PROCESAMIENTO DEL HABLA

Data Dinasty

EVIDENCIA 2

TSCDIA - ISPC

Procesamiento de Texto y Análisis de Sentimiento en Reseñas de IMDB

Índice

Informe

a) Comparación de Vocabularios

Se utilizaron tres métodos de tokenización para procesar las reseñas:

- **Tokenización Simple**: divide el texto por espacios. Es rápida pero imprecisa, ya que conserva signos de puntuación adheridos a las palabras.
- **Tokenización con NLTK (TweetTokenizer)**: segmenta con mayor precisión, separando signos y palabras, ideal para textos informales.
- **Tokenización con CountVectorizer (scikit-learn)**: genera un vocabulario único ordenado y elimina duplicados, aunque no conserva el orden ni los signos de puntuación.

Tamaños del vocabulario:

- Tokenización simple y NLTK producen listas de tokens con repeticiones.
- CountVectorizer genera un vocabulario con ≈ 28,000 palabras únicas (estimado).

Conclusión: CountVectorizer es más útil para tareas de modelado. TweetTokenizer es más adecuado para análisis lingüístico detallado.

b) Identificación de Palabras Más Frecuentes

Se identificaron las palabras más frecuentes usando tres enfoques: diccionarios, collections.Counter, y CountVectorizer. Las más comunes incluyen:

- "la", "fue", "muy", "película", "me", "buena", "gustó".

Análisis semántico:

- Muchas son **palabras de parada** (stopwords), como "la", "fue", "muy", que no aportan significado específico.
- También aparecen **palabras de contenido** como "película", "actuación", "historia".

Eliminar stopwords puede ayudar a enfocarse en términos con mayor valor informativo.

c) Impacto de la Limpieza (Regex)

Se aplicaron varias etapas de limpieza:

1. Eliminación de signos de puntuación.
2. Conversión a minúsculas.
3. Normalización de espacios.
4. Eliminación de acentos (opcional).

Impacto:

- Mejoró la consistencia del texto.
- Redujo la variabilidad (por ejemplo, "Película" y "película" ahora se consideran iguales).
- Facilitó el conteo de palabras y disminuyó ruido semántico.

Conclusión: La limpieza con regex es fundamental para un análisis preciso y consistente.

d) Análisis de Sentimiento (Extensión)

Utilizando las etiquetas de sentimiento, se entrenó un modelo BoW + clasificador. Se observó que:

- Las **reseñas positivas** contenían con más frecuencia palabras como "excelente", "increíble", "maravillosa".

- Las **negativas** incluían palabras como “aburrida”, “mala”, “terrible”.

Esto demuestra que las frecuencias de ciertas palabras pueden estar correlacionadas con la polaridad del sentimiento.

e) Evaluación de la Precisión del Modelo

Se entrenó un modelo de clasificación simple (BoW + modelo supervisado). Para evaluar el rendimiento, se utilizaron las métricas:

- **Accuracy**: proporción total de predicciones correctas.
- **Precision**: qué tan precisas son las predicciones positivas.
- **Recall**: qué proporción de los positivos reales se identificaron.
- **F1-score**: equilibrio entre precisión y recall.

Resultados esperados: Accuracy en torno al 80–85%, dependiendo de la división de datos y parámetros.

Puntos de Discusión

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

Método	Ventajas	Desventajas
Tokenización Simple	- Muy rápida y fácil de implementar. - Requiere pocos recursos.	- No distingue bien la puntuación. - No maneja contracciones o palabras compuestas. - Poco adecuada para textos informales.
NLTK TweetTokenizer	- Más precisa. - Maneja signos y tokens especiales. - Adaptada para textos informales.	- Más lenta. - Requiere instalación y configuración. - Puede generar tokens poco útiles (signos aislados).
CountVectorizer	- Integra tokenización y vectorización. - Genera vocabulario único. - Permite configurar filtros y n-gramas.	- No conserva el orden ni contexto. - Tokenización básica comparada con librerías especializadas. - No para análisis lingüístico detallado.

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

Impacto	Descripción
Reducción del ruido	Elimina caracteres especiales, signos, y espacios innecesarios, mejorando la calidad del texto.
Normalización	Homogeniza palabras (minúsculas, sin acentos), evitando variabilidad innecesaria.
Mejora del vocabulario	Reduce el tamaño del vocabulario y facilita el aprendizaje de modelos más robustos.
Incremento de precisión	Facilita la tokenización y conteo correcto, mejorando la calidad de representaciones como BoW.
Posibles riesgos	Limpieza excesiva puede eliminar información útil o introducir sesgos.

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

Limitación	Explicación
Pérdida de orden/contexto	No considera posición ni relaciones entre palabras, lo que afecta la interpretación del texto.
Sensibilidad a ruido	Stopwords o palabras irrelevantes pueden dominar la representación si no se filtran.
Alta dimensionalidad	Gran tamaño del vocabulario genera vectores dispersos y costosos de procesar.
No captura semántica	Trata sinónimos y palabras distintas como completamente diferentes.
No maneja polisemia	Una palabra con varios significados es representada igual en todos los contextos.

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

Escenario	Aplicación
Análisis de opiniones	Evaluar satisfacción y aspectos positivos/negativos en productos y servicios.
Monitoreo en redes sociales	Detectar tendencias, opinión pública y gestionar crisis reputacionales.
Filtros antispam y moderación	Identificar mensajes no deseados o inapropiados en plataformas digitales.
Sistemas de recomendación	Clasificar contenido según el sentimiento expresado por usuarios.
Atención al cliente automatizada	Clasificar consultas para mejorar la eficiencia y asignación en soporte.
Investigación de mercado	Analizar términos y lenguaje para diseñar campañas de marketing más efectivas.