Case: Movie Scheduling Using Topic Models

Team L: Sherine George George, Sherry Li, Wendy Kuo, Yuchen Zhu, Enrique Garcia

Date: January 29th, 2024

Q1: The first step taken was to analyze the top-20 tags associated with each topic and cross-referencing them with a list of 20 movies per topic. This is a classic example of where SME (Subject-Matter-Expertise) can be applied to extract actionable insights in a data-analysis project. We extracted the top 20-tags per topic, Exhibit 1, and defined the top 3 sub-categories that would encompass these tags. We color coded their representation in each topic list, with those not highlighted being neutral terms associated with actor names, director names, or other nouns. Finally, we used SME-inference to define a final category name for each topic as follows: (i) Thriller (ii) Espionage (iii) Animation (iv) Fantasy (v) Science Fiction (vi) Heroes (vii) Action (viii) Apocalyptic (ix) Historical (x) Comedy. It is noteworthy that some tags are present in multiple lists - suggesting that some tags/topics are not mutually exclusive.

Q2: By calculating the euclidean distance between the 10 topic scores for “The Maze Runner” and all other movies, we were able to effectively construct a “dissimilarity index” where movies with similar probability distributions across the topics are allocated a lower score. According to this index, the top 10 most similar movies are as follows: (i) The Twilight Saga: New Moon (ii) Daybreakers (iii) 28 Weeks Later (iv) The Conjuring (v) Underworld: Evolution (vi) 1408 (vii) Insidious (viii) The Hunger Games: Catching Fire (ix) Doomsday (x) Resident Evil: Extinction. Exhibit 2A highlights the relative probability distribution of the films above across the 10 topics. We can see that these top 10 movies all share a high probability (~55%) of being in the Apocalyptic category.

To validate our analysis, we examined the top 20 tags for these 10 films (Exhibit 2B). We can see that only 4 out of 20 of such tags describe “The Maze Runner”. From question 1, we have separated category (vii) into 3 subcategories: (1) Apocalyptic (2) Horror (3) Supernatural, which are slightly different from each other. It thus follows that the 20 tags are spread out across the 3 subcategories. With this information, we can see that using our “similarity index” might not be accurate as three different categories of films are grouped together under topic (viii). This is corroborated by the raw data provided, where the top 10 films have varying genres, creative types, ratings, and sources (Exhibit 2C). To improve this model, LDA analysis can be made with a greater number of topics such that the 3 types of subcategories can be separated.

Q3: After filtering launch dates, we compiled a list of movies released weekly in 2014 (Exhibit 3A). Combining these weekly launches with dissimilarity scores to "The Maze Runner," we calculated average dissimilarity scores for each week. Exhibit 3B presents the weekly average dissimilarity scores in 2014, revealing that weeks 13, 22, and 44 recorded the three highest average dissimilarity scores throughout the year. Additionally, we incorporated the dissimilarity score from the previous week into our analysis, creating a two-week rolling average dissimilarity, as shown in Exhibit 3C. This approach allows us to discern trends over a two-week interval. Notably, weeks 15, 24, and 44 emerge with the highest two-week rolling average dissimilarity scores, providing further insights into the evolving patterns.

However, relying solely on the weekly average dissimilarity may obscure critical information due to the possibility of a week with both the highest and lowest scores. To ensure a more informed evaluation, it is necessary to assess both extremes. Exhibit 3D illustrates the weekly maximum and minimum dissimilarity scores. Notably, weeks 13, 20, and 22 exhibit the highest maximum dissimilarity scores, while weeks 11, 12, 39, 46, and 49 represent the lowest. By combining these two metrics, we can provide the backbone of analysis for Q4.

Q4: In Q3, our analysis pointed to weeks 13, 20, and 22 as having the highest maximum dissimilarity scores, while weeks 13, 22, and 44 stood out with the highest average dissimilarity scores. Weeks 15, 24, and 44 were notable for the highest two-week rolling average dissimilarity scores. Conversely, weeks 11, 12, 39, 46, and 49 displayed the lowest minimum dissimilarity scores, suggesting periods to avoid for movie launches.

Considering these scores, potential weeks for releasing "The Maze Runner" include 13, 15, 20, 22, 24, and 44. Notably, weeks 13, 22, and 44 consistently emerged across various metrics. Upon closer examination, we opted to eliminate week 13 due to its proximity to weeks 11 and 12, which exhibited low minimum dissimilarity scores, potentially impacting overall performance.

This leaves us with weeks 22 (May 26 to June 1) and 44 (Oct 27 to Nov 2). Week 22, positioned at the start of summer, seems ideal for launching an action movie like 'The Maze Runner,' targeting the teenage audience who will be on their summer break. Releasing a movie during the summer ensures audience availability, as many people are on break, matches the positive vibe action movies are trying to create, and allows more promotional opportunities for marketing purposes. Additionally, week 44, coinciding with fall break and preceding Thanksgiving, presents itself as a suitable release window.

November also falls in the award season, which can maximize exposure and, therefore, increase the possibility of nominations for awards. Considering additional options, weeks 20 (May 12 to May 18) and 24 (Jun 9 to June 15) become viable third and fourth choices as they align with the summer season and have the highest average dissimilarity score and rolling dissimilarity score behind weeks 22 and 44. Week 20 can be beneficial for people in Northern States as schools tend to start the summer break earlier in mid-May. On the other hand, week 24 can be suitable for people in Southern States as schools tend to start the summer break in mid-June. In conclusion, our top four choices for weekly launch dates are weeks 22, 44, 20, and 24. An overview of the selection of specific weeks and their considerations can be seen in Exhibit 4.

Q5B: The K-means clustering analysis of tags primarily grouped movies into a single cluster (Cluster 2) including the Maze Runner - suggesting a broad similarity among a large number of movies as seen in Exhibit 5A. Further analysis on the popular tags in cluster 2, as shown in Exhibit 5B, provides no valuable information as it presents a vast variety of tags with limited relations to each other. With the exception of cluster 8, which consists of a cluster of animated movies, other clusters beyond cluster 2 have a limited selection of 1 to 2 movies only, which are not interpretable.

