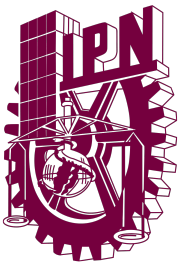

Instituto Politécnico Nacional

Escuela Superior de Cómputo

Práctica 2: Vectorización de Documentos

R E P O R T E

Materia:

Tecnologías de Lenguaje Natural

presenta:

Luis Fernando Rodríguez-Domínguez

Profesor:

Ituriel Enrique Flores Estrada

Objetivo:

Documentar y evidenciar el proceso de vectorización de documentos mediante técnicas de normalización y representación numérica.

Índice general

Capítulo 1

Introducción

El presente reporte documenta la realización de la Práctica 2 de la asignatura de Tecnologías de Lenguaje Natural, en la cual se aborda la vectorización de documentos a partir de un proceso de normalización y diferentes técnicas de representación numérica. En esta práctica se implementa un pipeline que integra diversas etapas de preprocesamiento de texto, con el objetivo de transformar el contenido original de los documentos en un formato que pueda ser analizado y comparado mediante técnicas de vectorización [1]. Para ello, se han aplicado métodos como la conversión a minúsculas, la eliminación de signos de puntuación y stop words, seguidos de distintas combinaciones de lematización, stemming y POS-tagging [2]. Cada uno de estos procesos tiene un impacto directo en la calidad y precisión del corpus final, permitiendo evaluar y contrastar las diferencias que cada técnica produce sobre el contenido semántico del texto.

La primera parte del proceso se centra en la normalización de los textos, en la que se ha considerado una variedad de técnicas (desde la simple lematización sin POS-tagging hasta combinaciones con stemming) para determinar cuál ofrece una mejor preservación de la información semántica [3]. La decisión final se tomó basándose en el análisis comparativo entre las opciones, justificando la selección del método que optimiza la integridad del contenido sin incurrir en una reducción excesiva del vocabulario original.

Posteriormente, se procede a extraer el vocabulario único y se genera un histograma que evidencia la frecuencia de los términos presentes en el corpus normalizado. Esta etapa es fundamental, ya que permite identificar y visualizar la distribución de palabras, estableciendo las bases para la siguiente fase del proyecto: la vectorización de documentos. Para la representación numérica, se han implementado técnicas como One Hot Encoding, Term Count, Probabilidad del Término (TF) y TF-IDF, cada una ofreciendo una perspectiva diferente sobre la relevancia y la frecuencia de los términos dentro del corpus [4, 5].

El desarrollo de este proceso no solo enfatiza la importancia de una correcta normalización del texto, sino que también destaca cómo la selección y el orden de aplicación de cada técnica de preprocesamiento afectan directamente a los resultados de la vectorización. Asimismo, se discuten las implicaciones de utilizar cada técnica en términos de precisión, reducción de dimensionalidad y capacidad para capturar la esencia semántica del documento.

Capítulo 2

Desarrollo

2.1. Parte 1: Normalización de Documentos

2.1.1. Objetivo

El objetivo de esta sección es aplicar y comparar diversas técnicas de normalización sobre un conjunto de documentos clínicos, evidenciando los resultados obtenidos en cada etapa del proceso.

2.1.2. Metodología y Pipeline de Normalización

Para abordar la normalización de los textos se implementó un pipeline estructurado en varias etapas. Cada etapa se ejecutó sobre el conjunto de documentos originales (ver Tabla 2.1), utilizando las siguientes técnicas:

1. **Texto Limpio:** Conversión de todas las palabras a minúsculas, eliminación de stop words y signos de puntuación. Este paso inicial garantiza que el análisis no se vea afectado por diferencias en mayúsculas o por caracteres irrelevantes.

2. **Lematización Simple (sin POS-tagging):** Se aplicó la función de lematización sin el uso de etiquetas gramaticales, transformando las palabras a su forma canónica y preservando la semántica del texto.

3. **Stemming:** Se redujeron las palabras a sus raíces mediante el algoritmo SnowballStemmer, lo que disminuye la variabilidad léxica pero puede truncar parte del significado original.

4. **Lematización seguida de Stemming:** Se ejecutó primero la lematización simple y, posteriormente, se aplicó el stemming. Esta combinación permite evaluar si la aplicación conjunta mejora la homogeneización o si el efecto del stemming prevalece.

5. **Stemming seguido de Lematización:** Se invierte el orden, aplicando el stemming antes que la lematización, para analizar si el cambio en la secuencia impacta en la preservación semántica.

6. **POS-Tagging seguido de Lematización:** Incorporando información gramatical mediante POS-tagging antes de la lematización, se pretende obtener una transformación más precisa a la forma canónica.

7. **POS-Tagging seguido de Stemming:** Se evalúa también la opción de aplicar el stemming tras asignar etiquetas gramaticales, para comparar con el método de stemming puro.

Cada uno de estos métodos se implementó en Python y se ejecutó sobre los documentos clínicos (Tabla 2.1). Como evidencia se muestran los resultados después de cada etapa, permitiendo comparar cómo varía el corpus según la técnica aplicada.

Doc. ID	Clinical Statement (Before pre-processing)
1	Metastasis pancreatic cancer. Acute renal failure, evaluate for hydronephrosis or obstructive uropathy.
2	Pancreatic cancer with metastasis. Jaundice with transaminitis, evaluate for obstruction process.
3	Breast cancer. Pancreatitis. No output from enteric tube, assess tube.

Tabla 2.1: Documentos para analizar.

2.1.3. Resultados de la Normalización

A continuación, se presenta un resumen de los resultados obtenidos:

a) Texto Limpio

Los documentos se tokenizaron y se eliminaron stop words y signos de puntuación. Por ejemplo, el Documento 1 quedó representado como:

```
['metastasis', 'pancreatic', 'cancer', 'acute', 'renal', '
   failure', 'evaluate', 'hydronephrosis', 'obstructive', '
   uropathy']
```

b) Lematización Simple

La aplicación de la lematización sin POS-tagging mostró que la mayoría de los términos ya se encontraban en su forma base. Se eligió esta opción para la vectorización, ya que preserva la integridad semántica. Los resultados fueron idénticos al paso de texto limpio, confirmando la adecuación del corpus.

c) Stemming

El proceso de stemming redujo significativamente la variabilidad de los términos, truncando palabras como "pancreatic" a "pancreat". Aunque esta técnica reduce la dimensionalidad, puede distorsionar la semántica original de ciertos términos.

d) Lematización → Stemming y e) Stemming → Lematización

Ambas combinaciones resultaron en efectos similares al stemming puro, evidenciando que la reducción agresiva impuesta por el stemming domina el proceso, sin aportar mejoras significativas en la preservación de la forma canónica.

f) POS-Tagging → Lematización y g) POS-Tagging → Stemming

La opción f) produjo resultados idénticos a la lematización simple, pero con la ventaja adicional de incorporar información gramatical, lo que en casos más complejos podría resultar beneficioso. La opción g) replicó los resultados del stemming puro.
Elección del Corpus: Aunque se evaluaron las opciones b) a e), se optó por la lematización simple (opción b)) para la vectorización, ya que preserva la forma completa de las palabras sin truncamientos, manteniendo la integridad semántica necesaria para un análisis robusto.

2.1.4. Evidencias Visuales

Para complementar el análisis, se generó el vocabulario extraído y un histograma de términos únicos a partir de los documentos normalizados. La Figura 2.1 muestra el histograma correspondiente, donde se evidencia la distribución de los tokens.

