pre_process

September 18, 2024

[2]:

```
import urllib.request
import time
import sys
import getopt
import pandas as pd
import numpy as np
import pickle
```

[28]:

```
%run ../utils.ipynb
```

[3]:

```
embSize = 300
d1_emb_size=20
d2_emb_size=20
trainFile='../data/GAD_merged_samples_mesh.csv'

# Replace with path of word embdding file
wefile = "../Dataset/embeddings/crawl-300d-2M.vec"
random_seed=1331
```

[8]:

```
import pandas as pd
from tabulate import tabulate
ftrain = '../data/GAD_merged_samples_mesh.csv'
with open(ftrain, 'r', encoding='latin1') as file:
    first_line = file.readline()
    print(first_line)
import pandas as pd

ftrain = '../data/GAD_merged_samples_mesh.csv'

# Specify the delimiter for tabs
df = pd.read_csv(ftrain)


# Display the first 10 rows of the DataFrame as a table
print(tabulate(df.head(4), headers='keys', tablefmt='grid'))
```

```
GAD_ID,associationType,geneSymbol,GAD_GENE_NAME,geneId,gene_mention,GENE_ENTITY_
OFFSET,diseaseName,disease_mention,DISEASE_ENTITY_OFFSET,raw_sentence,diseaseId
```

```
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|    |   GAD_ID | associationType    | geneSymbol   | GAD_GENE_NAME
|   geneId | gene_mention   | GENE_ENTITY_OFFSET   | diseaseName
| disease_mention   | DISEASE_ENTITY_OFFSET   | raw_sentence
| diseaseId    |
+====+==========+===================+==============+============================
=====+==========+================+======================+=======================
======================+===================+=========================+===========
================================================================================
==========================================================================+=====
=========+
|  0 |   116326 | Y                  | AGTR1        | Angiotensin II receptor,
type 1 |      185 | AT1R           | 80#84                | atherosclerosis,
coronary                   | CAD               | 159#162                 | This
study indicates a synergistic contribution of RAS genes (ACE I/D, AGT T/M, AT1R
T/C) and eNOS Glu298Asp polymorphisms to the development of the premature CAD. |
MESH:D003324 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  1 |   588219 | F                  | PALB2        | partner and localizer of
BRCA2  |    79728 | PALB2          | 4#9                  | breast cancer
| mutation          | 19#27                    | The PALB2 1592delT mutation has
a strong effect on familial breast cancer risk.
| nan          |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  2 |   127842 | Y                  | IL1A         | Interleukin 1, alpha
|     3552 | IL-1           | 30#34                | osteoarthritis
| OA                | 113#115                  | Our findings suggest that the
IL-1 gene cluster polymorphisms may play a significant role in the pathogenesis
of OA of the hip.                                     | MESH:D010003 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
```

```
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  3 |   154807 | F                 | TPMT         | Thiopurine
S-methyltransferase  |     7172 | TPMT             | 26#30                  |
azathioprine toxicity hepatitis, autoimmune | fibrosis          | 9#17
| Advanced fibrosis but not TPMT genotype or activity predicts azathioprine
toxicity in AIH.
| MESH:D005355 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
```

