**Drexel University**

**College of Computing and Informatics**

**INFO 371 – Data Mining Applications**

**Course Project**

**Due Date: Friday, Dec. 11, 2020**

**A. Team/Individual**

You canwork with another student or two on this project. Please inform the instructor of your team membership before working on the project.

**B. Tools**

Weka or other machine learning tools.

**C. Objective**

This project presents a practice opportunity for you to build classification models for sentiment analysis.

**D. Data Source**

The dataset is a tiny subset of the Rotten Tomatoes movie review dataset which is a corpus of movie reviews used for sentiment analysis. The original dataset is available at: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/

In the project, about 10% of the data is extracted and converted as the Weka’s arff format. You can download the dataset from the course shell: movie\_sentiment\_analysis\_small.arff

**E. Requirements:**

1. Build three types of models:
   1. KNN
   2. Naïve Bayes
   3. Decision Tree
2. For each type of model, use different values of its parameters to compare the performance. For example, under KNN, you should vary the value K and compare the results.
3. For each model, split the data into 80% and 20%. Using 80% for training the remaining 20% for test.
4. Remove unnecessary attributes before building model.
5. For text processing, use binary, TF, and TFIDF on the Phrase attribute.
6. Overall, you should have more than 3 models using the combinations of text processing, types of models, and different parameter values.
7. Record the time spent on building and testing.
8. Summarize the results nicely in tabular format.

**F. Discussion**

Compare the results from the different models, with different parameters, and perhaps using different variables. Which model appears to be better (best) given the stated problem and identified data? If it predicts correctly, what (in the model and/or in the data) are helpful? If it does not predict the correct answer, what (in the model and/or in the data) are missing?

**G. What to Hand In**

1. A well-structured report documenting the steps, data analyses, results, and discussions in MS Word or PDF format. Your report must contain the following content:
   1. For each step, describe the requirements and show the results including necessary screenshots. For each table and figure, you must add sufficient commentary to explain what the table or the figure is about.
   2. Detailed descriptions about the design of the models. Discussion about their pros and cons.
   3. Necessary intermedia steps or raw data to allow any readers to reproduce your results.
   4. Performance evaluation including confusion matrix, accuracy, precision, recall, and F1-Measure.
   5. A discussion section including the comparisons, analyses, and future considerations.
   6. A conclusion section describing lessons learned from the project.

**H. How to Hand In**

1. Please name your report file as **INFO371-project-yourFirstName-yourLastName.docx**.
2. Submit your report file through the course website in the **Blackboard Learn** system.

**I. When to Hand In**

1. Submit your report no later than **11:59pm** in the due date.
2. There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

**J. Written Presentation Requirements**

Images must be clear and legible. Assignments will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be either 11 or 12 point.