Index Stages:

1. Start timer and memory recorder
2. Parse config
3. Empty output directory
4. Initialize stopword\_sets and stemmer
   1. Efficient, easy reuse latter
5. Process file by file
6. Initialize an empty set to record the words in this file
   1. Because it’s bag of word model, we don’t count frequency.
7. Preprocess file
   1. Tokenized
   2. Remove non ascii
   3. Casefold
   4. Remove punctuaion
   5. Remove stop words
   6. Stem the word
8. If the word appear before in this doc, skip. Else add to block with key=word, value=doc\_id
9. Check the size of block, if it exceeds limit, write to file

Merging stage:

1. Open two file, split to get values.
2. Compare values and output to third file

62005 terms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Block size (KB) | Index Time | Merge Time (seconds) | Total Run time (seconds) | Memory (Top 3 memory consuming lines) |
| 1024 | 978.5 | 2.5 | 981.0 |  |
| 512 | 470.6 | 4.24 | 474.8 |  |
| 256 | 265.5 | 7.0 | 272.4 |  |
| 128 | 173.7 | 10.2 | 183.9 |  |
| 64 | 137.9 | 14.5 | 152.5 |  |
| 32 | 125.7 | 20.0 | 145.7 |  |
| 16 | 120.1 | 26.6 | 146.7 |  |
| 8 | 122.3 | 36.0 | 158.3 |  |
| 4 | 129.7 | 49.7 | 179.4 |  |
| 2 | 142.0 | 85.1 | 227.1 |  |
| 1 | 164.4 | 134.8 | 299.2 |  |