# Introduction

In this course work the team collectively got together trying to solve the IMDB problem using statistical Natural Language Processing techniques , we have to admit that the questions and their chronological sequence helped shape our thinking to come out with a good working framework to solve this challenge , to the best of our ability.

As we started to understand the data that we were given , visualizing it gave us valuable insights on the action plan that we took during the coursework, we became very clear on the data bias toward a specific score, the relevance of some features , and the diversity and magnitude of normalization that we need to do (HTML Tags, null values,duplicates..etc) , which we did in question 2 and we have documented several function in our code to deal with those issues.

After normalizing the data, we did a research on the best text vectorization technique we can use for this data, we experimented several techniques and several libraries to reach an optimal vectorization technique, we did document our findings and also tested them with our model in later questions.

On then, we tried to came up with a method to apply some of the techniques we have learned during the course to come out with a appropriate model , from estimators selections, Pipelines generation…etc, We tested that on the validation data set as a true measure of our model effectiveness

We finished this by applying Topic Modeling to the data and generating further insights , You will find our all our research summary detailed in section 6.

Note from the team : There were a great deal of learning and collaboration during the coursework, from applying principals we have learned in the class and researching things helped us conclude what is presented in this document . It was a great learning and experience for all team members