Procesamiento del Lenguaje Natural

Rodrigo S. Cortez Madrigal

Actividad Sumativa 2

Utilizando los textos de Hamlet, Julius Caesar, Othello y Macbeth, tokenizar, lematizar y obtener una matriz B-O-W. Después, obtener la distancia entre las palabras “Caesar” y “Brutus”.

In [1]:

```python
import numpy as np
import pandas as pd
import spacy
from sklearn.feature_extraction.text import CountVectorizer
import os
import plotly
from plotly import graph_objs as go
from plotly import express as px
from plotly.subplots import make_subplots
from sklearn.metrics.pairwise import cosine_similarity, euclidean_distances

import spacy

plotly.offline.init_notebook_mode(connected=False)
```

In [2]:

```python
# Cargar modelo de spaCy
nlp = spacy.load("en_core_web_sm")

# Función para preprocesar documentos
def preprocess_docs(docs, nlp):
    processed_docs = []
    for doc in docs:
        spacy_doc = nlp(doc)
        tokens = [token.lemma_ for token in spacy_doc if not token.is_stop and not token.is_punct]
        processed_docs.append(' '.join(tokens))
    return processed_docs

# Obtener archivos de texto en el directorio actual
files = [file for file in os.listdir('./') if file.endswith('.txt')]

# Leer y procesar los textos
texts = []
for play in files:
    with open(play, 'r', encoding='utf-8') as f:
        text = f.read()
        preprocessed_text = preprocess_docs([text], nlp)
        texts.append(preprocessed_text[0])

print(f'Found {len(texts)} plays')

# Crear la matriz BOW
vectorizer = CountVectorizer()
# Es lo mismo usar TfidfVectorizer(use_idf=False, norm=None)
bow_matrix = vectorizer.fit_transform(texts)
feature_names = vectorizer.get_feature_names_out()

# Crear un DataFrame para inspeccionar la matriz
df = pd.DataFrame(bow_matrix.toarray(), columns=feature_names, index=files)
```

```
Found 4 plays
```

In [7]:

```python
# Calcular matriz de distancias
distances = euclidean_distances(bow_matrix)
```

In [9]:

```python
distances
```

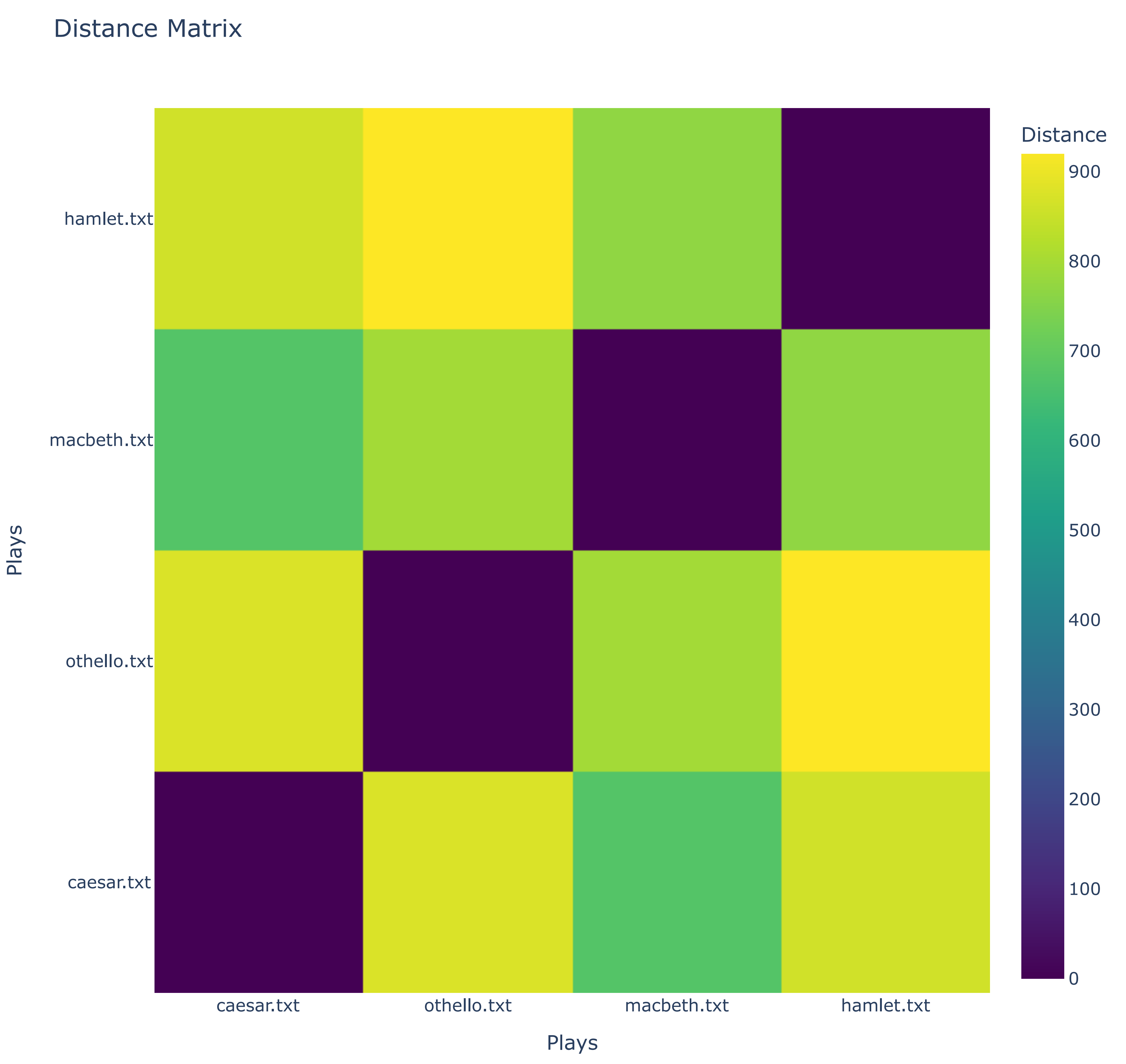
Out[9]:

```
array([[  0.        , 869.84883744, 668.9297422 , 856.49518387],
       [869.84883744,   0.        , 791.44677648, 919.80052185],
       [668.9297422 , 791.44677648,   0.        , 763.99803665],
       [856.49518387, 919.80052185, 763.99803665,   0.        ]])
```

In [10]:

```python
# Plot la matriz BOW

fig = make_subplots(rows=1, cols=1)
fig.add_trace(
    go.Heatmap(
        z=distances,
        x=files,
        y=files,
        colorscale='Viridis',
        colorbar=dict(title='Distance'),
    ),
    row=1, col=1
)
fig.update_layout(
    title='Distance Matrix',
    xaxis_title='Plays',
    yaxis_title='Plays',
    width=800,
    height=800,
)
fig.show()
```


Distance Matrix
Distance
hamlet.txt
macbeth.txt
othello.txt
caesar.txt
Plays
caesar.txt
othello.txt
macbeth.txt
hamlet.txt
Plays