**INFORMATION RETRIEVAL \_ ASSIGNMENT 3**

**Content**: Index Construction

1. PREPARING:

Download from Course Site:

- Dataset C50\_combined.zip

2. CONSTRUCTING INDEX

2.1. Remove stopwords (Using existing code)

2.2. Build vocabulary using 2 methods

* Blocked Sort-based Indexing with following parameters
  + **PARSE:** Parsing documents into records (termID, doc). Whenever the data size in memory is equal to a block size, write them into the file (call File\_Raw). Remembering to store another file for mapping term to termID.
  + **SORT:** Opening the File\_Raw, repectively loading an appropriate amount of records to memory for sorting and write the result into the File\_Sorted.
  + **MERGE:** Dividing the File\_Sorted into M portions. For each portion, using a buffer (having equal size to a disk block) to load data from that portion (called source buffers). One buffer is used to store the result (called des buffer). Whenever a source buffer is empty, filling it by data from disk. Whenever the des buffer is full, write it to the file (called File\_Index).
* Single-pass in-memory Indexing

3. QUERYING:

Using the existing code (from the previous class) along with the index built.