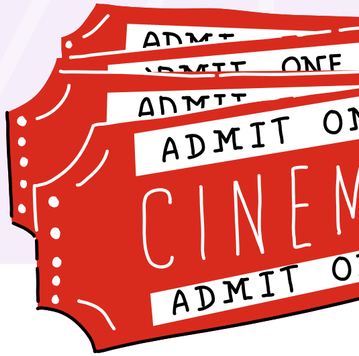


I like to movie it , movie it

- 110590450 歐佳昀
- 110590452 莊于潔

Movie search system optimized with named entity recognition (NER)

motivation

透過上課所學，了解到NLP不同的運作模式。

然而，除了理論知識，我們也想嘗試看看nlp如何在實際應用中運作，以及研究其中的差異

故我們選將ner與搜尋作結合，並且，嘗試建構電影搜尋、推薦的系統

Content

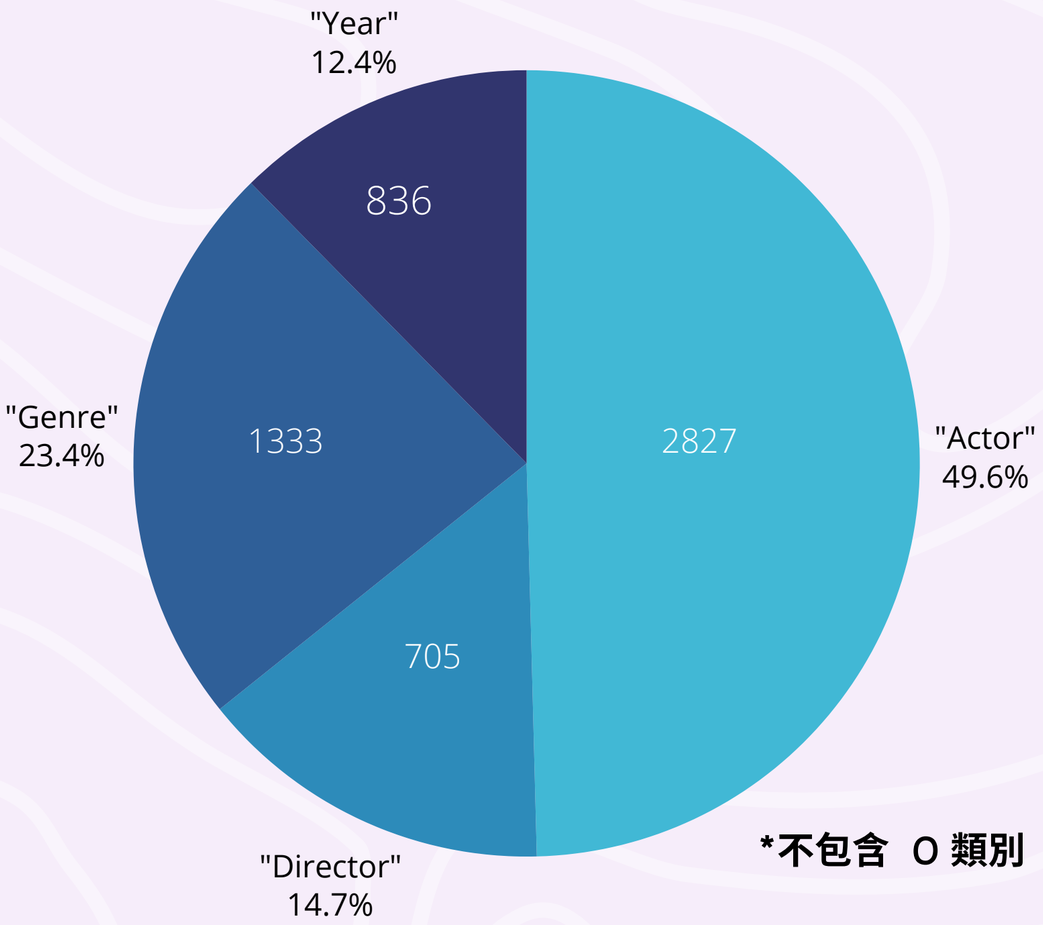
1. 分析NER 資料集
(Spark NLP - Hugging face dataset)

2. 分析movie資料集
(2022 - STR)
(Kaggle - IMDB Movies Dataset)

3. 嘗試實現搜尋引擎
(NER - with search engine)

4. 以Flask 框架配合實作
(Flask - python)

1. 分析 NER 資料集



Dataset : ttxy/ner_movie

去除資料量過少、與IMDB不匹配之類別

以及經綜合考量，Title類別不存在於Hugging face，故系統並不支援直接對Title進行查找，也許可以是未來更新目標

1. 分析 NER 資料集

spaCy

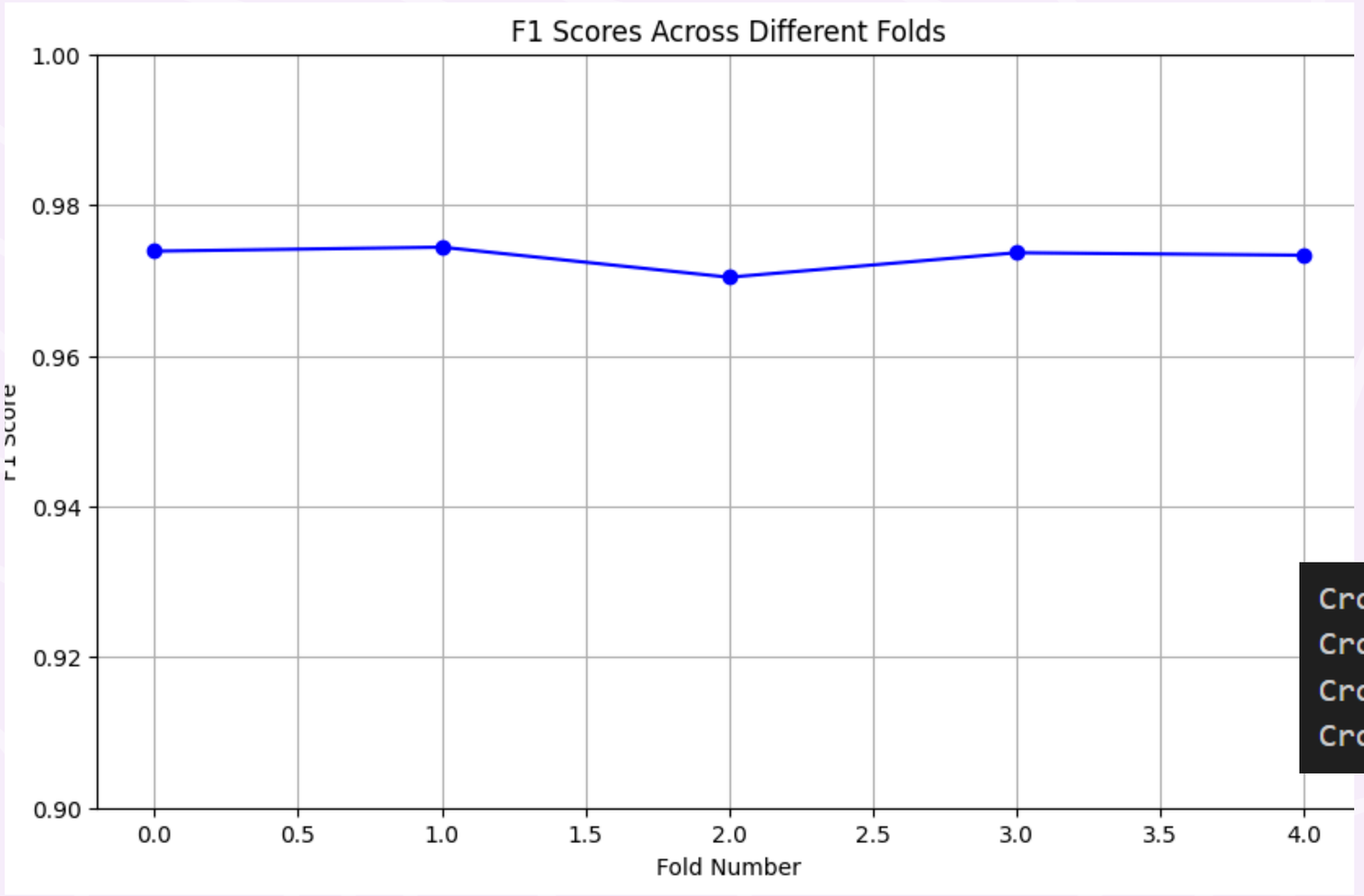
CRF

NLTK

```
==========================================  ==========================================  ==========================================

spaCy :                                     *CRF :                                      NLTK :
Accuracy: 0.781                             Accuracy: 0.974                             Accuracy:0.952
Weighted Average Precision: 0.883           Weighted Average Precision: 0.973           Weighted Average Precision: 0.956
Weighted Average Recall: 0.858              Weighted Average Recall: 0.974              Weighted Average Recall: 0.952
Weighted Average F1 Score: 0.870            Weighted Average F1 Score: 0.973            Weighted Average F1 Score: 0.952

==========================================  ==========================================  ==========================================
```

性能極好
須檢測是否具過擬和

1. 分析 NER 資料集



CRF

為避免過擬和，進行五折交叉，並依序輸出單次判斷結果，**並無過擬和問題**

```
Cross-validated Precision:  0.9729162486235718
Cross-validated Recall:  0.9735888537274064
Cross-validated F1 Score:  0.9731247206580292
Cross-validated Accuracy:  0.9735888537274064
```

1. 分析 NER 資料集

EX. Gone Girl', a 2014 mystery thriller directed by David Fincher and starring Ben Affleck, captivates audiences with its plot twists.

'Gone	Girl',	a	2014	mystery	thriller	directed	by	David	Fincher
			B-Year	B-Genre	I-Genre			B-Director	I-Director

and	starring	Ben	Affleck,	captivates	audiences	with	its	plot	twists.
		B-Actor	I-Actor						

2. 分析movie資料集

IMDB Top 250 Movies Dataset

IMDB 是最大的電影和電視節目線上資料庫之一，提供有關電影的全面信息，包括來自其龐大用戶群的評分和評論。 IMDB 評級被廣泛用作電影受歡迎程度和成功的基準。

該資料集包含截至 2021 年 IMDB 上評分最高的 250 部電影，提供了近期最受歡迎和評分最高的電影的快照。透過分析該資料集，可以深入了解電影產業，例如電影收視率、流行類型的趨勢。

2. 分析movie資料集

Column Description

- Title
- Year
- Rated
- Released
- Runtime
- Genre
- Director
- Writer
- Actors
- Plot
- Language
- Country
- Awards
- Poster
- Ratings.Source
- Ratings.Value
- Metascore
- imdbRating
- imdbVotes
- imdbID
- Type
- tomatoMeter
- tomatoImage
- tomatoRating
- tomatoReviews
- tomatoFresh
- tomatoRotten
- tomatoConsensus
- tomatoUserMeter
- tomatoUserRating
- tomatoUserReviews
- tomatoURL
- DVD
- BoxOffice
- Production
- Website
- Response

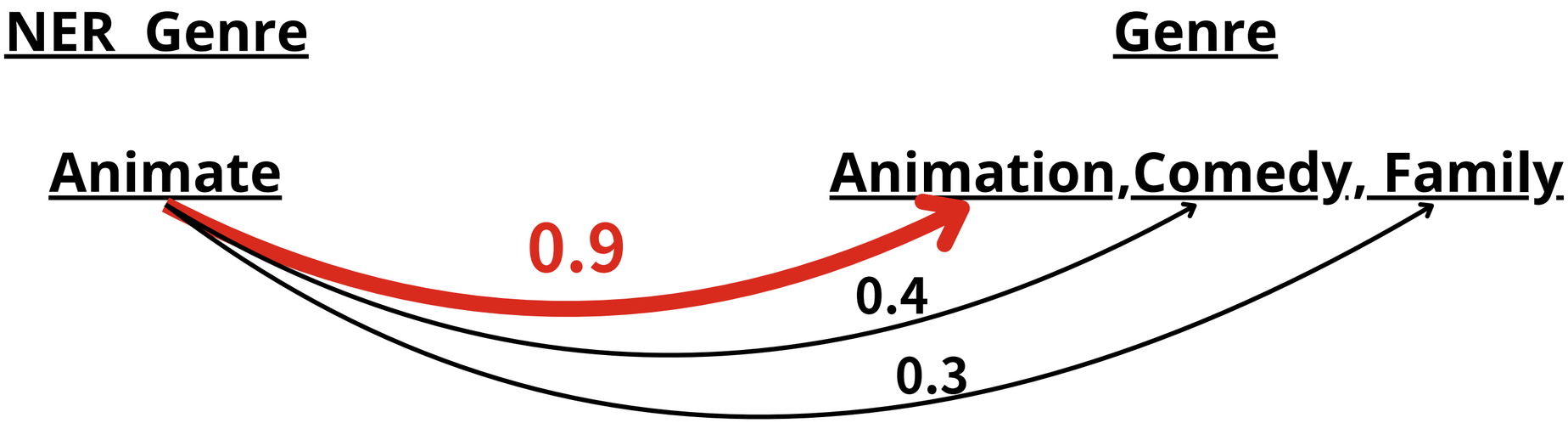
2. 分析movie資料集

對於搜尋系統 ，想完成的方向

1. **應用句子的實體類別，根據預測出的實體類別找出與該預測類別中相似度最高的結果**
2. **如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析**

