text-preprocessing-amazon-alexa

April 13, 2024

Text Preproceessing

[2]:
```
import pandas as pd
import nltk
import re
import numpy as np
```

[12]:
```
df = pd.read_csv('amazon_alexa.tsv', sep='\t')
df
```

[12]:

	rating	date	variation	verified_reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer…	1
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 …	1
4	5	31-Jul-18	Charcoal Fabric	Music	1
…	…	…	…	…	…
3145	5	30-Jul-18	Black Dot	Perfect for kids, adults and everyone in betwe…	1
3146	5	30-Jul-18	Black Dot	Listening to music, searching locations, check…	1
3147	5	30-Jul-18	Black Dot	I do love these things, i have them running my…	1
3148	5	30-Jul-18	White Dot	Only complaint I have is that the sound qualit…	1
3149	4	29-Jul-18	Black Dot	Good	1

[3150 rows x 5 columns]

[13]:
```
df.sample(5)
```

[13]:
```
      rating       date                     variation  \
2044       5  14-Jul-18                   Black  Plus
2178       5  30-Jul-18  Configuration: Fire TV Stick
1015       4  29-Jul-18             Sandstone Fabric
1812       4  29-Jul-18                   Black  Plus
2136       5  31-Jul-18  Configuration: Fire TV Stick

                                      verified_reviews  feedback
2044         Love, Love, Love my Amazon Echo Plus!!           1
2178  Works like a charm. Fast delivery and no probl…         1
1015  The sound quality is good just wish alexa coul…         1
1812  Great product I just wish it had some sort of …         1
2136  We love the Fire TV Stick and will love it eve…         1
```

[15]:
```
df['verified_reviews'][3]
```

[15]:
```
'I have had a lot of fun with this thing. My 4 yr old learns about dinosaurs, i
control the lights and play games like categories. Has nice sound when playing
music as well.'
```

Lowercasing

[17]:
```
df['verified_reviews'] = df['verified_reviews'].str.lower()
```

[8]:
```
df
```

[8]:
```
                                                review sentiment
0    one of the other reviewers has mentioned that …  positive
1    a wonderful little production. <br /><br />the…  positive
2    i thought this was a wonderful way to spend ti…  positive
3    basically there's a family where a little boy …  negative
4    petter mattei's "love in the time of money" is…  positive
..                                                 …         …
195  phantasm …class. phantasm ii…awesome. p…  negative
196  ludicrous. angelic 9-year-old annakin turns in…  negative
197  scotty (grant cramer, who would go on to star …  negative
198  if you keep rigid historical perspective out o…  positive
199  the film quickly gets to a major chase scene w…  negative

[200 rows x 2 columns]
```

Remove HTML Tage

[19]:
```
def remove_html(text):
    pattern = re.compile('<.*?>')
    return pattern.sub(r'',text)
```

[20]:
```
df['verified_reviews'] = df['verified_reviews'].apply(remove_html)
```

[21]:
```
df
```

[21]:
```
      rating       date         variation  \
0          5  31-Jul-18  Charcoal Fabric
1          5  31-Jul-18  Charcoal Fabric
2          4  31-Jul-18    Walnut Finish
3          5  31-Jul-18  Charcoal Fabric
4          5  31-Jul-18  Charcoal Fabric
…        …        …                …
3145       5  30-Jul-18        Black  Dot
3146       5  30-Jul-18        Black  Dot
3147       5  30-Jul-18        Black  Dot
3148       5  30-Jul-18        White  Dot
3149       4  29-Jul-18        Black  Dot

                                     verified_reviews  feedback
0                                        love my echo!         1
1                                            loved it!         1
2     sometimes while playing a game, you can answer…         1
3     i have had a lot of fun with this thing. my 4 …         1
4                                               music         1
…                                                 …       …
3145  perfect for kids, adults and everyone in betwe…         1
3146  listening to music, searching locations, check…         1
3147  i do love these things, i have them running my…         1
3148  only complaint i have is that the sound qualit…         1
3149                                             good         1

[3150 rows x 5 columns]
```

[22]:
```
df['verified_reviews'][3]
```

[22]:
```
'i have had a lot of fun with this thing. my 4 yr old learns about dinosaurs, i
control the lights and play games like categories. has nice sound when playing
music as well.'
```

Remove Urls

[23]:
```
def remove_url(text):
    pattern = re.compile(r'https?://(?:[-\w.]|(?:%[\da-fA-F]{2}))+')
    return pattern.sub(r'',text)
```

[24]:
```
df['verified_reviews'] = df['verified_reviews'].apply(remove_url)
```

[25]:
```
df
```

[25]:
```
      rating       date         variation  \
0          5  31-Jul-18  Charcoal Fabric
1          5  31-Jul-18  Charcoal Fabric
2          4  31-Jul-18    Walnut Finish
3          5  31-Jul-18  Charcoal Fabric
4          5  31-Jul-18  Charcoal Fabric
…        …        …                …
3145       5  30-Jul-18         Black  Dot
3146       5  30-Jul-18         Black  Dot
3147       5  30-Jul-18         Black  Dot
3148       5  30-Jul-18         White  Dot
3149       4  29-Jul-18         Black  Dot

                                      verified_reviews  feedback
0                                        love my echo!         1
1                                            loved it!         1
2     sometimes while playing a game, you can answer…         1
3     i have had a lot of fun with this thing. my 4 …         1
4                                                music         1
…                                                …       …
3145  perfect for kids, adults and everyone in betwe…         1
3146  listening to music, searching locations, check…         1
3147  i do love these things, i have them running my…         1
3148  only complaint i have is that the sound qualit…         1
3149                                             good         1

[3150 rows x 5 columns]
```

Remove Punctuation

[26]:
```
import string
exclude = string.punctuation
exclude
```

[26]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

[27]:
```
def remove_punctuation(text):
    for char in exclude:
        text = text.replace(char,'')
    return text
```

[28]:
```
df['verified_reviews'] = df['verified_reviews'].apply(remove_punctuation)
df
```

[28]:
```
      rating       date         variation  \
0          5  31-Jul-18  Charcoal Fabric
1          5  31-Jul-18  Charcoal Fabric
2          4  31-Jul-18    Walnut Finish
```

```
3          5  31-Jul-18  Charcoal Fabric
4          5  31-Jul-18  Charcoal Fabric
…      …        …                 …
3145       5  30-Jul-18        Black  Dot
3146       5  30-Jul-18        Black  Dot
3147       5  30-Jul-18        Black  Dot
3148       5  30-Jul-18        White  Dot
3149       4  29-Jul-18        Black  Dot

                                      verified_reviews  feedback
0                                         love my echo         1
1                                             loved it         1
2     sometimes while playing a game you can answer …         1
3     i have had a lot of fun with this thing my 4 y…         1
4                                                music         1
…                                                …       …
3145    perfect for kids adults and everyone in between         1
3146  listening to music searching locations checkin…         1
3147  i do love these things i have them running my …         1
3148  only complaint i have is that the sound qualit…         1
3149                                              good         1

[3150 rows x 5 columns]
```

[29]:
```
df['verified_reviews'][3]
```

[29]:
```
'i have had a lot of fun with this thing my 4 yr old learns about dinosaurs i
control the lights and play games like categories has nice sound when playing
music as well'
```

Spelling Correction

[30]:
```
from textblob import TextBlob
def spell_correction(text):
    text1 = TextBlob(text)
    correct_text = text1.correct()
    return str(correct_text)
```

[31]:
```
spell_correction(df['verified_reviews'][3])
df['verified_reviews'][3]
```

[31]:
```
'i have had a lot of fun with this thing my 4 yr old learns about dinosaurs i
control the lights and play games like categories has nice sound when playing
music as well'
```

[32]:
```
text2 = 'hii i an verry goodd at artting'
spell_correction(text2)
```

[32]:
```
'his i an very good at sitting'
```

Removal of StopWords

[33]:
```
from nltk.corpus import stopwords
```

[34]:
```
def remove_stopwords(text):
    stop_words = set(stopwords.words('english'))
    words = text.split()
    filtered_words = [word for word in words if word.lower() not in stop_words]
    filtered_text = ' '.join(filtered_words)
    return filtered_text
```

[35]:
```
df['verified_reviews'] = df['verified_reviews'].apply(remove_stopwords)
```

[36]:
```
df
```

[36]:

	rating	date	variation	verified_reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	love echo	1
1	5	31-Jul-18	Charcoal Fabric	loved	1
2	4	31-Jul-18	Walnut Finish	sometimes playing game answer question correct…	1
3	5	31-Jul-18	Charcoal Fabric	lot fun thing 4 yr old learns dinosaurs contro…	1
4	5	31-Jul-18	Charcoal Fabric	music	1
…	…	…	…	…	…
3145	5	30-Jul-18	Black Dot	perfect kids adults everyone	1
3146	5	30-Jul-18	Black Dot	listening music searching locations checking t…	1
3147	5	30-Jul-18	Black Dot	love things running entire home tv lights ther…	1
3148	5	30-Jul-18	White Dot	complaint sound quality isnt great mostly use …	1
3149	4	29-Jul-18	Black Dot	good	1

[3150 rows x 5 columns]

[37]:
```
df['verified_reviews'][3]
```

[37]:
```
'lot fun thing 4 yr old learns dinosaurs control lights play games like
categories nice sound playing music well'
```

Handling Emojis

[38]:
```
import emoji
text = 'programming in python is   '
print(emoji.demojize(text))
```

```
programming in python is :fire::fire::fire::fire:
```

[40]:
```
df['verified_reviews'][2044]
```

[40]:
```
'love love love amazon echo plus '
```

[41]:
```
def emojies(text):
    return emoji.demojize(text)
```

[42]:
```
df['verified_reviews'] = df['verified_reviews'].apply(emojies)
```

[43]:
```
df['verified_reviews'][2044]
```

[43]:
```
'love love love amazon echo plus:red_heart::red_heart:'
```

Tokenization

[46]:
```
from nltk.tokenize import word_tokenize
def tokenize_text(text):
    tokens = nltk.word_tokenize(text)
    return tokens
```

[47]:
```
df['verified_reviews'] = df['verified_reviews'].apply(tokenize_text)
```

[48]:
```
df['verified_reviews']
```

[48]:
```
0                                              [love, echo]
1                                                   [loved]
2       [sometimes, playing, game, answer, question, c…
3       [lot, fun, thing, 4, yr, old, learns, dinosaur…
4                                                   [music]
                                ...
3145                    [perfect, kids, adults, everyone]
3146    [listening, music, searching, locations, check…
3147    [love, things, running, entire, home, tv, ligh…
3148    [complaint, sound, quality, isnt, great, mostl…
3149                                                 [good]
Name: verified_reviews, Length: 3150, dtype: object
```

Steamming

[49]:
```
from nltk.stem import PorterStemmer
```

[50]:
```
def perform_stemming(text):
    stemmer = PorterStemmer()
    stemmed_words = [stemmer.stem(word) for word in text]
    stemmed_text = ' '.join(stemmed_words)
    return stemmed_text
```

[51]:
```
df['verified_reviews'] = df['verified_reviews'].apply(perform_stemming)
```

[52]:
```
df['verified_reviews'].sample()
```

[52]:
```
1245    love eas use conveni echo spot offer
Name: verified_reviews, dtype: object
```

[53]:
```
df['verified_reviews'][3]
```

[53]:
```
'lot fun thing 4 yr old learn dinosaur control light play game like categori
nice sound play music well'
```

Lemmatization

[54]:
```
from nltk.stem import WordNetLemmatizer
```

[55]:
```
word_net =  WordNetLemmatizer()
```

[56]:
```
sentence = ' study studying studied studies'
sentence_words = nltk.word_tokenize(sentence)
```

[57]:
```
for word in sentence_words:
    print(word , word_net.lemmatize(word , pos = 'v'))
```

```
study study
studying study
studied study
studies study
```

Bag of Words

[58]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv= CountVectorizer()
```

[59]:
```
bow=cv.fit_transform(df['verified_reviews'])
```

[52]:
```
#print(cv.vocabulary_)
```

[61]:
```
print(bow[0].toarray())
print(bow[23].toarray())
```

```
[[0 0 0 … 0 0 0]]
[[0 0 0 … 0 0 0]]
```

TF-IDF

[47]:
```
data_frame = pd.DataFrame({'text':['people watch netflix','netflix watch␣
↪netflix','people write comment','netflix write comment','palash watch␣
↪netflix','palash write comment']})
```

[48]:
```
data_frame
```

[48]:

	text
0	people watch netflix
1	netflix watch netflix
2	people write comment
3	netflix write comment
4	palash watch netflix
5	palash write comment

[50]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer()
tfidf.fit_transform(data_frame['text']).toarray()
```

[50]:
```
array([[0.        , 0.48380155, 0.        , 0.66871989, 0.56457928,
        0.        ],
       [0.        , 0.8637229 , 0.        , 0.        , 0.50396702,
        0.        ],
       [0.54209195, 0.        , 0.        , 0.64208461, 0.        ,
        0.54209195],
       [0.60474937, 0.51822427, 0.        , 0.        , 0.        ,
        0.60474937],
       [0.        , 0.48380155, 0.66871989, 0.        , 0.56457928,
        0.        ],
       [0.54209195, 0.        , 0.64208461, 0.        , 0.        ,
        0.54209195]])
```

[ ]: