pre_process

September 18, 2024

1 Preprocess Required for Generating Train Data

[1]:
```
import urllib.request
import time
import sys
import getopt
import pandas as pd
import numpy as np
import pickle
```

[2]:
```
%run ../utils.ipynb
```

[24]:
```
embSize = 200
ftrain='../data/SNP_train_data.txt'
ftest = "../data/SNP_test_data.txt"
# Replace with path of word embdding file
wefile = "../support/PubMed-and-PMC-w2v.bin"
random_seed=1331
```

2 Read Train Data

[18]:
```
samples= pd.read_csv(ftrain,sep='\t')
Tr_sent_contents, Tr_entity1_list, Tr_entity2_list, Tr_sent_lables,_ =␣
 ↪dataRead_snp(ftrain)
Tr_word_list, Tr_d1_list, Tr_d2_list =␣
 ↪get_wordList_and_distances_snp(Tr_sent_contents)
print ("train_size", len(Tr_word_list))
```

```
Input File Reading
train_size 935
```

3 Read Test Data

[19]:
```
Te_sent_contents, Te_entity1_list, Te_entity2_list, Te_sent_lables,_ =␣
 ↪dataRead_snp(ftest)
```

```
Te_word_list, Te_d1_list, Te_d2_list =␣
 ↪get_wordList_and_distances_snp(Te_sent_contents)

print ("test_size", len(Te_word_list))
```

```
Input File Reading
test_size 365
```

3.1 Prepare Lable Matrix

[20]:
```
set(Tr_sent_lables)
```

[20]:
```
{'negative', 'neutral', 'positive'}
```

[21]:
```
label_dict = {'negative':0, 'neutral':0,'positive':1}
Y_t = mapLabelToId_snp(Tr_sent_lables, label_dict)

Y_train = np.zeros((len(Y_t), 2))
for i in range(len(Y_t)):
    Y_train[i][Y_t[i]] = 1.0
```

[22]:
```
Y_te = mapLabelToId_snp(Te_sent_lables, label_dict)

Y_test = np.zeros((len(Y_te), 2))
for i in range(len(Y_te)):
    Y_test[i][Y_te[i]] = 1.0
```

4 Generate Word and Position Embedding Vectors

4.0.1 Word Embedding

[25]:
```
sent_list = sum([Tr_word_list, Te_word_list], [])
word_dict, word_to_id, id_to_word = word_mapping(sent_list)
print( "word dictonary length", len(word_dict))
# Word Embedding
word_vectors = readWordEmb(word_dict,id_to_word,word_to_id, wefile,␣
 ↪embSize,limit=2000000)
W_train =   mapWordToId(Tr_word_list, word_to_id)
```

```
Found 2775 unique words (45627 in total)
word dictonary length 2775
Reading word vectors
Loaded 2000000 pretrained embeddings.
number of unknown word in word embedding 475
```

[26]:
```
W_test =   mapWordToId(Te_word_list, word_to_id)
```

4.0.2 Position Embedding

[27]:
```
d1_dict = makeDistanceList([Tr_d1_list, Te_d1_list])
d2_dict = makeDistanceList([Tr_d2_list, Te_d2_list])

d1_train = mapWordToId_list(Tr_d1_list, d1_dict)
d2_train = mapWordToId_list(Tr_d2_list, d2_dict)
```

[28]:
```
d1_test = mapWordToId_list(Te_d1_list, d1_dict)
d2_test = mapWordToId_list(Te_d2_list, d2_dict)
```

4.0.3 Pad Embdding Vectors

[29]:
```
train_sent_lengths, test_sent_lengths = findSentLengths([Tr_word_list,␣
 ↪Te_word_list])
sentMax = max(train_sent_lengths +  test_sent_lengths)

#padding
W_train, d1_train, d2_train = paddData([W_train, d1_train, d2_train ], sentMax)
```

[30]:
```
type (W_test)
```

[30]:
```
list
```

[31]:
```
#padding
W_test, d1_test, d2_test = paddData([W_test, d1_test, d2_test ], sentMax)
```

[32]:
```
type (W_test)
```

[32]:
```
numpy.ndarray
```

5 Save Training data as a Pickle file

[33]:
```
#with open('train_and_test_data_sentences_snp_2classWiki.pickle', 'wb') as␣
 ↪handle:
with open('../data/pickles/train_and_test_data_sentences_snp_2class.pickle',␣
 ↪'wb') as handle:
    pickle.dump(W_train, handle)
    pickle.dump(d1_train, handle)
    pickle.dump(d2_train, handle)
    pickle.dump(Y_train, handle)
    pickle.dump(Tr_word_list, handle)

    pickle.dump(W_test, handle)
    pickle.dump(d1_test, handle)
    pickle.dump(d2_test, handle)
    pickle.dump(Y_test, handle)
```

```
    pickle.dump(Te_word_list, handle)

    pickle.dump(word_vectors, handle)
    pickle.dump(word_dict, handle)
    pickle.dump(d1_dict, handle)
    pickle.dump(d2_dict, handle)
    pickle.dump(label_dict, handle)
    pickle.dump(sentMax, handle)
```

6 test data statistics

[34]:
```
disease_list=np.array(Te_entity2_list)
disease_list=disease_list[:,0]
len(np.unique(disease_list))
labels=np.array(Te_sent_lables)
len(np.unique(disease_list[labels[:]=='negative']))
len(np.unique(disease_list[labels[:]=='neutral']))
len(np.unique(disease_list[labels[:]=='positive']))
gene_list=np.array(Te_entity1_list)
gene_list=gene_list[:,0]
len(np.unique(gene_list))
len(np.unique(gene_list[labels[:]=='negative']))
len(np.unique(gene_list[labels[:]=='neutral']))
len(np.unique(gene_list[labels[:]=='positive']))
len(labels[labels[:]=='negative'])
len(labels[labels[:]=='positive'])
len(labels[labels[:]=='neutral'])
```

[34]:
```
166
```

7 train data statistics

[35]:
```
disease_list=np.array(Tr_entity2_list)
disease_list=disease_list[:,0]
len(np.unique(disease_list))
labels=np.array(Tr_sent_lables)
len(np.unique(disease_list[labels[:]=='negative']))
len(np.unique(disease_list[labels[:]=='neutral']))
len(np.unique(disease_list[labels[:]=='positive']))
gene_list=np.array(Tr_entity1_list)
gene_list=gene_list[:,0]
len(np.unique(gene_list))
len(np.unique(gene_list[labels[:]=='negative']))
len(np.unique(gene_list[labels[:]=='neutral']))
```

```
len(np.unique(gene_list[labels[:]=='positive']))
len(labels[labels[:]=='negative'])
len(labels[labels[:]=='positive'])
len(labels[labels[:]=='neutral'])
```

[35]: 142

[ ]:

[ ]: