Project3 for the Biomedical Information Retrieval Course

Q56091079 李昱玟

- ## Environment
  - macOS
  - Python3
  - Flask
  - nltk
  - Bootstrap5
  - Torch 1.10.0
- ## Github
  - https://github.com/yyyyuwen/BIR-Course/tree/Project3
- ## 功能
  - 前次作業的功能另外做擴充
  - Word to Vector

Word2Vec是以詞向量的方式來表示語義，如果語義上有相似的單字，則在空間上距離也會很近，而 **Embedding**是一種將單字從原先的空間映射到新的多維空間上，也就是把原先詞所在空間嵌入到一個新的空間中。我選擇使用**Skip Gram model**來預測相關聯單字，並在最後利用**PCA (Principal Components Analysis)**做資料的視覺化顯示其關聯性。

Model Architecture:

```
SkipGram_Model(
  (embeddings): Embedding(14086, 600, max_norm=1)
  (linear): Linear(in_features=600, out_features=14086, bias=True)
)
# Input Layer : 1 x 14,086
# Hidden Layer : 14,000 x 600
# Output Layer : 600 x 14,086
```

  - Data pre-Processing

1. 讀檔

讀取4000篇.xml，取Title、Label、AbstractText

2. 將文章分段轉成Sentences，取Stopword

3. 將句子切成單字

4. Lemmatizer

首先先將各個單字做詞性標註，最後再將字還原回去。

5. 建立詞彙表

將各個單字建立詞彙表，並單獨標示編號。

6. 建立pair

將詞彙表的編號建立成pair，`window_size = 2`