Dr. Tony Diana
DATA 602 Introduction to Machine Learning

Practice Exercises | Lecture Week 9

Natural Language Processing

Use the train.tsv file at
https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data

- Load the data, show the first five rows, check the data, and provide information on the data.
- Show the distribution of review documents.
- Generate the document term matrix by using scikitlearn’s CountVectorizer. From nltk.tokenizer import RegexpTokenizer. ‘Sentiment’ is the data. Train, test and split.
- Import the MultinomialNB module and create a MNB classifier object using the MultinomialNB() function. Import scikit.learn metrics to compute the accuracy and fit the model before computing the accuracy. What is the outcome?
- Use TfidfVectorizer-transformed data and split it into training and test datasets.
- Build the text classification model using TF-IDF. First, import the MultinomialNB module and create the MNB classifier object using the MultinomialNB function. Fit the model on the training set and perform the prediction. Is the accuracy better?