資料結構與進階程式設計（108-2）
程式作業二

作業設計：孔令傑
國立臺灣大學資訊管理學系

繳交作業時，請至 PDOGS（`http://pdogs.ntu.im/judge/`）為第一、二題上傳一份 C++ 原始碼（以複製貼上原始碼的方式上傳）。每位學生都要上傳自己寫的解答。不接受紙本繳交；不接受遲交。

這份作業的截止時間是 ***3 月 24 日早上八點***。在你開始前，請閱讀課本的第 8 和 18 章[1]。為這份作業設計測試資料並且提供解答的是***龔汶佑***。

第一題

（40 分）這次作業我們將針對一個文字上的「分類問題」（classification problem）實做一個分類演算法（常被稱為分類器，classifier），透過分析句子的長度，來猜測一段文字的作者是男生還是女生（我們將男生以 1 表示，女生以 2 表示）。根據前人研究指出，由於女生相較於男生比較不喜歡引起衝突、講話比較委婉，因此女生寫的句子平均而言會比男生的長。根據「女生所寫的句子會比男生的長」這個假設，我們將針對已知作者性別的歷史資料做分析，去找出一個用來判別性別的句子平均長度臨界值。我們將先透過訓練資料集（Training Set）來找到在訓練資料集中最好的臨界值，再透過驗證資料集（Validation Set）來檢測分類器的效能。在本題中，我們只處理英文。

我們先透過訓練集來找出最好的臨界值，方法如下。首先，針對一個給定的段落，我們用標點符號為分隔，拆解出數個句子，接著算出平均句子長度（每個句子所包含的字數的平均）。以範例

I enjoy going to school very much, and I also like to make friends with others.

為例，平均句子長度為 $\frac{7+9}{2} = 8$（個字）。接下來，我們要找到可以「最好」地分類男女的整數臨界值，也就是找到一個數字，平均句長大於等於該數字的段落就分類為女生所寫，比該數字小的就分類為男生所寫，而「最好的臨界值」決定的性別與真實性別相比後的錯誤能比其他臨界值的都小（或平手）。舉例來說，在表 1 的例子中，若我們選擇 7 作為臨界值，將會依序把這六個段落的作者性別判定為 2、2、1、1、1、2，一共有 2 個錯誤（第 1、4 筆）；而若選擇 6 作為臨界值，會依序判定為 2、2、1、2、1、2，只會產生 1 個錯誤，因此 6 是比 7 更好的臨界值。

性別	段落	平均句長
1	I enjoy going to school very much, and I also like to make friends with others.	8
2	I think the question is very hard. I spend a lot of time on it.	7.5
1	How are you? I am fine.	3
2	I never ate that food. I felt that the taste will be bad.	6.5
1	I love that drink. Do you love it too?	4.5
2	The last time I went to the zoo was ten years ago.	12

表 1: 範例訓練資料集

[1]課本是 Deitel and Deitel 著的 *C++ How to Program: Late Objects Version* 第七版。

在本題中，你將被給定一個訓練資料集與一個驗證資料集，你將在訓練資料集中做上述運算並找出最佳臨界值，再去驗證資料集中將每個段落做分類，並計算驗證資料集中的分類錯誤總數。由於我們只選擇整數作為臨界值，所以你可以搜尋所有可能的整數臨界值，並選擇最好的那個。如果有複數個整數臨界值都一樣地最好，請挑其中最小的那個。[2]

輸入輸出格式

系統會提供數組測試資料，每組測試資料裝在一個檔案裡。每個檔案會有 $n+m+1$ 行，第一行包含兩個整數 n 和 m，分別代表訓練資料集與驗證資料集中的資料筆數。第二行起的 n 行為訓練集，包含一個數字（表示性別）以及一串文字，並以分號分開。其中，文字所包含的字元數介於 1 到 10000 之間，只包含英文字母（大寫及小寫）、介於 0 到 999 的整數、空格，以及「,.;!?」這五種標點符號。我們以標點符號做為兩個句子的分隔，且已知在一個句子中用空格當做單字的分隔、每句最後一定有一個標點符號、不會有兩個標點符號中間沒有單字、不會有兩個空格相連。第 $n+1$ 行起的 m 行為驗證集，格式與訓練集一樣。已知 $2 \leq n \leq 1000000$、$2 \leq m \leq 10000$。

