
Universidad
Tecmilenio®

Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Introducción al procesamiento de lenguaje natural

Introducción

Uno de los objetivos del procesamiento del lenguaje natural es buscar que las computadoras realicen tareas que utilicen el lenguaje, por ejemplo, comunicar a una computadora con una persona.
En este tema conocerás:

•Los componentes del lenguaje.
•El objetivo del procesamiento del lenguaje natural.
•La relación del procesamiento del lenguaje natural con otras disciplinas.
•La evolución del procesamiento del lenguaje natural en el tiempo.
•Aplicaciones del procesamiento del lenguaje natural.

Explicación

El procesamiento del lenguaje natural (PLN) resuelve tareas que permiten a las computadoras comprender nuestro lenguaje, por ejemplo, entender los sentimientos plasmados en un texto, reconociendo el habla o generando respuestas a preguntas específicas, entre otros.

Actualmente, existen diferentes aplicaciones de cómputo que utilizamos diariamente y se han creado a partir del procesamiento del lenguaje natural:

- Corrección gramatical (editores de texto).
- Máquinas de búsqueda (Google, Bing, Yahoo! o Wolfram Alpha).
- Asistentes (Siri, Alexa).
- Clasificadores de correo no deseado (Gmail u Outlook).
- Procesadores de noticias (Google o Yahoo!).
- Traducción de textos (Google Translate).
- IBM Watson.

Explicación

El lenguaje natural posee propiedades que pueden afectar el rendimiento de aplicaciones de cómputo construidas a partir de lenguajes formales, como los lenguajes de programación de computadoras. Estas propiedades son la **variación** y la **ambigüedad lingüística**. La primera se refiere a la posibilidad de usar diferentes palabras para comunicar una misma idea. La segunda sucede cuando una palabra o frase permite más de una interpretación. Ambas propiedades hacen que el procesamiento del lenguaje natural sea un problema complejo de resolver, ya que implica varios retos sintácticos, semánticos y pragmáticos.

Explicación

En 1948, Shannon aplicó modelos probabilísticos de procesos discretos de Markov a autómatas para el lenguaje (Shannon, 1948). Al tomar la idea de un proceso de Markov de estado finito del trabajo de Shannon, en 1956 Chomsky consideró por primera vez las máquinas de estado finito como una forma de caracterizar un tipo de gramática y definió un lenguaje de estado finito como un lenguaje generado por una gramática de estado finito.

Koenig (1946) desarrolló el espectrógrafo de sonido, además de una investigación fundamental en fonética instrumental, que sentó las bases para el trabajo posterior en el reconocimiento de voz; esto llevó a los primeros reconocedores de voz de máquina a principios de la década de 1950.

En el verano de 1956, John McCarthy, Marvin Minsky, Claude Shannon y Nathaniel Rochester reunieron a un grupo de investigadores para un taller de dos meses sobre lo que decidieron llamar inteligencia artificial (IA). La IA incluía investigadores que se enfocaban en algoritmos estocásticos y estadísticos (modelos probabilísticos y redes neuronales). El enfoque principal del nuevo campo fue el trabajo sobre el razonamiento y la lógica tipificados por el trabajo de Newell y Simon sobre el Teórico de Lógica y el Solucionador General de Problemas.

El paradigma de la **comprensión del lenguaje natural** comenzó con el sistema *SHRDLU* de Terry Winograd (1972), que simulaba un robot incrustado en un mundo de bloques de juguete en 1972. El programa era capaz de recibir comandos de texto en lenguaje natural, por ejemplo, mover un bloque rojo y colocarlo encima de un bloque verde más pequeño.

Se vieron avances en traducción automática en el año 2003 y el modelado de tópicos donde se demostró que es posible construir sistemas entrenados sin la necesidad de integrar datos anotados. Además, el costo generalizado y la dificultad de producir *corpus* anotados de manera confiable se convirtieron en un factor limitante en el uso de técnicas supervisadas en muchos problemas, así que es muy probable que la tendencia apunte hacia un incremento en el uso de técnicas no supervisadas en la resolución de problemas.

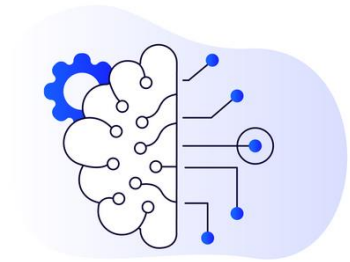
Actividad

01 ¿Qué componentes matemáticos y lingüísticos han estado siempre presentes a través de la historia en el procesamiento del lenguaje natural?

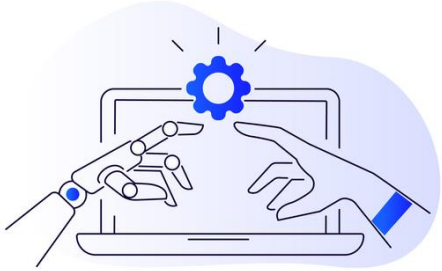
02 ¿Por qué consideras que aún no se ha logrado una IA que iguale al ser humano para comprender y profundizar en el lenguaje?

Cierre

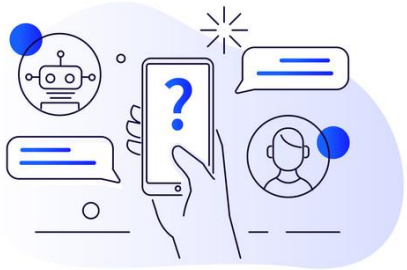
Desde el punto de vista técnico, el procesamiento del lenguaje natural es una tarea difícil de resolver, dada la propia complejidad del lenguaje y que implica un gran reto en ingeniería en términos de obtener mayor conocimiento en sistemas lingüísticos, aprendizaje automático, análisis morfológico, sintáctico y semántico, así como el tratamiento de la ambigüedad del lenguaje.



MACHINE LEARNING



INTERACTION



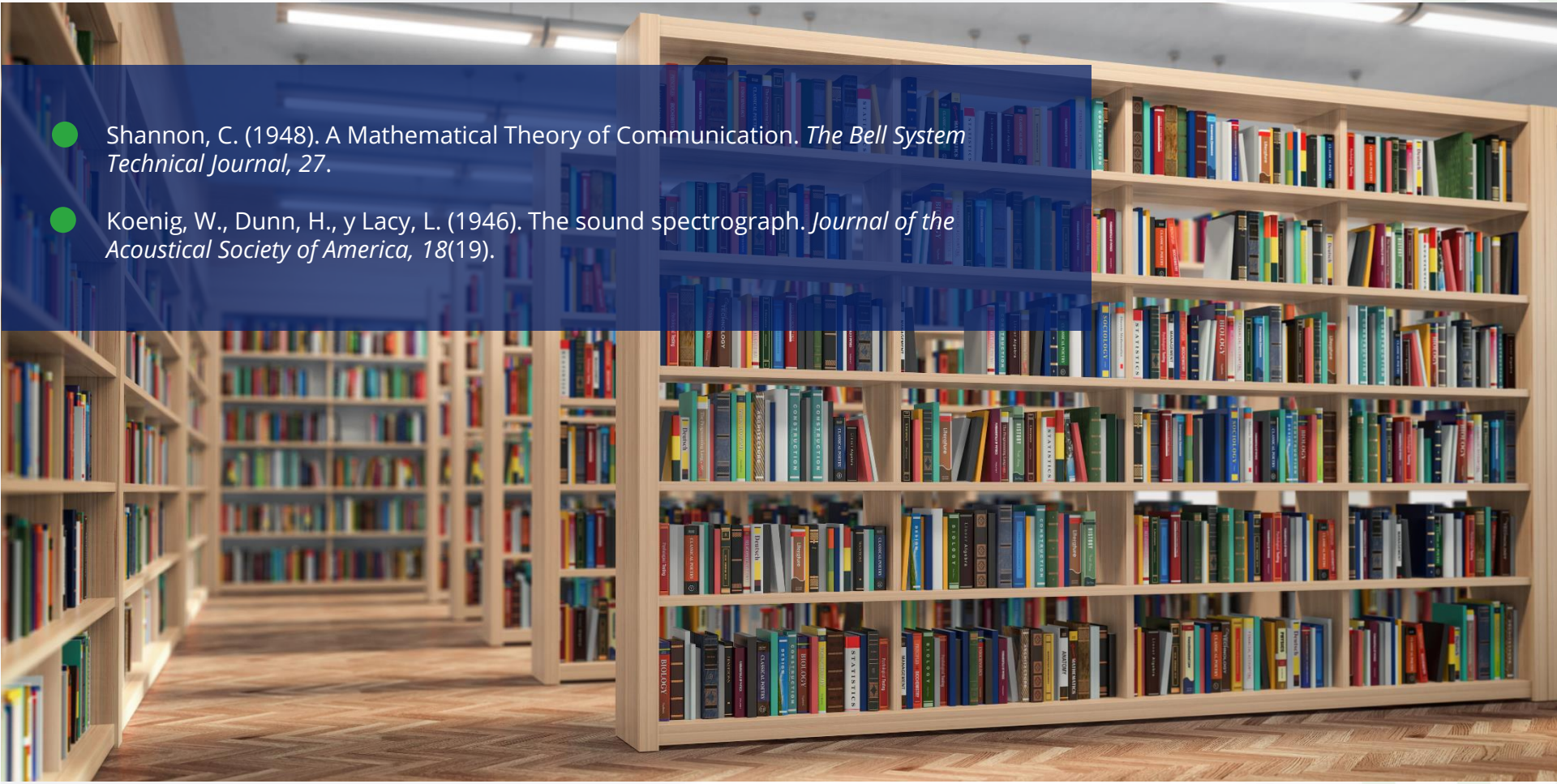
CHAT BOT

AI

No olvides que la tecnología del habla y del lenguaje se basa en modelos formales o representaciones del conocimiento del lenguaje en los niveles de fonología, fonética, morfología, sintaxis, semántica, pragmática y discurso y, para capturar este conocimiento, se utilizan máquinas de estado, sistemas de reglas formales, lógica y modelos probabilísticos.

Bibliografía

- Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*.
- Koenig, W., Dunn, H., y Lacy, L. (1946). The sound spectrograph. *Journal of the Acoustical Society of America, 18*(19).



Universidad Tecmilenio

Procesamiento de lenguaje natural y visión computacional

Lingüística computacional basada en corpus

Introducción

Un recurso léxico o, simplemente, léxico, es una colección de palabras y/o frases que contienen información asociada al texto, como partes de la oración (*Parts of the Speech, POS*) o acepciones de palabras.

En este tema conocerás:

Qué es un recurso léxico.

El concepto de corpus.

La importancia del uso de un corpus en PLN.

La librería NLTK para tareas de procesamiento de lenguaje natural.

Explicación

El trabajo práctico en el procesamiento de lenguaje natural generalmente requiere de una gran cantidad de datos, de un corpus. De manera simple, un corpus (en plural, corpora) es una colección de textos que se han recopilado con algún fin.
Los corpora se dividen en tres categorías según McEnery (2003): monolingües, comparables y paralelos. Los primeros, contienen información en un solo idioma. El corpus comparable incluye una variedad de corpus monolingües en diferentes idiomas, con un nivel similar de balance y representatividad. Los corpora paralelos incluyen textos originales en un idioma con traducciones de esos textos.

