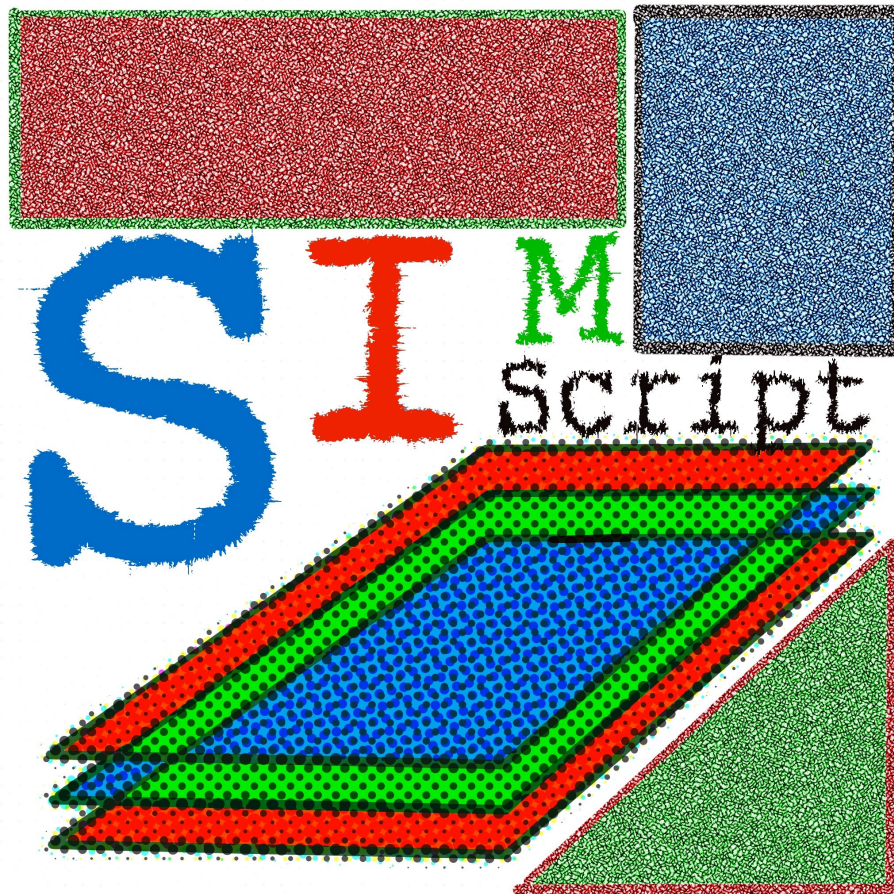




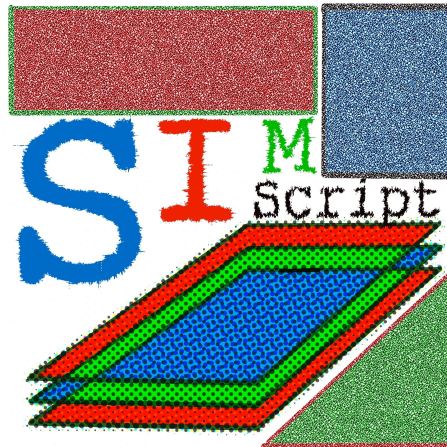
SIM_script

NLP analysis and choice assist for content providers of the
21st century

A digital tool for all „more-of-the-same“ industries

By
Michael Esser, PHD
Sergei Monakhov, PHD

Draft 2.00 / 02-25-2021



PITCH:

If you are a content provider and you want to keep your audience happy, don't try to improve the content creation with AI, improve the decision making with AI.

PITCH II

While you are waiting for AI to be able to write stories (probably never) why don't you use AI for what it's best at, looking for patterns and ranking material?

SIM_script

NLP analysis and choice assist for content providers of the 21st century

A digital tool for all „more-of-the-same“ industries

By
Michael Esser, PHD
Sergei Monakhov, PHD

Draft 2.00 / 02-25-2021

THE NEED

Actors in the entertainment industry are not known for conformity. Yet, there is a rare match among them: they all acknowledge that a script is the most important element in the success of a TV show or a feature film.

The hunger for content in the entertainment industry increases, new outlets enter the scene with no end in sight, and the competition for eyeballs gets tougher every day.

The entertainment industry will always strive for innovation and new ideas, but its core business relies on an uninterrupted supply of "more of the same," product of a defined type such as a specific TV show or a movie genre, that have shown to be successful.

There can never be too many action and adventure dramas, romance comedies, mystery thrillers or horror movies. Fans of "Friends," "Breaking Bad," or "The Walking Dead" wish their series never ended. Horror Fans will never complain about an oversupply of new movies.

"More of the same" is a strong incentive for consumers to return to a specific outlet. Yet, methods to guarantee content consistency are very expensive, unreliable or both.

***SIM_script* enables content providers to reach and maintain content consistency.**

THE CONVENTIONAL APPROACH

Screenplay supply and demand work like in other industries: when a TV show has good ratings, it triggers the content provider to order further seasons. When a movie genre proves successful at the box office, the provider wants to offer audiences more of this genre.

In both cases the challenge for the content provider is to ensure new products contain all the elements that made their predecessors successful.

The industry has an array of strategies to ensure the output of writing processes is in line with such content requirements:

- ★ Development experts analyze, generate specifications, hire and brief writers.
- ★ Writers, and/or teams, produce scripts under supervision of specialized creators.
- ★ Producers finance a process of creating, analyzing, revising scripts, until desired result is achieved .

These strategies are intended to ensure:

- ★ that a movie has all the characteristics of a particular genre.
- ★ that a new episode/season of a TV show contains all the elements that viewers expect.

These processes are based a number of assumptions:

- ★ what exactly audiences expect of a genre/series
- ★ what exact elements define a genre/series
- ★ how a script must be modified to conform to genre / show rules

What makes this task complex and difficult to tackle is that:

- ★ Parameters can not be checked against objective values
- ★ Development experts "know it when they see it."
- ★ Analyses, feedback and revisions have a shaky track record when it comes to improving the output according to specific criteria.

THE ***SIM_script*** APPROACH

SIM_script helps meet the need for reliable content supply according to a defined set of criteria

SIM_script's approach is simple:

Based on a set of good samples *SIM_script* will find more of it.

SIM_script searches and maps linguistic patterns in a sample set of templates (best in class), and captures the emotional arc.

SIM_script compares those patterns with findings in test candidates. It ranks test candidates according to their fidelity to the established ideal patterns in the set of the best in class.

SIM_script does what algorithms do best: It parses through large bodies of data in search of patterns and gives statistical feedback according to frequency, positioning, or interrelatedness of these findings.

SIM_script does ***not*** substitute or replace the experts, let alone the creators.

THE PLAYERS

For his PhD thesis in Media Sciences at TU Berlin, Michael Esser has researched methods to effectively support screenwriters to achieve output of a consistent quality according to a set of specifications, such as genre, tone, format, and style. The challenge of his three-year on/offline field experiment was to compare a set of strategies to deliver the best possible contribution to existing popular TV formats with minimal effort.

Sergei Monakhov, digital linguist at TU Jena, has developed an algorithm to detect social media postings by internet trolls through linguistic patterns. His NLP algorithm ranks social media posts according to the probability of this posts to be authored by trolls. The algorithm looks at the text only, not at metadata like address or pathway. The algorithm Sergei developed shows a high detection rate even on small numbers of samples.

Together, Sergei and Michael have realigned the troll detection tool. They have trained the algorithm to recognize patterns in a set of scripts. These sets can be the most successful scripts of a specific genre or scripts that belong to a specific TV show.

The resulting pattern library is then used to analyze and rank new material hitherto unknown to the algorithm against the best-of-its-class examples.

The result: ***SIM_script***

- ★ ***SIM_script*** develops parameters typical for a given genre or show format.
- ★ ***SIM_script*** evaluates new content on the basis of those linguistic patterns.
- ★ ***SIM_script*** ranks new content according to the evaluation and thus creates meaningful results.

THE THOUGHT PROCESS

Genre is where entertainment products (a movie, a TV show) meet the consumer's expectations. A genre movie has to contain certain elements that define the genre. Episodes of a TV series are a variant of the initial episode; some elements need to remain consistent.

In a highly developed industry, where hundreds of new products enter the market each day competing for eyeballs, genre or format are key in the sales pitch to the audience. The precise match of a genre or format with the consumer expectation is more important for their purchase decision than the cast.

Content producers and distributors who want to replicate a successful genre or add a new season to a TV series need to deliver "more of the same" to the consumer.

***SIM_script* helps content providers to make smart and informed decisions about future offers.**

Better decision making in content choice increases consumer satisfaction and loyalty.

SIM_script lives alongside the traditional approach analyzing or evaluating style, tone, structure, relevance of a given piece.

SIM_script gives decision makers the additional layer of insight into a content's specific genre affiliation that can make the big difference.

SIM_script complements the traditional approach, providing a ranked comparison according to a desired outcome

SIM_script delivers an assessment in a much shorter time, and with more fact based results.

SIM_script IN ACTION

Movies

- ★ ***SIM_script*** takes as input a set of samples of the most successful movies of the given genre: Best-of-its-class
- ★ ***SIM_script***'s analyses test candidates against the Best-of-its class.
- ★ ***SIM_script*** ranks test candidates according to their genre loyalty against the Best-of-its-class

TV series

- ★ ***SIM_script*** takes as input a set of samples of the most successful episodes of a TV Show.
- ★ ***SIM_script*** analyses test candidates against the Best-of-the-past.
- ★ ***SIM_script*** ranks test candidates according to their format loyalty against the Best-of-the-Past

HOW **SIM_script** WORKS

A screenplay is a set of instructions for a performance, written in simplified and codified language. The constituent elements, action description, character description, object description and dialogue, refer to a number of elements--characters and objects--and give instructions--dialogue, action, interaction--how these elements will relate. A scene as a subordinate order quantity puts one or more known elements in context with one or more new elements.

SIM_script divides screenplays into discrete chunks of text and looks for patterns across those constituent elements which are quantitatively measurable. These measures are to be interpreted as a measure of *topic relatedness* and *topic development*.

With a sitcom like "The Big Bang Theory" (see below: PROOF OF CONCEPT), for example by judging a certain episode or season good or bad, we assess whether it constitutes a holistic theatrical piece following in its unravelling some inherent logic of evolution, or if it is just a series of tediously repetitive or randomly combined scenes.

On the basis of the analysis of a number of "perfect" examples of a genre or a format and the comparison of new, unknown material, **SIM_script** ranks a screenplay according to the degree to which it reflects the ideal pattern.

What the algorithm really does (and what its limitations are) is to calculate a quantitative measure, for example of how target (most frequent) words are diluted with different context fillers in a given chunk of text.

USE CASES FOR ***SIM_script***

- ★ A ***content provider*** intends to produce a movie in a specific genre and has a number of offers on the table.
They test the offers against a set of Best-of-its-class movies in that genre
- ★ A ***streaming service*** needs to fill the gap of a TV series or movie franchise that has ended.
They test the offers against a simple set of Best-of-the-past episodes or movies
- ★ A ***TV network*** intends to produce a new season of a successful TV series.
They test the offers against a set of the Best-of-the-past episodes

PROOF OF CONCEPT

In order to prove ***SIM_script***'s ability to effectively rank screenplays and episodic scripts, we have undertaken a set of two tests. GENRE and SERIES

GENRE

- ★ Horror Movies

SERIES

- ★ "The Big Bang Theory"
- ★ "Friends"

The Test Setup:

- ★ Load text corpus (template material) of the genre or format
- ★ ***SIM_script*** produces an internal set of parameters and patterns

- ★ Load material to be compared to the defined probe
- ★ **SIM script** compares the corpus and the parable according to defined criteria.
- ★ **SIM script** produces a ranked result of probe's adherence to corpus

TEST CASE 1 GENRE

DETAILED TEST SETUP

We use a unique filtered dataset of 121 full-length horror movie scripts as training data and 15 movies as test data.

We use only **original screenplays** as were written by the author and submitted to evaluation by producers etc. The internet is full of other text datasets related to movies and series, for example subtitle files or so called movie transcripts, which are used for the process of dubbing. We do not consider subtitle files or dubbing transcripts valid entries for the data set because they are not used in the industry to evaluate content and/or as a basis for decision making.

To become part of the final dataset, a movie had to be a **financial success**. We measure success of a motion picture by box office revenues minus production budget.

The dataset for this project is compiled from several sources. We harvested screenplays from various sources on the Internet (see appendix A).

We used <https://www.the-numbers.com> for data about movie revenues and production budgets. (For a sample chart see appendix B)

The dataset of original screenplays was matched with the data extracted from <https://www.the-numbers.com> on revenues and budgets. <https://www.the-numbers.com> provides estimated production budget, domestic gross revenue (in the US), and worldwide gross revenue. (all in US dollars)

For a screenplay to become part of the dataset the result of the worldwide gross revenue minus the production budget had to be a positive number.

Worldwide gross revenue was available for all movies. Production budget estimates and worldwide gross revenues were available for all movies.

For completion of information the resulting dataset was cross-checked and complimented with data as listed on <https://www.imdb.com>.

This information included:

- IMDb genre definition of the movie (only movies that were **identified as belonging to the horror genre** were considered)
- date of release
- average IMDb user satisfaction rating from 1 (very bad) to 10 (excellent);
- critics satisfaction meta score from 0 (very bad) to 100 (excellent);
- rating count (number of individual assessments contributing to IMDb rating)
- number of awards (Oscars and other awards);
- name of the motion picture director
- runtime in minutes
- age appropriateness rating.

This generated a dataset like the following (see Table 1)

The dataset was then analyzed by **SIM_script** to create mapping, capture of the emotional arc etc in the best-of-its-class.

In the next step, we used scripts that were not previously known to **SIM_script** and compared them to the data generated on the basis of the best-of-its-class scripts.

SIM_script created a ranking of the new material according to mapping, emotional arch for the specific genre of horror, sentiment analysis, genre-skewed language, and how characters interact with each other.

The cross referencing of revenue/budget allowed for an additional unique opportunity to measure **probability of success**:

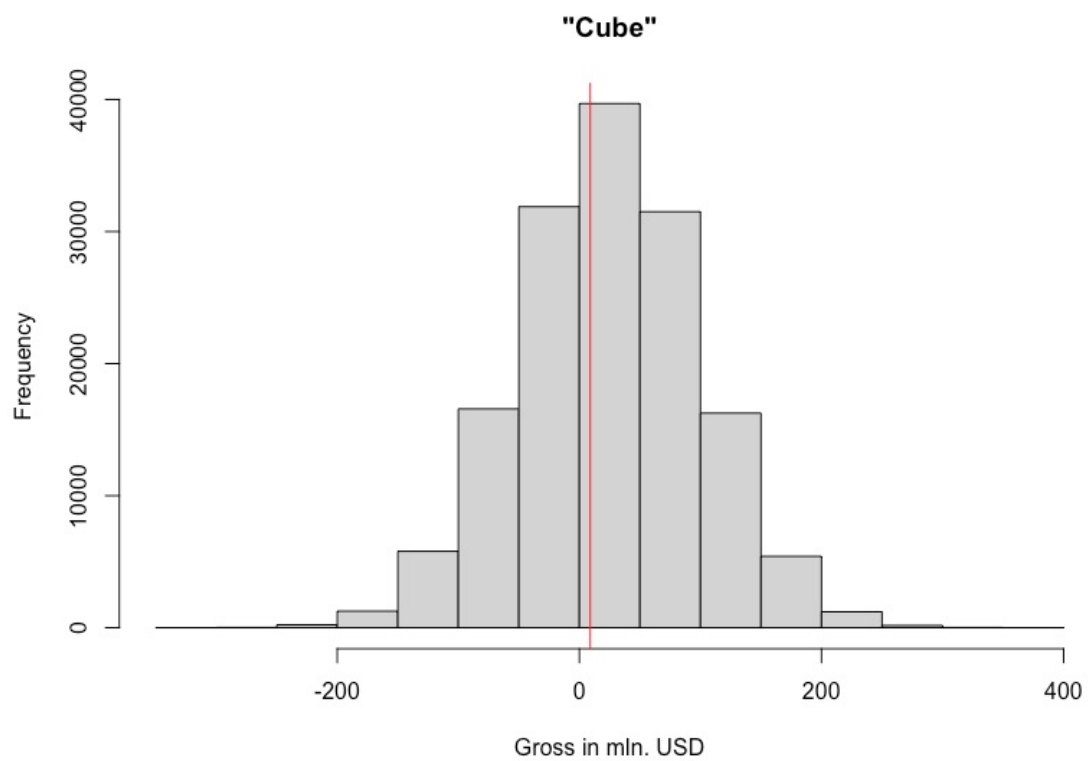
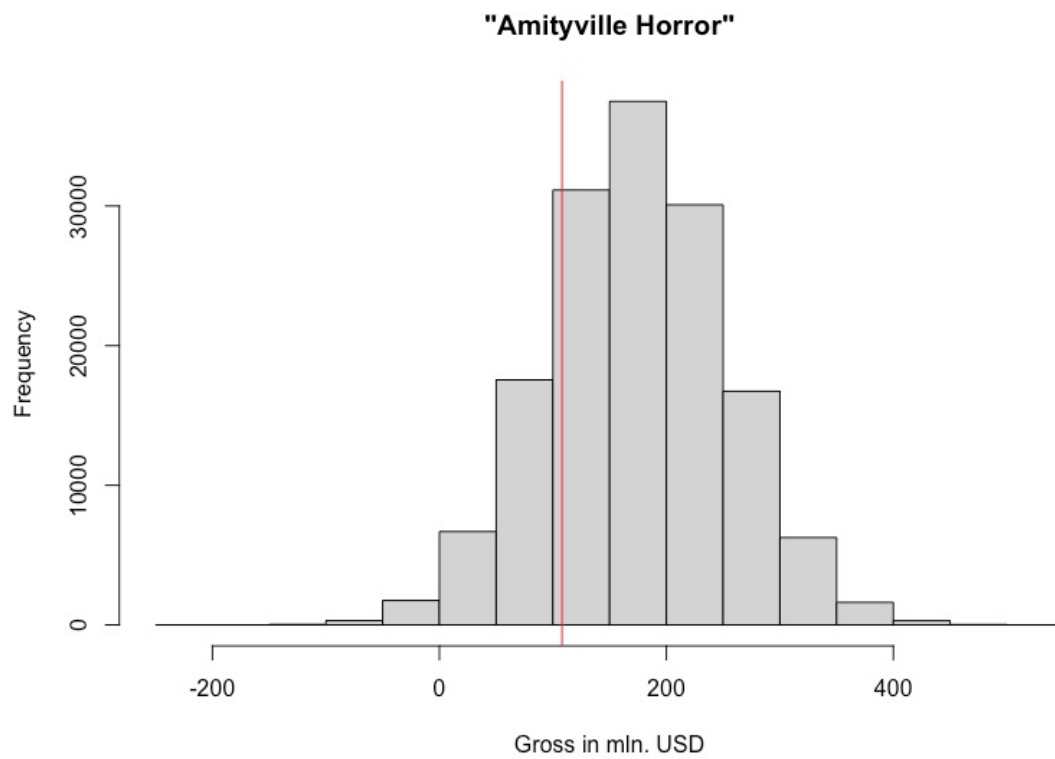
SIM_script not only ranked test candidate scripts provided for analysis in order of their probable success. As an additional outcome, ***SIM_script*** also generated a forecast of probable revenue against the investment of a set amount of dollar.

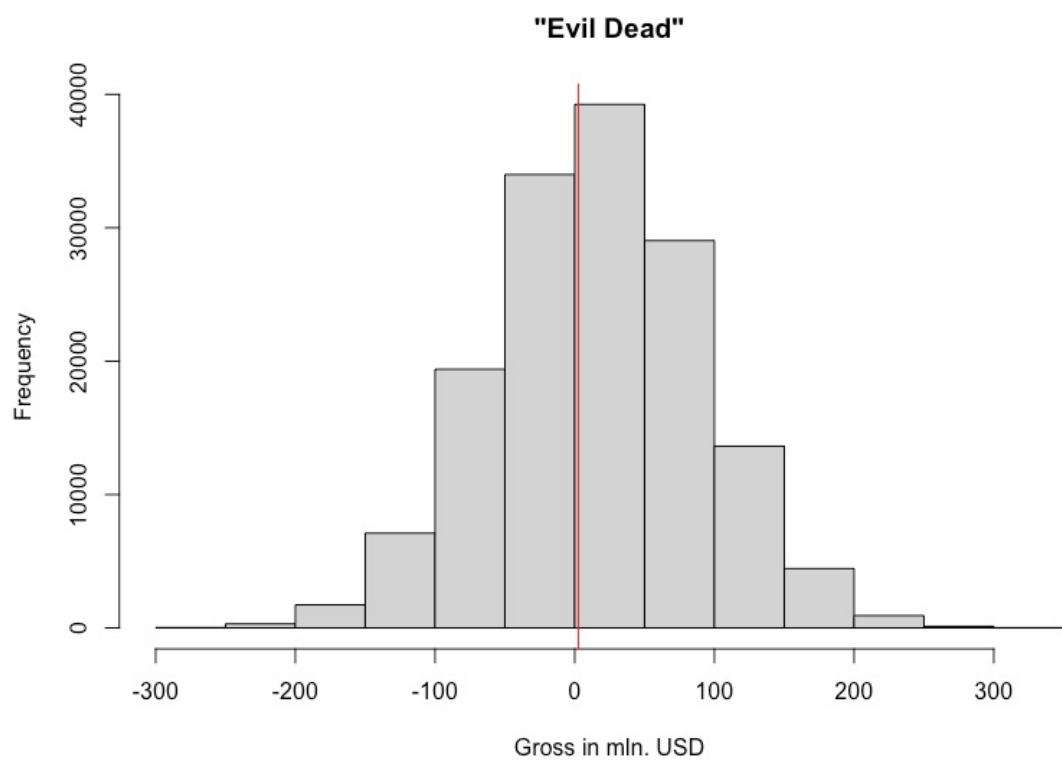
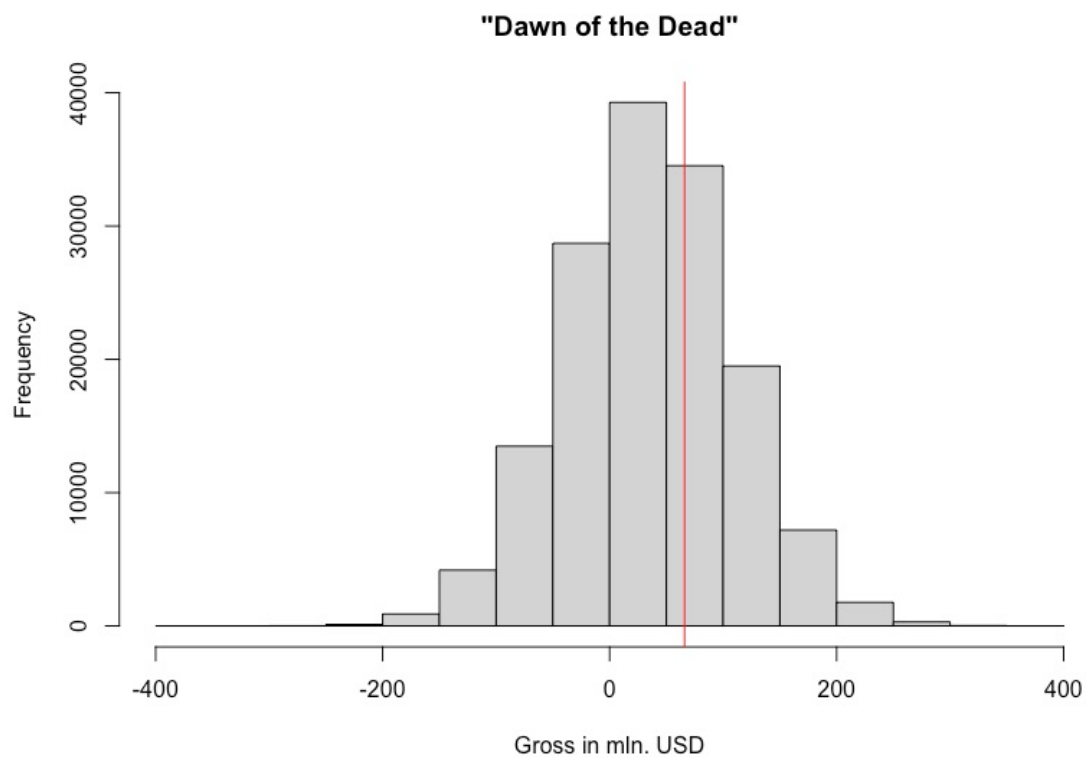
Table 1							
title	rating	budget	gross	profit	script coefficient	prob gross > 50 mln	accuracy
Alien 1	0.98	11	203.6	192.6	-2.094	0.53	1
Alien 2	0.97	18.5	183.3	164.8	3.272	0.88	1
Alien 3	0.43	50	159.8	109.8	-1.121	0.9	1
Alien 4	0.54	70	161.3	91.3	2.727	0.99	1
Alien 5	0.73	130	403.3	273.3	5.86	1	1
Alien 6	0.65	111	240.8	129.8	2.039	1	1
Alien vs Predator	0.21	70	177.4	107.4	-2.828	0.93	1
Jaws	0.98	9	472	463	-0.276	0.64	1
Jaws 2	0.58	30	208	178	-0.674	0.8	1
Jaws 3D	0.12	18	88	70	-2.037	0.61	1
Jaws the Revenge	0	23	51.9	28.9	-4.739	0.48	0
Nightmare on Elm Street 1	0.94	1.8	57	55.2	-2.577	0.4	0
Nightmare on Elm Street 2	0.43	3	30	27	-3.149	0.37	1
Nightmare on Elm Street 3	0.72	4.6	44.8	40.2	0.304	0.63	0
Nightmare on Elm Street 4	0.53	6.5	49.4	42.9	-2.402	0.46	1
Nightmare on Elm Street 5	0.31	8	22.1	14.1	-2.614	0.46	1
Nightmare on Elm Street 6	0.19	11	34.8	23.8	-2.132	0.53	0
Nightmare on Elm Street 7	0.8	8	19.7	11.7	-2.639	0.46	1
Scream	0.79	15	173	158	-1.786	0.6	1
Scream 2	0.81	24	172.3	148.3	1.915	0.87	1
Scream 3	0.4	40	161.8	121.8	-1.98	0.81	1
Pet Sematary	0.51	11.5	57.5	46	-2.369	0.52	1
Pet Sematary II	0.22	8	17.1	9.1	-3.096	0.43	1

