CSE 4/587 Data Intensive Computing
Homework #1 - MapReduce

1. The 25 most common words and the number of occurrences of each without removing the stop words are:

the 28419
and 15083
of 13583
to 12083
a 10078
in 7012
was 5082
that 4124
he 3997
I 3679
is 3553
for 3537
with 3377
his 3124
her 3110
had 3079
it 3018
as 2996
she 2793
be 2776
on 2561
you 2551
not 2464
The 2449
at 2425

```
(base) vinithavudhayagiri@Vinithas-Air Downloads % sort -nrk2 part-r-00000 |head -n 25
the     28419
and     15083
of      13583
to      12083
a       10078
in      7012
was     5082
that    4124
he      3997
I       3679
is      3553
for     3537
with    3377
his     3124
her     3110
had     3079
it      3018
as      2996
she     2793
be      2776
on      2561
you     2551
not     2464
The     2449
at      2425
```

2. The 25 most common words and the number of occurrences of each after removing the stop words are:

she 3820
are 1586
one 1483
de 1105
project 894
little 891
gutenberg 870
time 769
down 736
bunny 723
amelia 717
work 701
back 678
upon 651
man 651
may 650
know 643
good 582
old 578
two 559
er 558
go 535
come 535
great 532
long 512

```
(base) vinithavudhayagiri@Vinithas-Air Downloads % sort -nrk2 part-r|head -n 25
she     3820
are     1586
one     1483
de      1105
project 894
little  891
gutenberg       870
time    769
down    736
bunny   723
amelia  717
work    701
back    678
upon    651
man     651
may     650
know    643
good    582
old     578
two     559
er      558
go      535
come    535
great   532
long    512
```

3.

The total amount of bytes output by mappers before removing the stop words:
Map output bytes=4819675

```
        Total megabyte-milliseconds taken by all reduce tasks=10940416
Map-Reduce Framework
        Map input records=60870
        Map output records=494276
        Map output bytes=4819675
        Map output materialized bytes=1425602
        Input split bytes=1151
        Combine input records=494276
        Combine output records=98820
        Reduce input groups=54437
        Reduce shuffle bytes=1425602
        Reduce input records=98820
        Reduce output records=54437
        Spilled Records=197640
        Shuffled Maps =10
        Failed Shuffles=0
        Merged Map outputs=10
        GC time elapsed (ms)=1353
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=3164602368
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
```

The total amount of bytes output by mappers before removing the stop words:
Map output bytes=2787232

```
        Total megabyte-milliseconds taken by all reduce tasks=9382432
Map-Reduce Framework
        Map input records=60870
        Map output records=256928
        Map output bytes=2787232
        Map output materialized bytes=853757
        Input split bytes=1151
        Combine input records=256928
        Combine output records=60467
        Reduce input groups=27708
        Reduce shuffle bytes=853757
        Reduce input records=60467
        Reduce output records=27708
        Spilled Records=120934
        Shuffled Maps =10
        Failed Shuffles=0
        Merged Map outputs=10
        GC time elapsed (ms)=1523
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=3173515264
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=2948859
File Output Format Counters
        Bytes Written=301517
```

The one concrete way that would affect the performance of the application is to adjust the number of mappers and reducers used.

4.

The size of key space with stop words:
with stop words
54437
The size of key space without stop words:
without stop words
27708

The size of the key space gets reduced as the number of unique words are decreased, this is happened because the stop words from the input dataset are removed.

```
(base) vinithavudhayagiri@Vinithas-Air Downloads % cat part-r| cut -f1 | sort | uniq | wc -l
   27708
(base) vinithavudhayagiri@Vinithas-Air Downloads % cat part-r-00000| cut -f1 | sort | uniq | wc -l
   54437
```

5.

a. Each mapper will parse = 100 TB / 10 sites / 20 mappers per site = 0.5 TB = 500 GB= 5 * (10 ^11) Bytes.

b. As per question, the size of our key space after ignoring all is 25.

c. The maximum number of key-value pairs that could be communicated during the barrier between mapping and reducing is

25 keys * 200 mappers * S sites = 5000 * S key value pairs

d. The key-value pairs for each reducer are

Total key values/ Number of reducers = 5000/10 = 500

6.

Data Flow Diagram