2. 分析movie資料集

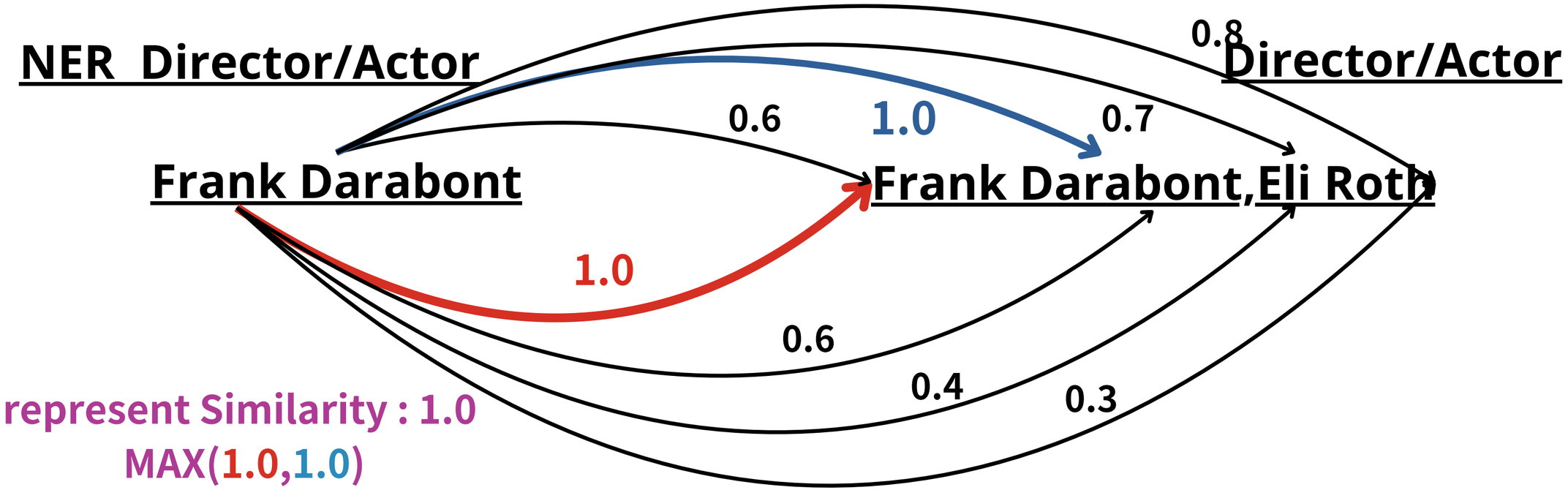
- 應用句子的實體類別，根據預測出的實體類別找出與該預測類別中相似度最高的結果，EX ：



represent Similarity : 0.9

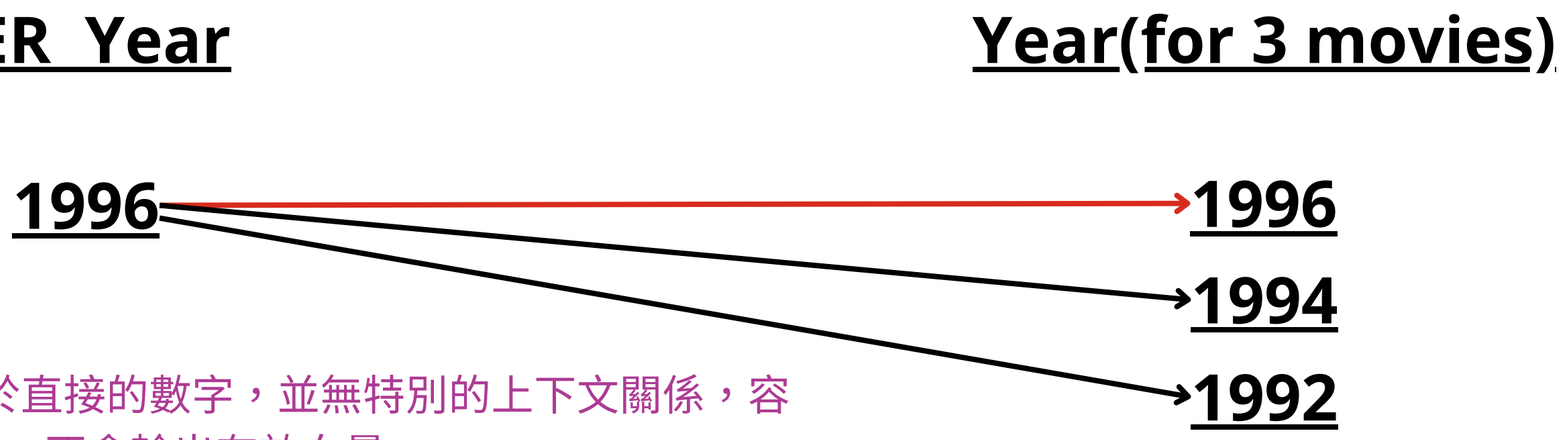
2. 分析movie資料集

- 應用句子的實體類別，根據預測出的實體類別找出與該預測類別中相似度最高的結果，EX ：



2. 分析movie資料集

- **應用句子的實體類別，根據預測出的實體類別找出與該預測類別中相似度最高的結果，EX ：**



然而，年份屬於直接的數字，並無特別的上下文關係，容易導致word2vec不會輸出有效向量
且使用上，使用者大都是希望直接搜尋該年份電影，故選擇直接查找年份是否相同的電影

2. 分析movie資料集

- **如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析**

如何判斷兩個句子的相關性 ？

根據上課所學習到的方法

Word embedding?

TF-IDF?

combine two method?

2. 分析movie資料集

- **如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析**

如何判斷兩個句子的相關性 ？

論文連結　文章連結

根據上課所學習到的方法

論文內容：TF-IDF 結合其他模型作句子相似度計算中的權重因子:
文章內容 : 獲取每個詞對應的詞向量，然後將所有的詞向量相加求平均，得到句子向量，最後計算兩個句子向量的餘弦值（餘弦相似度）。

2. 分析movie資料集

- 如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析

STR-2022

資料集連結

文章連結

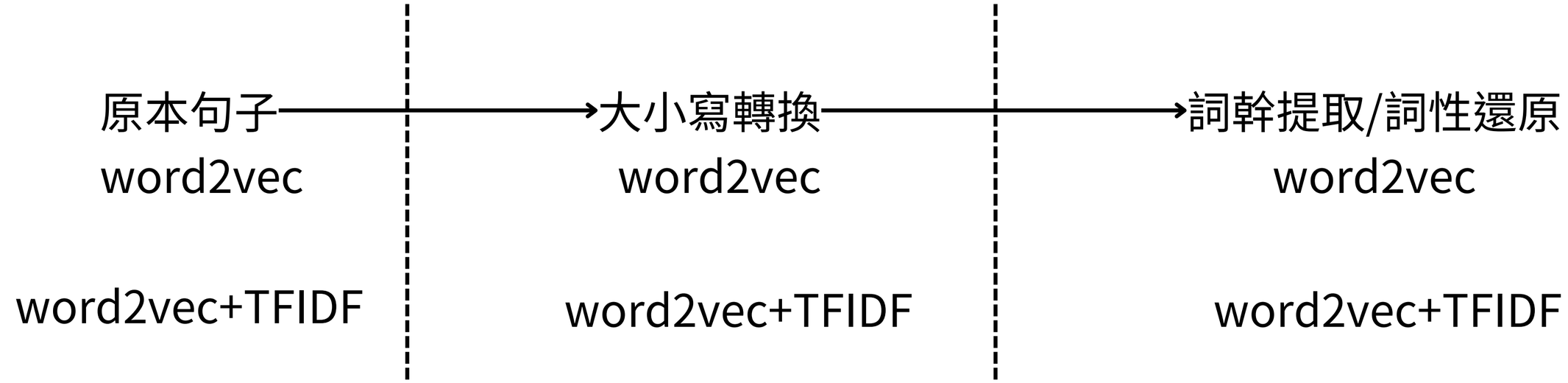
該數據集包含5500對英文句子配對，這些句子配對根據相關性等級進行評分和排名，範圍從0（最不相關）到1（最相關）。

是一工標註的句子與句子語義相關性數據集。包括5500對英文句子配對的細緻相關性評分，這些句子來自多樣化的來源，因此也具有多樣化的句子結構、不同程度的詞彙重疊和不同的形式風格。

2. 分析movie資料集

- **如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析**

多種不同資料預處理方式



2. 分析movie資料集

*Spearman 相關係數 ：0.3-0.5 弱相關 /0.5-0.7中等相關 0.7相關性高

原本句子	大小寫轉換	詞幹提取/詞性還原
word2vec	**word2vec**	**word2vec**
statistic=0.5215107379648181, pvalue=0.0	**statistic=0.598884800148964 5 pvalue: 0.0**	statistic= 0.596620593290864 9 pvalue: 0.0
word2vec+TFIDF	**word2vec+TFIDF**	**word2vec+TFIDF**
statistic=0.4321769190441943 pvalue=4.0480567737406395e-249	statistic=0.560636600581575, pvalue=0.0	statistic=0.5432093808323544, pvalue=0.0

2. 分析movie資料集

- **如同多數網站，在搜尋之後會顯示出像似性質的電影，故希望基於情節進行相似分析**

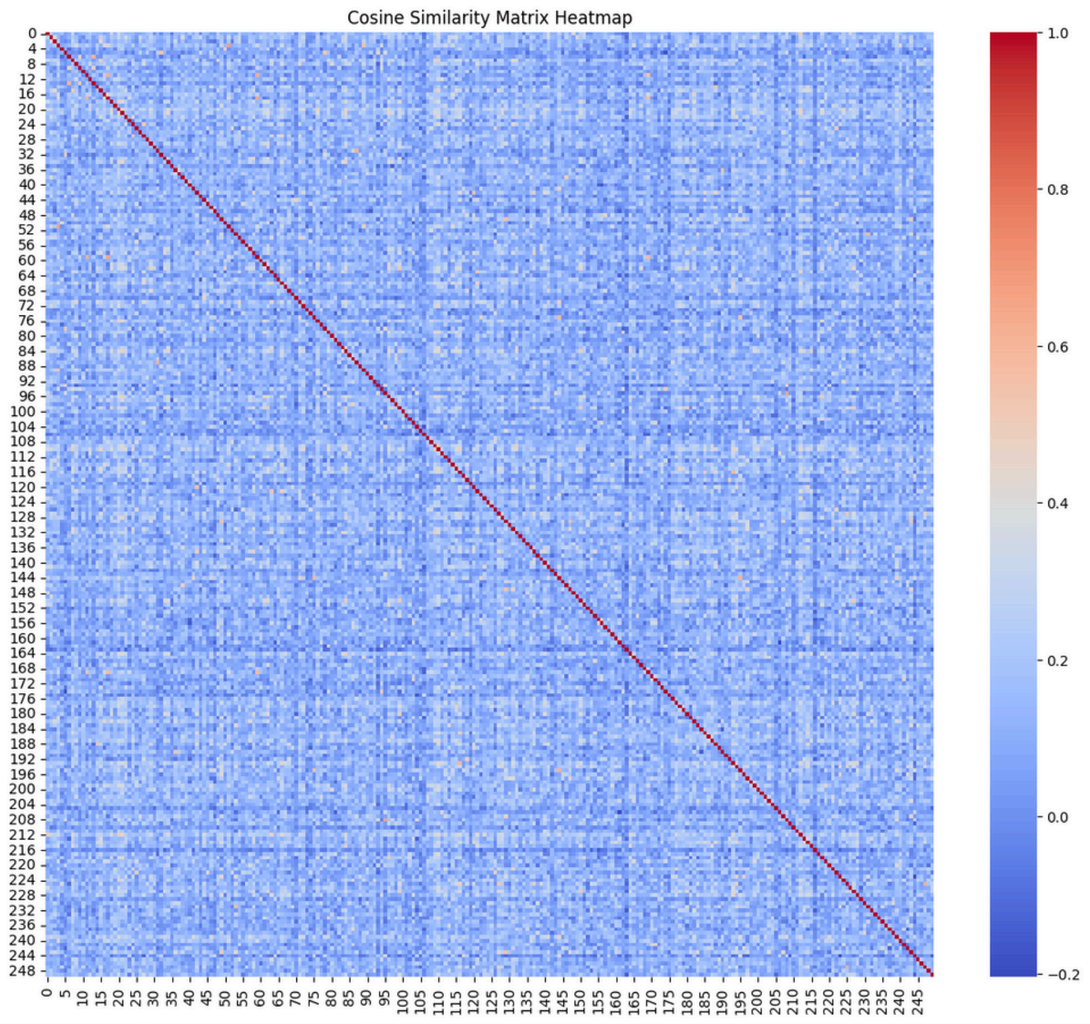
GPT-2
statistic= 0.26945093441564893 pvalue=4.098929343565402e-92
Bert/sentence-transformers/all-MiniLM-L6-v2
statistic= 0.8080134762506194 **pvalue=0.0**

*新增兩種高階模型

*資料處理：大小寫轉換、去除標點符號

不愧是BERT

2. 分析movie資料集

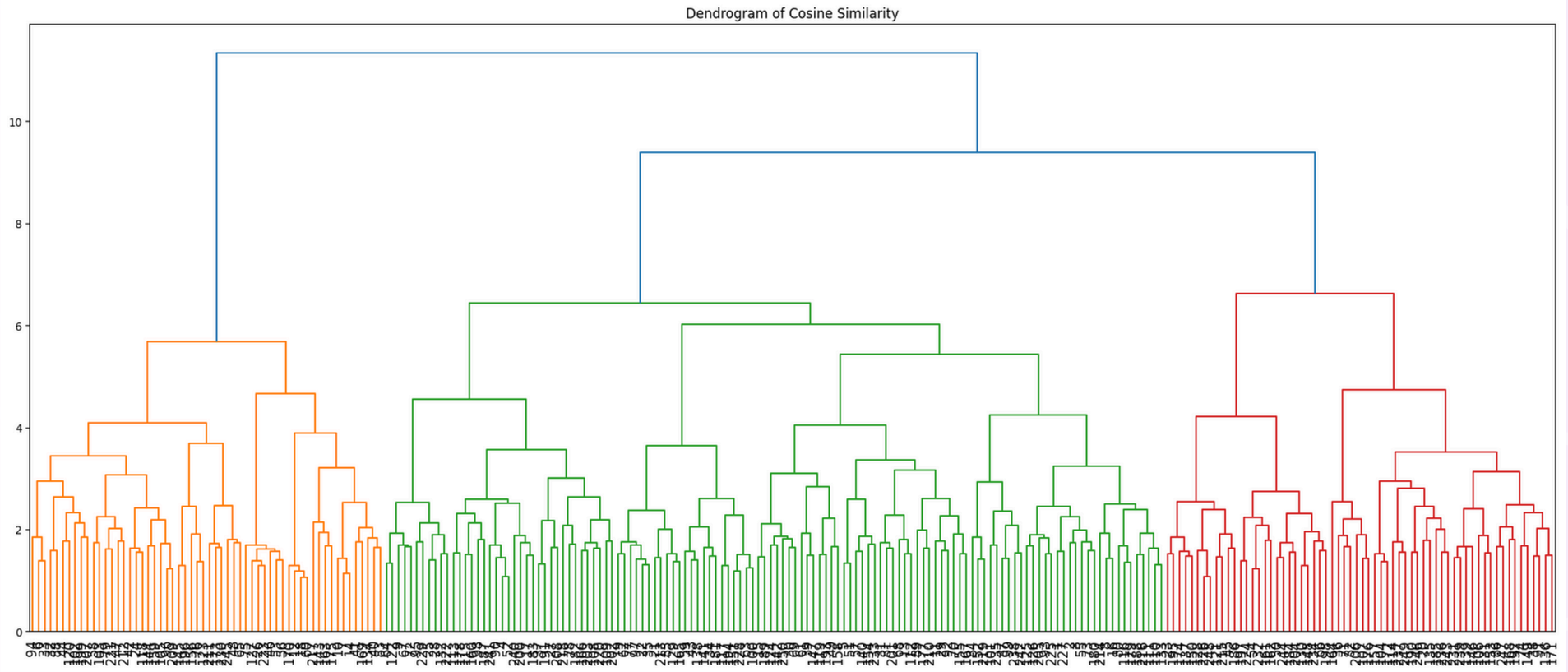


透過Bert/sentence-transformers/all-MiniLM-L6-v2

計算Movie 資料集 中 Plot 相似度，產生距離矩陣熱力圖

可以直觀感受大致上的相似度情況介於 0 ~ 0.5左右

2. 分析movie資料集



透過Elbow/
Silhouette
大致區分為2-3群左右

3. 分析movie資料集

IMDB Top 250 Movies Dataset

IMDB 是最大的電影和電視節目線上資料庫之一，提供有關電影的全面信息，包括來自其龐大用戶群的評分和評論。 IMDB 評級被廣泛用作電影受歡迎程度和成功的基準。

該資料集包含截至 2021 年 IMDB 上評分最高的 250 部電影，提供了近期最受歡迎和評分最高的電影的快照。透過分析該資料集，可以深入了解電影產業，例如電影收視率、流行類型的趨勢。

3. 嘗試實現搜尋引擎

資料查詢對象: IMDB Top 250 Movies Dataset

查詢目標 : Actors、Directors、Genre 、Year

3. 嘗試實現搜尋引擎


CRF模型導入
Bert/sentence-transformers/all-MiniLM-L6-v2
計算之相似度矩陣導入
word2vec(pre-train) 導入
輸入字串
word2vec計算
與該實體對象類別相似度
輸出結果
個電影與其簡介相似串接

3. 嘗試實現搜尋引擎

- 各類別資料處理

```
df['Genre_lower'] = df['Genre'].apply(lambda x: [i.strip().lower() for i in x.split(',')])
df['Actors_lower'] = df['Actors'].apply(lambda x: [i.strip().lower() for i in x.split(',')][:3])
df['Director_lower'] = df['Director'].apply(lambda x: [i.strip().lower() for i in x.split(',')])
df['Year_str'] = df['Year'].astype(str)

# 檢查是否正確應用
print(df[['Genre_lower', 'Actors_lower', 'Director_lower', 'Year_str']].head(3))
```

- NER 後之資料提取

去除 "O"

+ Code + Markdown

```
for i in predicted_labels:
    print(i)

print("======================")
non_o_labels = [(word, label) for word, label in predicted_labels if label != 'O']

for word, label in non_o_labels:
    print(f"({word}, {label})")
```

整理 B-I 關係

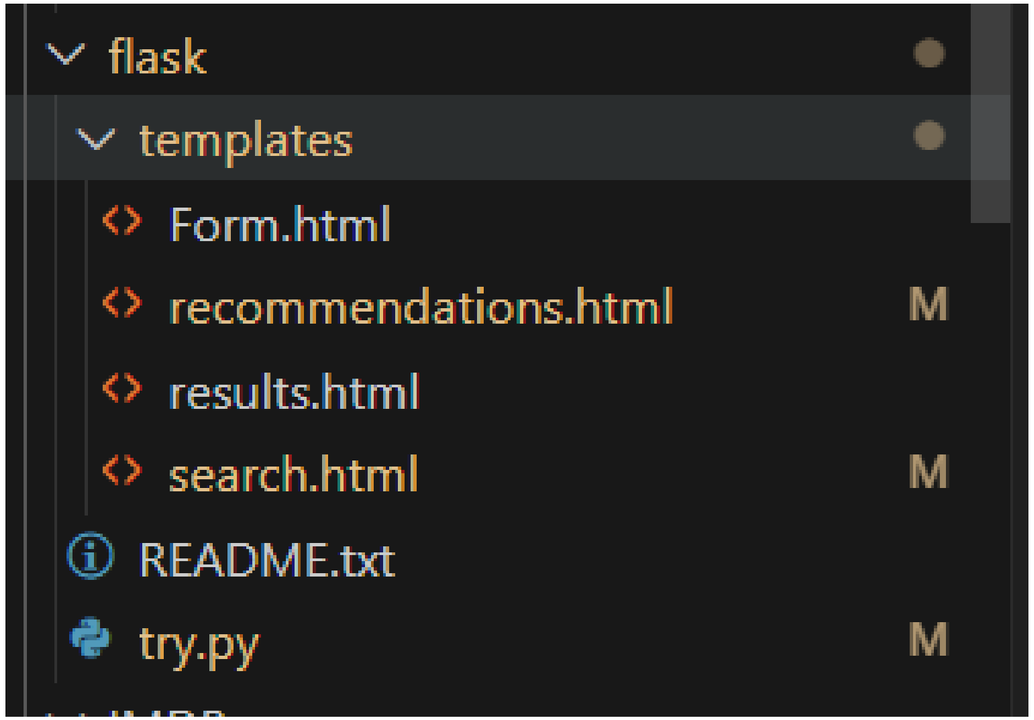
```
def extract_entities(predicted_labels, entity_type):
    entities = []
    current_entity = []

    for word, label in predicted_labels:
        if label == f"B-{entity_type}":
            if current_entity:
                entities.append(" ".join(current_entity))
                current_entity = []
            current_entity.append(word)
        elif label == f"I-{entity_type}" and current_entity:
            current_entity.append(word)
        elif label == "O" and current_entity:
            entities.append(" ".join(current_entity))
            current_entity = []

    if current_entity:
        entities.append(" ".join(current_entity))

    return entities
```

- **B-NP**：名詞短語的開頭
- **I-NP**：名詞短語的中間
- **O**：不是名詞短語

ex.
(Amy, B-Actor)
(Adams, I-Actor)

ex.
Actor: Amy Adams

4. 以Flask 框架配合實作



Demo

I like to movie it, movie it.

Search for Movies

a 2014 mystery thriller directed by David Fincher and starring Ben Affleck, captivates audiences with its plot twists.

submit

輸入搜尋語句

全部電影列表

ALL Movies

Slumdog Millionaire directed by Danny Boyle, Loveleen Tandan

Notorious directed by Alfred Hitchcock

The Killing directed by Stanley Kubrick

The Hustler directed by Robert Rossen

Sin City directed by Frank Miller, Robert Rodriguez, Quentin Tarantino

Arsenic and Old Lace directed by Frank Capra

I like to movie it, movie it.

顯示處理後的語句

Search Results for a 2014 mystery thriller directed by David Fincher and starring Ben Affleck captivates audiences with its plot twists

Back to Search

NER Results : **Actors** ['Ben Affleck'] **Director** ['David Fincher'] **Years** ['2014'] **Genres** ['mystery thriller']

經NER顯示的實體命名

Movie Title : Anatomy of a Murder
Director : Otto Preminger
Actors : James Stewart, Lee Remick, Ben Gazzara, Arthur O'Connell
Genre : Crime, Drama, Mystery
Year : 1959
Plot : In a murder trial, the defendant says he suffered temporary insanity after the victim raped his wife. What is the truth, and will he win his case?

See More

Movie Title : Gandhi
Director : Richard Attenborough
Actors : Ben Kingsley, Candice Bergen, Edward Fox, John Gielgud
Genre : Biography, Drama, History
Year : 1982
Plot : Gandhi's character is fully explained as a man of nonviolence. Through his patience, he is able to drive the British out of the subcontinent. And the

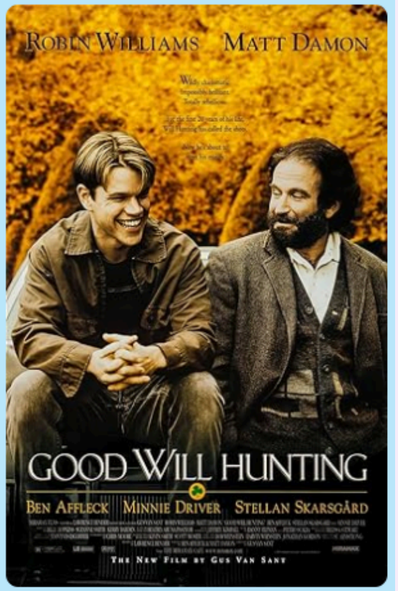
點擊進入

I like to movie it, movie it.

Good Will Hunting

Director: Gus Van Sant
Actors: Matt Damon, Ben Affleck, Stellan Skarsgård, John Mighton
Genre: Drama
Year: 1997
Rating: 8.3
Rated: R
Released: 09 Jan 1998
Runtime: 126 min
Language: English
Country: USA
Awards: Won 2 Oscars. Another 22 wins & 55 nominations.
Website:http://www.miramax.com/movie/good-will-hunting/
Plot: Will Hunting, a janitor at M.I.T., has a gift for mathematics, but needs help from a psychologist to find direction in his life.

詳細資訊



Recommendations - Base on Plot

A Beautiful Mind - Similary Score: 0.3454 - Genre: Biography, Drama
A Clockwork Orange - Similary Score: 0.342631 - Genre: Crime, Drama, Sci-Fi
Fight Club - Similary Score: 0.320579 - Genre: Drama
A Christmas Story - Similary Score: 0.312696 - Genre: Comedy, Family
Short Term 12 - Similary Score: 0.308563 - Genre: Drama

顯示相關電影

Back to Search

DEMO 影片

Colusion

分工表

姓名	學號	內容	配比
歐佳昀	110590450	模型訓練、專案架構、專案研究、前端實作	60%
莊于潔	110590452	專案發想、前端架構、模型參數調整	40%

END