CMP2003 Data Structures and Algorithms (C++)
Term Project

-- Top 10 Frequent Words—

1. Project Definition

You are expected to write a c++ console application which reads files from Reuters-21578 documents collection appeared on the Reuters newswire in 1987 and find Top 10 frequent words used in the newswire articles. The Reuters-21578 collection is distributed in 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contain 1000 documents, while the last (reut2-021.sgm) contains 578 documents.

Each article starts with an "open tag" of the form:

<REUTERS TOPICS=?? LEWISSPLIT=?? CGISPLIT=?? OLDID=?? NEWID=??>

where the ?? are filled in an appropriate fashion and ends with a "close tag" of the form:

</REUTERS>

Here is an example of these article entries in the file:

```
 <REUTERS ... >
 <DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>cocoa</D></TOPICS>
<PLACES><D>el-salvador</D><D>usa</D><D>uruguay</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN> ... </UNKNOWN>
<TEXT> ...
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE>    SALVADOR, Feb 26 - </DATELINE>
  <BODY>Showers continued throughout
    the week in the Bahia cocoa zone, alleviating the drought since
    ...
    ...
    Brazilian Cocoa Trade Commission after
    carnival which ends midday on February 27.
   Reuter
  &#3;</BODY></TEXT>
</REUTERS>
```

Your program must be able to read words in articles in between <BODY> … </BODY> tags and insert each unique word into a suitable data structure.

Stopwords

A list of stopwords is supplied in stopwords.txt file. You should not count these words.

Definition of a word:

For simplicity assume that any contiguous block of alphabetic characters (letters from “a” to “z”, both upper and lower case) which includes at most one single quotation mark between these letters is a word. According to this definition the following sentence in a article:

" if we don't have local agreements settled by Thursday", has the words: “if”, “we”, “don’t”, “have”, “local”, “agreements”, “settled”, “by”, “Thursday”.

Main Requirements:

After reading and processing is over, your program must print “top 10” most frequent words used in these articles in **descending order**.

Additionally, the total time elapsed from the beginning of your code to the end of printing top 10 must be calculated and printed at the end of the execution.

Here is an example output:

```
<word1>     <word count>

<word2>     <word count>

<word3>     <word count>

<word4>     <word count>

<word5>     <word count>.

   .
   .
   .
   .
   .
   .
<word10>    <word count>

Total Elapsed Time: X seconds
```