[17]:

```python
import pandas as pd
from tabulate import tabulate

# Path to the CSV file
ftrain = '../data/GAD_merged_samples_mesh.csv'

# Read the CSV file with the correct encoding
df = pd.read_csv(ftrain, encoding='latin1')

# Display the first 4 rows to check the structure
print("First 4 rows of the DataFrame:")
print(tabulate(df.head(4), headers='keys', tablefmt='grid'))

# Get unique labels from the 'associationType' column
labels = df['associationType'].unique()
print(f"\nUnique labels: {labels}")

# Filter and display samples for the label 'F'
target_label = 'Y'
if target_label in labels:
    filtered_df = df[df['associationType'] == target_label]
    print(f"\nSamples with label '{target_label}':")
    print(tabulate(filtered_df.head(8), headers='keys', tablefmt='grid'))
else:
    print(f"No samples found with label '{target_label}'")
```

```
First 4 rows of the DataFrame:
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
```

```
---------+
|    |   GAD_ID | associationType   | geneSymbol   | GAD_GENE_NAME
|   geneId | gene_mention   | GENE_ENTITY_OFFSET   | diseaseName
| disease_mention   | DISEASE_ENTITY_OFFSET   | raw_sentence
| diseaseId    |
+====+==========+===================+==============+============================
=====+==========+================+======================+=======================
======================+===================+=========================+===========
================================================================================
==========================================================================+=====
=========+
|  0 |   116326 | Y                 | AGTR1        | Angiotensin II receptor,
type 1 |      185 | AT1R           | 80#84                | atherosclerosis,
coronary                  | CAD               | 159#162                 | This
study indicates a synergistic contribution of RAS genes (ACE I/D, AGT T/M, AT1R
T/C) and eNOS Glu298Asp polymorphisms to the development of the premature CAD. |
MESH:D003324 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  1 |   588219 | F                 | PALB2        | partner and localizer of
BRCA2  |    79728 | PALB2          | 4#9                  | breast cancer
| mutation          | 19#27                   | The PALB2 1592delT mutation has
a strong effect on familial breast cancer risk.
| nan          |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  2 |   127842 | Y                 | IL1A         | Interleukin 1, alpha
|     3552 | IL-1           | 30#34                | osteoarthritis
| OA                | 113#115                 | Our findings suggest that the
IL-1 gene cluster polymorphisms may play a significant role in the pathogenesis
of OA of the hip.                                  | MESH:D010003 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+
|  3 |   154807 | F                 | TPMT         | Thiopurine
S-methyltransferase  |     7172 | TPMT           | 26#30                |
azathioprine toxicity hepatitis, autoimmune | fibrosis          | 9#17
```

```
| Advanced fibrosis but not TPMT genotype or activity predicts azathioprine
toxicity in AIH.
| MESH:D005355 |
+----+----------+-------------------+--------------+----------------------------
-----+----------+----------------+----------------------+-----------------------
----------------------+-------------------+-------------------------+-----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------+-----
---------+

Unique labels: ['Y' 'F' 'N' 'P']

Samples with label 'Y':
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
|    |   GAD_ID | associationType   | geneSymbol   | GAD_GENE_NAME
|   geneId | gene_mention     | GENE_ENTITY_OFFSET   | diseaseName
| disease_mention   | DISEASE_ENTITY_OFFSET   | raw_sentence
| diseaseId    |
+====+==========+===================+==============+============================
====================================================+==========+================
==+======================+===============================+===================+==
=======================+========================================================
================================================================================
================================================================================
================================================================================
================================================================================
=============================================================+==============+
|  0 |   116326 | Y                 | AGTR1        | Angiotensin II receptor,
type 1                                              |      185 | AT1R
| 80#84                | atherosclerosis, coronary     | CAD                 |
159#162                 | This study indicates a synergistic contribution of RAS
genes (ACE I/D, AGT T/M, AT1R T/C) and eNOS Glu298Asp polymorphisms to the
development of the premature CAD.
| MESH:D003324 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
|  2 |   127842 | Y                  | IL1A         | Interleukin 1, alpha
|     3552 | IL-1             | 30#34                  | osteoarthritis
| OA                | 113#115                 | Our findings suggest that the
IL-1 gene cluster polymorphisms may play a significant role in the pathogenesis
of OA of the hip.
| MESH:D010003 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
|  5 |   135175 | Y                  | SLC7A9       | Solute carrier family 7
(cationic amino acid transporter, y+ system), member 9 |    11136 | SLC7A9
| 193#199              | cystinuria                     | cystinuria        |
34#44                    | In summary, our results show that cystinuria is a
complex disease which is not only caused by mutations in SLC7A9 and SLC3A1, but
also influenced by other modifying factors such as variants in SLC7A9.
| MESH:D003555 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
| 11 |   152643 | Y                  | PON1         | Paraoxonase 1
|     5444 | PON1             | 0#4                    | stroke
| stroke            | 90#96                   | PON1 genetic variations are
associated with risk factors, severity, type and prognosis of stroke and
oxidative stress.
| MESH:D020521 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
-------------------------------------------------------------+--------------+
| 13 |   114502 | Y                  | A2M          | Alpha-2-macroglobulin
|        2 | A2M              | 27#30                | Alzheimer's Disease
| AD               | 111#113                 | our data suggests that the A2M D
allele is a modest risk factor for late-onset sporadic AD in Koreans, and the AD
risk conferred by the A2M D allele increases in APOE epsilon4 negative subjects.
| MESH:D000544 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
| 14 |   133817 | Y                  | RP1          | Retinitis pigmentosa 1
(autosomal dominant)                                    |     6101 | RP1
| 237#240              | retinitis pigmentosa          | RP                |
124#126                  | The de novo origin of an RP1 (Arg677ter) mutation in a
patient with simplex RP suggests that this common autosomal dominant RP mutation
can arise independently in the population and supports the hypothesis of a
mutational hotspot in the RP1 gene.
| MESH:D012174 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
| 17 |   127050 | Y                  | IFNG         | Interferon, gamma
|     3458 | interferon gamma | 78#94                | tuberculosis
| tuberculosis      | 152#164                 | This preferential binding
suggests that genetically determined variability in interferon gamma and
expression might be important for the development of tuberculosis.
| MESH:D014376 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
| 19 |   139782 | Y                 | HLA-DRB1     | major histocompatibility
complex, class II, DR beta 1                          |     3123 | DRB1
| 176#180              | Vogt-Koyanagi-Harada syndrome | VKH syndrome      |
253#265                 | (1) DRB1 * 0405 and DRB1 * 15 are closely associated
with the susceptibility to VKH syndrome, DRB1 * 0405 may be the major
susceptible gene and DRB1 * 15 may be the minor; (2) DRB1 * 14 and DRB1 * 08 are
negatively associated with the susceptibility to VKH syndrome, suggesting that
they may be the resistant genes; (3) DRB1 * 0405 is not related to the clinical
features, incidence of ocular complications as well as visual prognosis. |
MESH:D014607 |
+----+----------+-------------------+--------------+----------------------------
----------------------------------------------------+----------+----------------
--+----------------------+-------------------------------+-------------------+--
-----------------------+--------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------------------------------------------------------------+--------------+
```

