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 ( (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**

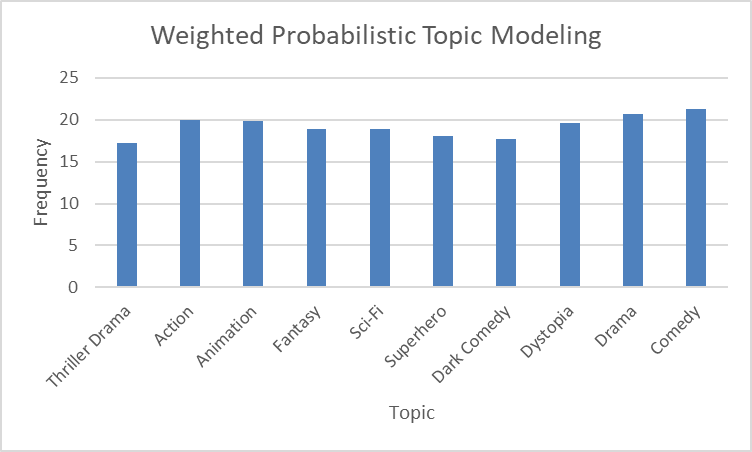


Figure 1

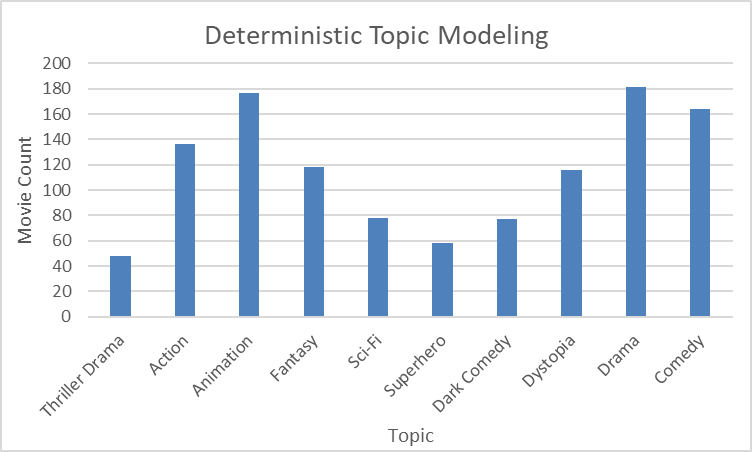


Figure 2

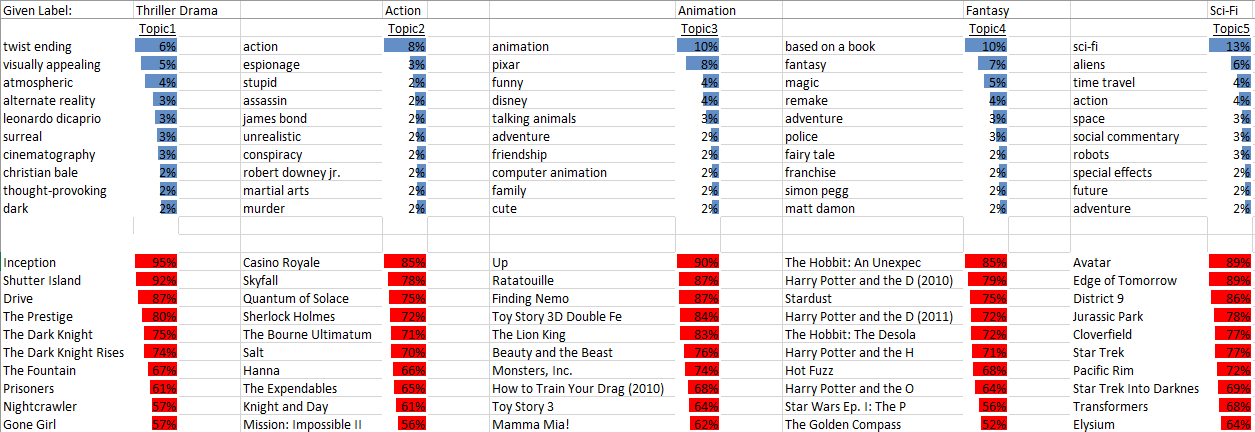


Figure 3

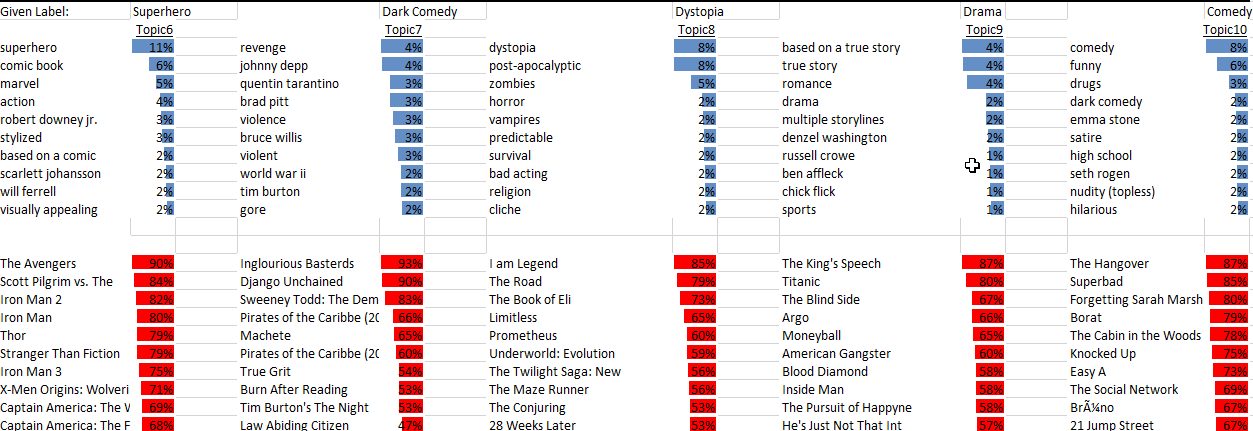


Figure 4

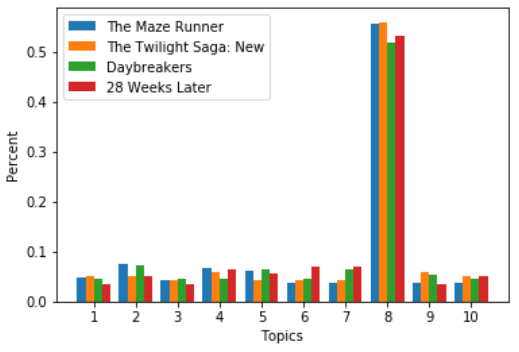


Figure 5

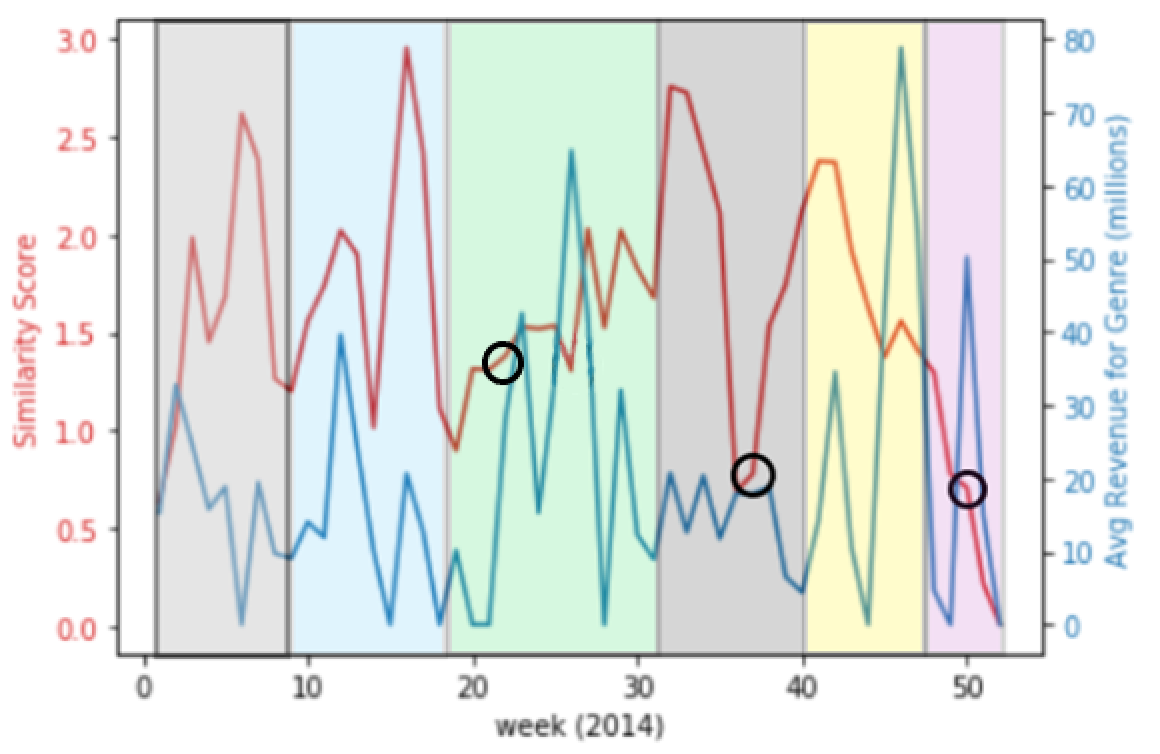


Figure 6

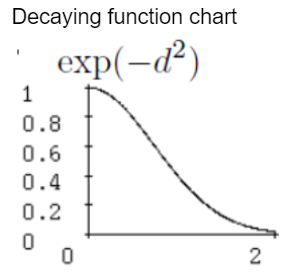


Figure 7

