Dharavath Ramdas

source code link:https: //github.com/dharavathramdas101/Machine-learning-Algorithms-Practical-implementation Notes link: https://www.linkedin.com/in/dharavath-ramdas-a283aa213/ (https://www.linkedin.com/in/dharavath-ramdas-a283aa213/)

NLP (Natural Language Processing) ¶

1.Text Preprocessing:
 1.Tokenization
 2.Text Preprocessing
 3.Stemming
 4.Limmatization
 5.StopWords
2.Words to vectors
 1.Bag of Words
 2.TF-IDF
 3.Word2Vector

1.Text Preprocessing:

In [2]:

```
paragraph ="""Narendra Damodardas Modi (Gujarati: ['nəɾendɾə dɑmodəɾ'dɑs 'modiː] (listen); born 17 September 1950)[a] is an Indian politic

Born and raised in Vadnagar, a small town in northeastern Gujarat, Modi completed his secondary education there. He was introduced to the
"""
```

In [3]:

```
paragraph
```

Out[3]:

```
"Narendra Damodardas Modi (Gujarati: ['nəɾendɾə dɑmodəɾ'dɑs 'modiː] (listen); born 17 September 1950)[a] is an Indian polit
ician serving as the 14th and current prime minister of India since 2014. Modi was the chief minister of Gujarat from 2001
to 2014 and is the Member of Parliament from Varanasi. He is a member of the Bharatiya Janata Party (BJP) and of the Rashtr
iya Swayamsevak Sangh (RSS), a right-wing Hindu nationalist paramilitary volunteer organisation. He is the first prime mini
ster to have been born after India's independence in 1947 and the second prime minister not belonging to the Indian Nationa
l Congress to have won two consecutive majorities in the Lok Sabha, or the lower house of India's parliament. He is also th
e longest serving prime minister from a non-Congress party.\n\nBorn and raised in Vadnagar, a small town in northeastern Gu
jarat, Modi completed his secondary education there. He was introduced to the RSS at age eight. He has drawn attention to h
aving to work as a child in his father's tea stall on the Vadnagar railway station platform, a description that has not bee
n reliably corroborated. At age 18, Modi was married to Jashodaben Chimanlal Modi, whom he abandoned soon after. He left hi
s parental home where she had come to live. He first publicly acknowledged her as his wife more than four decades later whe
n required to do so by Indian law, but has made no contact with her since. Modi has asserted he had travelled in northern I
ndia for two years after leaving his parental home, visiting a number of religious centres, but few details of his travels
have emerged. Upon his return to Gujarat in 1971, he became a full-time worker for the RSS. After the state of emergency wa
s declared by prime minister Indira Gandhi in 1975, Modi went into hiding. The RSS assigned him to the BJP in 1985 and he h
eld several positions within the party hierarchy until 2001, rising to the rank of general secretary.[b]\n"
```

In [4]:

```
import nltk
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
```

punkt

tokanization convert paragraph sentence to words
tokenize. punkt module. Punkt Sentence Tokenizer. This tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.

2.Tokanization:

In [6]:

```
nltk.download('punkt')
sentences = nltk.sent_tokenize(paragraph)
```

```
[nltk_data] Error loading punkt: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
```

In [7]:

```
print(sentences)
```

```
['Narendra Damodardas Modi (Gujarati: ['nərendrə dɑmodərˈdɑs ˈmodiː] (listen); born 17 September 1950)[a] is an Indian poli
tician serving as the 14th and current prime minister of India since 2014.', 'Modi was the chief minister of Gujarat from 2
001 to 2014 and is the Member of Parliament from Varanasi.', 'He is a member of the Bharatiya Janata Party (BJP) and of the
Rashtriya Swayamsevak Sangh (RSS), a right-wing Hindu nationalist paramilitary volunteer organisation.', "He is the first p
rime minister to have been born after India's independence in 1947 and the second prime minister not belonging to the India
n National Congress to have won two consecutive majorities in the Lok Sabha, or the lower house of India's parliament.", 'H
e is also the longest serving prime minister from a non-Congress party.', 'Born and raised in Vadnagar, a small town in nor
theastern Gujarat, Modi completed his secondary education there.', 'He was introduced to the RSS at age eight.', "He has dr
awn attention to having to work as a child in his father's tea stall on the Vadnagar railway station platform, a descriptio
n that has not been reliably corroborated.", 'At age 18, Modi was married to Jashodaben Chimanlal Modi, whom he abandoned s
oon after.', 'He left his parental home where she had come to live.', 'He first publicly acknowledged her as his wife more
than four decades later when required to do so by Indian law, but has made no contact with her since.', 'Modi has asserted
he had travelled in northern India for two years after leaving his parental home, visiting a number of religious centres, b
ut few details of his travels have emerged.', 'Upon his return to Gujarat in 1971, he became a full-time worker for the RS
S.', 'After the state of emergency was declared by prime minister Indira Gandhi in 1975, Modi went into hiding.', 'The RSS
assigned him to the BJP in 1985 and he held several positions within the party hierarchy until 2001, rising to the rank of
general secretary.', '[b]']
```

3.Stemming:

In [8]:

```
stemmer = PorterStemmer()
```

In [9]:

```
stemmer.stem('History')
```

Out[9]:

```
'histori'
```

In [10]:

```
stemmer.stem('ramdas')
```

Out[10]:

```
'ramda'
```

4.Lemmatization:

