module03-analyzing-text-content-natural-language-processing-30pct/3-3-2-computer-lab.qmd

List of Figures

List of Tables

summarize and visualize data

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os

# Load token_df from the TSV file
input_file_path = '/content/osm-cca-nlp/csv/token_data.tsv'
input_file_path = '/home/sol-nhl/rnd/d/quarto/osm-cca-nlp/csv/token_data.tsv'
token_df = pd.read_csv(input_file_path, sep='\t')

# Filter the DataFrame to keep only rows where the part of speech is 'NOUN'
noun_df = token_df[token_df['token_pos'] == 'NOUN']

# Group by the lemma and count the occurrences of each lemma
lemma_counts = noun_df['token_lemma'].value_counts().reset_index()

# Rename the columns for clarity
lemma_counts.columns = ['lemma', 'count']

# Get the 20 most frequent lemmas
top_lemmas = lemma_counts.head(20)

# Plot the 20 most frequent nouns using Seaborn
plt.figure(figsize=(10, 8))
sns.barplot(x='count', y='lemma', data=top_lemmas, palette='viridis')
plt.title('Top 20 Most Frequent Nouns')
plt.xlabel('Count')
plt.ylabel('Lemma')

# Save the figure to a PNG file
output_file_path = '/content/osm-cca-nlp/fig/token_noun.png'
output_file_path = '/home/sol-nhl/rnd/d/quarto/osm-cca-nlp/fig/token_noun.png'
plt.savefig(output_file_path)

# Display the plot
plt.show()
```