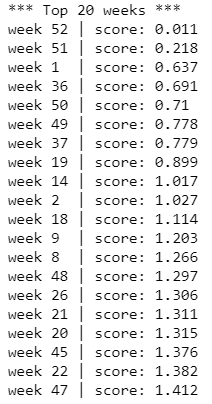


Figure 8

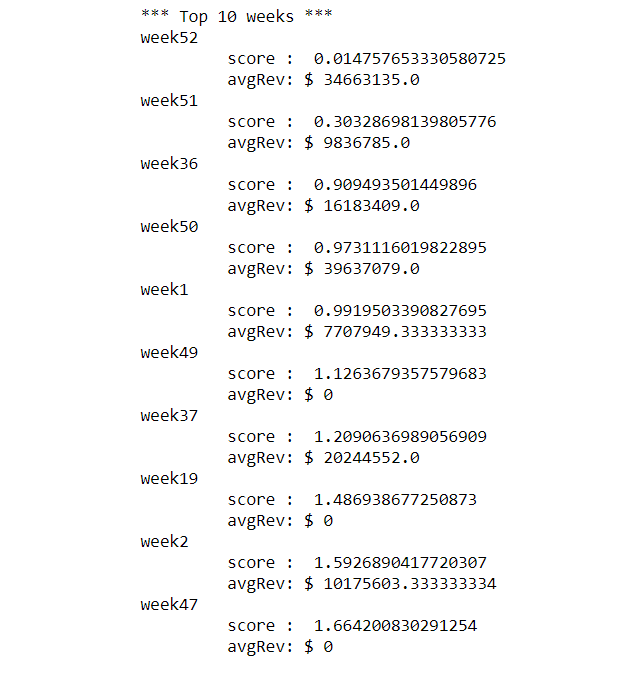


Figure 9

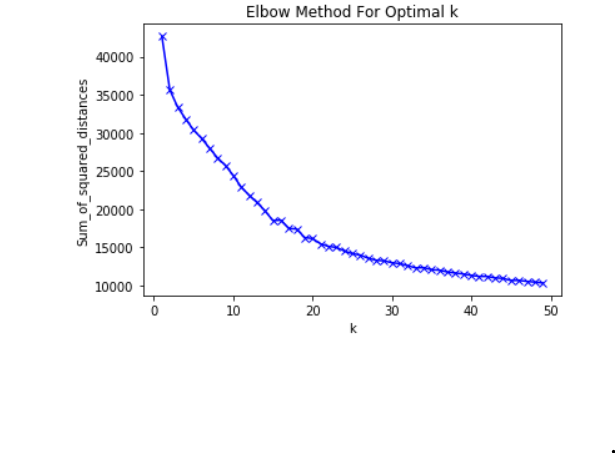


Figure 10

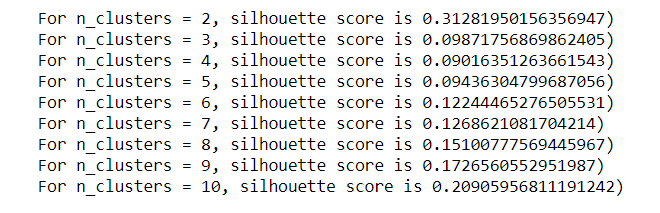


Figure 11

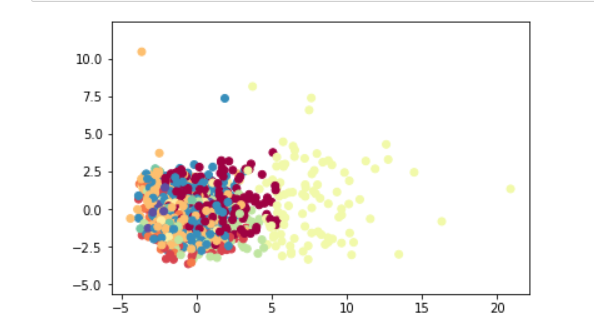


Figure 12