CSE 291: Homework 1

Instructions

In this assignment, you will train a text classifier using different feature representation techniques. You should submit the final writeup and code for this assignment. The writeup should be a pdf that includes experimental findings of the programming part. Please submit your solution **by October** *17*. Submissions should be made on **Gradescope (course code:** *2BY4J6*). Please complete homework **individually**. Please include the code of your solutions in the submission with a write-up describing how to run the code.

You will need two datasets for this homework that are included in the HW1 download files on piazza: **New York Times (NYT) news** and **AG News**.

NYT dataset contains a *text* column consisting of news articles and a *label* column indicating the category to which this article belongs. AG News has just the text column. For questions 1 and 2, use the logistic regression classifier.

The classifier should be trained and tested on the NYT dataset. Shuffle the NYT 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`).

1. **Bag Of Words (20 points):**

Train a text classifier using the following document representation techniques and report accuracy, macro-f1 score, and micro-f1 score on the test set.

(a) 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.

(b) 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.

(c) 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.

2. **Word2Vec (20 points):**

(a) Train a text classifier using the following document representation techniques using **100-dimensional** word vectors and report accuracy, macro-f1 score, and micro-f1 score on the test set. Compare and analyze their performance.

(i) Using publicly available pre-trained Glove embeddings as word vectors, a document vector is represented as an average of word vectors of its constituent words.

(ii) Train Word2Vec (e.g., use `gensim` package) on AGNews text data and use them as word vectors to compute document vectors by averaging word vectors of its constituent words.

(iii) Train Word2Vec on NYT text data and use them as word vectors to compute document vectors by averaging word vectors of its constituent words

(b) What are the disadvantages of averaging word vectors for the document representation? Describe an idea to overcome this. The document vectors should be formed using word vectors.
Note: *This is an open-ended question. Feel free to propose new ideas.*

CSE 291: Homework 1

3. **BERT (20 points):**

Fine-tune the BERT (bert-base-uncased) for text classification and report accuracy, macro f1-score, and micro f1-score. If you are using PyTorch, hugging face transformers is highly recommended for this task. While tokenizing, set the maximum length to 64 and fine-tune for 3 epochs.