

Measuring Movie Similarity through Movie Synopses

...

By:

Zack Christiansen

Advised by: Donna Calhoun and Grady Wright

Why Measure Movie Similarity?

Help find a show or movie to watch with minimal effort (Recommended Movies)

Netflix already measures:

- Viewing history and how you rated other titles
- Other members with similar tastes
- Genre, categories, actors, release year, etc.
- Time of day you're watching
- How long you watch
- Devices you watch on

More Like This



New

PG-13 2002



An FBI agent makes it his mission to put cunning con man Frank Abagnale Jr. behind bars. But Frank not only eludes capture, he revels in the pursuit.



94% Match

R 2015



A group of wily opportunists make a fortune off of the U.S. economic crash by sniffing out the situation in advance and betting against the banks.



New

R 1983



In a ruthless rise to Miami drug lord, a Cuban-born gangster descends into addiction, obsession and brutality, with grisly consequences.



91% Match

R 2018



With his formidable wife Lynne by his side, Dick Cheney gains power and shrewdly manipulates the U.S. vice presidency with explosive global consequences.



82% Match

R 2002



In 1995 Detroit, a talented rapper fights to prove himself on the local hip-hop scene. Eminem stars in this acclaimed drama loosely based on his life.






TV-MA 2018



With Frank out of the picture, Claire Underwood steps fully into her own as the first woman president, but faces formidable threats to her legacy.

Turning a Synopsis into a Vector

- Get rid of stop words (and, if, a, the, etc.)
- Cleansing synopsis of characters like : ' , . - !
- Count up the times each word is used in a synopsis
- Put these values into a dictionary where the word is the key and the word count is the value.
- $O(n)$

Key		Value
Frodo		3
Mountain		2
Ring		5

Getting Rid of Stop Words

```
1 import pandas as pd
2 n = 14544 #Cleaned dataset number of rows
3 file = pd.read_csv('Synopsis\Cleaned.csv', usecols = [1,2], nrows = n)
```

```
1 #Get only synopsis numpy array
2 desc = file.drop('title',axis=1)
3 desc = desc.to_numpy()
4 desc = desc.astype('str')
5
6 #Removing stop words from every description. Keep titles as is.
7 stopWords = ['the','to','a','from','as','the','with','and','in','of','but','for','it','was','have','on','that']
8 temp = []
9 for i in desc:
10     querywords = str(i).split()
11     resultWords=[word for word in querywords if word.lower() not in stopWords]
12     join = ' '.join(resultWords)
13     temp.append(join)
14
15 desc=temp
```

```
1 len(desc)
```

14544

Synopsis Before Cleaning (Lord of the Rings)

Early in the Second Age of Middle earth elven smiths forged nine Rings of Power for mortal men seven for the Dwarf Lords and three for the Elf Kings At the same time the Dark Lord Sauron made the One Ring to rule them all after learning the secrets of how to forge them from the Elves of Hollin a deviation from Tolkien's work in which Sauron taught ring lore to the Elves and forged all the rings except the three Elvish rings As the Last Alliance of Elves and Men fell the Ring fell into the hands of Prince Isildur from across the sea and after Isildur was killed by orcs the Ring lay at the bottom of the river Anduin Over time Sauron captured the nine Rings made for men and turned their owners into the Ringwraiths terrible beings who roamed the world searching for the One Ring The Ring was found by a Stoor named Dáagol whose friend Smāgol murdered him and stole it for himself The Ring warped Smāgol into a twisted gurgling wretch known only as Gollum Peter Woodthorpe and he wandered with it to a cave in the Misty Mountains Hundreds of years later the hobbit Bilbo Baggins Norman Bird accidentally discovered his precious Ring and took it back with him to the Shire

Years later during Bilbo's birthday celebrations in the Shire the wizard Gandalf William Squire tells him to leave the Ring for Frodo Baggins Christopher Guard Bilbo agrees and leaves the Shire Seventeen years pass during which Gandalf learns that the Shire is in danger evil forces have discovered that the Ring is in the possession of a Baggins Gandalf meets with Frodo to explain the Ring's history and the danger it poses to all of Middle earth Frodo leaves his home taking the Ring with him

He is accompanied by three hobbit friends Pippin Dominic Guard Merry Simon Chandler and Sam Michael Scholes After a narrow escape from the Ringwraiths pursuing them the hobbits eventually come to Bree where they meet Aragorn John Hurt who is first introduced to them as Strider a friend of Gandalf's who leads them the rest of the way to Rivendell Frodo is stabbed atop Weathertop mountain by the chief of the Ringwraiths with a knife imbued with evil magic Part of the knife stays inside him and he gets sicker as the journey progresses The Ringwraiths catch up with them shortly after they meet the elf Legolas Anthony Daniels and at a standoff at the ford of Rivendell the Ringwraiths are swept away by the enchanted river At Rivendell Frodo is healed by its lord Elrond He meets Gandalf again held captive by his fellow wizard Saruman Fraser Kerr who plans to join with Sauron but also wants the Ring for himself Bilbo Gandalf and the others argue about what should be done with the One Ring and Frodo volunteers to go to Mordor where the Ring can be destroyed Frodo sets off from Rivendell with eight companions Gandalf; Aragorn; Boromir son of the Steward of Gondor Michael Graham Cox; Legolas; Gimli the dwarf David Buck; and Frodo's original three hobbit companions

