# Homework #8: 中文斷詞(Chinese word segmentation)

Due Date: 2017/06/03 str.

## Instruction

Please turnin the program to PD.hw8; if overdue, turnin the program to PD.hw8.delay.
請將作業 turnin 至 PD.hw8；遲交請 turnin 至 PD.hw8.delay。

Please finish demo before 2017/06/17 str. (two weeks after the due date).
請於 2017/06/17 str. (due date 後兩週) 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

## Description

請撰寫一個中文斷詞程式，於工作站上編寫makefile編譯程式，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）來進行斷詞，並輸出結果到stdout。

## Requirement

1.Command: ./seg argv[1] argv[2]

1. argv[1]: 辭典檔，使用教育部國語辭典 (請到教材預覽下載 ”dic.txt”)
2. argv[2]: 文章檔案,如果沒有 argv[2] 則從 stdin 輸入文章

2.需建立Hash function 或 BST for dictionary

3.請使用正向長詞優先方法進行中文斷詞

4.Hint:

1.請加上-Wall -Wextra -Werror參數進行編譯。

2.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 (80%)

b. 使用BST 或 Hash (20%)

Bonus

c. 提早交K天加 k * 2 分. ex 6/1交作業而外加4分

d. 輸出文章內詞的出現頻率 (10%)

# Sample I/O

Execute: ./seg dic.text 8-1input
Remind: 執行程式後依序輸出格式需與下列完全相同。

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
```