Actividad Práctica para el Aprendizaje de Técnicas del Procesamiento del Habla Basadas en PNL

Introducción

La siguiente actividad práctica está diseñada para que los alumnos de la materia Técnicas del Procesamiento del Habla apliquen los conceptos fundamentales del Procesamiento del Lenguaje Natural (PNL) que han estudiado.
El objetivo principal es proporcionar una experiencia práctica y tangible con las técnicas de Regex (Expresiones Regulares), tokenización (simple y con NLTK y scikit-learn), conteo de palabras y la creación de la Bolsa de Palabras (BoW). A través de esta actividad, los estudiantes podrán comprender mejor cómo se manipula y analiza el texto para extraer información valiosa.

Selección del Conjunto de Datos

Para esta actividad, se propone utilizar un conjunto de datos de reseñas de películas en español. Este tipo de datos es adecuado porque contiene una variedad de expresiones lingüísticas, opiniones y sentimientos, lo que permite ilustrar de manera efectiva las técnicas de PNL. Un recurso útil para este propósito es el **"IMDB Dataset of 50K Movie Reviews (Spanish)"** disponible en Kaggle. Este conjunto de datos contiene 50,000 reseñas de películas traducidas al español, junto con su sentimiento asociado (positivo o negativo). La disponibilidad de este dataset en español lo hace directamente relevante para la práctica de las técnicas aprendidas en la materia. Otra opción podría ser un dataset generado a partir de reseñas de sitios como Tripadvisor o Google Maps.

Paso 1: Descarga e Inspección del Conjunto de Datos

1. **Descarga:** Deberán descargar el conjunto de datos seleccionado (por ejemplo, el archivo CSV del IMDB dataset en español desde Kaggle).

2. **Inspección Inicial:** Utilizando bibliotecas de Python como Pandas, deben cargar el archivo y realizar una inspección inicial. Esto incluye visualizar las primeras filas del dataset, identificar las columnas disponibles (por ejemplo, la reseña en español y el sentimiento), y entender la estructura general de los datos.
   Es importante que se familiaricen con el tipo de texto con el que van a trabajar.
   Por ejemplo, en el IMDB dataset, las columnas relevantes serían **review_es** (la reseña en español) y **sentimiento** (el sentimiento asociado).

Paso 2: Limpieza de Texto con Expresiones Regulares (Regex):

La limpieza del texto es un paso crucial en el preprocesamiento para asegurar que los datos sean consistentes y estén en un formato adecuado para el análisis posterior.
Las expresiones regulares son una herramienta poderosa para realizar esta tarea.

1. **Eliminación de Signos de Puntuación y Caracteres Especiales:**

    a) **Explicación:** Los signos de puntuación y otros caracteres especiales (como **@**, **#**, *****, etc.) generalmente no aportan información semántica relevante para el análisis de texto y pueden interferir con la tokenización y el conteo de palabras. Por lo tanto, es común eliminarlos.

    b) **Aplicación con Regex:** Deben utilizar la biblioteca **re** de Python para definir patrones de expresiones regulares que coincidan con los signos de puntuación y caracteres especiales presentes en las reseñas en español.
    Luego, usarán la función re.sub() para reemplazar estas coincidencias con un espacio en blanco o eliminarlas por completo.

2. **Conversión de Texto a Minúsculas:**

    a) **Explicación:** Convertir todo el texto a minúsculas es una práctica común para asegurar que las palabras se traten de manera uniforme, independientemente de su capitalización. Por ejemplo, "Hola" y "hola" deben considerarse la misma palabra.
    b) **Aplicación:** Deberán realizar esta conversión en la columna de reseñas limpias.

3. **Eliminación de Espacios en Blanco Adicionales:**

    a) **Explicación:** Pueden existir espacios en blanco redundantes entre las palabras o al principio y al final de las reseñas. Estos espacios no aportan valor y pueden ser eliminados para una mayor consistencia.

    b) **Aplicación con Regex:** Se puede utilizar expresiones regulares para reemplazar múltiples espacios en blanco con un solo espacio y eliminar los espacios al principio y al final de la cadena.