1 Read Data

[30]:
```
sent_contents, entity1_list, entity2_list,␣
 ↪sent_lables,gene_id_list,disease_id_list,gene_symbol_list =␣
 ↪dataRead_befree(trainFile)
```

```
Input File Reading
```

[31]:
```
word_list, distance1_list, distance2_list  =␣
 ↪get_wordList_and_distances_befree(sent_contents)
```

[32]:
```
print ("train_size", len(word_list))
```

```
train_size 5330
```

[33]:
```
sent_lengths= findSentLengths([word_list])
```

[34]:
```
sentMax =max(max(sent_lengths[:]))
```

1.1 Prepare Lable Matrix

[35]:
```
# Three Class

# Y : positive association
# N:  negative association
# F:  No Smmantic Association
```

```
label_dict = {'F':0, 'Y': 1, 'N': 2}


Y = mapLabelToId_befree(sent_lables, label_dict)
Y_train = np.zeros((len(Y), len(label_dict)))
for i in range(len(Y)):
    Y_train[i][Y[i]] = 1.0
```

2 Generate Word and Position Embedding Vectors

2.0.1 Word Embedding

[36]:

```
word_dict, word_to_id, id_to_word = word_mapping(word_list)
print( "word dictonary length", len(word_dict))
word_vectors = readWordEmb_fastText(word_dict,id_to_word,word_to_id, wefile,␣
 ↪embSize)
X_train =   mapWordToId(word_list, word_to_id)
```

```
Found 6766 unique words (139634 in total)
word dictonary length 6766
Reading word vectors
number of unknown word in word embedding 1525
number of known word in word embedding 5241
```

2.0.2 Position Embedding

[37]:

```
distance1_dict = makeDistanceList([distance1_list])
distance2_dict = makeDistanceList([distance2_list])
distance1_vectors = mapWordToId_list(distance1_list, distance1_dict)
distance2_vectors = mapWordToId_list(distance2_list, distance2_dict)
```

2.0.3 Pad Embdding Vectors

[38]:

```
X_train, distance1_vectors, distance2_vectors = paddData([X_train,␣
 ↪distance1_vectors, distance2_vectors ], sentMax,padd_num= 6765)
```

3 Save Prepared Data as Pickle File

[39]:

```
with open('../data/pickles/befree_3class_crawl-300d-2M.pickle', 'wb') as handle:
    pickle.dump(gene_id_list, handle)
    pickle.dump(gene_symbol_list, handle)
    pickle.dump(disease_id_list, handle)
    pickle.dump(X_train, handle)
    pickle.dump(distance1_vectors, handle)
    pickle.dump(distance2_vectors, handle)
    pickle.dump(Y_train, handle)
    pickle.dump(word_list, handle)
    pickle.dump(word_vectors, handle)
    pickle.dump(word_dict, handle)
    pickle.dump(distance1_dict, handle)
    pickle.dump(distance2_dict, handle)
    pickle.dump(label_dict, handle)
    pickle.dump(sentMax, handle)
```

[ ]:

```
```

[ ]:

```
```