Data Science for Business
Lecture #4
Movie Scheduling Solution

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

Movie Scheduling Assignment

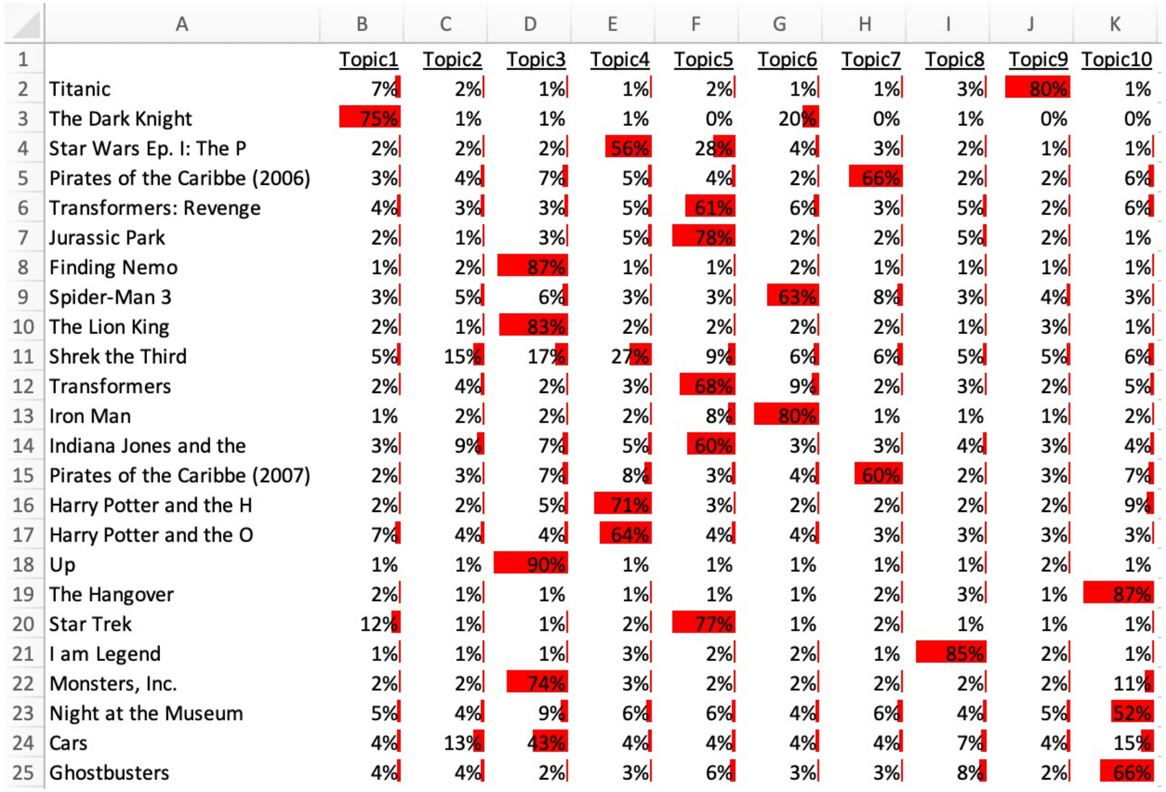
Background: Gotham Group (a movie studio) wants to release a new action movie called “The Maze Runner”. Traditionally movie studios use heuristics for scheduling releases. Namely, a rule of thumb is to not release two movies that are too similar to one another.

The problem statement: Use probabilistic clustering with the LDA algorithm to segment movies and use these segments to understand the movie marketplace. What would you recommend to your Gotham Group about what week to release?

Proposed Solution

1. Construct an LDA model to compare all movies based upon their keywords (we have 1153 movies x 962 terms)
2. Compute the "Euclidean distance" between movies based upon their "topics"
3. Compare our target movie "The Maze Runner" to every potential weekly launch date in 2014 (assume all movies released that week and the previous week are the set of competitors that we care about). Our recommendation will be to choose the week in which we have the "weakest" competition (e.g., choose the week where our release looks as different as possible from other films)

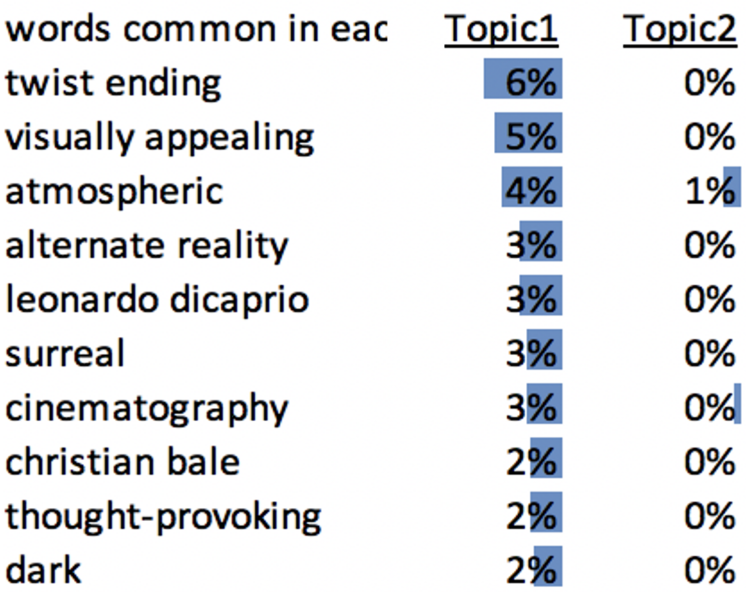
LDA Topic Scores Associated with Each Movie



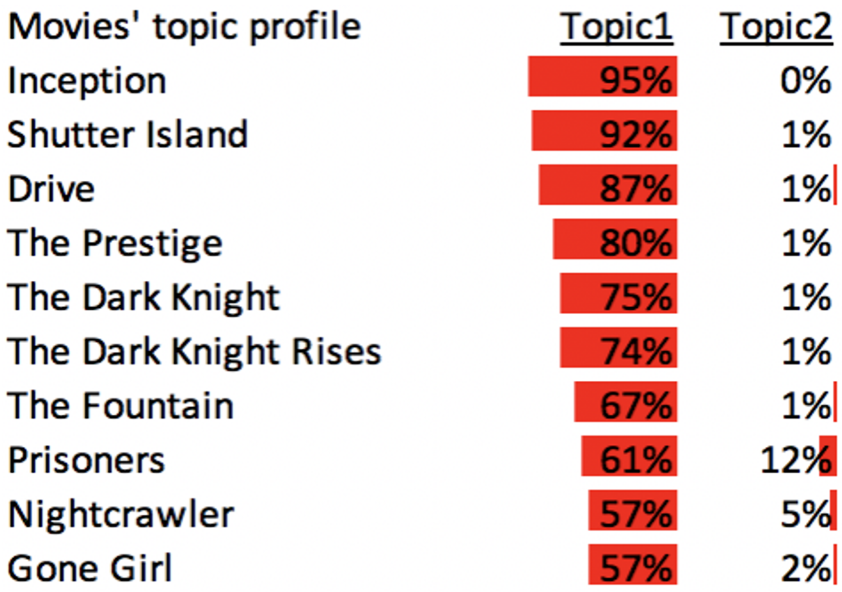
	A	B	C	D	E	F	G	H	I	J	K
1		Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
2	Titanic	7%	2%	1%	1%	2%	1%	1%	3%	80%	1%
3	The Dark Knight	75%	1%	1%	1%	0%	20%	0%	1%	0%	0%
4	Star Wars Ep. I: The P	2%	2%	2%	56%	28%	4%	3%	2%	1%	1%
5	Pirates of the Caribbe (2006)	3%	4%	7%	5%	4%	2%	66%	2%	2%	6%
6	Transformers: Revenge	4%	3%	3%	5%	61%	6%	3%	5%	2%	6%
7	Jurassic Park	2%	1%	3%	5%	78%	2%	2%	5%	2%	1%
8	Finding Nemo	1%	2%	87%	1%	1%	2%	1%	1%	1%	1%
9	Spider-Man 3	3%	5%	6%	3%	3%	63%	8%	3%	4%	3%
10	The Lion King	2%	1%	83%	2%	2%	2%	2%	1%	3%	1%
11	Shrek the Third	5%	15%	17%	27%	9%	6%	6%	5%	5%	6%
12	Transformers	2%	4%	2%	3%	68%	9%	2%	3%	2%	5%
13	Iron Man	1%	2%	2%	2%	8%	80%	1%	1%	1%	2%
14	Indiana Jones and the	3%	9%	7%	5%	60%	3%	3%	4%	3%	4%
15	Pirates of the Caribbe (2007)	2%	3%	7%	8%	3%	4%	60%	2%	3%	7%
16	Harry Potter and the H	2%	2%	5%	71%	3%	2%	2%	2%	2%	9%
17	Harry Potter and the O	7%	4%	4%	64%	4%	4%	3%	3%	3%	3%
18	Up	1%	1%	90%	1%	1%	1%	1%	1%	2%	1%
19	The Hangover	2%	1%	1%	1%	1%	1%	2%	3%	1%	87%
20	Star Trek	12%	1%	1%	2%	77%	1%	2%	1%	1%	1%
21	I am Legend	1%	1%	1%	3%	2%	2%	1%	85%	2%	1%
22	Monsters, Inc.	2%	2%	74%	3%	2%	2%	2%	2%	2%	11%
23	Night at the Museum	5%	4%	9%	6%	6%	4%	6%	4%	5%	52%
24	Cars	4%	13%	43%	4%	4%	4%	4%	7%	4%	15%
25	Ghostbusters	4%	4%	2%	3%	6%	3%	3%	8%	2%	66%

See “topic_allmovies.txt”

What is the meaning of Topic 1?

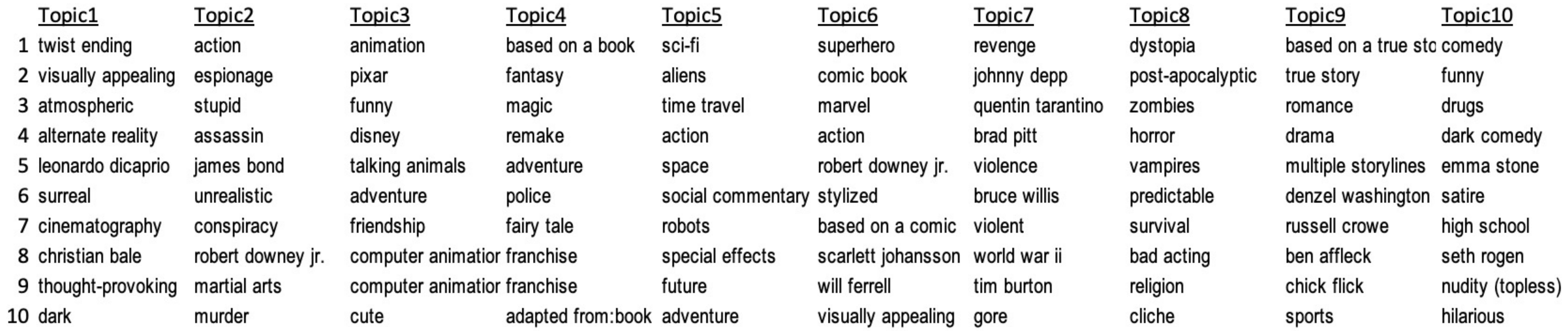


words common in eac	Topic1	Topic2
twist ending	6%	0%
visually appealing	5%	0%
atmospheric	4%	1%
alternate reality	3%	0%
leonardo dicaprio	3%	0%
surreal	3%	0%
cinematography	3%	0%
christian bale	2%	0%
thought-provoking	2%	0%
dark	2%	0%



Movies' topic profile	Topic1	Topic2
Inception	95%	0%
Shutter Island	92%	1%
Drive	87%	1%
The Prestige	80%	1%
The Dark Knight	75%	1%
The Dark Knight Rises	74%	1%
The Fountain	67%	1%
Prisoners	61%	12%
Nightcrawler	57%	5%
Gone Girl	57%	2%

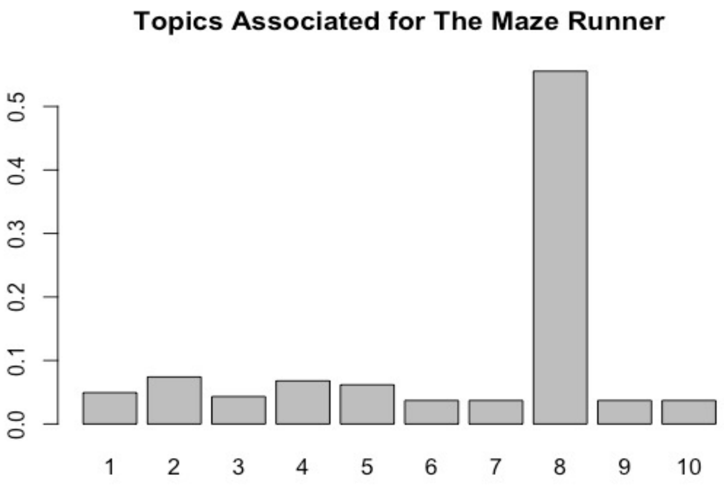
What do the Topics Mean?

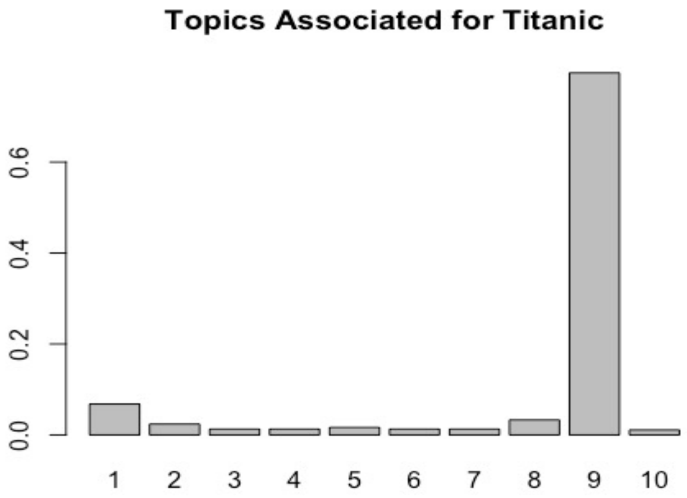


	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
1	twist ending	action	animation	based on a book	sci-fi	superhero	revenge	dystopia	based on a true sto	comedy
2	visually appealing	espionage	pixar	fantasy	aliens	comic book	johnny depp	post-apocalyptic	true story	funny
3	atmospheric	stupid	funny	magic	time travel	marvel	quentin tarantino	zombies	romance	drugs
4	alternate reality	assassin	disney	remake	action	action	brad pitt	horror	drama	dark comedy
5	leonardo dicaprio	james bond	talking animals	adventure	space	robert downey jr.	violence	vampires	multiple storylines	emma stone
6	surreal	unrealistic	adventure	police	social commentary	stylized	bruce willis	predictable	denzel washington	satire
7	cinematography	conspiracy	friendship	fairy tale	robots	based on a comic	violent	survival	russell crowe	high school
8	christian bale	robert downey jr.	computer animatior	franchise	special effects	scarlett johansson	world war ii	bad acting	ben affleck	seth rogen
9	thought-provoking	martial arts	computer animatior	franchise	future	will ferrell	tim burton	religion	chick flick	nudity (topless)
10	dark	murder	cute	adapted from:book	adventure	visually appealing	gore	cliche	sports	hilarious
label	atmospheric	action	animation	fantasy	sci.fi	superhero	revenge	survival	romance	comedy

See “topic_allterms.txt”

Proposed Solution
Does this make sense?


Topics Associated for The Maze Runner


Topics Associated for Titanic

The essence of our technique is that we represent every movie in a 10-dimensional (topic) space

Measuring the similarity (or dissimilarity) between movies

Use the Euclidean distance to measure the topic difference between movies

$$SimilarityScore = \sqrt{\sum_{i \in Topic\#} (Pr^{Movie1}(Topic_i) - Pr^{Movie2}(Topic_i))^2}$$

- The smaller score, the more similar.
  (The higher score is, the more dissimilar.)

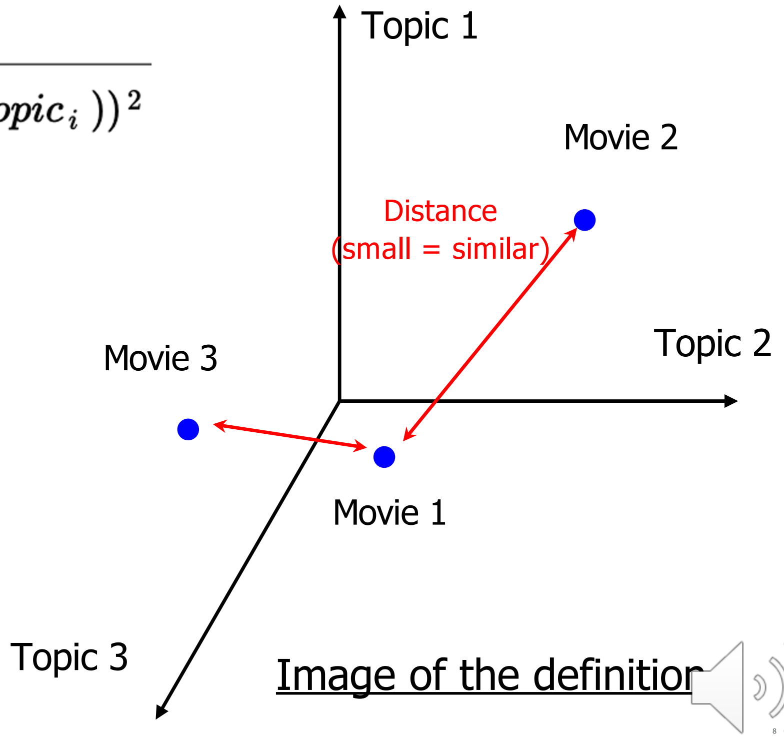
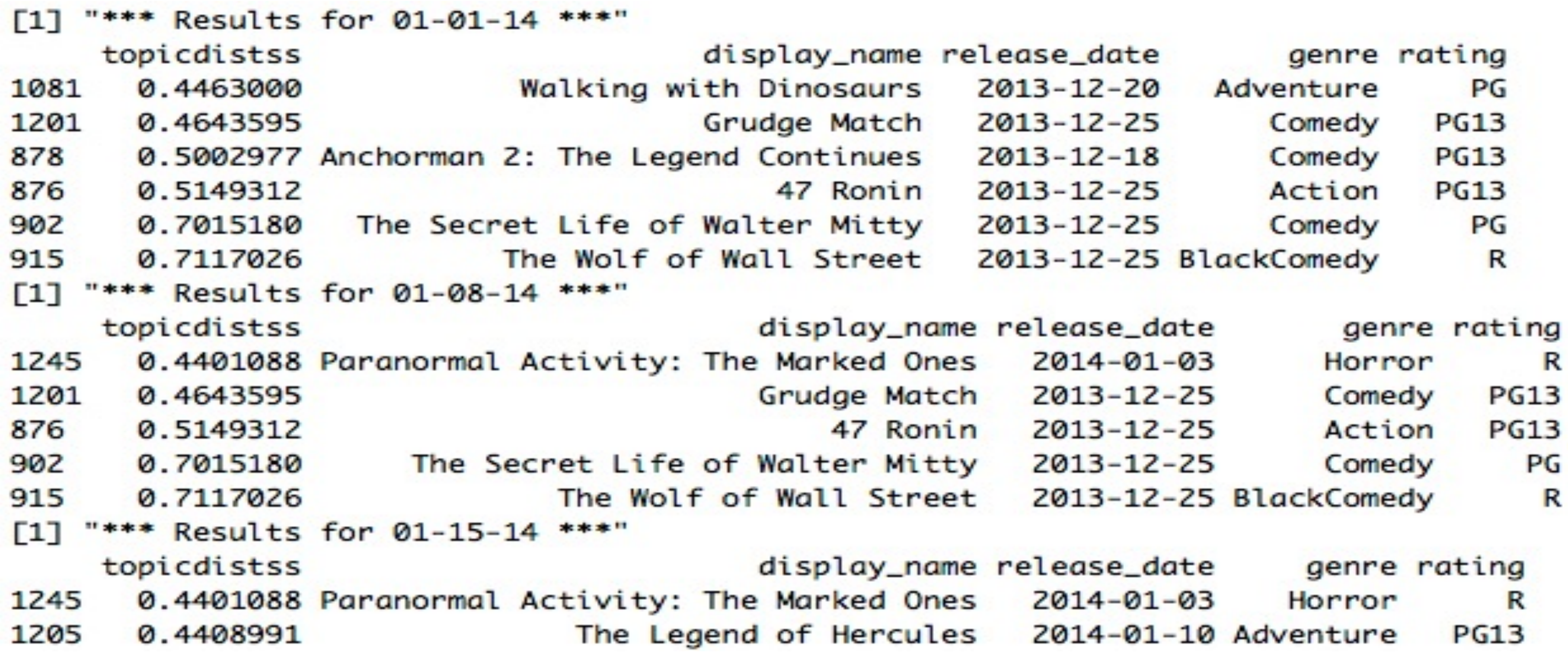