讀入上述資訊之後，請根據題意計算並輸出最好的臨界值，以及該臨界值在驗證集中的錯誤數，兩數以一個逗號隔開。舉例來說，若輸入為

```
8 3
1;I enjoy going to school a lot. And I also like to be nice to others.
2;I think the question is very hard. I spend a lot of time on it.
1;How are you? I am fine.
1;What is wrong with that machine?
2;Here you are.
2;I never ate that food. I felt that the taste will be bad.
1;I love that drink. Do you love it too?
2;The last time I went to the zoo was ten years ago.
1;No one knows her.
2;Where is the book I bought yesterday?
1;I think it is behind the desk.
```

則輸出應該為

```
5,1
```

你上傳的原始碼裡應該包含什麼

你的.cpp 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的 C++ 程式碼。當然，你應該寫適當的註解。針對這個題目，你可以使用任何方法。

[2]有些人的實作方式是在所有的平均句長中找尋致使錯誤最少的那個，但由於這題要找的是整數臨界值，所以請不要那麼做。

評分原則

這一題的 40 分會根據程式運算的正確性給分。PDOGS 會直譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。一筆測試資料佔 2 分。

第二題

（60 分）第二題的題目敘述和第一題一模一樣，但在第二題中，數字可能介於 0 到到 9,999,999 間，並包含正確的千分位符號，也可能會包含小數點，所以不能將數字中的千分位符號以及小數點當作句子的分隔。舉例來說，若句子為

I took an exam and I got 7.6. That made me lose 1,000 dollars because I had a bet with my friend.

則這段文字應該只被分為兩個句子。

輸入輸出格式

此題的輸入輸出格式和第一題一模一樣。

舉例來說，若輸入為

```
8 3
1;I enjoy going to school a lot. And I also like to be nice to others.
2;I think the question is very hard. I spend a lot of time on it.
1;How are you? I am fine.
1;What is wrong with that machine?
2;Here you are.
2;I never ate that food. I felt that the taste will be bad.
1;I love that drink. Do you love it too?
2;The last time I went to the zoo was ten years ago.
1;No one knows her.
2;Where is the book I bought yesterday?
2;I earned 1,000,000.99 dollars everyday due to my hard working.
```

則輸出應該為

```
5,0
```

請注意由於訓練資料集中的最佳臨界值是 5，若我們錯誤地把驗證資料集的第三筆資料切成四個句子，就會認為該段落的平均句長為 $\frac{13}{4} < 5$，並將其判定為男生所寫，這樣就會輸出 `5,1`；只有在我們正確地把該段落理解為一句時，我們才會將之判定為女生所寫，並且輸出 `5,0`。

你上傳的原始碼裡應該包含什麼

你的.cpp 原始碼檔案裡面應該包含讀取測試資料、做運算，以及輸出答案的 C++ 程式碼。當然，你應該寫適當的註解。針對這個題目，你***不可以***使用上課沒有教過的方法。

評分原則

- 這一題的其中 40 分會根據程式運算的正確性給分。PDOGS 會編譯並執行你的程式、輸入測試資料，並檢查輸出的答案的正確性。一筆測試資料佔 2 分。
- 這一題的其中 20 分會根據你所寫的程式的品質來給分。助教會打開你的程式碼並檢閱你的程式的運算邏輯、可讀性，以及可擴充性（順便檢查你有沒有使用上課沒教過的語法，並且抓抓抄襲）。請寫一個「好」的程式吧！