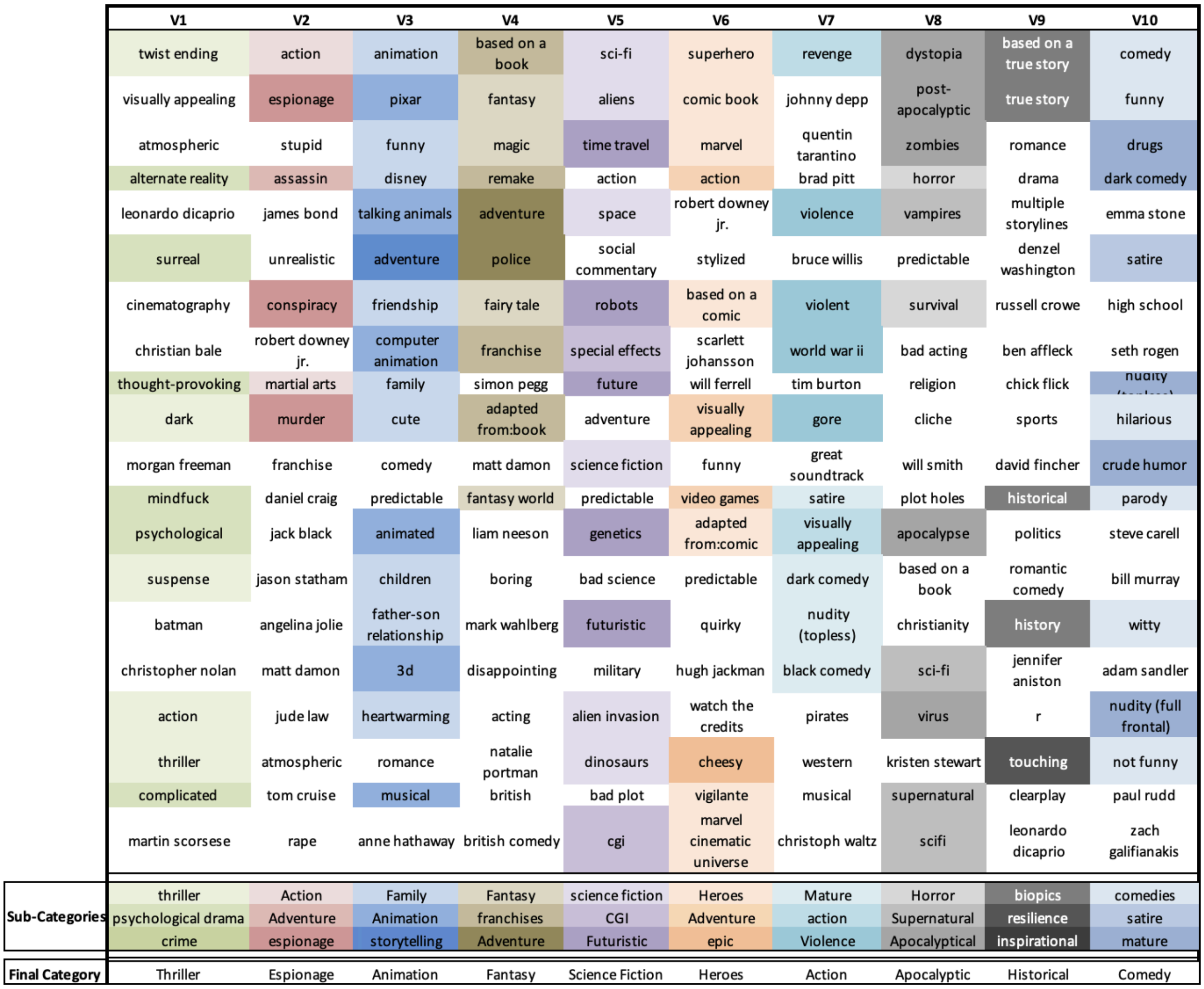
The gross generalization of the vast number of movies in cluster 2 also limits the usefulness of k-means as it limits the ability to differentiate and schedule movies effectively, as it provides a wide but shallow classification of movie attributes. In contrast, LDA topic modeling gave a more granular analysis, particularly with "The Maze Runner" being associated with Topic 8, which includes specific and relevant tags such as dystopia, post-apocalyptic, and zombies as shown in Exhibit 5C. This detailed thematic information from LDA is more conducive to understanding the distinct content of movies and can guide more precise marketing and scheduling decisions.

While K-means offered a macro-level view of movie attributes, LDA provided a micro-level thematic insight. For strategic planning in the movie industry, where understanding the nuanced preferences of audiences is key, LDA's detailed approach is more advantageous than the broader strokes of K-means clustering. To improve the k-means clustering, we can clean the data by grouping similar tags together and increase the number of clusters in k-means before grouping different clusters together.

—--------------- —--------------- —--------------- —--------------- —--------------- —---------------

Appendix

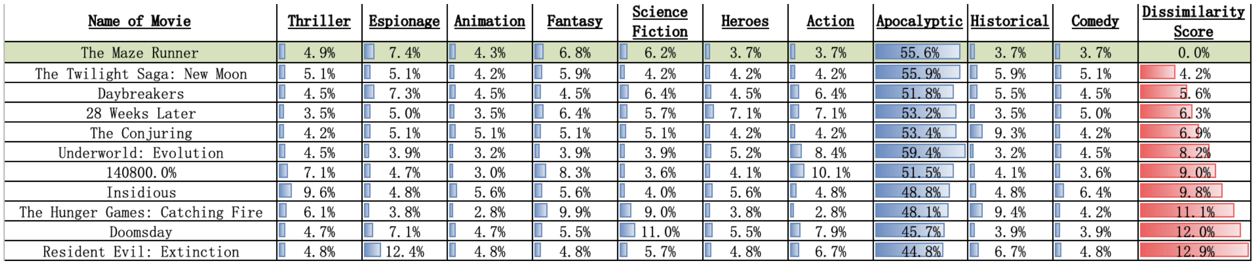
Exhibit 1: Top-20 Tags per Topic and Resulting Topic Definition



