MGTA 415 Homework 1

Instructions

In this assignment, you will pre-process the text and train a text classifier using different feature representation techniques.

You should submit the final write-up and code for this assignment. The write-up should be a PDF file that includes experimental findings and insights. Please submit your solution **by the beginning of the week 5 lecture (Feb 9).** Please complete the homework individually. Please include instructions on how to run the code and/or descriptions about your implementations if you think it will be helpful.

You will need one dataset for this homework (available on Canvas): **FinancialPhraseBank (FPB.csv).** This dataset contains the sentiments for financial news headlines from the perspective of a retail investor.

The classifier should be trained and tested on the FPB dataset. Shuffle the FPB data with random seed 42, and split it into training, validation, and test splits, with a 80/10/10% ratio (e.g., use `random state` in `sklearn.model_selection.train_test_split.`) Note that the data file is not encoded with the default character encoding `UTF-8`. You may need to specify the encoding to use `ISO-8859-1` when you load the file.

Problem 1: Text Pre-Processing (10 points)

Pre-process the text data using text preprocessing techniques such as tokenization, stemming, etc. Be sure to make each step clear in your code. In the writeup, show the results of tokenization and stemming for the first 5 sentences in the original FPB.csv file.

Problem 2: Bag Of Words (20 points)

Train a text classifier (e.g. logistic regression) using the following document representation techniques and report AUROC, macro-f1 score, and micro-f1 score on the test set. You are highly encouraged to pay attention to details. If your model receives a warning and fails to converge, think about ways to address it.

i. Each document is represented as a binary-valued vector of dimension equal to the size of the vocabulary. The value at an index is 1 if the word corresponding to that index is present in the document, else 0.

ii. A document is represented by a vector of dimension equal to the size of the vocabulary where the value corresponding to each word is its frequency in the document.

iii. Each document is represented by a vector of dimension equal to the size of the vocabulary where the value corresponding to each word is its tf-idf value.