Their attempt to cross the Misty Mountains is foiled by heavy snow and they are forced to take a path under the mountains via Moria Moria was an ancient dwarf kingdom but is now full of orcs and other evil creatures and Gandalf falls into an abyss while battling a balrog The remaining eight members of the Fellowship continue through the elf haven Lothlórien but Boromir tries to take the Ring from Frodo Frodo decides to leave the others behind and continue his quest alone although faithful Sam insists on accompanying him

Boromir is killed by orcs while trying to defend Merry and Pippin They are captured by the orcs who intend to take them to Isengard through the land of Rohan The hobbits escape and flee into Fangorn forest where they meet Treebeard John Westbrook a huge tree like creature Aragorn Gimli and Legolas find Merry and Pippin; they find small footprints and follow them into Fangorn Forest There they find Gandalf whom they believed had died in the mines of Moria The four ride to Rohan's capital Edoras where Gandalf persuades King Thāoden Philip Stone that his people are in danger Aragorn Gimli and Legolas then travel to the defensive fortification Helm's Deep

Frodo and Sam meanwhile discover Gollum stalking them and capture him Frodo pities him and lets him live in return for guidance to Mount Doom Gollum promises to lead them to a secret entrance to Mordor At Helm's Deep Thāoden's forces struggle to resist an onslaught of orcs sent by Saruman Gandalf arrives the next morning with the Riders of Rohan just in time destroying the orc army

Synopsis After Cleaning

Early Second Age Middle earth elven smiths forged nine Rings Power mortal men seven Dwarf Lords three Elf Kings same time Dark Lord Sauron made One Ring rule them all after learning secrets how forge them Elves Hollin deviation Tolkien s work which Sauron taught ring lore Elves forged all rings except three Elvish rings Last Alliance Elves Men fell Ring fell into hands Prince Isildur across sea after Isildur killed orcs Ring lay bottom river Anduin Over time Sauron captured nine Rings made men turned their owners into Ringwraiths terrible beings who roamed world searching One Ring Ring found Stoor named Déagol whose friend Sméagol murdered him stole himself Ring warped Sméagol into twisted gurgling wretch known only Gollum Peter Woodthorpe wandered cave Misty Mountains Hundreds years later hobbit Bilbo Baggins Norman Bird accidentally discovered precious Ring took back him Shire\\nYears later during Bilbo s birthday celebrations Shire wizard Gandalf William Squire tells him leave Ring Frodo Baggins Christopher Guard Bilbo agrees leaves Shire Seventeen years pass during which Gandalf learns Shire danger evil forces discovered Ring possession Baggins Gandalf meets Frodo explain Ring s history danger poses all Middle earth Frodo leaves home taking Ring him\\nHe accompanied three hobbit friends Pippin Dominic Guard Merry Simon Chandler Sam Michael Scholes After narrow escape Ringwraiths pursuing them hobbits eventually come Bree where meet Aragorn John Hurt who first introduced them Strider friend Gandalf s who leads them rest way Rivendell Frodo stabbed atop Weathertop mountain chief Ringwraiths knife imbued evil magic Part knife stays inside him gets sicker journey progresses Ringwraiths catch up them shortly after meet elf Legolas Anthony Daniels standoff ford Rivendell Ringwraiths swept away enchanted river Rivendell Frodo healed lord Elrond meets Gandalf again held captive fellow wizard Saruman Fraser Kerr who plans join Sauron also wants Ring himself Bilbo Gandalf others argue about what should done One Ring Frodo volunteers go Mordor where Ring can destroyed Frodo sets off Rivendell eight companions Gandalf; Aragorn; Boromir son Steward Gondor Michael Graham Cox; Legolas; Gimli dwarf David Buck; Frodo s original three hobbit companions\\nTheir attempt cross Misty Mountains foiled heavy snow forced take path under mountains via Moria Moria an ancient dwarf kingdom now full orcs other evil creatures Gandalf falls into an abyss while battling balrog remaining eight members Fellowship continue through elf haven Lothlórien Boromir tries take Ring Frodo Frodo decides leave others behind continue quest alone although faithful Sam insists accompanying him\\nBoromir killed orcs while trying defend Merry Pippin captured orcs who intend take them Isengard through land Rohan hobbits escape flee into Fangorn forest where meet Treebeard John Westbrook huge tree like creature Aragorn Gimli Legolas find Merry Pippin; find small footprints follow them into Fangorn Forest There find Gandalf whom believed had died mines Moria four ride Rohan s capital Edoras where Gandalf persuades King Théoden Philip Stone people danger Aragorn Gimli Legolas then travel defensive fortification Helm s Deep\\nFrodo Sam meanwhile discover Gollum stalking them capture him Frodo pities him lets him live return guidance Mount Doom Gollum promises lead them secret entrance Mordor Helm s Deep Théoden s forces struggle resist an onslaught orcs sent Saruman Gandalf arrives next morning Riders Rohan just time destroying orc army

Cosine Similarity

$$A \cdot B = \|A\| \|B\| \cos(\theta) \longrightarrow \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Using this kind of similarity, the number of times a word comes up in a synopsis matters.

Cosine Similarity

```
1 import math
2 import re #Regular expression operations for unpacking documents
3 from collections import Counter #Dictionary
4
5 #compiles words (\w) into a regular expression object
6 WORD = re.compile(r"\w+")
7
8 #Takes in two vectors, finds the angle between them using cosine
9 def cosSimilarity(vec1, vec2):
10     intersection = set(vec1.keys()) & set(vec2.keys())
11
12     #If there is no word that is in both descriptions, return zero
13     if not intersection:
14         return 0.0
15
16     numer = sum([vec1[x] * vec2[x] for x in intersection])
17
18     sum1 = sum([vec1[x] ** 2 for x in list(vec1.keys())])
19     sum2 = sum([vec2[x] ** 2 for x in list(vec2.keys())])
20     denom = math.sqrt(sum1) * math.sqrt(sum2)
21
22     return float(numer) / denom
23
24
25 #Counts every word in a document, returns a vector with the counts of each word
26 def textToVector(text):
27     words = WORD.findall(text)
28     return Counter(words)
```


Jaccard Similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Using this similarity measure, only the number of words two synopsis have in common are counted.

The word counts are not used with Jaccard similarity

Jaccard Similarity

```
def JaccardSimilarity(vec1, vec2):  
    intersection = len(list(set(vec1.keys()).intersection(vec2.keys())))  
  
    union = (len(list(vec1)) + len(list(vec2))) - intersection  
  
    return float(intersection) / union
```

Putting Similarity Calculations into a Dataset

- Compute the similarity between two movies for every combination of movies.
- The resulting matrix should have ones in the diagonals, be square, and be symmetric.
- This matrix is called a distance matrix
- A distance matrix was computed for Jaccard and Cosine similarity.

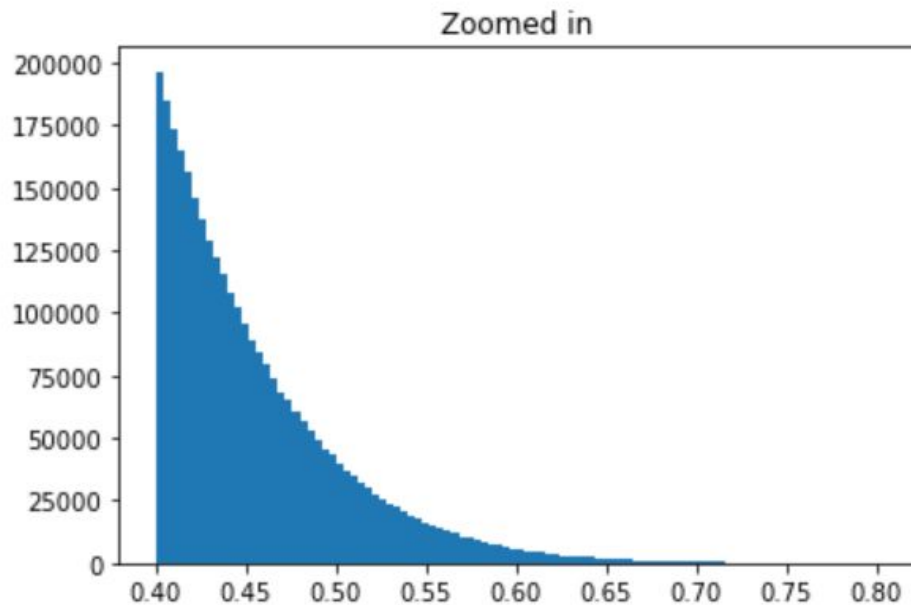
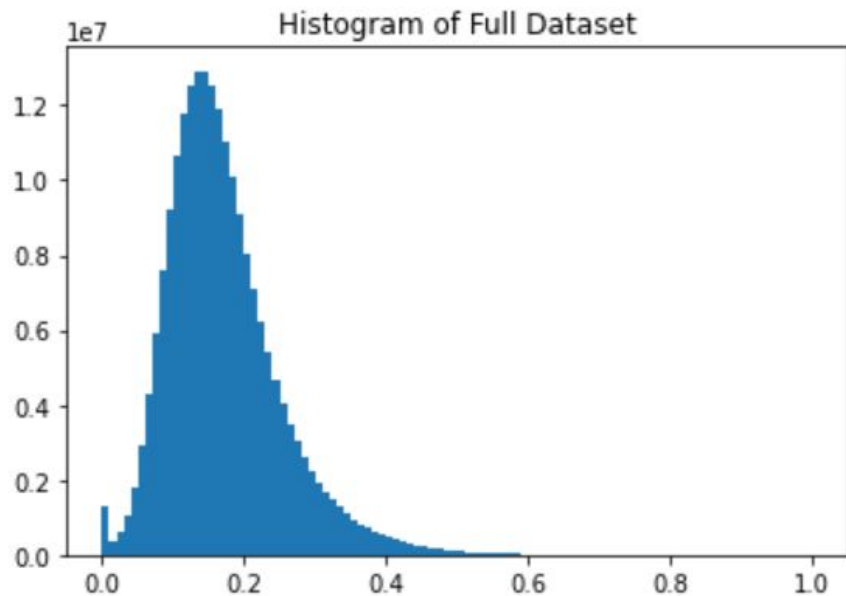
	Movie1	Movie2	Movie3	Movie4
Movie1	1	0.1749	0.483	0.6372
Movie2	0.1749	1	0.1276	0.3926
Movie3	0.483	0.1276	1	0.4274
Movie4	0.6732	0.3926	0.4274	1

Putting Similarity Calculations into a Dataset

```
1  #Does diagonal and loop
2  import numpy as np
3
4  matrix2 = np.empty((n,n),dtype=np.float32)
5  TtoV = [0]*(n)
6  for i in range(n):
7      TtoV[i] = textToVector(desc[i])
8  for i in range(n):
9      vec1 = TtoV[i]
10     for j in range(i,n):
11         vec2 = TtoV[j]
12         if(i==j):
13             temp=1
14         else:
15             temp = cosSimilarity(vec1,vec2)
16         matrix2[i][j] = temp
17         matrix2[j,i]=temp
```

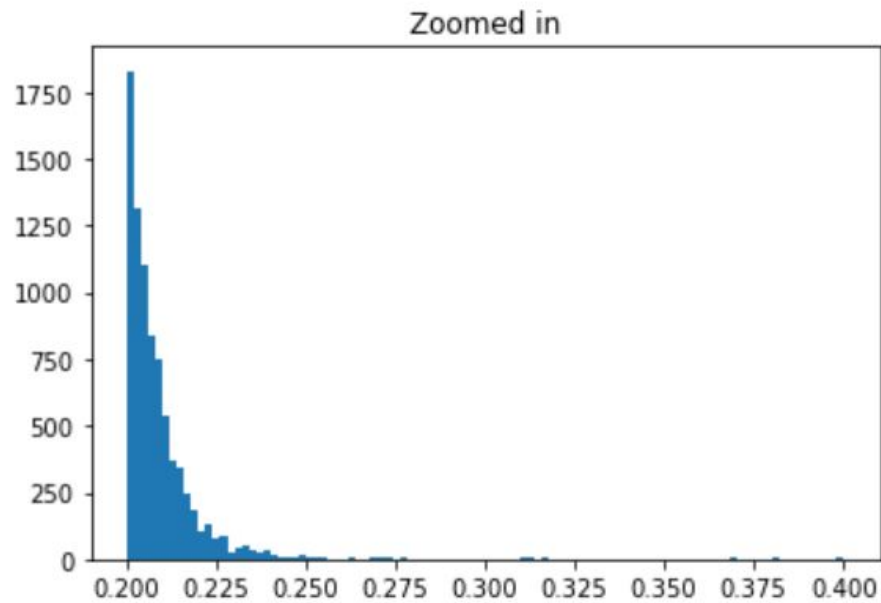
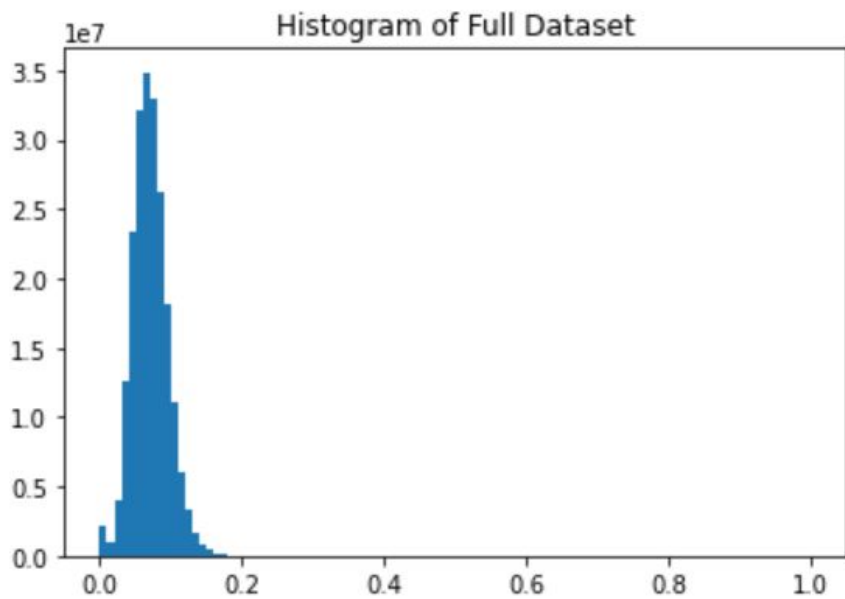
Cosine Similarity Results

- Original dataset had 14,544 movies in it
- Entire similarity dataset took 13 hours to calculate



Jaccard Similarity Results

- Entire dataset took almost 2 hours to complete

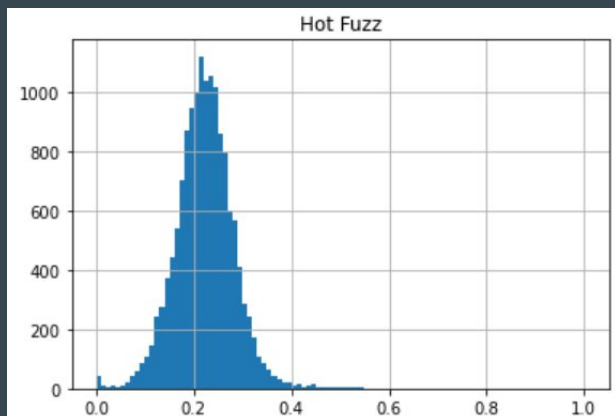


Differences for Specific Movies (Hot Fuzz)

Cosine Similarity

Hot Fuzz 1.0

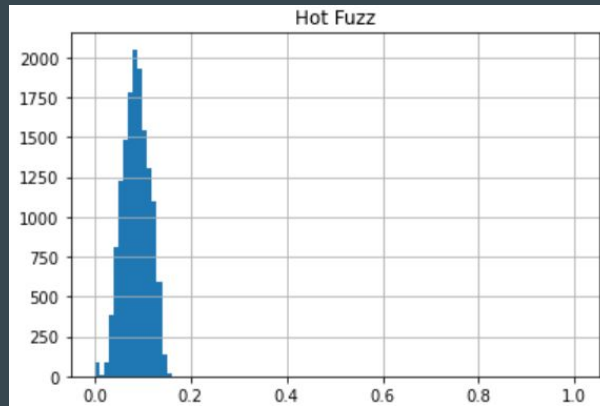
The Thin Man Goes Home 0.5695300102233887
R.I.P.D. 0.5580000281333923
Best Laid Plans 0.5488299727439879
The Butterfly Effect 2 0.5405600070953369
Everything Must Go 0.5393800139427185
30 Minutes or Less 0.5374600291252136
The January Man 0.5319799780845642
By the Gun 0.5228899717330933
What Women Want 0.5210000276565552
Death Sentence.1 0.516040027141571



Jaccard Similarity

Hot Fuzz 1.0

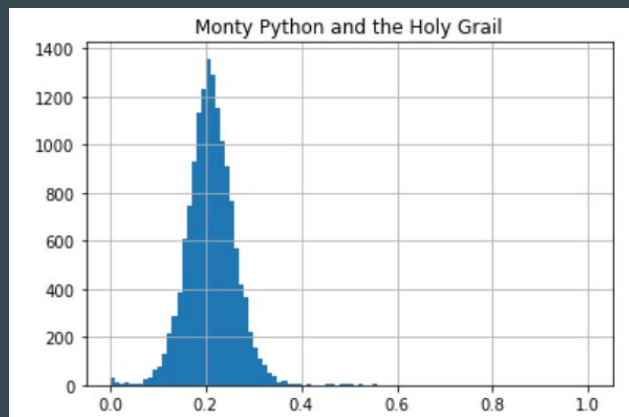
Super Troopers 0.16007000207901
Warm Bodies 0.15914000570774078
Amores perros 0.15836000442504886
Once Upon a Time in America 0.1556999981
American Pastoral 0.15408000349998474
A Horrible Way to Die 0.1540399938821792
Appaloosa 0.1536400020122528
Toxic 0.1532299965620041
Boyz n the Hood 0.1519400030374527
Water for Elephants 0.15146000683307648



Differences For Specific Movies (Monty Python)

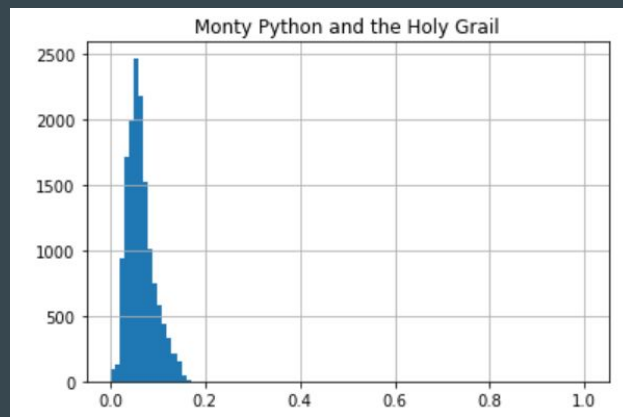
Cosine Similarity

Monty Python and the Holy Grail 1.0
Camelot 0.61976999904441833
Excalibur 0.6042699813842773
King Arthur 0.5773199796676636
Camelot.1 0.5573700070381165
King Arthur: Legend of the Sword 0.5533400177955627
The Hitchhiker's Guide to the Galaxy 0.552049994468689
The Sword in the Stone 0.5328199863433838
Arthur.1 0.5295000076293945
King Arthur.1 0.5257800221443176
Pennies from Heaven 0.5251399874687195



Jaccard Similarity

Monty Python and the Holy Grail 1.0
The Hunchback of Notre Dame.1 0.17237000167369845
Frozen 0.1687300056219101
The Chronicles of Narnia: The Lion the Witch and the Wardrobe 0.16805000
Tenkû no shiro Rapyuta 0.1642100065946579
The Beastmaster 0.1638599932193756
The Chronicles of Narnia: Prince Caspian 0.16369999945163727
Mononoke-hime 0.1632400006055832
Season of the Witch 0.16298000514507294
Tuck Everlasting 0.16243000328540802
Hornblower: The Even Chance 0.15880000591278076



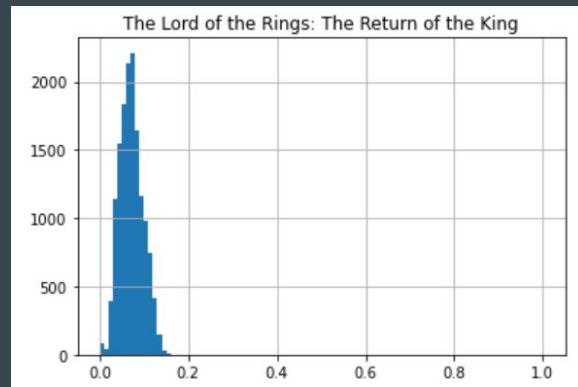
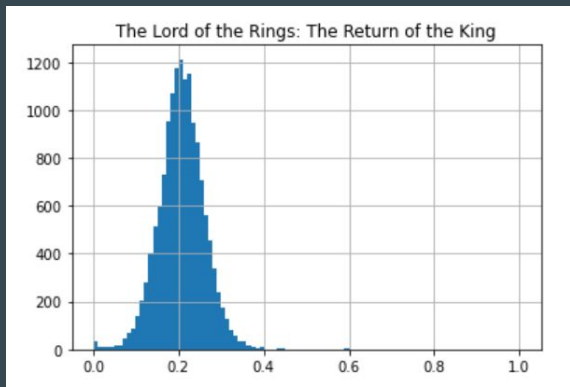
Differences for Specific Movies (LOTR Return of the King)

Cosine Similarity

The Lord of the Rings: The Return of the King 1.0
The Lord of the Rings: The Two Towers.1 0.6534600257873535
The Lord of the Rings: The Two Towers 0.6441299915313721
Lord of the Rings 0.5968400239944458
The Lord of the Rings 0.5924299955368042
The Lord of the Rings: The Fellowship of the Ring.1 0.578279972076416
The Lord of the Rings: The Fellowship of the Ring 0.5558199882507324
The Lord of the Rings: The Third Age 0.5495399832725525
Adulthood 0.4499799907207489
Transformers 0.4412899911403656
The Butterfly Effect 3: Revelations 0.4389500021934509

Jaccard Similarity

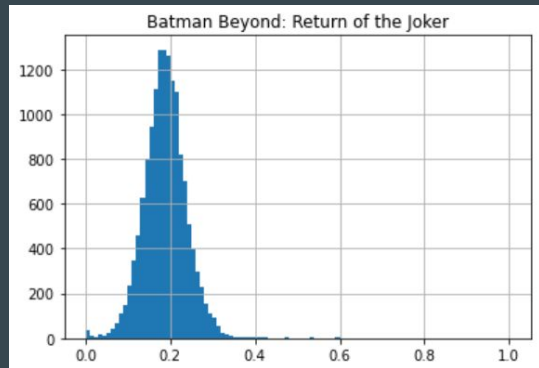
The Lord of the Rings: The Return of the King 1.0
The Lord of the Rings: The Third Age 0.18052999675273892
The Lord of the Rings: The Two Towers 0.17057999968528748
The Hobbit: The Battle of the Five Armies 0.15789000689983368
Mononoke-hime 0.15463000535964966
The Hobbit: An Unexpected Journey 0.1538500040769577
Troy 0.15127000212669373
Conan the Barbarian 0.150519996881485
300: Rise of an Empire 0.14938999712467196
Fury.1 0.14927999675273895
Dawn of the Planet of the Apes 0.1488800048828125



Differences Between Specific Movies (Batman Beyond)

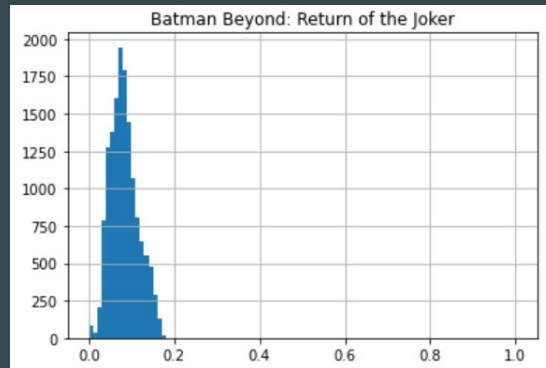
Cosine Similarity

Batman Beyond: Return of the Joker 1.0
The Dark Knight 0.6239200234413147
Batman 0.5931299924850464
Batman Beyond 0.590719997882843
Batman: The Killing Joke 0.5905200242996216
Batman Begins 0.5446299910545349
Batman: Mask of the Phantasm 0.5398100018501282
The Batman vs. Dracula 0.5311800241470337
Lego Batman: The Movie - DC Super Heroes Unite 0.5296800136566162
On the Waterfront 0.5118399858474731
Batman: Assault on Arkham 0.507860004901886



Jaccard Similarity

Batman Beyond: Return of the Joker 1.0
Four Brothers 0.1869100034236908
Batman Begins 0.1797100007534027
The Raid 2: Berandal 0.17680999636650085
Divergent 0.17646999657154086
My Bloody Valentine.1 0.1756799966096878
27 Dresses 0.17486000061035156
Legion 0.17483000457286835
V for Vendetta 0.1746399998664856
The Transporter Refueled 0.17417000234127045
It 0.17371000349521634



Differences Between Specific Movies (Nightmare on Elm St)

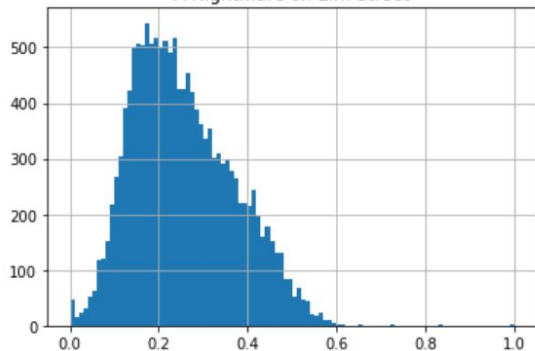
Cosine Similarity

```
A Nightmare on Elm Street 1.0
A Nightmare on Elm Street.2 1.0
A Nightmare on Elm Street.1 0.8343499898910522
A Nightmare on Elm Street 3: Dream Warriors 0.726999998
Nancy Drew 0.6569700241088867
The Shallows 0.6163300275802612
Blood of Dracula 0.6000199913978577
100 Feet 0.5989099740982056
The Collector.1 0.597350001335144
Wild 0.5971999764442444
Cría cuervos 0.5953999757766724
```

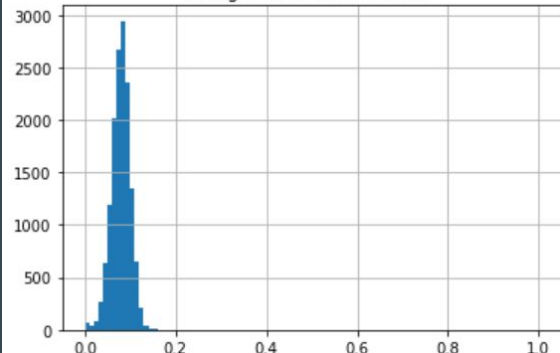
Jaccard Similarity

```
A Nightmare on Elm Street.2 1.0
A Nightmare on Elm Street 1.0
A Nightmare on Elm Street.1 0.1756799966096878
Hide and Go Shriek 0.15738999843597412
A Nightmare on Elm Street 4: The Dream Master 0.156599998474
Scream for Help 0.15253999829292295
Naked Vengeance 0.1505099982023239
Unhinged 0.15000000596046448
A Nightmare on Elm Street Part 2: Freddy's Revenge 0.1490900
The Hole 0.14821000397205353
Bleed 0.1481499969959259
```

A Nightmare on Elm Street



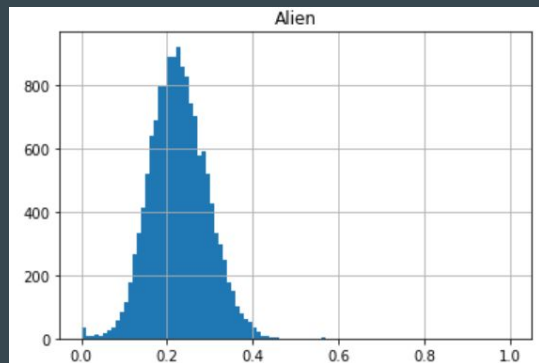
A Nightmare on Elm Street



Differences Between Specific Movies (Alien)

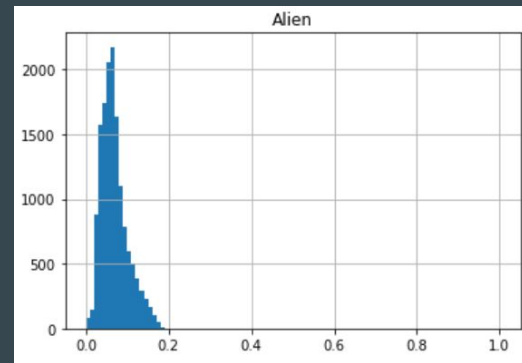
Cosine Similarity

```
Alien 1.0
Alien 2 - Sulla Terra 0.7306399941444397
Aliens 0.5982099771499634
Alien³ 0.5652899742126465
Alien: Resurrection 0.564490020275116
Evil Dead II 0.4958699941635132
Paris je t'aime 0.4585399925708771
How to Steal a Million 0.4564400017261505
Grindhouse 0.4489899873733521
The Librarian: The Curse of the Judas Chalice 0.444959
The Accountant 0.4443399906158447
```



Jaccard Similarity

```
Alien 1.0
Alien³ 0.2075600028038025
Aliens 0.20468999445438385
The Dig 0.19843000173568728
Pandorum 0.19199000298976895
Halloween 4: The Return of Michael Myers 0.181679
1408 0.1815000027418137
Solyaris 0.1808599978685379
Silent Hill 0.18080000579357147
Tenkû no shiro Rapyuta 0.18051999807357788
Oculus 0.1804399937391281
```

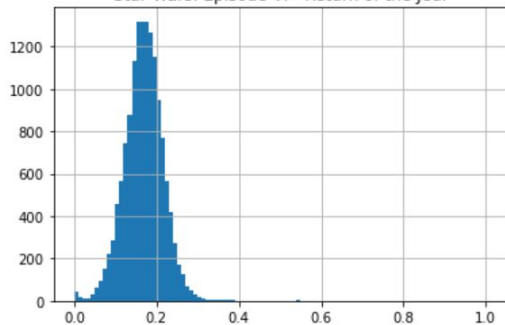


Differences Between Specific Movies (Star Wars)

Cosine Similarity

Star Wars: Episode VI - Return of the Jedi 1.0
Star Wars: Episode V - The Empire Strikes Back 0.7693799
Star Wars 0.7343500256538391
The Ewok Adventure 0.6502199769020081
Robot Chicken: Star Wars Episode III 0.5748900175094604
Death Race 2 0.5464400053024292
Robot Chicken: Star Wars Episode II 0.5463200211524963
Cool Hand Luke 0.5413699746131897
Robot Chicken: Star Wars 0.5359500050544739
The Skulls 0.51978999376297
Thunder Road 0.491320013999939

Star Wars: Episode VI - Return of the Jedi



Jaccard Similarity

Star Wars: Episode VI - Return of the Jedi 1.0
Star Wars: Episode V - The Empire Strikes Back 0.22754000
Star Wars 0.18998999893665314
Star Wars: Episode VII - The Force Awakens 0.172340005636
Star Wars: Episode III - Revenge of the Sith 0.1643799990
Rogue One 0.15982000529766086
Robot Chicken: Star Wars Episode III 0.15834000706672668
X-Men: Apocalypse 0.15588000416755676
Terra 0.15418000519275665
300: Rise of an Empire 0.15177999436855316
Troy 0.14972999691963196

Star Wars: Episode VI - Return of the Jedi

