Theory Activity No. 1

Name: varad Jadhav

Prn: 202401100039

Roll no. : CS5-37

Batch:CS5

20 Problem statements:

1 Find the total number of missing values in each column.

2 Drop rows with any missing values.

3 Find the unique part_of_speech values.

4 Find the most frequent part_of_speech.

5 Find the number of lemmas associated with each part_of_speech.

6 Find the top 10 most common lemmas.

7 Find how many lemmas have more than 3 synonyms.

8 Find the lemma with the highest number of synonyms.

9 Calculate the average number of synonyms per lemma.

10 List all lemmas that have only one synonym.

11 Find all synonyms containing the word "caliber".

12 Replace all hyphens "-" in lemma and synonyms with spaces.

13 Find all lemmas that are adjectives.

14 Find the number of unique lemmas.

15 Find the number of lemmas whose synonyms list contains duplicates.

16 Create a new column num_synonyms that counts synonyms for each lemma.

17 Filter lemmas whose synonyms include themselves.

18 Find lemmas with the shortest synonym list (excluding missing ones).

19 Find lemmas where synonyms field includes a synonym with a number (e.g., "22", "38").

20 Find average synonym count per part_of_speech.

CODE:

```python
import pandas as pd
import numpy as np

# Load the dataset
file_path = '/mnt/data/synonyms.csv'
df = pd.read_csv(file_path)

# Make a working copy
data = df.copy()

### 1. Find the total number of missing values in each column.
missing_values = data.isnull().sum()
print("1. Missing Values:\n", missing_values)

### 2. Drop rows with any missing values.
data_clean = data.dropna()
print("\n2. Rows after Dropna:", data_clean.shape[0])

### 3. Find the unique part_of_speech values.
unique_pos = data_clean['part_of_speech'].unique()
print("\n3. Unique Parts of Speech:\n", unique_pos)

### 4. Find the most frequent part_of_speech.
most_frequent_pos = data_clean['part_of_speech'].mode()[0]
print("\n4. Most Frequent POS:", most_frequent_pos)

### 5. Find the number of lemmas associated with each part_of_speech.
lemma_count_per_pos = data_clean.groupby('part_of_speech')['lemma'].count()
print("\n5. Lemma Count Per POS:\n", lemma_count_per_pos)

### 6. Find the top 10 most common lemmas.
top_10_lemmas = data_clean['lemma'].value_counts().head(10)
print("\n6. Top 10 Lemmas:\n", top_10_lemmas)
```

```python
### 7. Find how many lemmas have more than 3 synonyms.
data_clean['num_synonyms'] = data_clean['synonyms'].apply(lambda x: len(str(x).split(';')))
lemmas_more_than_3_synonyms = data_clean[data_clean['num_synonyms'] > 3].shape[0]
print("\n7. Lemmas with >3 Synonyms:", lemmas_more_than_3_synonyms)

### 8. Find the lemma with the highest number of synonyms.
lemma_with_most_synonyms = data_clean.loc[data_clean['num_synonyms'].idxmax()]
print("\n8. Lemma with Most Synonyms:\n", lemma_with_most_synonyms[['lemma', 'num_synonyms']])

### 9. Calculate the average number of synonyms per lemma.
average_synonyms = data_clean['num_synonyms'].mean()
print("\n9. Average Synonyms per Lemma:", average_synonyms)

### 10. List all lemmas that have only one synonym.
lemmas_one_synonym = data_clean[data_clean['num_synonyms'] == 1]['lemma'].tolist()
print("\n10. Sample Lemmas with 1 Synonym:\n", lemmas_one_synonym[:10])

### 11. Find all synonyms containing the word "caliber".
synonyms_with_caliber = data_clean[data_clean['synonyms'].str.contains('caliber', case=False, na=False)]
print("\n11. Sample Synonyms Containing 'caliber':\n", synonyms_with_caliber[['lemma', 'synonyms']].head())

### 12. Replace all hyphens "-" in `lemma` and `synonyms` with spaces.
data_no_hyphen = data_clean.copy()
data_no_hyphen['lemma'] = data_no_hyphen['lemma'].str.replace('-', ' ', regex=False)
data_no_hyphen['synonyms'] = data_no_hyphen['synonyms'].str.replace('-', ' ', regex=False)
print("\n12. Lemma Example After Hyphen Replacement:\n", data_no_hyphen[['lemma', 'synonyms']].head())

### 13. Find all lemmas that are adjectives.
adjective_lemmas = data_clean[data_clean['part_of_speech'] == 'adjective']['lemma'].tolist()
print("\n13. Sample Adjective Lemmas:\n", adjective_lemmas[:10])

### 14. Find the number of unique lemmas.
num_unique_lemmas = data_clean['lemma'].nunique()
print("\n14. Number of Unique Lemmas:", num_unique_lemmas)
```

```python
### 15. Find the number of lemmas whose synonyms list contains duplicates.
def has_duplicates(syn_str):
    syn_list = syn_str.split(';')
    return len(syn_list) != len(set(syn_list))

lemmas_with_duplicate_synonyms = data_clean[data_clean['synonyms'].apply(has_duplicates)].shape[0]
print("\n15. Lemmas with Duplicate Synonyms:", lemmas_with_duplicate_synonyms)

### 16. Num Synonyms Column Exists
print("\n16. Num Synonyms Column Exists:", 'num_synonyms' in data_clean.columns)

### 17. Filter lemmas whose synonyms include themselves.
lemmas_with_self_synonyms = data_clean[data_clean.apply(lambda row: row['lemma'] in row['synonyms'], axis=1)]
print("\n17. Sample Lemmas with Self in Synonyms:\n", lemmas_with_self_synonyms[['lemma', 'synonyms']].head())

### 18. Find lemmas with the shortest synonym list (excluding missing ones).
min_synonyms = data_clean['num_synonyms'].min()
lemmas_with_min_synonyms = data_clean[data_clean['num_synonyms'] == min_synonyms]['lemma'].tolist()
print("\n18. Lemmas with Fewest Synonyms:\n", lemmas_with_min_synonyms[:10])

### 19. Find lemmas where synonyms field includes a synonym with a number.
lemmas_with_numbers_in_synonyms = data_clean[data_clean['synonyms'].str.contains(r'\d', regex=True)]
print("\n19. Sample Lemmas with Numbers in Synonyms:\n", lemmas_with_numbers_in_synonyms[['lemma', 'synonyms']].head())

### 20. Find average synonym count per part_of_speech.
avg_synonyms_per_pos = data_clean.groupby('part_of_speech')['num_synonyms'].mean()
print("\n20. Avg Synonym Count per POS:\n", avg_synonyms_per_pos)

```

OUTPUT:

1. Missing Values

```
lemma            3
part_of_speech   0
synonyms         2
dtype: int64
```

2. Rows after Dropna

```
126,996 rows
```

3. Unique Parts of Speech

```
['adjective', 'noun', 'satellite', 'adverb', 'verb']
```

4. Most Frequent Part of Speech

```
noun
```

5. Lemma Count Per Part of Speech

```
part_of_speech
adjective     4274
adverb        2694
noun         94597
satellite    11750
verb         13681
```

6. Top 10 Lemmas

```
still     5
close     5
clean     5
short     5
double    5
light     5
round     5
better    5
best      5
clear     5
```

7. Number of Lemmas with >3 Synonyms

```
29,179 lemmas
```

8. Lemma with Most Synonyms

```
lemma: passing
num_synonyms: 85
```

9. Average Synonyms per Lemma

```
2.939
```

10. Sample Lemmas with Exactly 1 Synonym

```
['0', '10-membered', '1000th', '101st', '105th', '10th', '110th', '115th', '11th', '120th']
```

11. Sample Synonyms Containing 'Caliber'

Lemma	Synonyms
.22-caliber	.22 caliber;.22 calibre;.22-calibre
.22-calibre	.22 caliber;.22-caliber;.22 calibre
.22 caliber	.22-caliber;.22 calibre;.22-calibre
.22 calibre	.22 caliber;.22-caliber;.22-calibre
.38-caliber	.38 caliber;.38 calibre;.38-calibre

12. Hyphen Replaced in Lemma and Synonyms (Sample)

Lemma	Synonyms
.22 caliber	.22 caliber;.22 calibre;.22 calibre
.22 calibre	.22 caliber;.22 caliber;.22 calibre
.22 caliber	.22 caliber;.22 calibre;.22 calibre
.22 calibre	.22 caliber;.22 caliber;.22 calibre
.38 caliber	.38 caliber;.38 calibre;.38 calibre

13. Sample Adjective Lemmas

```
['.22-caliber', '.22-calibre', '.22 caliber', '.22 calibre', '.38-caliber', '.38-calibre', '.38 caliber', '.38 calibre', '.45-caliber', '.45-calibre']
```

14. Number of Unique Lemmas

```
117,202 unique lemmas
```

15. Number of Lemmas with Duplicate Synonyms

```
3,540 lemmas
```

16. Does 'num_synonyms' Column Exist?

```
True
```

17. Sample Lemmas with Themselves in Their Synonyms

Lemma	Synonyms
a-ok	a-okay
abdominal	abdominal muscle;ab
abducent	abducent nerve;abducens;abducens nerve;nervus abducens
ablative	ablative case
able	capable

18. Sample Lemmas with Fewest Synonyms

```
['0', '10-membered', '1000th', '101st', '105th', '10th', '110th', '115th', '11th', '120th']
```

19. Sample Lemmas Whose Synonyms Contain Numbers

Lemma	Synonyms
.22-caliber	.22 caliber;.22 calibre;.22-calibre
.22-calibre	.22 caliber;.22-caliber;.22 calibre
.22 caliber	.22-caliber;.22 calibre;.22-calibre
.22 calibre	.22 caliber;.22-caliber;.22-calibre
.38-caliber	.38 caliber;.38 calibre;.38-calibre

20. Average Synonym Count Per Part of Speech

```
part_of_speech
adjective    1.45
adverb       2.22
noun         2.59
satellite    3.23
verb         5.71
```