Word senses on WordNet — Wikipedia

NLP Lab 02

學號 ： 106062527 姓名： 孔啟熙

1. Implementation

主要使用了方法二：以中文作為sense去比對，以下為邏輯結構：

首先建立兩個dict:

（一）WordNet 中英對照：sense ID->中文
（二）Wiki Link：中文->Wiki Page ID

1. 每一組Sense list對應到的中文若有在 Wiki Link的dict中，則直接以sense中文輸出wiki page，且若一個sense有多個中文（如 汽車|轎車）則只有找到一個page就break loop.

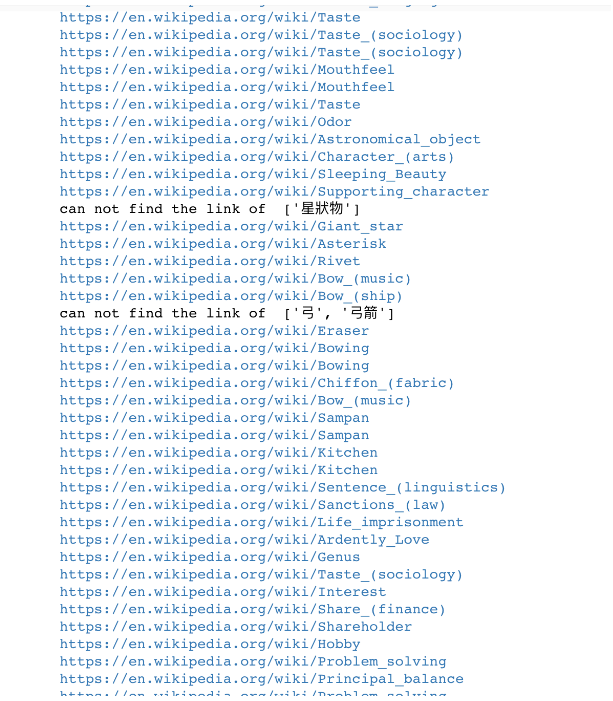
2. 若1. 有對到Wiki Link dict但是以中文輸出發生Key error，則再以Wiki Link Dict中對應的Page ID輸出wiki link，且若sense有多個中文如同1.

3. 若1. 2.皆失敗則表示sense的中文在Wiki Link Dict中不存在，則將sense中文與Wiki Link Dict中所有中文key做Word2Vec比較相似度，且若sense有多個中文則每筆都比較出最高的相似度輸出.

4. 若仍然無法配對到Page則將sense中文以結吧斷詞取出最後一段最具代表意義的詞再做Step 3 一次

Else: Can not match to Wiki

2. Result & Analyze



```
https://en.wikipedia.org/wiki/Taste
https://en.wikipedia.org/wiki/Taste_(sociology)
https://en.wikipedia.org/wiki/Taste_(sociology)
https://en.wikipedia.org/wiki/Mouthfeel
https://en.wikipedia.org/wiki/Mouthfeel
https://en.wikipedia.org/wiki/Taste
https://en.wikipedia.org/wiki/Odor
https://en.wikipedia.org/wiki/Astronomical_object
https://en.wikipedia.org/wiki/Character_(arts)
https://en.wikipedia.org/wiki/Sleeping_Beauty
https://en.wikipedia.org/wiki/Supporting_character
can not find the link of  ['星狀物']
https://en.wikipedia.org/wiki/Giant_star
https://en.wikipedia.org/wiki/Asterisk
https://en.wikipedia.org/wiki/Rivet
https://en.wikipedia.org/wiki/Bow_(music)
https://en.wikipedia.org/wiki/Bow_(ship)
can not find the link of  ['弓', '弓箭']
https://en.wikipedia.org/wiki/Eraser
https://en.wikipedia.org/wiki/Bowing
https://en.wikipedia.org/wiki/Bowing
https://en.wikipedia.org/wiki/Chiffon_(fabric)
https://en.wikipedia.org/wiki/Bow_(music)
https://en.wikipedia.org/wiki/Sampan
https://en.wikipedia.org/wiki/Sampan
https://en.wikipedia.org/wiki/Kitchen
https://en.wikipedia.org/wiki/Kitchen
https://en.wikipedia.org/wiki/Sentence_(linguistics)
https://en.wikipedia.org/wiki/Sanctions_(law)
https://en.wikipedia.org/wiki/Life_imprisonment
https://en.wikipedia.org/wiki/Ardently_Love
https://en.wikipedia.org/wiki/Genus
https://en.wikipedia.org/wiki/Taste_(sociology)
https://en.wikipedia.org/wiki/Interest
https://en.wikipedia.org/wiki/Share_(finance)
https://en.wikipedia.org/wiki/Shareholder
https://en.wikipedia.org/wiki/Hobby
https://en.wikipedia.org/wiki/Problem_solving
https://en.wikipedia.org/wiki/Principal_balance
https://en.wikipedia.org/wiki/Problem_solving
```

仍有兩筆無法匹配到，且有多個Sense配對到相同的Page，程式大約跑一分鐘半

3. Conclusion

本次作業我使用到的資料只有WN的中英對照以及 Wiki的中英、Page ID對照檔案以及Word2Vec Model ，做出來的成果相當吃力，過程中有想過使用wordnet提供的difinition，但不知是否有API能夠直接將英文句子和中文句子做相似度比對，或是英文句子間的相似度比對。而我沒有使用wiki首段的discription(e.sbst.txt)是因為需要額外的parse且檔案過大，讀入+Parse完塞進Dict可能會影響效能很多因此沒有使用。