4. **Manejo de Acentos (Opcional pero Recomendado):**

    a) **Explicación:** En español, los acentos diacríticos son importantes para distinguir palabras. Sin embargo, para ciertos análisis, puede ser útil normalizar las palabras eliminando o convirtiendo los acentos (por ejemplo, **"á"** a **"a"**).
    Esto puede ayudar a agrupar palabras que son fundamentalmente las mismas

pero se escriben con o sin acento.

b) **Aplicación con Regex (Básico):** Se podrían usar expresiones regulares para reemplazar caracteres acentuados específicos con sus equivalentes sin acento. Sin embargo, para un manejo más robusto, se recomienda mencionar el uso de bibliotecas especializadas como unidecode.

Actividad para los Alumnos: Aplicar estas técnicas de limpieza con Regex a una muestra de las reseñas de películas del dataset descargado. Deben mostrar el texto original y el texto después de cada paso de limpieza.

Paso 3: Tokenización de Texto:

La tokenización es el proceso de dividir un texto en unidades más pequeñas llamadas tokens, que generalmente son palabras o subpalabras. Este paso es fundamental para el análisis posterior, como el conteo de palabras y la creación de la Bolsa de Palabras.

1. **Tokenización Simple:**

   a) **Explicación:** La forma más sencilla de tokenizar es dividir el texto utilizando los espacios en blanco como delimitadores. Esto se puede lograr fácilmente con las cadenas en Python.

2. **Tokenización con NLTK:**

   a) **Instalación y Descarga de Recursos:** Instalar NLTK y descargar los recursos necesarios para el idioma español, específicamente el tokenizador **punkt.**

   b) **Tokenización de Palabras con NLTK:** Demostrar cómo utilizar la tokenización del texto en español.

      - **Manejo de Caracteres Específicos:** Mencionar que NLTK generalmente maneja bien los caracteres específicos del español como "¿" y "¡", separándolos como tokens individuales. Sin embargo, se puede referenciar la discusión sobre posibles ajustes para tokenizadores específicos del español si es necesario.

3. **Tokenización con scikit-learn:**

   a) **Introducción a CountVectorizer:** Presentar la clase CountVectorizer del módulo sklearn.feature_extraction.text.

b) **Tokenización como Parte de la Vectorización:** CountVectorizer no solo tokeniza el texto, sino que también lo convierte en una representación numérica (que se utilizará en el paso de BoW).

c) **Tokenización Básica con CountVectorizer:** Mostrar cómo inicializar y usar CountVectorizer para tokenizar texto en español. Explicar que por defecto, CountVectorizer tokeniza dividiendo por espacios en blanco y signos de puntuación.

d) **Personalización de la Tokenización (Opcional):** Mencionar brevemente que CountVectorizer permite utilizar tokenizadores personalizados si es necesario, por ejemplo, utilizando un tokenizador de NLTK dentro de CountVectorizer.

Actividad para los Alumnos: Tokenizar una muestra de las reseñas limpias utilizando los tres métodos (simple, NLTK y scikit-learn) y comparar los resultados. Deben identificar las diferencias en cómo cada método maneja la puntuación y los caracteres especiales.

Paso 4: Conteo de Palabras:

El conteo de palabras es esencial para comprender la frecuencia de los términos en un texto, lo que puede revelar información importante sobre su contenido.

1. **Métodos para el Conteo de Palabras en Python:** Demostrar diferentes maneras de contar la frecuencia de las palabras en las listas de tokens obtenidas en el paso anterior.

    a) **Método 4.1: Usando un diccionario:** Proporcionar un ejemplo de código en Python para iterar a través de la lista de tokens y almacenar los recuentos de palabras en un diccionario.

    b) **Método 4.2: Usando collections.Counter:** Presentar la clase Counter del módulo collections como una forma más eficiente de contar frecuencias.
    c) Proporcionar un ejemplo de código.

    d) **Método 4.3: Usando CountVectorizer (Implícito):** Señalar que CountVectorizer también cuenta implícitamente la frecuencia de las palabras como parte de su proceso. Mostrar cómo acceder a estos recuentos utilizando el atributo **vocabulary_.**

2. **Comparación de Resultados:** Deberán contar la frecuencia de las palabras para los tokens obtenidos de los diferentes métodos de tokenización (simple, NLTK, scikit-learn) y que comparen las N palabras más frecuentes.

3. **Tabla: Las N Palabras Más Frecuentes:** Incluir una tabla para mostrar las 10-20 palabras más frecuentes obtenidas de cada uno de los tres métodos de tokenización para una muestra de reseña. Esto demostrará visualmente las diferencias en la salida de cada método de tokenización.

Rank	Tokenización Simple	Tokenización NLTK	Tokenización scikit-learn
1	la	la	la
2	de	de	de
3	que	que	que
4	y	y	y
5	el	el	el
6	en	en	en
7	un	un	un
8	película	película	película
9	es	es	es
10	con	con	con
...	...	...	...

Actividad para los Alumnos: Contar la frecuencia de las palabras en una muestra de reseñas utilizando los diferentes métodos y completar una tabla comparativa de las principales palabras.

Paso 5: Creación de la Bolsa de Palabras (BoW):

El modelo de Bolsa de Palabras (Bag of Words or BoW) representa un texto como la colección de sus palabras, sin tener en cuenta la gramática ni el orden de las palabras, pero manteniendo un registro de la frecuencia de cada palabra.[15] Es una técnica fundamental para muchas tareas de clasificación de texto.

1. **Implementación de BoW con CountVectorizer:** Demostrar cómo utilizar CountVectorizer para crear una representación BoW de las reseñas de películas en español.

    a) **Inicialización:** Mostrar cómo inicializar **CountVectorizer.**

    b) **Ajuste y Transformación:** Explicar el método fit() para aprender el vocabulario del corpus y el método transform() para crear la matriz documento-término.

    c) **Obtención de la Matriz de Características:** Mostrar cómo obtener la matriz

dispersa resultante de la representación BoW.

d) **Acceso al Vocabulario:** Explicar cómo acceder al vocabulario (mapeo de palabras a índices) usando **vocabulary_.**

e) **Obtención de los Nombres de las Características:** Mostrar cómo obtener la lista de nombres de las características (palabras) usando **get_feature_names_out()**.

Actividad para los Alumnos: Crear una representación BoW de un subconjunto más grande de las reseñas de películas utilizando **CountVectorizer** y explorar el vocabulario y la matriz resultante. Deben intentar variar los parámetros de CountVectorizer como **ngram_range**.

Aplicaciones Prácticas y Análisis de lo estudiado:

Actividades:

En grupos de 4 alumnos máximo.

Código completo: notebook JupyterLab con script organizado en secciones, conteniendo todos los puntos solicitados sobre la base del dataset "IMDB Dataset of 50K Movie Reviews (Spanish)"

Informe breve (1–3 páginas) que responda:

a) **Comparación de vocabularios:** Comparar el tamaño y el contenido de los vocabularios generados por diferentes métodos de tokenización.

b) **Identificación de Palabras Más Frecuentes:** Identificar las palabras más frecuentes en todo el dataset y discutir su posible significado (¿son palabras de parada, palabras de contenido?).

c) **Impacto de la Limpieza:** Analizar cómo los pasos de limpieza de texto (Regex) afectaron la tokenización y el conteo de palabras.

d) **Análisis de Sentimiento (Extensión):** Si se utilizan las etiquetas de sentimiento, los alumnos podrían explorar la relación entre las frecuencias de las palabras y el sentimiento (¿tienen las reseñas positivas palabras frecuentes diferentes a las reseñas negativas?). Esto podría implicar el uso de la representación BoW como entrada para un clasificador de sentimiento simple.

e) **Evaluación de la Precisión del Modelo (Si se Realiza Análisis de Sentimiento):** Si

los alumnos construyeran un modelo simple de análisis de sentimiento utilizando la representación BoW, discutir cómo evaluar su precisión utilizando métricas como **accuracy**, **precisión**, **recall** y **F1-score**.

Puntos de Discusión:

Informe breve (1–3 páginas) que responda:

a) ¿Cuáles son las ventajas y desventajas de los diferentes métodos de tokenización?

b) ¿Cómo impacta la limpieza de texto en los resultados de las tareas de PNL?

c) ¿Cuáles son las limitaciones del modelo de Bolsa de Palabras?

d) ¿En qué escenarios del mundo real son útiles estas técnicas?

Conclusiones:

Esta actividad práctica proporciona una base sólida para que los alumnos comprendan y apliquen las técnicas fundamentales del Procesamiento del Lenguaje Natural. Al trabajar con un conjunto de datos reales de reseñas en español, los estudiantes pueden observar de primera mano cómo las expresiones regulares, la tokenización, el conteo de palabras y la Bolsa de Palabras son pasos esenciales en el análisis de texto.
La exploración de diferentes herramientas y métodos, así como la reflexión sobre las aplicaciones prácticas, fomentarán una comprensión más profunda de estos conceptos y su relevancia en el campo del procesamiento del habla y el lenguaje natural.

Obras citadas

1. IMDB Dataset of 50K Movie Reviews (Spanish) - Kaggle, fecha de acceso: mayo 13, 2025, https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish/data

2. IMDB Dataset of 50K Movie Reviews (Spanish) | Kaggle, fecha de acceso: mayo 13, 2025, https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish

3. Dataset for sentiment analysis in Spanish - Zenodo, fecha de acceso: mayo 13, 2025, https://zenodo.org/records/6425687

4. Cleaning Text with python and re - Stack Overflow, fecha de acceso: mayo 13, 2025,

https://stackoverflow.com/questions/55187374/cleaning-text-with-python-and-re

5. Using Regular Expressions for Data Cleaning in Python - Noble Desktop, fecha de acceso: mayo 13, 2025, https://www.nobledesktop.com/learn/python/using-regular-expressions-for-data-cleaning-in-python

6. Sentiment analysis in Spanish - Manugarri's blog, fecha de acceso: mayo 13, 2025, https://blog.manugarri.com/sentiment-analysis-in-spanish/

7. NLTK Tokenizing y Python - José Antonio Mora - WordPress.com, fecha de acceso: mayo 13, 2025, https://jantoniomora.wordpress.com/2017/06/26/nltk-tokenizing-y-python/

8. NLTK: tus primeros pasos con Procesamiento del Lenguaje Natural - Adictos al trabajo, fecha de acceso: mayo 13, 2025, https://adictosaltrabajo.com/2023/07/27/nltk-python/

9. Regex parser for a Spanish text - python - Stack Overflow, fecha de acceso: mayo 13, 2025, https://stackoverflow.com/questions/63882750/regex-parser-for-a-spanish-text

10. python 3.x - Spanish word tokeniser - Stack Overflow, fecha de acceso: mayo 13, 2025, https://stackoverflow.com/questions/41337363/spanish-word-tokeniser

11. Sample usage for tokenize - NLTK, fecha de acceso: mayo 13, 2025, https://www.nltk.org/howto/tokenize.html

12. word_tokenizer and Spanish · Issue #1558 · nltk/nltk - GitHub, fecha de acceso: mayo 13, 2025, https://github.com/nltk/nltk/issues/1558

13. How to tokenize non english language text in nlp - ProjectPro, fecha de acceso: mayo 13, 2025, https://www.projectpro.io/recipes/tokenize-non-english-language-text

14. CountVectorizer — scikit-learn 1.6.1 documentation, fecha de acceso: mayo 13, 2025, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

15. 34. Bag-of-Words Using Scikit Learn - GitHub, fecha de acceso: mayo 13, 2025, https://pages.github.rpi.edu/kuruzj/website_introml_rpi/notebooks/08-intro-nlp/03-scikit-learn-text.html

16. Sentiment Analysis Using Bag-of-Words — ENC2045 Computational Linguistics, fecha de acceso: mayo 13, 2025, https://alvinntnu.github.io/NTNU_ENC2045_LECTURES/nlp/ml-sklearn-classification.html

17. Efficiently count word frequencies in python - Stack Overflow, fecha de acceso: mayo 13, 2025, https://stackoverflow.com/questions/35857519/efficiently-count-word-frequencies-in-python

18. Find frequency of each word in a string in Python - GeeksforGeeks, fecha de acceso: mayo 13, 2025, https://www.geeksforgeeks.org/find-frequency-of-each-word-in-a-string-in-python/

19. python - Count frequency of words in a list and sort by frequency - Stack Overflow, fecha de acceso: mayo 13, 2025, https://stackoverflow.com/questions/20510768/count-frequency-of-words-in-a-list-and-sort-by-frequency

20. A simple Python script to count and rank the frequency of words in a text file, e.g. for verifying that you are kerning important pairs for specific content · GitHub, fecha de acceso: mayo 13, 2025, https://gist.github.com/arrowtype/1cbddcfe2fac1b0b6c8b547e7f561986

21. Finding the frequency of words in a text document - Python discussion forum, fecha de acceso: mayo 13, 2025, https://discuss.python.org/t/finding-the-frequency-of-words-in-a-text-document/52323

22. Working With Text Data — scikit-learn 1.4.2 documentation, fecha de acceso: mayo 13, 2025, https://scikit-learn.org/1.4/tutorial/text_analytics/working_with_text_data.html

23. sklearn.feature_extraction.text.TfidfVectorizer — documentación de scikit-learn - 0.24.1, fecha de acceso: mayo 13, 2025, https://qu4nt.github.io/sklearn-doc-es/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

24. sentiment-analysis-spanish - PyPI, fecha de acceso: mayo 13, 2025,

https://pypi.org/project/sentiment-analysis-spanish/

25. Sentiment Classification of Spanish Reviews: An Approach based on Feature Selection and Machine Learning Methods - Semantic Scholar, fecha de acceso: mayo 13, 2025, https://www.semanticscholar.org/paper/Sentiment-Classification-of-Spanish-Reviews%3A-An-on-Salas-Z%C3%A1rate-Paredes-Valverde/3ab1235519d06b0fec19bfb84ff4def49a698616

26. VerificadoProfesional/SaBERT-Spanish-Sentiment-Analysis - Hugging Face, fecha de acceso: mayo 13, 2025, https://huggingface.co/VerificadoProfesional/SaBERT-Spanish-Sentiment-Analysis

27. Sentiment analysis en críticas de películas mediante regresión logística, fecha de acceso: mayo 13, 2025, https://elmundodelosdatos.com/sentiment-analysis-criticas-peliculas-regresion-logistica/

28. Análisis de sentimientos de reseñas de películas (IMDb Dataset - 50k reseñas) - Toolify.ai, fecha de acceso: mayo 13, 2025, https://www.toolify.ai/es/ai-news-es/anlisis-de-sentimientos-de-reseas-de-pelculas-imdb-dataset-50k-reseas-2475369

29. Análisis de Sentimiento a nivel de documento en críticas de cine en español - Research in Computing Science, fecha de acceso: mayo 13, 2025, https://rcs.cic.ipn.mx/2020_149_8/Analisis%20de%20Sentimiento%20a%20nivel%20de%20documento%20en%20criticas%20de%20cine%20en%20espanol.pdf

30. Análisis de sentimiento de textos basado en opiniones de películas usando algoritmos de aprendizaje computacional, fecha de acceso: mayo 13, 2025, https://openaccess.uoc.edu/bitstream/10609/132328/7/jchulillaTFG0621memoria.pdf

31. Análisis de Sentimiento en Reseñas de Películas: Proyecto de Verano 2016 - Toolify.ai, fecha de acceso: mayo 13, 2025, https://www.toolify.ai/es/ai-news-es/anlisis-de-sentimiento-en-reseas-de-pelculas-proyecto-de-verano-2016-1748089

32. SentiMovie : Análisis de sentimiento de las reseñas de películas - RIA-UTN, fecha de acceso: mayo 13, 2025, https://ria.utn.edu.ar/items/623d8788-fb47-445c-acb9-5ef9e433e5d2

33. Top 7 Metrics to Evaluate Sentiment Analysis Models - Focal, fecha de acceso: mayo 13, 2025, https://www.getfocal.co/post/top-7-metrics-to-evaluate-sentiment-analysis-models

34. How to evaluate the performance of a sentiment analysis model? - Tencent Cloud, fecha de acceso: mayo 13, 2025, https://www.tencentcloud.com/techpedia/106761

35. Sentiment Accuracy: Explaining the Baseline and How to Test It - Lexalytics, fecha de acceso: mayo 13, 2025, https://www.lexalytics.com/blog/sentiment-accuracy-baseline-testing/

36. How to measure the accuracy of your sentiment analysis results? - ServiceNow, fecha de acceso: mayo 13, 2025, https://www.servicenow.com/community/performance-analytics-blog/how-to-measure-the-accuracy-of-your-sentiment-analysis-results/ba-p/2269161

37. What is sentiment analysis accuracy and why does it matter? - Neople, fecha de acceso: mayo 13, 2025, https://www.neople.io/glossary/sentiment-analysis-accuracy

38. 2.7 Evaluating a Sentiment Analysis Model — Practical NLP with Python - NLPlanet, fecha de acceso: mayo 13, 2025, https://www.nlplanet.org/course-practical-nlp/02-practical-nlp-first-tasks/07-evaluate-sentiment-analysis