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
```

[11]:
```
%run ../utils.ipynb
```

[2]:
```
embSize = 200

ftrain='../data/EUADR_target_disease.csv'
# Replace with path of word embdding file
#wefile = "/mnt/admin/GDA_backup/Dataset/embeddings/PubMed-and-PMC-w2v.bin"
wefile = "../support/PubMed-and-PMC-w2v.bin"
random_seed=1331
```

[11]:
```
import pandas as pd
from tabulate import tabulate
ftrain = '../data/EUADR_target_disease.csv'
with open(ftrain, 'r', encoding='latin1') as file:
    first_line = file.readline()
    print(first_line)
import pandas as pd

ftrain = '../data/EUADR_target_disease.csv'

# Specify the delimiter for tabs
df = pd.read_csv(ftrain, encoding='latin1', sep='\t')


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

```
"ASSOCIATION_TYPE"      "PMID"  "NUM_SENTENCE"  "ENTITY1_TEXT"  "ENTITY1_INI"
```

```
"ENTITY1_END"   "ENTITY1_TYPE"  "ENTITY2_TEXT"  "ENTITY2_INI"   "ENTITY2_END"
"ENTITY2_TYPE"  "SENTENCE"

+----+--------------------+----------+----------------+--------------------+-----
----------+--------------+-----------------------------+--------------------+---
-----------+--------------+----------------------+--------------------------
-------------------------------------------------------------------------------
-----------------------------------------+
|    | ASSOCIATION_TYPE   |     PMID |   NUM_SENTENCE | ENTITY1_TEXT       |
ENTITY1_INI |   ENTITY1_END | ENTITY1_TYPE                | ENTITY2_TEXT       |
ENTITY2_INI |   ENTITY2_END | ENTITY2_TYPE         | SENTENCE
|
+====+====================+==========+================+====================+=====
==========+==============+=============================+====================+===
===========+==============+======================+==========================
===============================================================================
=========================================+
|  0 | SA                 | 17241106 |             16 | LRP5               |
19 |            23 | Genes & Molecular Sequences | osteoporosis       |
80 |            92 | Diseases & Disorders | Our work supported LRP5 genetic
variants as possible susceptibility factors for osteoporosis and fractures in
humans.                                  |
+----+--------------------+----------+----------------+--------------------+-----
----------+--------------+-----------------------------+--------------------+---
-----------+--------------+----------------------+--------------------------
-------------------------------------------------------------------------------
-----------------------------------------+
|  1 | SA                 | 17241106 |             17 | rs491347           |
20 |            28 | SNP & Sequence variations   | osteoporosis       |
108 |           120 | Diseases & Disorders | Especially, the SNP rs491347 and
its strongly associated SNPs (e.g., rs1784235) could be important to human
osteoporosis phenotypes.                    |
+----+--------------------+----------+----------------+--------------------+-----
----------+--------------+-----------------------------+--------------------+---
-----------+--------------+----------------------+--------------------------
-------------------------------------------------------------------------------
-----------------------------------------+
|  2 | SA                 | 17241106 |             17 | rs1784235          |
69 |            78 | SNP & Sequence variations   | osteoporosis       |
108 |           120 | Diseases & Disorders | Especially, the SNP rs491347 and
its strongly associated SNPs (e.g., rs1784235) could be important to human
osteoporosis phenotypes.                    |
+----+--------------------+----------+----------------+--------------------+-----
----------+--------------+-----------------------------+--------------------+---
-----------+--------------+----------------------+--------------------------
-------------------------------------------------------------------------------
-----------------------------------------+
|  3 | FA                 | 18697826 |              0 | fetal haemoglobin  |
```

```
93 |           110 | Genes & Molecular Sequences | beta-thalassaemia |
132 |           149 | Diseases & Disorders | The HBS1L-MYB intergenic region on
chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin
level in carriers of beta-thalassaemia. |
+----+-------------------+----------+---------------+-------------------+-----
----------+--------------+-----------------------------+-------------------+---
-----------+--------------+----------------------+---------------------------
---------------------------------------------------------------------------
-------------------------------------------+
```

