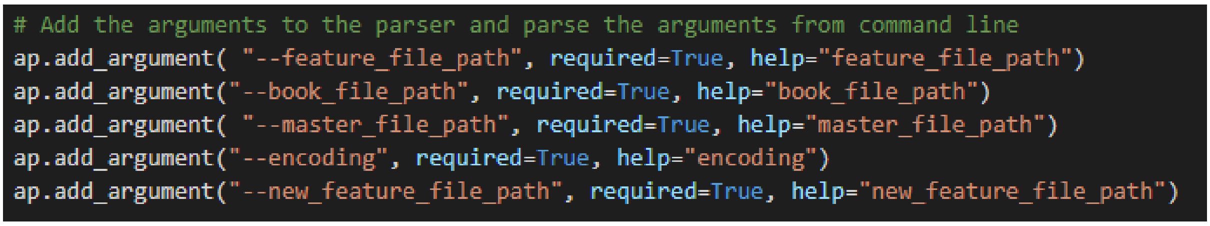
GENRE EXTRACTION:

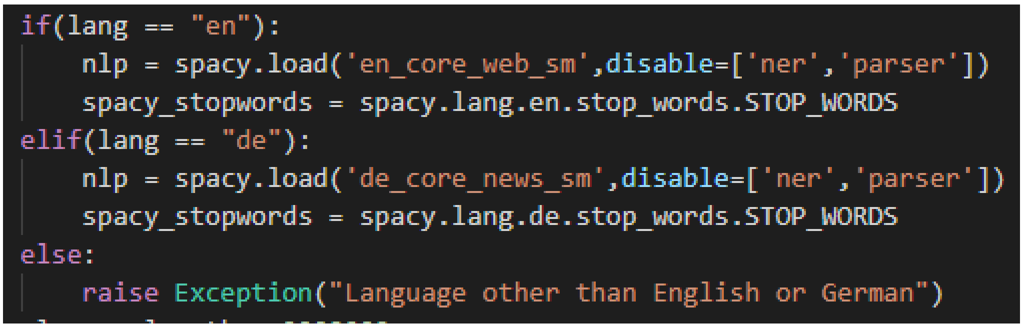
METHOD:

1) Call the python program from the java program via a batch script and by passing correct parameters. The parameters are given below.



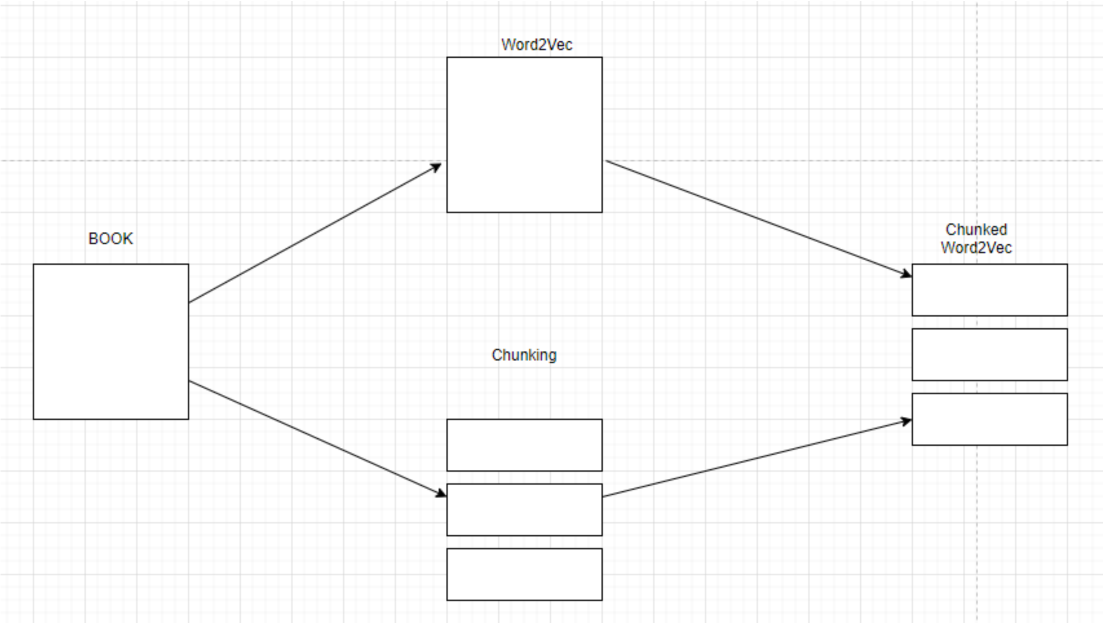
```
# Add the arguments to the parser and parse the arguments from command line
ap.add_argument( "--feature_file_path", required=True, help="feature_file_path")
ap.add_argument("--book_file_path", required=True, help="book_file_path")
ap.add_argument( "--master_file_path", required=True, help="master_file_path")
ap.add_argument("--encoding", required=True, help="encoding")
ap.add_argument("--new_feature_file_path", required=True, help="new_feature_file_path")
```

2) The next step is to pre-process the obtained HTML files. Pre-processing involves retrieving only the text from HTML page and then doing tokenisation and stop-word removal. Spacy packages were used for these purposes.

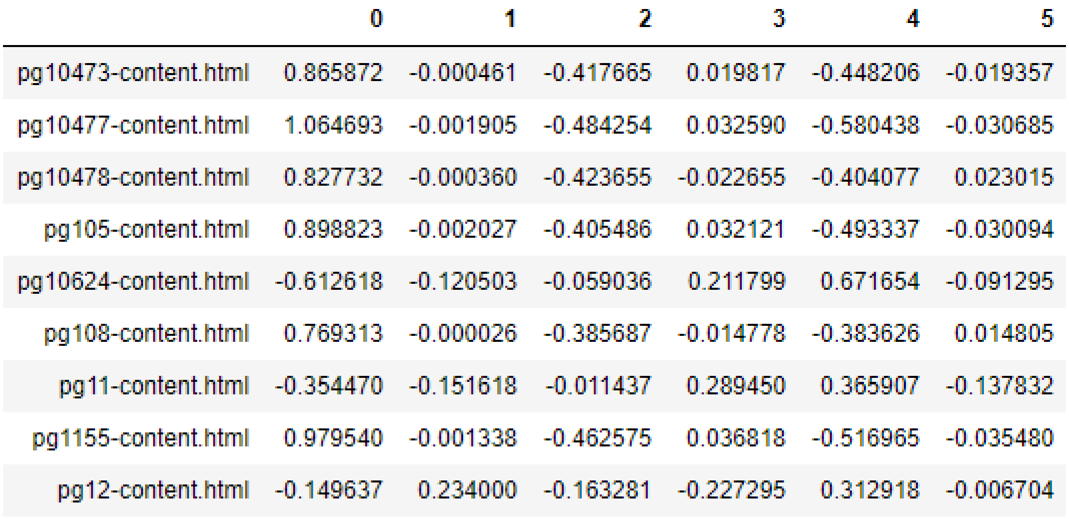


```
if(lang == "en"):
    nlp = spacy.load('en_core_web_sm',disable=['ner','parser'])
    spacy_stopwords = spacy.lang.en.stop_words.STOP_WORDS
elif(lang == "de"):
    nlp = spacy.load('de_core_news_sm',disable=['ner','parser'])
    spacy_stopwords = spacy.lang.de.stop_words.STOP_WORDS
else:
    raise Exception("Language other than English or German")
```

3) After the pre-processing is done, the book undergoes two different processes. First the book is sent to Word2Vec to be converted into word vectors of length 100. Different vector lengths were experimented with, and size of 100 was deemed to be optimal. The same thing was done with window sizes as well, and window size of 50 was deemed to be optimal. The book is also sent to the chunking process, which chunks the books into equal parts. The optimal chunk parts were deemed to be 3. All the word vectors present in a particular chunk is averaged to get a chunk vector of size 100 per chunk.

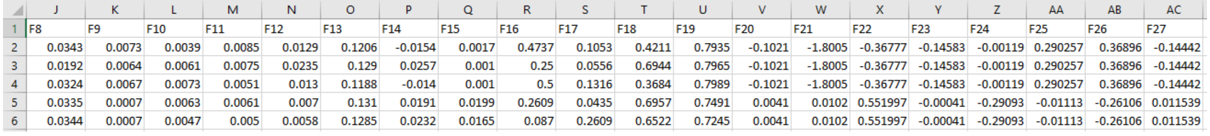


4) Now we have a vector of length 100 for each chunk, and we use PCA to reduce it from vector of length 100 to length 2. So vectors of length 2 per chunk, and since we have 3 chunks we end up with a vector of length of 6.



	0	1	2	3	4	5
pg10473-content.html	0.865872	-0.000461	-0.417665	0.019817	-0.448206	-0.019357
pg10477-content.html	1.064693	-0.001905	-0.484254	0.032590	-0.580438	-0.030685
pg10478-content.html	0.827732	-0.000360	-0.423655	-0.022655	-0.404077	0.023015
pg105-content.html	0.898823	-0.002027	-0.405486	0.032121	-0.493337	-0.030094
pg10624-content.html	-0.612618	-0.120503	-0.059036	0.211799	0.671654	-0.091295
pg108-content.html	0.769313	-0.000026	-0.385687	-0.014778	-0.383626	0.014805
pg11-content.html	-0.354470	-0.151618	-0.011437	0.289450	0.365907	-0.137832
pg1155-content.html	0.979540	-0.001338	-0.462575	0.036818	-0.516965	-0.035480
pg12-content.html	-0.149637	0.234000	-0.163281	-0.227295	0.312918	-0.006704

5) The above vectors are the result of the python program. This result is combined with the Feature Vectors excel file for further process. Features F22-F27 is the result of genre extraction.



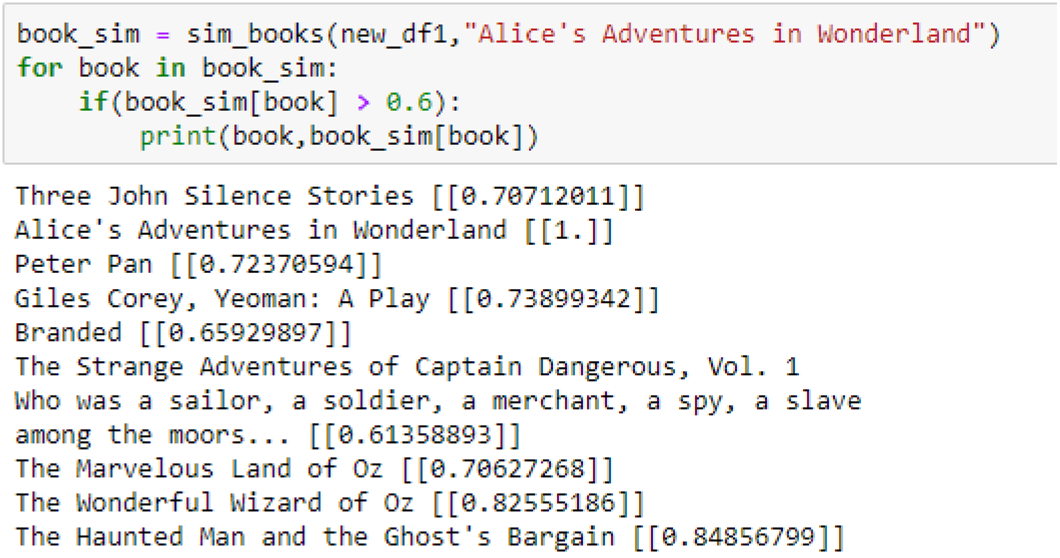
	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	F26	F27
2	0.0343	0.0073	0.0039	0.0085	0.0129	0.1206	-0.0154	0.0017	0.4737	0.1053	0.4211	0.7935	-0.1021	-1.8005	-0.36777	-0.14583	-0.00119	0.290257	0.36896	-0.14442
3	0.0192	0.0064	0.0061	0.0075	0.0235	0.129	0.0257	0.001	0.25	0.0556	0.6944	0.7965	-0.1021	-1.8005	-0.36777	-0.14583	-0.00119	0.290257	0.36896	-0.14442
4	0.0324	0.0067	0.0073	0.0051	0.013	0.1188	-0.014	0.001	0.5	0.1316	0.3684	0.7989	-0.1021	-1.8005	-0.36777	-0.14583	-0.00119	0.290257	0.36896	-0.14442
5	0.0335	0.0007	0.0063	0.0061	0.007	0.131	0.0191	0.0199	0.2609	0.0435	0.6957	0.7491	0.0041	0.0102	0.551997	-0.00041	-0.29093	-0.01113	-0.26106	0.011539
6	0.0344	0.0007	0.0047	0.005	0.0058	0.1285	0.0232	0.0165	0.087	0.2609	0.6522	0.7245	0.0041	0.0102	0.551997	-0.00041	-0.29093	-0.01113	-0.26106	0.011539

We have also performed an experiment on 50 books to check weather the vectors of similar genres has high L2 similarity.

RESULTS OF EXPERIMENT:

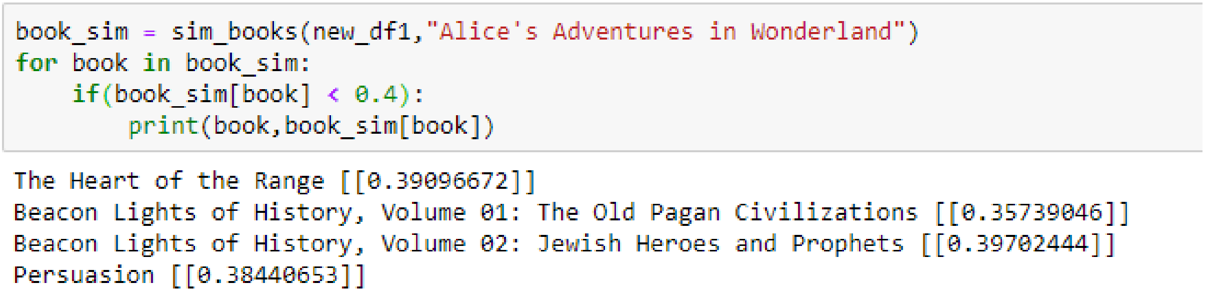
FANTASY:

QUERY BOOK: 11. ALICE IN WONDERLAND



```
book_sim = sim_books(new_df1,"Alice's Adventures in Wonderland")
for book in book_sim:
    if(book_sim[book] > 0.6):
        print(book,book_sim[book])
```

```
Three John Silence Stories [[0.70712011]]
Alice's Adventures in Wonderland [[1.]]
Peter Pan [[0.72370594]]
Giles Corey, Yeoman: A Play [[0.73899342]]
Branded [[0.65929897]]
The Strange Adventures of Captain Dangerous, Vol. 1
Who was a sailor, a soldier, a merchant, a spy, a slave
among the moors... [[0.61358893]]
The Marvelous Land of Oz [[0.70627268]]
The Wonderful Wizard of Oz [[0.82555186]]
The Haunted Man and the Ghost's Bargain [[0.84856799]]
```

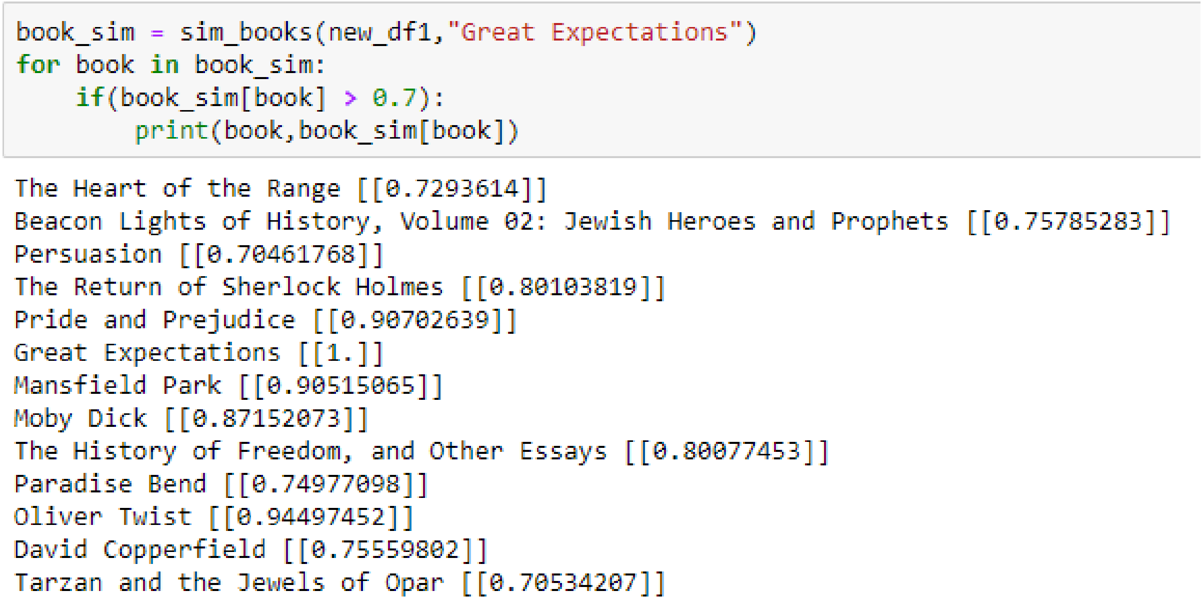


```
book_sim = sim_books(new_df1,"Alice's Adventures in Wonderland")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
The Heart of the Range [[0.39096672]]
Beacon Lights of History, Volume 01: The Old Pagan Civilizations [[0.35739046]]
Beacon Lights of History, Volume 02: Jewish Heroes and Prophets [[0.39702444]]
Persuasion [[0.38440653]]
```

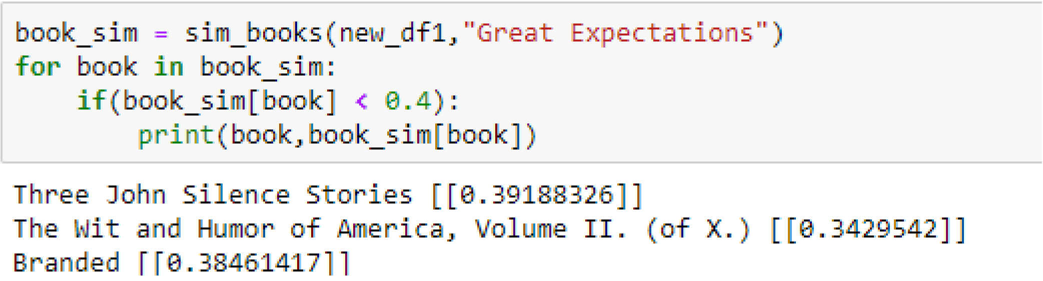
ADVENTURE:

1400. GREAT EXPECTATIONS



```
book_sim = sim_books(new_df1,"Great Expectations")
for book in book_sim:
    if(book_sim[book] > 0.7):
        print(book,book_sim[book])
```

```
The Heart of the Range [[0.7293614]]
Beacon Lights of History, Volume 02: Jewish Heroes and Prophets [[0.75785283]]
Persuasion [[0.70461768]]
The Return of Sherlock Holmes [[0.80103819]]
Pride and Prejudice [[0.90702639]]
Great Expectations [[1.]]
Mansfield Park [[0.90515065]]
Moby Dick [[0.87152073]]
The History of Freedom, and Other Essays [[0.80077453]]
Paradise Bend [[0.74977098]]
Oliver Twist [[0.94497452]]
David Copperfield [[0.75559802]]
Tarzan and the Jewels of Opar [[0.70534207]]
```

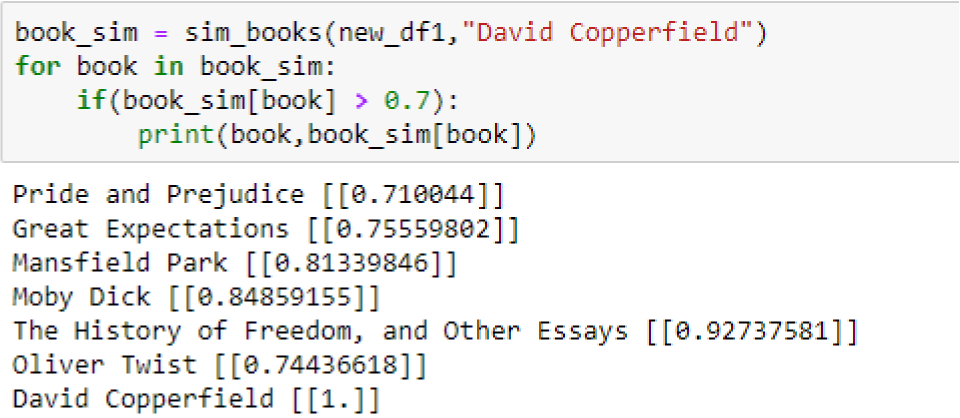


```
book_sim = sim_books(new_df1,"Great Expectations")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
Three John Silence Stories [[0.39188326]]
The Wit and Humor of America, Volume II. (of X.) [[0.3429542]]
Branded [[0.38461417]]
```

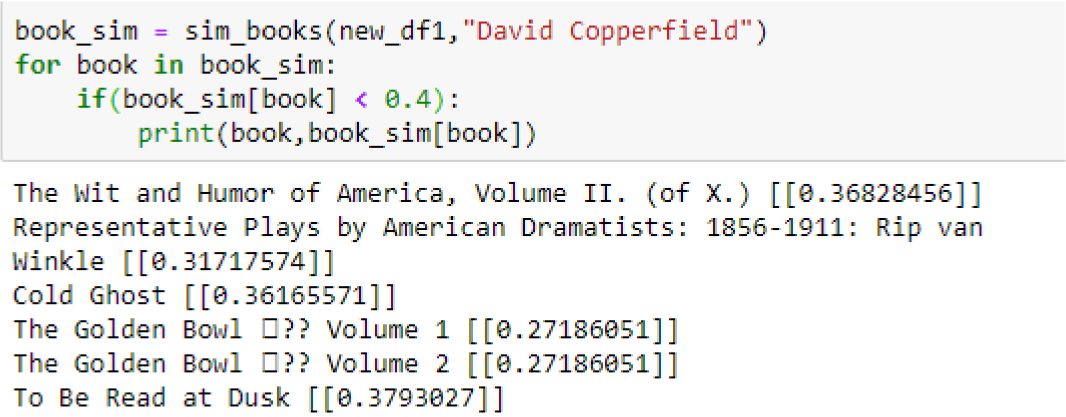
LITERARY:

766. DAVID COPPERFIELD



```
book_sim = sim_books(new_df1,"David Copperfield")
for book in book_sim:
    if(book_sim[book] > 0.7):
        print(book,book_sim[book])
```

```
Pride and Prejudice [[0.710044]]
Great Expectations [[0.75559802]]
Mansfield Park [[0.81339846]]
Moby Dick [[0.84859155]]
The History of Freedom, and Other Essays [[0.92737581]]
Oliver Twist [[0.74436618]]
David Copperfield [[1.]]
```

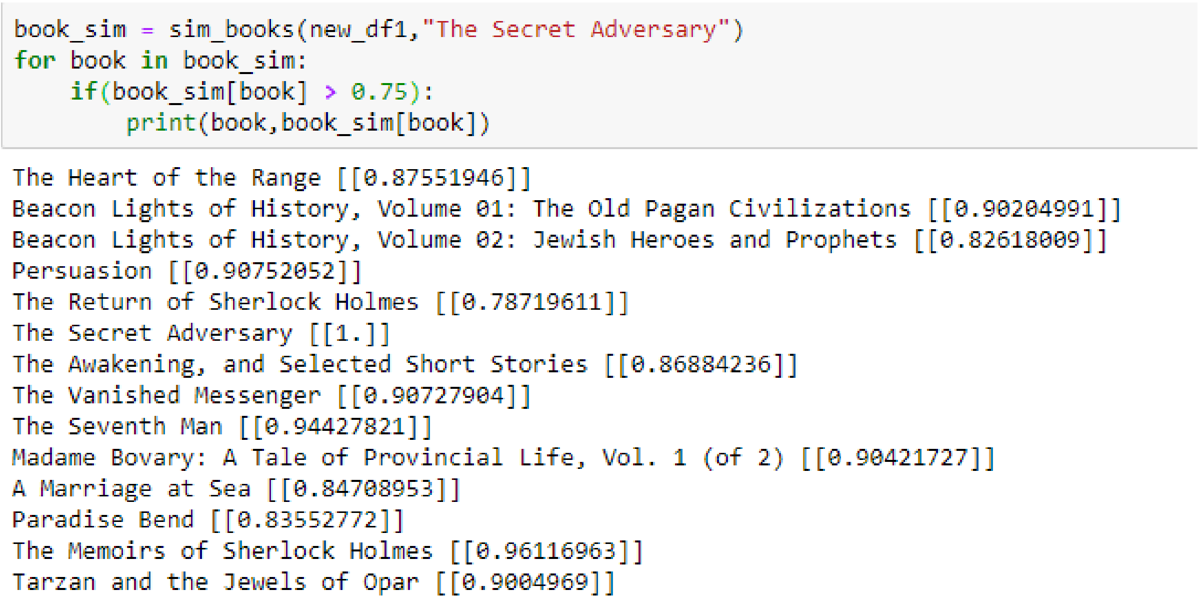


```
book_sim = sim_books(new_df1,"David Copperfield")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
The Wit and Humor of America, Volume II. (of X.) [[0.36828456]]
Representative Plays by American Dramatists: 1856-1911: Rip van
Winkle [[0.31717574]]
Cold Ghost [[0.36165571]]
The Golden Bowl □?? Volume 1 [[0.27186051]]
The Golden Bowl □?? Volume 2 [[0.27186051]]
To Be Read at Dusk [[0.3793027]]
```

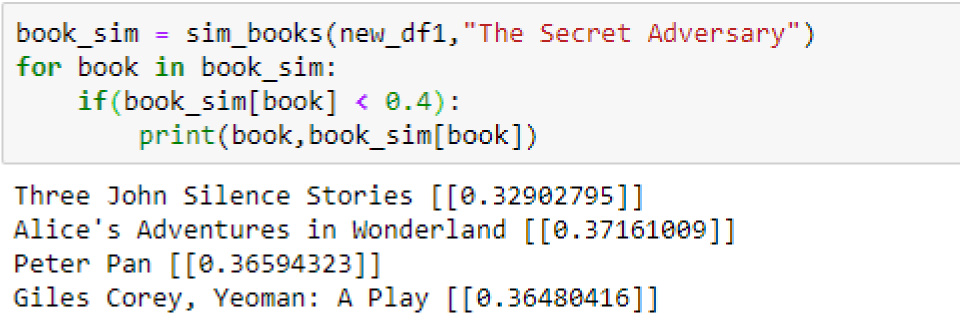
DETECTIVE AND MYSTERY:

1155. The Secret Adversary



```
book_sim = sim_books(new_df1,"The Secret Adversary")
for book in book_sim:
    if(book_sim[book] > 0.75):
        print(book,book_sim[book])
```

```
The Heart of the Range [[0.87551946]]
Beacon Lights of History, Volume 01: The Old Pagan Civilizations [[0.90204991]]
Beacon Lights of History, Volume 02: Jewish Heroes and Prophets [[0.82618009]]
Persuasion [[0.90752052]]
The Return of Sherlock Holmes [[0.78719611]]
The Secret Adversary [[1.]]
The Awakening, and Selected Short Stories [[0.86884236]]
The Vanished Messenger [[0.90727904]]
The Seventh Man [[0.94427821]]
Madame Bovary: A Tale of Provincial Life, Vol. 1 (of 2) [[0.90421727]]
A Marriage at Sea [[0.84708953]]
Paradise Bend [[0.83552772]]
The Memoirs of Sherlock Holmes [[0.96116963]]
Tarzan and the Jewels of Opar [[0.9004969]]
```

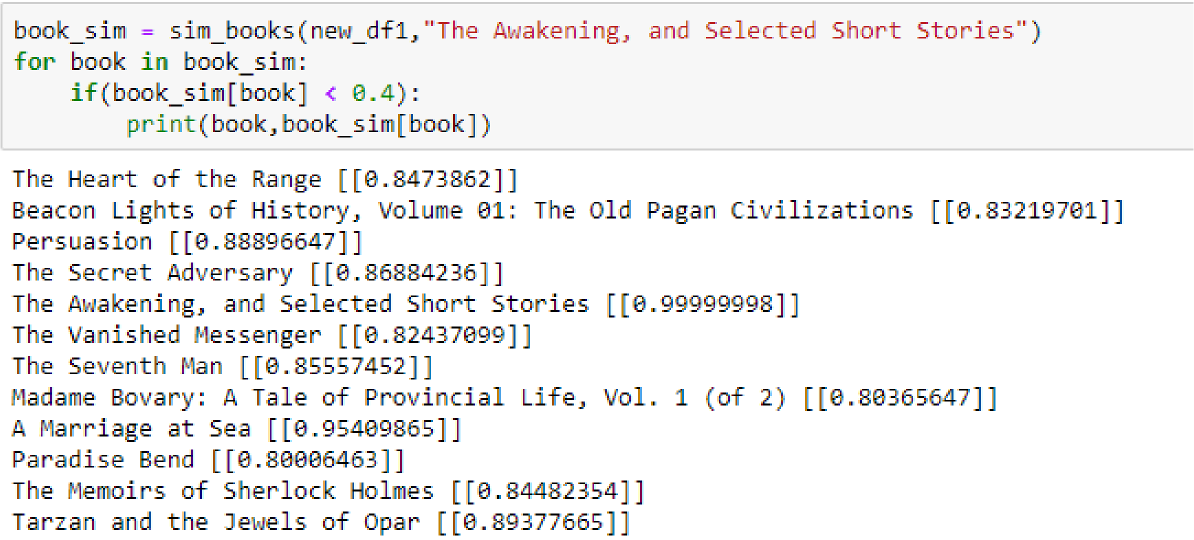


```
book_sim = sim_books(new_df1,"The Secret Adversary")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
Three John Silence Stories [[0.32902795]]
Alice's Adventures in Wonderland [[0.37161009]]
Peter Pan [[0.36594323]]
Giles Corey, Yeoman: A Play [[0.36480416]]
```

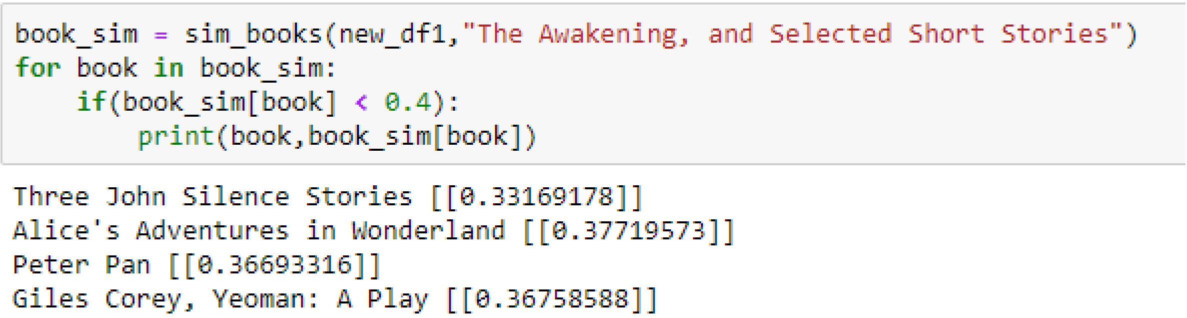
ROMANCE:

160. The Awakening, and Selected Short Stories



```
book_sim = sim_books(new_df1,"The Awakening, and Selected Short Stories")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
The Heart of the Range [[0.8473862]]
Beacon Lights of History, Volume 01: The Old Pagan Civilizations [[0.83219701]]
Persuasion [[0.88896647]]
The Secret Adversary [[0.86884236]]
The Awakening, and Selected Short Stories [[0.99999998]]
The Vanished Messenger [[0.82437099]]
The Seventh Man [[0.85557452]]
Madame Bovary: A Tale of Provincial Life, Vol. 1 (of 2) [[0.80365647]]
A Marriage at Sea [[0.95409865]]
Paradise Bend [[0.80006463]]
The Memoirs of Sherlock Holmes [[0.84482354]]
Tarzan and the Jewels of Opar [[0.89377665]]
```

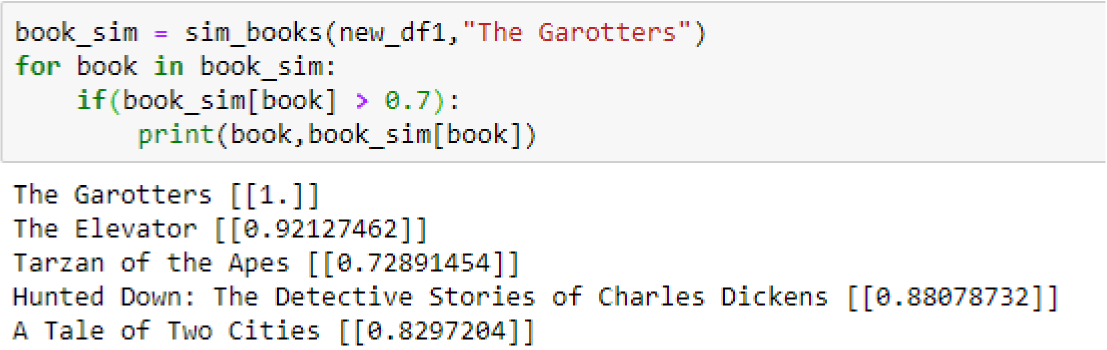


```python
book_sim = sim_books(new_df1,"The Awakening, and Selected Short Stories")
for book in book_sim:
    if(book_sim[book] < 0.4):
        print(book,book_sim[book])
```

```
Three John Silence Stories [[0.33169178]]
Alice's Adventures in Wonderland [[0.37719573]]
Peter Pan [[0.36693316]]
Giles Corey, Yeoman: A Play [[0.36758588]]
```

DRAMA:

3237. The Garrotters

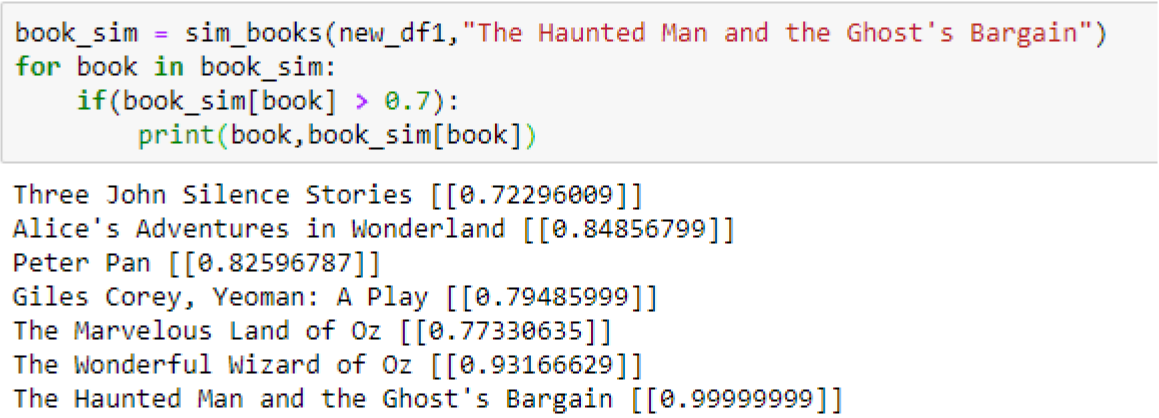


```python
book_sim = sim_books(new_df1,"The Garotters")
for book in book_sim:
    if(book_sim[book] > 0.7):
        print(book,book_sim[book])
```

```
The Garotters [[1.]]
The Elevator [[0.92127462]]
Tarzan of the Apes [[0.72891454]]
Hunted Down: The Detective Stories of Charles Dickens [[0.88078732]]
A Tale of Two Cities [[0.8297204]]
```

HORROR:

644. The Haunted Man and the Ghost's Bargain

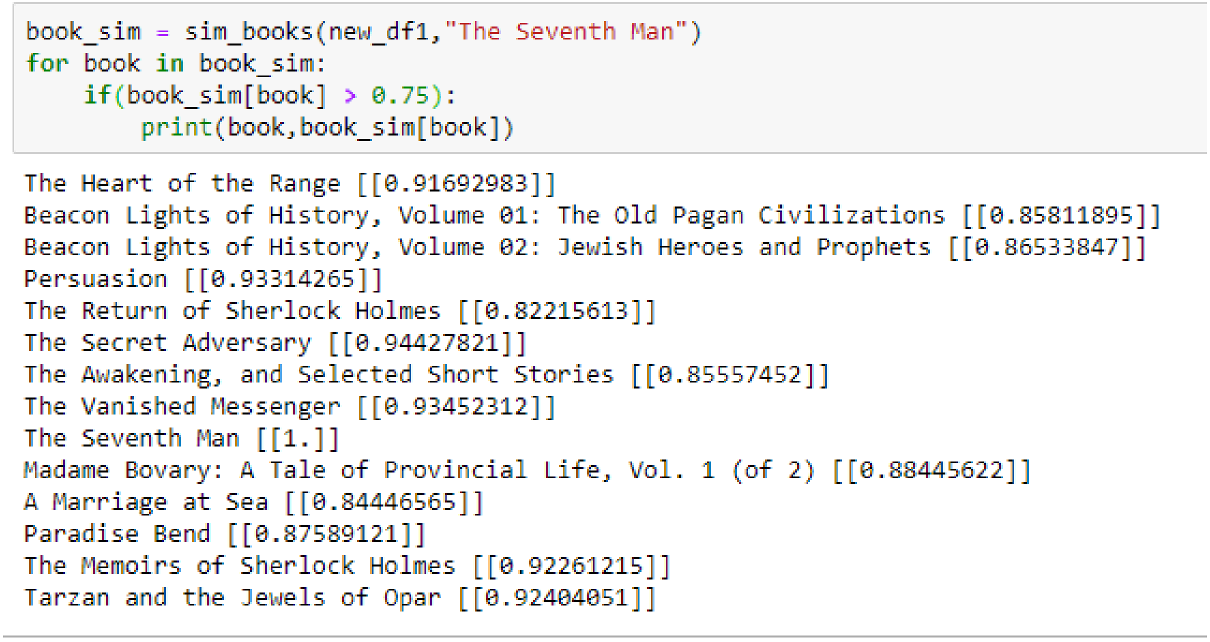


```python
book_sim = sim_books(new_df1,"The Haunted Man and the Ghost's Bargain")
for book in book_sim:
    if(book_sim[book] > 0.7):
        print(book,book_sim[book])
```

```
Three John Silence Stories [[0.72296009]]
Alice's Adventures in Wonderland [[0.84856799]]
Peter Pan [[0.82596787]]
Giles Corey, Yeoman: A Play [[0.79485999]]
The Marvelous Land of Oz [[0.77330635]]
The Wonderful Wizard of Oz [[0.93166629]]
The Haunted Man and the Ghost's Bargain [[0.99999999]]
```

WESTERN STORIES:

1897. The Seventh Man



```
book_sim = sim_books(new_df1,"The Seventh Man")
for book in book_sim:
    if(book_sim[book] > 0.75):
        print(book,book_sim[book])
```

```
The Heart of the Range [[0.91692983]]
Beacon Lights of History, Volume 01: The Old Pagan Civilizations [[0.85811895]]
Beacon Lights of History, Volume 02: Jewish Heroes and Prophets [[0.86533847]]
Persuasion [[0.93314265]]
The Return of Sherlock Holmes [[0.82215613]]
The Secret Adversary [[0.94427821]]
The Awakening, and Selected Short Stories [[0.85557452]]
The Vanished Messenger [[0.93452312]]
The Seventh Man [[1.]]
Madame Bovary: A Tale of Provincial Life, Vol. 1 (of 2) [[0.88445622]]
A Marriage at Sea [[0.84446565]]
Paradise Bend [[0.87589121]]
The Memoirs of Sherlock Holmes [[0.92261215]]
Tarzan and the Jewels of Opar [[0.92404051]]
```

Observations:

Similar genre books have a similarity rating upwards of 0.70 in most of the genres. Since the book genres are not mutually exclusive, books that have a small resembles also ends up getting good similarity. Example, Book “Alice Adventure in Wonderland” belong to fantasy genre. We have book “The Haunted Man and Ghost’s Bargain” of horror genre having a high similarity.