The goal of this assignment is to classify documents in a corpus. You will train a variety of tree-based 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.

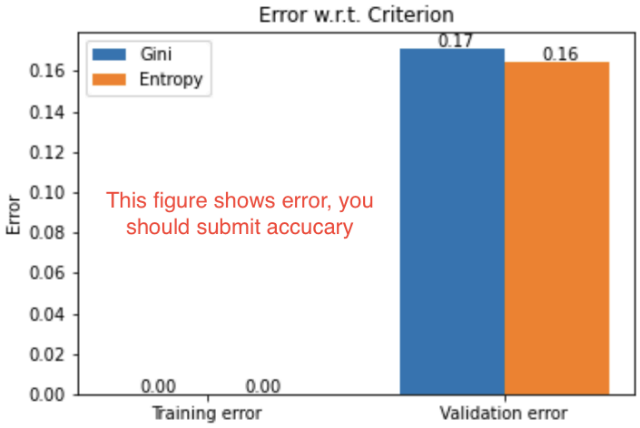
Dataset Description:

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 681 news articles. Each row is a document.
    - The raw file is a .csv with two columns: ArticleId,Text.
    - The labels are not provided

Your job:

1. Preprocess the raw training data. You can use your code from the HW0. Additionally, you can use the code from the posted solution from the HW0. You may also construct other features, such n-grams or keyword extractions. Feel free to use any other features you feel may be relevant.
2. Evaluate the decision tree model on your pre-processed data. (25pt)
    1. Randomly select 80% data instances as training, and the remaining 20% data instances as validation. Change the parameter setting on **criterion** ("gini", "entropy"}. Draw a bar chart showing the training accuracy and validation accuracy w.r.t. different parameter values. (5pt)
       Example:



This figure shows error, you should submit accucary

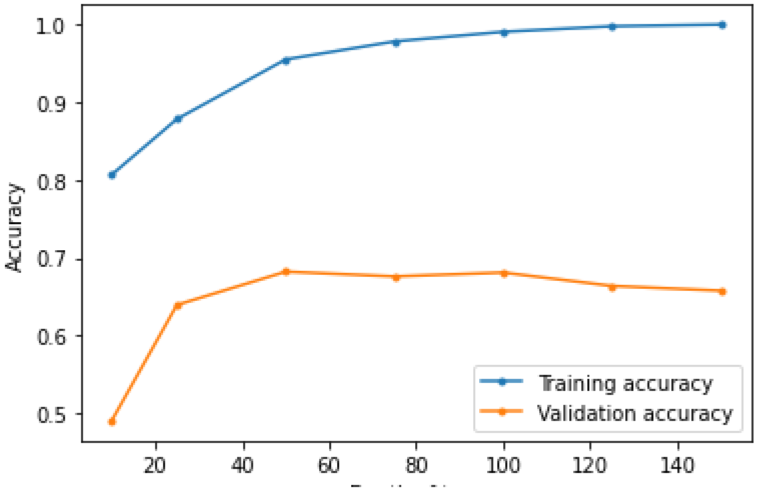
    2. Evaluate the decision tree using 5-fold cross-validation (see the [example code for a different task here]()) w.r.t min_samples_leaf:
        1. Report the average training and validation accuracy, and their standard deviation for different parameter values (organize the results in a table). (5pt)

           Example:

min_samples_leaf	training accuracy	testing accuracy
10	0.839	0.723
50	0.899	0.923
…	…	…
200	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 cvalidation accuracy. (5pt)

           Example:



    3. Evaluate the decision tree using 5-fold cross-validation w.r.t max_features:
        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 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)

3. Evaluate [random forests model]() on pre-processed training data. (25 pt)
    1. Describe your parameter setting. (5pt)
    2. Use 5-fold cross-validation to evaluate the performance w.r.t. the number of trees (n_estimators):
        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 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)

3. Use 5-fold cross-validation to evaluate the performance w.r.t. the minimum number of samples required to be at a leaf node (min_samples_leaf)
   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 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)

4. Predict the labels for the testing data (using raw training data and raw testing data). (50pt)
   1. Describe how you pre-process the data to generate features. (5pt)
   2. Describe how you choose the model and parameters. (5pt)
   3. Describe the performance of your chosen model and parameter on the training data. (5pt)
   4. The final classification models to be used in this question are limited to decision trees, random forests, and boosting trees (AdaBoost, or GradientBoostingTree). It is OK to use other models/methods to do feature engineering (e.g., using word embeddings). (35pt)
      1. Note that this question will be graded based on your accuracy on our test data. 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 Q4. 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 1 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.