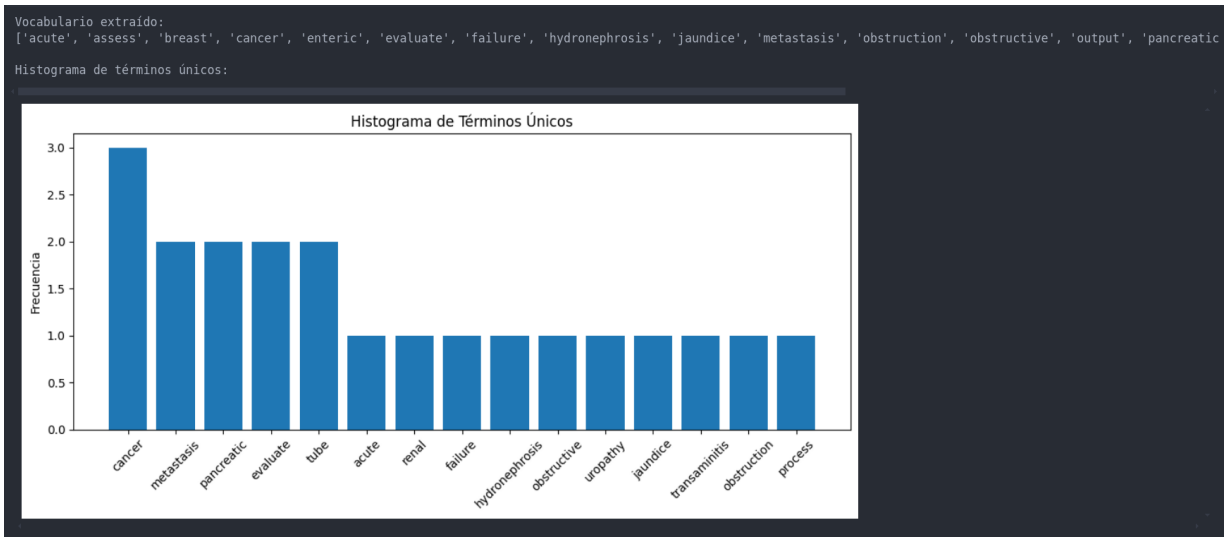

Figura 2.1: Histograma de términos únicos extraídos del corpus normalizado.

2.2. Parte 2: Vectorización de Documentos

2.2.1. Objetivo

El objetivo de esta sección es transformar el corpus normalizado (opción b: lematización simple) en representaciones numéricas mediante diversas técnicas de vectorización,

facilitando el análisis cuantitativo del contenido textual.

2.2.2. Metodología de Vectorización

Se implementaron cuatro técnicas de vectorización sobre el corpus seleccionado:

1. **One Hot Encoding (Term Presence):** Se genera una matriz binaria donde cada posición indica la presencia (1) o ausencia (0) de un término en cada documento.
2. **Term Count Representation:** Se cuenta la frecuencia de cada término en cada documento, proporcionando información detallada sobre la ocurrencia de cada token.
3. **Probabilidad del Término (TF):** Se calcula la frecuencia relativa de cada término en el documento, dividiendo la cantidad de ocurrencias por el total de tokens.
4. **TF-IDF:** Se aplica la técnica TF-IDF que combina la frecuencia del término en el documento (TF) y su importancia en el corpus (IDF). Esto permite resaltar términos relevantes que son menos frecuentes en el conjunto total.

Cada técnica se implementó utilizando las funciones de vectorización desarrolladas en Python, y se generaron las respectivas matrices y vocabularios. Los resultados se presentan a continuación.

2.2.3. Resultados de la Vectorización

a) One Hot Encoding

- **Vocabulario:** `['acute', 'assess', 'breast', 'cancer', 'enteric', 'evaluate', 'failure'`
- **Matriz de Documentos:** Cada documento se representa mediante un vector binario que indica la presencia o ausencia de los términos del vocabulario. Por ejemplo, el Documento 1 presenta:

$$[1,\ 0,\ 0,\ 1,\ 0,\ 1,\ 1,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1]$$

b) Term Count Representation

- **Vocabulario:** (Idéntico al de One Hot Encoding)
- **Matriz de Frecuencias:** Se observa que la mayoría de los términos aparecen una sola vez en cada documento, excepto casos específicos (por ejemplo, el término "tube" en el Documento 3 aparece dos veces).

c) Probabilidad del Término (TF)

- **Matriz TF:** Se normaliza la frecuencia de cada término dividiendo su conteo por el total de tokens del documento. Esto permite comparar la relevancia de los términos entre documentos de distintas longitudes.

d) TF-IDF

- **Vocabulario:** (Mismo conjunto de términos utilizado en las representaciones anteriores)
- **Matriz TF-IDF:** Se muestran los valores de TF-IDF para cada término en cada documento, lo que resalta la importancia relativa de términos específicos. Por ejemplo, en el Documento 1, términos como "acute" y "hydronephrosis" tienen un peso elevado, indicando su relevancia dentro del contexto.

