# W1

## 講什麼

數字人文兩條線索：古籍數字化、

計算機語言相關

字符編碼

## 计算机语言1

* 正則表達式（非完全圖靈的）
  + NFA
  + DFA
* eg
  + [說説脫说]文解字[注註]
  + 王念孫 王引之 終.且.
* python
  + 動態腳本語言

## 計算機語言2

* 標記語言
  + markdown
  + HTML
  + OOXML

## 要求

* 課堂表現20%
* 作業30%
* 報告30%
  + 用所學的技術方法解決研究中的問題（不需要太大）
* 成果20%
  + 1000-2000，不可超過3000字

## 有什麼用

### 案例一：詩經句式問題

* 搜索 詩經原始文獻
* 整理
* 分析 如何分析句式（如何數字數）
* 呈現 怎樣最為直觀

#### 搜集

* 何處更可靠
  + baidu/bing
  + ctext（中國哲學書電子書計劃）
  + 殆知阁
  + xueheng学衡数据
  + github
  + ⋯⋯
* 如何下載
  + 部分
  + 單一文件/頁面
* 以何種方式保存
  + 文件格式
    - docx（用python處理很麻煩，且不支持正則表達式）
    - html（標記語言的一種）
    - 某一筆記軟件
    - （筆記最好用輕量的工具記）
    - syncthing（去中央化的P to P），設備之間只要打開聯網就可以即時同步，無法做到設備一直打開可以買個服務器相互連接
    - txt
  + 其他
    - 文件編碼
    - 換行符
    - 命名方式
* 有心搜集自己研究領域的文本資源

#### 整理

* 去除羨餘
  + metadata
  + 代碼
  + 標示符號
  + 額外空白符
  + ⋯⋯
  + 保留需要
  + 正訛誤
* 統一格式
  + 部分區分
  + 卷題
  + 詩題
  + 尾題
  + 章句
  + 詩序
  + 詩每一章

#### 分析

* 編寫程序分析
  + 文本清理
  + 每一部分之區分

#### 呈現

* 直觀方便有助於說明問題

### 案例二：段校廣韻研究中的文本移录問題

* 段玉裁分部
  + 文本移录：早期（堅守同諧聲者必同部，中後期也加入審音的考慮）
  + 六書音韻表：中期
  + 說文段注“\*部”：晚期
* 檢查移录是否錯誤——辦法？ 眼看+耳聽

if mark：

print(“\a”) #發出一個bell

### 小結

* 搜集 整理 分析 呈現

## 工作應用

### 案例分析：格式轉換與錄入

* 常規方法
  + 對著錄入
  + 複製
  + date：人工分割
* 部分自動化
  + 快捷鍵
  + 鍵鼠紀錄
  + Date：人工分割
* 網絡有相關API（接口暴露）
* 全自動化

## 前期準備

### 文本編輯器1

* atom
* sublime text（插件可以用python寫）
* vscode

### 文本編輯器2

* vim
  + （win系統）gvim
  + neovim
  + vscode的vim插件會導致生僻字符出不來
  + 學習資源
    - cheatsheet
    - help
    - game
* GNU Emacs

### 正則表達式相關

* regex101
* 命令行工具
  + ripgrep
* 文本編輯器
  + sublime text（插件可以用python寫）
  + editpad lite
* python包
  + re（有限）
  + regex（全面）
  + 圖形化dngrep（可以針對文件夾使用）

### 虛擬機

* windows
  + hyper-V
  + WSl
* macos
  + parallels desktop
  + VMware
  + M1M2安裝win11的arm版本
* Linux
  + KVM
* 安裝配置虛擬機
  + 安裝自己喜歡的Linux發行版
    - ubuntu
    - manjaro
  + 安裝python
    - pyenv（區分不同包的環境避免互相影響）
    - pyenv- visualenv