INSTITUTO POLITÉCNICO NACIONAL
Escuela Superior de Cómputo

Tecnologías de Lenguaje Natural
Profesor: Ituriel Enrique Flores Estrada

Octubre 2024

PRÁCTICA 2

VECTORIZACIÓN DE DOCUMENTOS

Generar el reporte de un programa en Python (Jupyter Notebook) para vectorizar documentos a través de distintas técnicas. En específico, a partir del cuerpo de documentos de la Tabla 1 hacer lo indicado en cada uno de los siguientes numerales.

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

Tabla 1. Documentos para analizar.

1. Obtener los documentos resultado de una etapa de normalización. **(7 puntos)** En específico, después de:
   a. Convertir cada palabra a minúscula y de remover “stop-words” y signos de puntuación.
   b. Aplicar la técnica de “stemming”.
   c. Aplicar POS-Tagging
   d. Aplicar “lemmatization”.

   Nota: como evidencia se debe mostrar el resultado después de cada etapa.

2. Generar y mostrar el vocabulario de términos únicos extraídos de los documentos normalizados y también el histograma de tales términos.
   (6 puntos)

3. Generar los vectores para representar numéricamente cada documento de acuerdo con las siguientes técnicas:

   a. One Hot Encoding o “Term Presence”. **(07 puntos)**
   b. Cantidad de términos o “Term Count”. **(07 puntos)**
   c. Probabilidad del término. **(07 puntos)**

$$P(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ the\ corpus}{Total\ number\ of\ terms\ in\ the\ corpus}$$

d. Frecuencia de términos o “Term Frequency (TF)”. **(07 puntos)**

$$TF = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

e. Frecuencia inversa de documentos “Inverse Document Frequency (IDF)”. **(07 puntos)**

$$IDF = log\frac{Number\ of\ documents\ in\ corpus}{Number\ of\ documentos\ where\ term\ appers}$$

f. TDF-IDF. **(07 puntos)**

$$TDF - IDF = TDF * IDF$$

Análisis y conclusiones (25 puntos). Se deben contrastar los resultados obtenidos en los puntos 2 y 4 para identificar las diferencias entre el texto original y el documento resultado después de haber sido normalizado. En específico, se debe indicar y argumentar el efecto que pudo o no haber tenido cada una de las técnicas de normalización aplicadas sobre los respectivos documentos así como el orden en que se realizan.

Reporte formal y comentarios en código (20 puntos). El reporte debe ser formal por lo que **al menos** debe estar divido en secciones y nombrar todas las tablas e imágenes incluidas en el mismo. Además, cada sección debe describir su propósito, el resultado de las tareas ejecutadas sobre el texto o cuerpo de documentos, así como incluir la respectiva evidencia para comprobar dichos resultados. Asimismo, el código deberá incluir una cabecera y comentarios relevantes conforme a lo detallado en el documento “ReglasEvaluación” publicado al inicio del semestre.

Notas:

1. Las imágenes de código no son evidencia de funcionamiento.
2. En caso de no entregar reporte se asignará cero en la calificación de la práctica.