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
	twist ending	action	animation	based on a book	sci-fi	superhero	revenge	dystopia	based on a true story	comedy
	visually appealing	espionage	pixar	fantasy	aliens	comic book	johnny depp	post-apocalyptic	true story	funny
	atmospheric	stupid	funny	magic	time travel	marvel	quentin tarantino	zombies	romance	drugs
	alternate reality	assassin	disney	remake	action	action	brad pitt	horror	drama	dark comedy
	leonardo dicaprio	james bond	talking animals	adventure	space	robert downey jr.	violence	vampires	multiple storylines	emma stone
	surreal	unrealistic	adventure	police	social commentary	stylized	bruce willis	predictable	denzel washington	satire
	cinematography	conspiracy	friendship	fairy tale	robots	based on a comic	violent	survival	russell crowe	high school
	christian bale	robert downey jr.	computer animation	franchise	special effects	scarlett johansson	world war ii	bad acting	ben affleck	seth rogen
	thought-provoking	martial arts	family	simon pegg	future	will ferrell	tim burton	religion	chick flick	nudity (topless)
	dark	murder	cute	adapted from:book	adventure	visually appealing	gore	cliche	sports	hilarious
	morgan freeman	franchise	comedy	matt damon	science fiction	funny	great soundtrack	will smith	david fincher	crude humor
	mindfuck	daniel craig	predictable	fantasy world	predictable	video games	satire	plot holes	historical	parody
	psychological	jack black	animated	liam neeson	genetics	adapted from:comic	visually appealing	apocalypse	politics	steve carell
	suspense	jason statham	children	boring	bad science	predictable	dark comedy	based on a book	romantic comedy	bill murray
	batman	angelina jolie	father-son relationship	mark wahlberg	futuristic	quirky	nudity (topless)	christianity	history	witty
	christopher nolan	matt damon	3d	disappointing	military	hugh jackman	black comedy	sci-fi	jennifer aniston	adam sandler
	action	jude law	heartwarming	acting	alien invasion	watch the credits	pirates	virus	r	nudity (full frontal)
	thriller	atmospheric	romance	natalie portman	dinosaurs	cheesy	western	kristen stewart	touching	not funny
	complicated	tom cruise	musical	british	bad plot	vigilante	musical	supernatural	clearplay	paul rudd
	martin scorsese	rape	anne hathaway	british comedy	cgi	marvel cinematic universe	christoph waltz	scifi	leonardo dicaprio	zach galifianakis
Sub-Categories	thriller	Action	Family	Fantasy	science fiction	Heroes	Mature	Horror	biopics	comedies
	psychological drama	Adventure	Animation	franchises	CGI	Adventure	action	Supernatural	resilience	satire
	crime	espionage	storytelling	Adventure	Futuristic	epic	Violence	Apocalyptical	inspirational	mature
Final Category	Thriller	Espionage	Animation	Fantasy	Science Fiction	Heroes	Action	Apocalyptic	Historical	Comedy

*How to read table: The color code does not reflect intensity. Each column has three *Sub-Categories*, which were assigned three color shades. These color shades are then used with each “associated” tag in that column. Tags that are white are the “neutral” tags referred to in the response document.

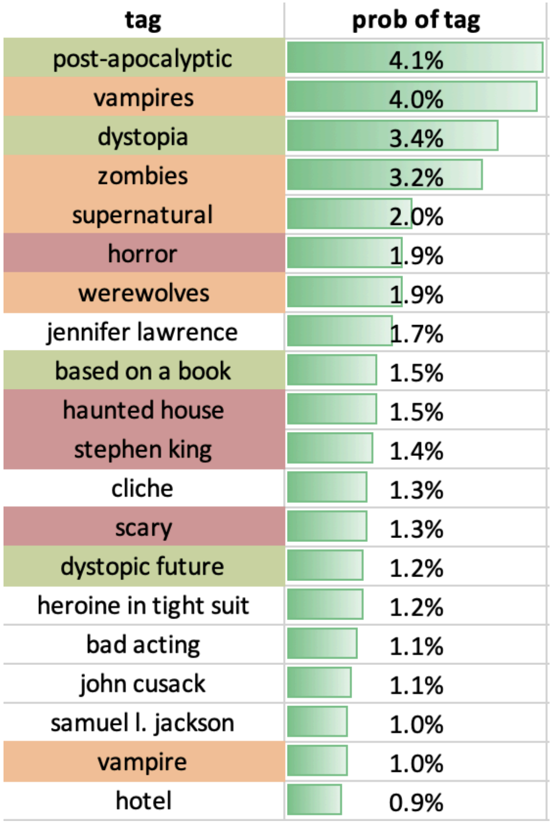
Exhibit 2A: Dissimilarity Scores & Probability Distribution of “The Maze Runner” and 10 most similar movies across 10 Topics



Name of Movie	Thriller	Espionage	Animation	Fantasy	Science Fiction	Heroes	Action	Apocalyptic	Historical	Comedy	Dissimilarity Score
The Maze Runner	4.9%	7.4%	4.3%	6.8%	6.2%	3.7%	3.7%	55.6%	3.7%	3.7%	0.0%
The Twilight Saga: New Moon	5.1%	5.1%	4.2%	5.9%	4.2%	4.2%	4.2%	55.9%	5.9%	5.1%	4.2%
Daybreakers	4.5%	7.3%	4.5%	4.5%	6.4%	4.5%	6.4%	51.8%	5.5%	4.5%	5.6%
28 Weeks Later	3.5%	5.0%	3.5%	6.4%	5.7%	7.1%	7.1%	53.2%	3.5%	5.0%	6.3%
The Conjuring	4.2%	5.1%	5.1%	5.1%	5.1%	4.2%	4.2%	53.4%	9.3%	4.2%	6.9%
Underworld: Evolution	4.5%	3.9%	3.2%	3.9%	3.9%	5.2%	8.4%	59.4%	3.2%	4.5%	8.2%
140800.0%	7.1%	4.7%	3.0%	8.3%	3.6%	4.1%	10.1%	51.5%	4.1%	3.6%	9.0%
Insidious	9.6%	4.8%	5.6%	5.6%	4.0%	5.6%	4.8%	48.8%	4.8%	6.4%	9.8%
The Hunger Games: Catching Fire	6.1%	3.8%	2.8%	9.9%	9.0%	3.8%	2.8%	48.1%	9.4%	4.2%	11.1%
Doomsday	4.7%	7.1%	4.7%	5.5%	11.0%	5.5%	7.9%	45.7%	3.9%	3.9%	12.0%
Resident Evil: Extinction	4.8%	12.4%	4.8%	4.8%	5.7%	4.8%	6.7%	44.8%	6.7%	4.8%	12.9%

*How to read table: The blue data bars represent a probability distribution of tags describing different movies across the 10 topics, while the red data bars represent the similarity scores of the movies when compared to “The Maze Runner”. Note that the lower the similarity score, the more similar the movie is to “The Maze Runner”.

