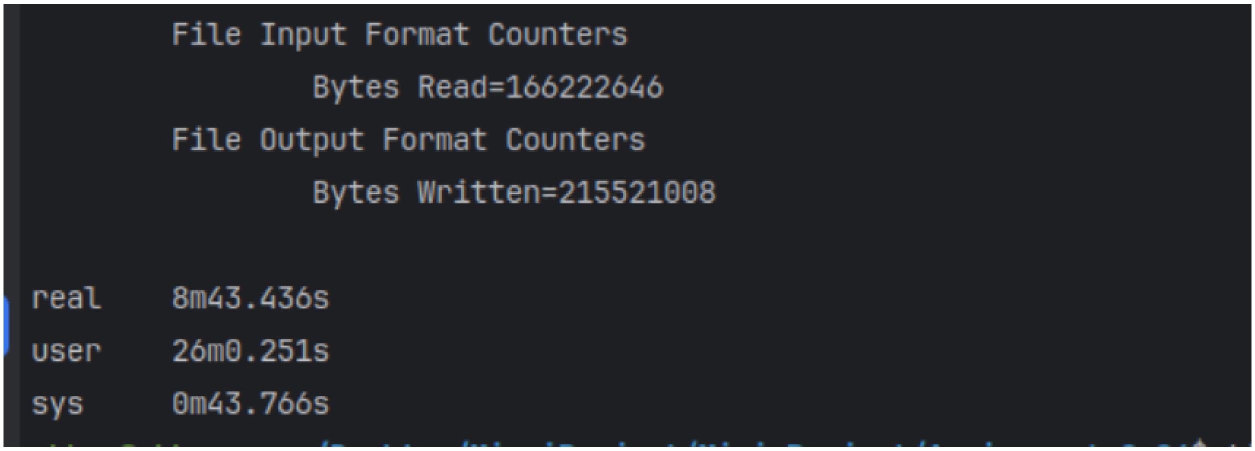
```
year,10 30
year,2  40
year,2010       46
year,2011       47
year,26 7
year,3  33
year,4  23
year,5  27
year,6  22
year,:  27
year,[  424
year,]  207
year,a  603
year,d  1
year,first      111
year,in 1089
year,new        72
year,other      26
year,s  19
year,state      17
year,the        2300
year,time       25
year,two        55
```

Part C: Constructing the Co-occurring Word Matrix (Stripes Approach)

For different values of D, execute the following:

For D = 1

```
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50Stripe /10000/ /outputforQ4P3OP01 1
```

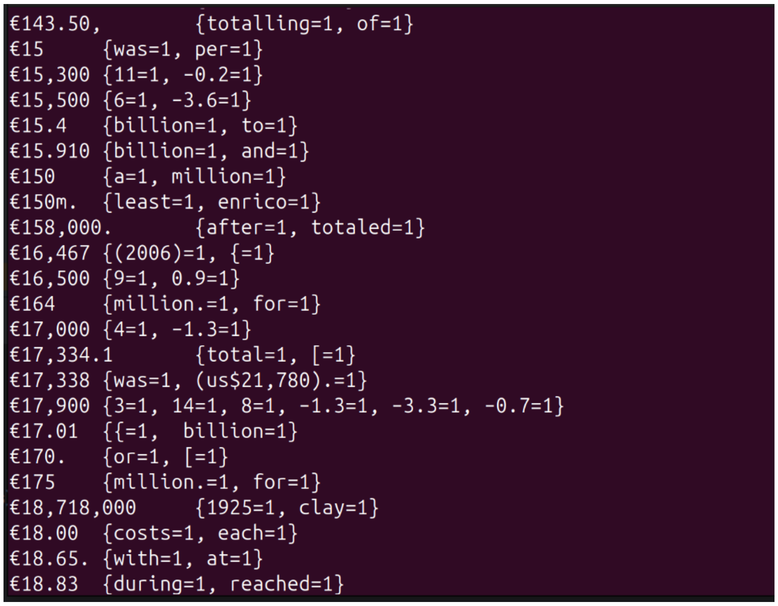


```
        File Input Format Counters
                Bytes Read=166222646
        File Output Format Counters
                Bytes Written=215521008

real    8m43.436s
user    26m0.251s
sys     0m43.766s
```

Check output:

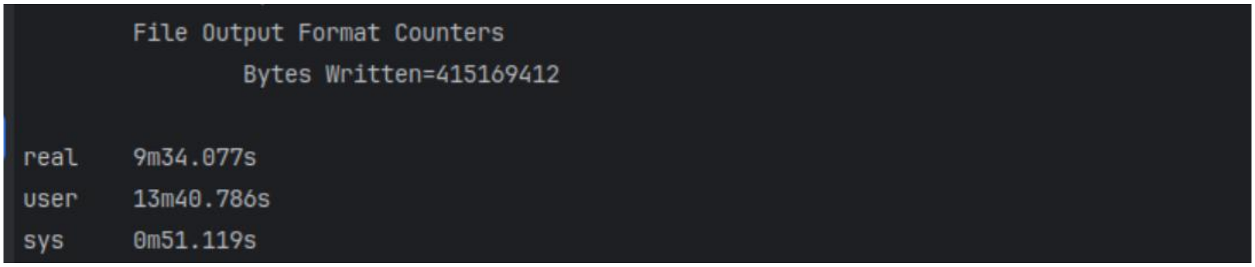
hadoop fs -cat /outputforQ4P3OP01/part-r-00000



```
€143.50,        {totalling=1, of=1}
€15     {was=1, per=1}
€15,300 {11=1, -0.2=1}
€15,500 {6=1, -3.6=1}
€15.4   {billion=1, to=1}
€15.910 {billion=1, and=1}
€150    {a=1, million=1}
€150m.  {least=1, enrico=1}
€158,000.       {after=1, totaled=1}
€16,467 {(2006)=1, {=1}
€16,500 {9=1, 0.9=1}
€164    {million.=1, for=1}
€17,000 {4=1, -1.3=1}
€17,334.1       {total=1, [=1}
€17,338 {was=1, (us$21,780).=1}
€17,900 {3=1, 14=1, 8=1, -1.3=1, -3.3=1, -0.7=1}
€17.01  {{=1,  billion=1}
€170.   {or=1, [=1}
€175    {million.=1, for=1}
€18,718,000     {1925=1, clay=1}
€18.00  {costs=1, each=1}
€18.65. {with=1, at=1}
€18.83  {during=1, reached=1}
```

For D = 2

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50Stripe /10000/ /outputforQ4P3OP02 2

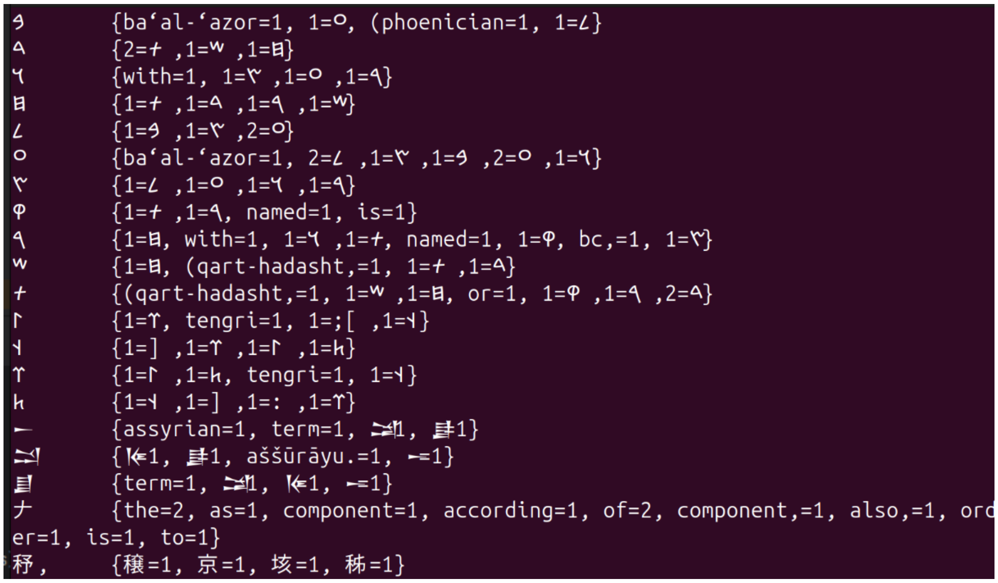


```
        File Output Format Counters
                Bytes Written=415169412

real    9m34.077s
user    13m40.786s
sys     0m51.119s
```

Check output:

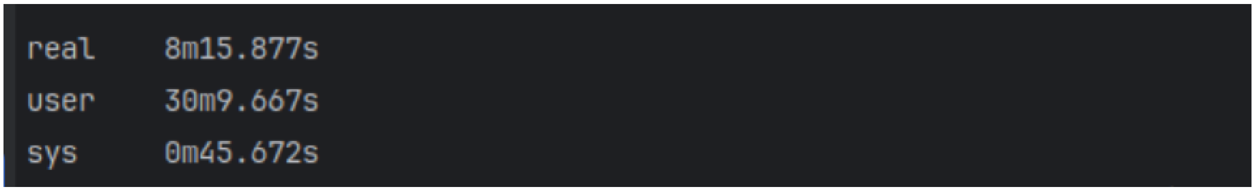
hadoop fs -cat /outputforQ4P3OP02/part-r-00000



```
𐤁       {baʻal-ʻazor=1, 1=𐤏, (phoenician=1, 1=𐤋}
𐤃       {2=𐤕 ,1=𐤔 ,1=𐤇}
𐤅       {with=1, 1=𐤆 ,1=𐤏 ,1=𐤓}
𐤇       {1=𐤕 ,1=𐤃 ,1=𐤓 ,1=𐤔}
𐤋       {1=𐤁 ,1=𐤆 ,2=𐤏}
𐤏       {baʻal-ʻazor=1, 2=𐤋 ,1=𐤆 ,1=𐤁 ,2=𐤏 ,1=𐤅}
𐤆       {1=𐤋 ,1=𐤏 ,1=𐤅 ,1=𐤓}
𐤒       {1=𐤕 ,1=𐤓, named=1, is=1}
𐤓       {1=𐤇, with=1, 1=𐤅 ,1=𐤕, named=1, 1=𐤒, bc,=1, 1=𐤆}
𐤔       {1=𐤇, (qart-hadasht,=1, 1=𐤕 ,1=𐤃}
𐤕       {(qart-hadasht,=1, 1=𐤔 ,1=𐤇, or=1, 1=𐤒 ,1=𐤓 ,2=𐤃}
𐰃       {1=𐱅, tengri=1, 1=;[ ,1=𐰭}
𐰭       {1=] ,1=𐱅 ,1=𐰃 ,1=𐰼}
𐱅       {1=𐰃 ,1=𐰼, tengri=1, 1=𐰭}
𐰼       {1=𐰭 ,1=] ,1=: ,1=𐱅}
𒀸       {assyrian=1, term=1, 𒋗=1, 𒊒=1}
𒋗       {𒀀=1, 𒊒=1, aššūrāyu.=1, 𒀸=1}
𒊒       {term=1, 𒋗=1, 𒀀=1, 𒀸=1}
ナ       {the=2, as=1, component=1, according=1, of=2, component,=1, also,=1, ord
er=1, is=1, to=1}
𥝱,     {穣=1, 京=1, 垓=1, 秭=1}
```

For D = 3

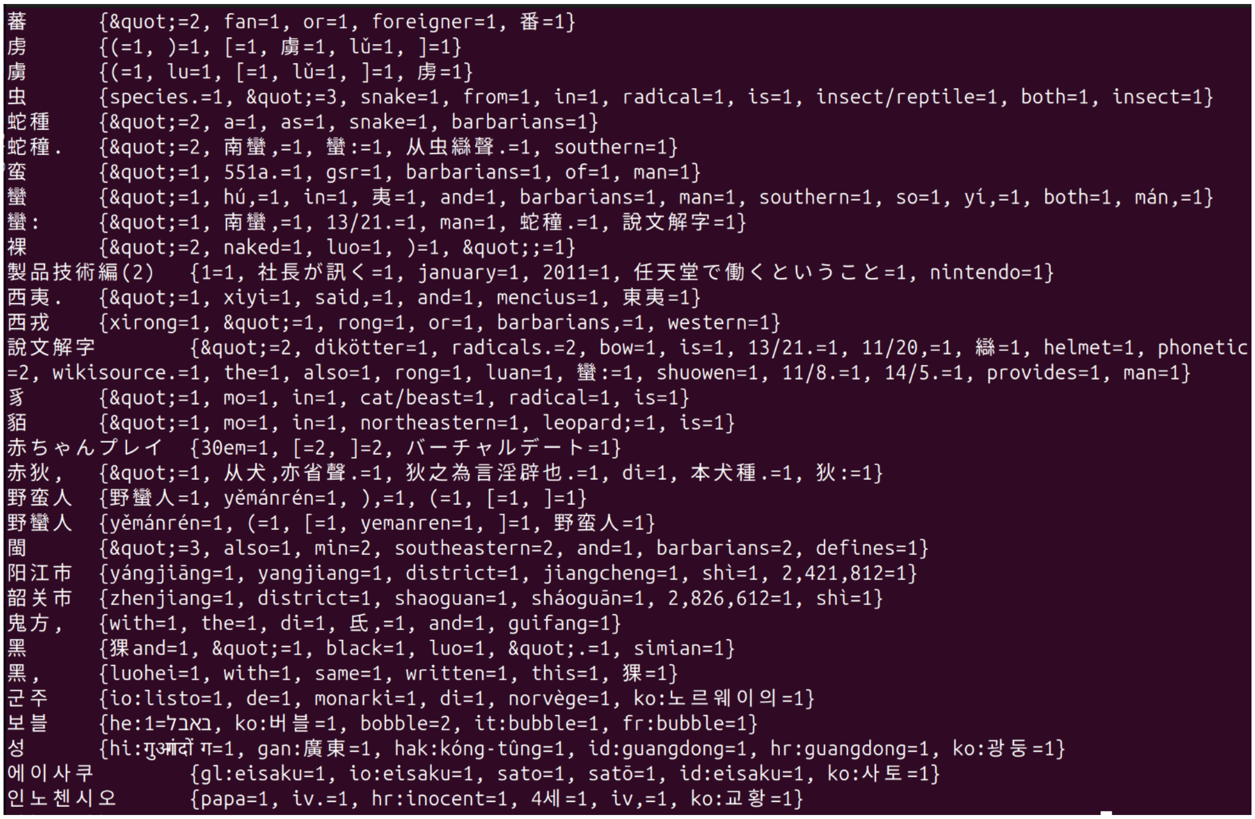
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50Stripe /10000/ /outputforQ4P3OP03 3



```
real    8m15.877s
user    30m9.667s
sys     0m45.672s
```

Check output:

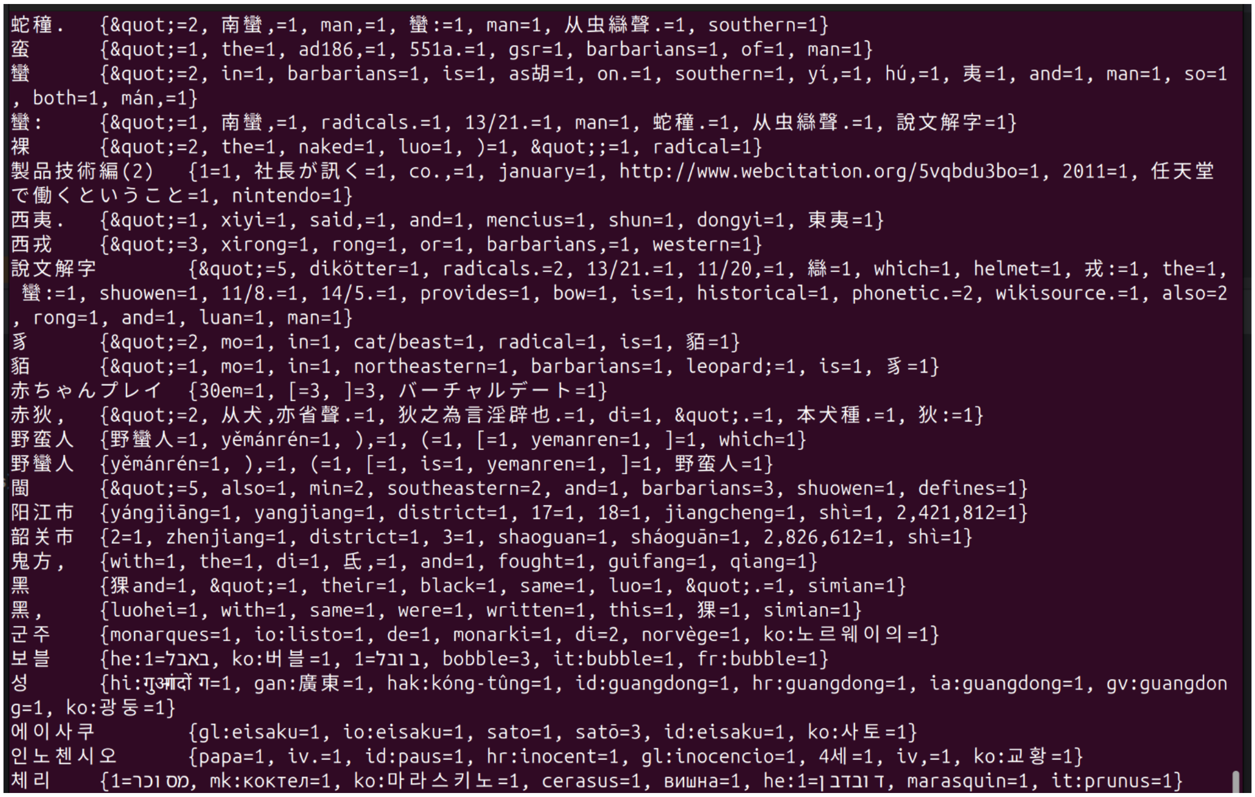
hadoop fs -cat /outputforQ4P3OP03/part-r-00000



```
蕃      {&quot;=2, fan=1, or=1, foreigner=1, 番=1}
虏      {(=1, )=1, [=1, 虜=1, lǔ=1, ]=1}
虜      {(=1, lu=1, [=1, lǔ=1, ]=1, 虏=1}
虫      {species.=1, &quot;=3, snake=1, from=1, in=1, radical=1, is=1, insect/reptile=1, both=1, insect=1}
蛇種    {&quot;=2, a=1, as=1, snake=1, barbarians=1}
蛇種.   {&quot;=2, 南蠻,=1, 蠻:=1, 从虫䜌聲.=1, southern=1}
蛮      {&quot;=1, 551a.=1, gsr=1, barbarians=1, of=1, man=1}
蠻      {&quot;=1, hú,=1, in=1, 夷=1, and=1, barbarians=1, man=1, southern=1, so=1, yí,=1, both=1, mán,=1}
蠻:     {&quot;=1, 南蠻,=1, 13/21.=1, man=1, 蛇種.=1, 說文解字=1}
裸      {&quot;=2, naked=1, luo=1, )=1, &quot;;=1}
製品技術編(2)   {1=1, 社長が訊く=1, january=1, 2011=1, 任天堂で働くということ=1, nintendo=1}
西夷.   {&quot;=1, xiyi=1, said,=1, and=1, mencius=1, 東夷=1}
西戎    {xirong=1, &quot;=1, rong=1, or=1, barbarians,=1, western=1}
說文解字        {&quot;=2, dikötter=1, radicals.=2, bow=1, is=1, 13/21.=1, 11/20,=1, 䜌=1, helmet=1, phonetic
=2, wikisource.=1, the=1, also=1, rong=1, luan=1, 蠻:=1, shuowen=1, 11/8.=1, 14/5.=1, provides=1, man=1}
豸      {&quot;=1, mo=1, in=1, cat/beast=1, radical=1, is=1}
貊      {&quot;=1, mo=1, in=1, northeastern=1, leopard;=1, is=1}
赤ちゃんプレイ  {30em=1, [=2, ]=2, バーチャルデート=1}
赤狄,   {&quot;=1, 从犬,亦省聲.=1, 狄之為言淫辟也.=1, di=1, 本犬種.=1, 狄:=1}
野蛮人  {野蠻人=1, yěmánrén=1, ),=1, (=1, [=1, ]=1}
野蠻人  {yěmánrén=1, (=1, [=1, yemanren=1, ]=1, 野蛮人=1}
閩      {&quot;=3, also=1, min=2, southeastern=2, and=1, barbarians=2, defines=1}
阳江市  {yángjiāng=1, yangjiang=1, district=1, jiangcheng=1, shì=1, 2,421,812=1}
韶关市  {zhenjiang=1, district=1, shaoguan=1, sháoguān=1, 2,826,612=1, shì=1}
鬼方,   {with=1, the=1, di=1, 氐,=1, and=1, guifang=1}
黑      {猓and=1, &quot;=1, black=1, luo=1, &quot;.=1, simian=1}
黑,     {luohei=1, with=1, same=1, written=1, this=1, 猓=1}
군주    {io:listo=1, de=1, monarki=1, di=1, norvège=1, ko:노르웨이의=1}
보블    {he:באבל=1, ko:버블=1, bobble=2, it:bubble=1, fr:bubble=1}
성      {hi:गुआंदों ग=1, gan:廣東=1, hak:kóng-tûng=1, id:guangdong=1, hr:guangdong=1, ko:광동=1}
에이사쿠        {gl:eisaku=1, io:eisaku=1, sato=1, satō=1, id:eisaku=1, ko:사토=1}
인노첸시오      {papa=1, iv.=1, hr:inocent=1, 4세=1, iv,=1, ko:교황=1}
```

For D = 4

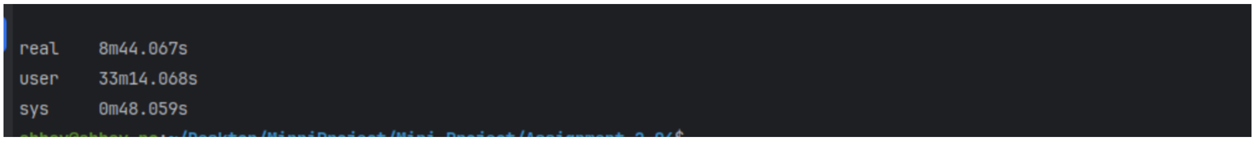
```
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50Stripe /10000/ /outputforQ4P3OP04 4
```



```
蛇穜.   {&quot;=2, 南蠻,=1, man,=1, 蠻:=1, man=1, 从虫䜌聲.=1, southern=1}
蛮      {&quot;=1, the=1, ad186,=1, 551a.=1, gsr=1, barbarians=1, of=1, man=1}
蠻      {&quot;=2, in=1, barbarians=1, is=1, as胡=1, on.=1, southern=1, yí,=1, hú,=1, 夷=1, and=1, man=1, so=1
, both=1, mán,=1}
蠻:     {&quot;=1, 南蠻,=1, radicals.=1, 13/21.=1, man=1, 蛇穜.=1, 从虫䜌聲.=1, 說文解字=1}
裸      {&quot;=2, the=1, naked=1, luo=1, )=1, &quot;;=1, radical=1}
製品技術編(2)   {1=1, 社長が訊く=1, co.,=1, january=1, http://www.webcitation.org/5vqbdu3bo=1, 2011=1, 任天堂
で働くということ=1, nintendo=1}
西夷.   {&quot;=1, xiyi=1, said,=1, and=1, mencius=1, shun=1, dongyi=1, 東夷=1}
西戎    {&quot;=3, xirong=1, rong=1, or=1, barbarians,=1, western=1}
說文解字        {&quot;=5, dikötter=1, radicals.=2, 13/21.=1, 11/20,=1, 䜌=1, which=1, helmet=1, 戎:=1, the=1,
 蠻:=1, shuowen=1, 11/8.=1, 14/5.=1, provides=1, bow=1, is=1, historical=1, phonetic.=2, wikisource.=1, also=2
, rong=1, and=1, luan=1, man=1}
豸      {&quot;=2, mo=1, in=1, cat/beast=1, radical=1, is=1, 貊=1}
貊      {&quot;=1, mo=1, in=1, northeastern=1, barbarians=1, leopard;=1, is=1, 豸=1}
赤ちゃんプレイ  {30em=1, [=3, ]=3, バーチャルデート=1}
赤狄,   {&quot;=2, 从犬,亦省聲.=1, 狄之為言淫辟也.=1, di=1, &quot;.=1, 本犬種.=1, 狄:=1}
野蛮人  {野蠻人=1, yěmánrén=1, ),=1, (=1, [=1, yemanren=1, ]=1, which=1}
野蠻人  {yěmánrén=1, ),=1, (=1, [=1, is=1, yemanren=1, ]=1, 野蛮人=1}
閩      {&quot;=5, also=1, min=2, southeastern=2, and=1, barbarians=3, shuowen=1, defines=1}
阳江市  {yángjiāng=1, yangjiang=1, district=1, 17=1, 18=1, jiangcheng=1, shì=1, 2,421,812=1}
韶关市  {2=1, zhenjiang=1, district=1, 3=1, shaoguan=1, sháoguān=1, 2,826,612=1, shì=1}
鬼方,   {with=1, the=1, di=1, 氐,=1, and=1, fought=1, guifang=1, qiang=1}
黑      {猓and=1, &quot;=1, their=1, black=1, same=1, luo=1, &quot;.=1, simian=1}
黑,     {luohei=1, with=1, same=1, were=1, written=1, this=1, 猓=1, simian=1}
군주    {monarques=1, io:listo=1, de=1, monarki=1, di=2, norvège=1, ko:노르웨이의=1}
보블    {he:1=באבל, ko:버블=1, 1=בובל, bobble=3, it:bubble=1, fr:bubble=1}
성      {hi:गुआंदों ग=1, gan:廣東=1, hak:kóng-tûng=1, id:guangdong=1, hr:guangdong=1, ia:guangdong=1, gv:guangdon
g=1, ko:광둥=1}
에이사쿠        {gl:eisaku=1, io:eisaku=1, sato=1, satō=3, id:eisaku=1, ko:사토=1}
인노첸시오      {papa=1, iv.=1, id:paus=1, hr:inocent=1, gl:inocencio=1, 4세=1, iv,=1, ko:교황=1}
체리    {1=ךכו סמ, mk:коктел=1, ko:마라스키노=1, cerasus=1, вишна=1, he:1=ןדבדוד, marasquin=1, it:prunus=1}
```

Check output:

hadoop fs -cat /outputforQ4P3OP04/part-r-00000



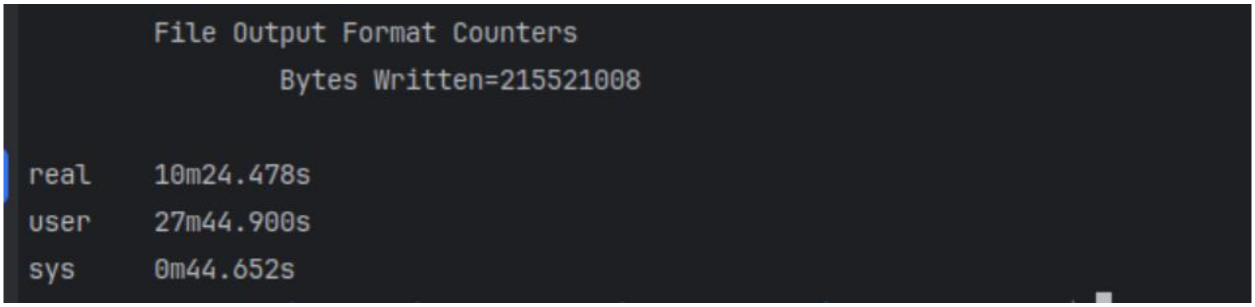
```
real    8m44.067s
user    33m14.068s
sys     0m48.059s
```

Part D: Local Aggregation (Comparison of Performance)

Run local aggregation using **both map-class level and map-function level**:

For D = 1

hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50P4 /10000/ /outputforQ4P4OP1 1

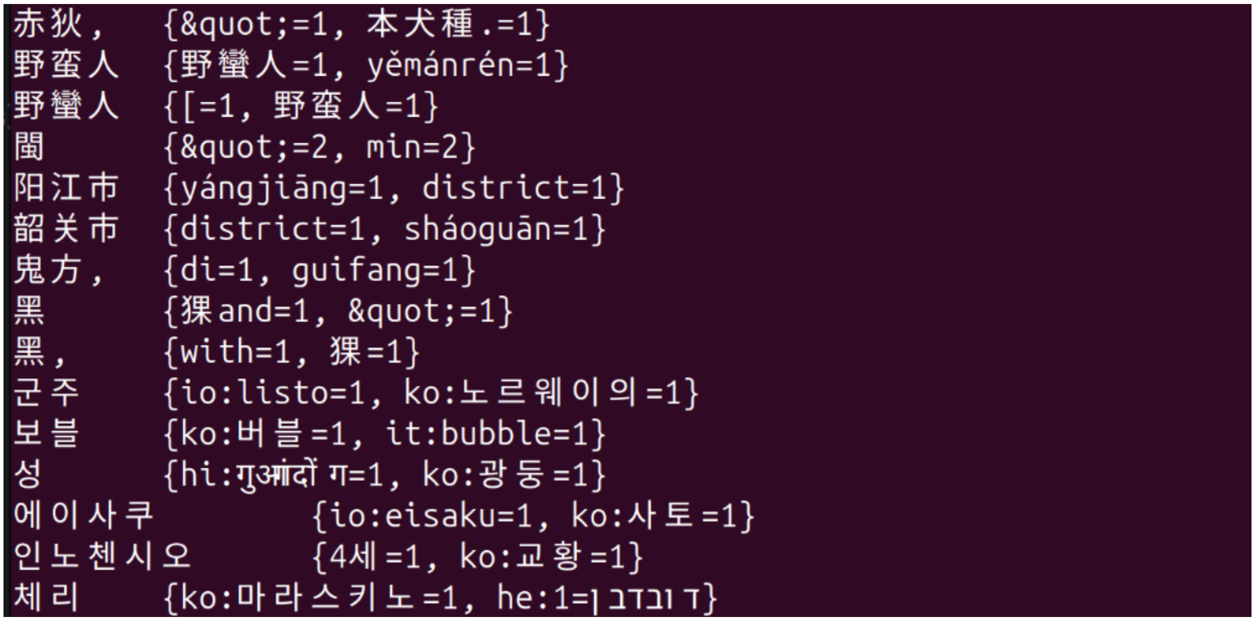


```
        File Output Format Counters
                Bytes Written=215521008

real    10m24.478s
user    27m44.900s
sys     0m44.652s
```

Check output:

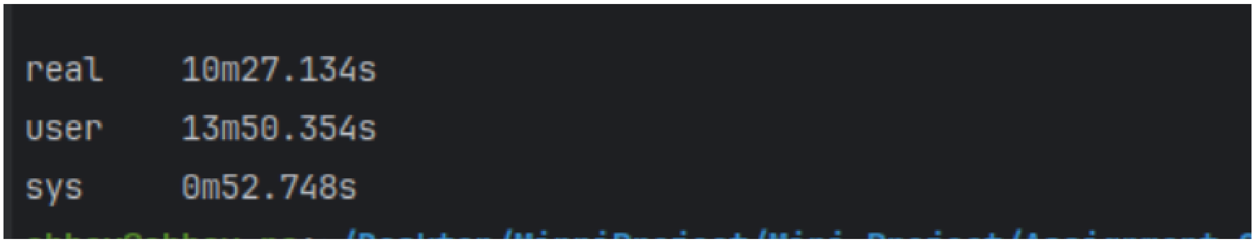
hadoop fs -cat /outputforQ4P4OP1/part-r-00000



```
赤狄，   {&quot;=1, 本犬種.=1}
野蛮人   {野蠻人=1, yěmánrén=1}
野蠻人   {[=1, 野蛮人=1}
閩       {&quot;=2, min=2}
阳江市   {yángjiāng=1, district=1}
韶关市   {district=1, sháoguān=1}
鬼方，   {di=1, guifang=1}
黑       {猓and=1, &quot;=1}
黑，     {with=1, 猓=1}
군주     {io:listo=1, ko:노르웨이의=1}
보블     {ko:버블=1, it:bubble=1}
성       {hi:गुआंदों ग=1, ko:광둥=1}
에이사쿠         {io:eisaku=1, ko:사토=1}
인노첸시오       {4세=1, ko:교황=1}
체리     {ko:마라스키노=1, he:1=ןבדבו ד}
```

For D = 2

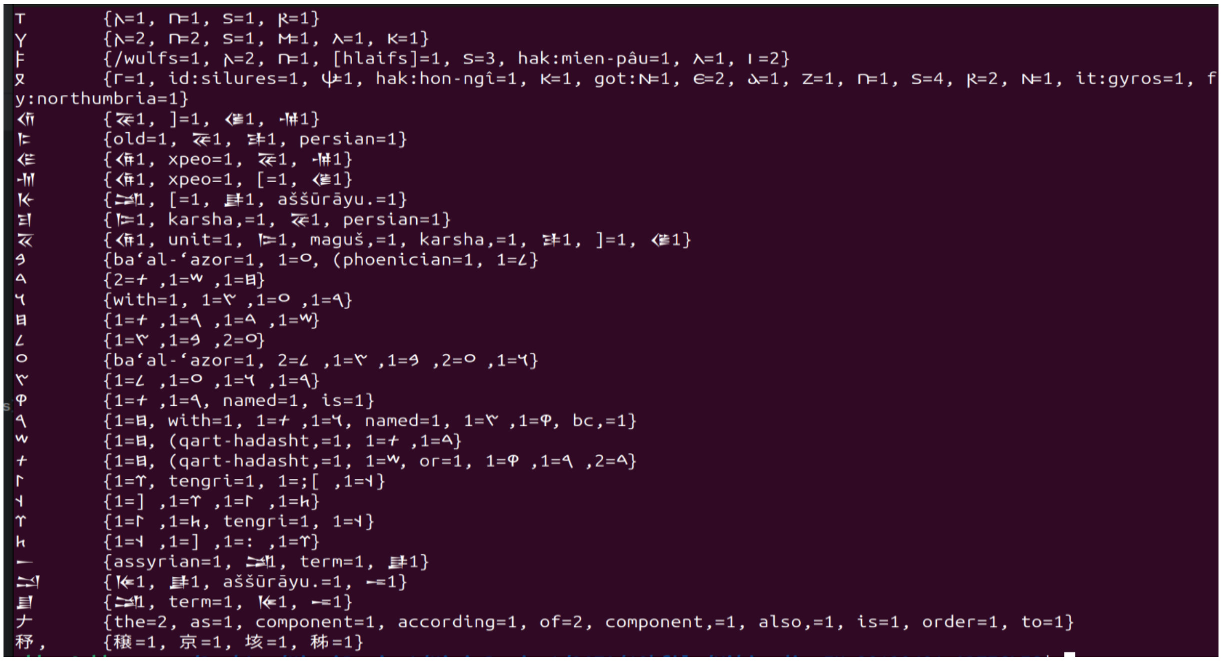
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50P4 /10000/ /outputforQ4P4OP2 2



```
real    10m27.134s
user    13m50.354s
sys     0m52.748s
```

Check output:

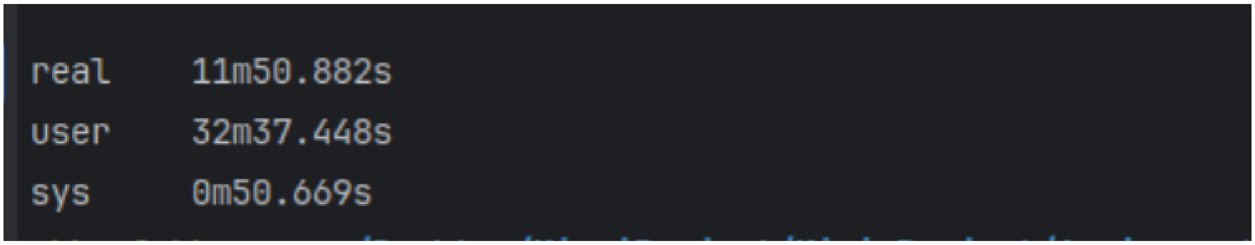
hadoop fs -cat /outputforQ4P4OP2/part-r-00000



```
𐍄       {𐌰=1, 𐌿=1, 𐍃=1, 𐍂=1}
𐍅       {𐌰=2, 𐌿=2, 𐍃=1, 𐌼=1, 𐌻=1, 𐌺=1}
𐍆       {/wulfs=1, 𐌰=2, 𐌿=1, [hlaifs]=1, 𐍃=3, hak:mien-pâu=1, 𐌻=1, 𐌹=2}
𐍉       {𐌲=1, id:silures=1, 𐌸=1, hak:hon-ngî=1, 𐌺=1, got:𐌽=1, 𐌴=2, 𐌳=1, 𐌶=1, 𐌿=1, 𐍃=4, 𐍂=2, 𐌽=1, it:gyros=1, f
y:northumbria=1}
𐎧       {𐎶=1, ]=1, 𐎨=1, 𐎴=1}
𐎠       {old=1, 𐎶=1, 𐎼=1, persian=1}
𐎨       {𐎧=1, xpeo=1, 𐎶=1, 𐎴=1}
𐎴       {𐎧=1, xpeo=1, [=1, 𐎨=1}
𐎡       {𐎰=1, [=1, 𐎢=1, aššūrāyu.=1}
𐎼       {𐎠=1, karsha,=1, 𐎶=1, persian=1}
𐎶       {𐎧=1, unit=1, 𐎠=1, maguš,=1, karsha,=1, 𐎼=1, ]=1, 𐎨=1}
𐤁       {baʿal-ʿazor=1, 1=𐤏, (phoenician=1, 1=𐤋}
𐤃       {2=𐤕 ,1=𐤔 ,1=𐤇}
𐤆       {with=1, 1=𐤊 ,1=𐤏 ,1=𐤓}
𐤇       {1=𐤕 ,1=𐤓 ,1=𐤃 ,1=𐤔}
𐤋       {1=𐤊 ,1=𐤁 ,2=𐤏}
𐤏       {baʿal-ʿazor=1, 2=𐤋 ,1=𐤊 ,1=𐤁 ,2=𐤏 ,1=𐤆}
𐤊       {1=𐤋 ,1=𐤏 ,1=𐤆 ,1=𐤓}
𐤒       {1=𐤕 ,1=𐤓, named=1, is=1}
𐤓       {1=𐤇, with=1, 1=𐤕 ,1=𐤆, named=1, 1=𐤊 ,1=𐤒, bc,=1}
𐤔       {1=𐤇, (qart-hadasht,=1, 1=𐤕 ,1=𐤃}
𐤕       {1=𐤇, (qart-hadasht,=1, 1=𐤔, or=1, 1=𐤒 ,1=𐤓 ,2=𐤃}
𐰭       {1=𐱅, tengri=1, 1=;[ ,1=𐰃}
𐰃       {1=] ,1=𐱅 ,1=𐰭 ,1=𐰼}
𐱅       {1=𐰭 ,1=𐰼, tengri=1, 1=𐰃}
𐰼       {1=𐰃 ,1=] ,1=: ,1=𐱅}
𐏐       {assyrian=1, 𐎰=1, term=1, 𐎢=1}
𐎰       {𐎡=1, 𐎢=1, aššūrāyu.=1, 𐏐=1}
𐎢       {𐎰=1, term=1, 𐎡=1, 𐏐=1}
ナ       {the=2, as=1, component=1, according=1, of=2, component,=1, also,=1, is=1, order=1, to=1}
秄，     {穣=1, 京=1, 垓=1, 秭=1}
```

