TALENTO TECH

TIC

Lección 3: Lematización

Universidad Tecnológica de Pereira

faceIT

Tiempo de ejecución: 2 horas

PLANTEAMIENTO DE LA SESIÓN	MATERIALES
La lematización, una técnica esencial en el procesamiento del lenguaje natural, se utiliza para reducir las palabras a su forma base, o lema, lo que simplifica el análisis y la interpretación del texto. A diferencia de la tokenización, que divide el texto en unidades más pequeñas como palabras o caracteres, la lematización considera la morfología y el contexto lingüístico para identificar la forma canónica de una palabra. Por ejemplo, verbos conjugados como "corriendo", "corre" y "corrió" se reducirían al lema "correr". Esto resulta fundamental en aplicaciones de donde la precisión y la comprensión del texto son cruciales, como en motores de búsqueda, sistemas de recomendación y análisis de sentimientos.	Conexión a Internet.

Tiempo de ejecución: 2 horas

PLANTEAMIENTO DE LA SESIÓN	MATERIALES
Existen varias herramientas y estrategias para llevar a cabo la lematización de manera efectiva. Una de las bibliotecas más utilizadas en Python para realizar lematización es NLTK (Natural Language Toolkit), que ofrece una amplia gama de funciones y algoritmos de, incluyendo la lematización. Además de NLTK, spaCy es otra biblioteca popular que proporciona capacidades avanzadas de procesamiento de texto, incluida la lematización. Estas herramientas utilizan algoritmos lingüísticos y modelos de aprendizaje automático entrenados en grandes corpus de texto para identificar y asignar los lemas correctos a las palabras en un texto.	

Tiempo de ejecución: 2 horas

PLANTEAMIENTO DE LA SESIÓN	MATERIALES
Al momento de aplicar la lematización, es importante considerar el contexto específico de la tarea y las características del texto. Por ejemplo, en algunos casos puede ser preferible conservar la información morfológica completa, mientras que en otros puede ser necesario simplificar las palabras a sus formas más básicas para facilitar el análisis. Además, la elección del algoritmo de lematización y el modelo de idioma adecuado puede influir en la precisión y el rendimiento del proceso. Por lo tanto, es crucial realizar pruebas y ajustes para encontrar la mejor estrategia de lematización para cada caso de uso específico en el procesamiento de texto.	

La lematización es como un "ajuste fino" de las palabras en el análisis de texto. Imagina que estás leyendo un libro y encuentras palabras como "corriendo" , "corre" y "correrá". En lugar de tratarlas como palabras distintas, la lematización las agrupa bajo su forma base, que en este caso sería "correr". Esto hace que el análisis sea mucho más claro y eficiente, ya que elimina la redundancia y simplifica la comprensión del texto.

La lematización ayuda a interpretar automáticamente los textos de varias maneras:

Reduce la variabilidad léxica: Al agrupar diferentes formas flexionadas bajo un mismo lema, disminuye la variedad de palabras únicas en un corpus textual. Esto simplifica el procesamiento y análisis del texto.

Facilita la identificación de relaciones semánticas: Las palabras con el mismo lema suelen tener un significado semántico relacionado. Agruparlas bajo un lema ayuda a identificar mejor estas relaciones de significado.

Mejora el rendimiento de modelos y algoritmos: Muchos modelos de procesamiento de lenguaje natural, como los utilizados en búsqueda de información, clasificación de textos, etc., funcionan mejor cuando las palabras están lematizadas, ya que pueden capturar mejor los patrones subyacentes.

Permite un mejor manejo de textos en diferentes idiomas: La lematización tiene en cuenta las reglas morfológicas específicas de cada idioma, lo que facilita el procesamiento de textos en distintos idiomas de manera más precisa.

Las herramientas de lematización, como NLTK y spaCy, son como asistentes lingüísticos que aplican reglas y modelos inteligentes para llevar a cabo esta tarea de manera precisa y rápida. Así que, básicamente, la lematización es una herramienta poderosa que ayuda a desentrañar el significado oculto dentro del texto.

Un ejemplo de lematización, puede ser el siguiente:

Palabra	Lematización
running	run
runs	run
ran	run
books	book
better	good
quickest	quick
better	good
cats	cat

Un ejemplo práctico, puede ser el siguiente, que propone una serie de palabras y lematiza algunas de ellas, a su expresión mínima.

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from tabulate import tabulate

# Descargar recursos necesarios de NLTK
nltk.download('punkt')
nltk.download('wordnet')

# Conjunto de comentarios de los clientes
comentarios = [
    "Los zapatos son muy cómodos y bonitos.",
    "La calidad de la tela de la camisa es excelente.",
    "No estoy satisfecho con el tamaño del bolso.",
```

```
    "Los productos de esta tienda son geniales.",
    "El color de la blusa no coincide con la imagen en línea."
]

# Inicializar el lematizador de palabras
lemmatizador = WordNetLemmatizer()

# Tokenizar cada comentario y lematizar las palabras
tabla_datos = []
for comentario in comentarios:
    tokens = word_tokenize(comentario)
    lematizados = [lemmatizador.lemmatize(token) for token in
tokens]
    tabla_datos.append([comentario, " ".join(lematizados)])
```

```python
# Imprimir tabla con comentarios originales y lematizados
print(tabulate(tabla_datos, headers=['Comentario Original',
'Comentario Lematizado'], tablefmt='grid'))
```

```
+-------------------------------------------------------+-------------
-----------------------------------------+

| Comentario Original                                   | Comentario
Lematizado                                |

+=======================================================+=============
=========================================+

| Los zapatos son muy cómodos y bonitos.                | Los zapatos son
muy cómodos y bonito .                    |
```

La calidad de la tela de la camisa es excelente.	La calidad de la tela de la camisa e excelente .
No estoy satisfecho con el tamaño del bolso.	No estoy satisfecho con el tamaño del bolso .

| Los productos de esta tienda son geniales. | Los productos de esta tienda son geniales . |

| El color de la blusa no coincide con la imagen en línea. | El color de la blusa no coincide con la imagen en línea . |

Un ejemplo que ilustra la importancia de la lematización es el análisis de sentimientos en redes sociales. En este contexto, mediante el análisis de los comentarios de los clientes sobre un producto en una plataforma en línea, la lematización de las palabras en los comentarios permite agrupar términos similares y reducir la dimensión del texto. Por ejemplo, todas las formas de la palabra "bueno", como "buenos" y "buenas", se lematizarían a "bueno". Esto proporciona una representación más precisa de la frecuencia de términos relevantes y ayuda a comprender mejor el sentimiento general de los clientes hacia el producto.

Otro ejemplo práctico se encuentra en el análisis de texto en el campo de la atención médica. Aquí, al procesar informes clínicos o notas de pacientes, la lematización resulta útil para identificar términos médicos clave y agrupar conceptos relacionados. Esto facilita el análisis y la extracción de información relevante para el diagnóstico y el tratamiento de los pacientes.

```
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

# Descargar recursos necesarios de NLTK
nltk.download('punkt')
nltk.download('wordnet')

# Crear un DataFrame con los datos de las reseñas
data = {
     'review_en': [
     "One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked...",
     "A wonderful little production. The filming technique is very unassuming...",
```

```
        "I thought this was a wonderful way to spend time on a too
hot summer weekend, sitting in the air conditioned theater..."
        # Agrega más reseñas aquí si es necesario
        ]
}
df = pd.DataFrame(data)
# Inicializar el lematizador de palabras
lemmatizer = WordNetLemmatizer()

# Función para lematizar texto
def lemmatize_text(text):
        tokens = word_tokenize(text)
        lemmatized_tokens = [lemmatizer.lemmatize(token) for token
in tokens]
        return ' '.join(lemmatized_tokens)
```

```
# Aplicar lematización a la columna de reseñas
df['lemmatized_review'] = df['review_en'].apply(lemmatize_text)


# Mostrar las primeras filas del DataFrame con las reseñas
lematizadas
print(df[['review_en', 'lemmatized_review']].head())
```

Este dataset, contiene reseñas en inglés de diversas películas o programas de televisión. Cada entrada en el conjunto de datos representa una reseña escrita por un usuario sobre una obra audiovisual específica. Las reseñas abordan diferentes aspectos de las obras, como trama, actuación, dirección, y pueden expresar opiniones positivas, negativas o neutrales sobre el contenido. También se puede encontrar el dataset en csv.

```
import nltk
nltk.download('wordnet')


import pandas as pd
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer


# Crear un DataFrame de ejemplo con comentarios y sentimientos
simulados
data = pd.DataFrame({
    'review': [
    "This movie is fantastic, I loved it!",
    "Terrible experience, the service was awful.",
    "The food at this restaurant is amazing.",
    "I hated every minute of this movie.",
```

```
        "Great customer service, highly recommended.",
        "The product arrived late and was damaged.",
        "I'm not sure how I feel about this book.",
        "The concert was mediocre, I expected more.",
        "Absolutely brilliant, can't wait to go again!",
        "Disappointing performance, wouldn't go back."
        ],
        'sentiment': ['positive', 'negative', 'positive',
'negative', 'positive',
                      'negative', 'neutral', 'negative', 'positive',
'negative']
```

```
})

# Inicializar el lematizador
lemmatizer = WordNetLemmatizer()

# Tokenización y lematización de los comentarios
data['tokenized_review'] = data['review'].apply(word_tokenize)
data['lemmatized_review'] =
data['tokenized_review'].apply(lambda tokens:
[lemmatizer.lemmatize(token) for token in tokens])

# Filtrar los comentarios por sentimiento
positive_comments = data[data['sentiment'] ==
'positive']['lemmatized_review']
negative_comments = data[data['sentiment'] ==
'negative']['lemmatized_review']
neutral_comments = data[data['sentiment'] ==
'neutral']['lemmatized_review']
```

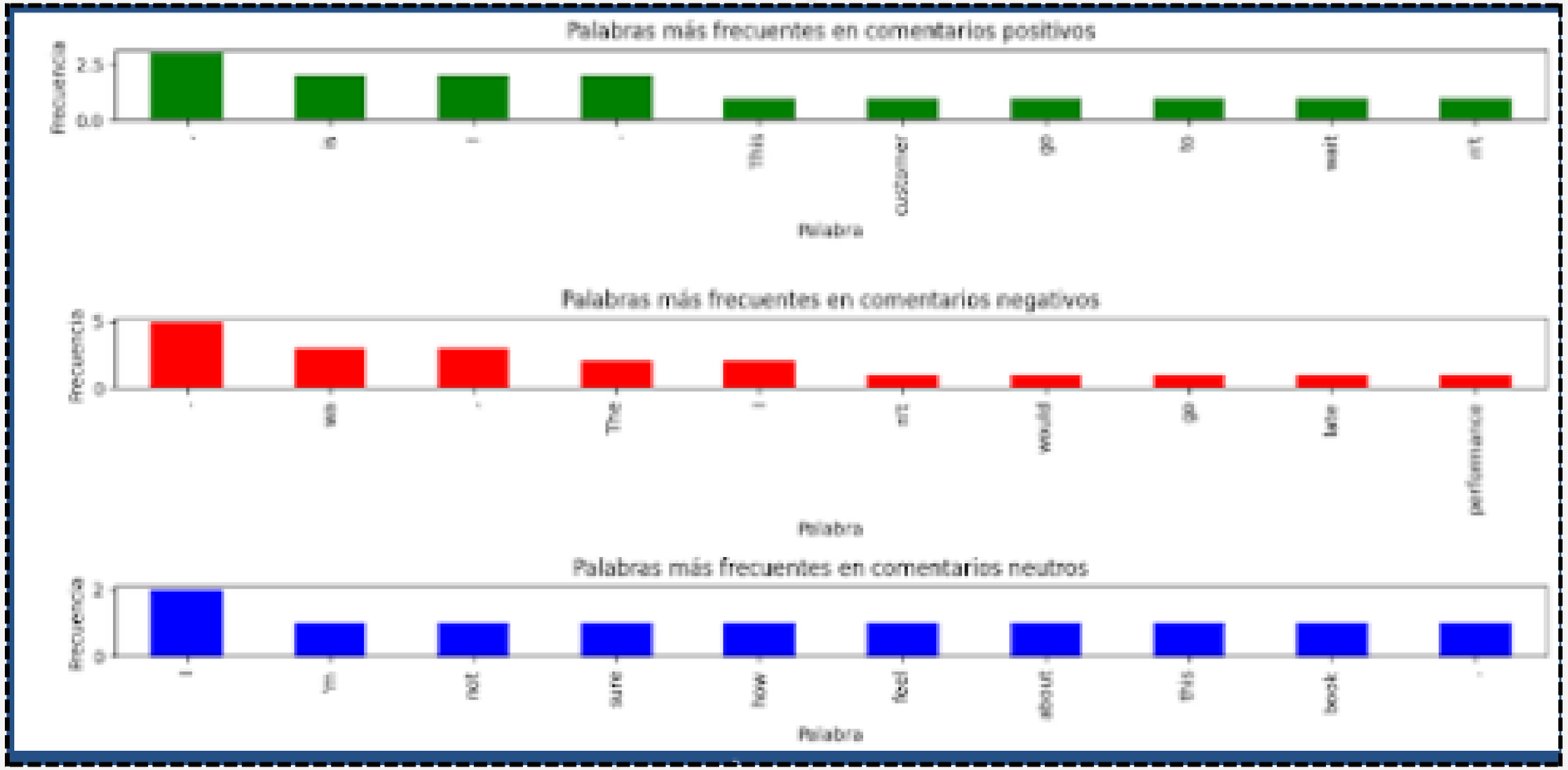
```
# Calcular la frecuencia de palabras en cada conjunto de comentarios
positive_word_freq = pd.Series([word for sublist in positive_comments for word in sublist]).value_counts()
negative_word_freq = pd.Series([word for sublist in negative_comments for word in sublist]).value_counts()
neutral_word_freq = pd.Series([word for sublist in neutral_comments for word in sublist]).value_counts()

# Graficar las palabras más frecuentes en comentarios positivos, negativos y neutros
plt.figure(figsize=(12, 6))

plt.subplot(3, 1, 1)
positive_word_freq.head(10).plot(kind='bar', color='green')
plt.title('Palabras más frecuentes en comentarios positivos')
plt.xlabel('Palabra')
plt.ylabel('Frecuencia')
```

```
plt.subplot(3, 1, 3)
neutral_word_freq.head(10).plot(kind='bar', color='blue')
plt.title('Palabras más frecuentes en comentarios neutros')
plt.xlabel('Palabra')
plt.ylabel('Frecuencia')

plt.tight_layout()
plt.show()
```


Palabras más frecuentes en comentarios positivos
Frecuencia
2.5
0.0
.
to
—
.
this
customer
go
so
wait
n't
Palabra
Palabras más frecuentes en comentarios negativos
Frecuencia
5
0
.
will
.
The
—
n't
would
go
late
performance
Palabra
Palabras más frecuentes en comentarios neutros
Frecuencia
2
0
—
to
not
sure
how
feel
about
this
book
.
Palabra

Actividad

A continuación, se proporciona un ejemplo práctico. El objetivo es identificar y lematizar los términos en la columna correspondiente, aplicando los conceptos aprendidos sobre lematización.

Comentario	Lematización
"Los zapatos son muy cómodos y bonitos."	Los zapatos son muy cómodo y bonito.
"La calidad de la tela de la camisa es excelente."	
"No estoy satisfecho con el tamaño del bolso."	
"Los productos de esta tienda son geniales."	
"El color de la blusa no coincide con la imagen en línea."	