220006

以自然語言處理擷取西域各國間之空間資料

西域傳 | 自然語言處理 | 漢朝地理

研究目標

自然語言處理

擷取史書文本中關於各國相對於某地點的距離與方位資料

統整成表格

整理文本中各國資料，並擷取該國戶、口、兵、風俗、河川等資料

研究方法：文本

提及西域的篇章

01 史記卷 123

02 漢書卷 96

03 後漢書卷 88

研究方法：選擇的模型

(.*)

傳統模型 + 正規表示式

- Jiayan 模型
- Stanza 模型
- Jieba 模型
- **HanLP 模型**

GPT-3.5

串接API，修改 prompt，產生表格

研究方法：擷取內容

擷取國名、治所、地點、里程、來源等

距離京城或地標的里程，如陽關、玉門、洛陽等

補述鄰國、河川、戶數、人口、戶口、兵力等

以表格的形式呈現，並製成圖片

目前結果

以範例文本爲例，各模型分詞的結果之評估指標表

	Jiayan	Jieba	Stanza	HanLP
TP	22	15	19	23
TP+FP	30	22	29	24
TP+FN	25	25	25	25
Precision	0.733	0.682	0.655	0.958
Recall	0.880	0.600	0.760	0.920
f1-score	0.800	0.638	0.704	**0.939**

使用 nltk 的 RegexpParser ，找到符合「描述地點的文法」的句子

目前結果：已知語法

1. 專有名詞 + 介詞/動詞 + 專有名詞 + 數詞 + 量詞
2. 專有名詞 + 介詞/動詞 + 專有名詞 + 方位詞
3. 專有名詞 + 介詞/動詞 + 專有名詞 + 方位詞 + 數詞 + 量詞
4. 代名詞 + 方位詞 + 數詞 + 量詞 + 動詞 + 專有名詞

目前結果：傳統模型

國名	相對地點	方位	里程
大宛國	長安		12550里
康居國	京	西	10600里
奄蔡	康居國	西北	2000里
大月氏	大宛	西南	
大夏國	嬀水	南	
大宛	匈奴	西南	
拘彌國	于窴		300里
玉門關	沙州壽昌縣	西	6里
蒲昌海	蒲類海	東	
鹽澤	長安		5000里
烏孫	大宛	東北	2000里

國名	相對地點	方位	里程
康居	大宛	西北	2000里
月氏	敦煌	以東	
祁連山	甘州	西南	
火山國	扶風	南東	
條枝	安息	西	數千里
女國	于窴國	南	2700里
于窴	京		9670里
大夏	大宛	西南	2000餘里
天竺	崑崙山	南	
大夏	漢		12000里
身毒國	大夏	東南	數千里

