
Universidad
Tecmilenio®

Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Modelos de lenguaje en n-gramas

Introducción

Las convenciones utilizadas en la escritura a veces no reflejan la organización lógica de los símbolos textuales, por ejemplo, algunos signos de puntuación se escriben antes de cierta palabra y otros después, aunque no formen parte de ella y esto depende del lenguaje y del estilo de escritura. En ocasiones, en una cadena de caracteres o texto no están claros los límites lógicos entre las palabras o símbolos de puntuación y eso puede complicar el análisis de textos escritos.

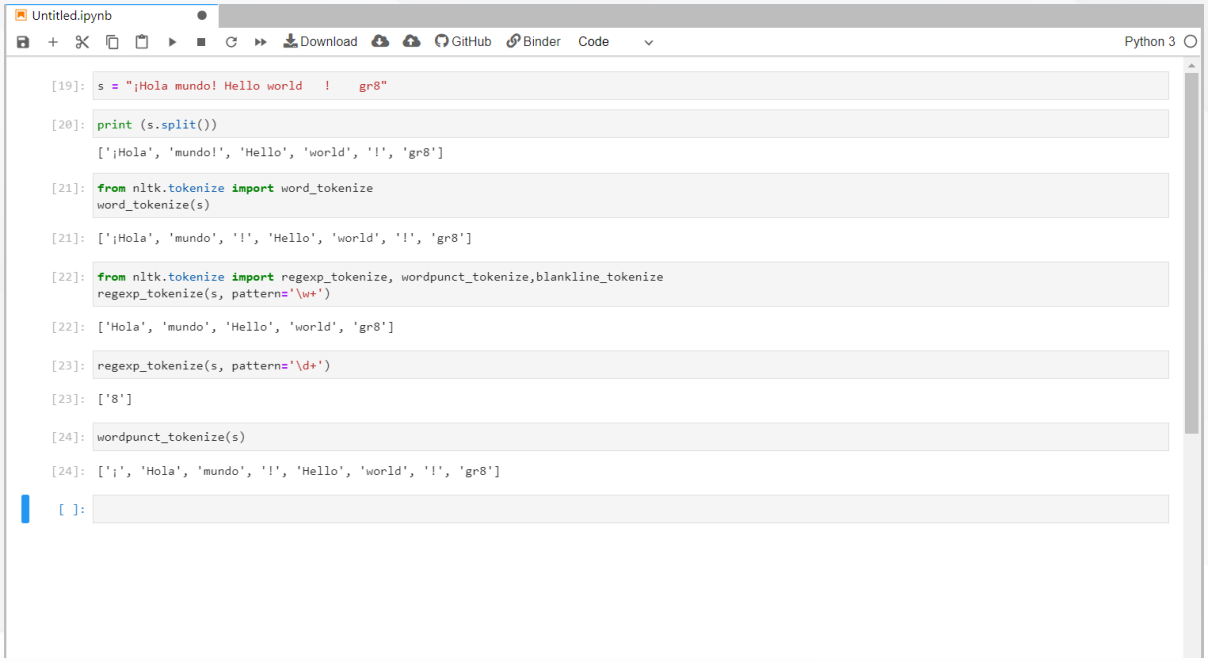
En este tema conocerás lo siguiente:

- La importancia de la estandarización de textos.
- El concepto de token y tokenización.
- Qué son los unigramas, bigramas y n-gramas.
- La diferencia entre un q-grama y un n-grama.
- La utilidad de los modelos de lenguaje.
- La diferencia entre un conjunto de prueba y uno de entrenamiento.

