ass1

October 29, 2024

This code scrapes data from the 20 Newsgroups dataset and extracts specific fields from each document using regular expressions.

[21]:
```
# import libraries

# import the dataset
from sklearn.datasets import fetch_20newsgroups

# import the necessary libraries
import re
import pandas as pd
```

[22]:
```
# Load the 20 Newsgroups dataset
newsgroups = fetch_20newsgroups(subset='train')

# Initialize lists to store extracted data
from_list = []
subject_list = []
summary_list = []
distribution_list = []
organization_list = []
keywords_list = []
lines_list = []
```

0.1 Define regex patterns:

A dictionary patterns stores the regular expressions for each field. Each pattern is designed to match the beginning of a line (^) followed by the field name and a colon, and then capture the rest of the line as the value ((.*)).

[23]:
```
# Define regex patterns for each field
patterns = {
    'From': re.compile(r'^From: (.*)', re.MULTILINE),
    'Subject': re.compile(r'^Subject: (.*)', re.MULTILINE),
    'Summary': re.compile(r'^Summary: (.*)', re.MULTILINE),
    'Distribution': re.compile(r'^Distribution: (.*)', re.MULTILINE),
    'Organization': re.compile(r'^Organization: (.*)', re.MULTILINE),
    'Keywords': re.compile(r'^Keywords: (.*)', re.MULTILINE),
    'Lines': re.compile(r'^Lines: (.*)', re.MULTILINE)
```

```
}
```

extract_field function: - This function takes a regex pattern and text as input. - It uses `pattern.search(text)` to find a match in the text. - If a match is found, it returns the captured group (the value of the field); otherwise, it returns `None`.

[24]:
```
# Function to extract field using regex
def extract_field(pattern: re.Pattern, text: str) -> str:
    match = pattern.search(text)
    return match.group(1) if match else None
```

[25]:
```
# Iterate over each document in the dataset
for text in newsgroups.data:
    try:
        # Extract data from each field and append to respective lists
        from_list.append(extract_field(patterns['From'], text))
        subject_list.append(extract_field(patterns['Subject'], text))
        summary_list.append(extract_field(patterns['Summary'], text))
        distribution_list.append(extract_field(patterns['Distribution'], text))
        organization_list.append(extract_field(patterns['Organization'], text))
        keywords_list.append(extract_field(patterns['Keywords'], text))
        lines_list.append(extract_field(patterns['Lines'], text))
    except Exception as e:
        # Handle any errors that occur during extraction
        print(f"Error processing document: {e}")
        from_list.append(None)
        subject_list.append(None)
        summary_list.append(None)
        distribution_list.append(None)
        organization_list.append(None)
        keywords_list.append(None)
        lines_list.append(None)
```

[26]:
```
# Create a DataFrame to store the extracted data
data = {
    'From': from_list,
    'Subject': subject_list,
    'Summary': summary_list,
    'Distribution': distribution_list,
    'Organization': organization_list,
    'Keywords': keywords_list,
    'Lines': lines_list
}
df = pd.DataFrame(data)

# Display the DataFrame
df.head()
```

[26]:

	From	Subject	Summary	Distribution	Organization	Keywords	Lines
0	lerxst@wam.umd.edu (where's my thing)	WHAT car is this!?	None	None	University of Maryland, College Park	None	15
1	guykuo@carson.u.washington.edu (Guy Kuo)	SI Clock Poll - Final Call	Final call for SI clock reports	None	University of Washington	SI,acceleration,clock,upgrade	11
2	twillis@ec.ecn.purdue.edu (Thomas E Willis)	PB questions…	None	usa	Purdue University Engineering Computer Network	None	36
3	jgreen@amber (Joe Green)	Re: Weitek P9000 ?	None	world	Harris Computer Systems Division	None	14
4	jcm@head-cfa.harvard.edu (Jonathan McDowell)	Re: Shuttle Launch Question	None	sci	Smithsonian Astrophysical Observatory, Cambrid…	None	23