***Background:***

Movie studios make a large number of decisions both pre- and post-production. The pre-production includes decisions about which movies to produce or “greenlight”, setting production budgets, staffing, and a host of creative decisions. The post-production includes decisions about when to release the movie, the release strategy (wide or rollout, international or domestic market first or simultaneous), the promotional strategy, and the marketing budget. Often the marketing budget can be close in magnitude to the production budgets, which means that post-production decisions are just as important as pre-production decisions.

While analytics has made inroads into the movie production process, studios have been notorious for their intuitive approach to decision making. In this exercise we consider how analytics could provide some insight into the scheduling decision for movie releases. Consider that a major studio schedules the release of 10-20 films per year and must maximize the performance of their films by gaining as much market share as possible, but also avoid cannibalizing their **own** **film’s sales**. Decisions about release date are heavily driven by two considerations: seasonality and competition. Release decisions are often guided by intuition and informal rules. For example, you would rarely see animated kids films and R-rated comedies open in the same week. Some rules of thumb are:

* Release the biggest movies during the busiest times of year like Summer and Christmas (or Winter). Summer tends to see the Release action-adventure movies during summer when students (teenagers) are on vacation and school is not in session leaving more free time for movie going. Christmas skews slightly towards more family-friend movies (as more families are together for holidays)
* Do not release major blockbusters in the same week.  You may see an animated kids film and an R-rated comedy open on the same weekend, but it would be rare to see two major romantic comedy’s open in the same week. To avoid this from happening studios will set their release schedules years in advance in hopes that other studios will keep similar films away from their opening dates. When you hear about studios “shelving” a movie, and not releasing it, that may mean that they believe the competition is stronger and then want to find a more favorable time to release. There are numerous trade publications like Boxoffice Pro, Boxoffice Mojo, The-Numbers, Boxoffice Analyst, Screentrade, Variety and Hollywood Reporter that are used to release information. For example, The-Numbers has movie releases posted for up to six years in advance, although the calendars for the next year are typically quite detailed while those for three or four years out only have big projects.
* January and February (Winter) is a slower time for movie theatres and studio tend to “dump” weaker films during this time when competition is softer. Sometimes January and February are referred to as “dump months” due to lower commercial and critical expectations.
* Animated fare comes out in March and April (Spring) because of spring breaks and the Easter holiday.
* August and September sometimes are also referred to as “dump months” as well. Although in years with intense box office competition from blockbusters some films may open in the first or second weekend of August to avoid competing with these films. One notable exception for Fall releases is that dramas and horror films (especially) tend to be released in the Fall.
* “Awards Season” generally refers to late September through early December when most “Oscar-contender films” tend to come out. Some Indie films and mid-budget films may aim towards award season with a limited release schedule and then go nationwide based upon the reception they get. Films shown after January 1 are ineligible for Oscars until the following year which means they are typically forgotten by the critics. A rare exception is “The Silence of the Lambs” which was released in January 1991 and went on to win the 1991 Academy Award for Best Picture.
* There are a few dates that studios shy away from. Specifically early September after Labor Day is back to school–which can depress movie attendance, and also the week after Thanksgiving is avoided since anything worth being released comes out before Thanksgiving to take advantage of the holiday.

Of course these are just rules of thumb, and there are many films that succeed that do not strictly follow them. For example, Hunger Games opened in April with over $400 million in box office revenue and “Alice in Wonderland” did over $300 million and was released March 3, 2010. Studios can also stretch, “Summer” releases by moving films into late April, like Universal’s 2011 release of “Fast Five”. Also, there can be other big events like the Olympics, World Cup, and presidential elections that can also come into play.

***Instructions:***

There are a complete set of data of all the movies and tags to recreate a topic model.  This script will estimate both a k-means and topic model.  To simplify your task an Excel spreadsheet that has all the relevant output from the topic model is provided.  This eliminates the need to execute the R script (which is long and time consuming).  This spreadsheet has the output for a topic model with 10 topics:

[topics\_analysis\_10.xlsxPreview the document](https://canvas.cmu.edu/courses/9526/files/3551601/download?verifier=8poeSHj0H3n1tAzD9qHDFXI58BufXttnbSTVvmJA&wrap=1)

Specifically, there are five sheets that provided in this spreadsheet (you only need the first two sheets along with the movie date releases in opus\_movies.txt for this assignment):

* "topics\_allterms.txt": conditional probabilities for Pr( word | topic ) = matrix of 962 terms (rows) x 10 topics (columns)
* "topics\_allmovies.txt": conditional probabilities for Pr( topic ) for each movies = matrix of 1153 movies (rows) x 10 topics( columns)
* "topics\_prediction.txt": example of predicted word count for "The Maze Runner" using the model compared with actual
* "opus\_movies.txt": original list of 1282 movies (some movies are dropped -- please ignore those that are missing topics)
* "opus\_movielens\_tags.txt": original tags for the 1282 movies (notice that each row is a tag X movie combination, along with the count of the tag)

*More background about the data and scripts:*

First, you are provided with a dataset of movie characteristics for all major releases in the US domestic market from 2006 through 2014:

[opus\_movies.txtPreview the document](https://canvas.cmu.edu/courses/9526/files/3508799/download?verifier=n73D9VA7OxzuEuhmIUPWLpxZoAXs9t5OK3nNzLAf&wrap=1)