For D = 3

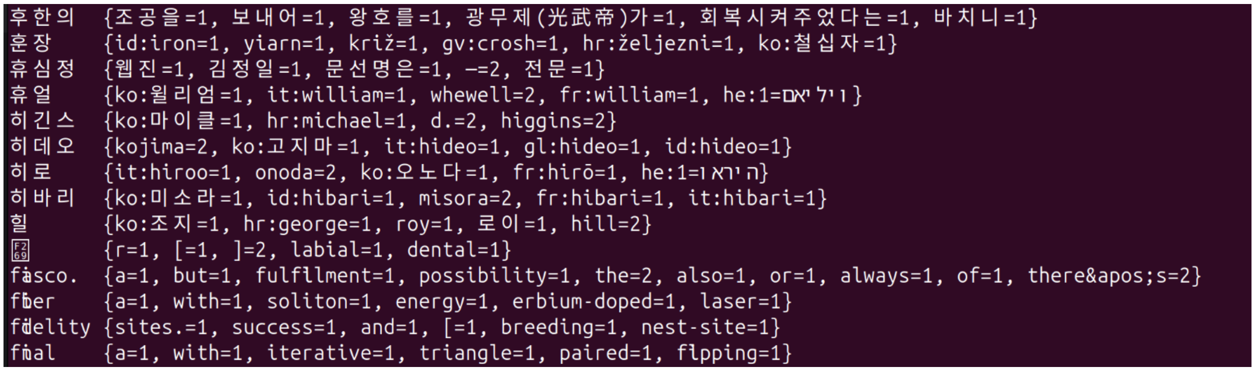
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50P4 /10000/ /outputforQ4P4OP3 3



```
real    11m50.882s
user    32m37.448s
sys     0m50.669s
```

Check output:

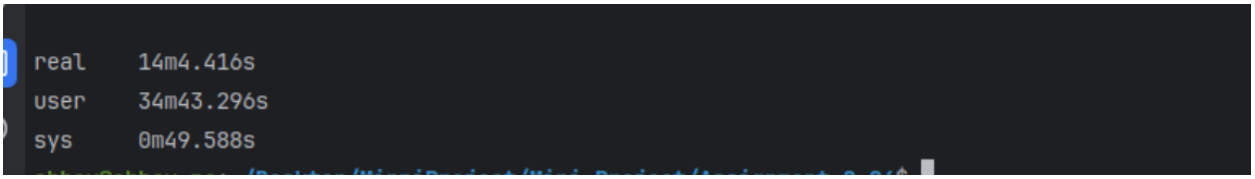
hadoop fs -cat /outputforQ4P4OP3/part-r-00000



```
후한의   {조공을=1, 보내어=1, 왕호를=1, 광무제(光武帝)가=1, 회복시켜주었다는=1, 바치니=1}
훈장     {id:iron=1, yiarn=1, križ=1, gv:crosh=1, hr:željezni=1, ko:철십자=1}
휴심정   {웹진=1, 김정일=1, 문선명은=1, –=2, 전문=1}
휴얼     {ko:윌리엄=1, it:william=1, whewell=2, fr:william=1, he:1=ויליאם}
히긴스   {ko:마이클=1, hr:michael=1, d.=2, higgins=2}
히데오   {kojima=2, ko:고지마=1, it:hideo=1, gl:hideo=1, id:hideo=1}
히로     {it:hiroo=1, onoda=2, ko:오노다=1, fr:hirō=1, he:1=היראו}
히바리   {ko:미소라=1, id:hibari=1, misora=2, fr:hibari=1, it:hibari=1}
힐       {ko:조지=1, hr:george=1, roy=1, 로이=1, hill=2}
       {r=1, [=1, ]=2, labial=1, dental=1}
fiasco.  {a=1, but=1, fulfllment=1, possibility=1, the=2, also=1, or=1, always=1, of=1, there&apos;s=2}
fiber    {a=1, with=1, soliton=1, energy=1, erbium-doped=1, laser=1}
fidelity {sites.=1, success=1, and=1, [=1, breeding=1, nest-site=1}
final    {a=1, with=1, iterative=1, triangle=1, paired=1, flpping=1}
```

For D = 4

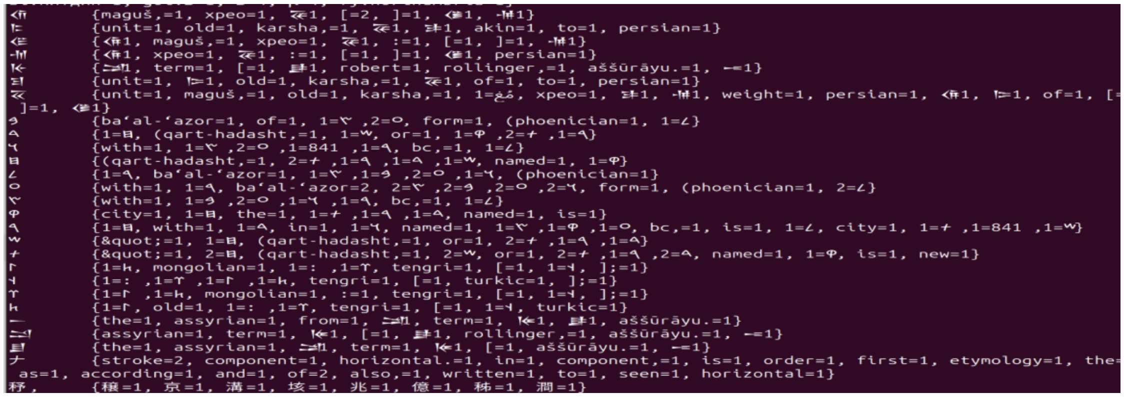
hadoop jar target/Assignment_2_Q4-1.0-SNAPSHOT-jar-with-dependencies.jar \
com.abhay.Top50P4 /10000/ /outputforQ4P4OP4 4



```
real    14m4.416s
user    34m43.296s
sys     0m49.588s
```

Check output:

hadoop fs -cat /outputforQ4P4OP4/part-r-00000



```
𐎶       {maguš,=1, xpeo=1, 𐏁=1, [=2, ]=1, 𐎥=1, 𐎢=1}
𐎣       {unit=1, old=1, karsha,=1, 𐏁=1, 𐎼=1, akin=1, to=1, persian=1}
𐎥       {𐎶=1, maguš,=1, xpeo=1, 𐏁=1, :=1, [=1, ]=1, 𐎢=1}
𐎢       {𐎶=1, xpeo=1, 𐏁=1, :=1, [=1, ]=1, 𐎥=1, persian=1}
𒊒       {𒋗=1, term=1, [=1, 𒊏=1, robert=1, rollinger,=1, aššūrāyu.=1, 𒀸=1}
𐎼       {unit=1, 𐎣=1, old=1, karsha,=1, 𐏁=1, of=1, to=1, persian=1}
𐏁       {unit=1, maguš,=1, old=1, karsha,=1, 1=غُ, xpeo=1, 𐎼=1, 𐎢=1, weight=1, persian=1, 𐎶=1, 𐎣=1, of=1, [=
 ]=1, 𐎥=1}
𐤁       {ba‘al-‘azor=1, of=1, 1=𐤆 ,2=𐤏, form=1, (phoenician=1, 1=𐤋}
𐤃       {1=𐤇, (qart-hadasht,=1, 1=𐤔, or=1, 1=𐤒 ,2=𐤕 ,1=𐤓}
𐤅       {with=1, 1=𐤆 ,2=𐤏 ,1=841 ,1=𐤓, bc,=1, 1=𐤋}
𐤇       {(qart-hadasht,=1, 2=𐤕 ,1=𐤓 ,1=𐤃 ,1=𐤔, named=1, 1=𐤒}
𐤋       {1=𐤓, ba‘al-‘azor=1, 1=𐤆 ,1=𐤁 ,2=𐤏 ,1=𐤅, (phoenician=1}
𐤏       {with=1, 1=𐤓, ba‘al-‘azor=2, 2=𐤆 ,2=𐤁 ,2=𐤏 ,2=𐤅, form=1, (phoenician=1, 2=𐤋}
𐤆       {with=1, 1=𐤁 ,2=𐤏 ,1=𐤅 ,1=𐤓, bc,=1, 1=𐤋}
𐤒       {city=1, 1=𐤇, the=1, 1=𐤕 ,1=𐤓 ,1=𐤃, named=1, is=1}
𐤓       {1=𐤇, with=1, 1=𐤃, in=1, 1=𐤅, named=1, 1=𐤆 ,1=𐤒 ,1=𐤏, bc,=1, is=1, 1=𐤋, city=1, 1=𐤕 ,1=841 ,1=𐤔}
𐤔       {&quot;=1, 1=𐤇, (qart-hadasht,=1, or=1, 2=𐤕 ,1=𐤓 ,1=𐤃}
𐤕       {&quot;=1, 2=𐤇, (qart-hadasht,=1, 2=𐤔, or=1, 2=𐤕 ,1=𐤓 ,2=𐤃, named=1, 1=𐤒, is=1, new=1}
𐰭       {1=𐰼, mongolian=1, 1=: ,1=𐱅, tengri=1, [=1, 1=𐰃, ];=1}
𐰃       {1=: ,1=𐱅 ,1=𐰭 ,1=𐰼, tengri=1, [=1, turkic=1, ];=1}
𐱅       {1=𐰭 ,1=𐰼, mongolian=1, :=1, tengri=1, [=1, 1=𐰃, ];=1}
𐰼       {1=𐰭, old=1, 1=: ,1=𐱅, tengri=1, [=1, 1=𐰃, turkic=1}
𒀸       {the=1, assyrian=1, from=1, 𒋗=1, term=1, 𒊒=1, 𒊏=1, aššūrāyu.=1}
𒋗       {assyrian=1, term=1, 𒊒=1, [=1, 𒊏=1, rollinger,=1, aššūrāyu.=1, 𒀸=1}
𒊏       {the=1, assyrian=1, 𒋗=1, term=1, 𒊒=1, [=1, aššūrāyu.=1, 𒀸=1}
ナ       {stroke=2, component=1, horizontal.=1, in=1, component,=1, is=1, order=1, first=1, etymology=1, the=
 as=1, according=1, and=1, of=2, also,=1, written=1, to=1, seen=1, horizontal=1}
秭 ,     {穣=1, 京=1, 溝=1, 垓=1, 兆=1, 億=1, 秭=1, 澗=1}
```