

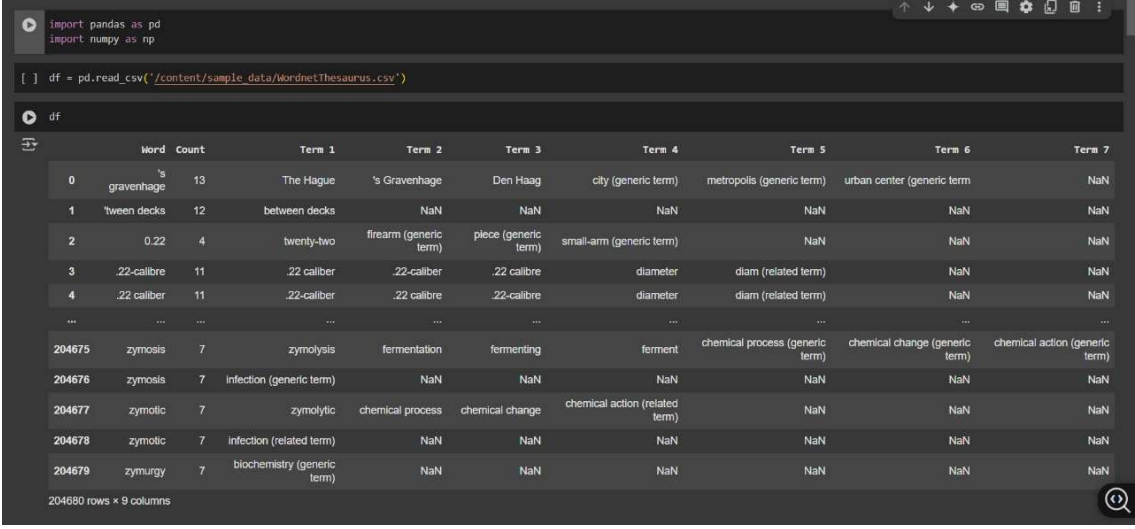
EDS THEORY ACTIVITY .1

Name: Mayuri Agrawal

Roll no: CS7-79

PRN: 202401110059

Dataset: Word Net



```
import pandas as pd
import numpy as np

df = pd.read_csv("../content/sample_data/wordnetthesaurus.csv")
```

df

	word	Count	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7
0	's	13	The Hague	's Gravenhage	Den Haag	city (generic term)	metropolis (generic term)	urban center (generic term)	NaN
1	'tween decks	12	between decks	NaN	NaN	NaN	NaN	NaN	NaN
2	0.22	4	twenty-two	firearm (generic term)	piece (generic term)	small-arm (generic term)	NaN	NaN	NaN
3	.22-calibre	11	.22 caliber	.22-caliber	.22 calibre	diameter	diam (related term)	NaN	NaN
4	.22 caliber	11	.22-caliber	.22 calibre	.22-calibre	diameter	diam (related term)	NaN	NaN
...
204675	zymosis	7	zymolysis	fermentation	fermenting	ferment	chemical process (generic term)	chemical change (generic term)	chemical action (generic term)
204676	zymosis	7	infection (generic term)	NaN	NaN	NaN	NaN	NaN	NaN
204677	zymotic	7	zymolytic	chemical process	chemical change	chemical action (related term)	NaN	NaN	NaN
204678	zymotic	7	infection (related term)	NaN	NaN	NaN	NaN	NaN	NaN
204679	zymurgy	7	biochemistry (generic term)	NaN	NaN	NaN	NaN	NaN	NaN

204680 rows x 9 columns

- 1) Display rows related to geographical locations based on specific keywords.

```
place_keywords = ['city', 'town', 'village', 'country', 'capital']
places = df[df.apply(lambda row: any(any(pk in str(x).lower() for pk in place_keywords) for x in row), axis=1)]
print(places)
```

	Word	Count	Term 1 \
0	's gravenhage	13	The Hague
435	aachen	6	Aachen
437	aalborg	7	Aalborg
439	aalost	6	Alost
446	aarhus	6	Arhus
...
204434	zimbabwean	10	Zimbabwean
204470	zion	4	Israel
204532	zomba	5	Zomba
204633	zurich	6	Zurich
204641	zydeco	6	country music (generic term)

	Term 2	Term 3 \
0	's Gravenhage	Den Haag
435	Aken	Aix la Chapelle
437	Alborg	city (generic term)
439	Aalost	town (generic term)
446	Aarhus	city (generic term)
...
204434	African country	African nation (related term)
204470	State of Israel	Yisrael
204532	city (generic term)	metropolis (generic term)
204633	city (generic term)	metropolis (generic term)
204641	country and western (generic term)	C and W (generic term)

	Term 4	Term 5 \
0	city (generic term)	metropolis (generic term)
435	city (generic term)	metropolis (generic term)
437	metropolis (generic term)	urban center (generic term)
439	NaN	NaN
446	metropolis (generic term)	urban center (generic term)
...
204434	NaN	NaN
204470	Zion	Sion
204633	urban center (generic term)	NaN

- 2) Identify the most common synonym across multiple term columns in the dataset.

```

top_words = df[['word', 'count']].sort_values(by='count', ascending=False).head(10)
print(top_words)

```

	word	count
19913	blood oxygenation level dependent functional m...	71
6097	american federation of labor and congress of i...	69
96220	international islamic front for jihad against ...	64
192336	united nations educational scientific and cult...	63
121034	national association of securities dealers aut...	63
139722	popular front for the liberation of palestine...	61
56230	earnings before interest taxes depreciation am...	60
192338	united nations office for drug control and cri...	59
15094	baron friedrich wilhelm ludolf gerhard augusti...	59
139719	popular democratic front for the liberation of...	56

3) Display rows where Term 2 is missing but Term 3 is filled to identify data inconsistencies.

```

words_with_5_synonyms = df.dropna(thresh=7)
print(words_with_5_synonyms[['Word', 'Term 1', 'Term 2', 'Term 3', 'Term 4', 'Term 5']])

```

	Word	Term 1	Term 2	Term 3	Term 4	Term 5
0	's gravenhage	The Hague	's Gravenhage	Den Haag	city (generic term)	metropolis (generic term)
3	.22-calibre	.22 caliber	.22-calibre	.22 calibre	diameter	diam (related term)
4	.22 caliber	.22-caliber	.22 calibre	.22-calibre	diameter	diam (related term)
5	.22 calibre	.22 caliber	.22-caliber	.22-calibre	diameter	diam (related term)
6	.38-caliber	.38 caliber	.38 calibre	.38-calibre	diameter	diam (related term)
...
204660	zygomycotina	Zygomycota	subdivision Zygomycota	Zygomycotina	subdivision Zygomycotina	division (generic term)
204663	zygophyllum fabago	bean caper	Syrian bean caper	Zygophyllum fabago	shrub (generic term)	bush (generic term)
204669	zyloprim	allopurinol	Zyloprim	medicine (generic term)	medication (generic term)	medicament (generic term)
204673	zymolysis	zymosis	fermentation	fermenting	ferment	chemical process (generic term)
204675	zymosis	zymolysis	fermentation	fermenting	ferment	chemical process (generic term)

[60828 rows x 6 columns]

4) Display words with missing entries in Term 1 to detect incomplete data.

```
[ ] missing_term1 = df[df['Term 1'].isnull()]
print(missing_term1['Word'])
```

```
197366 void
Name: Word, dtype: object
```

5) Extract all rows related to firearms based on keyword search.

```
firearm_related = df[df.apply(lambda row: any('firearm' in str(x).lower() for x in row), axis=1)]
print(firearm_related)
```

	Word	Count	\
2	0.22	4	
11765	atf	3	
12848	autoloader	10	
12863	automatic	9	
12864	automatic	9	
12870	automatic firearm	17	
12873	automatic pistol	16	
12874	automatic rifle	15	
24804	bureau of alcohol tobacco and firearms	38	
67692	firearm	7	
73600	garand	6	
73601	garand rifle	12	
84567	handgun	7	
109820	m-1	3	
109821	m-1 rifle	9	
109932	machine gun	11	
109942	machine rifle	13	
113224	mauser	6	
120171	muzzle loader	13	
135984	piece	5	
136923	pistol	6	
151083	repeater	8	
151087	repeating firearm	17	
153320	rifle	5	
153694	riot gun	8	
158806	scattergun	10	
161282	self-loader	11	
161484	semiautomatic	13	
161485	semiautomatic firearm	21	
161486	semiautomatic pistol	20	
164101	shooting iron	13	
164351	shotgun	7	
164862	side arm	8	
167566	small-arm	9	

6) Extract all rows mentioning 'city' in any column.

```
city_words = df[df.apply(lambda row: any('city' in str(x).lower() for x in row), axis=1)]
print(city_words[['Word', 'Term 1', 'Term 2']])
```

	Word	Term 1	Term 2
0	's gravenhage	The Hague	's Gravenhage
435	aachen	Aachen	Aken
437	aalborg	Aalborg	Alborg
446	aarhus	Arhus	Aarhus
470	abadan	Abadan	city (generic term)
...
204294	zaragoza	Zaragoza	Saragossa
204297	zaria	Zaria	city (generic term)
204300	zarqa	Az Zarqa	Zarqa
204532	zomba	Zomba	city (generic term)
204633	zurich	Zurich	city (generic term)

[1630 rows x 3 columns]

7) Display all synonyms of a given word from a dataset, handling case insensitivity.

```
given_word = "The Hague"
synonyms = df[df['Word'].str.lower() == given_word.lower()]
print(synonyms[['Term 1', 'Term 2', 'Term 3', 'Term 4', 'Term 5']])
```

	Term 1	Term 2	Term 3	Term 4	Term 5
182996	The Hague	's Gravenhage	Den Haag	city (generic term)	
182996	metropolis (generic term)				

8) Display synonyms u have listed per word.

```
given_word = "The Hague"
synonyms = df[df['Word'].str.lower() == given_word.lower()]
print(synonyms[['Term 1', 'Term 2', 'Term 3', 'Term 4', 'Term 5']])
```

	Term 1	Term 2	Term 3	Term 4	Term 5
182996	The Hague	's Gravenhage	Den Haag	city (generic term)	
182996	metropolis (generic term)				

9) Identify to find duplicate entries

```

duplicates = df[df.duplicated('Word')]
print(duplicates)

```

	Word	Count	Term 1 \
15	0	1	zero
17	1	1	one
21	10	2	ten
24	100	3	hundred
26	1000	4	thousand
...
204631	zulu	4	Zulu
204636	zurvanism	9	Zurvanism
204652	zygomatic	9	cheekbone
204676	zymosis	7	infection (generic term)
204678	zymotic	7	infection (related term)

	Term 2	Term 3 \
15	nought	cipher
17	I	ace
21	X	tenner
24	C	century
26	one thousand	M
...
204631	Nguni (generic term)	NaN
204636	theological doctrine (generic term)	heresy (generic term)
204652	zygomatic bone	malar
204676	NaN	NaN
204678	NaN	NaN

	Term 4	Term 5 \
15	cypher	digit (generic term)
17	single	unity
21	decade	large integer (generic term)
24	one C	centred
26	K	chiliad
...
204631	NaN	NaN
204636	unorthodoxy (generic term)	NaN

10) Counting words with with exactly three main synonyms.

11) Checking how many words have no synonyms .

12) Finding words that are numeric.

```
[ ] exactly_3_synonyms = df.dropna(subset=['Term 1', 'Term 2', 'Term 3']).dropna(subset=['Term 4', 'Term 5', 'Term 6', 'Term 7'], how='all')
print(len(exactly_3_synonyms))

93704

[ ] no_synonyms = df['Term 1'].isnull().sum()
percentage = (no_synonyms / len(df)) * 100
print(f"Percentage of words without synonyms: {percentage:.2f}%")

Percentage of words without synonyms: 0.00%

[ ] numeric_words = df[df['Word'].str.match(r'^\d', na=False)]
print(numeric_words[['Word', 'Count']])
```

	Word	Count
2	0.22	4
14	0	1
15	0	1
16	1	1
17	1	1
...
125612	24-Oct	5
161910	11-Sep	5
161989	11-Sep	5
161990	17-Sep	5
161991	29-Sep	5

[420 rows x 2 columns]

13) Finding non english word in df

```
[ ] import re

non_english = df[df['Word'].str.contains(r'^\x00-\x7F', na=False)]
print(non_english[['Word', 'Count']])
```

Empty DataFrame
Columns: [Word, Count]
Index: []

14) Code checks for rows in df where all terms are missing.

```

one_synonym = df[df[['Term 2', 'Term 3', 'Term 4', 'Term 5', 'Term 6', 'Term 7']].isnull().all(axis=1)]
print(one_synonym[['Word', 'Term 1']])

```

	Word	Term 1
1	'tween decks	between decks
175	24/7	uptime (generic term)
407	a cappella	unaccompanied (similar term)
415	a hundred times	hundredfold
417	a la carte	table d'hote (antonym)
...
204668	zygotic	cell (related term)
204670	zymase	enzyme (generic term)
204676	zymosis	infection (generic term)
204678	zymotic	infection (related term)
204679	zymurgy	biochemistry (generic term)

[25875 rows x 2 columns]

15) It finds and shows the rows where the 'Word' is identical to one of the synonym terms listed in 'Term 1' , 'Term 2' , or 'Term 3'.

```

identical_synonyms = df[df.apply(lambda row: any(row['Word'] == term for term in [row['Term 1'], row['Term 2'], row['Term 3']]), axis=1)]
print(identical_synonyms)

```

	Word	Count	Term 1 \
47	11-Nov	5	Martinmas
73	14-Jul	5	Bastille Day
82	15-Aug-45	5	V-J Day
106	17-Nov	5	Revolutionary Organization 17 November
314	06-Jun-44	5	D-day
352	08-May-45	5	V-E Day
357	11-Sep	5	11-Sep
359	11-Sep	5	11-Sep
9186	14-Apr	5	Pan American Day
12510	01-Aug	5	Lammas
12511	15-Aug	5	Assumption
12512	06-Aug	5	Transfiguration
47180	24-Dec	5	Christmas Eve
47281	31-Dec	5	New Year's Eve
47282	08-Dec	5	Immaculate Conception
60202	equine	6	equine
65933	02-Feb	5	Candlemas
65939	12-Feb	5	Lincoln's Birthday
65941	02-Feb	5	Groundhog Day
65942	22-Feb	5	Washington's Birthday
65943	29-Feb	5	leap day
66171	feline	6	feline
98356	01-Jan	5	Circumcision
98357	01-Jan	5	Solemnity of Mary
98358	01-Jan	5	New Year's Day
98360	20-Jan	5	Saint Agnes's Eve
98361	20-Jan	5	Inauguration Day
99842	01-Jul	5	Dominion Day

16) Displays the first few rows of a Data Frame that maps words to their counts and related terms, likely for text normalization or synonym analysis.



```
df.head()
```

	Word	Count	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7
0	's gravenhage	13	The Hague	's Gravenhage	Den Haag	city (generic term)	metropolis (generic term)	urban center (generic term)	NaN
1	'tween decks	12	between decks	NaN	NaN	NaN	NaN	NaN	NaN
2	0.22	4	twenty-two	firearm (generic term)	piece (generic term)	small-arm (generic term)	NaN	NaN	NaN
3	.22-calibre	11	.22 caliber	.22-caliber	.22 calibre	diameter	diam (related term)	NaN	NaN
4	.22 caliber	11	.22-caliber	.22 calibre	.22-calibre	diameter	diam (related term)	NaN	NaN

17) Filter the Data Frame to keep only rows where at least 7 columns are non-null, then displays words along with five of their synonyms.

18) Identify words or phrases that have at least five synonyms or closely related terms from a large textual dataset.

```

words_with_5_synonyms = df.dropna(thresh=7)
print(words_with_5_synonyms[['Word', 'Term 1', 'Term 2', 'Term 3', 'Term 4', 'Term 5']])

```

	Word	Term 1	Term 2	Term 3	Term 4	Term 5
0	's gravenhage	The Hague	's Gravenhage			
3	.22-calibre	.22 caliber	.22-calibre			
4	.22 caliber	.22-caliber	.22 calibre			
5	.22 calibre	.22 caliber	.22-calibre			
6	.38-calibre	.38 caliber	.38 calibre			
...
204660	zygomycotina	Zygomycota	subdivision Zygomycota			
204663	zygophyllum fabago	bean caper	Syrian bean caper			
204669	zyloprim	allopurinol	Zyloprim			
204673	zymolysis	zymosis	fermentation			
204675	zymosis	zymolysis	fermentation			
...
0		Den Haag	city (generic term)			
3		.22 calibre	diameter			
4		.22-calibre	diameter			
5		.22-calibre	diameter			
6		.38-calibre	diameter			
...
204660	Zygomycotina	subdivision Zygomycotina				
204663	Zygophyllum fabago	shrub (generic term)				
204669	medicine (generic term)	medication (generic term)				
204673	fermenting	ferment				
204675	fermenting	ferment				
...
0		metropolis (generic term)				
3		diam (related term)				
4		diam (related term)				
5		diam (related term)				
6		diam (related term)				
...
204660	division (generic term)					
204663	bush (generic term)					
204669	medicament (generic term)					
204673	chemical process (generic term)					
204675	chemical process (generic term)					

[68828 rows x 6 columns]

19) This code sorts a DataFrame by the "Count" column in descending order and displays the top 10 most frequent "Word" entries.

20) Display a tabular list showing the top 10 phrases along with their count values, sorted from highest to lowest.

```
top_words = df[['word', 'count']].sort_values(by='count', ascending=False).head(10)
print(top_words)
```

	word	count
19913	blood oxygenation level dependent functional m...	71
6097	american federation of labor and congress of i...	69
96220	international islamic front for jihad against ...	64
192336	united nations educational scientific and cult...	63
121034	national association of securities dealers aut...	63
139722	popular front for the liberation of palestine...	61
56230	earnings before interest taxes depreciation am...	60
192338	united nations office for drug control and cri...	59
15094	baron friedrich wilhelm ludolf gerhard augusti...	59
139719	popular democratic front for the liberation of...	56