[19]:
```python
import pandas as pd
from tabulate import tabulate

# Load the CSV file
ftrain = '../data/EUADR_target_disease.csv'
df = pd.read_csv(ftrain, encoding='latin1', sep='\t')

# Display the first line to understand the structure
with open(ftrain, 'r', encoding='latin1') as file:
    first_line = file.readline()
    print("First line of the file:", first_line)

# Display the column names to identify the one that corresponds to the entity␣
 ↪association type
print("Column names:", df.columns)

# Check for unique association types
unique_association_types = df['ASSOCIATION_TYPE'].unique()
print("Unique association types:", unique_association_types)

# Filter the DataFrame to exclude rows where the association type is NaN (if␣
 ↪applicable)
df_filtered = df[df['ASSOCIATION_TYPE'].notna()]

# Display the filtered rows
print(tabulate(df_filtered.head(15), headers='keys', tablefmt='grid'))
```

```
First line of the file: "ASSOCIATION_TYPE"      "PMID"  "NUM_SENTENCE"
"ENTITY1_TEXT"  "ENTITY1_INI"   "ENTITY1_END"   "ENTITY1_TYPE"  "ENTITY2_TEXT"
"ENTITY2_INI"   "ENTITY2_END"   "ENTITY2_TYPE"  "SENTENCE"

Column names: Index(['ASSOCIATION_TYPE', 'PMID', 'NUM_SENTENCE', 'ENTITY1_TEXT',
       'ENTITY1_INI', 'ENTITY1_END', 'ENTITY1_TYPE', 'ENTITY2_TEXT',
       'ENTITY2_INI', 'ENTITY2_END', 'ENTITY2_TYPE', 'SENTENCE'],
      dtype='object')
Unique association types: ['SA' 'FA' 'PA' nan]
+----+-------------------+----------+---------------+-----------------------
--------+--------------+--------------+-----------------------------+---------
```

```
-------------------------------+--------------+--------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
|    | ASSOCIATION_TYPE   |     PMID |   NUM_SENTENCE | ENTITY1_TEXT
|   ENTITY1_INI |   ENTITY1_END | ENTITY1_TYPE                | ENTITY2_TEXT
|   ENTITY2_INI |   ENTITY2_END | ENTITY2_TYPE         | SENTENCE
|
+====+====================+==========+================+=========================
========+===============+===============+=============================+=========
================================+===============+===============+================
======+=================================================================
=================================================================================
=================================================================================
=================================================================================
=================================================================================
=========================================+
|  0 | SA                 | 17241106 |             16 | LRP5
|            19 |            23 | Genes & Molecular Sequences | osteoporosis
|            80 |            92 | Diseases & Disorders | Our work supported LRP5
genetic variants as possible susceptibility factors for osteoporosis and
fractures in humans.
|
+----+--------------------+----------+----------------+-------------------------
--------+---------------+---------------+-----------------------------+---------
-------------------------------+---------------+---------------+----------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
|  1 | SA                 | 17241106 |             17 | rs491347
|            20 |            28 | SNP & Sequence variations   | osteoporosis
|           108 |           120 | Diseases & Disorders | Especially, the SNP
rs491347 and its strongly associated SNPs (e.g., rs1784235) could be important
to human osteoporosis phenotypes.
|
+----+--------------------+----------+----------------+-------------------------
--------+---------------+---------------+-----------------------------+---------
-------------------------------+---------------+---------------+----------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
-----------------------------------------+
|  2 | SA                 | 17241106 |             17 | rs1784235
|            69 |            78 | SNP & Sequence variations   | osteoporosis
|           108 |           120 | Diseases & Disorders | Especially, the SNP
rs491347 and its strongly associated SNPs (e.g., rs1784235) could be important
to human osteoporosis phenotypes.
|
+----+--------------------+----------+----------------+------------------------
--------+--------------+--------------+-----------------------------+---------
--------------------------------+--------------+---------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
|  3 | FA                 | 18697826 |              0 | fetal haemoglobin
|            93 |            110 | Genes & Molecular Sequences | beta-
thalassaemia                       |           132 |           149 | Diseases &
Disorders | The HBS1L-MYB intergenic region on chromosome 6q23 is a quantitative
trait locus controlling fetal haemoglobin level in carriers of beta-
thalassaemia.
|
+----+--------------------+----------+----------------+------------------------
--------+--------------+--------------+-----------------------------+---------
--------------------------------+--------------+---------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
|  4 | PA                 | 18697826 |              1 | HbF
|            19 |            22 | Genes & Molecular Sequences | HBB disorders
|            64 |            77 | Diseases & Disorders | Fetal haemoglobin (HbF)
level modifies the clinical severity of HBB disorders.
|
+----+--------------------+----------+----------------+------------------------
--------+--------------+--------------+-----------------------------+---------
--------------------------------+--------------+---------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
|  5 | PA                 | 18697826 |              1 | Fetal haemoglobin
```

