Implementation

For the first part, I wrote a method, *get_terms_into_dict()* to generate a dictionary. This dictionary has key as a term id and has a value list to store document ids later. It returns this dictionary. Then, I wrote a method, *make_idv_dictionary()* to append document ids into the dictionary, that is previously created, based on term ids. In this method, it traverses whole data, “crawl.DV.0” and “crawl.DV.1”. After that, I wrote *write_dict_to_textfile()*. This method writes the dictionary on “crawl.IDV” file.

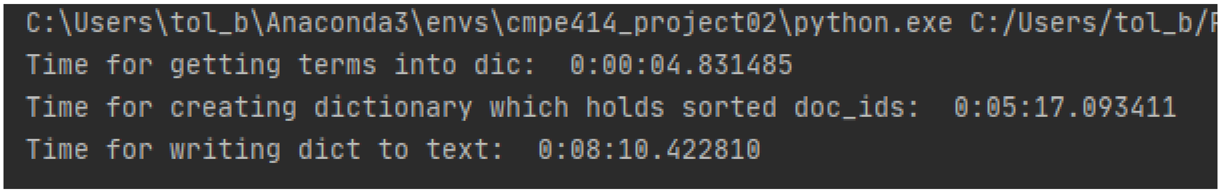
For the second part, I wrote *golomb_version()* method. In there, I calculated the b value with this formula, $b^A = 0.69 * \frac{N*n}{f}$. To do that, I get N, is the number of documents, n, is the number of distinct terms, and f, is the number of distinct (document, term) pairs. Then, I created “crawl.IDV.encoded.bin” file by using the Golomb encoding. To use the Golomb encoding, I write second python file, “golomb.py”. In this file, there is *golomb_encoding()* method and it calculates unary and truncated binary form of remainders of given number with respect to the given b value.

Conclusion

According to figure 1, the getting terms into the dictionary part is completed in 4.8 seconds in the first part. The appending sorted document ids into the dictionary which is previously created is completed in 5.17 minutes. Finally, the writing dictionary to the file “crawl.IDV” is completed in 8.10 minutes.

Figure 1

The time for part 1

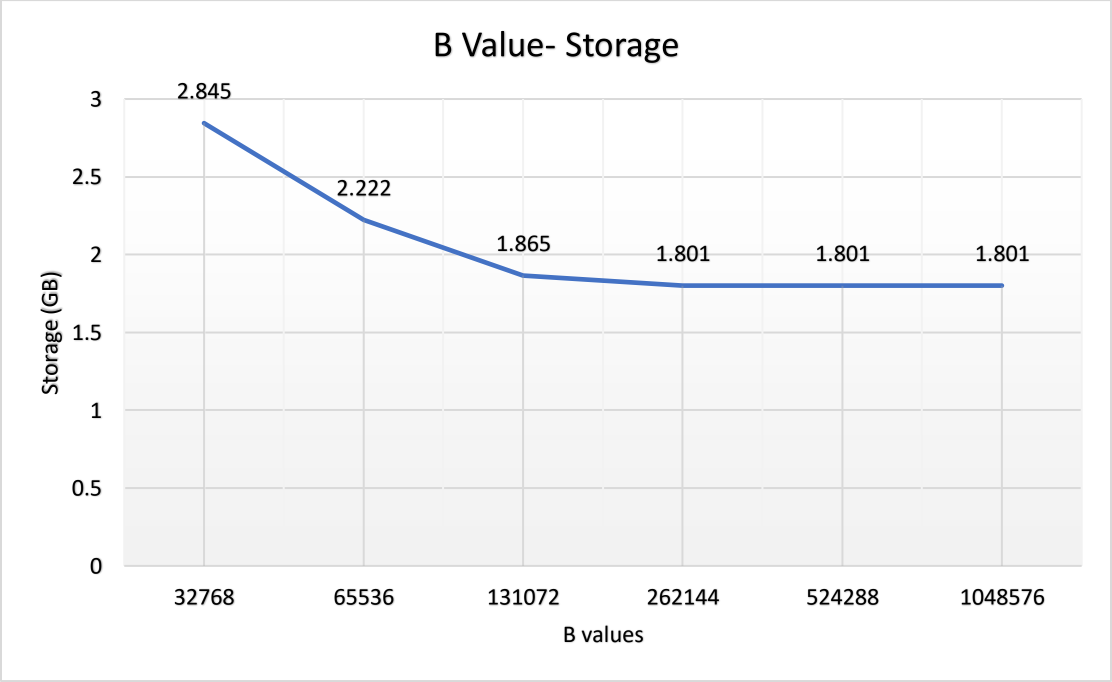


```
C:\Users\tol_b\Anaconda3\envs\cmpe414_project02\python.exe C:/Users/tol_b/
Time for getting terms into dic:  0:00:04.831485
Time for creating dictionary which holds sorted doc_ids:  0:05:17.093411
Time for writing dict to text:  0:08:10.422810
```

According to figure 2, in the second part, although the result of the equation of the optimal b value is 12406, the optimal b values that I found by trying are 2^{18} (262144), 2^{19} (524288), 2^{20} (1048576). I have used the value of 2^{18}.

Figure 2

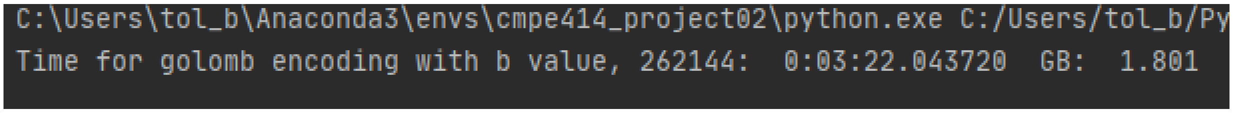
The b values with respect to the file size



According to figure 3, the golomb encoding part is completed in 3.22 minutes. To increase the performance of this part, I create a dictionary to store document ids as a key and its encoded version as a value. While it traverses “crawl.IDV” file, it uses this dictionary to exchange encoded version of this id instead of calculating all document ids’ encoded version when the code reads it.

Figure 3

The time for part 2



```
C:\Users\tol_b\Anaconda3\envs\cmpe414_project02\python.exe C:/Users/tol_b/Py
Time for golomb encoding with b value, 262144:  0:03:22.043720  GB:  1.801
```