2.2.4. Evidencias Visuales

Se generaron gráficos para visualizar las representaciones vectorizadas. Las Figuras 2.2, 2.3, 2.4 y 2.5 muestran, respectivamente, el vocabulario y las matrices correspondientes a cada técnica.

```
One Hot Encoding:
Vocabulario: ['acute' 'assess' 'breast' 'cancer' 'enteric' 'evaluate' 'failure'
 'hydronephrosis' 'jaundice' 'metastasis' 'obstruction' 'obstructive'
 'output' 'pancreatic' 'pancreatitis' 'process' 'renal' 'transaminitis'
 'tube' 'uropathy']
Vectores de documentos:
Documento 0: [1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1]
Documento 1: [0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0]
Documento 2: [0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0]
```

Figura 2.2: One Hot Encoding: Vocabulario y representación binaria de documentos.

```
Term Count Representation:
Vocabulario: ['acute' 'assess' 'breast' 'cancer' 'enteric' 'evaluate' 'failure'
 'hydronephrosis' 'jaundice' 'metastasis' 'obstruction' 'obstructive'
 'output' 'pancreatic' 'pancreatitis' 'process' 'renal' 'transaminitis'
 'tube' 'uropathy']
Vectores de documentos:
Documento 0: [1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1]
Documento 1: [0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0]
Documento 2: [0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 2, 0]
```

Figura 2.3: Representación por Term Count: Frecuencia de términos en cada documento.

```
Matriz de Probabilidad del Término (TF):
Vectores de documentos:
Documento 0: [0.1, 0.0, 0.0, 0.1, 0.0, 0.1, 0.1, 0.1, 0.0, 0.1, 0.0, 0.1, 0.0, 0.1, 0.0, 0.0, 0.1, 0.0, 0.0, 0.1]
Documento 1: [0.0, 0.0, 0.0, 0.125, 0.0, 0.125, 0.0, 0.0, 0.125, 0.125, 0.125, 0.0, 0.0, 0.125, 0.0, 0.125, 0.0, 0.125, 0.0, 0.0]
Documento 2: [0.0, 0.125, 0.125, 0.125, 0.125, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.125, 0.0, 0.125, 0.0, 0.0, 0.0, 0.25, 0.0]
```

Figura 2.4: Matriz de Probabilidad del Término (TF).

```
TF-IDF Representation:
Vocabulario: ['acute' 'assess' 'breast' 'cancer' 'enteric' 'evaluate' 'failure'
 'hydronephrosis' 'jaundice' 'metastasis' 'obstruction' 'obstructive'
 'output' 'pancreatic' 'pancreatitis' 'process' 'renal' 'transaminitis'
 'tube' 'uropathy']
Vectores de documentos:
Documento 0: [0.352, 0.0, 0.0, 0.208, 0.0, 0.267, 0.352, 0.352, 0.0, 0.267, 0.0, 0.352, 0.0, 0.267, 0.0, 0.0, 0.352, 0.0, 0.0, 0.352]
Documento 1: [0.0, 0.0, 0.0, 0.239, 0.0, 0.308, 0.0, 0.0, 0.405, 0.308, 0.405, 0.0, 0.0, 0.308, 0.0, 0.405, 0.0, 0.405, 0.0, 0.0]
Documento 2: [0.0, 0.327, 0.327, 0.193, 0.327, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.327, 0.0, 0.327, 0.0, 0.0, 0.0, 0.654, 0.0]
```

Figura 2.5: TF-IDF: Representación que pondera la relevancia de cada término en el corpus.

2.3. Parte 3: Análisis y Conclusiones de la Vectorización

2.3.1. Análisis Comparativo

El análisis de los resultados muestra diferencias significativas entre las técnicas de vectorización:

- **One Hot Encoding vs. Term Count:** Mientras que One Hot Encoding solo indica la presencia de términos, la representación por conteo ofrece información sobre la frecuencia, permitiendo una comprensión más detallada de la importancia relativa de cada palabra en el documento.
- **TF vs. TF-IDF:** La normalización a través de la probabilidad del término (TF) facilita la comparación entre documentos de diferente tamaño. Sin embargo, la técnica TF-IDF añade un nivel de ponderación que reduce la influencia de términos muy comunes, resaltando aquellos que son más informativos para cada documento.

2.3.2. Impacto del Pipeline de Normalización

La elección de la lematización simple (opción b)) como corpus de partida se fundamentó en su capacidad para preservar la integridad semántica de los documentos. Las evidencias obtenidas indican que, a diferencia del stemming, la lematización evita la reducción excesiva de los términos, lo cual es crucial para mantener la relevancia de la información en las fases de vectorización.

2.3.3. Síntesis

En síntesis, la práctica demuestra que:

1. La normalización del texto es un paso esencial para mejorar la calidad del análisis, ya que reduce la dispersión léxica y consolida la información.

2. La elección de la técnica de normalización impacta directamente en la eficacia de la vectorización; en este caso, la lematización simple ofrece un balance adecuado entre precisión y preservación semántica.

3. Las distintas técnicas de vectorización (One Hot Encoding, Term Count, TF y TF-IDF) proporcionan perspectivas complementarias del contenido textual, siendo cada una útil dependiendo del objetivo del análisis.

Este desarrollo permite no solo evidenciar el proceso de normalización y vectorización, sino también fundamentar la elección metodológica basada en resultados empíricos y comparativos, ofreciendo una base sólida para aplicaciones futuras en el análisis de datos textuales.

Nota: Todos los gráficos y evidencias presentados se obtuvieron a partir de la ejecución de scripts en Python, los cuales se encuentran documentados en el código fuente adjunto.

Capítulo 3

Conclusiones Generales

La presente práctica abordó aspectos fundamentales del procesamiento de lenguaje natural, enfocándose en la normalización de textos y en la vectorización de documentos clínicos. A continuación, se sintetizan las conclusiones derivadas del análisis y de los resultados obtenidos a lo largo del desarrollo de la práctica.

Parte 1: Normalización de Textos y Selección del Corpus

En esta sección se implementaron diversas técnicas de preprocesamiento, que incluyeron la conversión a minúsculas, eliminación de stop words, eliminación de signos de puntuación y, posteriormente, diferentes combinaciones de lematización, stemming y POS-tagging. Los hallazgos más relevantes son:

- La aplicación de la lematización simple (sin POS-tagging) permitió preservar la forma canónica de los términos, manteniendo la integridad semántica del corpus.
- Las técnicas que incorporaron stemming, ya sea de forma directa o combinada con lematización, redujeron significativamente la variabilidad léxica, aunque a costa de truncar palabras y perder parte del significado original.
- La comparación entre las distintas técnicas evidenció que, para el objetivo de vectorización, es preferible contar con un corpus que conserve el mayor nivel de detalle semántico; por ello se optó por la lematización simple.

El análisis demostró que un pipeline de normalización bien estructurado es esencial para obtener un corpus homogéneo y adecuado para el análisis cuantitativo. La evidencia gráfica (vocabularios y histogramas) respalda la efectividad del proceso, ya que se observa una reducción en la dispersión de tokens y una concentración de términos relevantes.

Parte 2: Vectorización de Documentos

En la segunda parte se transformó el corpus normalizado en representaciones numéricas mediante técnicas de vectorización, tales como One Hot Encoding, Term Count, Probabilidad del Término (TF) y TF-IDF. Las conclusiones derivadas de esta etapa son:

- **One Hot Encoding** ofrece una representación binaria simple que facilita la identificación de la presencia o ausencia de términos, aunque no refleja la frecuencia real de los mismos.
- **Term Count** proporciona una visión detallada de la ocurrencia de cada término en los documentos, lo que es esencial para el análisis de frecuencias.
- **TF** normaliza la frecuencia de términos, permitiendo comparar documentos de distinta longitud y ponderando la importancia relativa de cada palabra.
- **TF-IDF** combina la frecuencia local (dentro de cada documento) y la relevancia global (en el corpus completo), resaltando los términos que son más distintivos y útiles para análisis posteriores.

La integración de estas técnicas de vectorización evidencia que cada método aporta una perspectiva complementaria sobre el contenido textual. La selección de la técnica adecuada dependerá del objetivo del análisis, ya sea para tareas de clasificación, clustering o identificación de conceptos clave.

Finalmente, se concluye que la combinación de un proceso de normalización adecuado y la aplicación de diversas técnicas de vectorización no solo mejora la calidad del análisis textual, sino que también permite adaptar el enfoque a las necesidades específicas de cada proyecto. La práctica demuestra la aplicabilidad de estas herramientas y metodologías, ofreciendo una base sólida para el análisis semántico y cuantitativo en contextos de procesamiento de lenguaje natural.

Bibliografía

[1] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2019.

[3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[4] "spacy: Industrial-strength natural language processing." https://spacy.io, 2020. Accessed: 2025-04-04.

[5] F. Qi, Y. Shen, T. Fu, H. Bao, L. Zhao, H. Ma, and M. Sun, "Stanza: A python natural language processing toolkit for many human languages," in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1–8, 2020.