課程介紹與環境設定

台師大通識教育課程

文本分析與程式設計

授課：卓騰語言科技 _ PeterWolf

課程介紹:給分系統

- 作業 60%:每週都會有作業，請於當週六午夜12:00 以前提交至課程 Github 源碼倉中。https://github.com/PeterWolf-tw/NTNU_TextProcessing_2020
  - 專題 30%:其中 10% 由同學給分。另外 20% 由三個面向給分:
  - 創意程度 8%　(是新的應用嗎？還是新的方法？)
  - 社會服務 8%　(解決社會上的什麼問題？)
  - 實作可行性 4%（天馬行空以後，還是要考量怎麼落地的)
- 課堂討論 10%:課程簡報中同步提問、實作中提問、主動對其它專案提問，予以加分。

課程介紹:程式語言

Python3.6+

https://kopu.chat/2017/01/18/一小時python入門-part-1/

https://docs.python.org/zh-tw/3/tutorial/introduction.html#numbers

課程介紹:程式語言

Git & Github.com

https://blog.techbridge.cc/2018/01/17/learning-programming-and-coding-with-python-git-and-github-tutorial/

課程介紹:文本分析是什麼?

人工智慧

自然語言處理 NLP | 影像辨識 | 自動控制 | ...

Text Processing | Speech Processing

資訊檢索、文本探勘、語意分析、意圖分析、機器翻譯、文本分類、文本生成、情緒分析、自動語音辨識、自動語音合成...

...

中文裡，什麼是一個「字」？

英文：

lemma: cat = cats

wordform: cat != cats

中文：

蜻蜓、蚯蚓、車門

字 (word): 句子裡的獨立意義段落

字符 (character): 字碼表裡的獨立符號

詞 (phrase): 字+詞綴(構詞/句法)

符記 (token): 自定切分規格後的結果

詞彙 (lexicon): 字典中列出的獨立項目

詞條 (entry): 資料庫中列出的獨立項目

所以...這裡有幾個字？

英文：

I know uh a tech-guy and he's good at NLP. :)

中文：

啊我朋友就認識一個家裡養了綿羊的小朋友

語言是什麼？文字是什麼？

中文：

s w w w w w w w p w w w w

啊/我/朋友/就/認識/一個/家/裡/養了/綿羊/的/小朋友

1. 語言是「語音」+「意義(語意)」+「結構(句法)」
2. 文字是「語言的記號」
3. 文本分析的目標在處理「語言記號」(音/意/結構)，而不是單純的「符號分佈」而已

思考一下：
許多 NLP 的教學裡會提到「在自然語言處理中，詞是最小而且有意義單位」這句話是有問題的！

Quiz:

試算以下句子：

這個星期日本想往後山藥師佛寺去世人罕至處想一想自己的人生

字數：16

字符數：28

詞數：2

Assignment:

1. 在 Github.com 中建立自己的帳號，並通知助教你的帳號 email 以便後續作業提交。
2. 若你有電腦，請在自己的電腦上安裝 Python3.6+ 的版本，並成功在畫面中印出 "Hello World" 後，擷圖傳給助教。
3. 助教的聯絡方式：
    a. 林融 60621032L@gapps.ntnu.edu.tw
    b. 蘇洪寬 jacksugood@gmail.com