In [11]:

```
from nltk.stem import WordNetLemmatizer
```

In [12]:

```
lemmatizer = WordNetLemmatizer()
```

In [13]:

```
lemmatizer.lemmatize("ramdas")
```

Out[13]:

```
'ramdas'
```

In [17]:

```
lemmatizer.lemmatize("mangos")
```

Out[17]:

```
'mango'
```

In [18]:

```
len(sentences)
```

Out[18]:

```
16
```

import regular expression

In [20]:

```
import re
```

In [21]:

```
corpus=[]
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]',' ', sentences[i])
    review =review.lower()
    corpus.append(review)
```

In [22]:

```
corpus
```

Out[22]:

```
['narendra damodardas modi  gujarati    n  end   d mod   d s  modi    listen   born    september       a  is an indian poli
tician serving as the   th and current prime minister of india since      ',
 'modi was the chief minister of gujarat from      to      and is the member of parliament from varanasi ',
 'he is a member of the bharatiya janata party  bjp  and of the rashtriya swayamsevak sangh  rss   a right wing hindu natio
nalist paramilitary volunteer organisation ',
 'he is the first prime minister to have been born after india s independence in      and the second prime minister not bel
onging to the indian national congress to have won two consecutive majorities in the lok sabha  or the lower house of india
s parliament ',
 'he is also the longest serving prime minister from a non congress party ',
 'born and raised in vadnagar  a small town in northeastern gujarat  modi completed his secondary education there ',
 'he was introduced to the rss at age eight ',
 'he has drawn attention to having to work as a child in his father s tea stall on the vadnagar railway station platform  a
description that has not been reliably corroborated ',
 'at age     modi was married to jashodaben chimanlal modi  whom he abandoned soon after ',
 'he left his parental home where she had come to live ',
 'he first publicly acknowledged her as his wife more than four decades later when required to do so by indian law  but has
made no contact with her since ',
 'modi has asserted he had travelled in northern india for two years after leaving his parental home  visiting a number of
religious centres  but few details of his travels have emerged ',
 'upon his return to gujarat in       he became a full time worker for the rss ',
 'after the state of emergency was declared by prime minister indira gandhi in       modi went into hiding ',
 'the rss assigned him to the bjp in      and he held several positions within the party hierarchy until       rising to th
e rank of general secretary ',
 ' b ']
```

5.StopWords:

In [23]:

```
stopwords.words('english')
```

```
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
 'theirs',
 'themselves',
 'what',
 'which',
 'who',
 'whom',
 'this',
 'that',
```

In [24]:

```
# Stemming
```

In [26]:

```
for i in corpus:
    words = nltk.word_tokenize(i)
    for word in words:
        if word not in set(stopwords.words('english')):
            print(stemmer.stem(word))
```

```
narendra
damodarda
modi
gujarati
n
end
mod
modi
listen
born
septemb
indian
politician
serv
th
current
prime
minist
india
```

In [27]:

```
# Lemmatization
```

In [28]:

```
for i in corpus:
    words = nltk.word_tokenize(i)
    for word in words:
        if word not in set(stopwords.words('english')):
            print(lemmatizer.lemmatize(word))
```

```
narendra
damodardas
modi
gujarati
n
end
mod
modi
listen
born
september
indian
politician
serving
th
current
prime
minister
india
```

Apply Stopwords, lemmatize

In [30]:

```
import re
```

In [32]:

```
corpus = []
for i in range(len(sentences)):
    review = re.sub('[^a-zA-Z]'," ",sentences[i])
    review = review.lower()
    review =   review.split()
    review = [lemmatizer.lemmatize(word) for word in review if not word in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus.append(review)
```

Word to vectors:

In [34]:

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(binary=True)
```

In [35]:

```
X=cv.fit_transform(corpus)
```

In [37]:

```
cv.vocabulary_
```

Out[37]:

```
{'narendra': 71,
 'damodardas': 23,
 'modi': 70,
 'gujarati': 41,
 'end': 33,
 'mod': 69,
 'listen': 59,
 'born': 11,
 'september': 103,
 'indian': 50,
 'politician': 84,
 'serving': 104,
 'th': 114,
 'current': 22,
 'prime': 86,
 'minister': 68,
 'india': 49,
 'since': 106,
```

In [38]:

```
corpus[0]
```

Out[38]:

```
'narendra damodardas modi gujarati n end mod modi listen born september indian politician serving th current prime minister
india since'
```

In [39]:

```
corpus[1]
```

Out[39]:

```
'modi chief minister gujarat member parliament varanasi'
```

In [40]:

```
X[0]
```

Out[40]:

```
<1x132 sparse matrix of type '<class 'numpy.int64'>'
        with 18 stored elements in Compressed Sparse Row format>
```

In [41]:

```
X[0].toarray()
```

Out[41]:

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
        0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
        0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0,
        0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]],
      dtype=int64)
```

TF IDF:

In [45]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [46]:

```
cv = TfidfVectorizer(ngram_range=(1,1),max_features=10)
```

In [47]:

```
X = cv.fit_transform(corpus)
```

In [48]:

```
corpus[0]
```

Out[48]:

```
'narendra damodardas modi gujarati n end mod modi listen born september indian politician serving th current prime minister
india since'
```

In [49]:

```
corpus[1]
```

Out[49]:

```
'modi chief minister gujarat member parliament varanasi'
```

In [50]:

```
X[0].toarray()
```

Out[50]:

```
array([[0.38064993, 0.        , 0.38064993, 0.38064993, 0.        ,
        0.31757459, 0.5871889 , 0.        , 0.        , 0.34593707]])
```

In [ ]: