CS 6350
ASSIGNMENT _1_

Names of students in your group:

Kuei-Yu Tsai (kxt230002)

Number of free late days used: _0_

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

1. The Spirit of St. Francis de Sales by Jean-Pierre Camus:
   https://www.gutenberg.org/ebooks/9184

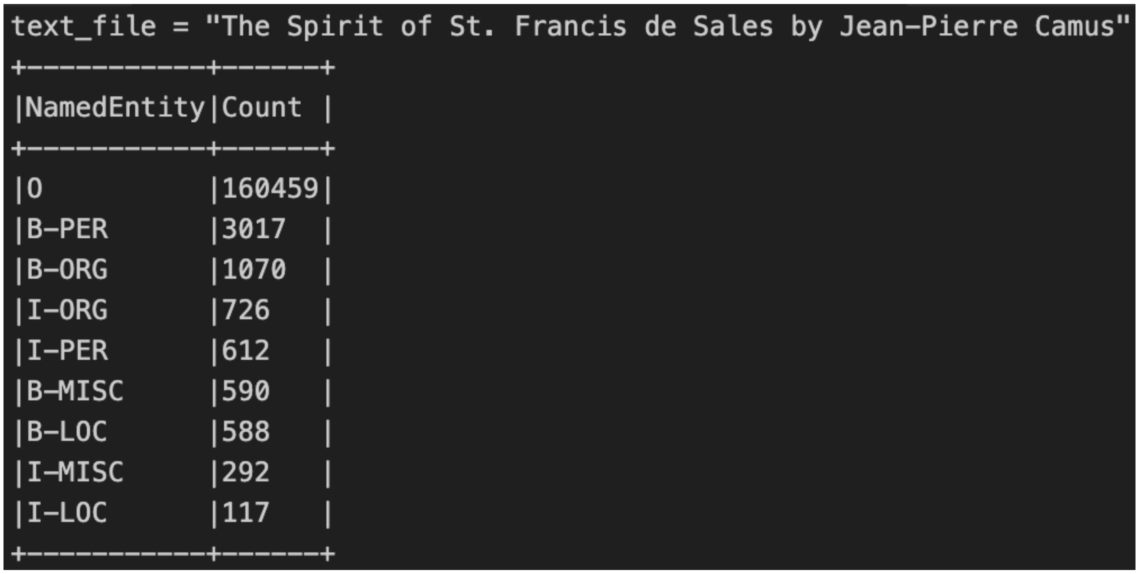
2. JohnSnowLabs Spark-NLP:
   https://github.com/JohnSnowLabs/spark-nlp

3. Glove Embeddings 6B 100:
   https://sparknlp.org/2020/01/22/glove_100d.html

4. NerDLModel:
   https://sparknlp.org/api/com/johnsnowlabs/nlp/annotators/ner/dl/NerDLModel$.html

5. Movie Summaries:
   http://www.cs.cmu.edu/~ark/personas/data/MovieSummaries.tar.gz

1. WordCount for Named Entities:
   https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3469934735814656/444921680346973/520102209841210/latest.html

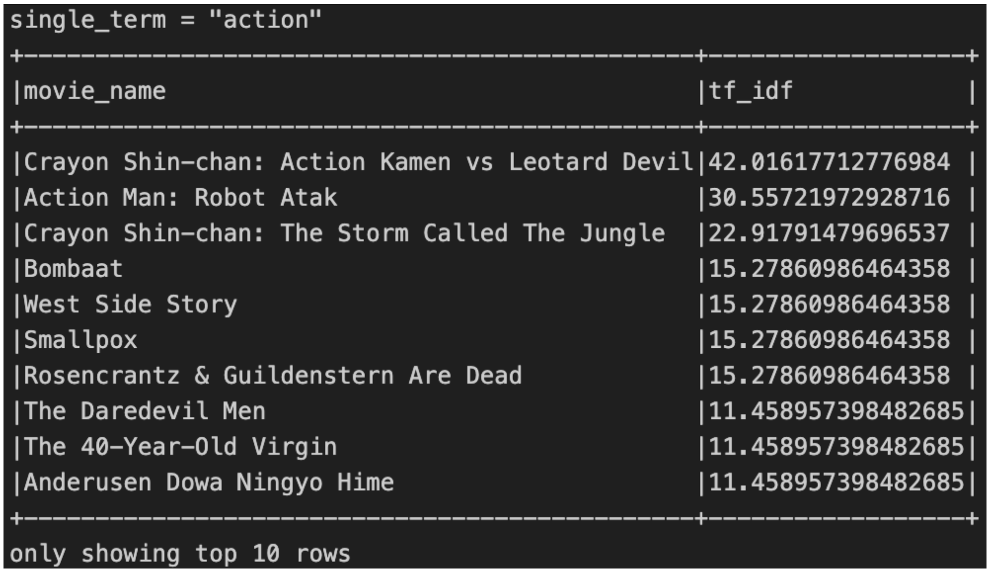


```
text_file = "The Spirit of St. Francis de Sales by Jean-Pierre Camus"
+-----------+------+
|NamedEntity|Count |
+-----------+------+
|O          |160459|
|B-PER      |3017  |
|B-ORG      |1070  |
|I-ORG      |726   |
|I-PER      |612   |
|B-MISC     |590   |
|B-LOC      |588   |
|I-MISC     |292   |
|I-LOC      |117   |
+-----------+------+
```

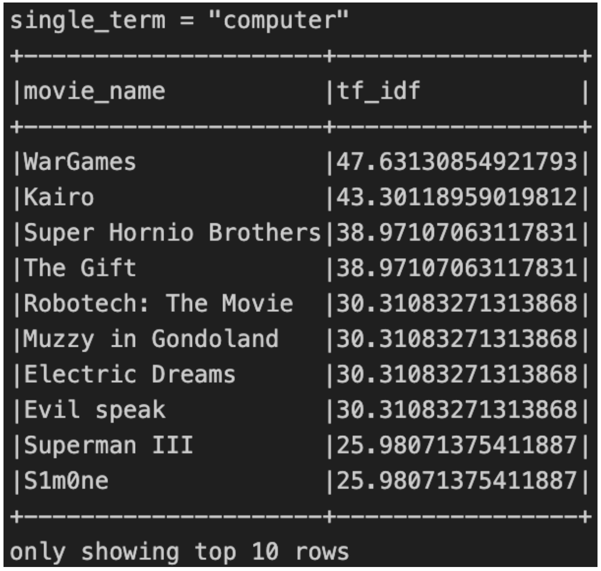
2. Search Engine for Movie Plot Summaries:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3469934735814656/1910473543019259/520102209841210/latest.html

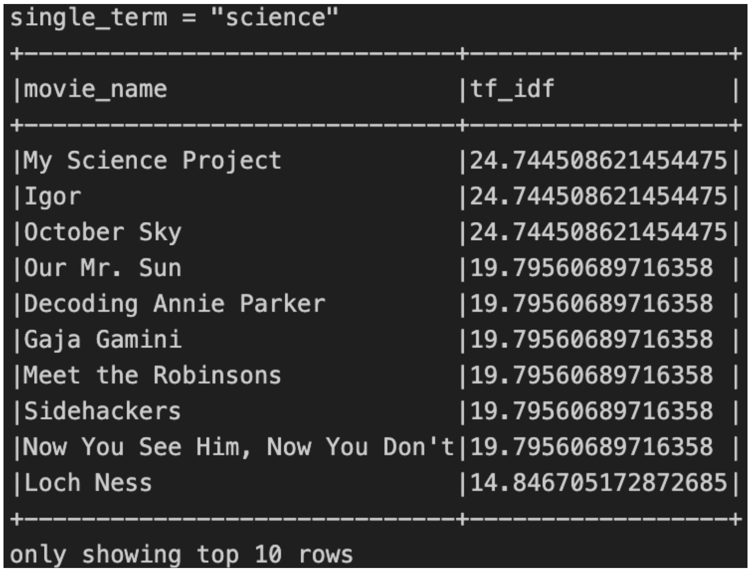
Single Terms:



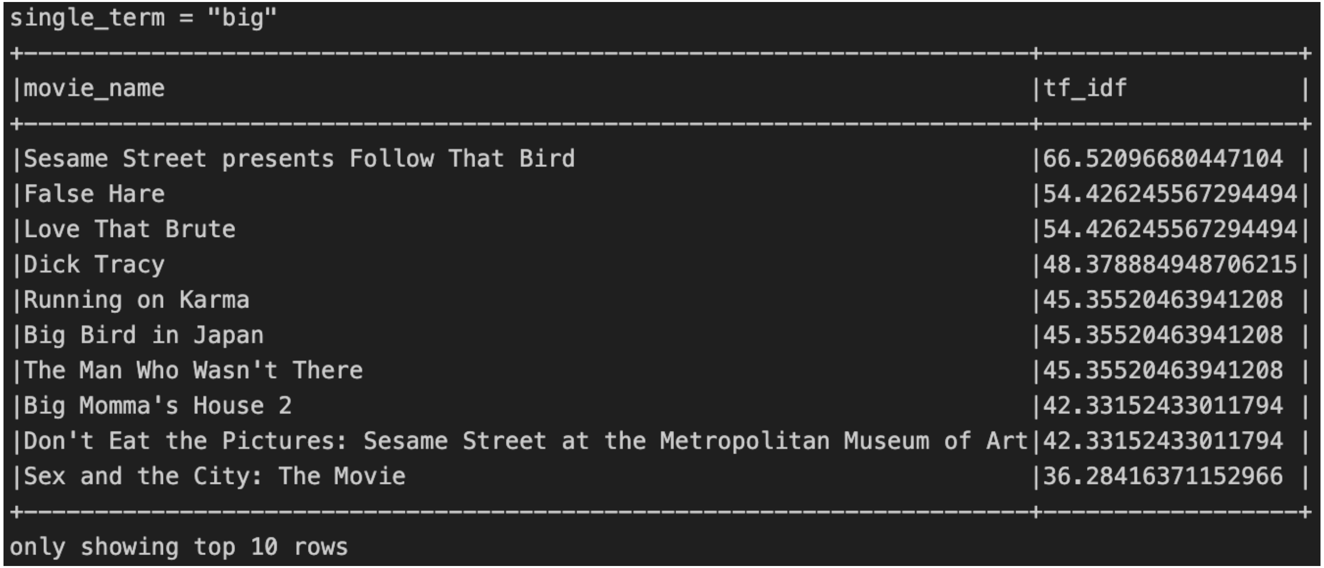
```
single_term = "action"
+-----------------------------------------------+------------------+
|movie_name                                     |tf_idf            |
+-----------------------------------------------+------------------+
|Crayon Shin-chan: Action Kamen vs Leotard Devil|42.01617712776984 |
|Action Man: Robot Atak                         |30.55721972928716 |
|Crayon Shin-chan: The Storm Called The Jungle  |22.91791479696537 |
|Bombaat                                        |15.27860986464358 |
|West Side Story                                |15.27860986464358 |
|Smallpox                                       |15.27860986464358 |
|Rosencrantz & Guildenstern Are Dead            |15.27860986464358 |
|The Daredevil Men                              |11.458957398482685|
|The 40-Year-Old Virgin                         |11.458957398482685|
|Anderusen Dowa Ningyo Hime                     |11.458957398482685|
+-----------------------------------------------+------------------+
only showing top 10 rows
```



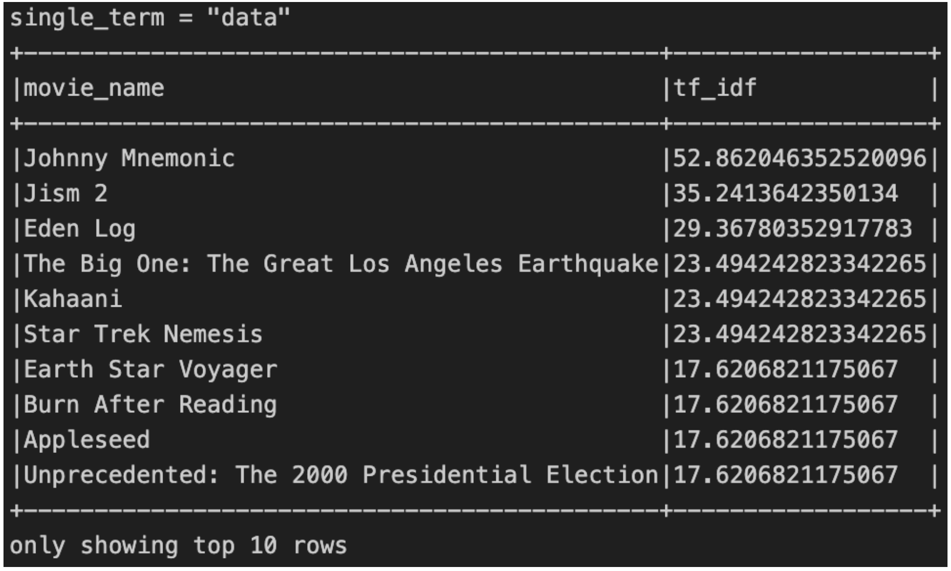
```
single_term = "computer"
+---------------------+-----------------+
|movie_name           |tf_idf           |
+---------------------+-----------------+
|WarGames             |47.63130854921793|
|Kairo                |43.30118959019812|
|Super Hornio Brothers|38.97107063117831|
|The Gift             |38.97107063117831|
|Robotech: The Movie  |30.31083271313868|
|Muzzy in Gondoland   |30.31083271313868|
|Electric Dreams      |30.31083271313868|
|Evil speak           |30.31083271313868|
|Superman III         |25.98071375411887|
|S1m0ne               |25.98071375411887|
+---------------------+-----------------+
only showing top 10 rows
```



```
single_term = "science"
+------------------------------+------------------+
|movie_name                    |tf_idf            |
+------------------------------+------------------+
|My Science Project            |24.744508621454475|
|Igor                          |24.744508621454475|
|October Sky                   |24.744508621454475|
|Our Mr. Sun                   |19.79560689716358 |
|Decoding Annie Parker         |19.79560689716358 |
|Gaja Gamini                   |19.79560689716358 |
|Meet the Robinsons            |19.79560689716358 |
|Sidehackers                   |19.79560689716358 |
|Now You See Him, Now You Don't|19.79560689716358 |
|Loch Ness                     |14.846705172872685|
+------------------------------+------------------+
only showing top 10 rows
```

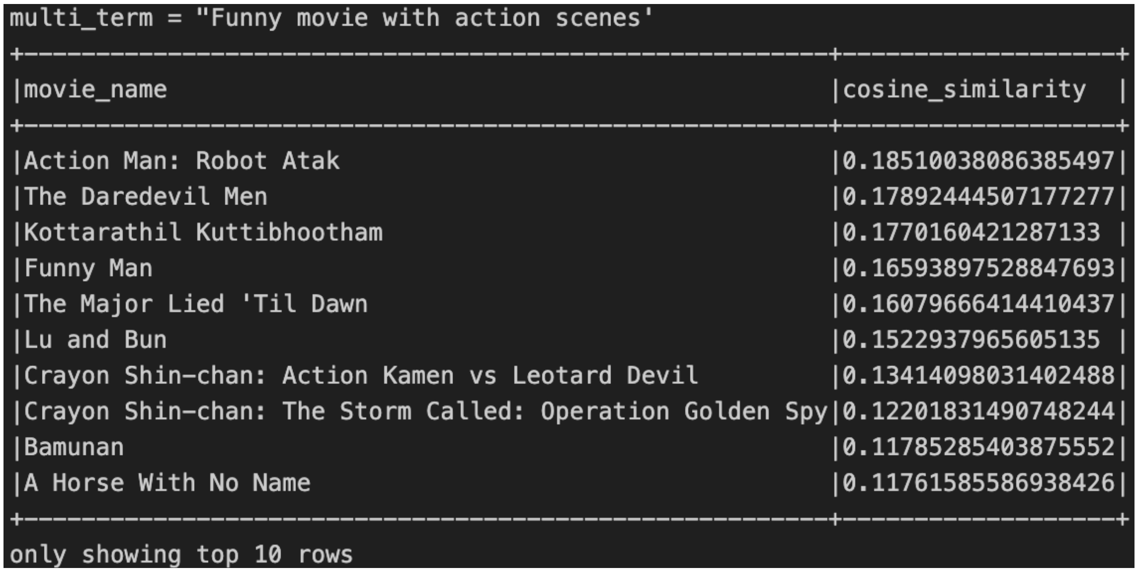


```
single_term = "big"
+-----------------------------------------------------------------------+------------------+
|movie_name                                                             |tf_idf            |
+-----------------------------------------------------------------------+------------------+
|Sesame Street presents Follow That Bird                                |66.52096680447104 |
|False Hare                                                             |54.426245567294494|
|Love That Brute                                                        |54.426245567294494|
|Dick Tracy                                                             |48.378884948706215|
|Running on Karma                                                       |45.35520463941208 |
|Big Bird in Japan                                                      |45.35520463941208 |
|The Man Who Wasn't There                                               |45.35520463941208 |
|Big Momma's House 2                                                    |42.33152433011794 |
|Don't Eat the Pictures: Sesame Street at the Metropolitan Museum of Art|42.33152433011794 |
|Sex and the City: The Movie                                            |36.28416371152966 |
+-----------------------------------------------------------------------+------------------+
only showing top 10 rows
```

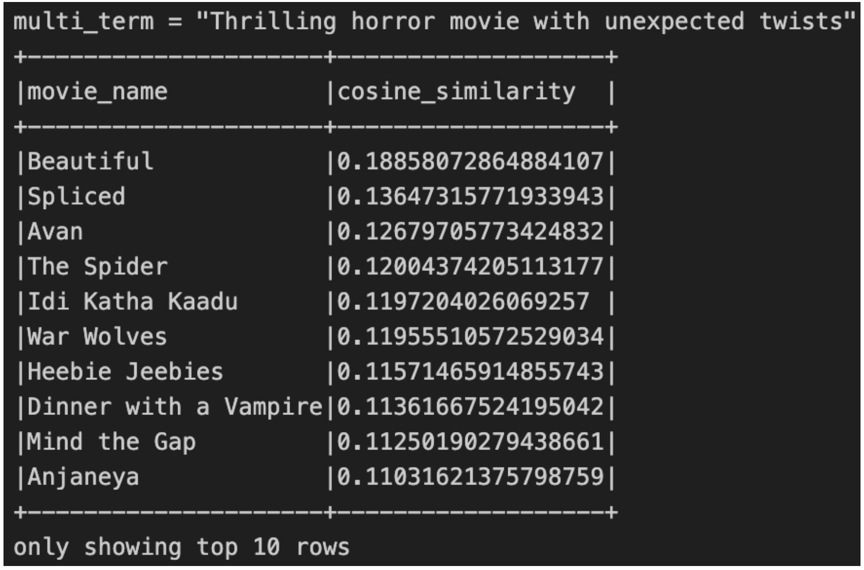


```
single_term = "data"
+---------------------------------------------+------------------+
|movie_name                                   |tf_idf            |
+---------------------------------------------+------------------+
|Johnny Mnemonic                              |52.862046352520096|
|Jism 2                                       |35.2413642350134  |
|Eden Log                                     |29.36780352917783 |
|The Big One: The Great Los Angeles Earthquake|23.494242823342265|
|Kahaani                                      |23.494242823342265|
|Star Trek Nemesis                            |23.494242823342265|
|Earth Star Voyager                           |17.6206821175067  |
|Burn After Reading                           |17.6206821175067  |
|Appleseed                                    |17.6206821175067  |
|Unprecedented: The 2000 Presidential Election|17.6206821175067  |
+---------------------------------------------+------------------+
only showing top 10 rows
```

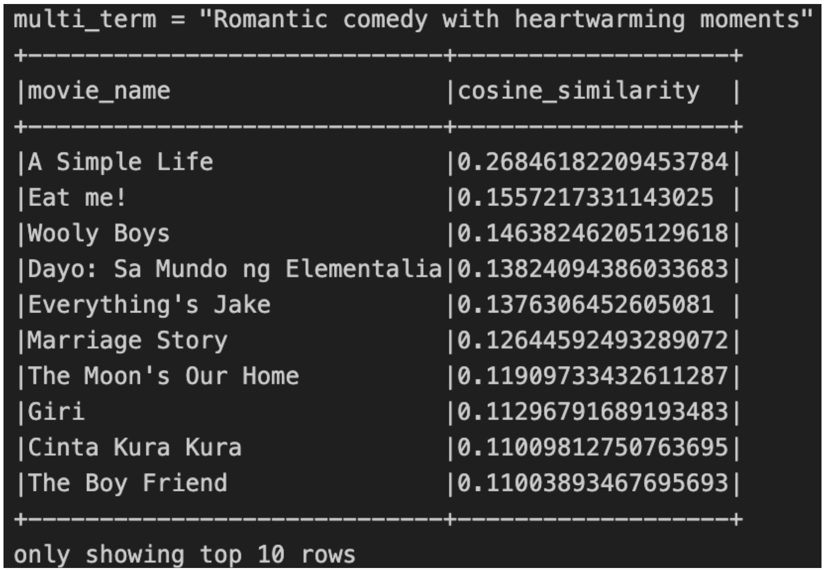
Multi Terms:



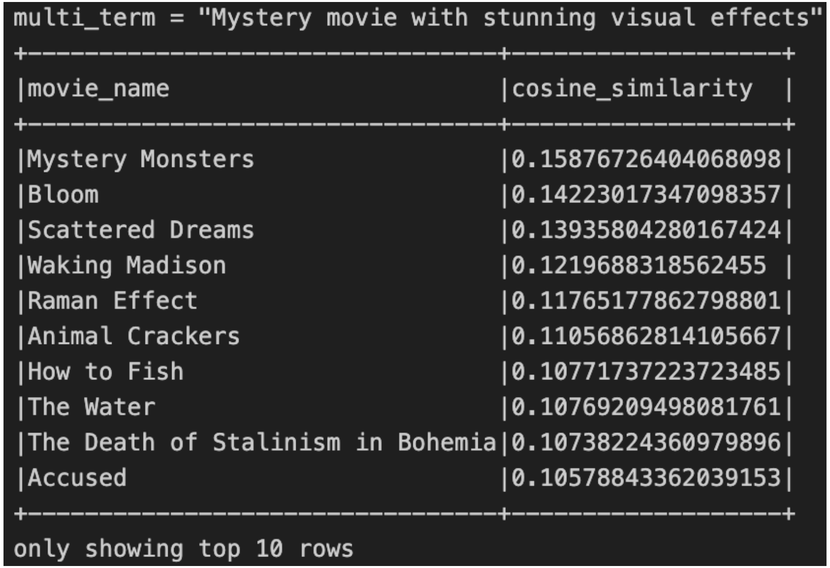
```
multi_term = "Funny movie with action scenes'
+--------------------------------------------------------+-------------------+
|movie_name                                              |cosine_similarity  |
+--------------------------------------------------------+-------------------+
|Action Man: Robot Atak                                  |0.18510038086385497|
|The Daredevil Men                                       |0.17892444507177277|
|Kottarathil Kuttibhootham                               |0.177016042128713 3|
|Funny Man                                               |0.16593897528847693|
|The Major Lied 'Til Dawn                                |0.16079666414410437|
|Lu and Bun                                              |0.1522937965605135 |
|Crayon Shin-chan: Action Kamen vs Leotard Devil         |0.13414098031402488|
|Crayon Shin-chan: The Storm Called: Operation Golden Spy|0.12201831490748244|
|Bamunan                                                 |0.11785285403875552|
|A Horse With No Name                                    |0.11761585586938426|
+--------------------------------------------------------+-------------------+
only showing top 10 rows
```



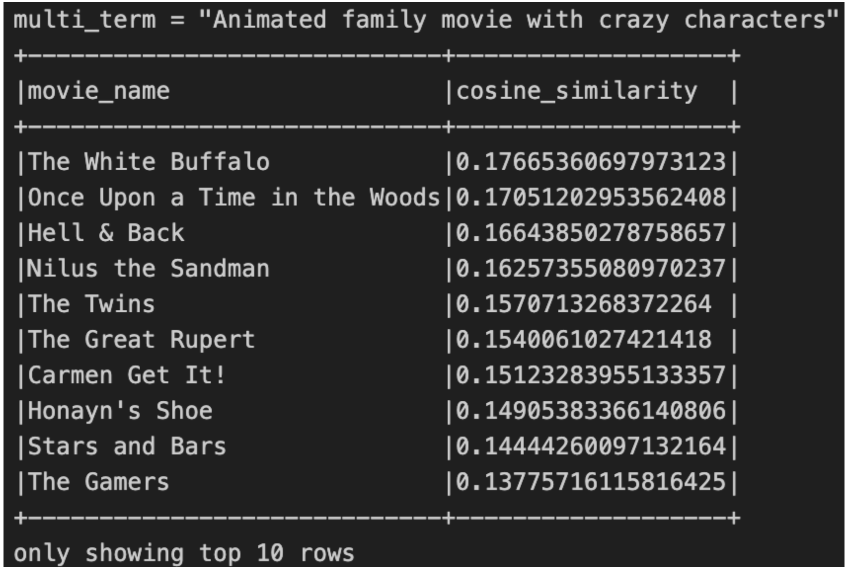
```
multi_term = "Thrilling horror movie with unexpected twists"
+---------------------+-------------------+
|movie_name           |cosine_similarity  |
+---------------------+-------------------+
|Beautiful            |0.18858072864884107|
|Spliced              |0.13647315771933943|
|Avan                 |0.12679705773424832|
|The Spider           |0.12004374205113177|
|Idi Katha Kaadu      |0.1197204026069257 |
|War Wolves           |0.11955510572529034|
|Heebie Jeebies       |0.11571465914855743|
|Dinner with a Vampire|0.11361667524195042|
|Mind the Gap         |0.11250190279438661|
|Anjaneya             |0.11031621375798759|
+---------------------+-------------------+
only showing top 10 rows
```



```
multi_term = "Romantic comedy with heartwarming moments"
+-----------------------------+-------------------+
|movie_name                   |cosine_similarity  |
+-----------------------------+-------------------+
|A Simple Life                |0.26846182209453784|
|Eat me!                      |0.1557217331143025 |
|Wooly Boys                   |0.14638246205129618|
|Dayo: Sa Mundo ng Elementalia|0.13824094386033683|
|Everything's Jake            |0.1376306452605081 |
|Marriage Story               |0.12644592493289072|
|The Moon's Our Home          |0.11909733432611287|
|Giri                         |0.11296791689193483|
|Cinta Kura Kura              |0.11009812750763695|
|The Boy Friend               |0.11003893467695693|
+-----------------------------+-------------------+
only showing top 10 rows
```



```
multi_term = "Mystery movie with stunning visual effects"
+---------------------------------+-------------------+
|movie_name                       |cosine_similarity  |
+---------------------------------+-------------------+
|Mystery Monsters                 |0.15876726404068098|
|Bloom                            |0.14223017347098357|
|Scattered Dreams                 |0.13935804280167424|
|Waking Madison                   |0.1219688318562455 |
|Raman Effect                     |0.11765177862798801|
|Animal Crackers                  |0.11056862814105667|
|How to Fish                      |0.10771737223723485|
|The Water                        |0.10769209498081761|
|The Death of Stalinism in Bohemia|0.10738224360979896|
|Accused                          |0.10578843362039153|
+---------------------------------+-------------------+
only showing top 10 rows
```



```
multi_term = "Animated family movie with crazy characters"
+-----------------------------+-------------------+
|movie_name                   |cosine_similarity  |
+-----------------------------+-------------------+
|The White Buffalo            |0.17665360697973123|
|Once Upon a Time in the Woods|0.17051202953562408|
|Hell & Back                  |0.16643850278758657|
|Nilus the Sandman            |0.16257355080970237|
|The Twins                    |0.1570713268372264 |
|The Great Rupert             |0.1540061027421418 |
|Carmen Get It!               |0.15123283955133357|
|Honayn's Shoe                |0.14905383366140806|
|Stars and Bars               |0.14444260097132164|
|The Gamers                   |0.13775716115816425|
+-----------------------------+-------------------+
only showing top 10 rows
```