Branch: **master** ▾ | **Find file** | **Copy path**

Unsupervised-Protein-Genes-Diseases-Extraction / Code /

Relation_Association_Classification.ipynb

tsantosh7 Loading models

`b9ec887` yesterday

1 contributor

Raw | **Blame** | **History**

`515 lines (514 sloc)  24.3 KB`

In [1]:
```
import gensim.models.keyedvectors as word2vec
from nltk.tokenize import RegexpTokenizer

from keras.models import Sequential
from keras.layers.core import Dense, Dropout
from keras.layers.embeddings import Embedding
from keras.preprocessing.sequence import pad_sequences
from keras.preprocessing.text import Tokenizer
from keras.layers import LSTM

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, roc_curve,  ro
c_auc_score, classification_report



import os
```

```
Using TensorFlow backend.
```

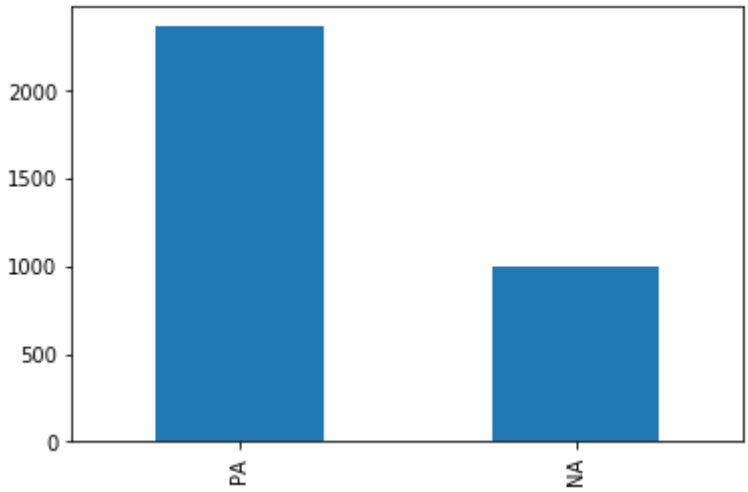
In [2]:
```
# get the current working directory
data_path = os.path.abspath(os.path.join(os.path.dirname( '_
_file__' ), '..', 'Datasets'))+'/'

# Although the dataset says csv, it is tab delimited. In add
ition to this, they have severe codels problems.
# So best to parse throught codes first.
# UnicodeDecodeError: 'utf-8' codec can't decode byte 0xfc i
n position 2: invalid start byte

#open for reading with "universal" type set

import codecs

doc_d_t = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EU
ADR_drug_target'+'.csv','rU','UTF-8')
EUADR_drug_target = pd.read_csv(doc_d_t, sep='\t', na_filter
= False)
EUADR_drug_target['CLASS'] = 'drug_gene'

doc_t_d = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EU
ADR_target_disease'+'.csv','rU','UTF-8',errors='ignore')
EUADR_target_disease = pd.read_csv(doc_t_d, sep='\t', na_fil
ter = False)
EUADR_target_disease['CLASS'] = 'gene_disease'

doc_d_d = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EU
ADR_drug_disease'+'.csv','rU','UTF-8')
EUADR_drug_disease = pd.read_csv(doc_d_d, sep='\t', na_filte
r = False)
EUADR_drug_disease['CLASS'] = 'drug_disease'


EUADR_temp = EUADR_drug_target.append(EUADR_target_disease).
append(EUADR_drug_disease)
```

```
EUADR_temp_1 = EUADR_temp[EUADR_temp['ASSOCIATION_TYPE'] ==
'PA']
EUADR_temp_2 = EUADR_temp[EUADR_temp['ASSOCIATION_TYPE'] ==
'NA']
EUADR_temp = EUADR_temp_1.append(EUADR_temp_2)
```

In [3]:
```
# Get GAD datasset

doc_t_d = codecs.open(data_path+'GAD_Corpus_IBIgroup/'+'GAD_
Y_N'+'.csv','rU','UTF-8',errors='ignore')
GAD_target_disease_Y_N = pd.read_csv(doc_t_d, sep='\t', na_f
ilter = False)
GAD_target_disease_Y_N['CLASS'] = 'gene_disease'

doc_t_d = codecs.open(data_path+'GAD_Corpus_IBIgroup/'+'GAD_
F'+'.csv','rU','UTF-8',errors='ignore')
GAD_target_disease_F = pd.read_csv(doc_t_d, sep='\t', na_fil
ter = False)
GAD_target_disease_F['CLASS'] = 'gene_disease'

GAD_temp = GAD_target_disease_Y_N
```

In [4]:
```
# get sentences and their associations

sentences = EUADR_temp['SENTENCE'].append(GAD_temp['GAD_CONC
LUSION'])
labels = EUADR_temp['ASSOCIATION_TYPE'].append(GAD_temp['GAD
_ASSOC'].apply(lambda x: x.replace('Y', 'PA').replace('N',
'NA').replace('F', 'FA')))

labels_count = labels.value_counts()
labels_count.plot(kind="bar")
print(labels.value_counts())
```

```
PA    2366
NA     994
dtype: int64
```



In [5]:
```
from sklearn import preprocessing
from keras.utils import np_utils

le = preprocessing.LabelEncoder()

le.fit(labels)
le.classes_
# le.inverse_transform()
```

```
# tokenizer.fit_transform()

y = le.transform(labels)
dummy_y = np_utils.to_categorical(y)
```

In [6]:
```
def format_sentences(a_list_of_sentences):

#     tkr = RegexpTokenizer('[a-zA-Z@]+')
    sentences_split = []

    for i, line in enumerate(a_list_of_sentences):
        #print(line)
        sent = str(line).split()
#         sent = tkr.tokenize(str(sent))
        sentences_split.append(sent)

    return sentences_split

sentences_split = format_sentences(sentences)
print(sentences_split[10])
```

```
['Inhibition', 'of', 'EGFR', 'kinase', 'activity', 'by', 'gef
itinib', 'causes', 'the', 'translocation', 'of', 'the', 'ABCG
2', 'drug', 'transporter', 'away', 'from', 'the', 'plasma',
'membrane,', 'resulting', 'in', 'a', 'concomitant', 'decreas
e', 'in', 'doxorubicin', 'extrusion', 'in', 'thyroid', 'cance
r', 'cell', 'lines.']
```

In [7]:
```
w2vModel = word2vec.KeyedVectors.load_word2vec_format('/hom
e/stirunag/pre-trained_word_embeddings/PubMed-and-PMC-w2v.bi
n', binary=True, limit=1000000)

#Convert words to integers
tokenizer = Tokenizer()
tokenizer.fit_on_texts(sentences_split)
X = tokenizer.texts_to_sequences(sentences_split)

#lenght of sentence to consider
maxlenth = 100
#add padding
X = pad_sequences(X, maxlen=maxlenth)
print(X.shape)

#create a embedding layer using PMC vectors (100000 words)
embedding_layer = Embedding(input_dim=w2vModel.wv.vectors.sh
ape[0], output_dim=w2vModel.wv.vectors.shape[1], weights=[w2
vModel.wv.vectors],
                            input_length=X.shape[1])
```

```
/home/stirunag/anaconda3/envs/KerasCPU/lib/python3.7/site-pac
kages/smart_open/smart_open_lib.py:398: UserWarning: This fun
ction is deprecated, use smart_open.open instead. See the mig
ration notes for details: https://github.com/RaRe-Technologie
s/smart_open/blob/master/README.rst#migrating-to-the-new-open
-function
  'See the migration notes for details: %s' % _MIGRATION_NOTE
S_URL
/home/stirunag/anaconda3/envs/KerasCPU/lib/python3.7/site-pac
kages/ipykernel_launcher.py:15: DeprecationWarning: Call to d
eprecated `wv` (Attribute will be removed in 4.0.0, use self
instead).
  from ipykernel import kernelapp as app
```

```
WARNING: Logging before flag parsing goes to stderr.
W0724 14:24:09.579116 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/backend/tensorflow_backend.py:74: The n
ame tf.get_default_graph is deprecated. Please use tf.compat.
v1.get_default_graph instead.


(3360, 100)
```

In [8]:
```
#create model

lstm_out = 80

model = Sequential()
model.add(embedding_layer)
model.add(LSTM(units=lstm_out))
# model.add(Dense(1, activation='softmax'))
model.add(Dense(2, activation='softmax'))
# model.compile(loss='binary_crossentropy', optimizer='ada
m', metrics=['accuracy'])
model.compile(loss='categorical_crossentropy', optimizer='ad
am', metrics=['accuracy'])
print(model.summary())
```

```
W0724 14:24:34.496708 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/backend/tensorflow_backend.py:517: The
name tf.placeholder is deprecated. Please use tf.compat.v1.pl
aceholder instead.

W0724 14:24:34.504989 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/backend/tensorflow_backend.py:4138: The
name tf.random_uniform is deprecated. Please use tf.random.un
iform instead.

W0724 14:24:34.524087 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/backend/tensorflow_backend.py:174: The
name tf.get_default_session is deprecated. Please use tf.comp
at.v1.get_default_session instead.

W0724 14:24:34.525708 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/backend/tensorflow_backend.py:181: The
name tf.ConfigProto is deprecated. Please use tf.compat.v1.Co
nfigProto instead.

W0724 14:24:37.476446 140163365340928 deprecation_wrapper.py:
119] From /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.
7/site-packages/keras/optimizers.py:790: The name tf.train.Op
timizer is deprecated. Please use tf.compat.v1.train.Optimize
r instead.


_________________________________________________________________
____
Layer (type)                 Output Shape              Param
#
=================================================================
====
embedding_1 (Embedding)      (None, 100, 200)          200000
000
```

```
000
_________________________________________________________________
____
lstm_1 (LSTM)                (None, 80)                89920
_________________________________________________________________
____
dense_1 (Dense)              (None, 2)                 162
=================================================================
====
Total params: 200,090,082
Trainable params: 200,090,082
Non-trainable params: 0
_________________________________________________________________
____
None
```

In [9]:
```
from sklearn.model_selection import train_test_split

#split dataset
# X_train, X_test, Y_train, Y_test = train_test_split(X, y,
 test_size= 0.1, random_state = 24, stratify=y)
X_train, X_test, Y_train, Y_test = train_test_split(X, dummy
_y, test_size= 0.1, random_state = 24, stratify=y)

#fit model
batch_size = 1024
model.fit(X_train, Y_train, epochs=25, verbose=1, batch_size
=batch_size)

#analyze the results
score, acc = model.evaluate(X_test, Y_test, verbose = 2, bat
ch_size=batch_size)
y_pred = model.predict(X_test)
```

```
W0724 14:25:05.567893 140163365340928 deprecation.py:323] Fro
m /home/stirunag/anaconda3/envs/KerasCPU/lib/python3.7/site-p
ackages/tensorflow/python/ops/math_grad.py:1250: add_dispatch
_support.<locals>.wrapper (from tensorflow.python.ops.array_o
ps) is deprecated and will be removed in a future version.
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.
where
/home/stirunag/anaconda3/envs/KerasCPU/lib/python3.7/site-pac
kages/tensorflow/python/ops/gradients_util.py:90: UserWarnin
g: Converting sparse IndexedSlices to a dense Tensor with 200
000000 elements. This may consume a large amount of memory.
  num_elements)
Epoch 1/25
3024/3024 [==============================] - 11s 4ms/step - l
oss: 0.6701 - acc: 0.6177
Epoch 2/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.6098 - acc: 0.7040
Epoch 3/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.6024 - acc: 0.7040
Epoch 4/25
3024/3024 [==============================] - 7s 2ms/step - lo
ss: 0.5870 - acc: 0.7040
Epoch 5/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.5681 - acc: 0.7040
```

```
ss: 0.5081 - acc: 0.7040
Epoch 6/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.5533 - acc: 0.7040
Epoch 7/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.5314 - acc: 0.7067
Epoch 8/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.4974 - acc: 0.7120
Epoch 9/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.4436 - acc: 0.7371
Epoch 10/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.3615 - acc: 0.8310
Epoch 11/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.3684 - acc: 0.8644
Epoch 12/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.3395 - acc: 0.8849
Epoch 13/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.2652 - acc: 0.9011
Epoch 14/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.2670 - acc: 0.8869
Epoch 15/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.2131 - acc: 0.9203
Epoch 16/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.1811 - acc: 0.9474
Epoch 17/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.1527 - acc: 0.9524
Epoch 18/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.1119 - acc: 0.9640
Epoch 19/25
3024/3024 [==============================] - 9s 3ms/step - lo
ss: 0.0995 - acc: 0.9676
Epoch 20/25
3024/3024 [==============================] - 8s 3ms/step - lo
ss: 0.0807 - acc: 0.9719
Epoch 21/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.0728 - acc: 0.9749
Epoch 22/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.0611 - acc: 0.9792
Epoch 23/25
3024/3024 [==============================] - 8s 2ms/step - lo
ss: 0.0556 - acc: 0.9805
Epoch 24/25
3024/3024 [==============================] - 8s 2ms/step - lo
```