Grading Criterion of Assignment 2

Total Score: 100 points

Part 1: Computing TF-IDF score of a term in a set of documents (25 points)

Task1:

Given a training set of documents (e.g., tweets), please implement a function to compute the TF-IDF score of each term in the training set.

What to Turn In:

(1) The source code of the function(s) to finish this task.

Grade Criterion:

1. Whether submitted code can run successfully.
   - Successfully: 25 points
   - Unsuccessfully:
     -- Term frequency computed correctly: 10 points
     -- Document frequency computed correctly: 10 points

Part 2: Generating TF-IDF representation of a tweet (25 points)

Task 2:

Given a tweet, please implement a function to generate the TF-IDF vector representation of the tweet using the document frequency computed on the training set in Task 1. (Note: you may also need to pre-process the raw tweets by lowercasing the text, removing stop words/numbers/mentions/URLs/special characters, lemmatizing words.)

What to Turn In:

(1) The source code of the function(s) to finish this task.

Grade Criterion:

1. Whether submitted codes can run successfully.
   - Successfully: 25 points
   - Unsuccessfully:
     -- Properly preprocess the raw tweet: 15 points
     -- Correctly generate the TF-IDF vector representation of the tweet: 10 points

Part 3: Training a classifier to predict the sentiment of a tweet (50 Points)

Task 3:

Using the TF-IDF representation of each tweet extracted in Task 2 to train a logistic regression classifier and predict the binary sentiment (i.e., positive/negative) of each tweet in the testing set. (Note: you may use existing libraries (e.g., sklearn) to normalize the TF-IDF representation and train your model.)

What to Turn In:

(1) The source code of the function(s) to finish this task.

(2) A csv file that contains the prediction results (i.e., storing your binary prediction results as a column in tweets_test.csv).

Grade Criterion:

1. Whether submitted codes can run successfully.
   - Successfully: 25 points
   - Unsuccessfully:
     -- Correctly normalize the input feature: 5 points
     -- Correctly train the classifier: 20 points

2. Whether the prediction results achieve reasonable performance (i.e., both the accuracy and F1 score are above 0.7).
   - Successfully: 25 points
   - Unsuccessfully:
     -- Accuracy or F1 score is above 0.7: 20 points
     -- Accuracy and F1 score are above 0.6: 15 points
     -- Accuracy or F1 score is above 0.6: 10 points
     -- Accuracy and F1 score are above 0.5: 5 points