Preprocesamiento, tokenización y extracción de información

Fabián Villena

Junio 2025

El análisis y minería de texto son habilidades fundamentales en el ámbito de la ciencia de datos. En esta tarea pondrá en práctica técnicas de preprocesamiento, tokenización, filtrado y extracción de información estructurada desde documentos en lenguaje natural.

Para este ejercicio, trabajará con un conjunto de archivos de texto que contienen descripciones clínicas reales, pero las técnicas que aplicará son extensibles a cualquier dominio con textos libres.

El equipo de tecnologías de la organización ha exportado cientos de archivos de texto plano (`.txt`), cada uno correspondiente a una hipótesis diagnóstica. Un ejemplo de contenido de archivo es el siguiente:

```
1 – TRASTORNO DE LA REFRACCIÓN, NO ESPECIFICADO
2
3  paciente de 71 años, con antecedentes de hta en tto,
       diabetes insulinodependiente, dislipidemia,
       hipotiroidismo en tto, enfermedad renal cronica etapa
        iii ,tabaquismo cronico importante, en febrero de
       este año lo suspendio. Refiere que tiene principios
       de Alzheimer y parkinson?????? NO SALE REGISTRO DE
       DIAGNOSTICOS. Refiere que necesita ic a oftalmologo.
       Tiene astigmatismo y miopia, ocupa lentes para ambos
       trastornos de viciorefraccion, refiere que hace 4
       meses que ve borroso utilizando lentes ópticos. Fue
       operada hace mas de 2 años por retinopatia diabetica
       en ambos ojos.
4
5 Al ex fisico: No observo ojo rojo. pupilas isocoricas y
      reactivas. no observo opacidades corneales. RFM
      presente. agudeza visual conservada.
```

El conjunto completo de textos está publicado aquí:

`https://doi.org/10.5281/zenodo.7555181`

Tarea Práctica

Siga los pasos y responda en un *Jupyter Notebook* usando Python y las bibliotecas que estime convenientes. Recuerde entregar un Notebook ya ejecutado que muestre los resultados, dado que su Notebook no va a ser ejecutado.

1. **Carga de los textos**

   a) Importe todos los archivos de texto (`.txt`) y cuente cuántos hay en total (**N** documentos).

   b) Para cada documento, registre la cantidad de caracteres. (1 punto)

   c) Asegúrese de que los caracteres especiales y con tildes se importen correctamente (codificación).

2. **Preprocesamiento y análisis léxico**

   a) Implemente y explique una función de preprocesamiento, por ejemplo: pasar a minúsculas, eliminar puntuación, normalización de espacios, etc. (`str` → `str`). (1 punto)

   b) Implemente una función de tokenización (`str` → `List[str]`) para dividir el texto en palabras. (1 punto)

   c) Elabore una lista de *stopwords* (palabras vacías) en español, y filtre las palabras del texto eliminando estas. (1 punto)

   d) Visualice las palabras más frecuentes en el corpus, idealmente con una *nube de palabras* (*wordcloud*). (1 punto)

3. **Extracción de información mediante patrones (expresiones regulares)**

   a) Usando expresiones regulares, detecte menciones de las siguientes enfermedades en los textos. Para cada una, diseñe un patrón que contemple formas alternativas o sinónimos. Se proveen los códigos CIE-10 y vínculos a Wikidata para su referencia (1 puntos):

I10	Hipertensión esencial (primaria)
E11	Diabetes mellitus tipo 2
C50	Neoplasia maligna de mama
N18	Enfermedad renal crónica
G30	Enfermedad de Alzheimer

   b) Para cada entidad, indique cuántos documentos la mencionan y calcule en todos los casos la prevalencia (documentos que la mencionan/N).