The following variables are included in this dataset:

|  |  |
| --- | --- |
| Variable | Description |
| Display\_name | Movie title |
| Odid | Opus movie identifier |
| Imdb\_id | IMDB movie identifier |
| movieId | Movie Lens movie identifier |
| genre | Genre of the movie (dummy variables have been created that indicate whether genre is Action, Adventure, BlackComedy, Comedy, Documentary, Drama, Horror, Musical, Performance, Romantic Comedy, Thriller Western |
| rating | MPAA rating (dummy variables have been created that indicate whether rating is G, PG, PG13, or R) |
| sequel | Indicator if movie |
| franchise | If the movie is a sequel then what franchise does it belong to? |

Second, for each movie you have a set of tags that describe characteristics of the movie that has been collected by MovieLens:

[opus\_movielens\_tags.txtPreview the document](https://canvas.cmu.edu/courses/9526/files/3508798/download?verifier=DA3EWedj8QnErKZRbNwFgVSNEVl3ZdK1woqyf7rO&wrap=1)

This dataset has a series of keywords or tags like "Action" or "Based on Novel" that are used to describe the movie along with how many times the keyword is used to describe the movie.

Third, you are provided an R script that can estimate a topic model (the default is 6 topics, but change to 10 or higher) and save the results to a file from which you can use Excel to complete your answer:

[Movie\_Analysis.R](https://canvas.cmu.edu/courses/9526/files/3508797/download?verifier=Re3tIy48aXlv7WwgAPug4VBKNR8g7YJxtwORYrPR&wrap=1" \o "Movie_Analysis.R" \t ")

A suggestion is that you use the similarity scores to find periods when your selected release does not have any close competitors.

There is an additional keyword dataset and another R script with help on computing variations on the models posted here:

[opus\_keywords.txtPreview the document](https://canvas.cmu.edu/courses/9526/files/3508775/download?verifier=II94on4LQMPARfrB0rlfDejU8o4Px0Qj96EJj2tg&wrap=1)

[Movie\_ExtendedAnalysis.R](https://canvas.cmu.edu/courses/9526/files/3508768/download?verifier=ApEnV34AFlKOaIFbHF0P5hbolcZSJkkt4jbjQgW9&wrap=1)

*Assignment Instructions:*

This exercise is to be completed in your group as listed on Canvas.

Please provide a clear, concise, and well organized essay that addresses at least the following questions.  You are free to address other issues in the case as well.  The intent of the assignment is to have you think critically about the business problem faced in the case and how it can be solved through data mining.  Analyze the quantitative material in the case to support your answers.  Spend most of your time in defining and defending your recommendation for what should be done.

Good answers may require assumptions of facts that may not be presented in the case.  You are welcome to make these assumptions, but please state these assumptions and briefly justify why that are reasonable.  Also, you may use whatever resources you can locate to provide further information about this industry or the web in general.  Please reference your sources.

Your response must be typed, double spaced, with one-inch margins, and a 10 to 12 point font size, and must be a PDF file.  This writeup must not exceed 3 pages in length.  You may attach exhibits, tables, and/or graphs to support your arguments.  These supporting materials must be referenced in the text and do not count toward the 3 page limit.

***Required:***

Your task is to aid the production company Gotham Group in deciding when to release the target movie, “The Maze Runner”, by comparing it to other films that are scheduled for release in 2014. You may assume that the schedule of the other releases is known (see the release dates given in "opus\_movies.txt"). The studio wants to avoid releasing the movie in a week in which it is too “similar” to others. What recommendation would you give for the release date of "The Maze Runner"?

To help answer this problem I suggest you do the following:

1. Interpret the 10 topics from the LDA analysis.  This will help you understand/label the topics and movie "clusters".
2. Prepare a measure that summarizes the similarity/dissimilarity between your target movie "The Maze Runner" and every other movie.  (Hint: try to compute the "Euclidean distance" between the 10 dimensional topic score for "The Maze Runner" and every other movie.)  What are some really similar movies to "The Maze Runner" -- does this model make sense?
3. Create a list of weekly launch dates for 2014.  Compute some measure of average similarity for each week.  (Hint: In excel you could create a lookup table for the distance between the "Maze Runner" and each movie released that week.)  You'll need to make some judgement about whether to average the similarity metric across movies within a week or just choose the closest movie each week.  Complicating the release is that most weeks the movie from the previous week are also likely to still be performing well at the box office, so you might want to consider both movies launched this week as well as those released the previous week. And potentially the following week??
4. Identify the best 3 or 4 choices of weekly launch dates.  Illustrate and explain your results and recommendation.

Optional: How would the results look different if you used a k-means or Genre to make this decision?  How would your results change if you used 15 or 20 topics?

**Rubric**

Movie Scheduling

| Movie Scheduling | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| This criterion is linked to a Learning Outcome Interpretation of LDA Topics | |  |  | | --- | --- | | 2.0 pts  Full Marks | 0.0 pts  No Marks | | 2.0 pts |
| This criterion is linked to a Learning Outcome Comparison of Release Dates | |  |  | | --- | --- | | 2.0 pts  Full Marks | 0.0 pts  No Marks | | 2.0 pts |
| This criterion is linked to a Learning Outcome Recommendation of release  Is your recommendation compelling and convincing? Well supported? Visually appealing? | |  |  | | --- | --- | | 4.0 pts  Full Marks | 0.0 pts  No Marks | | 4.0 pts |
| This criterion is linked to a Learning Outcome Metric for computing similarity/dissimilarity of movies | |  |  | | --- | --- | | 2.0 pts  Full Marks | 0.0 pts  No Marks | | 2.0 pts |
| Total Points: 10.0 | | | |