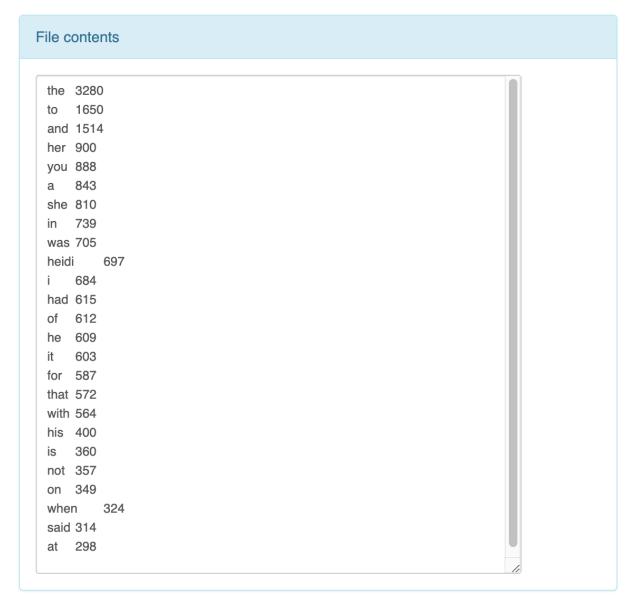
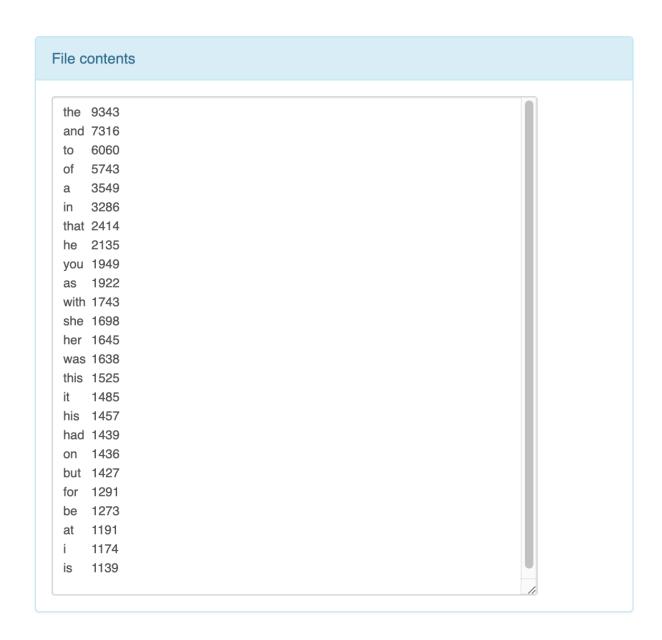
1. What are the 25 most common words and the number of occurrences of each when you do not remove stopwords? **[10pt]**

Heidi.txt:

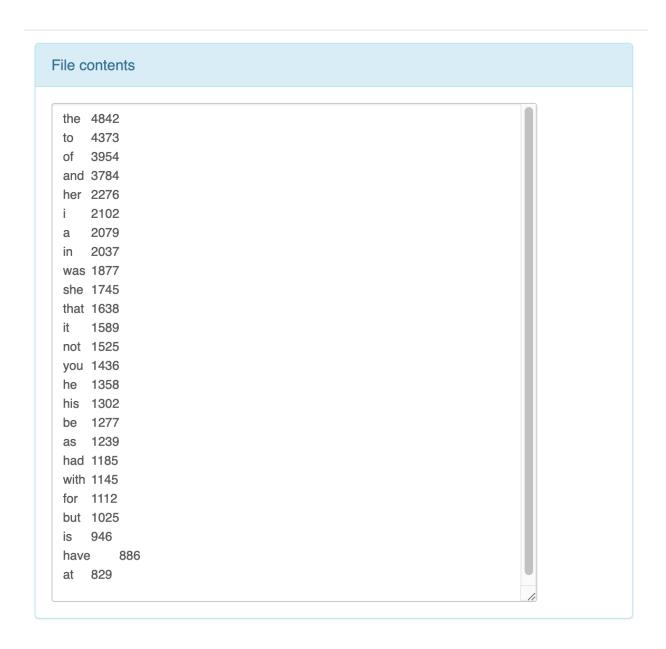


Close

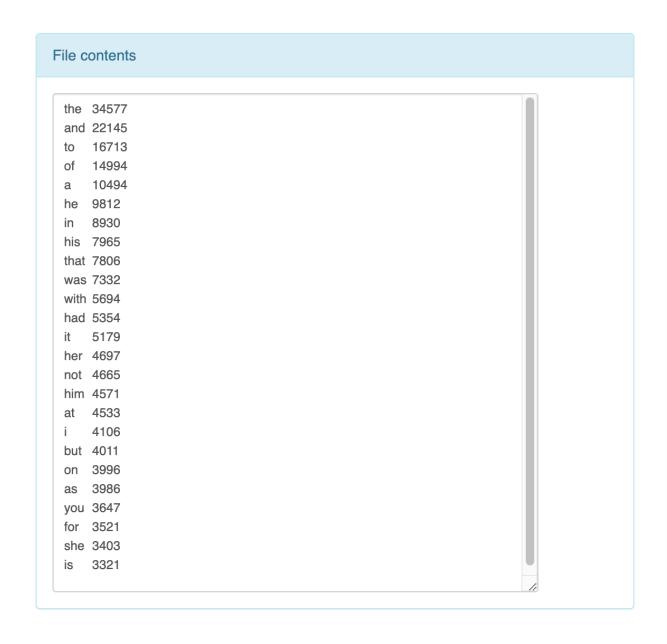
dream_of_red_chamber.txt



pride_and_prejudice.txt



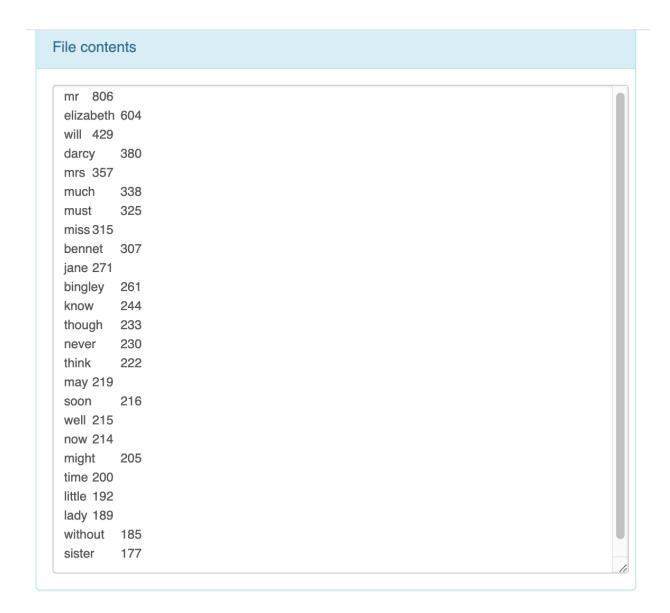
war_and_peace.txt



2. What are the 25 most common words and the number of occurrences of each when you do remove stopwords? [10pt]

Heidi.txt

File contents heidi 697 peter 243 child 229 now 221 210 come clara 208 grandfather 192 old 187 little 180 go 164 oh 146 must 143 grandmother 142 see 139 time 122 sesemann 115 will 113 day 110 asked 103 get 101 mr 100 uncle 95 away 94 never 91 tell 90



```
File contents
gatsby 194
tom 175
daisy
       147
like 120
back
      109
came
        108
man 106
little 103
just 100
know
now 97
dont 96
went
        91
project 90
gutenberg
            88
eyes
      86
got 85
see 84
old 83
looked 82
time 81
away 76
way 75
get 74
new 73
```

 $dream_of_red_chamber.txt$

File contents chia 1061 800 paoy lady 727 will 584 come 548 feng 448 time 438 upon 437 go 426 two 400 day 346 came 318 also 313 old 305 now 298 well 295 like 292 chin 286 words 268 whole 265 way 262 back 249 taiy 246 mrs 244 family 237

File contents prince 1886 pierre 1784 now 1303 natsha 1092 will 1064 andrew 1039 time 921 princess 915 face 891 french 872 went 859 know 841 eyes 820 old 803 room 766 thought 764 men 760 go 752 well 735 like 733 chapter 732 see 730 rostv 715 began 714 moscow 707

3. Based on the output of your application, how does removing stop words affect the total amount of bytes output by your mappers? Name one concrete way that this would affect the performance of your application. **[10pt]**

After removing stopwords, the total amount of bytes output significantly decrease. And the time decrease, so increase the performance.

4. Based on the output of your application, what is the size of your keyspace with and without removing stopwords? How does this correspond to the number of stopwords you have chosen to remove? [10pt]

The size of keyspace is 25. It related to it because I remove stopwords.

5. Let's now assume you were going to run your application on the entirety of Project Gutenberg. For this question, assume that there are **100TB** of input data, the data is spread over **10 sites**, and each site has **20 mappers**. Assume you ignore all but the **25 most common words** that you listed in question 2. Furthermore, assume that your combiners have been run optimally so that each combiner will output at most 1 key-value pair per key. **[4x5=20pt]**

a. How much data will each mapper have to parse?

100TB/10 sites*20 mapper=500 GB

b. What is the size of your keyspace?

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

key-value pair=sites number* mapper number per site *the size of keyspace =10*20 *25=5000

d. Assume you are running one reducer per site. On average, how many key-value pairs will each reducer have to handle?

key-value pairs per reducer=key-value pair / reducer number =5000/10=500

6. Draw the data flow diagram for question 5. The diagram should be similar to the diagram shown in the lecture. On your diagram, label the specific quantities you got for 5a,b,c, and d. **[10pt]**

