**DSBA/MBAD 6211, Fall 2020**

**Final Exam-Part II Text Mining**

**Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\*\*\*6:35 PM-7:25 PM\*\*\***

**\*\*\* Please submit Part II before 7:25 PM \*\*\***

**\*\*\*Please DO NOT share the dataset as well as questions with anyone\*\*\***

moviedata is a movie dataset randomly collected from the IMDB website. Please analyze synopses of those movies and answer the following questions (36 points).

* ID: ID for movies randomly sampled from the IMDB website
* Synopsis: The synopsis of each movie
* Title: The title of each movie
* MPAA Rating: Viewer rating
* Year: The year when each movie was released in theaters
* Viewer rating: IMDB user rating of each movie

**Variable and model naming requirements:**

* + Please include your ***name initials*** to the data frame names as well as model names in your R coding.
  + Please instance, in my coding, I would name the corpus as ***KZcorpus***, and name the data-term matrix as ***KZdfm*** etc.

**Canvas submission.** You need to submit ***two separate*** documents via the Canvas Final Exam Part II submission link:

* Word document: please provide your answers in the Word document, and copy/paste your R codes at the end of the document.
* R coding file.

**Questions:**

1. After removing punctuations, default stop words, and stemming, what are the top 25 features? (2 points)
2. Do you want to remove any additional user-defined stop words? If so, please create the ***stopword1*** list, list those words here, and remove them from your DFM. (2 points)
3. A movie critics further examines the dataset, and recommend removal of the following commonly used terms: (4 points)

* *film, movi, play, even, just, go, get, like, time, make, charact, scene, show, 1, 2, year, come, may, john*
* Please create the ***stopword2*** list, and use dfm\_remove to further remove those common terms from your DFM.

1. Please create a word cloud with the 150 most used terms. (5 points)

* Please summarize a few common things covered by the movie synopses
* Paste the visualization below.

1. Please further remove highly infrequent terms. (5 points)

* Specifically, we only keep terms that appear in the entire corpus at least 15 times, and appear in at least 5 different movies.
* What is the dimensionality of the trimmed DFM?

1. Which movie is most similar to Movie 91 (“Batman & Robin”), based on the correlation similarity measure? (5 points)
2. What are the top 8 terms that are most related to “school”, based on the cosine similarity measure. (5 points)
3. Please perform topic modeling with 4 topics, and keep 8 most relevant terms per topic. (8 points)

* Provide the term/beta plots for four topics.
* Try your best to summarize those four topics.