Explicación

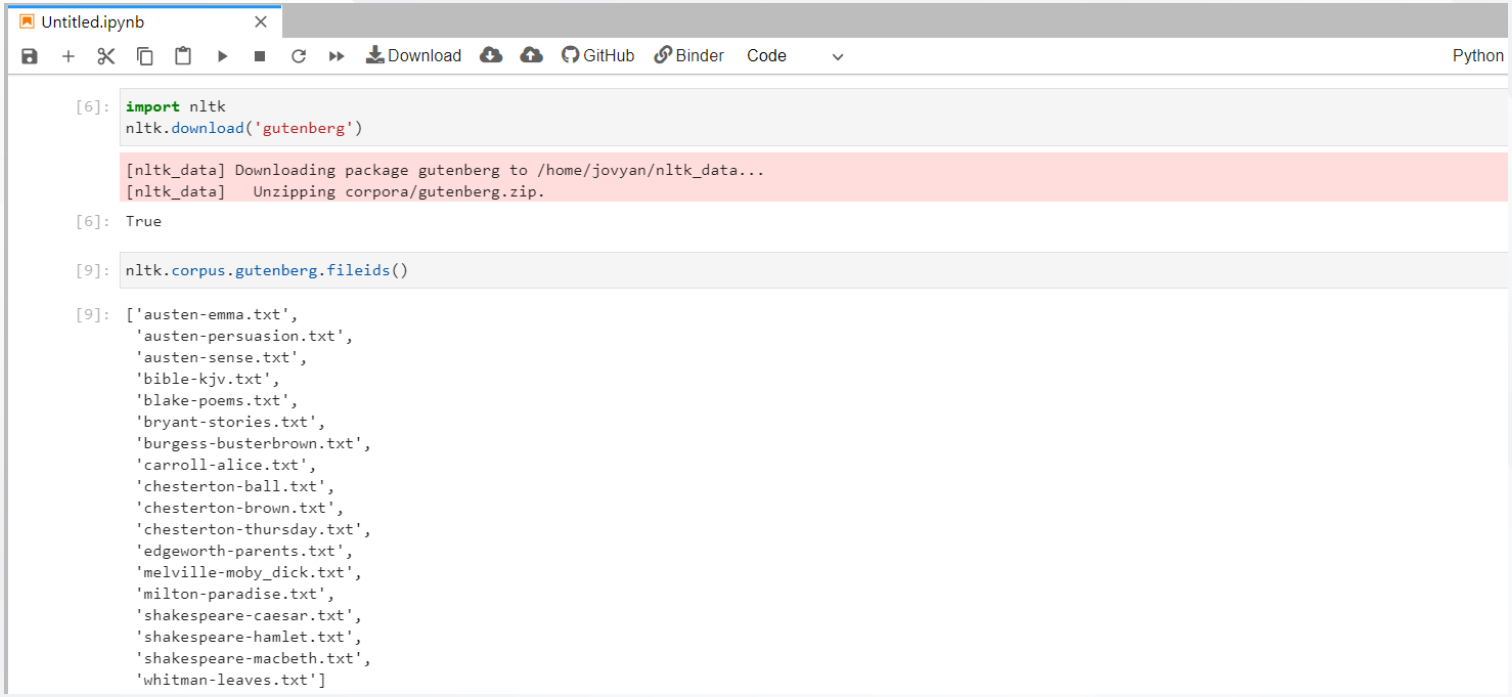
El Natural Language Toolkit (NLTK) fue creado en 2001 como parte de un curso de lingüística computacional del Departamento de Ciencias de la Computación e Información en la Universidad de Pennsylvania. Como lo puedes ver en la tabla 1, NLTK está compuesto de varios módulos que corresponden al tipo de tarea de procesamiento de lenguaje que resuelve.

