資料結構 程式作業(一) Word Frequency Counter 報告

學號：0016046
姓名：蔡佩珊

實作方法描述：

(1) 將文章內標點符號去除，並轉換所有大寫字母成小寫：
  (1.1) 用 istream& getline (istream& is, string& str);存取 input file，
    以'\n'為依據取 string 直到文章結束。
  (1.2) 每取一段 string 用 size_t string::find (char c);找出標點/大寫字母，
    將標點/大寫字母用 string& string::replace (size_t pos1, size_t n1, const char* s);
    取代成空白字元/小寫字母。

⇨ 全文皆小寫。
⇨ 全文無標點，皆以空白字元為分開單詞/數字依據。

(2) 整理處理過後的文章的單詞，統計每一個相異單詞的內容及次數：
  (2.1) 將文章 string 用 explicit istringstream (const string);轉 istream。
    用 istream::operator>>string 存取文章，以空白字元為依據取 string 直到文章結束。
  (2.2) 用一 integer 變數計算截至目前相異單詞的種類數量。
    用包含 string/integer 變數的結構陣列記錄每個單詞/單詞已出現數量。
    (2.2.1) 若某單詞第一次出現，則加入結構陣列。
      若某單詞非第一次出現，則在結構陣列找到該單詞並且上修數量。

⇨ 全文相異單詞的種類數量已確定。
⇨ 全文相異單詞及其個別數量已統計完畢，記錄於結構陣列。

(3) 依據相異單詞的個別數量進行排序，由多至少：
  (3.1) 對記錄相異單詞的結構陣列用 bubble-sort 進行排序。

⇨ 記錄相異單詞的結構陣列由數量多至少排序。

(4) 輸出使用者要求出現次數前 k 名的單詞及其數量，以及程式執行時間
  (4.1) 輸出出現次數前 k 名。
  (4.2) 輸出程式執行時間

⇨ 標準輸出結果
⇨ 程式達成目標，結束

測試結果：

(test1)

Input file: e3 平台公布之測資 => test1_hw1_new.txt		
(略)		
Execution:		
./0016046_hw1 test1_hw1_new.txt 1	./0016046_hw1 test1_hw1_new.txt 5	./0016046_hw1 test1_hw1_new.txt 10
Output:		
(the,67) 0	(the,67) (of,45) (and,36) (to,25) (a,23) 0.0078125	(the,67) (of,45) (and,36) (to,25) (a,23) (that,17) (in,17) (we,16) (for,13) (new,12) 0.0078125

(test2)

Input file: DS_Project_1.pdf 之 Example1
A skunk sat on a stump. The skunk thought the stump stunk; the stump thought the skunk stunk.
Execution:
./0016046_hw1 test2.txt 2
Output:
(the,4) (skunk,3) (stump,3) 0

(test3)

Input file: DS_Project_1.pdf 之 Example2
I wish to wish the wish you wish to wish, but if you wish the wish the witch wishes, I won't wish the wish you wish to wish.
Execution:
./0016046_hw1 test3.txt 3
Output:
(wish,11) (the,4) (to,3) (you,3) 0