祝平次的教學網站

語料分析與文史研究

課程資訊：

- 科號：10920CL 535200 ○科目名稱：語料分析與文史研究
- 上課時間：T7T8T9 ○教室：線上
- 授課老師：祝平次 ○助教：
- 學分數：3 ○研究室時間：R56，F56(利用遠距視訊，也可以另約時段。)
- 電郵信箱：ptchu@mx.nthu.edu.tw ○校內電話：42742
- **本課程將以遠距方式進行。**

一、課程說明

近三、四十年來大量的古代漢籍被數位化，怎麼利用這樣的數位文本來進行研究，已是中研讀中國古代文史學者不可迴避的課題。本課程就是為了回應這樣的課題而開設，以使修習本課程的同學，在修習完畢後，能夠利用基本的數位人文方法進行研究。課程設計的前半以建立自己的全文檢索、學會正則表示式、應用 MSExcel 強大的計算功能為主；後半則以台灣大學生傳系關河嘉團隊所開發的語料分析軟體《庫博》為主，來分析《全唐文》或《全唐詩》，藉由實作來了解數位人文的一些基本方法，和其中可能的問題。研究現當代的同學，也可以帶入自己分析的文本，來學習這些方法。

二、指定用書

林富士：《數位人 文白皮書》。

- 本課程所使用的數位工具/程式可以在下列網站取得：
  - Freeplane：
  - **https://www.freeplane.org/wiki/index.php/Home**
  - Notepad++：
  - **https://notepad-plus-plus.org/zh/**
  - AntConc：
  - **http://www.laurenceanthony.net/software.html**
  - Corpro 庫博中文語料庫-分析工具：
  - **http://dh.lis.ntu.edu.tw/resource.html**

§ MSExcel （校園授權軟體）

三、參考書籍

國立臺灣大學出版中心：《數位人文研究叢書》。

§ 本課程的輔助教學錄影，可以在 Youtube 透過搜尋檢得。

四、教學方法

以教師講授後，實際操作為主。期中，同學需準備進行之前工具的測試。每一週會有作業安排；每一週的開始，討論同學的作業情況。

五、教學進度

週數/日期 課程內容

(01) 02/23 課程介紹：數位人文簡介

. 介紹：本週將說明課程名稱的意義與問題、概略介紹本學期的課程，並說明本課程的課程安排。由於本課程會以遠距的方式進行，所以本週也會花一點時間確認遠距設備的使用。本週課程介紹，會很快地瀏覽這學期會用到數位工具，粗略地介紹它們能夠幫助我們做到的事情。也會介紹一些其它人文學科常用到的數位工具，具體地讓修課同學了解數位人文的可能性。

(02) 03/02 Freeplane 與網頁

- Freeplane 心智圖軟體：示範、說明與實做。
    - 心智圖裏面蘊涵了分析、綜合的兩種研究方式，也幫助我們對於所要分析、論述的課題，可以有一種全景式的掌握。Freeplane 除了做樹狀式的展開，也可以重新綜彙分支，進行總結；再繼續展開。對於研究議題的腦力激盪、議題整理都很有幫助。
    - 在這一節課程裏，同學將被要求一個多功能的心智圖，能夠呈現出資料的結構、分析綜合的關係、內部以及外部的超連結，並能夠利用心智圖進行類似報告。
- 標記語言：認識網頁的真相。
    - 標記語言，是目前處理數位資料幾種最常用的方式之一。本節課程的目的，一方面讓同學對網頁的超文本格式(HyperText Markup Language, HTML)有一基本的認識，一方面也藉由實做，讓同學了解網頁連結的機制。
    - 在這一節課程裏，同學將被要求一個資料來裏，做出五個互相連結的網頁。
    - 另外，現在有很多資料，往往是從網頁上下載；認識網頁的組成也有助於之後課程的資料處理。
- 作業：
    - 請想個題目，完成一張簡單的心智圖。
    - 請利用你學校的帳號（office 365 或 gmail）完成一個網站，裏面至少要有 5 個網頁，這 5 個網頁必須要有超連結加以連結在一起；亦即在每一張網頁裏，必須要有超連結連到其它 4 張網頁。
        - 從網站上下載自己建置的網頁，並觀察自己網頁的原始檔，寫下自己的發現或問題。

(03) 03/09 NotePad++（一）：中文內碼、全文檢索與正規表達式

- Notepad++是一款非常受歡迎的文字編輯軟體，我們將利用它來學會

- 1)如何利用 Notepad++來轉換中文檔案的內碼，
- 2)如何建立自己的全文檢索，
- 3)進行「規則運算式」(Regular Expression)的複雜檢索，
- 4)利用「規則運算式」的取代功能來清理資料。

· 概念：迴避字元(escape character)。

- 作業：
  - 利用 NotePad++建置自己的全文檢索，裏面至少要包括兩個文字檔
  - 想出 5 個關於規則運算式的問題，以便下次上課使用。

(04) 03/16 NotePad++（二）：資料處理與 MSExcel 的使用 。

- 在這一節課程裏，同學將被要求完成說明示範時所教授的各種功能，以具備清理數位文本的能力，以便進行進一步的文本分析。
- 並且進一步利用 Excel 來統計分析處理過後的資料。
- 作業：
  - 每個人先找兩本有電子文本的書完成它們的字頻統計。
  - 再將兩本書的字頻統計加以比較。

(05) 03/23 03/16 NotePad++與 MSExcel（三）

- 這個禮拜，我們要總體回顧 NotePad++，加以討論並實做一些測試問題。
- 作業：
  - 閱讀 CText.org 線上平台的使用說明。
  - 也鼓勵大家觀看 CText.org 的教學錄影。

(06) 03/30 **CText.org 線上平台與數位分析工具** （一）

- 資料比對是文史工作者必須常常面對的問題、從事的工作，也與我們種種對文本的判斷有所關聯。而電腦的特長正在於快速的資料比對，全文檢索就是一個很好的例子。在這兩週的課程中，我們將利用 Notepad++處理資料，然後利用 CText.org 來統計、分析資料，並將統計分析的結果視覺化為圖表或網絡關係。視覺化不但會影響到我們的研究視角，也是教學上的利器，更是文史科系與其它學科進行交流最好的橋樑。而如何藉由處理資料、分析資料、統計資料，完成從文字到數字、表格與圖表的轉換，正是本次工作坊要完成的目標。
  - 概念：N-Gram、文本異同比較。

        - 講義連結
    - 作業：每個人在下星期提供 5 件你想要利用這個平台完成的事。

(07) 04/06 校引陙活動週。

(08) 04/13 CText.org 線上平台與數位分析工具（二）
- 這個禮拜我們將實做 CText.org 平台的數位工具。

(09) 04/20 期中測試
- 每位修課同學各自跟授課教師約測試時間。

(10) 04/27 庫博一：安裝與概介
- 本節將請台灣大學闕河嘉老師介紹她利用文本分析所做出來的研究例，以讓同學了解文本分析在現今社會可能的研究對象。
- 作業：
    - 閱瀏庫博使用說明書。
    - 想出一個你要利用庫博來進行研究的學術議題，以便下週討論。

(11) 05/04 庫博二：與研究議題的扣合
- 每位修課同學，介紹自己要研究的議題，並與同學討論。
- 作業：
    - 依據與同學討論結果，來改進自己對於研究議題的設想。
    - 觀看庫博的教學錄影。
    - 此後的作業，都是利用庫博來整理自己要研究的文本與思考怎麼利庫博來完成期末報告。

(12) 05/11 庫博三：功能與實做一
- 詳細逐步地介紹庫博的功能，並進行實做。

(13) 05/18 庫博四：功能與實做二
- 詳細逐步地介紹庫博的功能，並進行實做。

(14) 05/25 庫博五：期中報告。
- 同學進行研究議題的初步報告。

(15) 06/01 庫博五：期中報告。
- 同學進行研究議題的初步報告。

(16) 06/08 庫博六：綜合討論。
- 對於本學期課程進行總體的回顧。

(17) 06/15 期末報告一

(18) 06/22 期末報告二

六、成績考核

- 平時成績百分之 30，包括上課參與度、對於軟體功能操作的熟悉度。
- 期末報告頁分之 70，博士班同學應完成一篇以應用庫博與課程介紹的其它軟體進行研究的報告。碩士班同學，可以選擇同博士班一樣的研究報告，或選定一個經授課教師同意的文本進行探勘報告。

七、講義位址

- https://sites.google.com/a/ptc.cl.nthu.edu.tw/courses/