In [1]:

```
import json
```

In [2]:

```
with open('output.json', 'r') as f:
    json_data = json.load(f)
```

In [3]:

```
import pandas as pd
from pandas import json_normalize
```

In [4]:

```
# First, normalize the 'subjects' part of the JSON
subjects_df = json_normalize(
    data=json_data['data'],
    record_path=['subjects'],
    meta=['Year', 'Term', 'Term Description'],
    errors='ignore'
)

# Initialize an empty DataFrame to hold all the classes
all_classes_df = pd.DataFrame()

subjects_df.head(10)
```

Out[4]:

	value	descr	descrformal	classes	Year	Term	Term Description
0	AAS	Asian American Studies	Asian American Studies	[{'titleShort': 'Intro To Asian American Hist'...	2014	FA14	Fall 2014
1	AEM	Applied Economics & Management	Applied Economics & Management	[{'titleShort': 'FWS:Fd Systems in Devlpng Wrl...	2014	FA14	Fall 2014
2	AEP	Applied & Engineering Physics	Applied & Engineering Physics	[{'titleShort': 'Laser & Photonics', 'titleLon...	2014	FA14	Fall 2014
3	AGSCI	Agriculture Sciences	Agricultural Sciences	[{'titleShort': 'Exploring AGSCI Careers', 'ti...	2014	FA14	Fall 2014
4	AIRS	Air Force Science	Aerospace Studies	[{'titleShort': 'Foundations of US Air Force I...	2014	FA14	Fall 2014
5	AIS	American Indian Studies	American Indian Studies	[{'titleShort': 'Indigenous North America', 't...	2014	FA14	Fall 2014
6	ALS	Agriculture & Life Sciences	Agriculture & Life Sciences	[{'titleShort': 'Leadership and GSL', 'titleLo...	2014	FA14	Fall 2014
7	AMST	American Studies	American Studies	[{'titleShort': 'FWS:Amer Cities-Global Econom...	2014	FA14	Fall 2014
8	ANSC	Animal Science	Animal Science	[{'titleShort': 'Domestic Animal Biology', 'ti...	2014	FA14	Fall 2014
9	ANTHR	Anthropology	Anthropology	[{'titleShort': 'FWS: Anthropology of Sport', ...	2014	FA14	Fall 2014

In [5]:
```
#The below lines of code normalizes with respect to classes

# Initialize an empty DataFrame to hold all the classes
all_classes_df = pd.DataFrame()

# Iterate through the 'subjects' DataFrame to normalize and append classes
for index, row in subjects_df.iterrows():
    classes_df = json_normalize(row['classes'])
    classes_df['subject_value'] = row['value']
    classes_df['subject_descr'] = row['descr']
    classes_df['subject_descrformal'] = row['descrformal']
    classes_df['Year'] = row['Year']
    classes_df['Term'] = row['Term']
    classes_df['Term Description'] = row['Term Description']

    # Append the classes to the all_classes_df
    #all_classes_df = all_classes_df.append(classes_df, ignore_index=True)
    all_classes_df = pd.concat([all_classes_df, classes_df], ignore_index=T
```

In [6]:
```
all_classes_df.head(50)
```

Out[6]:

	titleShort	titleLong	description	subject_value	subject_descr	su
0	Intro To Asian American Hist	Introduction to Asian American History	An introductory history of Chinese, Japanese, ...	AAS	Asian American Studies	
1	Asians in the Americas	Asians in the Americas: A Comparative Perspective	The common perception of ethnicity is that it ...	AAS	Asian American Studies	
2	Asian American Women's Hist	Asian American Women's History	This course examines the experiences and repre...	AAS	Asian American Studies	
3	Independent Study	Independent Study		AAS	Asian American Studies	

In [7]:
```
print("Total rows:",len(all_classes_df))
```

```
Total rows: 89590
```

In [8]:

```
#Declaring an order of columns
new_order = [
    'Year',
    'Term',
    'Term Description',
    'subject_value',
    'subject_descr',
    'subject_descrformal',
    'titleShort',
    'titleLong',
    'description'
]

all_classes_df = all_classes_df[new_order]
```

