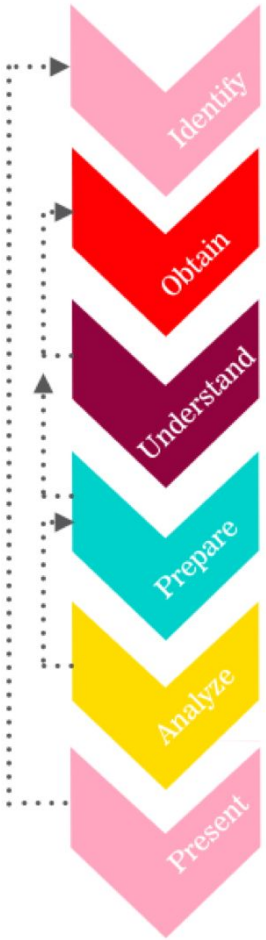
The Greatest Films of All Time

How to use multiple data sources to find a movie you will love

Owen Temple

ANALYTICS WORKFLOW



IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

OBTAIN THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

UNDERSTAND THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

PREPARE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean and combine data
- ☐ Create necessary derived columns from the data (new data)

ANALYZE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Visualize and transform data

PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

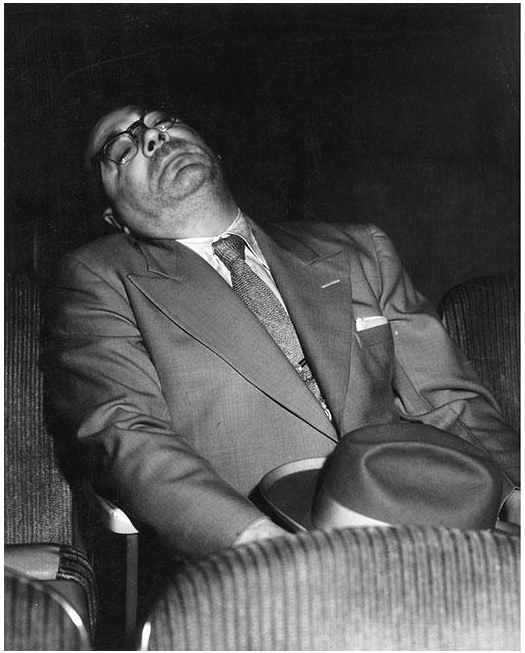
GA

The Problem

What movie is worth spending 90 to 120 minutes on?

The Problem

What movie is worth spending 90 to 120 minutes on?



The Problem

What movie is worth spending 90 to 120 minutes on?

Can we combine and use multiple “greatest films” lists and IMDB ratings to create a reliable movie recommendation?

What decades are “greatest movies” more likely to be from?

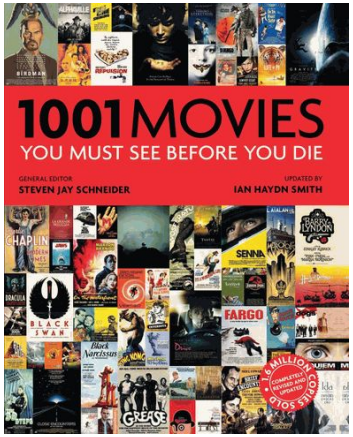
What directors make the “greatest films of all time”?

What genres are included in the greatest films?

The Problem: Multiple Sources

AFI's 100 YEARS 100 MOVIES

WGAW WRITERS GUILD OF AMERICA, WEST



the guardian

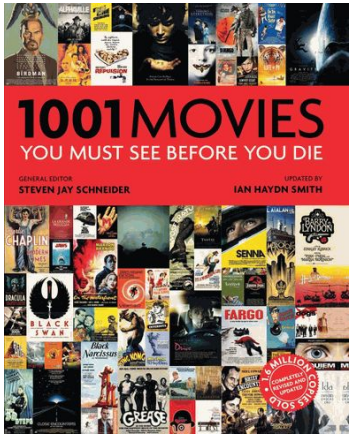
THE INTERNATIONAL FILM MAGAZINE
Sight&Sound

IMDb
The Internet Movie Database

The Problem: Multiple Sources

AFI's 100 YEARS 100 MOVIES
10th ANNIVERSARY EDITION

WGAW
WRITERS GUILD OF AMERICA, WEST



the guardian

INTERNATIONAL FILM MAGAZINE
Sight&Sound

IMDb
The Internet Movie Database

The Problem: Multiple Sources

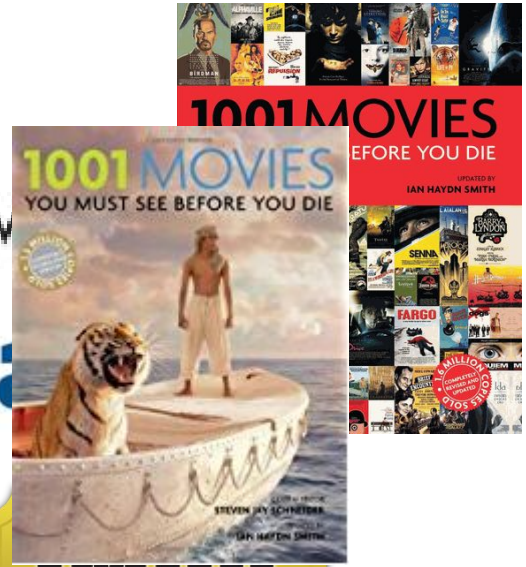
AFI's 100 YEARS 100 MOVIES 10th ANNIVERSARY EDITION

WGAW WRITERS GUILD OF AMERICA, W

the guardian

INTERNATIONAL FILM MAGAZINE

Sight&Sound



IMDb

The Internet Movie Database

The Problem: Multiple Sources

The Problem: Multiple Sources

The Problem: Multiple Sources

WRITERS GUILD OF AMERICA, W

theguard

INTERNATIONAL FILM MAGAZINE

Sight&Sound

The Problem: Multiple Sources

The Problem: Multiple Sources

Gather Titles using Copy/Paste and API

OMDb API

The Open Movie Database

curl://

Primary Key: Find IMDB IDs

Remove Duplicates, Wrong Titles

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	imdb_id	rank_afc_20(	rank_afc_19	rank_wga_2(	rank_ss_201	included_in_	included_in_	title	year	rated	released	runtime	genre	director	writer	actors	plot	language	country	awards	poster	metascore	imdb_rating	imdb_votes
2	tt0077559						TRUE	Five Deadly V	1978	R	12-Aug-78	97 min	Action	Cheh Chang	Cheh Chang,	Sheng Chiang	A dying teach	Mandarin	Hong Kong	N/A	https://imag	N/A	7.2	4,202
3	tt0082971	66	60	42		TRUE	TRUE	Raiders of th	1981	PG	12-Jun-81	115 min	Action, Adve	Steven Spielt	Lawrence Ka	Harrison For	Archaeologis	English, Gerr	USA	Won 4 Oscar	http://ia.me(	85	8.5	665,774
4	tt0026174						TRUE	Captain Bloo	1935	NOT RATED	28-Dec-35	119 min	Action, Adve	Michael Curt	Rafael Sabati	Errol Flynn, (	After being v	English, Fren	USA	N/A	http://ia.me(	N/A	7.8	9,922
5	tt0096895						TRUE	Batman	1989	PG-13	23-Jun-89	126 min	Action, Adve	Tim Burton	Bob Kane (Ba	Michael Keat	The Dark Kni	English, Fren	USA, UK	Won 1 Oscar	http://ia.me(	69	7.6	271,082
6	tt0103285						TRUE	Once Upon a	1991	R	15-Aug-91	134 min	Action, Adve	Hark Tsui	Yiu-ming Leu	Jet Li, Biao Yi	Set in late 19	Cantonese, E	Hong Kong	5 wins & 4 n(	https://imag	N/A	7.4	14,037
7	tt0017925	18			34		TRUE	The General	1926	UNRATED	24-Feb-27	67 min	Action, Adve	Clyde Bruckn	Buster Keato	Buster Keato	When Union	English	USA	2 wins & 1 n(	https://imag	N/A	8.2	54,036
8	tt0046414						TRUE	Beat the Dev	1953	APPROVED	17-Dec-53	89 min	Action, Adve	John Huston	Claud Cockb	Humphrey Bo	On their way	English, Italia	UK, USA, Ital	1 win.	https://imag	N/A	6.6	6,206
9	tt0087332						TRUE	Ghostbusters	1984	PG	8-Jun-84	105 min	Action, Adve	Ivan Reitmar	Dan Aykroyd	Bill Murray, [	Three forme(	English	USA	Nominated f	https://imag	67	7.8	275,143
10	tt0120188						TRUE	Three Kings	1999	R	1-Oct-99	114 min	Action, Adve	David O. Rus	John Ridley (	George Cloo	In the afterm	English, Arab	USA	8 wins & 19 r	http://ia.me(	82	7.1	138,366
11	tt0901487						TRUE	The Good, th	2008	R	17-Jul-08	130 min	Action, Adve	Jee-woon Kir	Jee-woon Kir	Kang-ho Son	The story of	Korean, Man	South Korea	8 wins & 19 r	http://ia.me(	69	7.3	26,319
12	tt0053125	55	40	21		TRUE	TRUE	North by Nor	1959	APPROVED	26-Sep-59	136 min	Action, Adve	Alfred Hitchc	Ernest Lehm	Cary Grant, E	A hapless Ne	English	USA	Nominated f	http://ia.me(	N/A	8.4	221,404
13	tt0006206					TRUE	TRUE	Les vampires	1915	NOT RATED	23-Nov-16	421 min	Action, Adve	Louis Feuillad	Louis Feuillad	Musidora, _E	An intrepid r	French	France	1 nominatior	https://imag	N/A	6.8	3,139
14	tt0120737	50				TRUE	TRUE	The Lord of t	2001	PG-13	19-Dec-01	178 min	Action, Adve	Peter Jackso	J.R.R. Tolkier	Alan Howard	A meek Hobb	English, Sind	New Zealand	Won 4 Oscar	https://imag	92	8.8	1,252,518

