Text Data Pre-Processing

August 1, 2024

importing the dependencies.

-import re regular expression library, for scanning and going through some texts in the documents

import numpy as np #for makingarrays import pandas as pd #for making data frames and data frames are structured table

-from nltk.corpus import stopwords -corpus –> it mean some text contents e.g paragraph, documents. -corpus –> is basilally a collection of words. -nltk –. natural tool kit; it contains several functions and methods that we use for text processing.

importing means to import the whole library from means to import specific functions we need from a library

NOTES

stopwords are words that can be repeated alot of times in a paragraph of a document. It dosent covers much meaning.

streaming; is the processof reducing a word to its rootword.

[1]:
```
import numpy as np
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
```

[2]:
```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\DONATUS\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[2]:
```
True
```

[3]:
```
#printing the  stopwards
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
```

```
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
"wouldn't"]
```

Data Pre-Processing Steps

[4]:
```
#Loading the dataset to pandas DataFrame
news_data = pd.read_csv('fake_news_dataset.csv')
```

[5]:
```
#first 5 rows of the dataset
news_data.head()
```

[5]:

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let…	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let…	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - …	Daniel J. Flynn	Ever get the feeling your life circles the rou…	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, …	1
3	3	15 Civilians Killed In Single US Airstrike Hav…	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr…	1
4	4	Iranian woman jailed for fictional unpublished…	Howard Portnoy	Print \nAn Iranian woman has been sentenced to…	1

0 –> real news 1 –> fake news

[25]:
```
#to know the total data points

news_data.shape
```

[25]:
```
(20800, 6)
```

[7]:
```
#checking for missing values
news_data.isnull().sum()
```

[7]:
```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

[8]:
```
#replacing missing values with null string.  na = not available
news_data = news_data.fillna('')
```

[9]:
```
#checking for missing values
news_data.isnull().sum()
```

[9]:
```
id        0
title     0
author    0
text      0
label     0
dtype: int64
```

[10]:
```
#merging the authur name and news title
news_data['content'] = news_data['author']+ ' ' +news_data['title']
```

[11]:
```
news_data.head()
```

[11]:

	id	title	author	text	label	content
0	0	House Dem Aide: We Didn’t Even See Comey’s Let…	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let…	1	Darrell Lucus House Dem Aide: We Didn’t Even S…
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - …	Daniel J. Flynn	Ever get the feeling your life circles the rou…	0	Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo…
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, …	1	Consortiumnews.com Why the Truth Might Get You…
3	3	15 Civilians Killed In Single US Airstrike Hav…	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr…	1	Jessica Purkiss 15 Civilians Killed In Single …
4	4	Iranian woman jailed for fictional unpublished…	Howard Portnoy	Print \nAn Iranian woman has been sentenced to…	1	Howard Portnoy Iranian woman jailed for fictio…

[12]:
```
#seperating fetures from target.
x = news_data.drop(columns='label', axis = 1)
y = news_data['label']
```

[13]:
```
print(x)
```

```
          id                                                 title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let…
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - …
2          2                  Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav…
4          4  Iranian woman jailed for fictional unpublished…
…        …                                                 …
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S…
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -…
20797  20797  Macy’s Is Said to Receive Takeover Approach by…
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal…
20799  20799                          What Keeps the F-35 Alive

                                       author  \
0                                Darrell Lucus
1                              Daniel J. Flynn
2                           Consortiumnews.com
3                              Jessica Purkiss
4                               Howard Portnoy
…                                          …
20795                            Jerome Hudson
20796                          Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                              Alex Ansary
20799                            David Swanson

                                                    text  \
0      House Dem Aide: We Didn’t Even See Comey’s Let…
1      Ever get the feeling your life circles the rou…
2      Why the Truth Might Get You Fired October 29, …
3      Videos 15 Civilians Killed In Single US Airstr…
4      Print \nAn Iranian woman has been sentenced to…
…                                                    …
20795  Rapper T. I. unloaded on black celebrities who…
20796  When the Green Bay Packers lost to the Washing…
20797  The Macy’s of today grew from the union of sev…
20798  NATO, Russia To Hold Parallel Exercises In Bal…
20799    David Swanson is an author, activist, journa…

                                                 content
0      Darrell Lucus House Dem Aide: We Didn’t Even S…
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo…
```

```
2      Consortiumnews.com Why the Truth Might Get You…
3      Jessica Purkiss 15 Civilians Killed In Single …
4      Howard Portnoy Iranian woman jailed for fictio…
…                                                    …
20795  Jerome Hudson Rapper T.I.: Trump a ’Poster Chi…
20796  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma…
20797  Michael J. de la Merced and Rachel Abrams Macy…
20798  Alex Ansary NATO, Russia To Hold Parallel Exer…
20799            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
```

[14]:
```
print(y)
```

```
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

[15]:
```
#streaming using PorterStemmer
port_stem = PorterStemmer()
```

[16]:
```
#creating a function to do the steming procedure

def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not␣
 ↪word in stopwords.words('english')]
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content
```

[17]:
```
news_data['content'] = news_data['content'].apply(stemming)
```

[18]:
```
print(news_data['content'])
```

```
0        darrel lucu hous dem aid even see comey letter…
1        daniel j flynn flynn hillari clinton big woman…
2                   consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri…
```

```
4        howard portnoy iranian woman jail fiction unpu…
                                 …
20795    jerom hudson rapper trump poster child white s…
20796    benjamin hoffman n f l playoff schedul matchup…
20797    michael j de la merc rachel abram maci said re…
20798    alex ansari nato russia hold parallel exercis …
20799                              david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

[ ]:

[19]:
```
#seperating the data to its coressponing features and target
x = news_data['content'].values
y = news_data['label'].values
```

[20]:
```
print(x)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' …
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay
new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

[21]:
```
print(y)
```

```
[1 0 1 … 0 1 1]
```

[22]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

# Assuming your original text data is stored in a variable called␣
 ↪'original_text_data'
vectorizer = TfidfVectorizer()
#vectorizer.fit(x)

X = vectorizer.fit_transform(news_data)
```

[23]:
```
print(x)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' …
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay
new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

[ ]:

[ ]: