CSC498 – Introduction to Natural Language Processing
Final Project: Affective Movie Classifier and Recommender System
Due date: Thursday December 17 - Midnight
Submission: Blackboard upload.

1. Build an Affective Movie Classifier and Recommender System.

In the present assignment, you are required to build an Affective movie classifier and recommender system, that is, a system that first, classifies/predicts movies based on emotional reviews, and second, recommends movies based on people's emotional opinions and the movie's genre.

We'll call the system **Emovie**. You will work with the dataset given to you on the blackboard. The assignment consists of the following modules:

Module #1: Load and Preprocess text data.

In the first module, load the dataset "**movie_reviews.csv**", posted for you. The dataset consists of 3 columns: "**review | sentiment | genre | title**":

1. Preprocess and normalize your text:
    a. Tokenize
    b. Remove punctuations, stopwords, non-alphanumeric characters, etc…
    c. Lower case the entire text.

Module #2: Classifying

In the second module you will create a classifier using Naïve Bayes learning model. To that end:

1. Create the feature set, using TF-IDF word vectorizer, from the '**review**' text.
2. Standardize your dataset
3. Print the dataset.
4. Split the dataset into training and testing set.
5. Create, fit and train the classifier
6. Predict the sentiments of the testing set.
7. Evaluate your classifier's sentiment predictions using the accuracy measure.
8. Plot the confusion matrix.

Module #3: Recommending Movies based on Genre and Emotion.

In this third module, you will create a movie recommender system based on sentiment and genre. To that end:

1. Create a new feature set using TF-IDF, based on both columns: “sentiment” and “genre”.
2. Compute the cosine similarity matrix that consists of the similarity values between the movies, example:

$$
\begin{array}{c} \\ Movie_1 \\ Movie_2 \\ Movie_3 \\ \vdots \\ Movie_n \end{array}
\begin{array}{c}
\begin{array}{ccccc} Movie_1 & Movie_2 & Movie_3 & \cdots & Movie_n \end{array} \\
\left(\begin{array}{ccccc}
1 & 0.158 & 0.138 & \ldots & 0.056 \\
0.158 & 1 & 0.367 & \ldots & 0.056 \\
0.138 & 0.367 & 1 & \ldots & 0.049 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0.056 & 0.056 & 0.049 & \ldots & 1
\end{array}\right)
\end{array}
$$

3. Finally, create a function that takes as input the title of a movie, and returns a list of movie titles that are similar to the title you have input, i.e., that have the highest TF-IDF values. Example:
   Input = “Wuthering Heights”
   Returns = “Never let me go” (and more…)
   ……

All the best!