| 0 | 17 | Genes & Molecular Sequences | HBB disorders
| 64 | 77 | Diseases & Disorders | Fetal haemoglobin (HbF) level modifies the clinical severity of HBB disorders.
|

| 6 | FA | 18697826 | 0 | HBS1L-MYB
| 4 | 13 | Genes & Molecular Sequences | beta-thalassaemia | 132 | 149 | Diseases & Disorders | The HBS1L-MYB intergenic region on chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin level in carriers of beta-thalassaemia.
|

| 7 | FA | 18697826 | 0 | 6q23
| 46 | 50 | Genes & Molecular Sequences | beta-thalassaemia | 132 | 149 | Diseases & Disorders | The HBS1L-MYB intergenic region on chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin level in carriers of beta-thalassaemia.
|

| 8 | FA | 18697826 | 7 | HbF
| 87 | 90 | Genes & Molecular Sequences | HBB disorders
| 196 | 209 | Diseases & Disorders | Functional studies to

unravel the biological significance of this region in regulating HbF production is clearly indicated, which may lead to new strategies to modify the disease course of severe HBB disorders.
|
+----+------------------+----------+----------------+------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 9 | PA | 18322311 | 0 | TNFalpha gene
polymorphisms | 16 | 43 | Genes & Molecular Sequences
| multiple sclerosis | 69 | 87 |
Diseases & Disorders | IL-1, IL-1R and TNFalpha gene polymorphisms in Iranian patients with multiple sclerosis.
|
+----+------------------+----------+----------------+------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 10 | FA | 18322311 | 7 | IL-1alpha TT -889
| 35 | 52 | SNP & Sequence variations | multiple
sclerosis | 406 | 424 | Diseases &
Disorders | On the other hand the frequency of IL-1alpha TT -889 genotype (p=0.028), IL-1R C pst1 1970 allele (p=0.0001) and CC genotype (p=0.00006), TNFalpha G -308 allele (p=0.0002) and GG genotype (p=0.000001) decreased significantly in the patients versus normal subjects.These results suggest that polymorphic variations of these pro-inflammatory cytokines may play an important role in susceptibility of Iranian multiple sclerosis patients. |
+----+------------------+----------+----------------+------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 11 | PA | 18322311 | 7 | TNFalpha G -308 allele
| 138 | 160 | Genes & Molecular Sequences | multiple

sclerosis | 406 | 424 | Diseases & Disorders | On the other hand the frequency of IL-1alpha TT -889 genotype (p=0.028), IL-1R C pst1 1970 allele (p=0.0001) and CC genotype (p=0.00006), TNFalpha G -308 allele (p=0.0002) and GG genotype (p=0.000001) decreased significantly in the patients versus normal subjects.These results suggest that polymorphic variations of these pro-inflammatory cytokines may play an important role in susceptibility of Iranian multiple sclerosis patients. |

| 12 | PA | 18322311 | 0 | IL-1 | 0 | 4 | Genes & Molecular Sequences | multiple sclerosis | 69 | 87 | Diseases & Disorders | IL-1, IL-1R and TNFalpha gene polymorphisms in Iranian patients with multiple sclerosis. |

| 13 | PA | 18347176 | 0 | T393C polymorphism of the GNAS1 | 127 | 158 | SNP & Sequence variations | oropharyngeal | 37 | 50 | Diseases & Disorders | Overall and relapse-free survival in oropharyngeal and hypopharyngeal squamous cell carcinoma are associated with genotypes of T393C polymorphism of the GNAS1 gene. |

| 14 | PA | 18347176 | 0 | T393C polymorphism of

```
the GNAS1 |           127 |           158 | SNP & Sequence variations    |
hypopharyngeal squamous cell carcinoma |            55 |            93 |
Diseases & Disorders | Overall and relapse-free survival in oropharyngeal and
hypopharyngeal squamous cell carcinoma are associated with genotypes of T393C
polymorphism of the GNAS1 gene.
|
+----+--------------------+----------+----------------+-------------------------
--------+--------------+--------------+----------------------------+---------
--------------------------------+--------------+---------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+
```

[16]:
```python
import pandas as pd
from tabulate import tabulate

# Load the CSV file
ftrain = '../data/EUADR_target_disease.csv'
df = pd.read_csv(ftrain, encoding='latin1', sep='\t')

# Display the first line to understand the structure
with open(ftrain, 'r', encoding='latin1') as file:
    first_line = file.readline()
    print("First line of the file:", first_line)

# Display the column names to identify the one that corresponds to the entity␣
 ↪association type
print("Column names:", df.columns)

# Get unique association types and their counts
association_type_counts = df['ASSOCIATION_TYPE'].value_counts()
print("Association types and their counts:")
print(tabulate(association_type_counts.reset_index(), headers=['Association␣
 ↪Type', 'Count'], tablefmt='grid'))

# If you want to see a sample of rows for each association type
# Display the first 10 rows for each unique association type
for assoc_type in df['ASSOCIATION_TYPE'].unique():
    print(f"\nSample rows for association type '{assoc_type}':")
    sample_df = df[df['ASSOCIATION_TYPE'] == assoc_type].head(10)
    print(tabulate(sample_df, headers='keys', tablefmt='grid'))
```

```
First line of the file: "ASSOCIATION_TYPE"      "PMID"  "NUM_SENTENCE"
"ENTITY1_TEXT"  "ENTITY1_INI"   "ENTITY1_END"   "ENTITY1_TYPE"  "ENTITY2_TEXT"
"ENTITY2_INI"   "ENTITY2_END"   "ENTITY2_TYPE"  "SENTENCE"
```

```
Column names: Index(['ASSOCIATION_TYPE', 'PMID', 'NUM_SENTENCE', 'ENTITY1_TEXT',
       'ENTITY1_INI', 'ENTITY1_END', 'ENTITY1_TYPE', 'ENTITY2_TEXT',
       'ENTITY2_INI', 'ENTITY2_END', 'ENTITY2_TYPE', 'SENTENCE'],
      dtype='object')
Association types and their counts:
+----+--------------------+---------+
|    | Association Type   |   Count |
+====+====================+=========+
|  0 | PA                 |     213 |
+----+--------------------+---------+
|  1 | FA                 |      93 |
+----+--------------------+---------+
|  2 | SA                 |      30 |
+----+--------------------+---------+

Sample rows for association type 'SA':
+-----+--------------------+----------+----------------+----------------+-------
--------+---------------+-----------------------------+------------------------
---+---------------+---------------+----------------------+--------------------
-----------------------------------------------------------------------------
-------------------------------------------------------+
|     | ASSOCIATION_TYPE   |     PMID |   NUM_SENTENCE | ENTITY1_TEXT   |
ENTITY1_INI |   ENTITY1_END | ENTITY1_TYPE                | ENTITY2_TEXT
|   ENTITY2_INI |   ENTITY2_END | ENTITY2_TYPE         | SENTENCE
|
+=====+====================+==========+================+================+=======
========+===============+=============================+========================
===+===============+===============+======================+====================
=============================================================================
=======================================================+
|   0 | SA                 | 17241106 |             16 | LRP5           |
19 |            23 | Genes & Molecular Sequences | osteoporosis               |
80 |            92 | Diseases & Disorders | Our work supported LRP5 genetic
variants as possible susceptibility factors for osteoporosis and fractures in
humans.                                    |
+-----+--------------------+----------+----------------+----------------+-------
--------+---------------+-----------------------------+------------------------
---+---------------+---------------+----------------------+--------------------
-----------------------------------------------------------------------------
-------------------------------------------------------+
|   1 | SA                 | 17241106 |             17 | rs491347       |
20 |            28 | SNP & Sequence variations   | osteoporosis               |
108 |           120 | Diseases & Disorders | Especially, the SNP rs491347 and
its strongly associated SNPs (e.g., rs1784235) could be important to human
osteoporosis phenotypes.                        |
+-----+--------------------+----------+----------------+----------------+-------
--------+---------------+-----------------------------+------------------------
```

```
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|   2 | SA                | 17241106 |             17 | rs1784235      |
69 |            78 | SNP & Sequence variations   | osteoporosis               |
108 |           120 | Diseases & Disorders | Especially, the SNP rs491347 and
its strongly associated SNPs (e.g., rs1784235) could be important to human
osteoporosis phenotypes.                        |
+-----+-------------------+----------+----------------+----------------+-------
--------+--------------+-----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  23 | SA                | 17593927 |             11 | XRCC1          |
0 |             5 | Genes & Molecular Sequences | gastric cancer             |
61 |            75 | Diseases & Disorders | XRCC1 genotyping might make tailor
chemotherapy possible for gastric cancer patients treated with oxaliplatin-based
chemotherapy.                           |
+-----+-------------------+----------+----------------+----------------+-------
--------+--------------+-----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  32 | SA                | 19035720 |              6 | rs42524        |
4 |            11 | SNP & Sequence variations   | IAs                        |
79 |            82 | Diseases & Disorders | The rs42524 polymorphism of COL1A2
could be a genetic risk factor for sporadic IAs among individuals of Chinese Han
ethnicity.                              |
+-----+-------------------+----------+----------------+----------------+-------
--------+--------------+-----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  43 | SA                | 19110417 |              2 | CYP1A1         |
56 |            62 | Genes & Molecular Sequences | HCC                        |
70 |            73 | Diseases & Disorders | However, the contribution of common
genetic variants in CYP1A1 to the HCC risk in Chinese populations has not been
thoroughly investigated.                |
+-----+-------------------+----------+----------------+----------------+-------