Explicación

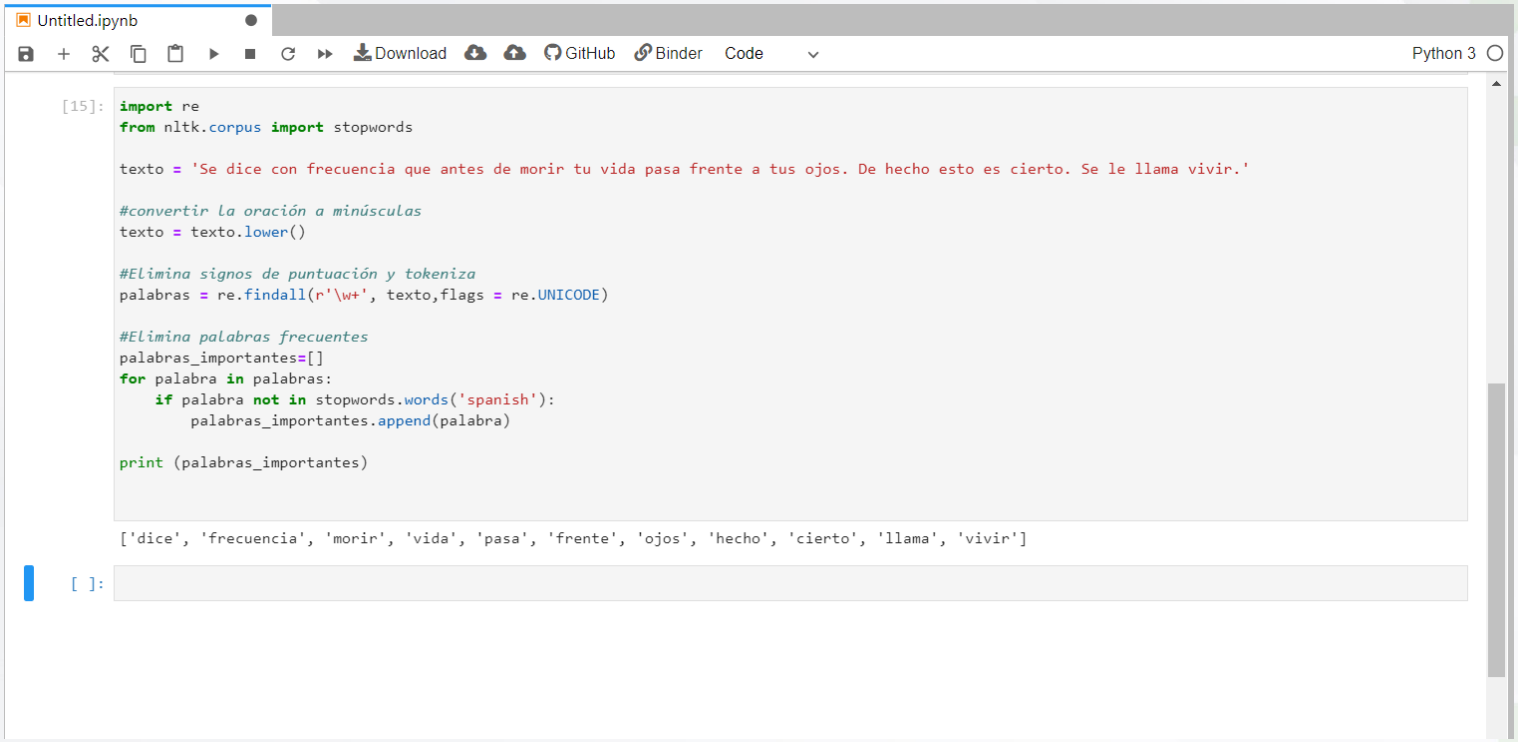
Es posible someter a los textos a un **proceso de estandarización**, el cual puede ayudar a mejorar la obtención de las entidades individuales (las palabras) y con ello, realizar el análisis del mensaje dentro del texto.
La **palabra** o **token** es la unidad mínima utilizada en procesamiento de textos y definirla depende de la tarea que se esté realizando. Por lo general, cada palabra se define como una unidad separada por espacios o signos de puntuación, sin embargo, los signos de puntuación pueden estimarse o no como palabras.

Los **tokens** son, por tanto, la ocurrencia de las palabras en un texto y la ***tokenización*** es el proceso que permite encontrar todas las palabras dentro de él. El conjunto de todas las palabras posibles dentro de un texto se conoce como **vocabulario**.
En la figura 1 se muestra la tokenización del texto: ¡Hola mundo! Hello world ! gr8", utilizando diferentes instrucciones.



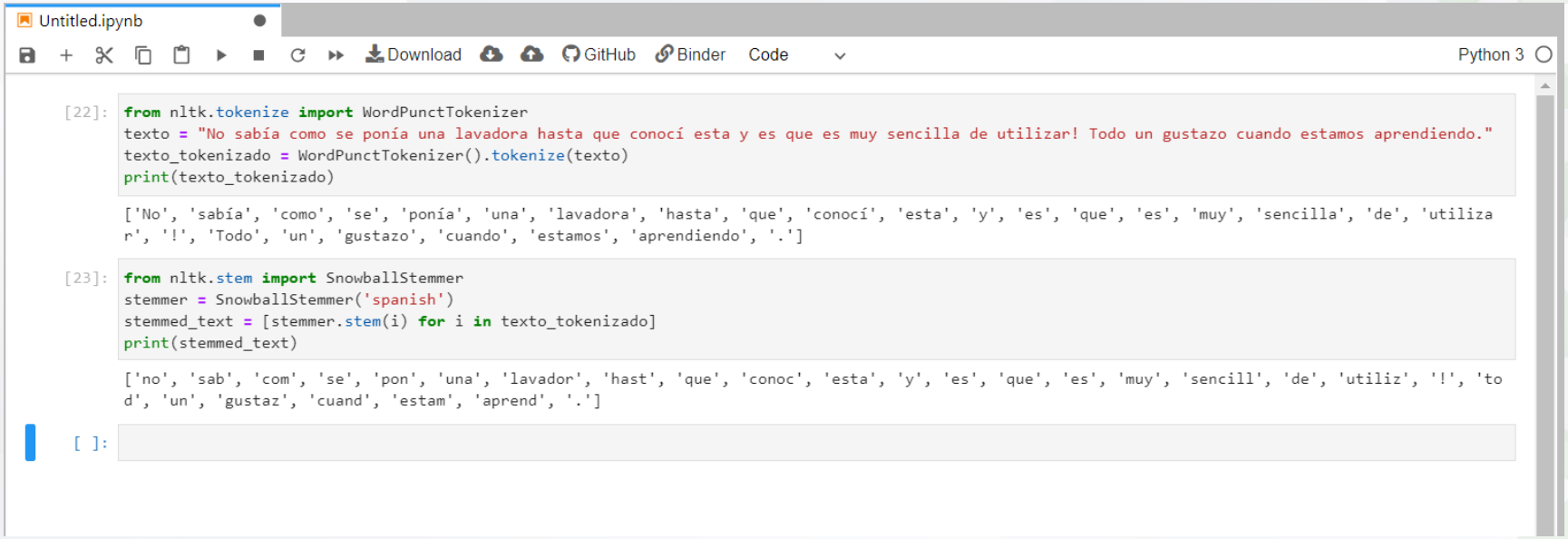
Explicación

La eliminación de las palabras frecuentes puede ayudar a construir un vocabulario que contenga solo aquellas palabras que le den significado o valor al texto. En la figura se observa un ejemplo de eliminación de palabras frecuentes en un texto, se procesa independientemente si está escrito en minúsculas o mayúsculas. Nota que las palabras “se, con, que, antes, de, tu, a, tus, esto, es, le” no forman parte del vocabulario final del ejemplo.



Explicación

En ocasiones se requiere extraer las raíces de las palabras y construir el vocabulario con ellas, así es posible eliminar sufijos como “ando” o “iendo”. El ***stemming*** consiste en eliminar y reemplazar los sufijos de la raíz de la palabra. Este proceso tiene sus reglas dependiendo del idioma, en español se eliminan sufijos de acciones (ar, er, ir, ía, es, etc.), terminaciones plurales, géneros, entre otros. El proceso de extracción regresa una palabra sin la terminación o sufijos.

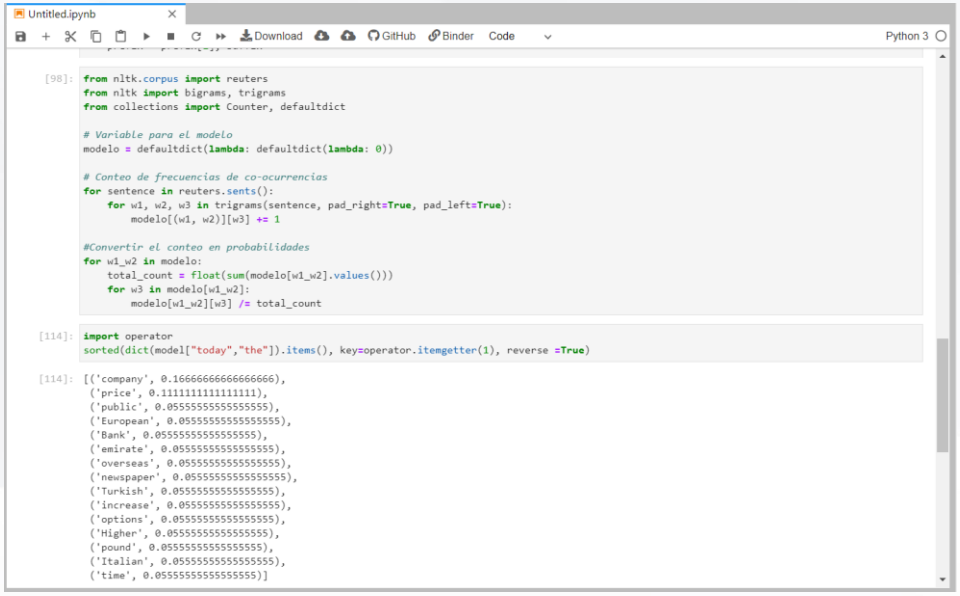


Explicación

En general, el **n-grama** se puede definir como la secuencia de N elementos de una muestra determinada de texto o voz. Los elementos pueden ser letras o palabras. Algunos autores consideran que un **n-grama** es una secuencia de N palabras y un **q-grama** una secuencia de Q caracteres (incluidas las letras). Si el n-grama está formado solo por una palabra se llama **unigrama**, si está compuesto por dos, **bigrama**, y si está formado por tres elementos, **trigrama**.

Un **modelo de lenguaje** debe ser capaz de capturar la mayor cantidad de información posible para describir al lenguaje en términos de esa información. Es posible obtener un modelo estadístico de lenguaje basado en n-gramas. El modelo debe ser capaz de predecir la siguiente palabra en una oración, es decir, $P(w|h)$, la probabilidad de que una palabra *w* suceda dada una historia *h*.

En la figura se ve una implementación del modelo utilizando el corpus Reuters disponible en NLTK. El ejemplo utiliza trigramas para construir el modelo. Como resultado, se determina que la palabra más probable después del texto "*today the*" es "*company*" seguida de "*price*".



