Joe Standerfer, Spriha Gupta, Jasmine Kaur, Daniel Lesser

Marketing Analytics 95-832

Assignment 2: Movie Topic Modeling

Gotham Group is preparing to launch The Maze Runner, a new dystopian film based on James Dashner's 2009 novel. Our group has been tasked with identifying the most opportune week of 2014 to launch the movie. Heuristic approaches have traditionally been used to choose release dates, taking advantage of industry knowledge regarding what time of the year is ideal for certain movie genres. In this case study, we have employed a topic modeling approach as well as a K-means approach. Using movie reviews, our analysis identifies which movies are categorically similar to Maze Runner, what the optimal release weeks have been for the genre over time, and what specific week Gotham Group should release The Maze Runner in 2014. Our analysis shows that releasing in week 37 will both maximize expected revenue and minimize competition from similar movies.

Latent Dirichlet Allocation (LDA) was used to generate topic probabilities for a subset of movies released between 2006 and 2014. Ten topics were generated using tags from movie reviews, with each movie having a probabilistic distribution across the topics. This distribution was relatively uniform (Figure 1). However, if one were to take a deterministic approach and assign only the highest probability topic to each movie, the distribution is more skewed (Figure 2). In this case, we can see that Animation, Drama, Comedy, and Action are the most frequent topics.

LDA does not provide topic labels by default, so one required step in our approach was to assign qualitative labels to each topic based on the output of the model. To determine the appropriate labels, we reviewed the conditional probability of each tag showing up given each topic as well as the probability of each topic showing up given each movie (Figure 3 and 4). Based on the term and topic frequencies, we assigned the following labels to topics one through ten: Thriller Drama, Action, Animation, Fantasy, Sci-Fi, Dark Comedy, Dystopia, Drama, Comedy. The Maze Runner falls within the Dystopia topic, which is often

described with words such as dystopia, post-apocalyptic, zombies and horror. This topic includes movies such as I am Legend, The Road, The Book of Eli, and Limitless.

We calculated the similarity between our target movie The Maze Runner and every other movie based on Euclidean distance. The top five similar movies were The Twilight: New Moon, Daybreakers, 28 Weeks Later, The Conjuring and Underworld: Evolution (Figure 5). While Twilight is usually characterized as a vampire love story, it is also about the fight for survival which draws a parallel to the storyline of The Maze Runner series. These movies are representations of dystopic futures that include elements of horror and are more representative of topic eight than the other topics. Overall, the model is a good indicator of movie topics.

Using the weekly launch dates for movies in 2014, we computed the average similarity for each week using a decaying function. Each movie's score was penalized depending on whether it was released two weeks before or one week after that week ($e^{-weeksApart^2}$) (Figure 7). We chose this version of a decaying function as it should reflect best how movie demand would behave during opening weeks. The distance for each week was the sum of the similarities for all the individual movies.

In addition to computing a similarity score for each week, several other industry factors were taken into consideration. For instance, there is an abundance of horror films released in the fall around Halloween, indicating that movie preference varies by season. In Figure 6, we identify which weeks of the year have the best average opening weekend revenue for dystopian movies. While the data is relatively sparse due to only including nine years of data, the end of the year as well as the Fall appear to be the most successful times for dystopian movies to be released.

As a result of our analysis, the best week to release The Maze Runner appears to be week 37, which coincides with Labor Day weekend. There were several factors that led to this decision. First, week 37 had a very low similarity score compared to the rest of the year, meaning The Maze Runner wouldn't

face stiff competition from other similar movies released around that date. Being a holiday weekend, initial viewership will be high. Finally, there are no other blockbusters being released around that date.

Other possible selections include weeks 50, 51, and 22. Weeks 50 and 51 were attractive due to low similarity scores and high historical revenue. However, The Maze Runner does not appear to be a family-oriented film, which is what typically performs well in the holiday season. Week 22 was another candidate because of the low similarity score and lack of other major releases. Coming at the start of summer, it is also a good time of year for the genre. Most other major movies have chosen weeks later in the summer months.

Topic modeling with 15 topics instead of 10 returned similar results to our original analysis, though weeks 50-52 were slightly more attractive options in this approach. Figure 9 shows the top 10 weeks using a 15-topics distribution model. Based on the various approaches taken, week 37 remains the most ideal option available.

As an alternative to topic modeling, K-means was used to determine movie clusters. The dataset includes a wide range of features and as a result we used Variance Inflation Scores to eliminate highly correlated variables. The elbow method did not yield a definitive optimal k (Figure 10), so we also reviewed the Silhouette score metric (Figure 11). K=10 appears to be an optimal value, with the highest score of 0.2. The scores plateaued on higher value of K.

A simple Euclidean distance measure was used to identify which cluster was furthest away from The Maze Runner's cluster. The release dates of the movies in this segment were then viewed as potential release dates due to the high dissimilarities. According to this analysis, the ideal week to release would have been week 27, during the 4th of July holiday. However, this conclusion did not align with our topic modeling approach or historical revenue estimates for the genre. In addition, applying PCA to the resulting clusters (Figure 12) did not result in distinct groups, limiting the interpretability of the model. As a result, we would not suggest using K-means as an approach to address this problem.

Appendix

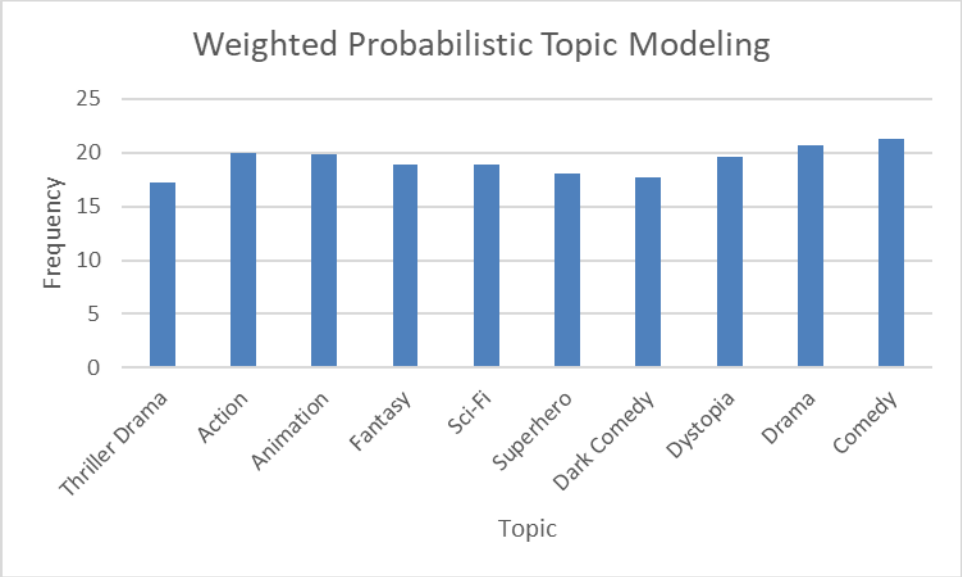

Figure 1

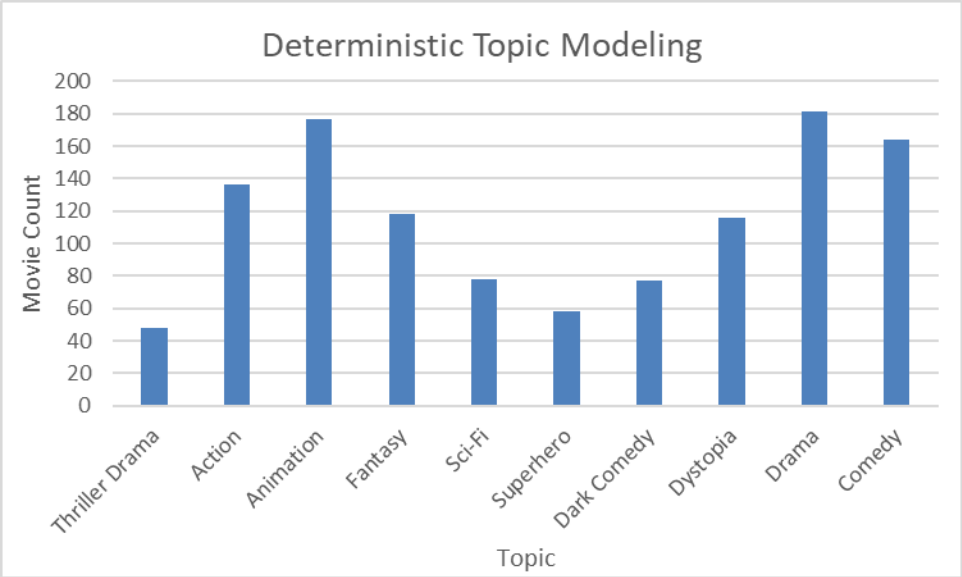

Figure 2

Given Label:	Thriller Drama		Action		Animation		Fantasy		Sci-Fi
	Topic1		Topic2		Topic3		Topic4		Topic5
twist ending	6%	action	8%	animation	10%	based on a book	10%	sci-fi	13%
visually appealing	5%	espionage	3%	pixar	8%	fantasy	7%	aliens	6%
atmospheric	4%	stupid	2%	funny	4%	magic	5%	time travel	4%
alternate reality	3%	assassin	2%	disney	4%	remake	4%	action	4%
leonardo dicaprio	3%	james bond	2%	talking animals	3%	adventure	3%	space	3%
surreal	3%	unrealistic	2%	adventure	2%	police	3%	social commentary	3%
cinematography	3%	conspiracy	2%	friendship	2%	fairy tale	2%	robots	3%
christian bale	2%	robert downey jr.	2%	computer animation	2%	franchise	2%	special effects	2%
thought-provoking	2%	martial arts	2%	family	2%	simon pegg	2%	future	2%
dark	2%	murder	2%	cute	2%	matt damon	2%	adventure	2%
Inception	95%	Casino Royale	85%	Up	90%	The Hobbit: An Unexpec	85%	Avatar	89%
Shutter Island	92%	Skyfall	78%	Ratatouille	87%	Harry Potter and the D (2010)	79%	Edge of Tomorrow	89%
Drive	87%	Quantum of Solace	75%	Finding Nemo	87%	Stardust	75%	District 9	86%
The Prestige	80%	Sherlock Holmes	72%	Toy Story 3D Double Fe	84%	Harry Potter and the D (2011)	72%	Jurassic Park	78%
The Dark Knight	75%	The Bourne Ultimatum	71%	The Lion King	83%	The Hobbit: The Desola	72%	Cloverfield	77%
The Dark Knight Rises	74%	Salt	70%	Beauty and the Beast	76%	Harry Potter and the H	71%	Star Trek	77%
The Fountain	67%	Hanna	68%	Monsters, Inc.	74%	Hot Fuzz	68%	Pacific Rim	72%
Prisoners	61%	The Expendables	65%	How to Train Your Drag (2010)	68%	Harry Potter and the O	64%	Star Trek Into Darknes	69%
Nightcrawler	57%	Knight and Day	61%	Toy Story 3	64%	Star Wars Ep. I: The P	56%	Transformers	68%
Gone Girl	57%	Mission: Impossible II	56%	Mamma Mia!	62%	The Golden Compass	52%	Elysium	64%

Figure 3

Given Label:	Superhero		Dark Comedy		Dystopia		Drama		Comedy
	Topic6		Topic7		Topic8		Topic9		Topic10
superhero	11%	revenge	4%	dystopia	8%	based on a true story	4%	comedy	8%
comic book	6%	johnny depp	4%	post-apocalyptic	8%	true story	4%	funny	6%
marvel	5%	quentin tarantino	3%	zombies	5%	romance	4%	drugs	3%
action	4%	brad pitt	3%	horror	2%	drama	2%	dark comedy	2%
robert downey jr.	3%	violence	3%	vampires	2%	multiple storylines	2%	emma stone	2%
stylized	3%	bruce willis	3%	predictable	2%	denzel washington	2%	satire	2%
based on a comic	2%	violent	3%	survival	2%	russell crowe	1%	high school	2%
scarlett johansson	2%	world war ii	2%	bad acting	2%	ben affleck	1%	seth rogen	2%
will ferrell	2%	tim burton	2%	religion	2%	chick flick	1%	nudity (topless)	2%
visually appealing	2%	gore	2%	cliche	2%	sports	1%	hilarious	2%
The Avengers	90%	Inglourious Basterds	93%	I am Legend	85%	The King's Speech	87%	The Hangover	87%
Scott Pilgrim vs. The	84%	Django Unchained	90%	The Road	79%	Titanic	80%	Superbad	85%
Iron Man 2	82%	Sweeney Todd: The Dem	83%	The Book of Eli	73%	The Blind Side	67%	Forgetting Sarah Marsh	80%
Iron Man	80%	Pirates of the Caribbe (2(	66%	Limitless	65%	Argo	66%	Borat	79%
Thor	79%	Machete	65%	Prometheus	60%	Moneyball	65%	The Cabin in the Woods	78%
Stranger Than Fiction	79%	Pirates of the Caribbe (2(	60%	Underworld: Evolution	59%	American Gangster	60%	Knocked Up	75%
Iron Man 3	75%	True Grit	54%	The Twilight Saga: New	56%	Blood Diamond	58%	Easy A	73%
X-Men Origins: Wolveri	71%	Burn After Reading	53%	The Maze Runner	56%	Inside Man	58%	The Social Network	69%
Captain America: The V	69%	Tim Burton's The Night	53%	The Conjuring	53%	The Pursuit of Happyne	58%	BrÃ¼no	67%
Captain America: The F	68%	Law Abiding Citizen	47%	28 Weeks Later	53%	He's Just Not That Int	57%	21 Jump Street	67%

Figure 4

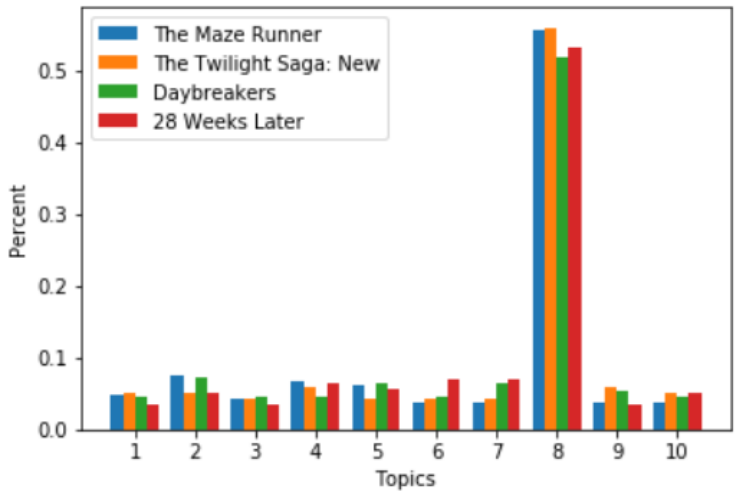

Figure 5

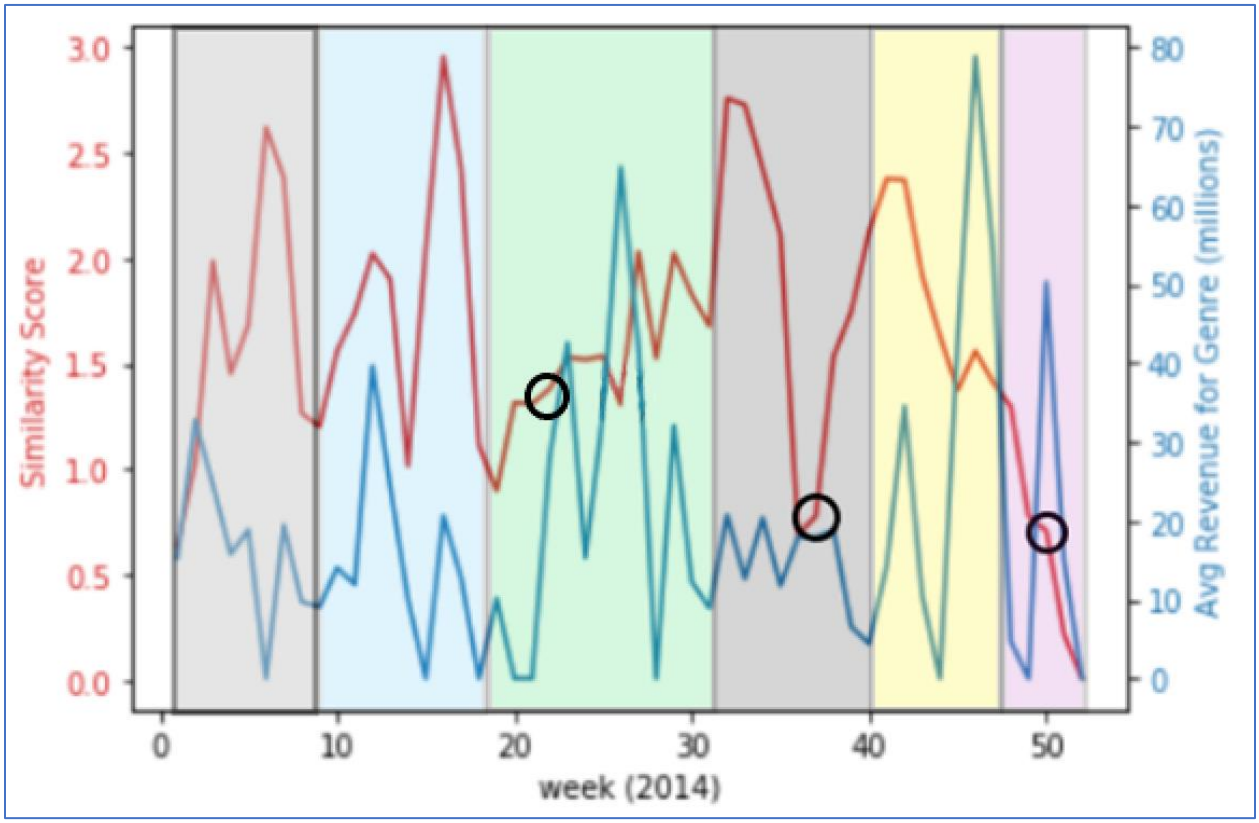

Figure 6

Decaying function chart

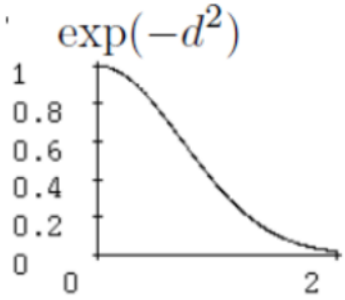

Figure 7

```
*** Top 20 weeks ***
week 52 | score: 0.011
week 51 | score: 0.218
week 1  | score: 0.637
week 36 | score: 0.691
week 50 | score: 0.71
week 49 | score: 0.778
week 37 | score: 0.779
week 19 | score: 0.899
week 14 | score: 1.017
week 2  | score: 1.027
week 18 | score: 1.114
week 9  | score: 1.203
week 8  | score: 1.266
week 48 | score: 1.297
week 26 | score: 1.306
week 21 | score: 1.311
week 20 | score: 1.315
week 45 | score: 1.376
week 22 | score: 1.382
week 47 | score: 1.412
```

Figure 8

```
*** Top 10 weeks ***
week52
         score :  0.014757653330580725
         avgRev: $ 34663135.0
week51
         score :  0.30328698139805776
         avgRev: $ 9836785.0
week36
         score :  0.909493501449896
         avgRev: $ 16183409.0
week50
         score :  0.9731116019822895
         avgRev: $ 39637079.0
week1
         score :  0.9919503390827695
         avgRev: $ 7707949.333333333
week49
         score :  1.1263679357579683
         avgRev: $ 0
week37
         score :  1.2090636989056909
         avgRev: $ 20244552.0
week19
         score :  1.486938677250873
         avgRev: $ 0
week2
         score :  1.5926890417720307
         avgRev: $ 10175603.333333334
week47
         score :  1.664200830291254
         avgRev: $ 0
```

Figure 9

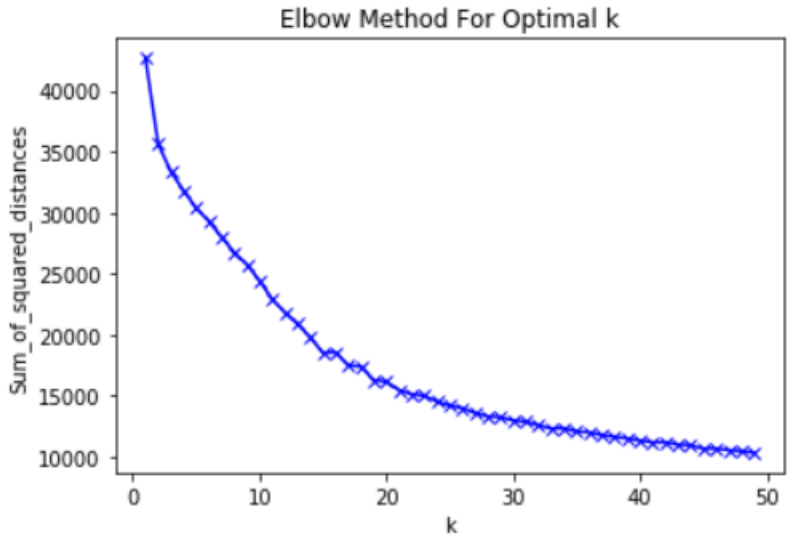

Figure 10

```
For n_clusters = 2, silhouette score is 0.31281950156356947)
For n_clusters = 3, silhouette score is 0.09871756869862405)
For n_clusters = 4, silhouette score is 0.09016351263661543)
For n_clusters = 5, silhouette score is 0.09436304799687056)
For n_clusters = 6, silhouette score is 0.12244465276505531)
For n_clusters = 7, silhouette score is 0.1268621081704214)
For n_clusters = 8, silhouette score is 0.15100777569445967)
For n_clusters = 9, silhouette score is 0.1726560552951987)
For n_clusters = 10, silhouette score is 0.20905956811191242)
```

Figure 11

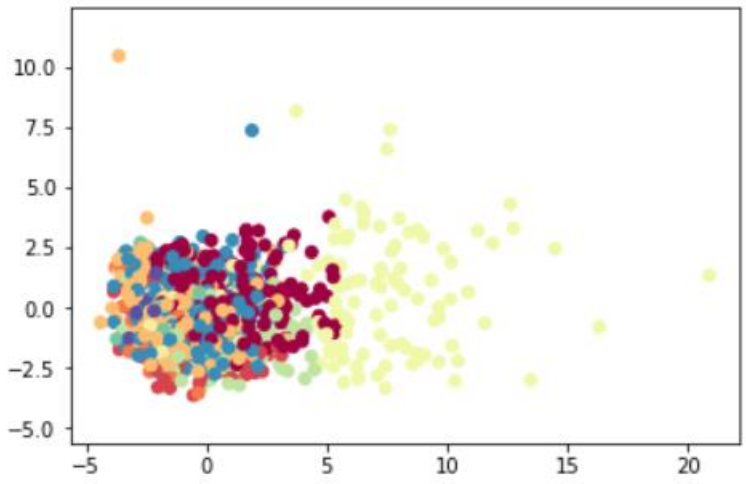

Figure 12