Exhibit 2B: Top 20 tags describing top 10 films most similar to “The Maze Runner”

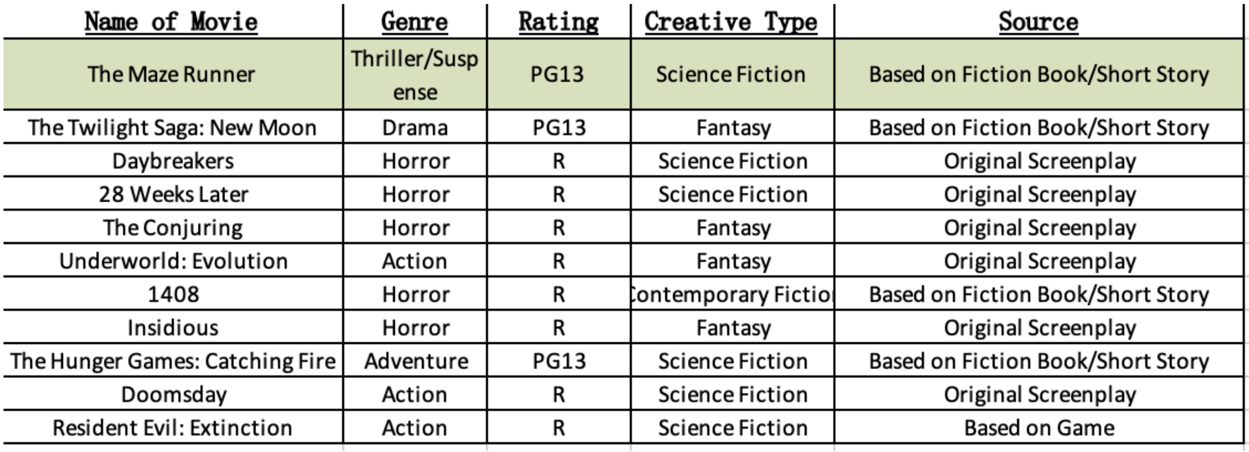


tag	prob of tag
post-apocalyptic	4.1%
vampires	4.0%
dystopia	3.4%
zombies	3.2%
supernatural	2.0%
horror	1.9%
werewolves	1.9%
jennifer lawrence	1.7%
based on a book	1.5%
haunted house	1.5%
stephen king	1.4%
cliche	1.3%
scary	1.3%
dystopic future	1.2%
heroine in tight suit	1.2%
bad acting	1.1%
john cusack	1.1%
samuel l. jackson	1.0%
vampire	1.0%
hotel	0.9%

*How to read table: The green data bars represent a probability distribution of tags describing the top 10 movies most similar to “The Maze Runner” (lowest similarity score). The color coding of the tags describes the following information:

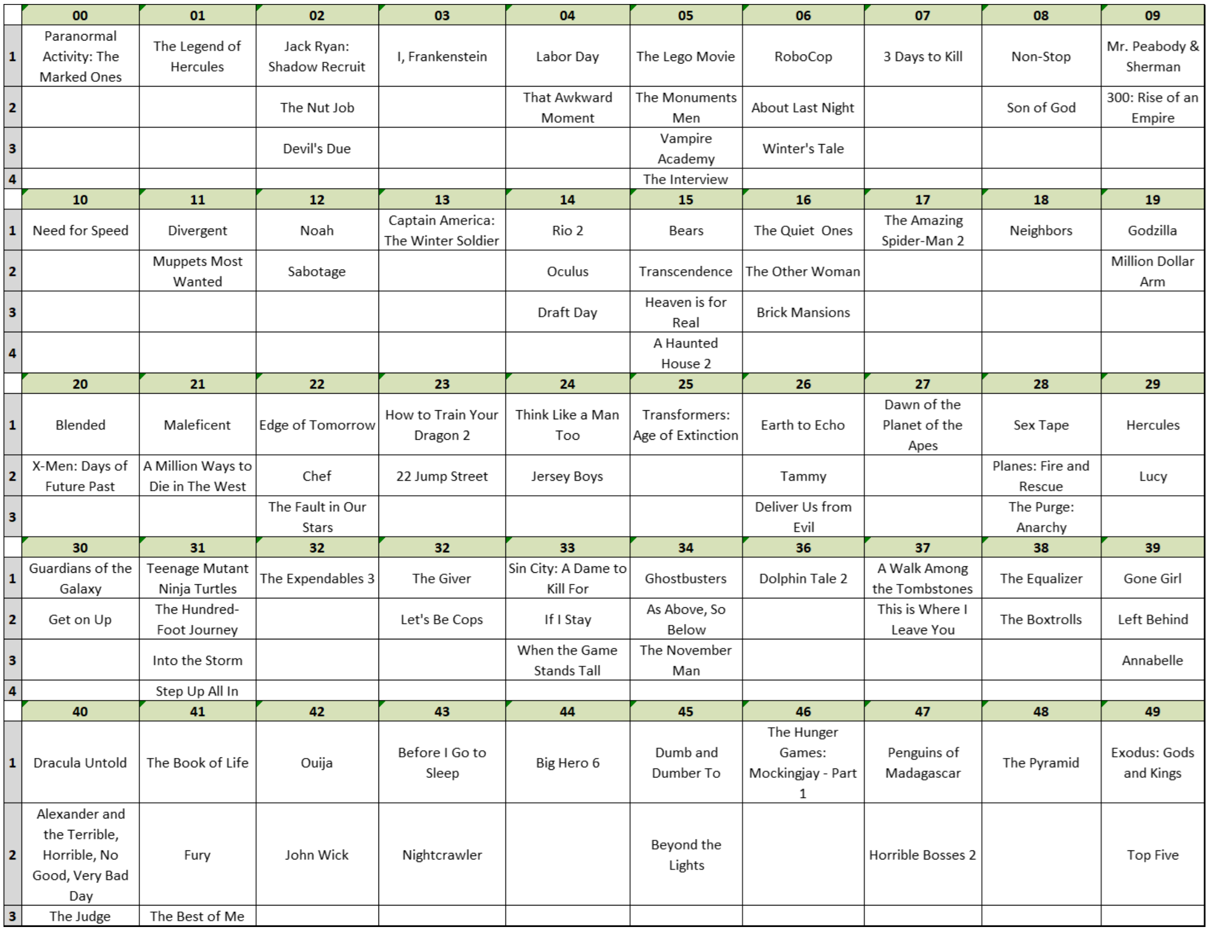
- Green: tags that describe “The Maze Runner” (sci-fi; dystopian nature)
- Orange: tags that describe supernatural movies with monsters
- Red: tags that describe horror films

Exhibit 2C: Description of “The Maze Runner” and its top 10 most similar movies



Name of Movie	Genre	Rating	Creative Type	Source
The Maze Runner	Thriller/Suspense	PG13	Science Fiction	Based on Fiction Book/Short Story
The Twilight Saga: New Moon	Drama	PG13	Fantasy	Based on Fiction Book/Short Story
Daybreakers	Horror	R	Science Fiction	Original Screenplay
28 Weeks Later	Horror	R	Science Fiction	Original Screenplay
The Conjuring	Horror	R	Fantasy	Original Screenplay
Underworld: Evolution	Action	R	Fantasy	Original Screenplay
1408	Horror	R	Contemporary Fictio	Based on Fiction Book/Short Story
Insidious	Horror	R	Fantasy	Original Screenplay
The Hunger Games: Catching Fire	Adventure	PG13	Science Fiction	Based on Fiction Book/Short Story
Doomsday	Action	R	Science Fiction	Original Screenplay
Resident Evil: Extinction	Action	R	Science Fiction	Based on Game

Exhibit 3A: Movies in 2014 categorized by Weekly Launch Dates



	00	01	02	03	04	05	06	07	08	09
1	Paranormal Activity: The Marked Ones	The Legend of Hercules	Jack Ryan: Shadow Recruit	I, Frankenstein	Labor Day	The Lego Movie	RoboCop	3 Days to Kill	Non-Stop	Mr. Peabody & Sherman
2			The Nut Job		That Awkward Moment	The Monuments Men	About Last Night		Son of God	300: Rise of an Empire
3			Devil's Due			Vampire Academy	Winter's Tale			
4						The Interview				

	10	11	12	13	14	15	16	17	18	19
1	Need for Speed	Divergent	Noah	Captain America: The Winter Soldier	Rio 2	Bears	The Quiet Ones	The Amazing Spider-Man 2	Neighbors	Godzilla
2		Muppets Most Wanted	Sabotage		Oculus	Transcendence	The Other Woman			Million Dollar Arm
3					Draft Day	Heaven is for Real	Brick Mansions			
4						A Haunted House 2				

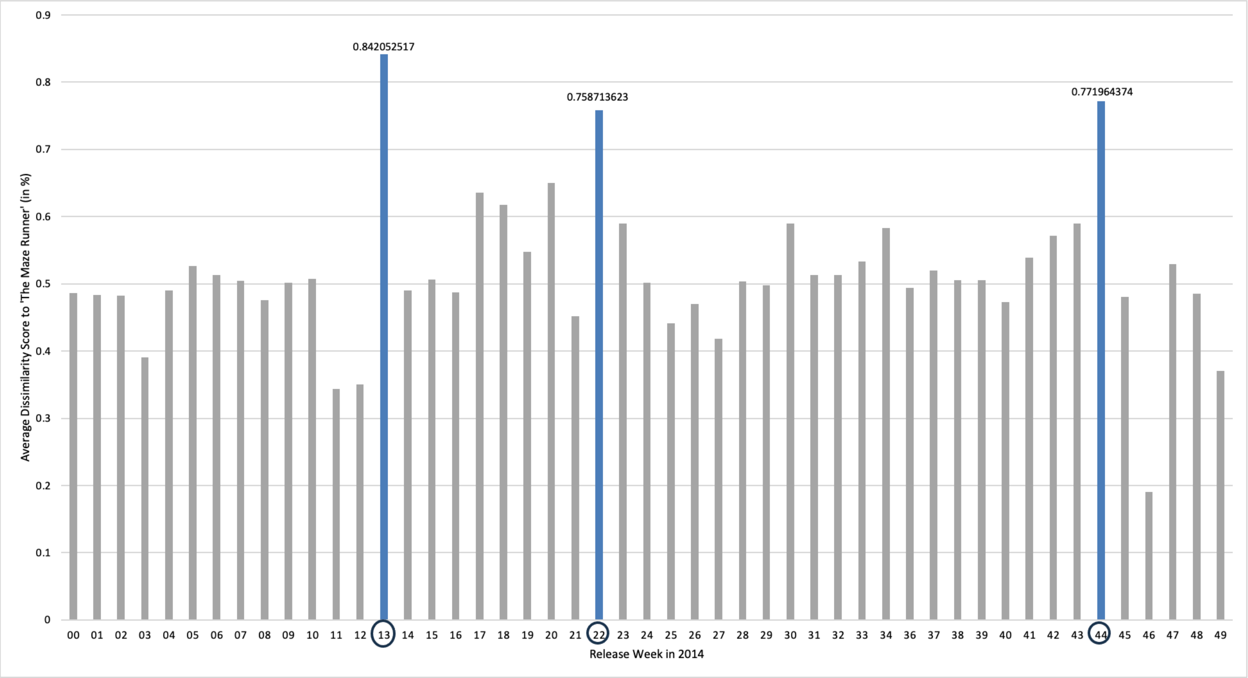
	20	21	22	23	24	25	26	27	28	29
1	Blended	Maleficent	Edge of Tomorrow	How to Train Your Dragon 2	Think Like a Man Too	Transformers: Age of Extinction	Earth to Echo	Dawn of the Planet of the Apes	Sex Tape	Hercules
2	X-Men: Days of Future Past	A Million Ways to Die in The West	Chef	22 Jump Street	Jersey Boys		Tammy		Planes: Fire and Rescue	Lucy
3			The Fault in Our Stars				Deliver Us from Evil		The Purge: Anarchy	

	30	31	32	32	33	34	36	37	38	39
1	Guardians of the Galaxy	Teenage Mutant Ninja Turtles	The Expendables 3	The Giver	Sin City: A Dame to Kill For	Ghostbusters	Dolphin Tale 2	A Walk Among the Tombstones	The Equalizer	Gone Girl
2	Get on Up	The Hundred-Foot Journey		Let's Be Cops	If I Stay	As Above, So Below		This is Where I Leave You	The Boxtrolls	Left Behind
3		Into the Storm			When the Game Stands Tall	The November Man				Annabelle
4		Step Up All In								

	40	41	42	43	44	45	46	47	48	49
1	Dracula Untold	The Book of Life	Ouija	Before I Go to Sleep	Big Hero 6	Dumb and Dumber To	The Hunger Games: Mockingjay - Part 1	Penguins of Madagascar	The Pyramid	Exodus: Gods and Kings
2	Alexander and the Terrible, Horrible, No Good, Very Bad Day	Fury	John Wick	Nightcrawler		Beyond the Lights		Horrible Bosses 2		Top Five
3	The Judge	The Best of Me								

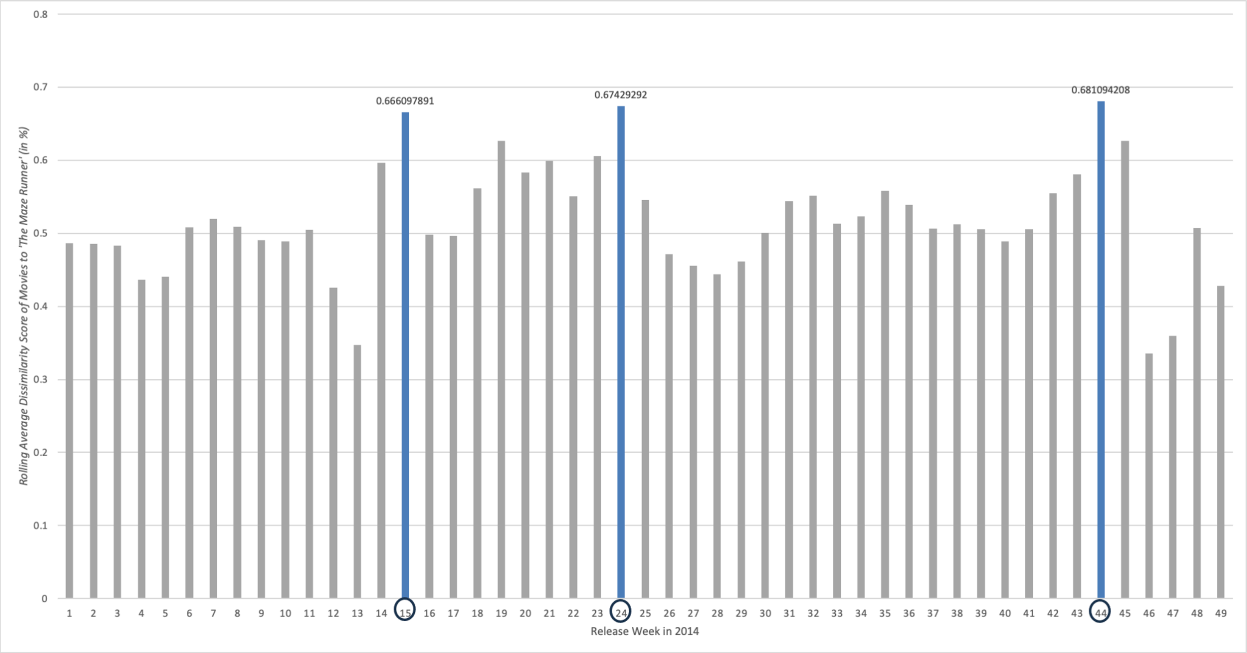
*How to read table: The table summarizes the movies launched in 2014 by weekly launch dates. The coding of weeks follows the ISO format, where week 00 represents Jan 1st to Jan 3rd, and each corresponding number after represents a week after.

Exhibit 3B: Average Weekly Dissimilarity Score of Movies to “The Maze Runner”



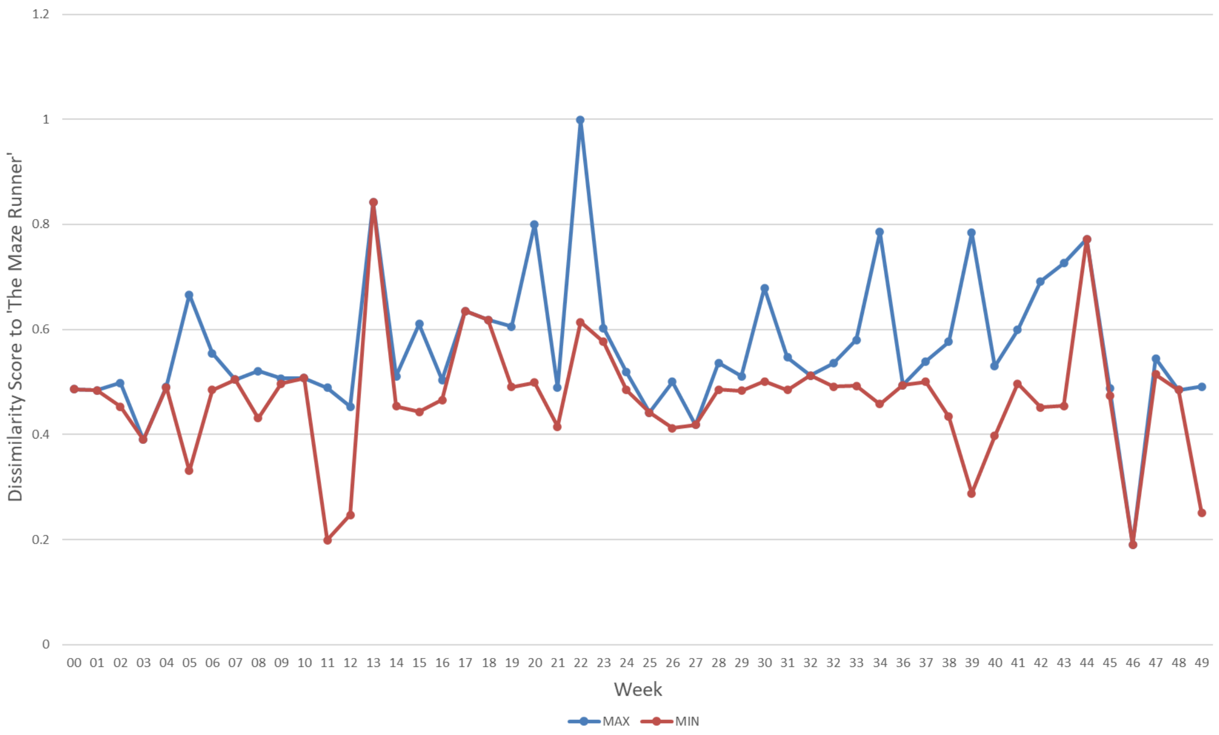
*How to read graph: The graph is about comparing the dissimilarity scores of movies to “The Maze Runner”, categorized by weeks. Each bar represents the average dissimilarity score of movies in that week. The top 3 scores are highlighted in blue.

Exhibit 3C: Two-Week Rolling Average Dissimilarity Score of Movies to “The Maze Runner”



*How to read graph: The graph is about comparing the two-week rolling dissimilarity scores of movies to “The Maze Runner”, categorized by weeks. Each bar represents the average dissimilarity score of movies in that week. The top 3 scores are highlighted in blue.

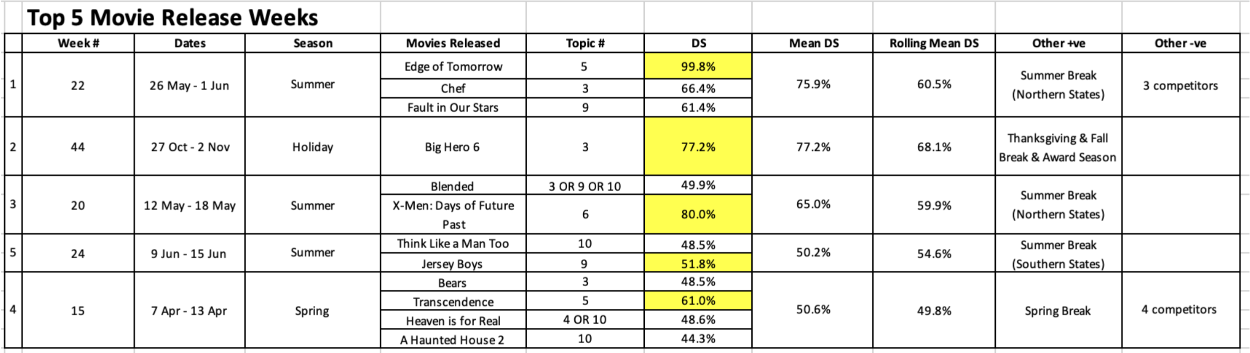
Exhibit 3D: Comparison of Dissimilarity Scores of selected movies to “The Maze Runner”



*How to read graph: The graph is about comparing the dissimilarity scores of movies to “The Maze Runner”, categorized by weeks. Each dot represents the index score of one movie in 2014, and we have selected only the movies with the highest and lowest score in each week.

- Red line: Minimum distance of a certain movie in specific week
- Blue line: Maximum number of a certain movie in specific week

Exhibit 4: Table describing Top 5 weeks of choice and their respective considerations/ metrics



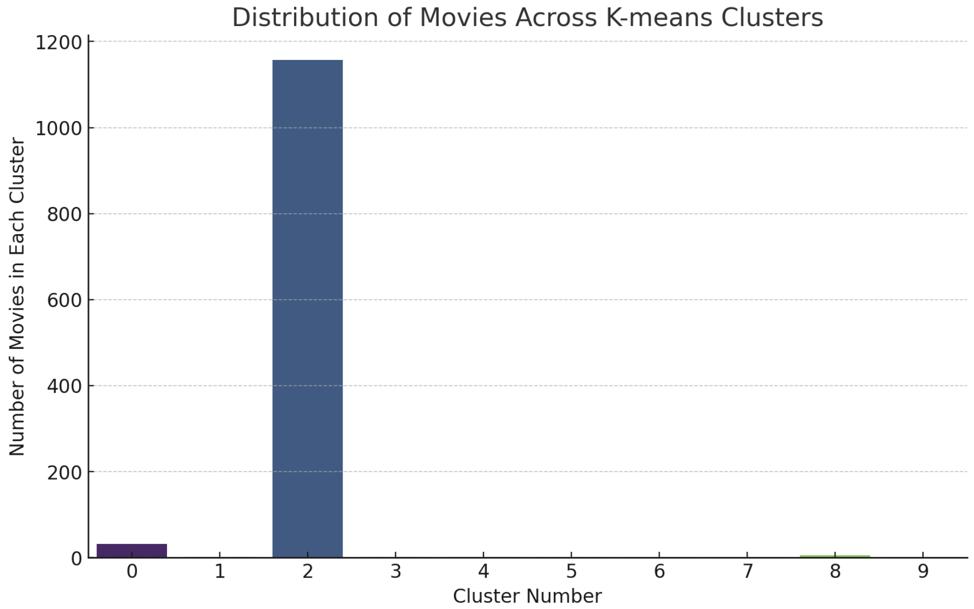
Top 5 Movie Release Weeks

	Week #	Dates	Season	Movies Released	Topic #	DS	Mean DS	Rolling Mean DS	Other +ve	Other -ve
1	22	26 May - 1 Jun	Summer	Edge of Tomorrow	5	99.8%	75.9%	60.5%	Summer Break (Northern States)	3 competitors
				Chef	3	66.4%				
				Fault in Our Stars	9	61.4%				
2	44	27 Oct - 2 Nov	Holiday	Big Hero 6	3	77.2%	77.2%	68.1%	Thanksgiving & Fall Break & Award Season	
3	20	12 May - 18 May	Summer	Blended	3 OR 9 OR 10	49.9%	65.0%	59.9%	Summer Break (Northern States)	
				X-Men: Days of Future Past	6	80.0%				
5	24	9 Jun - 15 Jun	Summer	Think Like a Man Too	10	48.5%	50.2%	54.6%	Summer Break (Southern States)	
				Jersey Boys	9	51.8%				
4	15	7 Apr - 13 Apr	Spring	Bears	3	48.5%	50.6%	49.8%	Spring Break	4 competitors
				Transcendence	5	61.0%				
				Heaven is for Real	4 OR 10	48.6%				
				A Haunted House 2	10	44.3%				

*How to read table: The graph shows the comparison of dates, topics, and dissimilarity score between different weeks.