--------+--------------+-----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  50 | SA                | 19110417 |             11 | CYP1A1         |
54 |            60 | Genes & Molecular Sequences | HCC                        |
97 |           100 | Diseases & Disorders | Our results suggested that common
genetic variants in CYP1A1 may modulate the risk of developing HCC in the study
population, particularly in non-smokers. |
```

```
+-----+-------------------+---------+---------------+---------------+-------
--------+--------------+----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  86 | SA                | 18708184 |             1 | NRG3          |
51 |           55 | Genes & Molecular Sequences | schizophrenia             |
65 |           78 | Diseases & Disorders | The study investigated the possible
association of NRG3 gene and schizophrenia in a Han Chinese population.
|
+-----+-------------------+---------+---------------+---------------+-------
--------+--------------+----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
|  98 | SA                | 19098911 |             6 | NLRP3         |
31 |           36 | Genes & Molecular Sequences | Crohn's disease           |
130 |          145 | Diseases & Disorders | These results suggest that the
NLRP3 region is also implicated in the susceptibility of more common
inflammatory diseases such as Crohn's disease.         |
+-----+-------------------+---------+---------------+---------------+-------
--------+--------------+----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+
| 102 | SA                | 19115152 |             4 | MDR1          |
32 |           36 | Genes & Molecular Sequences | inflammatory bowel disease |
93 |          119 | Diseases & Disorders | Therefore, the mutations of the MDR1
gene are thought to be related with the pathogenesis of inflammatory bowel
disease.                                  |
+-----+-------------------+---------+---------------+---------------+-------
--------+--------------+----------------------------+-----------------------
---+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------+

Sample rows for association type 'FA':
+----+-------------------+---------+---------------+----------------------+-
-------------+--------------+----------------------------+-------------------
----+--------------+--------------+---------------------+--------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
---------------+
|    | ASSOCIATION_TYPE   |     PMID |   NUM_SENTENCE | ENTITY1_TEXT         |
ENTITY1_INI |   ENTITY1_END | ENTITY1_TYPE               | ENTITY2_TEXT
```

```
|   ENTITY2_INI |   ENTITY2_END | ENTITY2_TYPE         | SENTENCE
|
+====+==================+==========+================+======================+=
=============+===============+============================+==================
====+===============+==============+======================+==================
==========================================================================
==========================================================================
==========================================================================
==========================================================================
==========================================================================
================+
|  3 | FA                 | 18697826 |              0 | fetal haemoglobin     |
93 |           110 | Genes & Molecular Sequences | beta-thalassaemia     |
132 |           149 | Diseases & Disorders | The HBS1L-MYB intergenic region on
chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin
level in carriers of beta-thalassaemia.
|
+----+------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
---------------+
|  6 | FA                 | 18697826 |              0 | HBS1L-MYB             |
4 |            13 | Genes & Molecular Sequences | beta-thalassaemia     |
132 |           149 | Diseases & Disorders | The HBS1L-MYB intergenic region on
chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin
level in carriers of beta-thalassaemia.
|
+----+------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
----------------------------------------------------------------------------
---------------+
|  7 | FA                 | 18697826 |              0 | 6q23                  |
46 |            50 | Genes & Molecular Sequences | beta-thalassaemia     |
132 |           149 | Diseases & Disorders | The HBS1L-MYB intergenic region on
chromosome 6q23 is a quantitative trait locus controlling fetal haemoglobin
level in carriers of beta-thalassaemia.
|
+----+------------------+----------+----------------+----------------------+-
```

```
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
---------------+
|  8 | FA                 | 18697826 |              7 | HbF                  |
87 |            90 | Genes & Molecular Sequences | HBB disorders         |
196 |           209 | Diseases & Disorders | Functional studies to unravel the
biological significance of this region in regulating HbF production is clearly
indicated, which may lead to new strategies to modify the disease course of
severe HBB disorders.
|
+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
---------------+
| 10 | FA                 | 18322311 |              7 | IL-1alpha TT -889     |
35 |            52 | SNP & Sequence variations   | multiple sclerosis    |
406 |           424 | Diseases & Disorders | On the other hand the frequency of
IL-1alpha TT -889 genotype (p=0.028), IL-1R C pst1 1970 allele (p=0.0001) and CC
genotype (p=0.00006), TNFalpha G -308 allele (p=0.0002) and GG genotype
(p=0.000001) decreased significantly in the patients versus normal
subjects.These results suggest that polymorphic variations of these pro-
inflammatory cytokines may play an important role in susceptibility of Iranian
multiple sclerosis patients. |
+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
---------------+
| 15 | FA                 | 18347176 |              9 | T393C                |
4 |             9 | SNP & Sequence variations   | hypopharyngeal cancer |
130 |           151 | Diseases & Disorders | The T393C SNP could be considered
as a genetic marker to predict the clinical course of patients suffering from
oropharyngeal and hypopharyngeal cancer.
|
```

+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----

---------------+
| 17 | FA | 18347176 | 3 | T393C |
28 | 33 | SNP & Sequence variations | oropharyngeal |
121 | 134 | Diseases & Disorders | The prognostic value of the T393C
SNP was evaluated in an unselected series of patients treated with curative
intent for oropharyngeal and hypopharyngeal squamous cell carcinomas, including
all tumor stages with different therapeutic regimens.
|
+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----

---------------+
| 18 | FA | 17430902 | 3 | Wnt |
69 | 72 | Genes & Molecular Sequences | colon cancer |
25 | 37 | Diseases & Disorders | However, the majority of colon
cancer cells have deregulation of the Wnt/beta-catenin pathway.
|
+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----

---------------+
| 19 | FA | 17430902 | 0 | Krüppel-like factor 5 |
87 | 108 | Genes & Molecular Sequences | colon cancer |
51 | 63 | Diseases & Disorders | Lysophosphatidic acid facilitates
proliferation of colon cancer cells via induction of Krüppel-like factor 5.
|
+----+--------------------+----------+----------------+----------------------+-
-------------+---------------+----------------------------+------------------
----+---------------+--------------+----------------------+------------------
----

```
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---------------+
| 28 | FA                 | 19097922 |              1 | CYP4F2                |
45 |            51 | Genes & Molecular Sequences | MI                    |
139 |           141 | Diseases & Disorders | This study assessed associations
between the CYP4F2 gene and myocardial infarction (MI), using a haplotype-based
case-control study of 234 MI patients and 248 controls genotyped for 5 single-
nucleotide polymorphisms (rs3093105, rs3093135, rs1558139, rs2108622,
rs3093200).
|
+----+--------------------+----------+----------------+-----------------------+-
-------------+---------------+---------------+----------------------------+------------------
----+---------------+---------------+----------------------+--------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---------------+

Sample rows for association type 'PA':
+----+--------------------+----------+----------------+------------------------
--------+---------------+---------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+-----------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-----------------------------------------+
|    | ASSOCIATION_TYPE   |     PMID |   NUM_SENTENCE | ENTITY1_TEXT
|   ENTITY1_INI |   ENTITY1_END | ENTITY1_TYPE                | ENTITY2_TEXT
|   ENTITY2_INI |   ENTITY2_END | ENTITY2_TYPE         | SENTENCE
|
+====+====================+==========+================+========================
========+===============+===============+============================+=========
================================+===============+===============+================
======+=================================================================
=================================================================================
=================================================================================
=================================================================================
=================================================================================
=========================================+
|  4 | PA                 | 18697826 |              1 | HbF
|            19 |            22 | Genes & Molecular Sequences | HBB disorders
```

| 64 | 77 | Diseases & Disorders | Fetal haemoglobin (HbF) level modifies the clinical severity of HBB disorders.
|

| 5 | PA | 18697826 | 1 | Fetal haemoglobin
| 0 | 17 | Genes & Molecular Sequences | HBB disorders
| 64 | 77 | Diseases & Disorders | Fetal haemoglobin (HbF) level modifies the clinical severity of HBB disorders.
|

| 9 | PA | 18322311 | 0 | TNFalpha gene polymorphisms | 16 | 43 | Genes & Molecular Sequences | multiple sclerosis | 69 | 87 | Diseases & Disorders | IL-1, IL-1R and TNFalpha gene polymorphisms in Iranian patients with multiple sclerosis.
|

| 11 | PA | 18322311 | 7 | TNFalpha G -308 allele | 138 | 160 | Genes & Molecular Sequences | multiple sclerosis | 406 | 424 | Diseases & Disorders | On the other hand the frequency of IL-1alpha TT -889 genotype (p=0.028), IL-1R C pst1 1970 allele (p=0.0001) and CC genotype (p=0.00006), TNFalpha G -308 allele (p=0.0002) and GG genotype (p=0.000001) decreased significantly in the patients versus normal subjects.These results suggest that

polymorphic variations of these pro-inflammatory cytokines may play an important role in susceptibility of Iranian multiple sclerosis patients. |
+----+--------------------+----------+----------------+------------------------
--------+---------------+---------------+-----------------------------+---------
--------------------------------+---------------+---------------+---------------
------+---

---+
| 12 | PA | 18322311 | 0 | IL-1
| 0 | 4 | Genes & Molecular Sequences | multiple
sclerosis | 69 | 87 | Diseases &
Disorders | IL-1, IL-1R and TNFalpha gene polymorphisms in Iranian patients with
multiple sclerosis.
|
+----+--------------------+----------+----------------+------------------------
--------+---------------+---------------+-----------------------------+---------
--------------------------------+---------------+---------------+---------------
------+---

---+
| 13 | PA | 18347176 | 0 | T393C polymorphism of
the GNAS1 | 127 | 158 | SNP & Sequence variations |
oropharyngeal | 37 | 50 |
Diseases & Disorders | Overall and relapse-free survival in oropharyngeal and
hypopharyngeal squamous cell carcinoma are associated with genotypes of T393C
polymorphism of the GNAS1 gene.
|
+----+--------------------+----------+----------------+------------------------
--------+---------------+---------------+-----------------------------+---------
--------------------------------+---------------+---------------+---------------
------+---

---+
| 14 | PA | 18347176 | 0 | T393C polymorphism of
the GNAS1 | 127 | 158 | SNP & Sequence variations |
hypopharyngeal squamous cell carcinoma | 55 | 93 |
Diseases & Disorders | Overall and relapse-free survival in oropharyngeal and
hypopharyngeal squamous cell carcinoma are associated with genotypes of T393C
polymorphism of the GNAS1 gene.

|
+----+-------------------+----------+----------------+-------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 16 | PA | 18347176 | 9 | T393C
| 4 | 9 | SNP & Sequence variations | oropharyngeal
| 112 | 125 | Diseases & Disorders | The T393C SNP could be
considered as a genetic marker to predict the clinical course of patients
suffering from oropharyngeal and hypopharyngeal cancer.
|
+----+-------------------+----------+----------------+-------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 20 | PA | 17430902 | 2 | beta-catenin
| 99 | 111 | Genes & Molecular Sequences | colon cancer
| 57 | 69 | Diseases & Disorders | A recent study showed
that LPA-mediated proliferation of colon cancer cells requires activation of
beta-catenin.
|
+----+-------------------+----------+----------------+-------------------------
--------+--------------+--------------+----------------------------+---------
-------------------------------+--------------+---------------+----------------
------+---

---+
| 21 | PA | 18930278 | 6 | CYP3A5 A6986G genotype
*3/*3 | 29 | 57 | Genes & Molecular Sequences |
leukocytopenia | 211 | 225 |
Diseases & Disorders | On multivariate analysis the CYP3A5 A6986G genotype *3/*3
(OR 8.205, 95% CI 1.616-41.667, p = 0.011) and smaller number of treatment
cycles (OR 0.156, 95% CI 0.037-0.659, p = 0.011) were independent factors for
leukocytopenia (grade 3 or greater) throughout the period of chemotherapy.
|

```
+----+-------------------+----------+----------------+------------------------
--------+---------------+---------------+-----------------------------+---------
-------------------------------+--------------+--------------+---------------
------+-----------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------------------------------------------
-----------------------------------------+

Sample rows for association type 'nan':
+--------------------+--------+----------------+----------------+---------------
+--------------+----------------+---------------+--------------+-------------
--+----------------+------------+
| ASSOCIATION_TYPE   | PMID   | NUM_SENTENCE   | ENTITY1_TEXT   | ENTITY1_INI
| ENTITY1_END   | ENTITY1_TYPE   | ENTITY2_TEXT   | ENTITY2_INI   | ENTITY2_END
| ENTITY2_TYPE   | SENTENCE   |
+====================+========+================+================+===============
+==============+================+===============+==============+=============
==+================+============+
+--------------------+--------+----------------+----------------+---------------
+--------------+----------------+---------------+--------------+-------------
--+----------------+------------+
```

[18]:
```
import pandas as pd

# Load the CSV file
ftrain = '../data/EUADR_target_disease.csv'
df = pd.read_csv(ftrain, encoding='latin1', sep='\t')

# Display the column names to identify the one that corresponds to the entity␣
 ↪association type
print("Column names:", df.columns)

# Check if 'SA' exists in the 'ASSOCIATION_TYPE' column
association_types = df['ASSOCIATION_TYPE'].unique()
if 'NA' in association_types:
    print("Association type 'SA' is present in the data.")
else:
    print("Association type 'SA' is not present in the data.")
```

```
Column names: Index(['ASSOCIATION_TYPE', 'PMID', 'NUM_SENTENCE', 'ENTITY1_TEXT',
       'ENTITY1_INI', 'ENTITY1_END', 'ENTITY1_TYPE', 'ENTITY2_TEXT',
       'ENTITY2_INI', 'ENTITY2_END', 'ENTITY2_TYPE', 'SENTENCE'],
      dtype='object')
Association type 'SA' is not present in the data.
```

2 Read Data

[13]:
```
from tabulate import tabulate
Tr_sent_contents, Tr_entity1_list, Tr_entity2_list, Tr_sent_lables = utils.
 ↪dataRead_befree_EUADR(ftrain)
# Convert the lists to a DataFrame
df = pd.DataFrame({
    'Sent Contents': Tr_sent_contents,
    'Entity 1': Tr_entity1_list,
    'Entity 2': Tr_entity2_list,
    'Sent Labels': Tr_sent_lables
})


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

Tr_word_list, Tr_d1_list, Tr_d2_list  = utils.
 ↪get_wordList_and_distances_befree(Tr_sent_contents)

df = pd.DataFrame({
    'word list': Tr_word_list,
    'd1 list': Tr_d1_list,
    'd2 list': Tr_d2_list,
})


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

print ("train_size", len(Tr_word_list))
```

```
Input File Reading
train_size 355
```

2.1 Prepare Lable Matrix

[17]:
```
# Y : is positive association
# N: is negative association
label_dict = {'FA':0, 'NA':0,'PA':1,'SA':1}

Y_t = mapLabelToId_befree_EUADR(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
```

3 Generate Word and Position Embedding Vectors

3.0.1 Word Embedding

[18]:
```
word_dict, word_to_id, id_to_word = word_mapping(Tr_word_list)

print( "word dictonary length", len(word_dict))

# Word Embedding
word_vectors = readWordEmb(word_dict,id_to_word,word_to_id, wefile,␣
 ↪embSize,limit=1356)
W_train =   mapWordToId(Tr_word_list, word_to_id)
print("W_train",len(W_train))
print("word_vectors",len(word_vectors))
```

```
Found 1355 unique words (10062 in total)
word dictonary length 1355
Reading word vectors
Loaded 1356 pretrained embeddings.
number of unknown word in word embedding 814
W_train 355
word_vectors 1355
```

3.0.2 Position Embedding

[19]:
```
d1_dict = makeDistanceList([Tr_d1_list])
d2_dict = makeDistanceList([Tr_d2_list])
d1_train = mapWordToId_list(Tr_d1_list, d1_dict)
d2_train = mapWordToId_list(Tr_d2_list, d2_dict)
```

3.0.3 Pad Embdding Vectors

[20]:
```
train_sent_lengths= findSentLengths([Tr_word_list])

sentMax =max(max(train_sent_lengths[:]))

W_train, d1_train, d2_train = paddData([W_train, d1_train, d2_train ],␣
 ↪sentMax,padd_num=1354)
print("sentMax",sentMax)
print("W_train",len(W_train))
print("d1_train",len(d1_train))
print("d2_train",len(d2_train))
```

```
sentMax 102
W_train 355
d1_train 355
d2_train 355
```

4 Save Prepared Data as Pickle File

[21]:
```
with open('../data/pickles/befree_EUADR_2class_PubMed-and-PMC-w2v.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(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)
```

[ ]:

[ ]:

[ ]: