How to run the code

1. Before running the program, edit `.config` files under the `data` folder as you see fit.
2. Decide the phase of program (train/test).
    1. For training, run `python3 question_classifier.py train -config [configuration_file_path]`
    2. For testing, run `python3 question_classifier.py test -config [configuration_file_path]`
3. Train and validation sets are split as the program run, so you only need one data set to provide train and validation sets.

About configurations

Configuration files

- `bow.config`: Configuration file of FFNN with bag of words.
- `bilstm.config`: Configuration file of FFNN with BiLSTM.
- `bow_bilstm.config`: Configuration file of FFNN with bag of words and BiLSTM.
- `bow_ffnn_ens.config`: Configuration file of FFNN with bag of words and ensemble method.
- `bilstm_ffnn_ens.config`: Configuration file of FFNN with bag of words and ensemble method.
- `bow_bilstm_ens.config`: Configuration file of FFNN with bag of words, BiLSTM and ensemble method.

Fields

- `path_data`: The path to the data used for training and validation. We divide the data into training and validation sets within the program with the ratio of 9:1.
- `path_test`: The path to the data used for testing.
- `model`: Can only be `bow`, `bilstm`, `bow_bilstm`, `bow_ens`, `bilstm_ens` and `bow_bilstm_ens`.
- `path_model`: The path of model storage.
- `ensemble_size`: The number of ensemble models.
- `min_words`: The minimum occurrences of words for it to be enlisted in vocabulary. `0` stands for include all appeared words.
- `freeze`: Whether to freeze the word embeddings during model training process.
- `from_pretrained`: Whether to use pretrained word embeddings.
- `early_stopping`: Set the early stopping threshold to prevent overfitting.
- `epoch`: The number of epochs in training process.
- `path_pre_emb`: The path to pre-trained embeddings.
- `hidden_size`: The size of the hidden layer in FFNN.
- `word_embedding_dim`: The dimension of word embeddings.
- `bilstm_hidden_size`: The size of hidden layer of BiLSTM.
- `batch_size`: The size of each batch in each epoch.
- `lr_param`: The learning rate parameter.
- `sgd_momentum`: The momentum of learning rate.

- `path_eval_result`: The path of evaluation output.

Evaluation results

- F1-score of each class
- Actual and predicted labels and whether the predictions are correct.
- Accuracy of the model
- Confusion matrix
- Micro and Macro F1-score