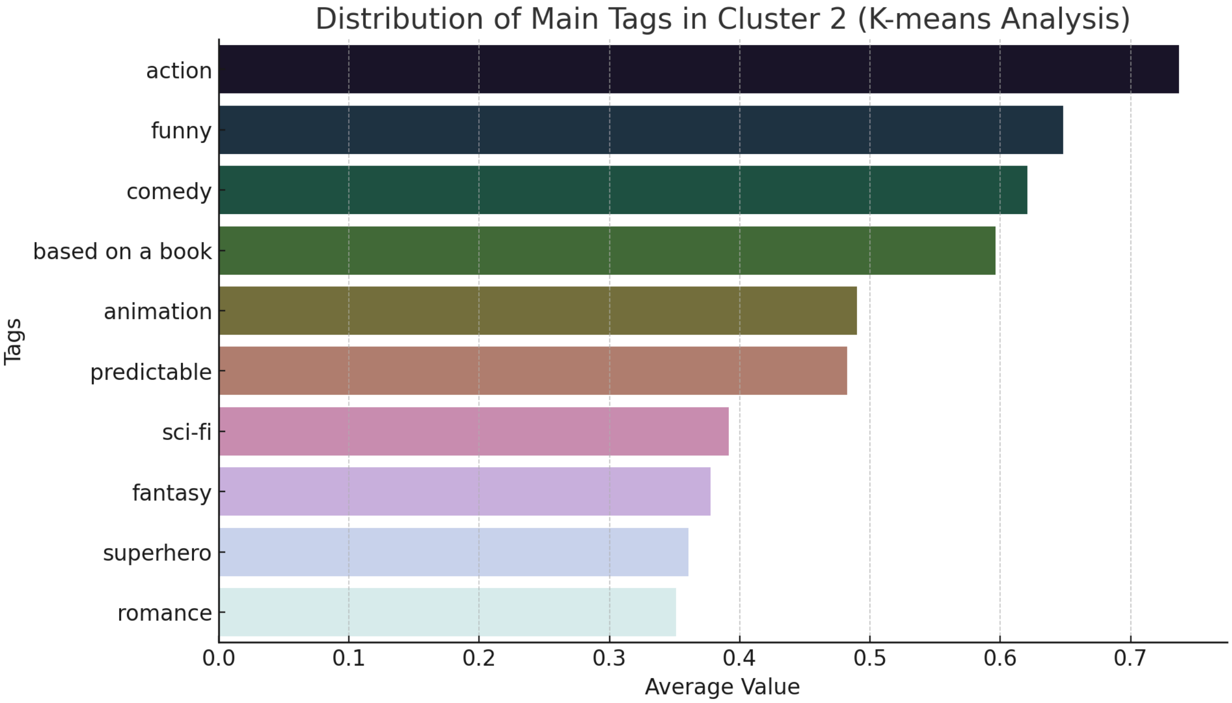
- DS = Dissimilarity Score
- Rolling Mean DS: The average dissimilarity score of a certain week and its previous week.

Exhibit 5A: Distribution of Movies across K Means Clusters



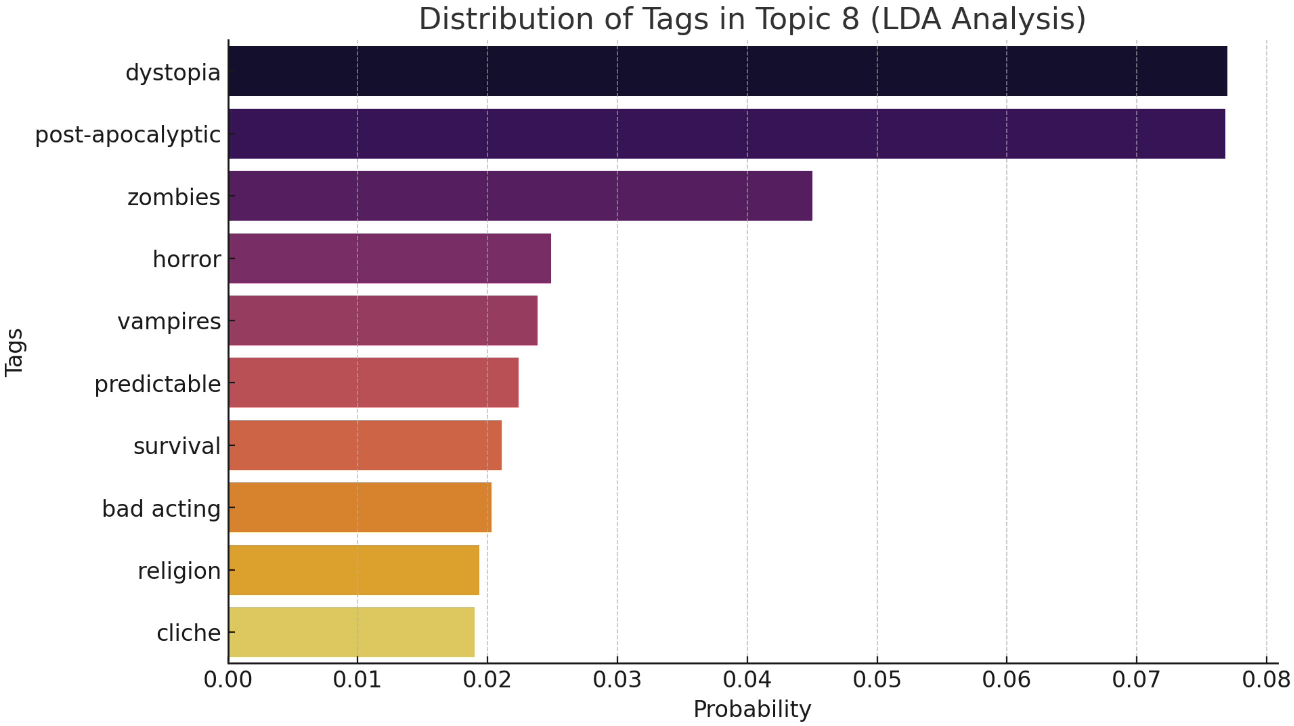
The bar chart showing the distribution of movies across K-means clusters indicates that a vast majority of movies fall into Cluster 2. This would be unhelpful for differentiation because it suggests that most movies share a similar set of attributes, leading to them being grouped together.

Exhibit 5B: Distribution of Main Tags in Cluster 2



The further analysis into the distribution of tags within Cluster 2 reveals a broad range of attributes: action, funny, comedy, based on a book, animation, predictable, sci-fi, fantasy, superhero, and romance. This indicates that Cluster 2 encompasses a wide variety of movies, which might suggest that the clustering parameters could be too broad or that the data itself is not granular enough to distinguish between finer nuances.

Exhibit 5C: Distribution of Main Tags in Cluster 2



The distribution of tags within Topic 8 shows a strong association with tags like dystopia, post-apocalyptic, zombies, horror, and vampires. These tags offer a more specific insight into the type of movies that would fall under the same category as "The Maze Runner."