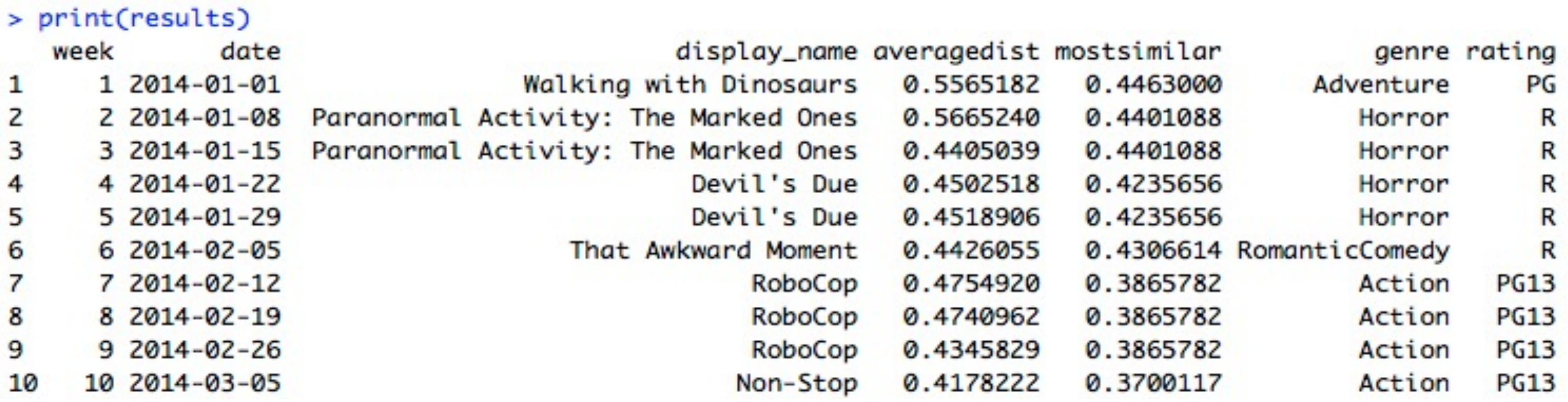
Image of the definition

For each compute “similarity” between “The Maze Runner” and other movies for that week



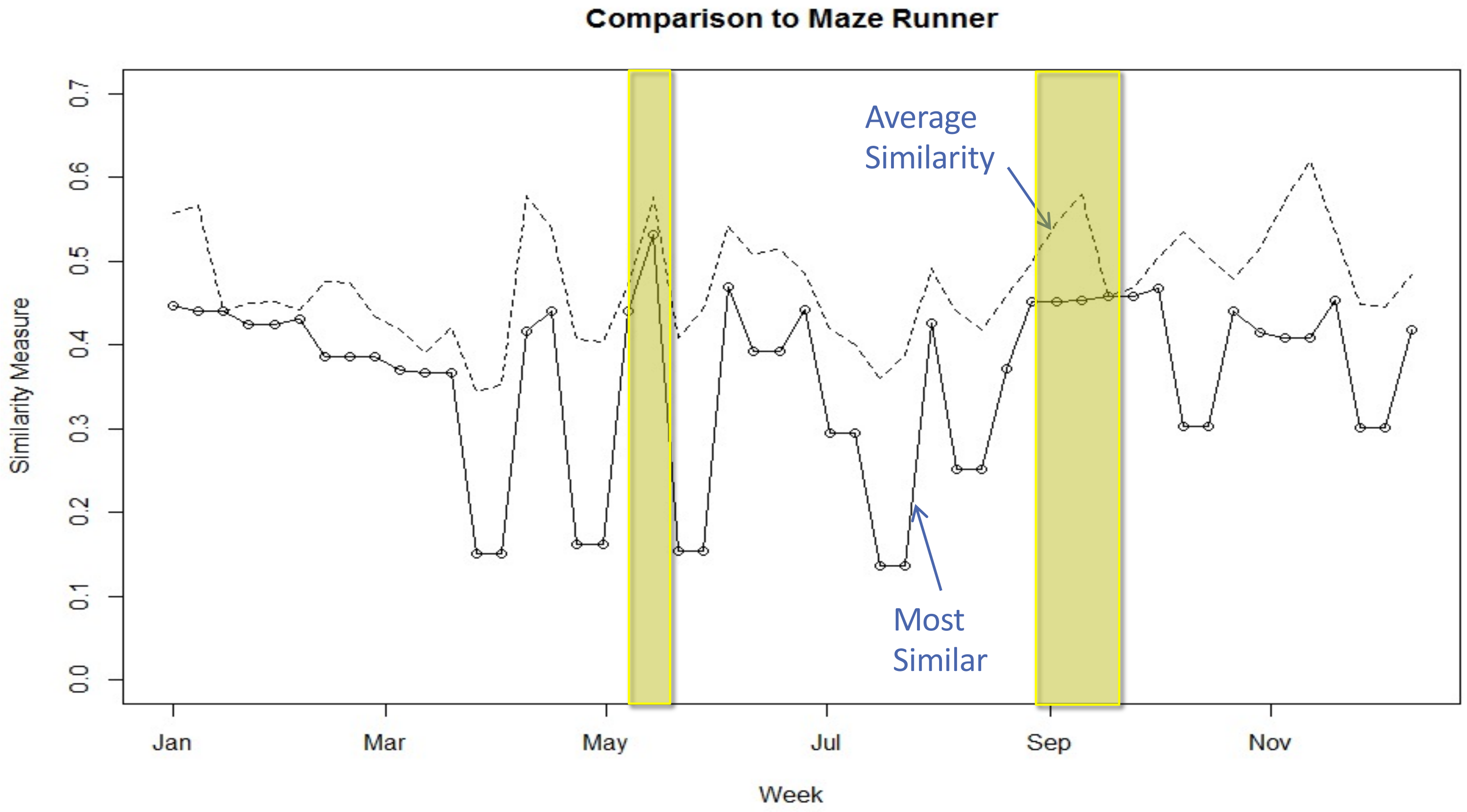
```
[1] "*** Results for 01-01-14 ***"
     topicdistss                      display_name release_date       genre rating
1081   0.4463000            Walking with Dinosaurs   2013-12-20   Adventure     PG
1201   0.4643595                      Grudge Match   2013-12-25      Comedy   PG13
878    0.5002977 Anchorman 2: The Legend Continues   2013-12-18      Comedy   PG13
876    0.5149312                          47 Ronin   2013-12-25      Action   PG13
902    0.7015180   The Secret Life of Walter Mitty   2013-12-25      Comedy     PG
915    0.7117026           The Wolf of Wall Street   2013-12-25 BlackComedy      R
[1] "*** Results for 01-08-14 ***"
     topicdistss                         display_name release_date       genre rating
1245   0.4401088 Paranormal Activity: The Marked Ones   2014-01-03      Horror      R
1201   0.4643595                         Grudge Match   2013-12-25      Comedy   PG13
876    0.5149312                             47 Ronin   2013-12-25      Action   PG13
902    0.7015180      The Secret Life of Walter Mitty   2013-12-25      Comedy     PG
915    0.7117026              The Wolf of Wall Street   2013-12-25 BlackComedy      R
[1] "*** Results for 01-15-14 ***"
     topicdistss                         display_name release_date     genre rating
1245   0.4401088 Paranormal Activity: The Marked Ones   2014-01-03    Horror      R
1205   0.4408991               The Legend of Hercules   2014-01-10 Adventure   PG13
```

Compute “similarity” score by week

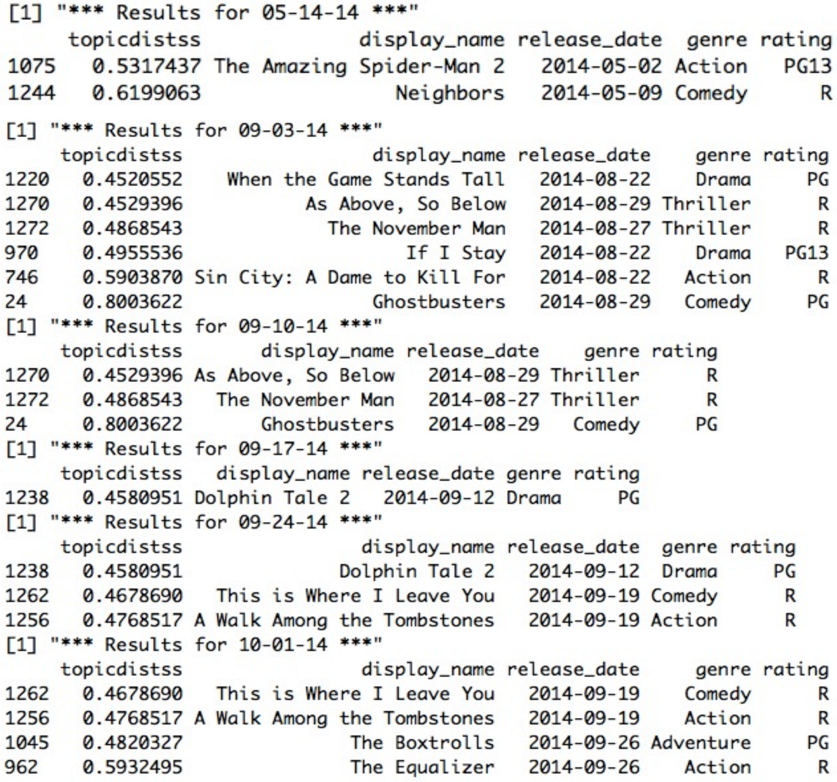


```
> print(results)
   week       date                     display_name averagedist mostsimilar         genre rating
1     1 2014-01-01           Walking with Dinosaurs   0.5565182   0.4463000     Adventure     PG
2     2 2014-01-08 Paranormal Activity: The Marked Ones   0.5665240   0.4401088        Horror      R
3     3 2014-01-15 Paranormal Activity: The Marked Ones   0.4405039   0.4401088        Horror      R
4     4 2014-01-22                       Devil's Due   0.4502518   0.4235656        Horror      R
5     5 2014-01-29                       Devil's Due   0.4518906   0.4235656        Horror      R
6     6 2014-02-05               That Awkward Moment   0.4426055   0.4306614 RomanticComedy      R
7     7 2014-02-12                          RoboCop   0.4754920   0.3865782        Action   PG13
8     8 2014-02-19                          RoboCop   0.4740962   0.3865782        Action   PG13
9     9 2014-02-26                          RoboCop   0.4345829   0.3865782        Action   PG13
10   10 2014-03-05                         Non-Stop   0.4178222   0.3700117        Action   PG13
```

Proposed Solution
Week by week comparison with all other movie releases



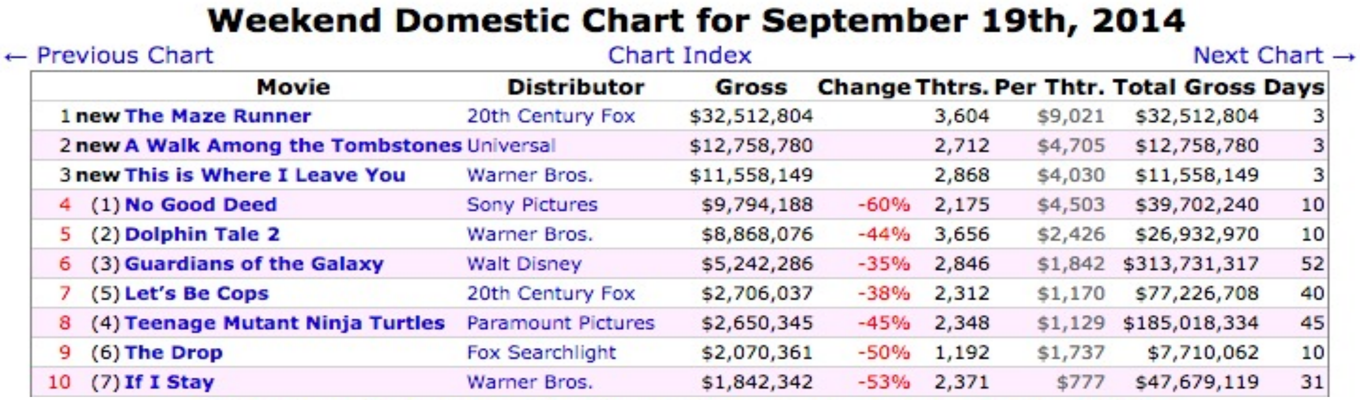
Proposed Solution
Some Good Dates



```
[1] "*** Results for 05-14-14 ***"
      topicdistss                display_name release_date  genre rating
1075    0.5317437 The Amazing Spider-Man 2   2014-05-02 Action   PG13
1244    0.6199063                  Neighbors   2014-05-09 Comedy      R
[1] "*** Results for 09-03-14 ***"
     topicdistss                 display_name release_date    genre rating
1220   0.4520552    When the Game Stands Tall   2014-08-22    Drama     PG
1270   0.4529396          As Above, So Below   2014-08-29 Thriller      R
1272   0.4868543            The November Man   2014-08-27 Thriller      R
970    0.4955536                   If I Stay   2014-08-22    Drama   PG13
746    0.5903870 Sin City: A Dame to Kill For   2014-08-22   Action      R
24     0.8003622                 Ghostbusters   2014-08-29   Comedy     PG
[1] "*** Results for 09-10-14 ***"
     topicdistss       display_name release_date    genre rating
1270   0.4529396 As Above, So Below   2014-08-29 Thriller      R
1272   0.4868543   The November Man   2014-08-27 Thriller      R
24     0.8003622       Ghostbusters   2014-08-29   Comedy     PG
[1] "*** Results for 09-17-14 ***"
     topicdistss  display_name release_date genre rating
1238   0.4580951 Dolphin Tale 2   2014-09-12 Drama     PG
[1] "*** Results for 09-24-14 ***"
     topicdistss                display_name release_date  genre rating
1238   0.4580951              Dolphin Tale 2   2014-09-12  Drama     PG
1262   0.4678690    This is Where I Leave You   2014-09-19 Comedy      R
1256   0.4768517 A Walk Among the Tombstones   2014-09-19 Action      R
[1] "*** Results for 10-01-14 ***"
     topicdistss                display_name release_date     genre rating
1262   0.4678690    This is Where I Leave You   2014-09-19    Comedy      R
1256   0.4768517 A Walk Among the Tombstones   2014-09-19    Action      R
1045   0.4820327               The Boxtrolls   2014-09-26 Adventure     PG
962    0.5932495               The Equalizer   2014-09-26    Action      R
```

Movie Scheduling Problem

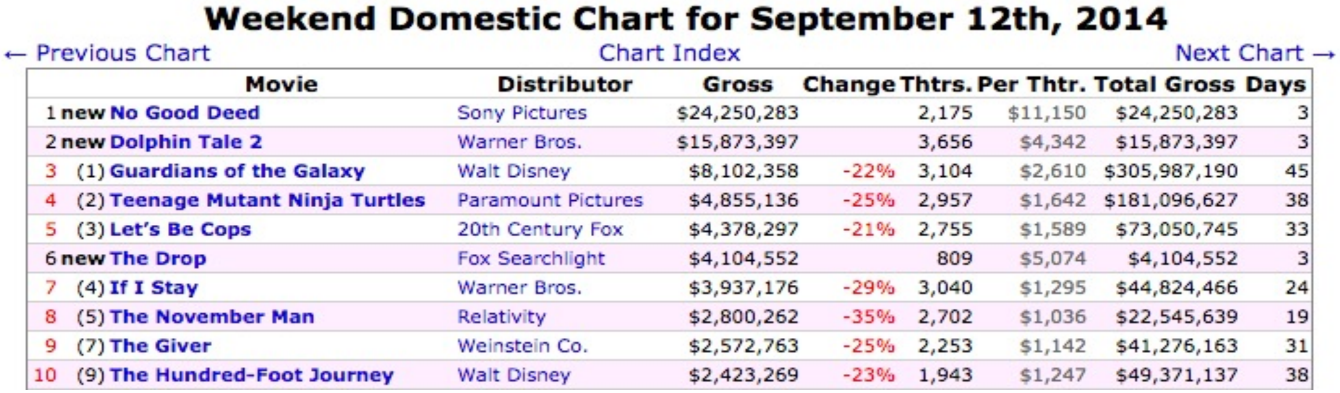
Is September 19 the ‘optimal’ release week for “The Maze Runner”? (previous week)



Weekend Domestic Chart for September 19th, 2014

← Previous Chart | Chart Index | Next Chart →

		Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1	new	The Maze Runner	20th Century Fox	$32,512,804		3,604	$9,021	$32,512,804	3
2	new	A Walk Among the Tombstones	Universal	$12,758,780		2,712	$4,705	$12,758,780	3
3	new	This is Where I Leave You	Warner Bros.	$11,558,149		2,868	$4,030	$11,558,149	3
4	(1)	No Good Deed	Sony Pictures	$9,794,188	-60%	2,175	$4,503	$39,702,240	10
5	(2)	Dolphin Tale 2	Warner Bros.	$8,868,076	-44%	3,656	$2,426	$26,932,970	10
6	(3)	Guardians of the Galaxy	Walt Disney	$5,242,286	-35%	2,846	$1,842	$313,731,317	52
7	(5)	Let’s Be Cops	20th Century Fox	$2,706,037	-38%	2,312	$1,170	$77,226,708	40
8	(4)	Teenage Mutant Ninja Turtles	Paramount Pictures	$2,650,345	-45%	2,348	$1,129	$185,018,334	45
9	(6)	The Drop	Fox Searchlight	$2,070,361	-50%	1,192	$1,737	$7,710,062	10
10	(7)	If I Stay	Warner Bros.	$1,842,342	-53%	2,371	$777	$47,679,119	31



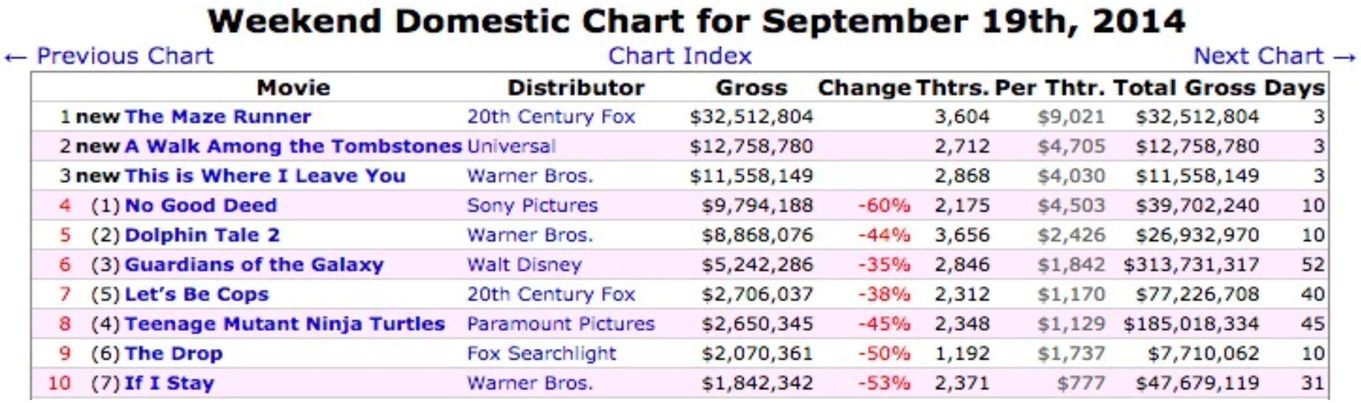
Weekend Domestic Chart for September 12th, 2014

← Previous Chart | Chart Index | Next Chart →

		Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1	new	No Good Deed	Sony Pictures	$24,250,283		2,175	$11,150	$24,250,283	3
2	new	Dolphin Tale 2	Warner Bros.	$15,873,397		3,656	$4,342	$15,873,397	3
3	(1)	Guardians of the Galaxy	Walt Disney	$8,102,358	-22%	3,104	$2,610	$305,987,190	45
4	(2)	Teenage Mutant Ninja Turtles	Paramount Pictures	$4,855,136	-25%	2,957	$1,642	$181,096,627	38
5	(3)	Let’s Be Cops	20th Century Fox	$4,378,297	-21%	2,755	$1,589	$73,050,745	33
6	new	The Drop	Fox Searchlight	$4,104,552		809	$5,074	$4,104,552	3
7	(4)	If I Stay	Warner Bros.	$3,937,176	-29%	3,040	$1,295	$44,824,466	24
8	(5)	The November Man	Relativity	$2,800,262	-35%	2,702	$1,036	$22,545,639	19
9	(7)	The Giver	Weinstein Co.	$2,572,763	-25%	2,253	$1,142	$41,276,163	31
10	(9)	The Hundred-Foot Journey	Walt Disney	$2,423,269	-23%	1,943	$1,247	$49,371,137	38

Movie Scheduling Problem

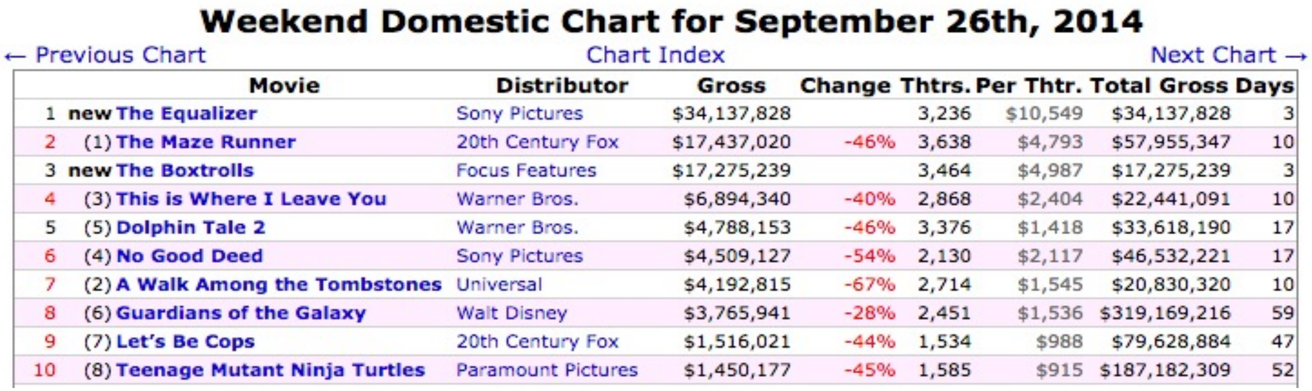
Is September 19 the ‘optimal’ release week for “The Maze Runner”? (later week)



Weekend Domestic Chart for September 19th, 2014

← Previous Chart Chart Index Next Chart →

Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1 new The Maze Runner	20th Century Fox	$32,512,804		3,604	$9,021	$32,512,804	3
2 new A Walk Among the Tombstones	Universal	$12,758,780		2,712	$4,705	$12,758,780	3
3 new This is Where I Leave You	Warner Bros.	$11,558,149		2,868	$4,030	$11,558,149	3
4 (1) No Good Deed	Sony Pictures	$9,794,188	-60%	2,175	$4,503	$39,702,240	10
5 (2) Dolphin Tale 2	Warner Bros.	$8,868,076	-44%	3,656	$2,426	$26,932,970	10
6 (3) Guardians of the Galaxy	Walt Disney	$5,242,286	-35%	2,846	$1,842	$313,731,317	52
7 (5) Let’s Be Cops	20th Century Fox	$2,706,037	-38%	2,312	$1,170	$77,226,708	40
8 (4) Teenage Mutant Ninja Turtles	Paramount Pictures	$2,650,345	-45%	2,348	$1,129	$185,018,334	45
9 (6) The Drop	Fox Searchlight	$2,070,361	-50%	1,192	$1,737	$7,710,062	10
10 (7) If I Stay	Warner Bros.	$1,842,342	-53%	2,371	$777	$47,679,119	31



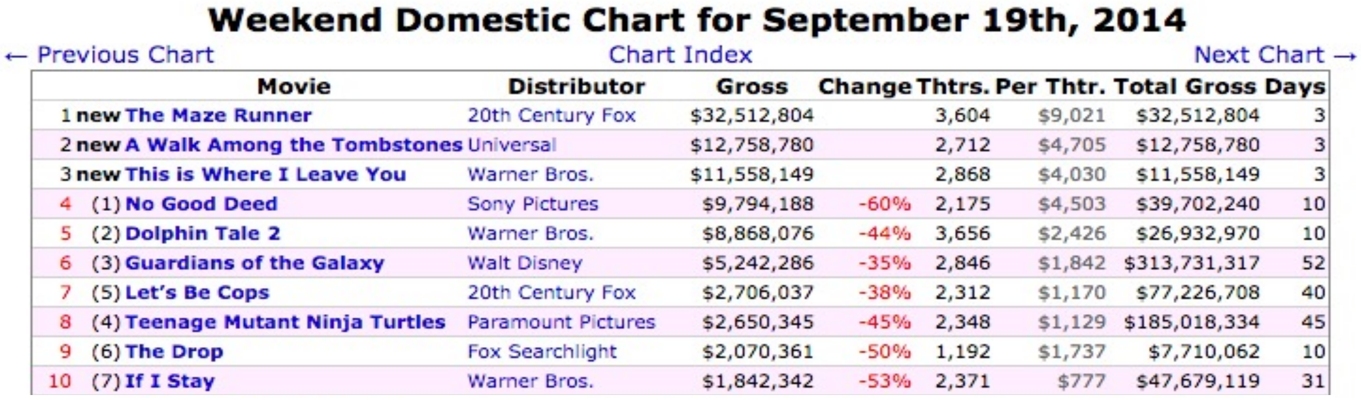
Weekend Domestic Chart for September 26th, 2014

← Previous Chart Chart Index Next Chart →

Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1 new The Equalizer	Sony Pictures	$34,137,828		3,236	$10,549	$34,137,828	3
2 (1) The Maze Runner	20th Century Fox	$17,437,020	-46%	3,638	$4,793	$57,955,347	10
3 new The Boxtrolls	Focus Features	$17,275,239		3,464	$4,987	$17,275,239	3
4 (3) This is Where I Leave You	Warner Bros.	$6,894,340	-40%	2,868	$2,404	$22,441,091	10
5 (5) Dolphin Tale 2	Warner Bros.	$4,788,153	-46%	3,376	$1,418	$33,618,190	17
6 (4) No Good Deed	Sony Pictures	$4,509,127	-54%	2,130	$2,117	$46,532,221	17
7 (2) A Walk Among the Tombstones	Universal	$4,192,815	-67%	2,714	$1,545	$20,830,320	10
8 (6) Guardians of the Galaxy	Walt Disney	$3,765,941	-28%	2,451	$1,536	$319,169,216	59
9 (7) Let’s Be Cops	20th Century Fox	$1,516,021	-44%	1,534	$988	$79,628,884	47
10 (8) Teenage Mutant Ninja Turtles	Paramount Pictures	$1,450,177	-45%	1,585	$915	$187,182,309	52

Movie Scheduling Problem

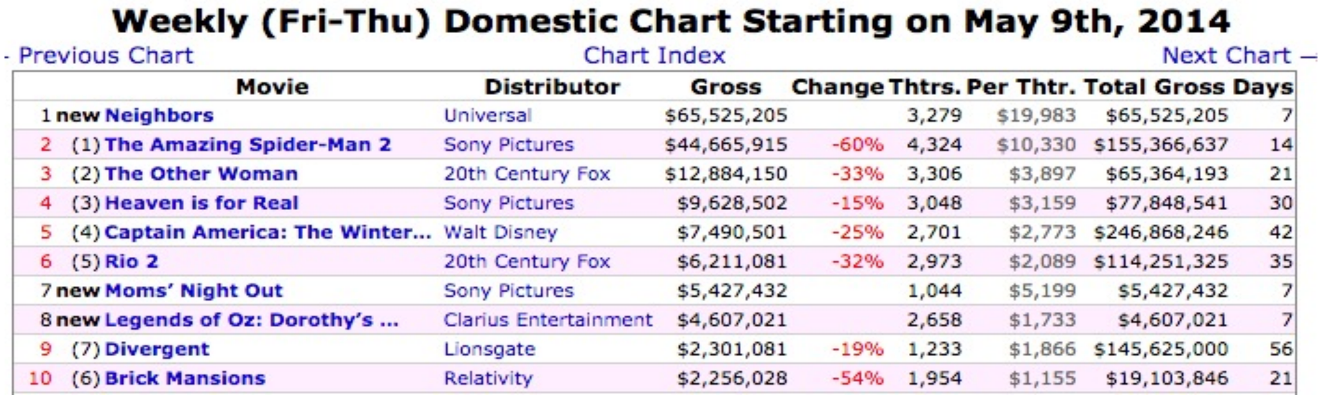
Is September 19 the 'optimal' release week for "The Maze Runner"? (late spring)



Weekend Domestic Chart for September 19th, 2014

← Previous Chart | Chart Index | Next Chart →

		Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1	new	The Maze Runner	20th Century Fox	$32,512,804		3,604	$9,021	$32,512,804	3
2	new	A Walk Among the Tombstones	Universal	$12,758,780		2,712	$4,705	$12,758,780	3
3	new	This is Where I Leave You	Warner Bros.	$11,558,149		2,868	$4,030	$11,558,149	3
4	(1)	No Good Deed	Sony Pictures	$9,794,188	-60%	2,175	$4,503	$39,702,240	10
5	(2)	Dolphin Tale 2	Warner Bros.	$8,868,076	-44%	3,656	$2,426	$26,932,970	10
6	(3)	Guardians of the Galaxy	Walt Disney	$5,242,286	-35%	2,846	$1,842	$313,731,317	52
7	(5)	Let's Be Cops	20th Century Fox	$2,706,037	-38%	2,312	$1,170	$77,226,708	40
8	(4)	Teenage Mutant Ninja Turtles	Paramount Pictures	$2,650,345	-45%	2,348	$1,129	$185,018,334	45
9	(6)	The Drop	Fox Searchlight	$2,070,361	-50%	1,192	$1,737	$7,710,062	10
10	(7)	If I Stay	Warner Bros.	$1,842,342	-53%	2,371	$777	$47,679,119	31



Weekly (Fri-Thu) Domestic Chart Starting on May 9th, 2014

- Previous Chart | Chart Index | Next Chart –

		Movie	Distributor	Gross	Change	Thtrs.	Per Thtr.	Total Gross	Days
1	new	Neighbors	Universal	$65,525,205		3,279	$19,983	$65,525,205	7
2	(1)	The Amazing Spider-Man 2	Sony Pictures	$44,665,915	-60%	4,324	$10,330	$155,366,637	14
3	(2)	The Other Woman	20th Century Fox	$12,884,150	-33%	3,306	$3,897	$65,364,193	21
4	(3)	Heaven is for Real	Sony Pictures	$9,628,502	-15%	3,048	$3,159	$77,848,541	30
5	(4)	Captain America: The Winter...	Walt Disney	$7,490,501	-25%	2,701	$2,773	$246,868,246	42
6	(5)	Rio 2	20th Century Fox	$6,211,081	-32%	2,973	$2,089	$114,251,325	35
7	new	Moms' Night Out	Sony Pictures	$5,427,432		1,044	$5,199	$5,427,432	7
8	new	Legends of Oz: Dorothy's ...	Clarius Entertainment	$4,607,021		2,658	$1,733	$4,607,021	7
9	(7)	Divergent	Lionsgate	$2,301,081	-19%	1,233	$1,866	$145,625,000	56
10	(6)	Brick Mansions	Relativity	$2,256,028	-54%	1,954	$1,155	$19,103,846	21

Key Takeaways

- LDA models work well with large dimensional data that may not be “dense” like we find in k-means clustering
- Prototypical examples involve large textual data sets (e.g., large vocabularies), but can be applied to any instance where probabilistic clustering like clickstream data, customer call center transcripts, or purchases
- Instead of thinking about understanding text we can represent language as a “bag of words” or represent word counts in a high dimensional vector space. Machine learning pushes both in terms of large-p (numbers of variables) as well as large-n (numbers of observations)
- An approach to analyzing unstructured data is to represent our data as a “big” matrix and look for patterns