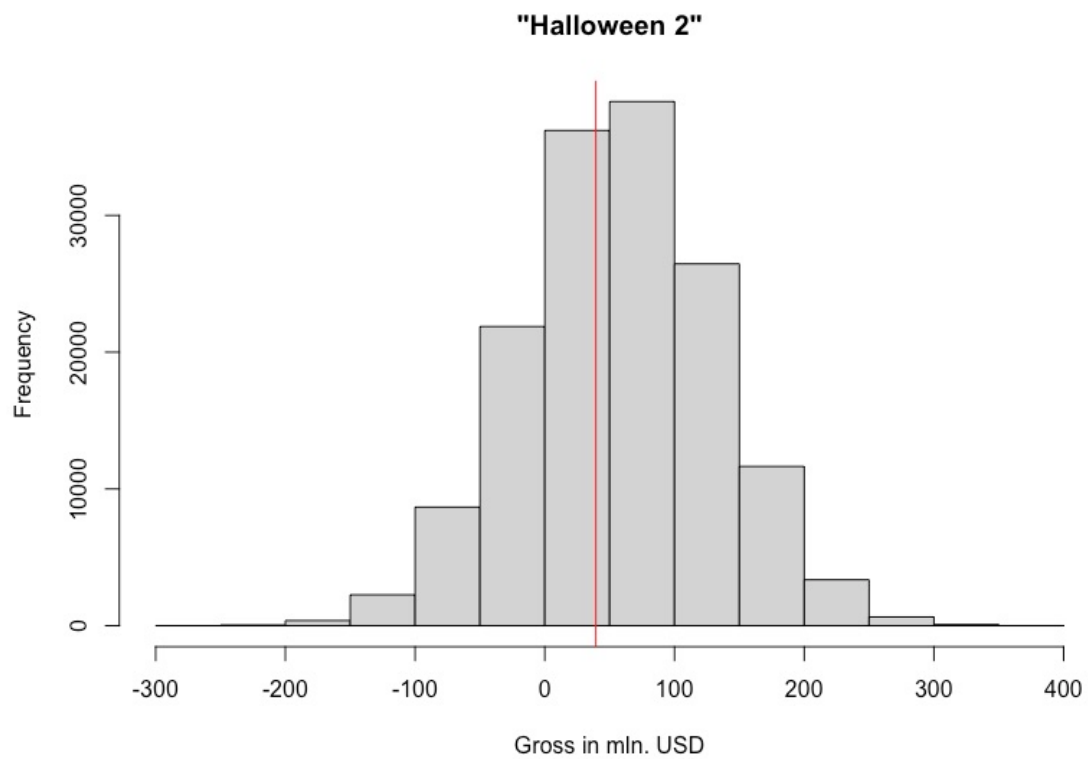
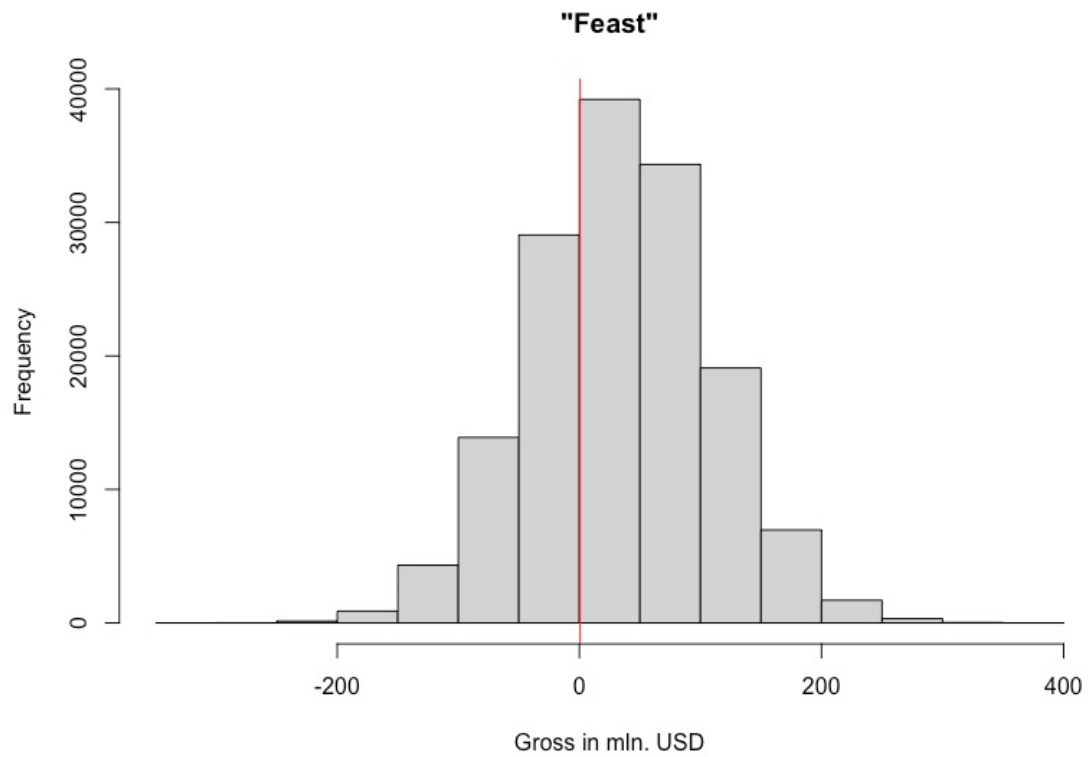
Labeling Table 1:

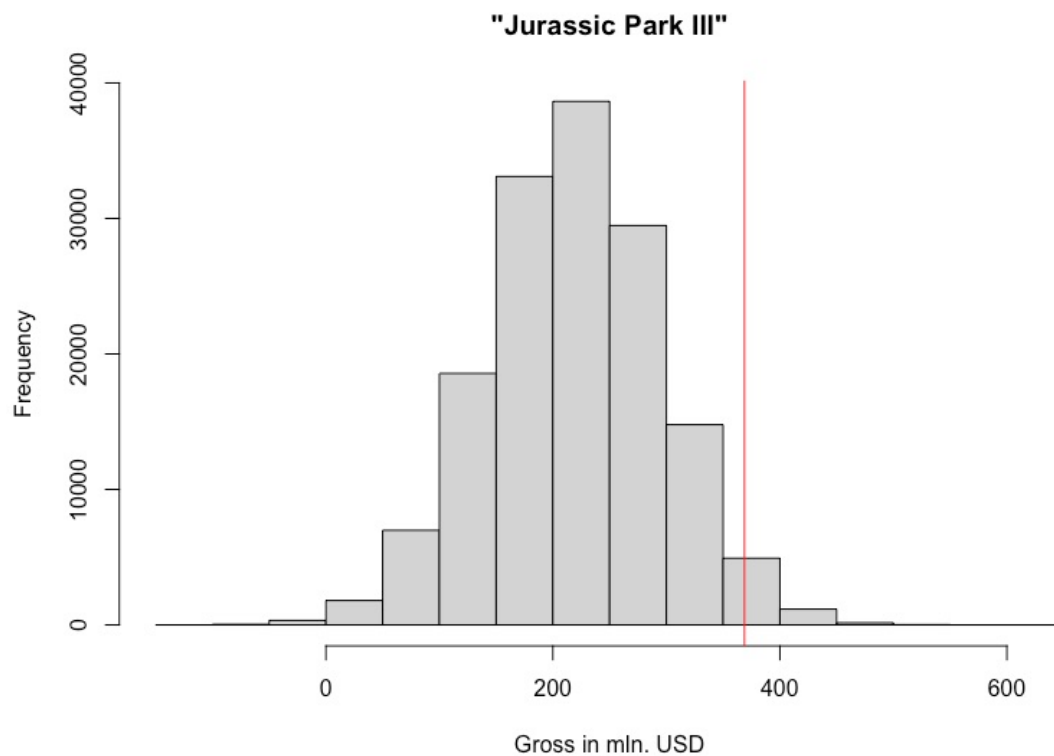
rating:	average Rotten Tomatoes user satisfaction rating from 1 (excellent) to 0 (bad). Average by count (number of individual assessments contributing to Rotten Tomatoes rating); number of user reviews, number of awards (Oscars and other awards)
budget:	budget for production of movie as reported on www.the-numbers.com . Budgets as reported on the-numbers.com include overall production costs + all costs incurred for P&A (Prints and Advertising)
gross:	gross worldwide revenue. This number includes all box office results as well as profits generated from sales of distribution and transmission rights
profit	= gross - budget
script_coefficient (-5 to 5):	is the coefficient indicating how much correlation SIM_script found between a given screenplay and the model/pattern created by its analysis of the best in class. A high coefficient indicates a high adherence to what SIM_script considers a good example of a horror movie screenplay
prob gross > 50 min:	indicates the probability SIM_script deducts that a screenplay will make more money than its production has cost. Every candidate script that is above a factor of 50 (>50) is considered a probable winner
accuracy (0 or 1):	where 0 means that SIM_script's prediction was not accurate: the movie has grossed either less than \$50M or more than \$50M while the model assigned more than 50% probability to the opposite outcome..

Illustrations 1-5 (below) show the results generated (*grey bars*) by SIM_script and the actual revenue the given movie generated. (*red line*)









RESULTS

SIM_script predicts with high probability (frequency) in the case of "Amityville Horror," "Cube," "Dawn of the Dead," "Evil Dead," "Feast", and "Halloween 2" a worldwide gross (*grey bar*) that is close to the actual outcome (*red line*).

In the case of "Jurassic Park" which we used as verification case, the actual worldwide gross (red line at about \$380M) was far beyond the prediction of ***SIM_script***.

Though the actual revenue was deemed highly improbable by our model, it had no doubt that the movie would be a huge commercial success: most of the probability mass in the histogram is centered around \$200+M. That is a result which is in alignment with ***SIM_script's*** prediction ability; the algorithm is trained on the specific genre horror and Jurassic Park is a mixed genre movie as it contains also elements of the genres Family Drama and Adventure.

TEST CASE 2: SERIES

A: BIG BANG THEORY

<i>Genre</i>	<i>Sitcom</i>
<i>Created by</i>	<i>Chuck Lorre, Bill Prady</i>
<i>No. of seasons</i>	<i>12</i>
<i>No. of episodes</i>	<i>279</i>
<i>Original network</i>	<i>CBS</i>
<i>Original release</i>	<i>September 24, 2007 – May 16, 2019</i>

We took the transcripts of the first 9 seasons of "The Big Bang Theory" and analyzed them just as we would analyze tweets, having made only a couple of tweaks to account for the different text structure and genre.

We were able to identify a number of episodes that were marked by the algorithm as 'troll-like', which, in our case, can be seen as a sign of their scripts' high quality

Ranking of scripts:

<u>No Episode</u>	<u>Season</u>	<u>Title</u>
1	1	"Pilot"
2	1	"The Big Bran Hypothesis"
9	1	"The Cooper-Hofstadter Polarization"
11	1	"The Pancake Batter Anomaly"
9	2	"The White Asparagus Triangulation"
21	2	"The Vegas Renormalization"
8	3	"The Adhesive Duck Deficiency"
11	3	"The Maternal Congruence"
13	3	"The Bozeman Reaction"
14	3	"The Einstein Approximation"
2	4	"The Cruciferous Vegetable Amplification"
3	4	"The Zazzy Substitution"
16	4	"The Cohabitation Formulation"
19	4	"The Zarnecki Incursion"
20	4	"The Herb Garden Germination"
21	4	"The Agreement Dissection"
24	4	"The Roommate Transmogrification"
12	5	"The Shiny Trinket Maneuver"
8	9	"The Mystery Date Observation"
22	9	"The Fermentation Bifurcation"

It was also possible to arrange all seasons in a hierarchy from best to worst according to the same criteria (the first number is the ordinal number of the season, the second one is the quantitative measure of its textual quality; plus or minus signs are the results of categorical classification whether the season as a whole is good enough):

Categorical

Classification Season Measure of Quality

(+)	4	1.0408163265306123,
(+)	1	0.5151515151515151,
(+)	2	0.4492753623188406,
(+)	3	0.2578616352201258,
(-)	7	0.14942528735632185,
(-)	8	0.12994350282485875,
(-)	9	0.11731843575418995,
(-)	6	0.04712041884816754,
(-)	5	0.0101010101010102

Overall, these results seem intuitively clear and compatible with the general opinion that after its fourth season, the show was on the decline.

We also analyzed the numbers of US viewers provided in Wikipedia.

With the 1st, the 3rd and the 4th season, the episodes marked by the algorithm as 'troll-like' attracted significantly greater audience than those marked as 'genuine'. (Source: Wikipedia)

Season	Timeslot (ET)	Episodes	First aired		Last aired		TV season	Viewership rank	Avg. viewers (millions)	18–49 rank	Avg. 18–49 rating
			Date	Viewers (millions)	Date	Viewers (millions)					
1	Monday 8:30pm (1–8) Monday 8:00pm (9–17)	17	September 24, 2007	9.52 ^[110]	May 19, 2008	7.34 ^[111]	2007–08	68	8.31	46	3.3/8 ^[112]
2	Monday 8:00 pm	23	September 22, 2008	9.36 ^[113]	May 11, 2009	9.81 ^[114]	2008–09	40	10.03	N/A	N/A
3	Monday 9:30 pm	23	September 21, 2009	12.96 ^[115]	May 24, 2010	14.78 ^[116]	2009–10	12	14.22	5	5.3/13 ^[117]
4		24	September 23, 2010	14.04 ^[118]	May 19, 2011	11.30 ^[119]	2010–11	13	13.21	7	4.4/13 ^[120]
5	Thursday 8:00pm	24	September 22, 2011	14.30 ^[121]	May 10, 2012	13.72 ^[122]	2011–12	8	15.82	6	5.5/17 ^[123]
6		24	September 27, 2012	15.66 ^[124]	May 16, 2013	15.48 ^[125]	2012–13	3	18.68	2	6.2/19 ^[126]
7		24	September 26, 2013	18.99 ^[127]	May 15, 2014	16.73 ^[128]	2013–14	2	19.96 ^[129]	2	6.2/20 ^[104]
8	Monday 8:00pm ^[a] Thursday 8:00pm	24	September 22, 2014	18.08 ^[135]	May 7, 2015	14.64 ^[136]	2014–15	2	19.05	4	5.6/17 ^[137]
9		24	September 21, 2015	18.20 ^[138]	May 12, 2016	14.73 ^[139]	2015–16	2	20.36	3	5.8/19 ^[140]
10		24	September 19, 2016	15.82 ^[141]	May 11, 2017	12.99 ^[142]	2016–17	2	18.99	3	4.9/19 ^[143]
11		24	September 25, 2017	17.65 ^[144]	May 10, 2018	15.51 ^[145]	2017–18	1	18.63	5	4.4 ^[146]
12		24	September 24, 2018	12.92 ^[147]	May 16, 2019	18.52 ^[148]	2018–19	2	17.31	6	3.6 ^[149]

B: FRIENDS

<i>Genre</i>	<i>Sitcom</i>
<i>Created by</i>	<i>David Crane, Marta Kauffman</i>
<i>No. of seasons</i>	<i>10</i>
<i>No. of episodes</i>	<i>236</i>
<i>Original network</i>	<i>NBC</i>
<i>Original release</i>	<i>September 22, 1994 – May 6, 2004</i>

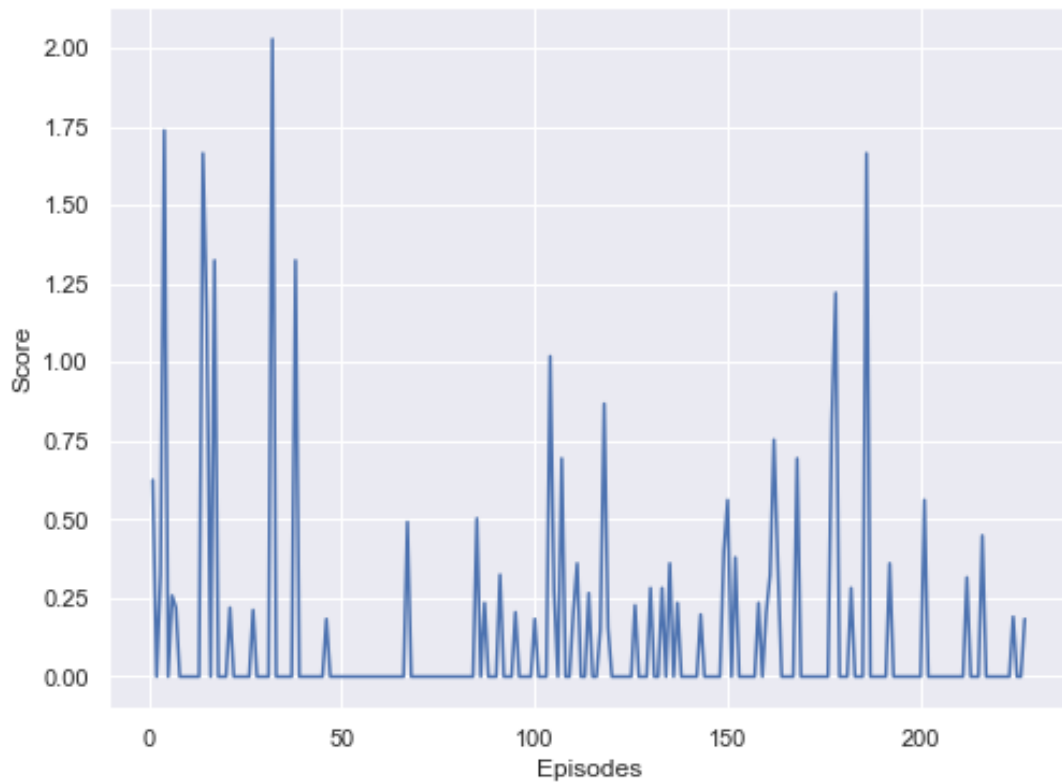
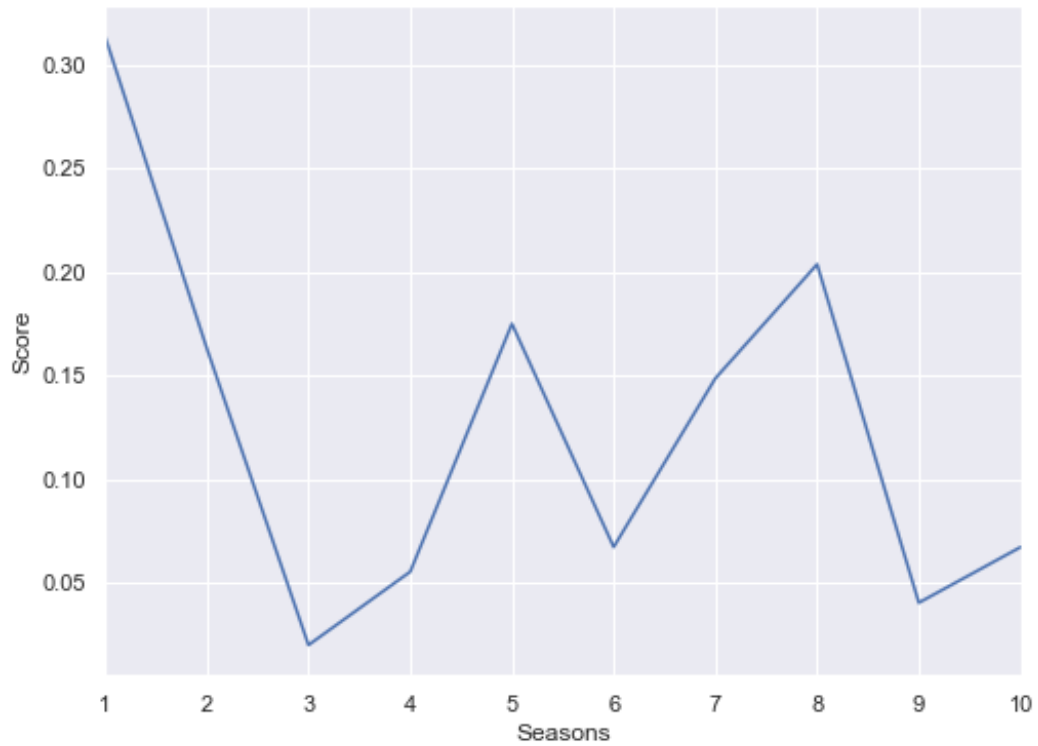
Repeating the same procedure for the TV series "Friends" shows results are largely compatible with what ratings and fan websites have shown:

- ★ Overall context analysis shows that the quality of the scripts in terms of character, dialogue and plot complexity is much better and much more even than in "The Big Bang Theory" (refer to graph "Score" which shows larger deflections --up to a score of 2.0 for the famous episode 44)
- ★ "Friends" best seasons in terms of adherence to the original concept are those in the middle (from season 5th to 8th). (refer to graph "Seasons")

The result makes sense considering that the writing of Friends was a true team effort, while Big Bang Theory was very much defined by its creator Chuck Lorre. **SIM_scripts'** finding confirms in a way that if a series is truly the outcome of a collaborative effort, the overall results tend to be better.

"Friends"

Charts Seasons + Episodes



Season Chart (Wikipedia) + Ranking of SIM-script

Season	Timeslot	Season premiere	Season finale	TV season	Rank	Viewers (in millions)	Most-watched episode	
							Title	Viewers (in millions)
1	Thursday 8:30 pm (1–16) Thursday 9:30 pm (17–24)	September 22, 1994	May 18, 1995	1994–95	8	24.3 ^[97]	"The One Where Rachel Finds Out"	31.3 ^[98]
2		September 21, 1995	May 16, 1996	1995–96	3	30 ^[97]	"The One After the Superbowl"	52.9 ^[98]
3		September 19, 1996	May 15, 1997	1996–97	4	24.9 ^[97]	"The One Where Chandler Can't Remember Which Sister"	29.8 ^[citation needed]
4		September 25, 1997	May 7, 1998	1997–98	4	24.0 ^[97]	"The One with Ross's Wedding"	31.6 ^[98]
5		September 24, 1998	May 20, 1999	1998–99	2	23.5 ^[97]	"The One After Ross Says Rachel"	30.9 ^[98]
6	Thursday 8:00 pm	September 23, 1999	May 18, 2000	1999–2000	5	20.7 ^[97]	"The One with the Proposal"	30.7 ^[98]
7		October 12, 2000	May 17, 2001	2000–01	5	20.2 ^[97]	"The One with Monica and Chandler's Wedding"	30.1 ^[98]
8		September 27, 2001	May 16, 2002	2001–02	1	24.5 ^[97]	"The One Where Rachel Has a Baby"	34.9 ^[98]
9		September 26, 2002	May 15, 2003	2002–03	2	21.8 ^[97]	"The One Where No One Proposes"	34.0 ^[98]
10		September 25, 2003	May 6, 2004	2003–04	4	22.8 ^[97]	"The Last One"	52.5 ^[4]

CONCLUSION

	Season	Summary	Stars	Votes	category	measure
31	2	With help from Chandler and Joey, Ross makes a list of pros and cons to decide whether	8.50	3581	troll	2.0303
3	1	Joey and Chandler take Ross to a hockey game to take his mind off the anniversary of the	8.10	4468	troll	1.7397
13	1	As Valentine's Day approaches; Ross and his date end up at the same restaurant as	8.30	3866	troll	1.6667
185	8	Monica and Phoebe bet on when Rachel's baby will be born as the due date comes and	8.30	2662	troll	1.6667
16	1	Rachel switches identities with Monica so she can use her health insurance. Meanwhile,	8.50	3883	troll	1.3256
37	2	Monica becomes infatuated with a friend of her parents when she caters a party for him.	8.90	3928	troll	1.3256
177	8	Chandler's very curious about a locked closet, but Monica won't tell him what is inside.	8.20	2654	troll	1.2222
14	1	Monica cooks a gourmet meal for a restaurateur who is looking for a new chef, however	8.20	3915	troll	1.1505
103	5	Joey finds himself constantly covering for Chandler and Monica's affair. Ross becomes	9.10	3913	troll	1.0202
117	5	The gang go to Las Vegas to visit Joey. Chandler is angry when he finds out Monica had	8.90	3504	troll	0.8692
176	8	Monica turns Chandler on to the concept of taking a bath to relax; Ross and Rachel toss	8.60	2800	troll	0.8182
161	7	Monica and Chandler are getting married in four weeks and have to write their vows, but	7.50	2832	troll	0.7544
106	5	Monica becomes annoyed at how Chandler sucks up to his boss by mimicking his boss's	8.30	2949	troll	0.6949
167	8	Ross and Rachel tell everyone about the night they were together. They have two different	9.30	4114	troll	0.6949
0	1	NA	8.30	4148	troll	0.6260
149	7	Monica makes candy to get to know her neighbours, Rachel and Tag try to hide their	8.10	2753	troll	0.5625
200	9	Phoebe and Joey purposely set Ross and Rachel up on bad dates in the hopes they'll get	8.20	2642	troll	0.5625
84	4	After seeing Kathy's play, Chandler becomes convinced she is cheating on him with her	8.20	3006	genuine	0.5038
66	3	Monica's aunt dies and leaves her an antique dollhouse that she doesn't want Phoebe to	8.10	3134	troll	0.4925
215	10	Ross applies for a paleontology grant and finds out Charlie's ex-boyfriend reviews the	8.20	2853	troll	0.4493
148	7	Ross struggles to name all 50 states in a game. Joey compromises Rachel's chances with	8.20	2772	genuine	0.3986
151	7	Chandler refuses to give back a cheesecake accidentally delivered to his apartment;	8.60	2936	troll	0.3793
110	5	Joey has a dream about Monica and becomes convinced he is in love with her.	8.60	3174	troll	0.3605
134	6	Ross jeopardizes his career when he starts dating a student from one of his	8.30	2870	troll	0.3605
162	7	Ross and Rachel hit the freeway together when Monica's Porsche shows up again;	8.40	2822	genuine	0.3605
191	9	Phoebe will celebrate her birthday by having a dinner with her friends at a fancy	8.50	2799	genuine	0.3605
2	1	Monica becomes irritated when everyone likes her new boyfriend more than she does.	8.20	4605	troll	0.3245
90	4	Ross proposes to Emily. Monica and Rachel try to win their apartment back from Chandler	8.70	3168	troll	0.3245
160	7	Rachel's college friend can't remember a scandalous kiss from the past; Chandler and	8.40	2894	troll	0.3245
211	10	Rachel and Joey fear for Ross's state of mind after he finds out about their relationship.	8.80	3366	genuine	0.3158
129	6	Chandler and Ross argue over a joke. Meanwhile, Joey reluctantly takes a waiter job at	8.10	2829	genuine	0.2821
132	6	When Rachel hears Barry and Mindy are getting divorced, the gang wonder how different	8.50	3067	troll	0.2821
181	8	Everyone prepares for the Gellers' 35th wedding anniversary in Long Island; Monica	8.20	2699	genuine	0.2821
104	5	A bored Ross wreaks havoc in Chandler and Joey's lives. Rachel is uncomfortable with	8.20	2680	troll	0.2720

The value of such an analysis tool is evident: **SIM_script** enables content providers faced with the task to satisfy audiences' expectations with "more of the same," material that fulfills this audience's expressed content preferences, is now able to select among a number of possibilities and get a recommendation that reflects on the closeness of the content to the audiences expressed preference.

NEXT STEPS

We will seek collaboration with a screenwriting consultancy and competition platform.

They will provide anonymized feature film screenplays which have been submitted either to a script competitions or to a script review service.

The scripts need to be classified by the professional readers to be of the genre horror.

All scripts have been independently assessed and rated by professional script readers. This produces a collection of scores for each script, resulting in an overall Review Score.

We run the scripts through **SIM_script**'s algorithm, and will pull out data from each script, including sentiment analysis, genre-skewed language, and how characters interact with each other. (see also Procedure Analysis) (See Timeline, APPENDIX C)

VERIFYING RESULTS

1. In a first step the data generated by **SIM_script** (a ranking of the analyzed screenplays) will be compared to the score given by the script readers.

The degree of matching between **SIM_script**'s scoring and the scoring by the professional readers will be a first indicator to verify the results generated by **SIM_script**.

This verification is significant because readers are widely used in the industry and their judgement is a significant gatekeeper for a script to enter into further consideration rounds or not.

2. In a second step we will closely monitor the further development regarding specific scripts (those ranked highest by SIM_script) and evaluate the correlation between scripts that proceed to further rounds of consideration.

These rounds are usually

- author finds representation
- screenplay is being offered by a reputable agency
- screenplay/writer find interest by production or distribution companies
- production/distribution attach talent (cast director, other writers)
- production/distribution find financing
- movie/series enter in pre-production

The first three steps (representation, offering by agency, interest by production company) are often completed in a relative short time period (1-12 months). Within 12 months step two of the verification of significant indicators about the viability of SIM_script's results will be concluded.

3. In a third step we will present the results of step one and two to a content provider (production, distribution, streamer, network). We will ask them to provide us with anonymized material submitted to them for consideration.

This will either be done in the specific genre horror, or, if the content provider is considering movies/series of a different genre, we will retrain **SIM_script** on that specific genre and conduct the next steps on the genre indicated by the content provider.

Parallel to the traditional evaluation process that takes place within the content provider company or organization, we will run the material through **SIM_script**'s algorithm and provide the company with the findings.

Next, the results of the in-house evaluation and SIM_script's results are compared.

Finally, **SIM_script**'s results will be compared with the product, movie or show.

Within 12-24 months step three of the verification of significant indicators about the viability of ***SIM_script***'s results will be concluded.

LOS ANGELES
JENA
MARCH 2021

APPENDIX A

The Screenplay downloads

[The Internet Movie Script Database \(IMSDb\)](#)
[50 Best Screenplays To Read And Download In Every Genre](#)
[Browse the Best Free Movie Scripts and PDFs | Screenplay Database](#)
[movie scripts: drew's meat and potatoes](#)
[SimplyScripts - Complete Listing of Movie Scripts and Screenplays on the 'Net](#)
[Awesome Movie Scripts and Screenplays](#)
[Daily Script - Movie Scripts and Movie Screenplays - Movie Script A to M](#)
[All Movie Scripts at The Screenplay Database](#)
[Movie Scripts and Screenplays Web Site](#)
[99 Drama Scripts That Screenwriters Can Download and Study - ScreenCraft](#)
[Free Movie Screenplays A-M – The Script Savant](#)
[Index of /scripts](#)
[Full List - Script PDF](#)
[Screenplay Download: The Top Ten Horror Screenplays - Bulletproof Screenwriting™](#)
[Script Downloads Archives - The Script Lab](#)
[Free Screenplays - Learn How To Be A Screenwriter](#)
[Screenplay Library](#)
[2018 Screenplays For Your Consideration - An Ongoing List | LA Screenwriter](#)
[Download 70 Screenplays Legally and Free - ScreenCraft](#)
[Scripts On Screen - The Movie Script Search Engine](#)
[Free Movie Scripts for Actors and Writers](#)
[Script Library - Writersroom](#)
[65 TV Pilot Scripts That Screenwriters Should Study - ScreenCraft](#)
[Script City | Reel Hollywood Script City - Movie Scripts || TV Scripts || Treatments](#)
[All Movie Scripts, Screenplays, TV Scripts, Screenwriting-Script Fly](#)
[Film Scripts, Screenplays & Images - Film Studies - LibGuides at Barnard College](#)
[HorrorLair - Horror and Suspense Movie Scripts](#)
[Scripts Archive - David Koepp](#)
[SimplyScripts » The Archives](#)
[Screenplays & Scripts - Film and Television - Research Guides at UCLA Library](#)
[Welcome to Download Screenplays | The Scribe Writers Room](#)
[Free Screenplays Archives](#)
<https://s3-us-west-2.amazonaws.com/screenplays-pdf/Movie-Screenplay-PDFs/Get-out-.pdf>
[Daily Script - Movie Scripts and Movie Screenplays - Movie Script A to M](#)
[All Movie Scripts at The Screenplay Database](#)
[Ridley Scott Film's Screenplay PDF Collection \(Download\) - Indie Film Hustle®](#)
[Home - Film Scripts Online - LibGuides at ProQuest](#)

APPENDIX B

The Numbers Chart Sample

Movie Report

[← Return to Report Builder](#)

Active criteria

Genre: Horror

Max Budget (mUSD): 400

Min Release Year: 1975

Max Release Year: 2021

Min Runtime (min.): 75

Released	Title	Production Budget	Domestic Box Office	Worldwide Box Office
1 Feb 12, 2010	The Wolfman	\$150,000,000	\$62,189,884	\$142,634,358
2 Dec 14, 2007	I am Legend	\$150,000,000	\$256,393,010	\$585,532,684
3 May 19, 2017	Alien: Covenant	\$97,000,000	\$74,262,031	\$238,521,247
4 Feb 9, 2001	Hannibal	\$87,000,000	\$165,092,266	\$350,100,280
5 Jul 23, 1999	The Haunting	\$80,000,000	\$91,188,905	\$180,188,905
6 Aug 20, 2004	Exorcist: The Beginning	\$78,000,000	\$41,814,863	\$43,957,541
7 Feb 13, 1998	Sphere	\$73,000,000	\$37,068,294	\$50,168,294
8 Jun 17, 1994	Wolf	\$70,000,000	\$65,011,757	\$131,011,757
9 Sep 6, 2019	It: Chapter Two	\$70,000,000	\$211,593,228	\$469,566,806
10 Nov 19, 1999	Sleepy Hollow	\$70,000,000	\$101,068,340	\$207,068,340
11 Jun 22, 2012	Abraham Lincoln: Vampire Hunter	\$67,500,000	\$37,519,139	\$137,489,730
12 Aug 28, 2020	The New Mutants	\$67,000,000	\$23,855,569	\$44,614,898
13 Sep 14, 2012	Resident Evil: Retribution	\$65,000,000	\$42,345,531	\$240,647,629
14 Jul 28, 1999	Deep Blue Sea	\$60,000,000	\$73,648,228	\$165,048,228
15 May 13, 2011	Priest	\$60,000,000	\$29,136,626	\$84,154,026
16 Aug 15, 1997	Event Horizon	\$60,000,000	\$26,673,242	\$26,673,953
17 Oct 16, 2015	Goosebumps	\$58,000,000	\$80,069,458	\$158,893,300
18 Sep 10, 2010	Resident Evil: Afterlife	\$57,500,000	\$60,128,566	\$295,874,190
19 Oct 17, 1997	Devil's Advocate	\$57,000,000	\$61,007,424	\$153,007,424
20 Sep 20, 2002	Eye See You	\$55,000,000	\$79,161	\$1,807,990
21 Oct 16, 2015	Crimson Peak	\$55,000,000	\$31,090,320	\$74,966,854
22 Apr 6, 2007	Grindhouse	\$53,000,000	\$25,031,037	\$50,187,789
23 Apr 21, 2006	Silent Hill	\$50,000,000	\$46,982,632	\$94,704,227
24 Mar 18, 2005	The Ring Two	\$50,000,000	\$75,941,727	\$161,941,727
25 Sep 30, 2011	Dream House	\$50,000,000	\$21,302,340	\$41,642,166
26 Nov 11, 1994	Interview with the Vampire	\$50,000,000	\$105,264,608	\$223,564,608
27 Sep 10, 2004	Resident Evil: Apocalypse	\$50,000,000	\$50,740,078	\$125,168,734
28 Apr 12, 2019	Hellboy	\$50,000,000	\$21,903,748	\$53,956,480

APPENDIX C

The Timeline

