6.- Lenguaje Natural

martes, 5 de marzo de 2024 11:46 p. m.

POS Tagging (Etiquetado gramatical de partes del discurso):
POS Tagging es un proceso en el procesamiento del lenguaje natural (PLN) que implica asignar una etiqueta gramatical a cada palabra en un texto según su función y significado en una oración. Estas etiquetas representan partes del discurso como sustantivos, verbos, adjetivos, adverbios, etc. El etiquetado POS es esencial para muchos otros procesos de PLN, como el análisis sintáctico, la traducción automática y la extracción de información. Se usa para comprender mejor la estructura gramatical de un texto y extraer información semántica útil.

POS Tagging a menudo se utiliza en conjunto con otros procesos de PLN, como el análisis sintáctico y la extracción de información. Por ejemplo, en un sistema de análisis de sentimientos, el POS Tagging puede ayudar a identificar los sustantivos y los adjetivos que están relacionados con las opiniones expresadas en un texto.

Stemming:
Stemming es un proceso de normalización de palabras en el procesamiento del lenguaje natural. Consiste en reducir una palabra a su raíz o base, eliminando afijos como sufijos, prefijos o infijos. El objetivo del stemming es reducir las palabras a su forma básica para mejorar la recuperación de información y el análisis de texto. Esto significa que palabras con la misma raíz tendrán la misma representación, lo que puede ser útil para tareas como la búsqueda de información o la agrupación de texto.

Por ejemplo, aplicar stemming a las palabras "corriendo", "corrió" y "corre" las reduciría todas a la forma base "corre". Un algoritmo comúnmente utilizado para el stemming es el algoritmo de Porter.

Lematización:
La lematización es un proceso similar al stemming en el procesamiento del lenguaje natural, pero más sofisticado. A diferencia del stemming, que simplemente elimina afijos para reducir una palabra a su forma base, la lematización tiene en cuenta el significado de la palabra y la convierte en su forma canónica o lema. El lema es la forma base de una palabra que se encuentra en un diccionario. La lematización utiliza reglas gramaticales y conocimiento lingüístico para realizar esta conversión de manera más precisa que el stemming.

Por ejemplo, la lematización convertiría las palabras "corriendo", "corrió" y "corre" en el lema "correr", reconociendo que todas tienen el mismo significado básico. Este enfoque es útil cuando se requiere una mayor precisión en el análisis de texto, especialmente en aplicaciones donde la comprensión del significado exacto de las palabras es crucial, como la traducción automática o el análisis de sentimientos.

Lemmatization --> Lema --> Raiz gramatical de la palabra

Steamming --> /prefijos/sufijos --> Palabras flexionales

POS Tagging (Part of Speech Tagging).

POS Tagging se realiza en conjunto con lematización para la efectividad del método.

======== Normalización de texto ==========
- Lematización
- Stemming
- POS Tagging
- Remover Stop Words
- Signos de puntuación

(Expresiones regulares) --> Puede entrar (no se usa para la práctica)

Expresiones regulares:

^ Inicio de texto
$ fin de texto
[^] negación
[ ] Rango de números o letras
( ) Grupo
* 0 o más caracteres
? Si existe o no existe un carácter
+ Uno o más caracteres
. Cualquier carácter
\w Capturar palabras
\d Capturar dígitos
{ } Conjunto de grupos previamente identificado
\s Espacios en blanco
\S Que no haya espacios
\D Que no haya dígitos
\w Que no haya palabras
\. Encontrar el punto
\? Encontrar una interrogación
\+ Encontrar la suma
\\ Encontrar un backslash
[x | y |] Que se encuentre x o y o …

Formas de operación de las expresiones regulares
Lazy
---> Greedy

Toma símbolos con su significado literal
---> Toma " " con significado especial dentro de RegEx

/^.*([0-2][1-9]|3[0-1]).*([0][1-9]|1[0-2]).*(19\d\d|20\d[0-5]).*([3,4,5,7]\d{6}\d{10}).*$/gm

Dia 01 del 02 del año 1900 con numero de tarjeta :41234567891234958
Dia 31 del 11 del año 2024 con numero de tarjeta :41234567891234958
ssss31sss12sss1900ssss36345678912349587
dia 31 del 12 del año 2000 con el numero de tarjeta 41234567891234958

Ejemplo:
La direccion del alumno con boleta 2022630769 es
micorreo@ipn.mx y vive en: direccion

Capturar boletas con las expresiones regulares:
Versión difícil:
"La direccion … " [0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]

Versión con los auxiliares :

1.-	^La direccion del alumno con boleta \d{10} es \S+@ipn\.mx y vive en: .+$
2.-	/^La direccion del alumno con boleta [0-9]{10} es [\w.-]+@ipn\.mx y vive en: [\w\s]+$/gm
3.-	/^.+ \d{10} .+ \S+@ipn\.mx .+$/gm

[^a - Z] ---> Que no tenga de la a hasta la Z (todas la letras en minúsculas y mayúsculas)