Titles sourced from American Film Institute, Writer's Guild of America, Sight & Sound Critic's Survey, The Guardian, and Barron's 1001 Movies to See Before You Die

The columns indicate either rank or inclusion on these lists, imdbID, and other details on the film and its release.

rank_afc_2007

- Film's rank position from 1 to 100 on American Film Institute's 2007 edition of The Greatest American Films of All Time (http://www.afi.com/100Years/movies10.aspx)

rank_afc_1998

- Film's rank position from 1 to 100 on American Film Institute's 1998 edition of The Greatest American Films of All Time (http://www.afi.com/100years/movies.aspx)

rank_wga_2005

- Film's rank position from 1 to 101 on Writers Guild of America's 2005 edition of The 101 Greatest Screenplays (http://www.wga.org/writers-room/101-best-lists/101-greatest-screenplays/list)

rank_ss_2012

- Film's rank position from 1 to 50 on The Sight & Sound 2012 edition of The Top 50 Greatest Films of All Time (http://www.bfi.org.uk/news/50-greatest-films-all-time)

included_in_guardian_list

- True indicates film is included in The Guardian's 2010 edition of The Greatest Films of All Time (https://www.theguardian.com/news/datablog/2010/oct/16/greatest-films-of-all-time)

included_in_1001_book

- True indicates film is included in any edition of the book 1,001 Movies to See Before You Die, 2003 through 2015 (http://1001films.wikia.com/wiki/The_Books)

master_list_greatest_films.csv

- imdb_id
- rank_afc_2007
- rank_afc_1998
- rank_wga_2005
- rank_ss_2012
- included_in_guardian_list
- included_in_1001_book
- title
- year
- rated
- released
- runtime
- genre
- director
- writer
- actors
- plot
- language
- country
- awards
- poster
- metascore
- imdb_rating
- imdb_votes
- alternate_title

Calculate Fields

Dummy-1001 Book

```
IF ([Included In 1001 Book] = TRUE)
THEN 1
ELSE 0
END
```

The calculation is valid. Sheets Affected Apply OK

Calculate Fields

Dummy-1001 Book

```
IF ([Included In 1001 Book] = TRUE)
THEN 1
ELSE 0
END
```

The calculation is valid. Sheets Affected Apply OK

Dummy-AFC 2007

```
IF ISNULL([Rank Afc 2007]) THEN 0
ELSE 1
END
```

The calculation is valid. Sheets Affected Apply OK

Calculate Fields

Dummy-1001 Book

```
IF ([Included In 1001 Book] = TRUE)
THEN 1
ELSE 0
END
```

The calculation is valid. Sheets Affected

Dummy-AFC 2007

```
IF ISNULL([Rank Afc 2007]) THEN 0
ELSE 1
END
```

The calculation is valid. Sheets Affected Apply OK

List of Lists Score

```
[Dummy-1001 Book]+
[Dummy-AFC 1998]+
[Dummy-AFC 2007]+
[Dummy-SS 2012]+
[Dummy-WGA 2005]+
[Dummy-Guardian List]
```

The calculation is valid. Sheets Affected Apply OK

Visualization in Tableau

Limitations

Recommendations are limited by quality of the list, but somewhat easy to add new lists (through Union in Tableau) as desired

More recent movies underrepresented

Next Steps

Now that I have a reliable primary key (IMDB ID), I can link to other databases with movie data

Post data to data.world

A Chance Encounter

Obtain

data.world preview

Robert Lewis
Assistant Professor

Stan Richards School of
Advertising & Public Relations

Obtain

data.world *preview*

Robert Lewis

Assistant Professor

Stan Richards School of
Advertising & Public Relations

Owen, cool data! Check out my profile ... maybe you'll find my dataset useful? There might be some components that you could take from my data and add to yours.

Obtain

data.world preview

Film subtitles' linguistic inquiry word count (liwc), sentiment, and IMDb meta-data, N > 30k

1 file

Filter Sort

6-7-2016 without formulas GenresAndFinanceAndCertsAndSentiment....
Request more info

Explore

GenresAndFinanceAndCertsAndSent | certifications | genres

MovieName	MovieYear	MovieImdbID	UserRank	SubDownloadsCnt
Jan Dara	2001	282628	bronze member	4094
Forgetting the Girl	2012	1492842	No data.	276
One Desire	1955	48448	gold member	83
Genevieve	1953	45808	No data.	200
Rough Night in Jericho	1967	62218	administrator	437

Showing 1-5 of 30,493 rows, 929 columns See all

Switch to column overview

Understand

LIWC

Linguistic Inquiry and Word Count

Moral Foundations Theory was created by a group of social and cultural psychologists (see us here) to understand why morality varies so much across cultures yet still shows so many similarities and recurrent themes. In brief, the theory proposes that several innate and universally available psychological systems are the foundations of "intuitive ethics." Each culture then constructs virtues, narratives, and institutions on top of these foundations, thereby creating the unique moralities we see around the world, and conflicting within nations too. The five foundations for which we think the evidence is best are:

1) Care/harm: This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.
2) Fairness/cheating: This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy. [Note: In our original conception, Fairness included concerns about equality, which are more strongly endorsed by political liberals. However, as we reformulated the theory in 2011 based on new data, we emphasize proportionality, which is endorsed by everyone, but is more strongly endorsed by conservatives]
3) Loyalty/betrayal: This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
4) Authority/subversion: This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
5) Sanctity/degradation: This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

- TotalWords
- HarmVirtue
- HarmVirtue-ratio
- HarmVice
- HarmVice-ratio
- FairnessVirtue
- FairnessVirtue-ratio
- FairnessVice
- FairnessVice-ratio
- IngroupVirtue
- IngroupVirtue-ratio
- IngroupVice
- IngroupVice-ratio
- AuthorityVirtue
- AuthorityVirtue-ratio
- AuthorityVice
- AuthorityVice-ratio
- PurityVirtue
- PurityVirtue-ratio
- PurityVice
- PurityVice-ratio
- MoralityGeneral
- MoralityGeneral-ratio
- random_number
- Negative
- Negative-ratio
- Positive
- Positive-ratio
- Function
- Function-ratio
- Pronoun
- Pronoun-ratio
- Ppron
- Ppron-ratio
- I
- I-ratio
- We
- We-ratio
- You
- You-ratio

- # SheHe
- # SheHe-ratio
- # They
- # They-ratio
- # Ipron
- # Ipron-ratio
- # Article
- # Article-ratio
- # Prep
- # Prep-ratio
- # Auxverb
- # Auxverb-ratio
- # Adverb
- # Adverb-ratio
- # Conj
- # Conj-ratio
- # Negate
- # Negate-ratio
- # Verb
- # Verb-ratio
- # Adj
- # Adj-ratio
- # Compare
- # Compare-ratio
- # Interrog
- # Interrog-ratio
- # Number
- # Number-ratio
- # Quant
- # Quant-ratio
- # Affect
- # Affect-ratio
- # Posemo
- # Posemo-ratio
- # Negemo
- # Negemo-ratio
- # Anx
- # Anx-ratio
- # Anger
- # Anger-ratio
- # Sad
- # Sad-ratio
- # Social
- # Social-ratio
- # Family
- # Family-ratio
- # Friend
- # Friend-ratio

- # Female
- # Female-ratio
- # Male
- # Male-ratio
- # CogProc
- # CogProc-ratio
- # Insight
- # Insight-ratio
- # Cause
- # Cause-ratio
- # Discrep
- # Discrep-ratio
- # Tentat
- # Tentat-ratio
- # Certain
- # Certain-ratio
- # Differ
- # Differ-ratio
- # Percept
- # Percept-ratio
- # See
- # See-ratio
- # Hear
- # Hear-ratio
- # Feel
- # Feel-ratio
- # Bio
- # Bio-ratio
- # Body
- # Body-ratio
- # Health
- # Health-ratio
- # Sexual
- # Sexual-ratio
- # Ingest
- # Ingest-ratio
- # Drives
- # Drives-ratio
- # Affiliation
- # Affiliation-ratio
- # Achieve
- # Achieve-ratio
- # Power
- # Power-ratio
- # Reward
- # Reward-ratio
- # Risk
- # Risk-ratio

- # FocusPast
- # FocusPast-ratio
- # FocusPresent
- # FocusPresent-ratio
- # FocusFuture
- # FocusFuture-ratio
- # Relativ
- # Relativ-ratio
- # Motion
- # Motion-ratio
- # Space
- # Space-ratio
- # Time
- # Time-ratio
- # Work
- # Work-ratio
- # Leisure
- # Leisure-ratio
- # Home
- # Home-ratio
- # Money
- # Money-ratio
- # Relig
- # Relig-ratio
- # Death
- # Death-ratio
- # Informal
- # Informal-ratio
- # Swear
- # Swear-ratio
- # Netspeak
- # Netspeak-ratio
- # Assent
- # Assent-ratio
- # Nonflu
- # Nonflu-ratio

Moral Words in Movie Subtitles Counted

```
%
01          HarmVirtue
02          HarmVice
03          FairnessVirtue
04          FairnessVice
05          IngroupVirtue
06          IngroupVice
07          AuthorityVirtue
08          AuthorityVice
09          PurityVirtue
10          PurityVice
11          MoralityGeneral
%
```

(More) Problems

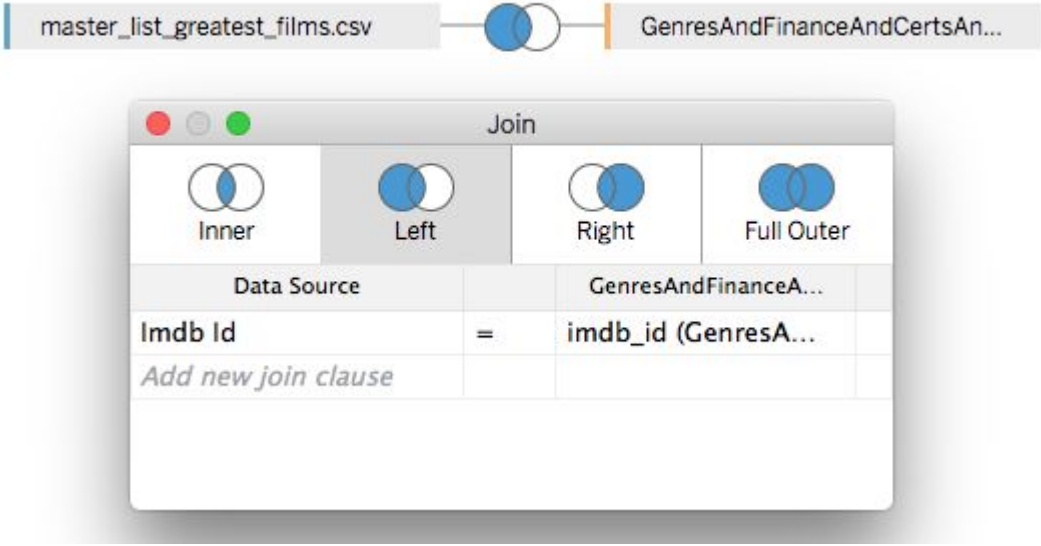
Can we learn about human values through the spoken content in the greatest films of all time?

Are there changes over time in word usage in the greatest films of all time?

What directors are more likely to talk about certain moral issues?

What film titles are particularly concerned with different kinds of moral dilemmas?

Combine Data Sets



Subtitles Matched to Greatest Films

	Subtitles Available	Total
Dummy-1001 Book	970	1,177
Dummy-AFC 1998	98	100
Dummy-AFC 2007	98	100
Dummy-Guardian List	156	179
Dummy-SS 2012	39	52
Dummy-WGA 2005	99	101
Number of Records	1,000	1,211

Visualization in Tableau

Limitations

Much of movie storytelling and moral issues are visual, told in pictures

Subtitle transcription quality varies, especially for dialogue that is translated

Next Steps

The recommendation tool/viz is posted on Tableau Public at https://public.tableau.com/profile/owen.temple

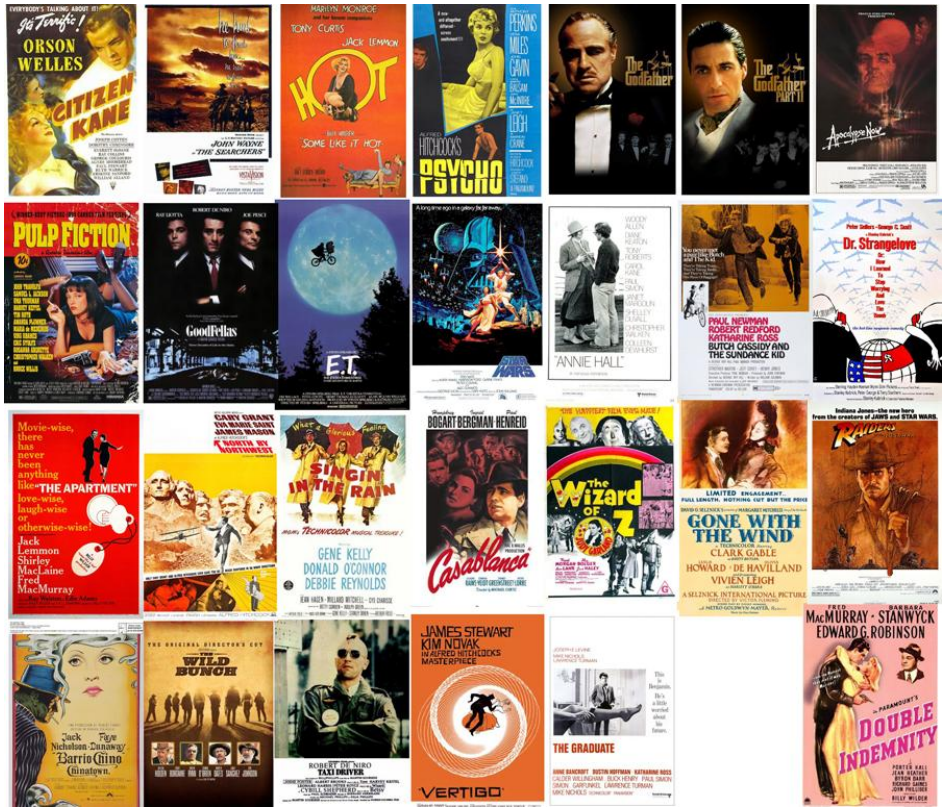
Look for other trusted greatest films lists to integrate

Begin testing recommendations firsthand

Greatest Movie Movie Quote

Present

C3PO: “It’s nothing sir. Old data. Pay it no mind”

Recommendation Based on ‘Old Data’

Present