Tarea de procesamiento de lenguaje	Módulo NLTK	Funcionalidad
Acceso a corpus	corpus	Interfaces estandarizadas para corpus y léxicos
Procesamiento de cadenas de texto	tokenize, stem	División de cadenas de texto (*tokenizadores*) y extracción de raíces de palabras (*stemmers*)
Descubrimiento de colocación	collocations	Prueba t, chi-cuadrada, información mutua puntual
Etiquetado POS (partes de la oración)	tag	n-gramas, HMM, TnT, backoff
Aprendizaje automático	classify, cluster, tbl	Árboles de decisión, máxima entropía, Bayes, k-means
Fragmentación	chunk	Expresiones regulares, n-gramas y entidades nombradas
Analizadores sintácticos	parse, ccg	Gráficos, basados en características, unificación, probabilísticos y dependencias
Interpretación semántica	sem, inference	Validación de modelos, lógica de primer orden y cálculo lambda
Métricos de evaluación	metrics	Precisión y recall
Probabilidad y estimaciones	probability	Distribuciones de frecuencias y distribuciones de probabilidad
Aplicaciones	app, chat	Chatbots, navegador WordNet, analizadores sintácticos
Lingüística	toolbox	Manipulación de datos en formato Toolbox de SIL

Para poder utilizar NLTK, es necesario instalarlo previamente en el ambiente de trabajo. Crea un nuevo notebook en Jupyter e instala NLTK.

Explicación

Para acceder a este corpus, se debe cargar la librería NLTK previamente instalada y descargar el corpus del Proyecto Gutenberg. Adicional a ello, es posible enlistar los archivos incluidos en el corpus.



El Corpus Brown fue el primer corpus electrónico de un millón de palabras en inglés, creado en 1961 en la Universidad de Brown. Contiene textos de 500 fuentes y las fuentes se han categorizado por género, como noticias, editoriales, etc.

Explicación

Una forma de identificar cuáles son las palabras que proporcionan más información sobre un texto es contando las veces que la palabra se utiliza dentro del texto. La distribución de frecuencias, indica la periodicidad de cada elemento de un vocabulario en un texto. El término distribución indica cómo se reparte el número total de palabras en el texto.

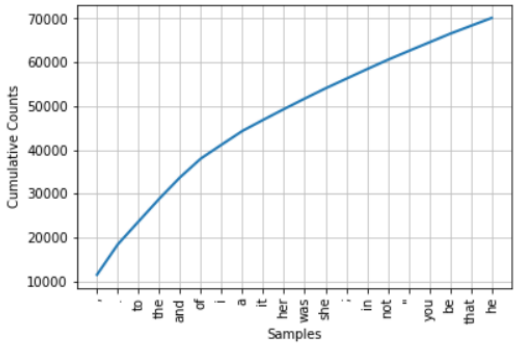
Es posible que de las 20 palabras más frecuentes se pueda inferir algo sobre el texto, por ejemplo, las palabras *her* y *she* aparecen cerca de 2500 veces cada una y eso puede sugerir que en el texto se habla de una persona del sexo femenino.

Si se grafica la distribución de frecuencias acumulada, se observa que tan solo esas 20 palabras más frecuentes son casi la mitad del libro.

recursos-lexicos.ipynb

Download GitHub Binder Code Python 3

```
[25]: dist_frec.plot(20, cumulative=True)
```




```
[25]: <matplotlib.axes._subplots.AxesSubplot at 0x7f156a4450d0>

[28]: int(len(palabras)/2)

[28]: 96213

[ ]:
```

Actividad

01 ¿Cuáles son los corpora y recursos léxicos que serían útiles para analizar la forma de escribir entre autores, los periodos en los que se escribieron ciertos textos o el género del autor que lo escribió?

02 Te seleccionan para construir un analizador de sentimientos cuyo objetivo es predecir la opinión (negativa o positiva) expresada en un texto. ¿Qué información deberá tener el analizador para etiquetar un texto como positivo o negativo? ¿Cuál sería la entrada del analizador? ¿Cuál sería su salida?

Cierre

En este tema se ha presentado el alcance de la información textual a la que se puede tener acceso, utilizando un corpus, con algunas instrucciones de Python y la librería NLTK.