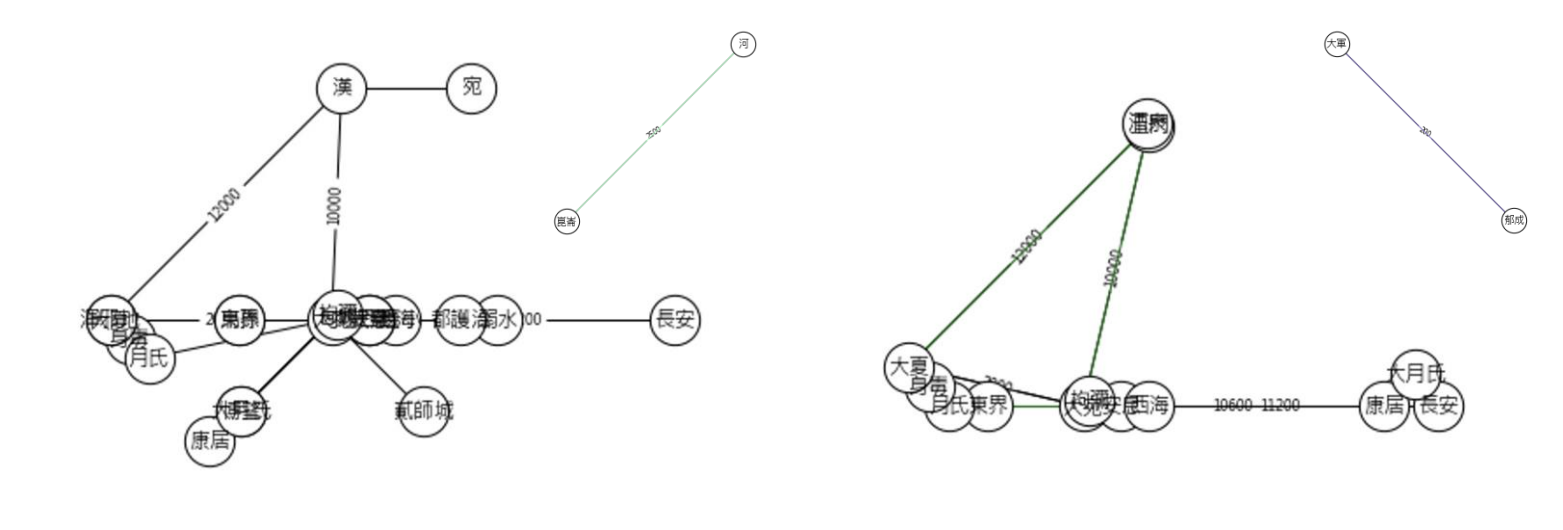
目前結果：GPT-3.5

國名	相對地點	方位	里程
大宛	長安	西	12550里
大宛	都護治	西	--
大宛	大月氏	東北	--
大宛	大月氏	北	--
大宛	康居	南	--
月氏	大宛	西	數十日
大宛	康居	西	10600里
康居	大月氏	西南	1600里
大宛	玉門	西南	10000里
拘彌	大宛	東北	300里
鹽澤	于窴	西	5000里
安息	長安	西	11200里
安息	阿蠻	--	3400里
安息	斯賓	--	3600里
安息	羅國	--	960里

國名	相對地點	方位	里程
安息	西海	西	1000里
安息	弱水	西	--
大秦	弱水	西	未知
大夏	大宛	西南	2000里
大夏	身毒	東南	1000里
身毒	月氏	東南	1000里
天竺	崑崙山	南	--
王舍	靈鷲山	--	40里
安息	西海	西	1000里
蜀	身毒	--	不遠
匈奴	祁連山	西	--
南山	金城	西	--
烏孫	匈奴西邊小國	西	--
大夏	烏孫	西	--
烏孫	渾邪地	東	--

國名	相對地點	方位	里程
酒泉	玉門	西北	118里
安息	東界	東	數千里
烏孫	安息	西	--
大宛	貳師城	東南	--
小國	鹽水	西	--
郁成	貳師將軍所在地	東	--
王都	郁成	東	--
敦煌	貳師將軍所在地	東	--
敦煌	郁成	西	--
大軍	郁成	東南	200里
宛	漢	東	--
外國	宛西	西	--
敦煌	鹽水	西	--
河	崑崙	東北	2500里
河	積石山	東	--

目前結果：建立圖片



預定進度

期中

改良圖形（邊、節點）

納入更多文本

優化建圖算法

提升準確度

預定進度

期末

利用 GIS 疊圖分析，並對照現代之地圖做對應

前往西域的實際路線，評估路徑的合理性

推測當代單位的實際長度

延伸不限於西域，自動處理文本得到地理的資訊

目前結論

- 本研究比較了四種能處理文言文的模型，模型 HanLP 表現最佳
- 成功使用正規表示式來篩選句子
- 加入 GPT-3.5 提升本研究
- 可擷取出國家名、相對位置與距離等資料，未來將會擷取更多資訊
- 最終，研究目標是將文本轉爲電腦可讀取的資料，並使用地理資訊系統進行分析，瞭解西域各國的相對位置，並與現代對比
- 未來可延伸至更多歷史文本，擴充至世界各地的文獻

Thank You

編號：220006