The goal of this assignment is to classify documents in a corpus. You will train a variety of linear/neural network models and evaluate each one using 5-fold cross-validation. Using your best-performing model, you will run inference on a test set and submit the predicted labels.

Dataset description:

You will use the news dataset from Quiz. As before, the dataset contains five categories (sport, business, politics, entertainment, tech). The task is to classify documents into one of these five categories. You will be provided with the following datasets:

- Raw training data (link) with labels:
  - The dataset contains the raw text of 1000 news articles and the article category. Each row is a document.
  - The raw file is a .csv with three columns: ArticleId, Text, Category
  - The "Category" column are the labels you will use for training
- Raw test data (link) without labels
  - This dataset contains the raw text of 735 news articles. Each row is a document.
  - The raw file is a .csv with two columns: ArticleId,Text.
  - The labels are not provided

Example codes for Pytorch and Tensorflow.

Your job:

1. Preprocess the raw training data. You can use the code from Homework 1. You are required to construct other features, such n-grams or keyword extractions. (15pt)
   a. Run Neural Networks with the 2-hidden layers, each has 128 neurons, extracting features by CountVectorizer() as the original features. Use 5-fold cross-validation to evaluate the performance.
   b. Feature exploration. Use other features like TFIDF, or any word embeddings provided by other packages like GloVe with gensim, or BERT. Use 5-fold cross-validation to evaluate the performance of your Neural Network.
   c. Describe how you generate features. (5pt)
   d. Report the average training and validation accuracy, and their standard deviation for different feature construction (organize the results in a table). (5pt)

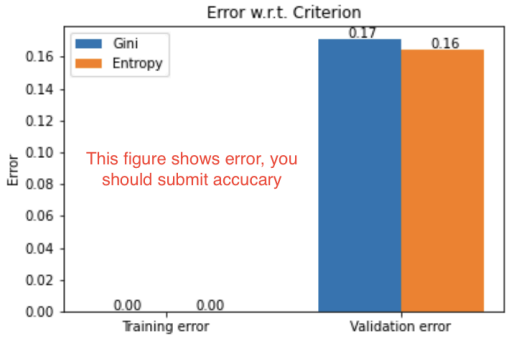
      Example:

Feature method	training accuracy	testing accuracy
CountVectorizer() [Required]	0.839	0.723
GloVe *	0.899	0.923
…	…	…
BERT *	0.702	0.792

      * Besides CounteVectorizer, At least one additional feature method should be included.

   e. Draw a bar figure showing the training and validation result, x-axis should be the parameter values, y-axis should be the training and validation accuracy. (5pt)

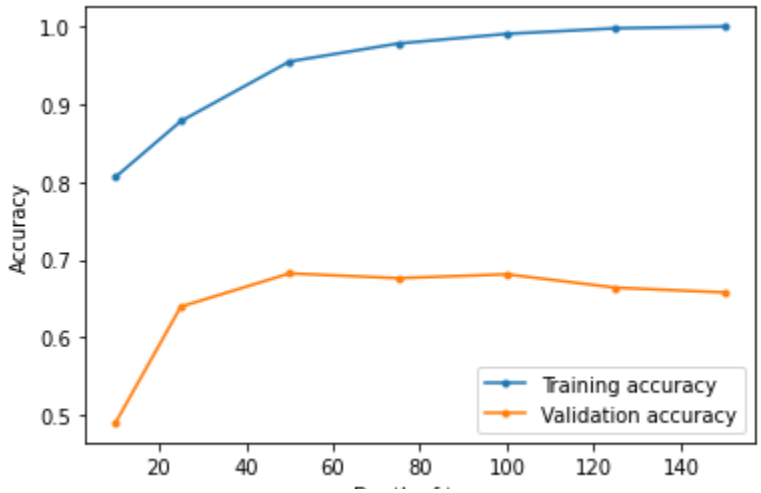
      Example:



2. Explore the Neural Network model on pre-processed training data. (25pt)
    a. Describe your parameter setting. (5pt)
    b. Use 5-fold cross-validation to evaluate the performance w.r.t. the learning rates (η), you could use the feature engineering method that has the best performance from Question 1. Recommended candidate values: [0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1]
        1. Report the average training and validation accuracy, and their standard deviation for different parameter values (organize the results in a table). (5pt)
        Example:

Learning rates	training accuracy	testing accuracy
0.0001	0.839	0.723
0.0003	0.899	0.923
…	…	…
0.1	0.702	0.792

        2. Draw a line figure showing the training and validation result, x-axis should be the parameter values, y-axis should be the training and validation accuracy. (5pt)
        Example:



    c. Use 5-fold cross-validation to evaluate the performance w.r.t. optimizers, you could use the feature engineering method that has the best performance from Question 1.

Recommended candidate values: [SGD, Adam, RMSprop] (see PyTorch or Tensorflow)

1. Report the average training and validation accuracy, and their standard deviation for different parameter values (organize the results in a table). (5pt)
2. Draw a bar figure showing the training and validation result, x-axis should be the parameter values, y-axis shodenseuld be the training and validation accuracy. (5pt)

3. Predict the labels for the testing data (using raw training data and raw testing data). (60pt)
   a. Describe how you pre-process the data to generate features. (5pt)
   b. Describe how you choose the model and parameters. (5pt)
   c. Describe the performance of your chosen model and parameter on the training data. (5pt)
   d. The final classification models to be used in this question are limited to random forest, neural networks, and ensemble methods. It is OK to use other models to do feature engineering. (45pt)
      1. Note that this question will be graded based on your accuracy. You should try to think of better features and try different models and parameters in order to get a higher accuracy.

What to submit:

You need to submit three files:

1. code.ipynb - The notebook containing all the code for the questions. Please do not include notebook cells that had no use randomly. For each cell in the notebook, you should include a description of what it does. This will help improve your code writing skills in general.

2. description.pdf - The description of the results for all questions

3. labels.csv, this is the predicted labels for Q3. Each row of the file will be a comma-separated string denoting the article ID and predicted label. For example, if the predicted label for article number 2 is politics, then the row in the file would be “2,politics”. **Make sure that your .csv file does not have a header row.**

Note:

- Remember to submit the three files by clicking on "add another file" in Canvas, instead of submitting one zipped file of the aforementioned three files.
- Submit all the files to HW 2 on CANVAS.
- If there is any question about the assignment, please email the TA.
- Late submission penalty will be strictly enforced (see syllabus). Assignment should be completed independently: Submissions after the deadline but less than 24 hours late are accepted but penalized 10%, and submissions more than 24 hours but less than 48 hours late are penalized 30%. No submissions are accepted more than 48 hours late.