英文小說人物網路圖生成及分析

presentation, code: 吳承鋮（組長）

ppt 製作：陳信宇（組員）

專案目標

- 數位人文 (digital humanity) 中的鳥瞰 (distant reading) 在以統計模型，自然語言處理的方式萃取人文要素，例如作者風格，創作背景，角色分配（主角群，敵方群 ... 組成）
- 我們認為鳥瞰對於中小學生有教育意義，讓他們以可以發現文學作品中的要素，例如不同故事性的文學（如群像劇跟後宮劇）通常會有顯著的人際網路不同，但同一個作者做的不同類型故事性的文學可能會在鳥瞰中某個提取的元素中相同

專題說明

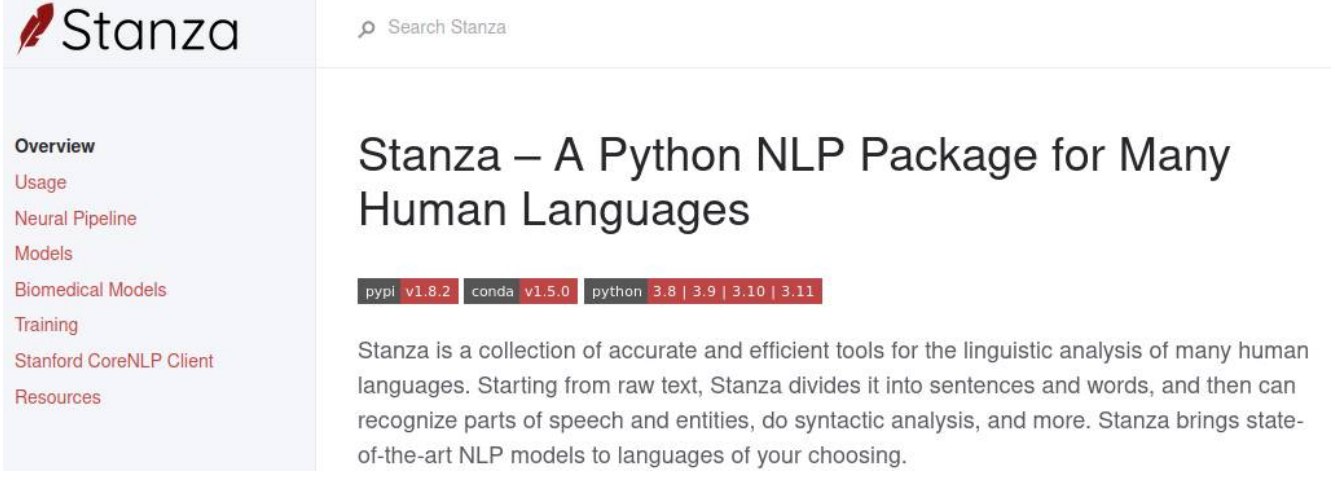
- 本文發想自” Unsupervised cluster analyses of character networks in fiction:Community structure and centrality”(2018)，在鳥瞰 (distant reading) 的領域上，提供新的 clustering method-MSC，僅需要角色間的距離，可以數據化人物結構
- 因為角色間的距離能做的分析不只有 MSC，本專案致力於小說輸入到輸出角色間的距離，留下後續選擇分析方法的自由。並且提供了可視化界面，讓沒有程式基礎的中小學生也能透過此專案文字分析。

Pipeline

- NLP.py: 以 stanza NLP 提取 text.txt 中角色
- character_select.py: 讓使用者用 gradio 輸入參數及調整角色，生成 lex program，執行後生成 raw distance matrix（角色間距離）
- threshold.py: 將 distance matrix 中的距離化成 histogram，讓使用者選擇截止距離

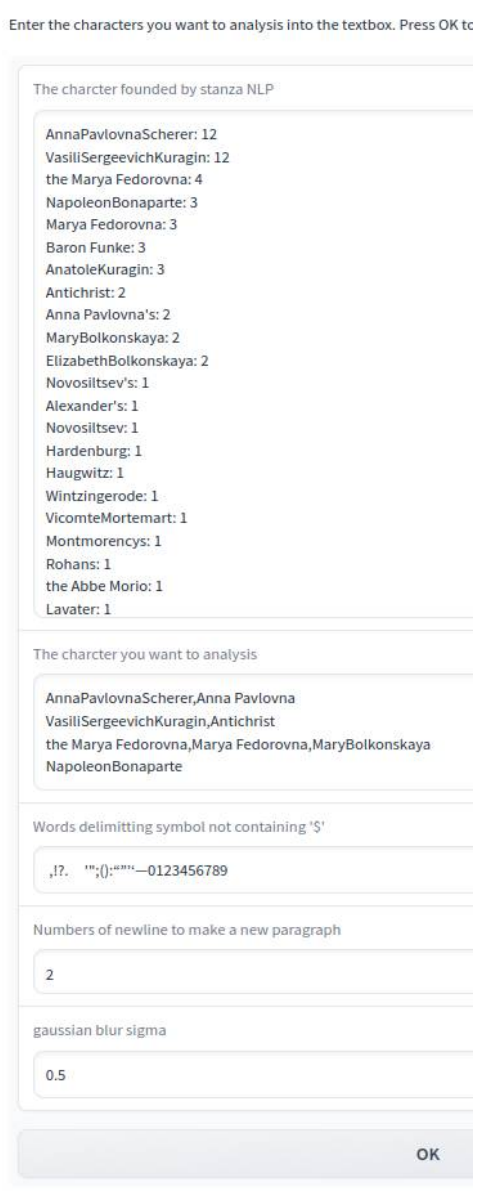
NLP.py

- input:python3 NLP.py (要分析的檔案)
- output:namelist0.txt namelist1.txt(統計 NLP.py 抓出的角色出場次數)



character_select.py

- textbox1:NLP 找到的（附出場次數）
- textbox2:user 想分析的
- delimitting symbol(s): 配合 lexcode.l
- 以幾個 \n 當換行
- gaussian blur sigma: 距離處理參數
- 按下 OK，即產出距離矩陣



aliasing problem

- “ 要怎麼在找 Mary 時不要找到 Marya 的 Mary?”
- “ 要怎麼讓 Princess Marya, Marya, Marya Fedorovna 認為是同一人” ?
- 我們使用 lex 可以同時解決以上問題

lexcode.l

- 程式碼由 character_select.py 生成後立即被執行
- 如圖，”　,NapoleonBonaparte?” 會 return 4
- lex 是用於編譯器設計的詞法分析器

```
oksyms [ ,!?.   '";():“”’‘–0123456789\n$]
%%
{oksyms}"AnnaPavlovnaScherer"{oksyms} return 1;
{oksyms}"Anna Pavlovna"{oksyms} return 1;
{oksyms}"VasiliSergeevichKuragin"{oksyms} return 2;
{oksyms}"Antichrist"{oksyms} return 2;
{oksyms}"the Marya Fedorovna"{oksyms} return 3;
{oksyms}"Marya Fedorovna"{oksyms} return 3;
{oksyms}"MaryBolkonskaya"{oksyms} return 3;
{oksyms}"NapoleonBonaparte"{oksyms} return 4;
```

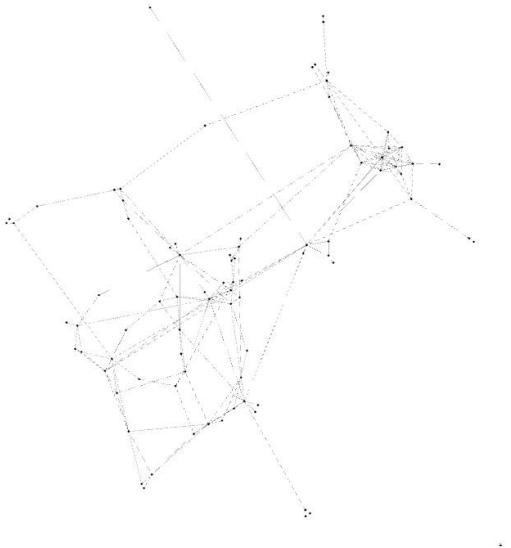
```
AnnaPavlovnaScherer,Anna Pavlovna
VasiliSergeevichKuragin,Antichrist
the Marya Fedorovna,Marya Fedorovna,MaryBolkonskaya
NapoleonBonaparte
```

distance calculating

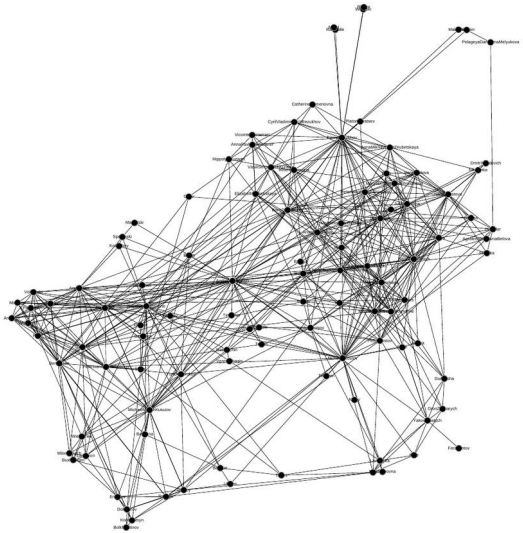
- 用 N dim vector 表示角色，列出在 N 個段落中的出場次數，而後經過 gaussian blur(sigma 由使用者輸入)
- 距離由**段落出現相關係數**控制
- distance=sqrt(1/max(1e-4,corr(v[i],v[j])))
- 使用相關係數，代表同時出場多，或同時出場少
- 以 tone 來更準確的判斷拉近 / 拉遠?

Threshold setting

- 若 distance>threshold，此邊會被捨去
- 若捨去不重要的邊，會使角色分佈變得更有架構



threshold=3



threshold=6

Threshold setting

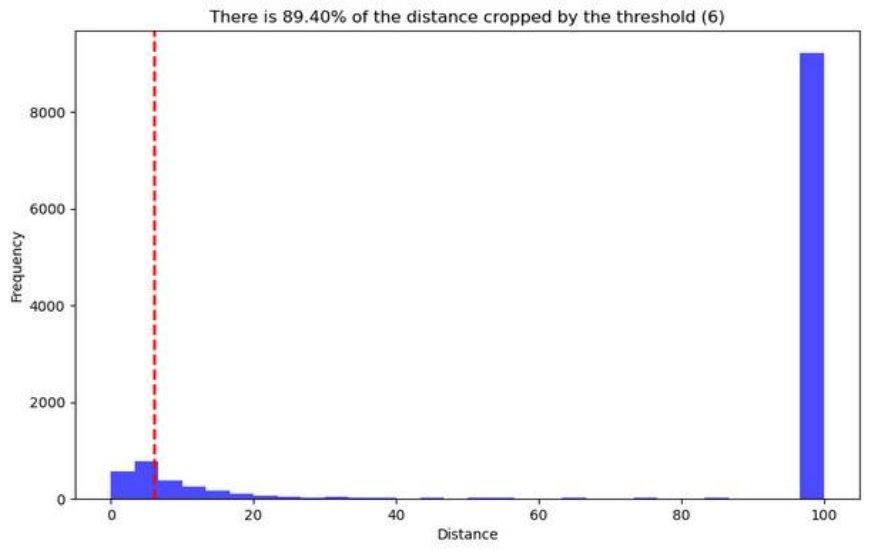
Threshold

6

Clear

Submit

Histogram of Distances



Summary

There is 89.40% of the distance greater than the threshold 6

Flag

Final output

- character_info.csv
- distances_list
- name_label.csv
- distmtx.csv

Character	Node Degree	Appearance	Same Appearance Count
1	11	1886	1
2	11	1213	1
3	10	1181	1
4	11	1122	1
5	5	719	1
6	8	736	1
7	3	540	1
8	9	449	1

source	target	id	Distance
1	2	1	2.813868
1	3	2	2.509118
1	12	3	2.439912
1	16	4	2.465
1	18	5	2.642532
1	19	6	2.431343

	Standard	Standard
1	Id	label
2	1	PierreBezukhov
3	2	AndrewBolkonski
4	3	NatashaRostova
5	4	NicholasRostov
6	5	MaryBolkonskaya
7	6	NapoleonBonaparte

Future work

- 1.Graph visualization
- 2.clustering
- 3.log(Appearance)-log(same appearance count)
- 4.GNN
- 以下用戰爭與和平（全 16 書）演示

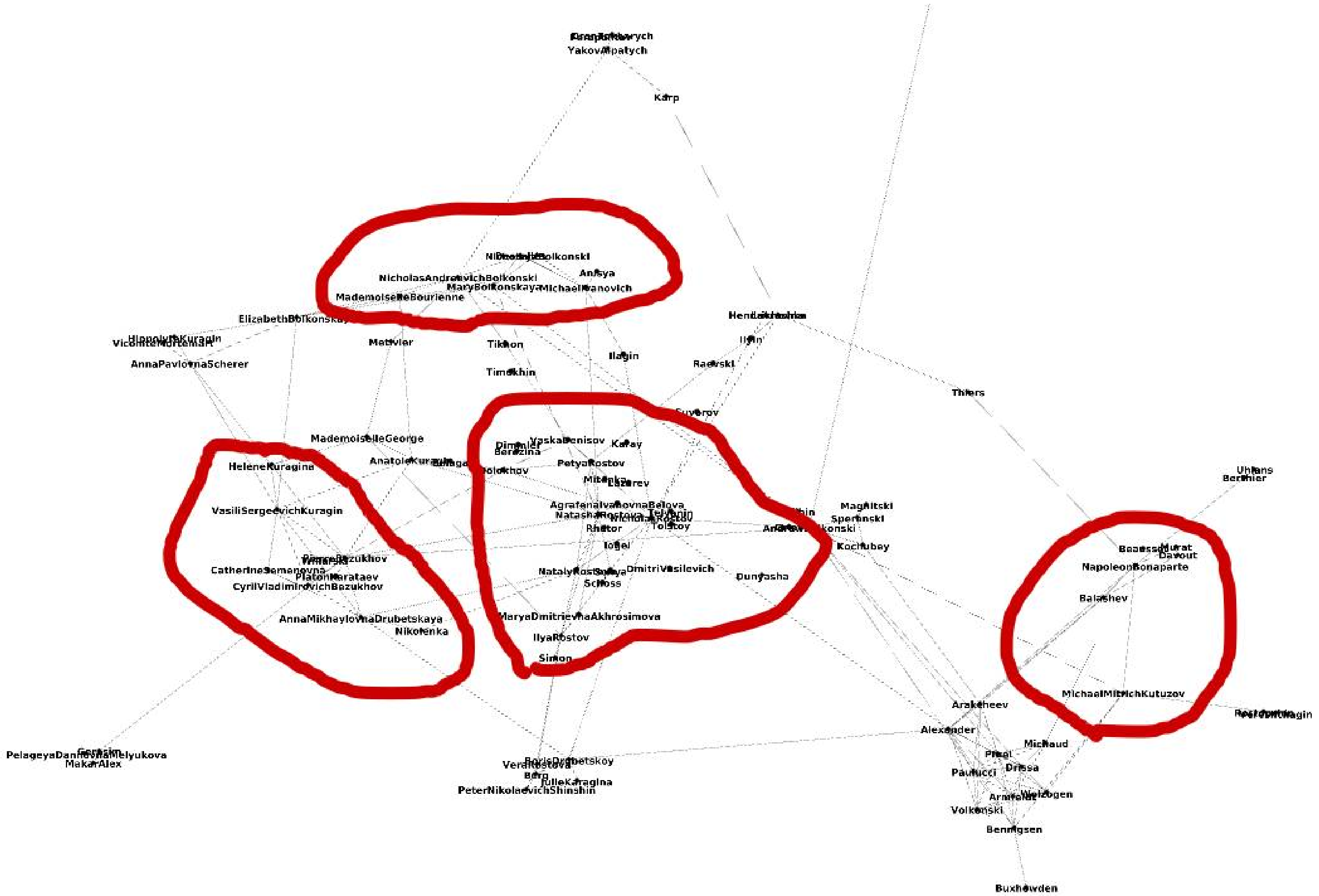
  https://www.gutenberg.org/files/2600/2600-0.txt

Graph visulization-ForceAtlas2

- ForceAtlas2
- 使用 2D 空間的帶電粒子 / 彈簧模型，力平衡後得到平面呈現圖，通常這個方法最能呈現以段落出現相關係數得到的距離

$$F_a(n_1,n_2) = d(n_1,n_2)$$

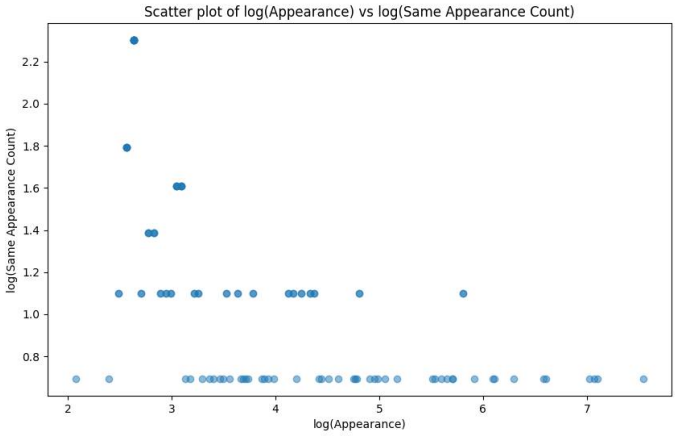
$$F_r(n_1,n_2) = k_r \frac{(deg(n_1)+1)(deg(n_2)+1)}{d(n_1,n_2)}$$


YakovAlpatych
Karp
NicholasAndreevichBolkonski
MaryBolkonskaya
MichaelIvanovich
Anisya
MademoiselleBourienne
ElizabethBolkonskaya
HippolyteKuragin
AnnaPavlovnaScherer
Tikhon
Timokhin
Ilagin
Raevski
Thiers
Suvorov
MademoiselleGeorge
HeleneKuragina
VasiliSergeevichKuragin
AnatoleKuragin
VaskaDenisov
Karay
PetyaRostov
Lazarev
AgrafenaIvanovnaBelova
NatashaRostova
Rhetor
Iogel
Tolstoy
CatherineSemenovna
PlatonKarataev
CyrilVladimirovichBazukhov
AnnaMikhaylovnaDrubetskaya
Nikolenka
DmitriVasilevich
Schloss
Dunyasha
MaryaDmitrievnaAkhrosimova
IlyaRostov
Simon
Magnitski
Speranski
Kochubey
Murat
Davout
NapoleonBonaparte
Balashev
Uhlans
Berthier
MichaelMitrichKutuzov
Arakcheev
Alexander
Michaud
Pfuel
Drissa
Paulucci
Wolzogen
Volkonski
Bennigsen
Buxhowden
PeterNikolaevichShinshin
Berg
JulieKaragina

log(Appearance)-log(same appearance count)

- 對於出場次數 (A) 少的人，有相同出場次數 (f) 的人會變多，滿足 $f \propto A^{-\gamma}$.

戰爭與和平



紅樓夢

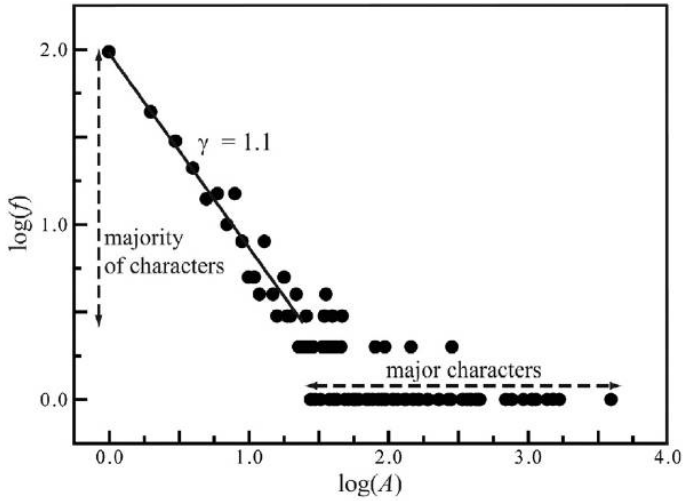


Fig. 1. Frequency (f) in the number of character appearances (A) in *Dream of the Red Chamber*. Here A denotes the counting of appearances of each character in the novel, and f denotes the number of characters with A appearances. The solid line is a line of best fit, given by the equation $f \propto A^{-\gamma}$.

clustering

- 我們可以用 clustering method: MSC, K means

  用鳥瞰的方式達成對角色的分群
- 探討 close reading 的人為 cluster 跟 unsupervised 的 cluster 優劣

Graph Neural Network(GNN)

- GNN 有 3 大常見任務
- 1. node classification （預測角色種類）
- 2.edge prediction （預測小說走向）
- 3.graph classification （歸納小說種類）

展望

- stanza 換成更好的 NLP
- 用情緒處理獲得更好的距離呈現（可以 custom）
- lex 換成其他 tokenizer
- 爬取大量文章用 GNN 做 graph classification