Homework #10: 中文斷詞(Chinese word segmentation)

Due Date: 2018/06/19 Tue.

Instruction

Please turnin the program to PD.hw10; if overdue, turnin the program to PD.hw10.delay.
請將作業 turnin 至 PD.hw10；遲交請 turnin 至 PD.hw10.delay。

Please finish demo before 2018/06/29 Fri. .
請於 2018/06/29 Fri. 前完成 demo。

Please contact pdta@gais.cs.ccu.edu.tw if any problem shall be encountered.
若有任何問題，請來信 pdta@gais.cs.ccu.edu.tw。

Identifying yourself and having proper signature are essential for TAs to reply.
請務必於信中表明身份，並於信末署名，以利助教群可以即時回覆。

Environment

CSIE workstations 系上工作站
csie0.cs.ccu.edu.tw
csie2.cs.ccu.edu.tw (外系學生請使用此機器)

Description

請撰寫一個中文斷詞程式,程式執行後從file或stdin讀入資料,請建立辭典資料結構(Hash 或 BST 或 Array)來進行斷詞,並輸出結果到stdout。

Requirement

- Command: ./hw10 argv[1] argv[2]
  argv[1]: 辭典檔,使用教育部國語辭典或其他辭典(教育部國語辭典，可到授課教材下載 ”dic.txt”)
  argv[2]: 文章檔案,如果沒有 argv[2] 則從 stdin 輸入文章
- 需建立Hash function 或 BST for dictionary
- 請使用正向長詞優先方法進行中文斷詞。

Hint

- 請加上-Wall -Wextra -Werror參數進行編譯。
- 文章中出現辭典裡不存在的單字,則單字直接印出來即可。

Grading Policy

a. 實現中文斷詞 (50%)。

b. 使用 Hash (50%) 或 BST(40%) 或 Array(20%)。

Sample I/O

Execute: gcc [*filename*.c] -o hw10
Execute:./hw10 [*dicname*] input
Remind:依提供的辭典不同，output可能會有所不同

8-1 input

```
使用長辭優先方法進行中文斷詞
```

8-1 output

```
使用
長辭
優先
方法
進行
中文
斷詞
```