# Assignment 2 – Topic Modeling and Sentiment Analysis

CIS8045 Unstructured Data Management · MM2 · AY2023/2024 Spring

In this assignment, you will focus on training a LDA model for topic modeling and using TF-IDF to build a machine learning model for sentiment classification.

The dataset **IMDB_Reviews.csv** is given. The data was collected from movie rating website, IMDB (www.imdb.com). The dataset is about movie reviews. The sample contains around 50K movie reviews. The data has two columns: the textual movie review (**review**) and the sentiment of review (**sentiment**).

Please create a Jupyter Notebook on your own and complete all your works there. After loading the data in the notebook, select a random sample on your own for the assignment tasks, which means every student will work on a unique random sample and submit a Jupiter Notebook file with unique codes/outputs. The random sample should include 5,000 movie reviews. Please refer to Assignment 1 for the method of random sampling.

You will complete two tasks in this assignment.

### Task 1. Topic Modeling

In the first task, train a topic model with the provided data to discover the topics in movie reviews. This task includes the following requirements:

1.1 Make sure you preprocess the text before training the topic model.
1.2 Train the LDA model with at least 10 topics. Print out the list of topics and at least 20 keywords for each topic (i.e., topic-to-words distribution).
1.3 Pick at least 5 topics from your LDA model, and visualize each topic with word cloud. In a separate word document, include the list of keywords and visuals of the picked topics, and interpret/explain the meaning of each topic.

### Task 2. Use TF-IDF for Sentiment Analysis

In the second task, build a machine learning model to predict the sentiment of movie reviews (positive or negative). This task includes the following requirements:

2.1 Make sure your text is preprocessed before building TF-IDF representation.
2.2 Build TF-IDF representation from the review text and split the training and testing set by 80%/20%. Then training a machine learning model on the training set to predict sentiment of review text. Use a machine learning model covered in the class.
2.3 Make predictions and evaluate your model on the testing set. Print your classification report to show the prediction metrics.

## DELIVERABLE

1. A Jupiter Notebook file containing the completed code and the printed results.
   - Given the random sampling process in this assignment, the probability is extremely low that two students will get the same set of topics and prediction metrics. If two notebooks produce the same results, we assume that both students did not do independent work.
   - If helpful, you can add necessary comments to explain your code and/or results.
2. A word document containing your picked topics from the LDA model, with the keywords, visualizations and interpretations of these topics.

Submit your notebook in iCollege before **March 31, 11:59 pm**. Name your notebook as **{LastName}_{FirstName}.ipynb** and word document as **{LastName}_{FirstName}.docx**

Please refer to the course syllabus for the late submission policy, plagiarism policy, and AI-aided tool policy.