Cleaning

February 9, 2024

0.1 Importing Libraries

[ ]:
```
import pandas as pd
import re # REGEX
import nltk # NLP
```

0.2 Initial Visualization

[ ]:
```
df = pd.read_csv('hotel_reviews.csv')

df.head()
```

[ ]:

	Review	Rating
0	nice hotel expensive parking got good deal sta…	4
1	ok nothing special charge diamond member hilto…	2
2	nice rooms not 4* experience hotel monaco seat…	3
3	unique, great stay, wonderful time hotel monac…	5
4	great stay great stay, went seahawk game aweso…	5

[ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20491 entries, 0 to 20490
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Review  20491 non-null  object
 1   Rating  20491 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 320.3+ KB
```

0.3 Cleaning Data

0.3.1 Removing Duplicates

[ ]:
```
df.drop_duplicates(subset=['Review'], inplace=True)
```

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

[ ]: (20491, 2)

0.3.2 Removing Special Characters

[ ]:
```
def remove_special_char(text):

  first_step = re.sub('www\S+', '', text)

  second_step = first_step.lower()

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

  final_step = re.sub('[0-9]', '', third_step)

  return final_step
```

[ ]:
```
df['Review'] = df['Review'].apply(remove_special_char)
df.head()
```

[ ]:

	Review	Rating
0	nice hotel expensive parking got good deal sta…	4
1	ok nothing special charge diamond member hilto…	2
2	nice rooms not experience hotel monaco seattl…	3
3	unique great stay wonderful time hotel monaco …	5
4	great stay great stay went seahawk game awesom…	5

0.3.3 Removing Stopwords

[ ]:
```
nltk.download('stopwords')
stopwords_list = nltk.corpus.stopwords.words('english')

def remove_stopwords(text):
    word_list = text.split()

    new_phrase = ''

    for word in word_list:
        if word not in stopwords_list:
            new_phrase = new_phrase + ' ' + word

    return new_phrase
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/vasconcellos/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[ ]:
```
df['Review'] = df['Review'].apply(remove_stopwords)
df.head()
```

[ ]:

	Review	Rating
0	nice hotel expensive parking got good deal st…	4
1	ok nothing special charge diamond member hilt…	2
2	nice rooms experience hotel monaco seattle go…	3
3	unique great stay wonderful time hotel monaco…	5
4	great stay great stay went seahawk game aweso…	5

0.3.4 Stemming

[ ]:
```
nltk.download('rslp')
Stem = nltk.stem.RSLPStemmer()

def stemming(text):
    word_list = text.split()
    new_phrase = ''

    for word in word_list:
        radical = Stem.stem(word)
        new_phrase = new_phrase + ' ' + radical

    return radical
```

```
[nltk_data] Downloading package rslp to
[nltk_data]     /home/vasconcellos/nltk_data…
[nltk_data]   Package rslp is already up-to-date!
```

[ ]:
```
df['Review'] = df['Review'].apply(stemming)
df.head()
```

[ ]:

	Review	Rating
0	night	4
1	seattl	2
2	going	3
3	stay	5
4	tell	5

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

[ ]:
```
(20491, 2)
```

0.4 Exporting

[ ]:
```
df.to_csv('cleaned.csv')
```