AILA-AI-for-Legal-Assisatance-2019_bm25_doc2vec

April 25, 2021

1 AILA: Artificial Intelligence for Legal Assisatance

1.1 Similar Case Matching

1.1.1 We are given a dataset consisting of 2914 prior cases and a test dataset of 50 queries. We need to retrieve the most similar prior case for each of the queries.

[52]:
```
# This Python 3 environment comes with many helpful analytics libraries␣
↪installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/
↪docker-python

#Imports
import glob
import functools
import datetime as dt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import random
import re
import numpy as np
import pandas as pd

# Input data files are available in the read-only "../input/" directory

import os
#for dirname, _, filenames in os.walk('/kaggle/input/legalai/Object_casedocs'):
#    for filename in filenames:
#        print(os.path.join(dirname, filename))
# the above code will list all files under the input directory
```

1.2 Data Handling

1.2.1 We first wrap up all the text files into a single csv(comma separated file)

[93]:

```
import glob
import csv

read_files = glob.glob('/kaggle/input/legalai/Object_casedocs/*')

with open("object_casedocs.csv", "w") as outfile:
    w=csv.writer(outfile)
    for f in read_files:
        with open(f, "r") as infile:
            w.writerow([" ".join([line.strip() for line in infile])])

lst_arr = os.listdir('/kaggle/input/legalai/Object_casedocs/')
df_filename = pd.DataFrame(lst_arr, columns = ['Name'])
df_filename
```

[93]:

```
           Name
0      C757.txt
1     C1092.txt
2     C1985.txt
3       C39.txt
4     C2055.txt
…           …
2909   C792.txt
2910  C2775.txt
2911   C924.txt
2912   C432.txt
2913   C244.txt

[2914 rows x 1 columns]
```

[107]:

```
evaluate = pd.read_csv('/kaggle/input/legalai/relevance_judgments_priorcases.
 ↪txt', delimiter = " ", header = None)
evaluate.columns = ["Query_Number", "Q0", "Document" ,"Relevance"]
evaluate=evaluate.drop(columns=["Q0"])
evaluate
```

[107]:

```
      Query_Number Document  Relevance
0          AILA_Q1     C168          0
1          AILA_Q1     C382          0
2          AILA_Q1     C428          0
3          AILA_Q1     C949          0
4          AILA_Q1    C2303          0
…             …        …         …
```

```
145695     AILA_Q50    C1367          0
145696     AILA_Q50    C2079          0
145697     AILA_Q50    C2066          0
145698     AILA_Q50    C1951          0
145699     AILA_Q50    C1111          0

[145700 rows x 3 columns]
```

1.2.2 A Glimpse about how the data inside the csv file looks!

[54]:
```
df = pd.read_csv('object_casedocs.csv',header=None)
df.columns = ["Text"]
df
```

[54]:
```
                                                  Text
0     L. Laxmikanta v State by Superintendent of Pol…
1     Homi Rajvansh v State of Maharashtra and other…
2     Direct Recruit Class Ii Engineering OfficersAs…
3     Rajinder Kumar Kindra v Delhi Administration T…
4     Kalyan and Others v State of Uttar Pradesh Sup…
…                                                  …
2909  Haryana State Cooperative Labour and others v …
2910  State of Karnataka v Chikkahottappa Alias Vara…
2911  Kilari Malakondiaah @ Malayadri and Others v S…
2912  Kanthimathy Plantations Pvt- Limited v State O…
2913  Union of India and Others v K. P. Prabhakaran …

[2914 rows x 1 columns]
```

[55]:
```
df = pd.concat([df_filename,df], axis = 1)
df
```

[55]:
```
           Name                                              Text
0      C757.txt  L. Laxmikanta v State by Superintendent of Pol…
1     C1092.txt  Homi Rajvansh v State of Maharashtra and other…
2     C1985.txt  Direct Recruit Class Ii Engineering OfficersAs…
3       C39.txt  Rajinder Kumar Kindra v Delhi Administration T…
4     C2055.txt  Kalyan and Others v State of Uttar Pradesh Sup…
…           …                                                …
2909   C792.txt  Haryana State Cooperative Labour and others v …
2910  C2775.txt  State of Karnataka v Chikkahottappa Alias Vara…
2911   C924.txt  Kilari Malakondiaah @ Malayadri and Others v S…
2912   C432.txt  Kanthimathy Plantations Pvt- Limited v State O…
2913   C244.txt  Union of India and Others v K. P. Prabhakaran …

[2914 rows x 2 columns]
```

1.2.3 Let us get some basic information about the data

[56]:
```
len(df)
```

[56]:
```
2914
```

[57]:
```
df.shape
```

[57]:
```
(2914, 2)
```

[58]:
```
df.info
```

[58]:
```
<bound method DataFrame.info of            Name
Text
0      C757.txt  L. Laxmikanta v State by Superintendent of Pol…
1     C1092.txt  Homi Rajvansh v State of Maharashtra and other…
2     C1985.txt  Direct Recruit Class Ii Engineering OfficersAs…
3       C39.txt  Rajinder Kumar Kindra v Delhi Administration T…
4     C2055.txt  Kalyan and Others v State of Uttar Pradesh Sup…
…           …                                                 …
2909   C792.txt  Haryana State Cooperative Labour and others v …
2910  C2775.txt  State of Karnataka v Chikkahottappa Alias Vara…
2911   C924.txt  Kilari Malakondiaah @ Malayadri and Others v S…
2912   C432.txt  Kanthimathy Plantations Pvt- Limited v State O…
2913   C244.txt  Union of India and Others v K. P. Prabhakaran …

[2914 rows x 2 columns]>
```

1.3 Text preprocessing techniques: Cleansing the data

1.3.1 1. Convert to lowercase, remove punctuation and special characters, using RegeX and strip

1.3.2 2. Remove stopwords

1.3.3 3. Stemming

1.3.4 4. Lemmatization

[59]:
```
import re
#Convert lowercase remove punctuation and Character and then strip
text = df.iloc[0]
print(text)
text = re.sub(r'[^\w\s]', '', str(text).lower().strip())
txt = text.split()
print(txt)
```

```
Name                                              C757.txt
Text    L. Laxmikanta v State by Superintendent of Pol…
Name: 0, dtype: object
['name', 'c757txt', 'text', 'l', 'laxmikanta', 'v', 'state', 'by',
'superintendent', 'of', 'pol', 'name', '0', 'dtype', 'object']
```

[60]:

```
#remove stopwords
import nltk
lst_stopwords = nltk.corpus.stopwords.words("english")
txt = [word for word in txt if word not in lst_stopwords]
print(txt)
```

```
['name', 'c757txt', 'text', 'l', 'laxmikanta', 'v', 'state', 'superintendent',
'pol', 'name', '0', 'dtype', 'object']
```

[61]:

```
#stemming
ps = nltk.stem.porter.PorterStemmer()
print([ps.stem(word) for word in txt])
```

```
['name', 'c757txt', 'text', 'l', 'laxmikanta', 'v', 'state', 'superintend',
'pol', 'name', '0', 'dtype', 'object']
```

[62]:

```
#Lemmetization
nltk.download('wordnet')
lem = nltk.stem.wordnet.WordNetLemmatizer()
print([lem.lemmatize(word) for word in txt])
```

```
[nltk_data] Downloading package wordnet to /usr/share/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
['name', 'c757txt', 'text', 'l', 'laxmikanta', 'v', 'state', 'superintendent',
'pol', 'name', '0', 'dtype', 'object']
```

1.4 Preprocessing the data: Apply these techniques on all records of the dataset

[63]:

```
#to apply all the technique to all the records on dataset
def utils_preprocess_text(text, flg_stemm=True, flg_lemm =True,␣
 ↪lst_stopwords=None ):
    text = re.sub(r'[^\w\s]', '', str(text).lower().strip())

    #tokenization(convert from string to List)
    lst_text = text.split()

    #remove stopwords
    if lst_stopwords is not None:
        lst_text = [word for word in lst_text if word not in
                    lst_stopwords]
```

```
     #stemming
    if flg_stemm == True:
        ps = nltk.stem.porter.PorterStemmer()
        lst_text = [ps.stem(word) for word in lst_text]

    #Lemmentization
    if flg_lemm == True:
        lem = nltk.stem.wordnet.WordNetLemmatizer()
        lst_text = [lem.lemmatize(word) for word in lst_text]

    # back to string from list
    text = " ".join(lst_text)
    return text

df['clean_text'] = df['Text'].apply(lambda x: utils_preprocess_text(x,␣
 ↪flg_stemm = False, flg_lemm=True))
```

1.5 A Glimpse into the cleansed data!

[64]:

```
#df
df
```

[64]:

```
           Name                                               Text  \
0      C757.txt  L. Laxmikanta v State by Superintendent of Pol…
1     C1092.txt  Homi Rajvansh v State of Maharashtra and other…
2     C1985.txt  Direct Recruit Class Ii Engineering OfficersAs…
3       C39.txt  Rajinder Kumar Kindra v Delhi Administration T…
4     C2055.txt  Kalyan and Others v State of Uttar Pradesh Sup…
…          …                                                …
2909   C792.txt  Haryana State Cooperative Labour and others v …
2910  C2775.txt  State of Karnataka v Chikkahottappa Alias Vara…
2911   C924.txt  Kilari Malakondiaah @ Malayadri and Others v S…
2912   C432.txt  Kanthimathy Plantations Pvt- Limited v State O…
2913   C244.txt  Union of India and Others v K. P. Prabhakaran …

                                             clean_text
0     l laxmikanta v state by superintendent of poli…
1     homi rajvansh v state of maharashtra and other…
2     direct recruit class ii engineering officersas…
3     rajinder kumar kindra v delhi administration t…
4     kalyan and others v state of uttar pradesh sup…
…                                                 …
2909  haryana state cooperative labour and others v …
2910  state of karnataka v chikkahottappa alias vara…
2911  kilari malakondiaah malayadri and others v sta…
2912  kanthimathy plantation pvt limited v state of …
```

```
2913  union of india and others v k p prabhakaran su…

[2914 rows x 3 columns]
```

[65]:

```
train = df["clean_text"]
train
```

[65]:

```
0       l laxmikanta v state by superintendent of poli…
1       homi rajvansh v state of maharashtra and other…
2       direct recruit class ii engineering officersas…
3       rajinder kumar kindra v delhi administration t…
4       kalyan and others v state of uttar pradesh sup…
                              …
2909    haryana state cooperative labour and others v …
2910    state of karnataka v chikkahottappa alias vara…
2911    kilari malakondiaah malayadri and others v sta…
2912    kanthimathy plantation pvt limited v state of …
2913    union of india and others v k p prabhakaran su…
Name: clean_text, Length: 2914, dtype: object
```

[67]:

```
import texthero as hero
```

1.5.1 Creating a test dataframe using the Query file

[68]:

```
test = pd.read_csv("/kaggle/input/legalai/Query_doc.txt",delimiter =␣
 ↪"|",header=None)
test.columns = ["AILA","NAN", "Query"]
test=test.drop(columns=["AILA","NAN"])
```

[69]:

```
test
```

[69]:

```
                                                Query
0   The appellant on February 9, 1961 was appointe…
1   The appellant before us was examined as prime …
2   This appeal arises from the judgment of the le…
3   The Petitioner was married to the Respondent N…
4   This appeal is preferred against the judgment …
5   On 19.3.1999, SI P1 along Ct. P2 went to Villa…
6   This criminal appeal is directed against the j…
7   This appeal, by special leave, has been prefer…
8   The complainant P1 filed a Special Leave Petit…
9   The four appellants, along with P1 son of P2, …
10  The detenu P1, a French national, at the relev…
11  The petitioner has been under detention pursua…
12  This is an appeal with a certificate granted b…
13  P1 is before us being aggrieved by and dissati…
```

```
14    The appellants are five in number and they hav…
15    The appellant P1 is convicted by the Additiona…
16    facts of the matter, as is evident from the pr…
17    These appeals involve a pure question of law a…
18    This appeal is preferred by the appellants aga…
19    This appeal by special leave is directed again…
20    Challenge in this appeal is to the judgment of…
21    Assailing the legal acceptability of the judgm…
22    The petitioner is a firm carrying on business …
23    These appeals are directed against the judgmen…
24    These appeals involving common questions of la…
25    The hearing before us now relates to certain o…
26    Appellant before us was detained. He is the Ma…
27    Challenge in this appeal is to the judgment of…
28    This appeal has been preferred against the jud…
29    That the deceased P1 got married to P2, the 2n…
30    This appeal by special leave is directed again…
31    On 9th May, 2004, the marriage of the daughter…
32    This is an appeal by special leave from the ju…
33    These writ petitions are filed as Public Inter…
34    Two appellants, who are brothers, along with t…
35    Interpretation and/or application of Medical B…
36    Appellants call in question legality of the ju…
37    The appellant herein is a Senior Manager in a …
38    Challenge in this appeal is to the order of a …
39    Having been selected by the Public Service Com…
40    Appellant calls in question legality of the ju…
41    This appeal arises out of the judgment dated 2…
42    Transfer Petition have been filed to transfer …
43    This petition is by the State directed against…
44    The appellants were tried for offences on the …
45    In this appeal by special leave the sole appel…
46    Challenge in this appeal is to the judgment of…
47    Whether sanction is required to initiate crimi…
48    Appellant was a Patwari working at village V1 …
49    A peculiar feature of this appeal by special l…
```

1.6 Cleanse the test data

1.6.1 We use the same methods as above to cleanse the test data

[70]:

```
test['Query_processed'] = test['Query'].apply(lambda x:␣
 ↪utils_preprocess_text(x, flg_stemm = False, flg_lemm=True))
```

[71]:

```
test
```

```
[71]:                                               Query  \
     0   The appellant on February 9, 1961 was appointe…
     1   The appellant before us was examined as prime …
     2   This appeal arises from the judgment of the le…
     3   The Petitioner was married to the Respondent N…
     4   This appeal is preferred against the judgment …
     5   On 19.3.1999, SI P1 along Ct. P2 went to Villa…
     6   This criminal appeal is directed against the j…
     7   This appeal, by special leave, has been prefer…
     8   The complainant P1 filed a Special Leave Petit…
     9   The four appellants, along with P1 son of P2, …
     10  The detenu P1, a French national, at the relev…
     11  The petitioner has been under detention pursua…
     12  This is an appeal with a certificate granted b…
     13  P1 is before us being aggrieved by and dissati…
     14  The appellants are five in number and they hav…
     15  The appellant P1 is convicted by the Additiona…
     16  facts of the matter, as is evident from the pr…
     17  These appeals involve a pure question of law a…
     18  This appeal is preferred by the appellants aga…
     19  This appeal by special leave is directed again…
     20  Challenge in this appeal is to the judgment of…
     21  Assailing the legal acceptability of the judgm…
     22  The petitioner is a firm carrying on business …
     23  These appeals are directed against the judgmen…
     24  These appeals involving common questions of la…
     25  The hearing before us now relates to certain o…
     26  Appellant before us was detained. He is the Ma…
     27  Challenge in this appeal is to the judgment of…
     28  This appeal has been preferred against the jud…
     29  That the deceased P1 got married to P2, the 2n…
     30  This appeal by special leave is directed again…
     31  On 9th May, 2004, the marriage of the daughter…
     32  This is an appeal by special leave from the ju…
     33  These writ petitions are filed as Public Inter…
     34  Two appellants, who are brothers, along with t…
     35  Interpretation and/or application of Medical B…
     36  Appellants call in question legality of the ju…
     37  The appellant herein is a Senior Manager in a …
     38  Challenge in this appeal is to the order of a …
     39  Having been selected by the Public Service Com…
     40  Appellant calls in question legality of the ju…
     41  This appeal arises out of the judgment dated 2…
     42  Transfer Petition have been filed to transfer …
     43  This petition is by the State directed against…
     44  The appellants were tried for offences on the …
     45  In this appeal by special leave the sole appel…
```

```
46  Challenge in this appeal is to the judgment of…
47  Whether sanction is required to initiate crimi…
48  Appellant was a Patwari working at village V1 …
49  A peculiar feature of this appeal by special l…

                                      Query_processed
0   the appellant on february 9 1961 wa appointed …
1   the appellant before u wa examined a prime wit…
2   this appeal arises from the judgment of the le…
3   the petitioner wa married to the respondent no…
4   this appeal is preferred against the judgment …
5   on 1931999 si p1 along ct p2 went to village v…
6   this criminal appeal is directed against the j…
7   this appeal by special leave ha been preferred…
8   the complainant p1 filed a special leave petit…
9   the four appellant along with p1 son of p2 wer…
10  the detenu p1 a french national at the relevan…
11  the petitioner ha been under detention pursuan…
12  this is an appeal with a certificate granted b…
13  p1 is before u being aggrieved by and dissatis…
14  the appellant are five in number and they have…
15  the appellant p1 is convicted by the additiona…
16  fact of the matter a is evident from the prese…
17  these appeal involve a pure question of law a …
18  this appeal is preferred by the appellant agai…
19  this appeal by special leave is directed again…
20  challenge in this appeal is to the judgment of…
21  assailing the legal acceptability of the judgm…
22  the petitioner is a firm carrying on business …
23  these appeal are directed against the judgment…
24  these appeal involving common question of law …
25  the hearing before u now relates to certain ob…
26  appellant before u wa detained he is the manag…
27  challenge in this appeal is to the judgment of…
28  this appeal ha been preferred against the judg…
29  that the deceased p1 got married to p2 the 2nd…
30  this appeal by special leave is directed again…
31  on 9th may 2004 the marriage of the daughter o…
32  this is an appeal by special leave from the ju…
33  these writ petition are filed a public interes…
34  two appellant who are brother along with their…
35  interpretation andor application of medical be…
36  appellant call in question legality of the jud…
37  the appellant herein is a senior manager in a …
38  challenge in this appeal is to the order of a …
39  having been selected by the public service com…
40  appellant call in question legality of the jud…
```

```
41   this appeal arises out of the judgment dated 2…
42   transfer petition have been filed to transfer …
43   this petition is by the state directed against…
44   the appellant were tried for offence on the al…
45   in this appeal by special leave the sole appel…
46   challenge in this appeal is to the judgment of…
47   whether sanction is required to initiate crimi…
48   appellant wa a patwari working at village v1 i…
49   a peculiar feature of this appeal by special l…
```

2 BM-25 ranking

[72]:

```
!pip install rank_bm25
```

```
Requirement already satisfied: rank_bm25 in /opt/conda/lib/python3.7/site-
packages (0.2.1)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(from rank_bm25) (1.19.5)
```

[73]:

```
from rank_bm25 import BM25Okapi

query_array_processed = [0]*50

corpus_array_processed = [0]*2914

train_array=df.iloc[:,1:].values

for i in range(2914):
    corpus_array_processed[i] = train_array[i][0]

query_array=test.iloc[:,1:].values

#test["Query_processed"]
#test.values(columns=[test["Query_processed"]])
#query_array[49][0]

for i in range(50):
    query_array_processed[i] = query_array[i][0]
```

[87]:

```
train_array=df.iloc[:,1:].values
tokenized_corpus = [doc.split(" ") for doc in corpus_array_processed]
```

[75]:

```
bm25 = BM25Okapi(tokenized_corpus)
bm25
```

```
[75]: <rank_bm25.BM25Okapi at 0x7fa74f185610>
```

```
[125]: name = df["Name"]
       name = name.str.rstrip('.txt')
       name
```

```
[125]: 0        C757
       1       C1092
       2       C1985
       3         C39
       4       C2055
               …
       2909     C792
       2910    C2775
       2911     C924
       2912     C432
       2913     C244
       Name: Name, Length: 2914, dtype: object
```

```
[126]: bm25.get_top_n(corpus_array_processed[4].split(" "), name, n=10)
```

```
[126]: ['C2055',
        'C241',
        'C6',
        'C4',
        'C822',
        'C1511',
        'C1096',
        'C63',
        'C1855',
        'C1357']
```

```
[129]: evaluate = evaluate.loc[evaluate['Relevance'] == 1]
       evaluate
```

```
[129]:        Query_Number Document  Relevance
       1192         AILA_Q1      C14          1
       2274         AILA_Q1       C9          1
       3076         AILA_Q2      C27          1
       3676         AILA_Q2      C22          1
       6033         AILA_Q3       C1          1
       …                  …      …         …
       140861      AILA_Q49      C38          1
       142203      AILA_Q49      C76          1
       142450      AILA_Q49      C92          1
       143069      AILA_Q50      C27          1
       143844      AILA_Q50      C22          1
```

```
[195 rows x 3 columns]
```

[174]:

```
# retrieved = bm25.get_top_n(query_array_processed[i].split(" "), name, n=10)
# relevant = evaluate.loc[evaluate['Query_Number'] ==␣
 ↪"AILA_Q"+str(i+1)]["Document"]
```

[175]:

```
count = 0
for i in range(50):
    for j in bm25.get_top_n(query_array_processed[i].split(" "), name, n=10):
        for k in evaluate.loc[evaluate['Query_Number'] ==␣
 ↪"AILA_Q"+str(i+1)]["Document"]:
            if (j==k):
                count=count+1

print(count)
```

```
20
```

[132]:

```
Precision = count/500
Recall = count/195

print(Precision)
print(Recall)
```

```
0.04
0.10256410256410256
```

3 Doc2Vec

[136]:

```
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from nltk.tokenize import word_tokenize
```

[137]:

```
#corpus_array_processed
tagged_data = [TaggedDocument(words=word_tokenize(_d.lower()), tags=[str(i)])␣
 ↪for i, _d in enumerate(corpus_array_processed)]
```

[141]:

```
model = Doc2Vec(tagged_data, vector_size=20, window=2, min_count=1, workers=4,␣
 ↪epochs = 100)
```

[142]:

```
model.save("test_doc2vec.model")
```

[143]:

```
model= Doc2Vec.load("test_doc2vec.model")
```

[183]:

```
count = 0
for i in range(50):
    for j in model.docvecs.most_similar(positive=[model.
 ↪infer_vector(word_tokenize(query_array_processed[i]))],topn=10)[0][0]:
        temp = evaluate.loc[evaluate['Query_Number'] ==␣
 ↪"AILA_Q"+str(i+1)]["Document"]
        for k in temp.str.replace('C', ''):
            if (j==k):
                count=count+1

print(count)
```

```
/opt/conda/lib/python3.7/site-packages/ipykernel_launcher.py:3:
DeprecationWarning: Call to deprecated `docvecs` (The `docvecs` property has
been renamed `dv`.).
  This is separate from the ipykernel package so we can avoid doing imports
until
```

```
6
```

[184]:

```
Precision = count/500
Recall = count/195

print(Precision)
print(Recall)
```

```
0.012
0.03076923076923077
```