

Coursework 1: Question classification description

This readme contain 3 sections: Project Structure, Running of code, Description for each function.

1. Project Structure

```
.
├── document                // description of coursework
│   ├── readme.md          // instruction. You are here.
│   └── readme.pdf         // PDF form for readme.
├── data
│   ├── bow_model.pt       // After training bow model, store model
in there
│   ├── bilstm_model.pt    // After training bilstm model, store
model in there
│   ├── bow.config         // bow model configuration
│   ├── bilstm.config      // bilstm model configuration
│   └── output.txt         // After testing the model, store the
results in here.
│   ├── train_5500.txt     // dataset
│   └── glove.small.txt    // glove pre-train word embedding
├── src                    // source code
│   └── question_classifier.py // source code of system
└── .
```

2. How to use the code

To run this system, you must change the directory to `src` folder.

2.1 Training and testing bow model

To train the `bow` model, you can use the command below:

```
python3 question_classifier.py train -config ../data/bow.config
```

Then you can test the bow model after the above step, you can use the command below:

```
python3 question_classifier.py test -config ../data/bow.config
```

2.2 Training and testing bilstm model

To train the `bilstm` model, you can use the command below:

```
python3 question_classifier.py train -config ../data/bilstm.config
```

Then you can test the bilstm model after the above step, you can use the command below:

```
python3 question_classifier.py test -config ../data/bilstm.config
```

3. Description for each function

Here, we only describe some main functions, the detailed information can be seen from the comments in the source code.

```
def load_dataset(data_path):  
    '''  
    1. This function load the dataset from  
    https://cogcomp.seas.upenn.edu/Data/QA/QC.  
    And do some preprocessing: data cleaning, removing stop words,  
    refactoring data structure.  
    Then return preprocessed data_set. A example in data_set is:  
    后面添加一个例子  
    '''  
  
def build_random_vocabulary(word_appear_times):  
    '''  
    2. This function build the vocabulary for the randomly initialize word  
    embeddings method. It also add word: #UKN# in vocabulary.  
    You can set word_appear_times value to select words that appearing at  
    least  
    kword_appear_times times in the training set. Then function return  
    vocabulary.  
    A example data in vocabulary is:  
    后面添加例子  
    '''  
  
def load_glove(glove_path):  
    '''  
    3. This function load the glove.small.txt pre-trained embeddings, and  
    pruning  
    pretrained embeddings by removing the words that do not appear in the  
    dataset.  
    Then function return the vocabulary for glove pretraining word  
    embeddings.  
    A example data in glove vocabulary is:  
    后面添加例子  
    '''  
  
def spliteDataset(validation_size):  
    '''  
    4. This function splite dataset into train and test subset. Default
```

```
validation_size=0.9
    which means 9 portions are for training, and the other is for testing.
    Then function return the train_set and test_set
    '''

def word_embedding(is_Pretrain,is_pre_freeze):
    '''
    5. This function realize two kinds of wording embedding by using
    pytorch:
    First kind is randomly initialize it by using function below:
    embedding=nn.Embedding(VOCAB_SIZE, WORD_DIM)
    Second kind is using glove pretrained weights. You can set whether
    using freeze:
    embedding = nn.Embedding.from_pretrained(weights,
    freeze=is_pre_freeze)
    Both of them return the vector representation of a word.
    '''

def make_bow_vector(sentence, vocabulary):
    '''
    6. This function turns the sentence into a vector form by adding up
    the vectors for all
    the words and divide by the number of words in the sentence.
    It returns the vector for a sentence.
    '''

def bow_train(bow_model):
    '''
    7. This function trains bow model.
    详细的讲解
    '''

def bow_test():
    '''
    8. This function tests bow model.
    详细的讲解
    '''

def bilstm_train(bilstm_model):
    '''
    9. This function trains bilstem model.
    详细的讲解
    '''

def bilstm_test():
    '''
    10. This function tests bilstm model.
    详细的讲解
    '''

def select_operation():
    '''
    11. This function checks what command you input in the terminal to
```

```
determine whether training or testing what model.  
'''
```