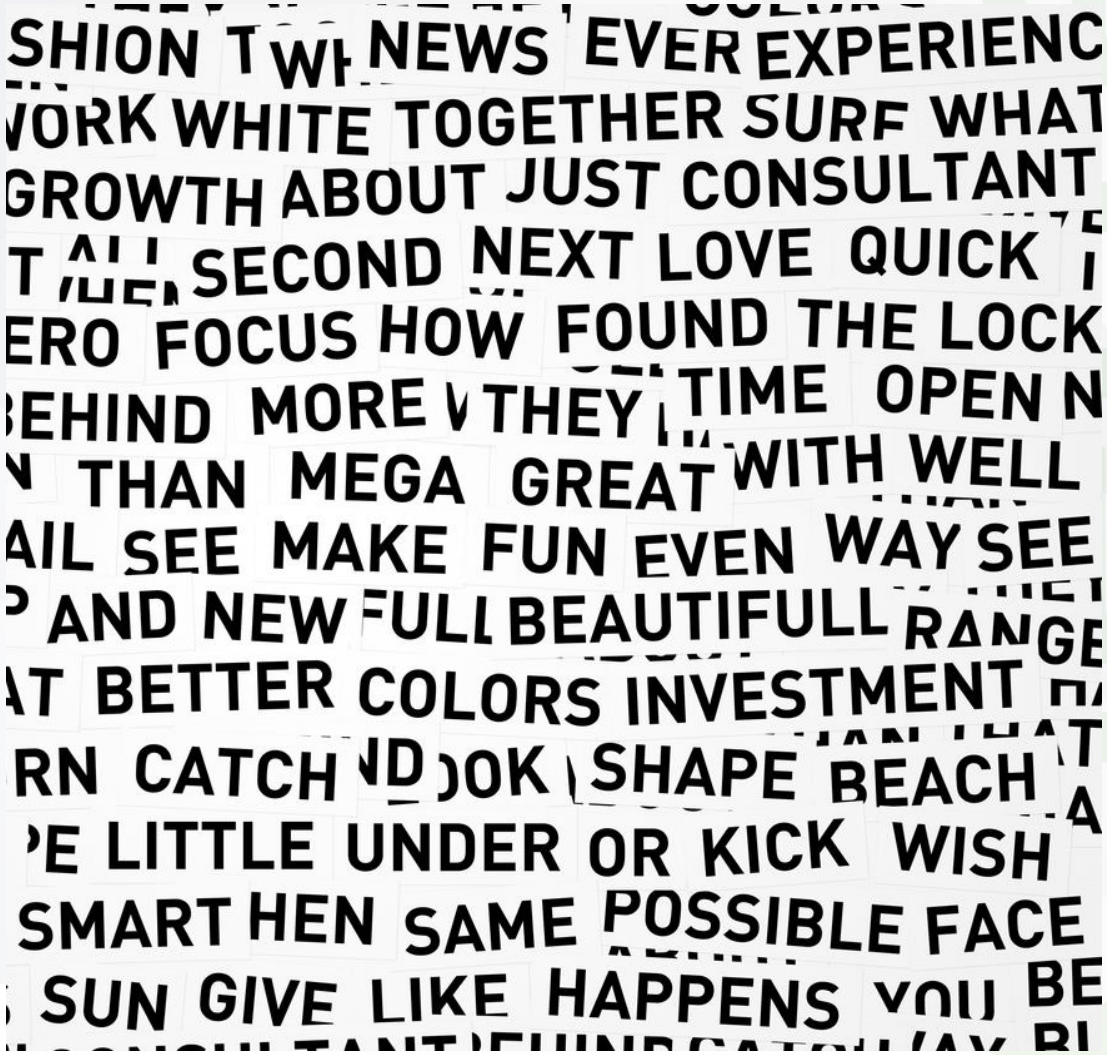
Introducción

01 ¿Cuál es la importancia del uso de n-gramas en el PLN?

02 ¿El proceso de tokenización y n-gramas es un proceso que realizamos mentalmente los seres humanos durante una conversación casual?

Cierre

En este tema se presentó el modelado del lenguaje y los n-gramas, que son dos de las herramientas más utilizadas en el procesamiento de lenguaje natural.
Los modelos de lenguaje asignan un valor de probabilidad a una oración o secuencia de palabras y con ello es posible predecir la palabra que sigue en un texto.



Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Clasificación de texto

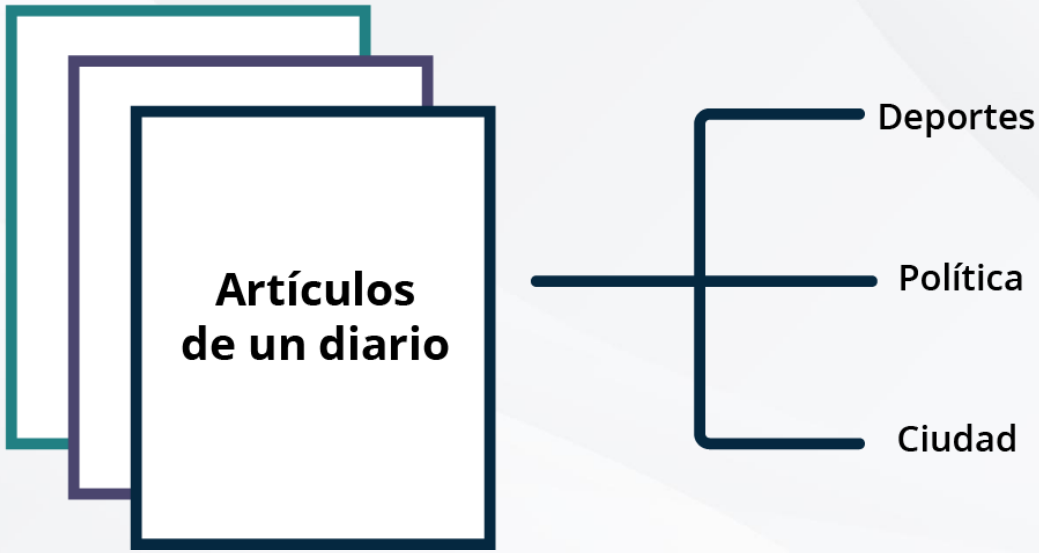
Introducción

La clasificación de textos es una de las tareas más importantes y típicas del aprendizaje automático supervisado. La asignación de categorías a documentos que pueden ser una página web, un libro de biblioteca, artículos multimedia, una galería, etc., tiene muchas aplicaciones, por ejemplo, filtrado de spam, enrutamiento de correo electrónico, análisis de opiniones o sentimientos, entre otros.

En este tema conocerás:
El proceso de clasificación de textos.
Los algoritmos más comunes utilizados para clasificar textos.
El clasificador bayesiano.
Los pasos necesarios para entrenar a un clasificador bayesiano.

Explicación

Existen distintos tipos de algoritmos de aprendizaje automático: supervisado, no supervisado, por refuerzo y semisupervisado. La clasificación de textos es un ejemplo de tarea de aprendizaje automático supervisado, ya que un conjunto de datos etiquetado y formado por documentos de texto y sus etiquetas se utiliza para entrenar a un clasificador.



Explicación

Un clasificador ingenuo bayesiano (*Naive Bayes,* por su nombre en inglés) o simplemente bayesiano está basado en el teorema de Bayes. Este clasificador asume que la presencia o ausencia de alguna característica particular no está relacionada con la presencia o ausencia de alguna otra. De tal manera que estas características contribuyen de forma independiente a la probabilidad de que tal objeto pertenezca a una clase independientemente de la presencia o ausencia de las otras características. Por ejemplo, una fruta puede clasificarse como un limón si es de color verde, redonda y aproximadamente de 4 cm de diámetro.

El teorema de Bayes permite calcular la probabilidad de que cierto objeto con ciertas características pertenezca a una clase dado cierto conocimiento *a priori*:
P(clase|objeto) = (P(objeto|clase) * P(clase)) / P(objeto)
Donde P(clase|objeto) es la probabilidad de una clase dado el objeto con ciertas características.

Explicación

El proceso de clasificación de textos incluye los siguientes pasos:
Preparación de los datos: en este paso se selecciona el conjunto de datos con el que se trabajará y se realiza un preprocesamiento básico al texto para después separarlo en el conjunto de entrenamiento y de prueba.
Extracción de características: el texto se transforma en elementos característicos que se usarán en el proceso de clasificación.
Entrenamiento del modelo: el clasificador se entrena con los datos etiquetados.
Evaluación del modelo: el clasificador se evalúa y, en caso de ser necesario, se realizan ajustes para mejorar el resultado.
El primer ejemplo trata de clasificar nombres de personas según el género. Se utilizará la librería NLTK (Bird, Klein y Loper, 2009) y el corpus *names*. En la figura se muestra la descarga de los datos y una breve exploración sobre ellos. En total hay 7,944 nombres clasificados como masculinos o femeninos.

Actividad

01 Descarga el corpus "*wine dataset*" dentro de scikit-learn y explóralo. El conjunto de datos contiene los resultados del análisis químico de vinos cultivados en la misma región en Italia, pero derivados de tres cultivares diferentes, consta de 13 atributos y 3 tipos de vino.

02 Separa las columnas de los datos en variables dependientes e independientes, es decir, atributos y etiquetas de clase.
Utiliza el código siguiente para separar los datos en el conjunto de entrenamiento y de prueba.

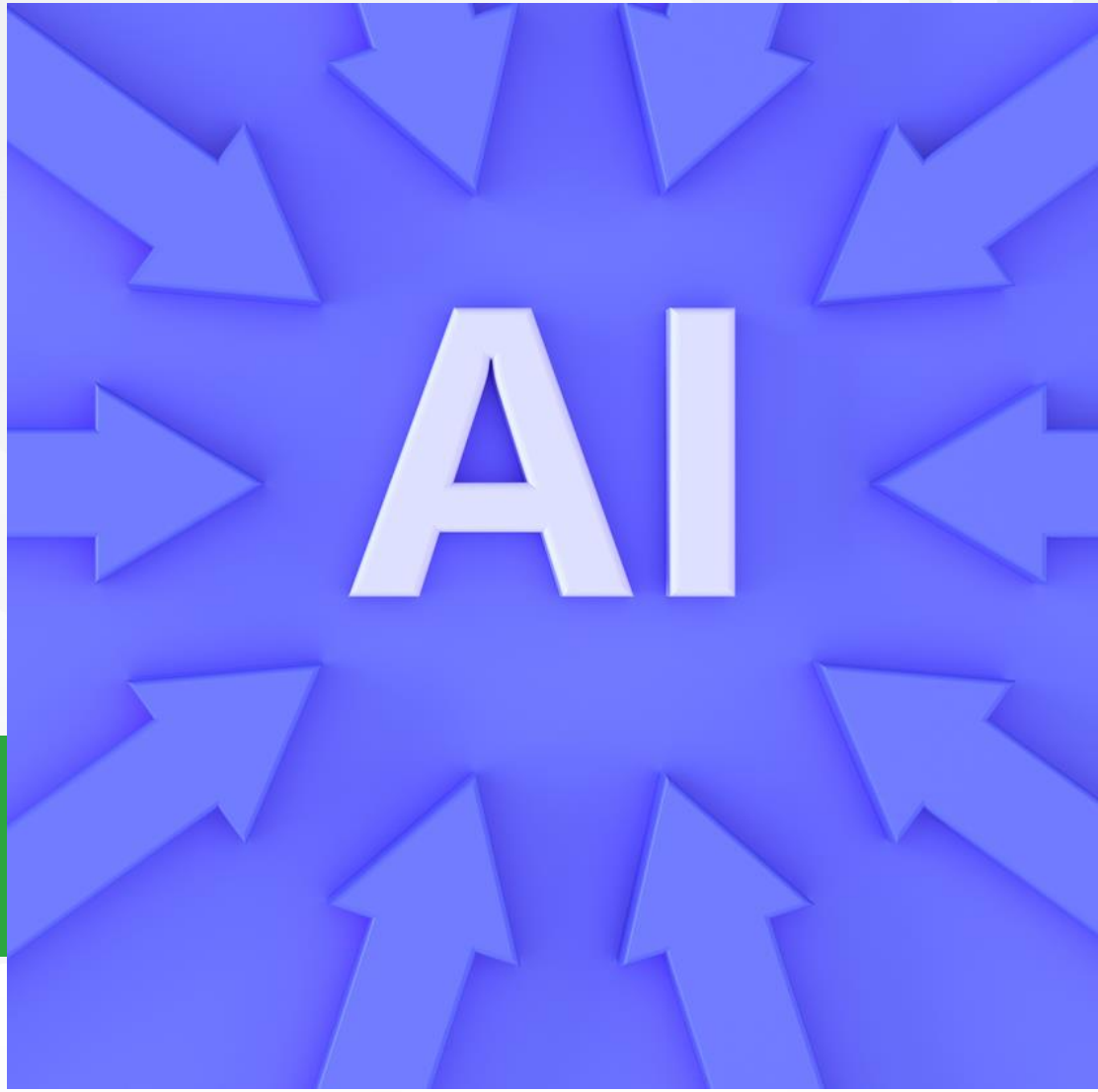
```
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(wine.data,
wine.target, test_size=0.3,random_state=109)
```

03 Genera el modelo utilizando el conjunto de datos de entrenamiento.

04 Con el modelo construido, realiza predicciones utilizando el conjunto de datos de prueba y almacena los resultados en una variable nombrada *y_pred*.

Cierre

Muchas de las tareas en procesamiento de lenguaje natural se pueden tratar como problemas de clasificación. La categorización de textos, donde el objetivo es asignar una clase a un texto particular, incluye tareas como análisis de sentimientos, detección de spam, identificación del lenguaje, identificación del sexo de un autor, entre otros.
No olvides que los clasificadores supervisados utilizan corpus previamente etiquetados que les permiten construir modelos que predicen la categoría o etiqueta de una entrada en función de ciertas características. De igual forma, recuerda que para entrenar a un clasificador supervisado se debe separar el total de los datos en un conjunto de entrenamiento y otro de pruebas.



Bibliografía

- Bird, S., Klein, E., y Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. Estados Unidos: O'Reilly Media.