DATA SCIENCE

Javier Lezama

Incluit

Mayo - Junio de 2019

PNL

El procesamiento del lenguaje natural abreviado PLN, o NLP del idioma inglés Natural Language Processing es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural, es decir, de las lenguas del mundo. El PLN no trata de la comunicación por medio de lenguas naturales de una forma abstracta, sino de diseñar mecanismos para comunicarse que sean eficaces computacionalmente que se puedan realizar por medio de programas que ejecuten o simulen la comunicación. Los modelos aplicados se enfocan no solo a la comprensión del lenguaje de por sí, sino a aspectos generales cognitivos humanos y a la organización de la memoria. El lenguaje natural sirve solo de medio para estudiar estos fenómenos. Hasta la década de 1980, la mayoría de los sistemas de PLN se basaban en un complejo conjunto de reglas diseñadas a mano.

A partir de finales de 1980, sin embargo, hubo una revolución en PLN con la introducción de algoritmos de aprendizaje automático para el procesamiento del lenguaje.

PNL

El experimento de Georgetown en 1954 involucró traducción automática de más de sesenta oraciones del Ruso al Inglés.

Dificultades en el procesamiento del lenguaje natural

Ambigüedad

Las lenguas naturales son inherentemente ambiguas en diferentes niveles: En el nivel léxico, una misma palabra puede tener varios significados, y la selección del apropiado se debe deducir a partir del contexto oracional o conocimiento básico. Muchas investigaciones en el campo del procesamiento de lenguajes naturales han estudiado métodos de resolver las ambigüedades léxicas mediante diccionarios, gramáticas, bases de conocimiento y correlaciones estadísticas.

En el nivel pragmático, una oración, a menudo, no significa lo que realmente se está diciendo. Elementos tales como la ironía tienen un papel importante en la interpretación del mensaje.

Para resolver estos tipos de ambigüedades y otros, el problema central en el PLN es la traducción de entradas en lenguas naturales a una representación interna sin ambigüedad, como árboles de análisis.

Detección de separación entre las palabras.
En la lengua hablada no se suelen hacer pausas entre palabra y palabra. El lugar en el que se debe separar las palabras a menudo depende de cuál es la posibilidad que mantenga un sentido lógico tanto gramatical como contextual. En la lengua escrita, lenguas como el chino mandarín tampoco tienen separaciones entre las palabras.
Recepción imperfecta de datos.
Acentos extranjeros, regionalismos o dificultades en la producción del habla, errores de mecanografiado o expresiones no gramaticales, errores en la lectura de textos mediante OCR.

Componentes

* Análisis morfológico.
* Análisis sintáctico.
* Análisis semántico.
* Análisis pragmático.
* Planificación de la frase.
* Generación de la frase.

Las principales tareas de trabajo en el PLN

* Síntesis del discurso.
* Análisis del lenguaje.
* Comprensión del lenguaje.
* Reconocimiento del habla.
* Síntesis de voz.
* Generación de lenguajes naturales.
* Traducción automática.
* Respuesta a preguntas.
* Recuperación de la información.
* Extracción de la información.

El gran desafío

NLP es considerado uno de los grandes retos de la inteligencia artificial ya que es una de las tareas más complicadas y desafiantes: ¿cómo comprender realmente el significado de un texto? ¿cómo intuir neologísmos, irónias, chistes ó poesía? Si la estrategia/algoritmo que utilizamos no sortea esas dificultades de nada nos servirán los resultados obtenidos.

Modelos, maquetas y el mundo

PLN

En NLP no es suficiente con comprender meras palabras, se deberá comprender al conjunto de palabras que conforman una oración, y al conjunto de lineas que comprenden un párrafo. Dando un sentido global al análisis del texto/discurso para poder sacar buenas conclusiones.
Nuestro lenguaje está lleno de ambigüedades, de palabras con distintas acepciones, giros y diversos significados según el contexto. Esto hace que el NLP sea una de las tareas más difíciles de dominar.

* Resumen de textos: El algoritmo deberá encontrar la idea central de un artículo e ignorar lo que no sea relevante.
* ChatBots: deberán ser capaces de mantener una charla fluida con el usuario y responder a sus preguntas de manera automática.
* Generación automática de keywords y generación de textos siguiendo un estilo particular
* Reconocimiento de Entidades: encontrar Personas, Entidades comerciales o gubernamentales ó Países, Ciudades, marcas
* Análisis de Sentimientos: deberá comprender si un tweet, una review o comentario es positivo ó negativo y en qué magnitud (ó neutro). Muy utilizado en Redes Sociales, en política, opiniones de productos y en motores de recomendación.
* Traducción automática de Idiomas
* Clasificación automática de textos en categorías pre-existentes ó a partir de textos completos, detectar los temas recurrentes y crear las categorías.

Ejemplos

* ChatBots: Banca, Recursos humanos, Turismo, Ecommerce, Salud,etc
Jarvis, de Facebook.
Siri, de Apple.
Alexa, de Amazon.
Google Now.
Cortana, de Microsoft.

Ejemplos

* Corrección:
Grammarly

Ejemplos

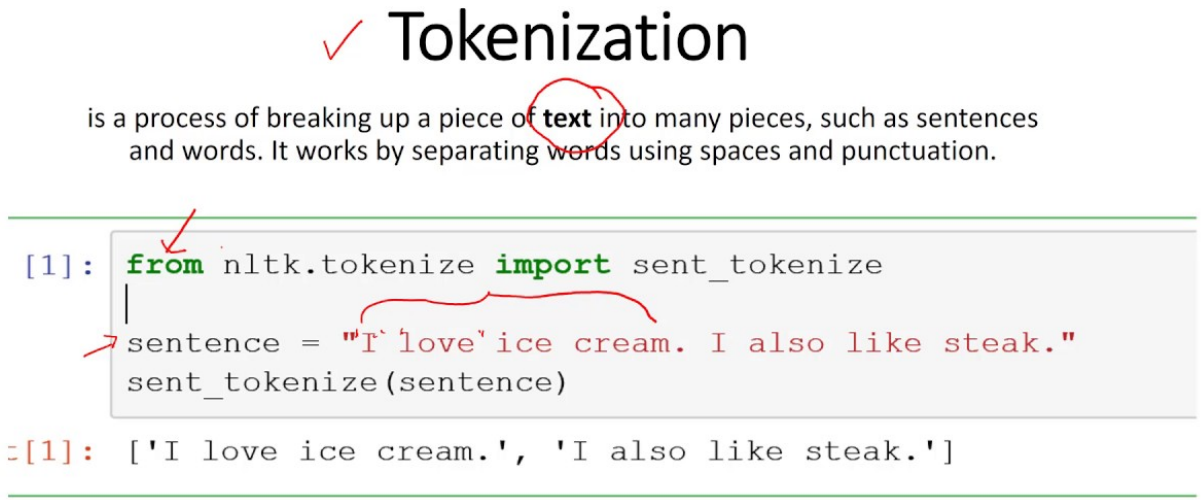
* Análisis de Sentimientos:
https://deepmoji.mit.edu/

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización

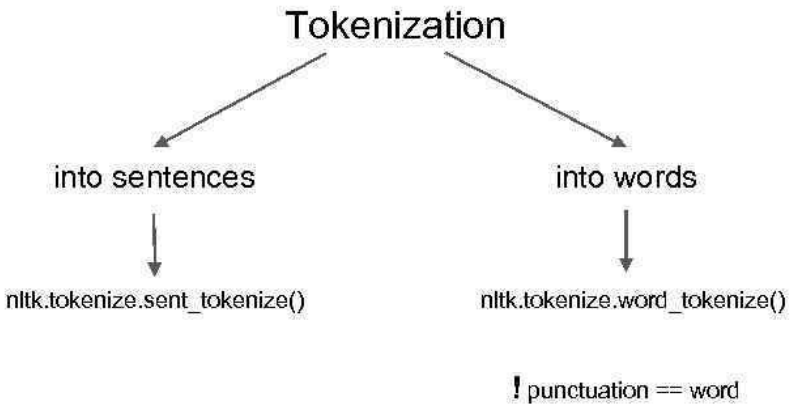


Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización

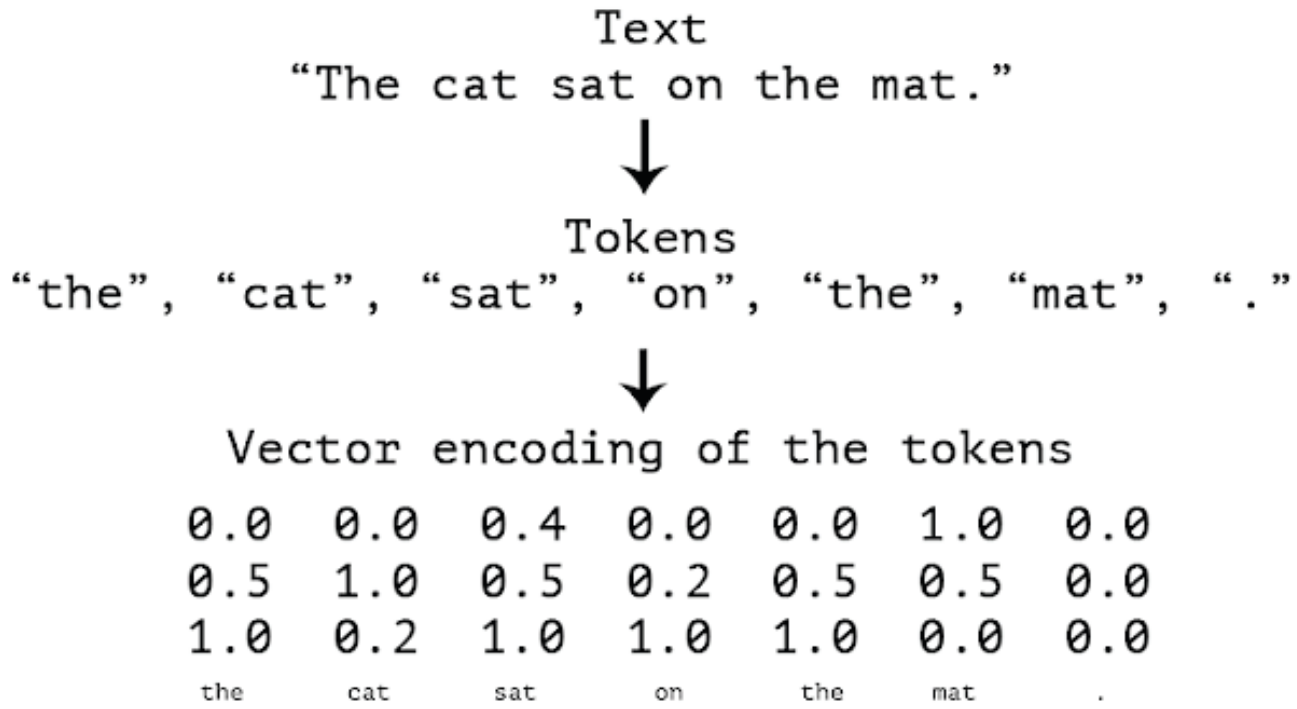


Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Tokenización



Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Normalización

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Frecuencia de palabras

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* Frecuencia de palabras

Técnicas

* Natural Language Tooikit - NLTK
* import nltk
* n-gramas o palabras compuestas