1.

Buffer size	**File size**
10	2221494
100	1909980
1000	1790771
1000	1785591

2. I will go through the test file and store all the needed vocabulary into a hashset and the size of the hashset will be dom (x).

3.

Firstly, go through each row in the input file and split and document id and content by the tab, and then combine the document id with each content.

The input format should be:

D[1] \t content[1] content[2] content[3] …. Content[n]

…

D[k] \t content[1] content[2] content[3] ….

The output should be:

Content[1], D[1]

…

Content[n], D[k]

Secondly, sort the output by the first word and output the MergeCount file.

The input and output format should be:

Content[1], D[1]

…

Content[n], D[k]

Thirdly, merge the same lines in the MergeCount file, which means delete the same lines in the MergeCount file.

The input format should be:

Content[1], D[1]

…

Content[n], D[k]

The output format should be:
(content[n], d[k])