nlp

October 27, 2024

[1]:
```
pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (3.8.1)
Requirement already satisfied: click in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (from nltk) (2023.10.3)
Requirement already satisfied: tqdm in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (from nltk) (4.66.4)
Requirement already satisfied: colorama in c:\users\sanket
upadhyay\anaconda3\lib\site-packages (from click->nltk) (0.4.6)
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
paragraph ="""Narendra Damodardas Modi[a] (born 17 September 1950)[b] is an␣
↪Indian politician serving as the current prime minister of India since 26␣
↪May 2014. Modi was the chief minister of Gujarat from 2001 to 2014 and is␣
↪the member of parliament (MP) for 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␣
↪longest-serving prime minister outside the Indian National Congress.[4]
```

```
Modi was born and raised in Vadnagar in northeastern Gujarat, where he␣
 ↪completed his secondary education. He was introduced to the RSS at the age␣
 ↪of eight. At the age of 18, he was married to Jashodaben Modi, whom he␣
 ↪abandoned soon after, only publicly acknowledging her four decades later␣
 ↪when legally required to do so. Modi became a full-time worker for the RSS␣
 ↪in Gujarat in 1971. The RSS assigned him to the BJP in 1985 and he rose␣
 ↪through the party hierarchy, becoming general secretary in 1998.[c] In 2001,␣
 ↪Modi was appointed Chief Minister of Gujarat and elected to the legislative␣
 ↪assembly soon after. His administration is considered complicit in the 2002␣
 ↪Gujarat riots,[d] and has been criticised for its management of the crisis.␣
 ↪According to official records, a little over 1,000 people were killed,␣
 ↪three-quarters of whom were Muslim; independent sources estimated 2,000␣
 ↪deaths, mostly Muslim.[13] A Special Investigation Team appointed by the␣
 ↪Supreme Court of India in 2012 found no evidence to initiate prosecution␣
 ↪proceedings against him.[e] While his policies as chief minister were␣
 ↪credited for encouraging economic growth, his administration was criticised␣
 ↪for failing to significantly improve health, poverty and education indices␣
 ↪in the state.[f]"""
```

[3]:
```
paragraph
```

[3]:
```
'Narendra Damodardas Modi[a] (born 17 September 1950)[b] is an Indian politician
serving as the current prime minister of India since 26 May 2014. Modi was the
chief minister of Gujarat from 2001 to 2014 and is the member of parliament (MP)
for 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 longest-serving prime minister outside the
Indian National Congress.[4]\n\nModi was born and raised in Vadnagar in
northeastern Gujarat, where he completed his secondary education. He was
introduced to the RSS at the age of eight. At the age of 18, he was married to
Jashodaben Modi, whom he abandoned soon after, only publicly acknowledging her
four decades later when legally required to do so. Modi became a full-time
worker for the RSS in Gujarat in 1971. The RSS assigned him to the BJP in 1985
and he rose through the party hierarchy, becoming general secretary in 1998.[c]
In 2001, Modi was appointed Chief Minister of Gujarat and elected to the
legislative assembly soon after. His administration is considered complicit in
the 2002 Gujarat riots,[d] and has been criticised for its management of the
crisis. According to official records, a little over 1,000 people were killed,
three-quarters of whom were Muslim; independent sources estimated 2,000 deaths,
mostly Muslim.[13] A Special Investigation Team appointed by the Supreme Court
of India in 2012 found no evidence to initiate prosecution proceedings against
him.[e] While his policies as chief minister were credited for encouraging
economic growth, his administration was criticised for failing to significantly
improve health, poverty and education indices in the state.[f]'
```

[11]:
```
import nltk
from nltk.stem import PorterStemmer
```

```
from nltk.corpus import stopwords
```

[12]:
```
## tokenization convert paragraph and sentences into woords
nltk.download('punkt')
sentences=nltk.sent_tokenize(paragraph)
```

```
[nltk_data] Downloading package punkt to C:\Users\SANKET
[nltk_data]     UPADHYAY\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

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

```
['Narendra Damodardas Modi[a] (born 17 September 1950)[b] is an Indian
politician serving as the current prime minister of India since 26 May 2014.',
'Modi was the chief minister of Gujarat from 2001 to 2014 and is the member of
parliament (MP) for 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 longest-serving
prime minister outside the Indian National Congress.', '[4]\n\nModi was born and
raised in Vadnagar in northeastern Gujarat, where he completed his secondary
education.', 'He was introduced to the RSS at the age of eight.', 'At the age of
18, he was married to Jashodaben Modi, whom he abandoned soon after, only
publicly acknowledging her four decades later when legally required to do so.',
'Modi became a full-time worker for the RSS in Gujarat in 1971.', 'The RSS
assigned him to the BJP in 1985 and he rose through the party hierarchy,
becoming general secretary in 1998.', '[c] In 2001, Modi was appointed Chief
Minister of Gujarat and elected to the legislative assembly soon after.', 'His
administration is considered complicit in the 2002 Gujarat riots,[d] and has
been criticised for its management of the crisis.', 'According to official
records, a little over 1,000 people were killed, three-quarters of whom were
Muslim; independent sources estimated 2,000 deaths, mostly Muslim.', '[13] A
Special Investigation Team appointed by the Supreme Court of India in 2012 found
no evidence to initiate prosecution proceedings against him.', '[e] While his
policies as chief minister were credited for encouraging economic growth, his
administration was criticised for failing to significantly improve health,
poverty and education indices in the state.', '[f]']
```

[17]:
```
stemmer=PorterStemmer()
```

[18]:
```
stemmer.stem('drinking')
```

[18]:
```
'drink'
```

[19]:
```
stemmer.stem('history')
```

[19]:
```
'histori'
```

[27]:
```
from nltk.stem import WordNetLemmatizer
```

[28]:
```
lemmatizer=WordNetLemmatizer()
```

[32]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to C:\Users\SANKET
[nltk_data]     UPADHYAY\AppData\Roaming\nltk_data…
```

[32]:
```
True
```

[35]:
```
from nltk.stem import WordNetLemmatizer

# Initialize the lemmatizer
lemmatizer = WordNetLemmatizer()

# Lemmatize the word
print(lemmatizer.lemmatize('sleeping'))
```

```
sleeping
```

[39]:
```
import re
corpus=[]
for i in range(len(sentences)):
    review=re.sub('[^a-zA-Z]','',sentences[i])
    review=review.lower()
    corpus.append(review)
```

[40]:
```
corpus
```

[40]:
```
['narendradamodardasmodiabornseptemberbisanindianpoliticianservingasthecurrentpr
imeministerofindiasincemay',
'modiwasthechiefministerofgujaratfromtoandisthememberofparliamentmpforvaranasi',
 'heisamemberofthebharatiyajanatapartybjpandoftherashtriyaswayamsevaksanghrssari
ghtwinghindunationalistparamilitaryvolunteerorganisation',
 'heisthelongestservingprimeministeroutsidetheindiannationalcongress',
 'modiwasbornandraisedinvadnagarinnortheasterngujaratwherehecompletedhissecondar
yeducation',
 'hewasintroducedtotherssattheageofeight',
 'attheageofhewasmarriedtojashodabenmodiwhomheabandonedsoonafteronlypubliclyackn
owledgingherfourdecadeslaterwhenlegallyrequiredtodoso',
 'modibecameafulltimeworkerfortherssingujaratin',
 'therssassignedhimtothebjpinandherosethroughthepartyhierarchybecominggeneralsec
retaryin',
 'cinmodiwasappointedchiefministerofgujaratandelectedtothelegislativeassemblysoo
nafter',
```

```
 'hisadministrationisconsideredcomplicitinthegujaratriotsdandhasbeencriticisedfo
ritsmanagementofthecrisis',
 'accordingtoofficialrecordsalittleoverpeoplewerekilledthreequartersofwhomweremu
slimindependentsourcesestimateddeathsmostlymuslim',
 'aspecialinvestigationteamappointedbythesupremecourtofindiainfoundnoevidencetoi
nitiateprosecutionproceedingsagainsthim',
 'ewhilehispoliciesaschiefministerwerecreditedforencouragingeconomicgrowthhisadm
inistrationwascriticisedforfailingtosignificantlyimprovehealthpovertyandeducatio
nindicesinthestate',
 'f']
```

[48]:
```
stopwords.words('english')
```

[48]:
```
['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"]
```

[49]:
```
from nltk.tokenize import word_tokenize

# Download required NLTK data
nltk.download('stopwords')
nltk.download('punkt')

# Initialize the stemmer
stemmer = PorterStemmer()

# Example corpus
corpus = ["This is an example sentence.", "Stemming words in NLTK is helpful."]

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

```
exampl
sentenc
.
stem
word
nltk
help
.
[nltk_data] Downloading package stopwords to C:\Users\SANKET
[nltk_data]     UPADHYAY\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to C:\Users\SANKET
[nltk_data]     UPADHYAY\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
```

[53]:
```
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))
```

```
This
```

```
example
sentence
.
Stemming
word
NLTK
helpful
.
```

[54]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer()
```

[55]:
```
X=cv.fit_transform(corpus)
```

[56]:
```
cv.vocabulary_
```

[56]:
```
{'this': 8,
 'is': 4,
 'an': 0,
 'example': 1,
 'sentence': 6,
 'stemming': 7,
 'words': 9,
 'in': 3,
 'nltk': 5,
 'helpful': 2}
```

[57]:
```
corpus[0]
```

[57]:
```
'This is an example sentence.'
```

[58]:
```
X[0].toarray()
```

[58]:
```
array([[1, 1, 0, 0, 1, 0, 1, 0, 1, 0]], dtype=int64)
```

[ ]:
```

```