CMPUT497/501 Intro. to NLP
Assignment 5: Classifying BBC Articles

Part 1: Classify Text Type.

Write and document a program that trains a Naive Bayes classifier (Chapter 4) for the 5-class classification problem at hand. Your program will process texts and classify them as belonging to one of the following classes: business, entertainment, politics, sport, or tech.

Your program should take trainBBC.csv as input, learn the model, and print the training accuracy (using 3-fold cross validation). The program should print the data splits (which data points belong to each fold). It should also print the accuracy of the classifier on the test set (testBBC.csv).

Your program should output a csv file (named output_[inputFilename].csv) containing the following information: original label, classifier-assigned label, text. As an example, the output file should be named output_testBBC.csv when running the testBBC data through your classifier.

Report a confusion matrix with precision and recall as in Fig 4.5. Also, report the aggregated (pooled) micro-averaged and macro-averaged precision as in Fig 4.6.

Document and justify your design decisions*.* The description in the textbook leaves some of the options for building an NB classifier and feature selection open (e.g., tokenization, model parameters, handling stop words, and handling unknown words). You are expected to make choices with respect to these issues, explain those choices, and justify them in your **report**.

Part 2: Error Analysis.

Take a look at text excerpts that are incorrectly classified. Which classes of text have a tendency to be misclassified as another (e.g., are tech articles misclassified as business)? Do the misclassified texts share common attributes? If so, what are they?

Take a look at the output from running evalBBC.csv through your classifier. Do the assigned labels make sense? When they don't, do the texts share attributes with the testBBC cases that were misclassified?

Note: the questions listed above are not exhaustive.

Data:

- trainBBC.csv
- testBBC.csv
- evalBBC.csv

All csv files contain 2 columns. The first, called category, specifies the category to which the text belongs. The second, called text, contains the text you are supposed to classify. The category column in the evalBBC.csv file is blank.

You may need to perform some preprocessing on the data to facilitate classification.

Never use testBBC.csv to train your models.

What to Submit:

- Your code and a readme file so that we can run your code on the lab machines
- The output from when you test your model using testBBC.csv
    - Each sample should be output with its predicted label and true label
- The output from when you run your model on evalBBC.csv
    - Each sample should be output with its predicted label and true label (which will be blank as the input file does not contain a true label)

- A report detailing your findings
  - This report should be in the same format as your project report. The style files are provided under the project section of eClass.
  - This report should contain one section for each of the assignment parts
  - All of the decisions that you make should be justified within the report
  - Please use appropriate sub-headings so that it is easier for the marker to find key information