Assignment 3 (20 ptos)

Due Date: Jan 31st, 23:59pm, BR time.

This is a followup on the second assignment. Using the HTML pages that compose the mini-collection crawled in Assignment 2, build an inverted index that includes postings for the terms in the documents (i.e., for each term occurring in a document, all positions of occurrence should be annotated in the index). There is no need to use compression.

Given your collection is small, your index should fit in main memory which simplifies the design and implementation of your indexer program. Once done, you should write a report explaining what you have done and providing basic statistics on the inverted index produced, such as:

- Size in Kbytes of your mini-collection
- Size in Kbytes of your whole inverted index
- Size of the vocabulary (number of distinct terms)
- Average size of each inverted list as the number of postings per term

You might also include additional statistics that shed light on particularities of your index, statistics that you find interesting or revealing.