In [9]:

```
all_classes_df.head(50)
```

Out[9]:

	Year	Term	Term Description	subject_value	subject_descr	subject_descrformal	
0	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Studies	In Ar
1	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Studies	A
2	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Studies	Asia W
3	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Studies	Indepe

In [10]:

```
print("Total rows:",len(all_classes_df))
```

```
Total rows: 89590
```

In [11]:

```
# Drop the 'subject_descrformal' column in place
all_classes_df.drop('subject_descrformal', axis=1, inplace=True)
```

```
In [12]: all_classes_df.head(50)
```

Out[12]:

	Year	Term	Term Description	subject_value	subject_descr	titleShort	tit
0	2014	FA14	Fall 2014	AAS	Asian American Studies	Intro To Asian American Hist	Introdu Asian Ar
1	2014	FA14	Fall 2014	AAS	Asian American Studies	Asians in the Americas	Asian Ame Comp Pers
2	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Women's Hist	Asian Ar Women's
3	2014	FA14	Fall 2014	AAS	Asian American Studies	Independent Study	Indep

```
In [34]: print("Total after dropping column rows:",len(all_classes_df))
```

```
Total after dropping column rows: 89577
```

In [13]:

```
# Define a dictionary with the old column names as keys and new column name
rename_dict = {
    'subject_value': 'Subject',
    'subject_descr': 'SubjectDescription',
    'titleShort': 'courseTitle',
    'titleLong': 'courseTitleLong',
    'description': 'courseDescription'
}

# Rename the columns using the rename method
all_classes_df.rename(columns=rename_dict, inplace=True)

# Now df has the columns renamed
all_classes_df.head(20)
```

Out[13]:

	Year	Term	Term Description	Subject	SubjectDescription	courseTitle	courseTitleLo
0	2014	FA14	Fall 2014	AAS	Asian American Studies	Intro To Asian American Hist	Introduction Asian Americ Histc
1	2014	FA14	Fall 2014	AAS	Asian American Studies	Asians in the Americas	Asians in tl Americas: Comparati Perspecti
2	2014	FA14	Fall 2014	AAS	Asian American Studies	Asian American Women's Hist	Asian Americ Women's Histc
3	2014	FA14	Fall 2014	AAS	Asian American Studies	Independent Study	Independe Stu
4	2014	FA14	Fall 2014	AEM	Applied Economics & Management	FWS:Fd Systems in Devlpng Wrld	FWS:Fo Systems In Tl Developi World: Hea
5	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Foundations of Entrep & Bus	Foundations Entrepreneursh and Busine
6	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Econ of Env & Nat Resources	An Introducti to the Economi of Environmer
7	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Cont Controversies in Global	Contempora Controversies the Glob Econor
8	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Spreadsheet Modeling	Spreadshe Modeling f Manageme and Economi
9	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Introductory Statistics	Introducto Statisti
10	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Business Mngmnt&Organization	Busine Manageme and Organizati
11	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Financial Accounting	Financ Accounti
12	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Financial Accounting For Dyson	Financ Accounting F Dyson Majc
13	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Finance	Finan
14	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Marketing	Marketi
15	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Managerial Economics I	Manager Economic
16	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Management Communication	Manageme Communicati

	Year	Term	Term Description	Subject	SubjectDescription	courseTitle	courseTitleLo
17	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Farm Business Management	Farm Busine Manageme
18	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Business Law I	Business La
19	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Digital Business Strategy	Digital Busine Strate

In [14]:

```
!pip install pandas spacy
import spacy
import nltk
from nltk.stem.porter import PorterStemmer

# Load the spaCy model
nlp = spacy.load("en_core_web_sm")
# Initialize stemmer.
stemmer = PorterStemmer()
```

```
WARNING: Ignoring invalid distribution -mpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -sspec (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution -mportlib-metadata (c:\programdata
\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\sit
e-packages)
WARNING: Ignoring invalid distribution -mpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -sspec (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution -mportlib-metadata (c:\programdata
\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\sit
e-packages)
WARNING: Ignoring invalid distribution -mpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -sspec (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution -mportlib-metadata (c:\programdata
\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\sit
e-packages)
WARNING: Ignoring invalid distribution -mpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -umpy (c:\programdata\anaconda3\lib
\site-packages)
WARNING: Ignoring invalid distribution -sspec (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution -mportlib-metadata (c:\programdata
\anaconda3\lib\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\sit
e-packages)
WARNING: You are using pip version 21.1.2; however, version 23.3.1 is avai
lable.
You should consider upgrading via the 'C:\ProgramData\Anaconda3\python.exe
-m pip install --upgrade pip' command.
```

```
Requirement already satisfied: pandas in c:\programdata\anaconda3\lib\site
-packages (2.0.3)
Requirement already satisfied: spacy in c:\programdata\anaconda3\lib\site-
packages (3.7.2)
Requirement already satisfied: numpy>=1.20.3 in c:\programdata\anaconda3\l
ib\site-packages (from pandas) (1.24.3)
Requirement already satisfied: pytz>=2020.1 in c:\programdata\anaconda3\li
b\site-packages (from pandas) (2023.3.post1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\programdata\an
aconda3\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: tzdata>=2022.1 in c:\programdata\anaconda3
\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\si
te-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\programdata\anaco
nda3\lib\site-packages (from spacy) (2.0.8)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\programda
ta\anaconda3\lib\site-packages (from spacy) (3.0.12)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\programdata\a
naconda3\lib\site-packages (from spacy) (3.3.0)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\programdata\anaco
nda3\lib\site-packages (from spacy) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\programdata\a
naconda3\lib\site-packages (from spacy) (2.0.10)
Requirement already satisfied: setuptools in c:\programdata\anaconda3\lib
\site-packages (from spacy) (68.0.0)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\programdata
\anaconda3\lib\site-packages (from spacy) (1.0.10)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\programdata\ana
conda3\lib\site-packages (from spacy) (3.0.9)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
c:\programdata\anaconda3\lib\site-packages (from spacy) (2.4.2)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\programdata
\anaconda3\lib\site-packages (from spacy) (6.4.0)
Requirement already satisfied: packaging>=20.0 in c:\programdata\anaconda3
\lib\site-packages (from spacy) (23.1)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\programdata\anac
onda3\lib\site-packages (from spacy) (0.9.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\programdata\anaco
nda3\lib\site-packages (from spacy) (4.65.0)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in c:\programdata\anac
onda3\lib\site-packages (from spacy) (0.3.3)
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in c:\programdata\anaco
nda3\lib\site-packages (from spacy) (8.2.1)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\programda
ta\anaconda3\lib\site-packages (from spacy) (1.0.5)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\programdata\a
naconda3\lib\site-packages (from spacy) (2.31.0)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\programdata\anac
onda3\lib\site-packages (from spacy) (1.1.2)
Requirement already satisfied: jinja2 in c:\programdata\anaconda3\lib\site
-packages (from spacy) (3.1.2)
Requirement already satisfied: pydantic-core==2.10.1 in c:\programdata\ana
conda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spac
y) (2.10.1)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\programdata
\anaconda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->s
pacy) (4.8.0)
Requirement already satisfied: annotated-types>=0.4.0 in c:\programdata\an
aconda3\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spac
y) (0.6.0)
```

```
Requirement already satisfied: idna<4,>=2.5 in c:\programdata\anaconda3\li
b\site-packages (from requests<3.0.0,>=2.13.0->spacy) (3.4)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\programdata
\anaconda3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (2.0.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\programdata\anacon
da3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in c:\programdata\anacon
da3\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy) (2023.7.22)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\programdata\anacon
da3\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\programdata
\anaconda3\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy) (0.1.3)
Requirement already satisfied: colorama in c:\programdata\anaconda3\lib\si
te-packages (from tqdm<5.0.0,>=4.38.0->spacy) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\programdata\anaco
nda3\lib\site-packages (from typer<0.10.0,>=0.3.0->spacy) (8.1.7)
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in c:\programda
ta\anaconda3\lib\site-packages (from weasel<0.4.0,>=0.1.0->spacy) (0.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\programdata\anaconda3
\lib\site-packages (from jinja2->spacy) (2.1.1)
```

In [16]:

```python

def clean_text(text):
    """
    This function returns the cleaned text using spacy

    Args: text to be cleaned

    Returns: Cleaned text and count of words in it
        type: string, int
    """
    # If the text is None or empty, return an empty string
    if text is None or text == '':
        return 'EmptyString', 0
    # Parse the sentence using the loaded 'en' model object `nlp`
    doc = nlp(text)

    # Tokenize and remove stop words, punctuation, and perform lemmatizatio
    clean_tokens = [token.lemma_.lower() for token in doc if not token.is_s
    cleaned_text = " ".join(clean_tokens)
    # Get the word count
    word_count = len(clean_tokens)

    # Re-join tokens into a single string
    return cleaned_text, word_count
```

In [17]:
```
#The above will apply clean_text function to courseDescription column

cleaned_data = all_classes_df['courseDescription'].apply(clean_text)


# Split the tuples into two lists - one for cleaned text and one for word c
cleaned_texts = [item[0] for item in cleaned_data]
word_counts = [item[1] for item in cleaned_data]

# Assign the cleaned texts and word counts to their respective columns
all_classes_df['cleaned_courseDescription'] = cleaned_texts  # Assuming the
all_classes_df['NoOfWords'] = word_counts
```

```
Before cleaning:  The common perception of ethnicity is that it is a "n
atural" and an inevitable consequence of cultural difference. "Asians"
overseas, in particular, have won repute as a people who cling tenaciou
sly to their culture and refuse to assimilate into their host societies
and cultures. But, who are the "Asians?" On what basis can we label "As
ians" an ethnic group? Although there is a significant Asian presence i
n the Caribbean, the category "Asian" itself does not exist in the Cari
bbean. What does this say about the nature of categories that label and
demarcate groups of people on the basis of alleged cultural and phenoty
pical characteristics? This course will examine the dynamics behind gro
up identity, namely ethnicity, by comparing and contrasting the multicu
ltural experience of Asian populations in the Caribbean and the United
States. Ethnographic case studies will focus on the East Indian and Chi
nese experiences in the Caribbean and the Chinese, Korean, Japanese, Fi
lipino, and Indian experiences in the United States.

After cleaning:  common perception ethnicity natural inevitable consequ
ence cultural difference asians overseas particular win repute people c
ling tenaciously culture refuse assimilate host society culture asians
basis label asians ethnic group significant asian presence caribbean ca
tegory asian exist caribbean nature category label demarcate group peop
```

In [18]:
```
all_classes_df.head(10)
```

3	2014	FA14	Fall 2014	AAS	Asian American Studies	Independent Study	Independent Study	
4	2014	FA14	Fall 2014	AEM	Applied Economics & Management	FWS:Fd Systems in Devlpng Wrld	FWS:Food Systems In The Developing World: Heal...	\
5	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Foundations of Entrep & Bus	Foundations of Entrepreneurship and Business	I
6	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Econ of Env & Nat Resources	An Introduction to the Economics of Environmen...	
7	2014	FA14	Fall 2014	AEM	Applied Economics & Management	Cont Controversies in Global	Contemporary Controversies in the Global Economy	T
					Applied Economics	Spreadsheet	Spreadsheet Modeling for	

In [19]:
```
#The below lines of code just check whether the cleaned_text fuction has be

# Filter the DataFrame where 'Year' is 2023 and 'Subject' is 'AAP'
filtered_df = all_classes_df[(all_classes_df['Year'] == '2014') & (all_clas

#Print all the 'cleaned_courseDescription' from the filtered DataFrame
for description in filtered_df['cleaned_courseDescription']:
    print(description)
```

```
introductory history chinese japanese asian indians filipinos koreans unit
ed states mid century major theme include racism resistance labor migratio
n community formation imperialism struggle equality
common perception ethnicity natural inevitable consequence cultural differ
ence asians overseas particular win repute people cling tenaciously cultur
e refuse assimilate host society culture asians basis label asians ethnic
group significant asian presence caribbean category asian exist caribbean
nature category label demarcate group people basis allege cultural phenoty
pical characteristic course examine dynamic group identity ethnicity compa
re contrast multicultural experience asian population caribbean united sta
tes ethnographic case study focus east indian chinese experience caribbean
chinese korean japanese filipino indian experience united states
course examine experience representation asian american woman century pres
ent explore life context immigrant woman woman bear questions identity pow
er heart course investigate intertwine nature race gender nation pay parti
cular attention practice history seek well understanding scholar recover h
istory population render invisible traditional method inquiry course mater
ial include numerous primary source addition scholarship variety disciplin
e history literature sociology anthropology
EmptyString
```

In [22]:
```
# Group by 'Subject' and 'Year', then join the 'cleaned_courseDescription'
# and get the first (or unique) 'SubjectDescription' for each group.
grouped_new = all_classes_df.groupby(['Subject', 'Year']).agg({
    #'cleaned_courseDescription': ' '.join,
    'cleaned_courseDescription': [' '.join, 'count'],
    'SubjectDescription': 'first'  # Assuming it's the same for each group,
}).reset_index()
```

In [24]:
```
grouped_new.columns = ['Subject', 'Year', 'Grouped_Subject_Description', 'N
```

In [26]:

```
new_order = [
    'Year',
    'Subject',
    'SubjectLongForm',
    'Grouped_Subject_Description',
    'NoOfClasses'
]

grouped_new = grouped_new[new_order]

grouped_new.head(20)
```

Out[26]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses
0	2020	AAP	Architecture, Art, and Plannin	EmptyString	1
1	2021	AAP	Architecture, Art, and Plannin	EmptyString	1
2	2023	AAP	Architecture, Art, and Plannin	topics tba create justice worlds examine struc...	2
3	2014	AAS	Asian American Studies	introductory history chinese japanese asian in...	4
4	2015	AAS	Asian American Studies	course examine historical contemporary issue a...	13
5	2016	AAS	Asian American Studies	course examine historical contemporary issue a...	14
6	2017	AAS	Asian American Studies	course examine historical contemporary issue a...	9
7	2018	AAS	Asian American Studies	course introduce student historical contempora...	10
8	2019	AAS	Asian American Studies	interdisciplinary course offer introduction st...	15
9	2020	AAS	Asian American Studies	interdisciplinary course offer introduction st...	12
10	2021	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9
11	2022	AAS	Asian American Studies	interdisciplinary course offer introduction st...	19
12	2023	AAS	Asian American Studies	course introduce variety writing asian north a...	13
13	2024	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9
14	2014	AEM	Applied Economics & Management	like subsistence farmer develop world choice c...	75
15	2015	AEM	Applied Economics & Management	introduction cost accounting emphasize applica...	163
16	2016	AEM	Applied Economics & Management	course develop data drive model base approach ...	205
17	2017	AEM	Applied Economics & Management	course develop data drive model base approach ...	199
18	2018	AEM	Applied Economics & Management	course develop data drive model base approach ...	218
19	2019	AEM	Applied Economics & Management	course develop data drive model base approach ...	229

In [27]:

```
print(len(grouped_new))
print(len(all_classes_df))
```

```
2022
89590
```

In [28]:

```
grouped_new.to_csv('Grouped_Subject_Description.csv', index=False)
```

In [29]:

```python
import pandas as pd
```

In [30]:

```python
# Read the CSV file back into a DataFrame. It is Cleaned Dataset
grouped_new_read = pd.read_csv('Grouped_Subject_Description.csv')
```

In [31]:

```python
grouped_new_read.head(20)
```

Out[31]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses
0	2020	AAP	Architecture, Art, and Plannin	EmptyString	1
1	2021	AAP	Architecture, Art, and Plannin	EmptyString	1
2	2023	AAP	Architecture, Art, and Plannin	topics tba create justice worlds examine struc...	2
3	2014	AAS	Asian American Studies	introductory history chinese japanese asian in...	4
4	2015	AAS	Asian American Studies	course examine historical contemporary issue a...	13
5	2016	AAS	Asian American Studies	course examine historical contemporary issue a...	14
6	2017	AAS	Asian American Studies	course examine historical contemporary issue a...	9
7	2018	AAS	Asian American Studies	course introduce student historical contempora...	10
8	2019	AAS	Asian American Studies	interdisciplinary course offer introduction st...	15
9	2020	AAS	Asian American Studies	interdisciplinary course offer introduction st...	12
10	2021	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9
11	2022	AAS	Asian American Studies	interdisciplinary course offer introduction st...	19
12	2023	AAS	Asian American Studies	course introduce variety writing asian north a...	13
13	2024	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9
14	2014	AEM	Applied Economics & Management	like subsistence farmer develop world choice c...	75
15	2015	AEM	Applied Economics & Management	introduction cost accounting emphasize applica...	163
16	2016	AEM	Applied Economics & Management	course develop data drive model base approach ...	205
17	2017	AEM	Applied Economics & Management	course develop data drive model base approach ...	199
18	2018	AEM	Applied Economics & Management	course develop data drive model base approach ...	218
19	2019	AEM	Applied Economics & Management	course develop data drive model base approach ...	229

In [32]:
```
print(len(grouped_new_read))
```

```
2022
```

In [1]:
```
# Function to count occurrences of business-related words
business_keywords = {'business', 'startup', 'entrepreneurship', 'entreprene

def count_business_words(description):
    """
    This function returns the count of business related words

    Args: text to be analysed

    Returns: Returns count of business related words
        type: int
    """
    # Tokenize the description into words
    words = description.split()
    count = len(words)
    # Count how many words are in the business_keywords set
    business_related_words = sum(word.lower() in business_keywords for word
    percentageOfBusinessWords = (business_related_words/count)*100
    return count, business_related_words, round(percentageOfBusinessWords,2
```

In [34]:
```
#Checking if the dataframe is imported correctly.

# Filter the DataFrame where 'Year' is 2023 and 'Subject' is 'AAP'
grouped_filtered_df_new = grouped_new_read[(grouped_new_read['Year'] == 201

#print(grouped_filtered_df_new)

#Print all the 'cleaned_courseDescription' from the filtered DataFrame
for description in grouped_filtered_df_new['Grouped_Subject_Description']:
    print(description)
```

```
introductory history chinese japanese asian indians filipinos koreans unit
ed states mid century major theme include racism resistance labor migratio
n community formation imperialism struggle equality common perception ethn
icity natural inevitable consequence cultural difference asians overseas p
articular win repute people cling tenaciously culture refuse assimilate ho
st society culture asians basis label asians ethnic group significant asia
n presence caribbean category asian exist caribbean nature category label
demarcate group people basis allege cultural phenotypical characteristic c
ourse examine dynamic group identity ethnicity compare contrast multicultu
ral experience asian population caribbean united states ethnographic case
study focus east indian chinese experience caribbean chinese korean japane
se filipino indian experience united states course examine experience repr
esentation asian american woman century present explore life context immig
rant woman woman bear questions identity power heart course investigate in
tertwine nature race gender nation pay particular attention practice histo
ry seek well understanding scholar recover history population render invis
ible traditional method inquiry course material include numerous primary s
ource addition scholarship variety discipline history literature sociology
anthropology EmptyString
```

In [35]:

```
data = grouped_new_read['Grouped_Subject_Description'].apply(count_business

totalCount = [item[0] for item in data]
businessCount = [item[1] for item in data]
percentageOfBusiness = [item[2] for item in data]

# Assign the cleaned texts and word counts to their respective columns
grouped_new_read['TotalWords'] = totalCount
grouped_new_read['businessCount'] = businessCount
grouped_new_read['%OfBusinessWords'] = percentageOfBusiness

grouped_new_read.head(20)
```

Out[35]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses	TotalWords
0	2020	AAP	Architecture, Art, and Plannin	EmptyString	1	1
1	2021	AAP	Architecture, Art, and Plannin	EmptyString	1	1
2	2023	AAP	Architecture, Art, and Plannin	topics tba create justice worlds examine struc...	2	17
3	2014	AAS	Asian American Studies	introductory history chinese japanese asian in...	4	162
4	2015	AAS	Asian American Studies	course examine historical contemporary issue a...	13	605
5	2016	AAS	Asian American Studies	course examine historical contemporary issue a...	14	666
6	2017	AAS	Asian American Studies	course examine historical contemporary issue a...	9	539
7	2018	AAS	Asian American Studies	course introduce student historical contempora...	10	491
8	2019	AAS	Asian American Studies	interdisciplinary course offer introduction st...	15	874
9	2020	AAS	Asian American Studies	interdisciplinary course offer introduction st...	12	662
10	2021	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9	545
11	2022	AAS	Asian American Studies	interdisciplinary course offer introduction st...	19	1088
12	2023	AAS	Asian American Studies	course introduce variety writing asian north a...	13	557
13	2024	AAS	Asian American Studies	interdisciplinary course offer introduction st...	9	399
14	2014	AEM	Applied Economics & Management	like subsistence farmer develop world choice c...	75	2410
15	2015	AEM	Applied Economics & Management	introduction cost accounting emphasize applica...	163	6182
16	2016	AEM	Applied Economics & Management	course develop data drive model base approach ...	205	8337
17	2017	AEM	Applied Economics & Management	course develop data drive model base approach ...	199	8133
18	2018	AEM	Applied Economics & Management	course develop data drive model base approach ...	218	9064
19	2019	AEM	Applied Economics & Management	course develop data drive model base approach ...	229	9482

# Retrieve all rows with the highest 'businessCount'

In [36]:

```python
# Retrieve all rows with the highest 'businessCount'
rows_with_max_business = grouped_new_read.nlargest(1, 'businessCount')

# Display the rows
rows_with_max_business.head()
```

Out[36]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses	TotalWor
1280	2023	NBA	Grad Mgmt Business Admin	combine classroom session international experi...	162	103

# Retrieve all rows with the highest '%ofBusinessWords'

In [37]:

```python
# Retrieve all rows with the highest '%ofBusinessWords'
rows_with_max_business_percent = grouped_new_read.nlargest(1, '%OfBusinessW

# Display the rows
rows_with_max_business_percent.head()
```

Out[37]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses	TotalWor
1294	2020	NBAT	Grad Mgmt Business Admin CT	course provide understanding emergence develop...	1	3

# Retrieve all rows with the highest 'Total number of words'

In [38]:

```python
# Retrieve all rows with the highest 'Total number of words'

rows_with_max_words = grouped_new_read.nlargest(1, 'TotalWords')

rows_with_max_words.head()
```

Out[38]:

	Year	Subject	SubjectLongForm	Grouped_Subject_Description	NoOfClasses	TotalWor
1132	2023	LAW	Law	major segment trial explore opening statement ...	379	193

Use Bag of Words

In [39]:

```python
#Using Bag of Words
from sklearn.feature_extraction.text import CountVectorizer
import matplotlib.pyplot as plt
```

In [40]:
```
# First, create the vectorizer with the business-related words only
business_terms = ['business', 'startup', 'entrepreneurship', 'entrepreneur'
vectorizer = CountVectorizer(vocabulary=business_terms)

# Apply the vectorizer to the Grouped_Subject_Description column
X = vectorizer.fit_transform(grouped_new_read['Grouped_Subject_Description'
```

In [41]:
```
# Convert the result to a DataFrame
business_words_df = pd.DataFrame(X.toarray(), columns=vectorizer.get_featur
```

In [43]:
```

# Now, combine this with your original dataframe
bag_of_words_df = pd.concat([grouped_new_read, business_words_df], axis=1)

bag_of_words_df.head(15)
```

```
 7  2018   AAS       Asian American        course introduce student   10
                            Studies      historical contempora...

 8  2019   AAS       Asian American   interdisciplinary course offer   15
                            Studies             introduction st...

 9  2020   AAS       Asian American   interdisciplinary course offer   12
                            Studies             introduction st...

10  2021   AAS       Asian American   interdisciplinary course offer    9
                            Studies             introduction st...

11  2022   AAS       Asian American   interdisciplinary course offer   19     1
                            Studies             introduction st...

12  2023   AAS       Asian American  course introduce variety writing   13
                            Studies               asian north a...

13  2024   AAS       Asian American   interdisciplinary course offer    9
                            Studies             introduction st...

14  2014   AEM  Applied Economics    like subsistence farmer develop   75     2
                    & Management              world choice c...
```