Ingeniería Informática

FACULTAD DE INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

OPTATIVA

Recuperación Avanzada de Información

Dr. J. Federico Medrano
@jfedemedrano

Unidad N° 4 – Parte 1

Temas a desarrollar

- Realimentación y expansión de consultas.
- Técnicas de representación de documentos: TF-IDF, LSI/LSA, WDM.
- Procesamiento del Lenguaje Natural (PLN).
- Resumen automático de documentos.

Expansión de consultas (*Query Expansion*-QE)

- Un Sistema de Recuperación de Información se compone a partir de diferentes módulos, uno de ellos es aquel que permite expandir las consultas ingresadas por el usuario a fin de **ampliar el espectro de las búsquedas** y de documentos a los que pueda acceder el SRI
- Permite la incorporación de diversos términos a la consulta original.
- Como resultado se obtiene un nuevo conjunto de consultas con términos adicionales, denominadas expansiones, estas nuevas consultas serán ejecutadas sobre las fuentes de datos del SRI generando conjuntos de resultados que luego serán unificados y procesados, generando un único listado para el usuario

Expansión de consultas

- Se identifican diferentes métodos para realizar expansión de consultas, cada uno de ellos haciendo uso de técnicas y herramientas diferentes como ser: tesauros, diccionarios, sistemas expertos, entre otros.
- Uno de los problemas en la construcción de un método de expansión de consultas para un SRI, es el tratamiento que se realice del lenguaje en el que el usuario escribe las consultas (estrategias):
  - ❑Emparejamiento de términos entre consultas y documentos sin traducción.
  - ❑Traducción de las consultas al idioma de los documentos.
  - ❑Traducción de los documentos al idioma de las consultas.
  - ❑Traducción de los documentos y las consultas a un lenguaje común.

Expansión de consultas

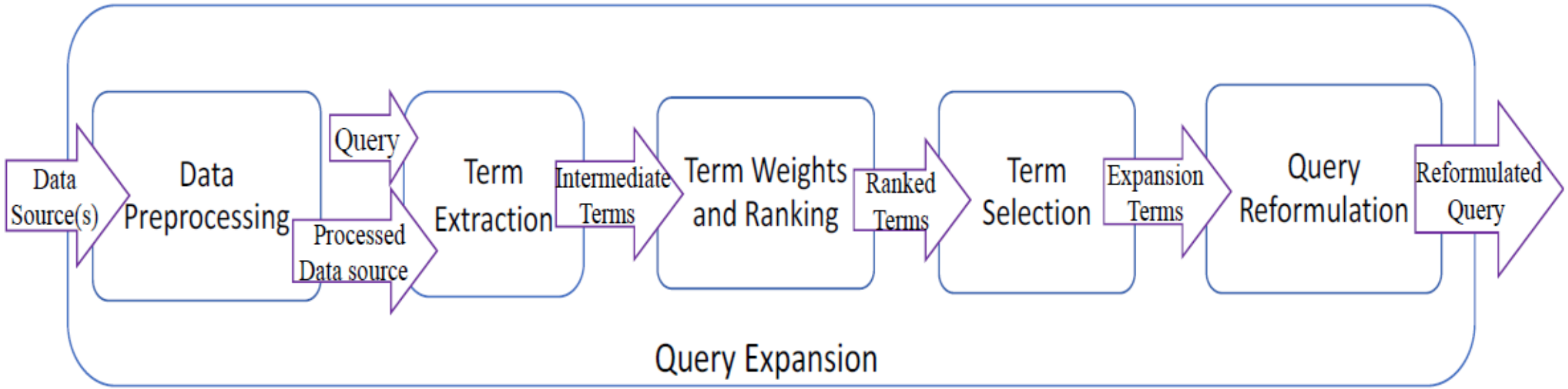
- La expansión de consultas reformula la consulta original del usuario para mejorar la efectividad de la recuperación de información.
- Una consulta está formada por *n* términos. $Q = \{t_1, t_2, ..., t_i, t_{i+1}, ..., t_n\}$

La consulta reformulada puede tener 2 componentes: nuevos términos $T'=\{t'_1, t'_2, ..., t'_m\}$ del origen de datos D y palabras vacías a eliminar $T''= \{t_{i+1}, t_{i+2}, ..., t_n\}$. La consulta reformulada quedaría:

$$Q_{exp} = (Q - T'') \cup T'$$
$$= \{t_1, t_2, ..., t_i, t'_1, t'_2, ..., t'_m\}$$

El aspecto clave de *QE* es el conjunto T': conjunto de nuevos términos significativos agregados a la consulta original del usuario para recuperar documentos más relevantes y reducir la ambigüedad.

Proceso


Data Source(s)
Data Preprocessing
Query
Processed Data source
Term Extraction
Intermediate Terms
Term Weights and Ranking
Ranked Terms
Term Selection
Expansion Terms
Query Reformulation
Reformulated Query
Query Expansion

LSI-LSA

- **Latent Semantic Analysis**, Análisis de la Semántica Latente, Análisis Semántico Latente.
- **Latent Semantic Indexing**, Indexado de la Semántica Latente
- El LSA es un tipo de análisis computacional que permite determinar y cuantificar la similitud semántica entre piezas textuales -sean palabras, documentos o palabras y documentos- de un corpus de textos pertenecientes a un mismo **dominio de conocimiento**.
- Para ello, el sistema computacional del LSA sigue un algoritmo matemático que tiene como centro a la técnica de factorización lineal conocida como descomposición de valores singulares (**SVD**, sigla del inglés **Singular Value Decomposition**), a partir de la cual se genera una representación vectorial del corpus o espacio semántico en cuya conformación y posterior utilización reconocemos dos supuestos lingüísticos acerca del significado: (1) el significado es contextualmente dependiente y (2) en el uso contextual hay relaciones de similitud semántica que están latentes.

LSI-LSA

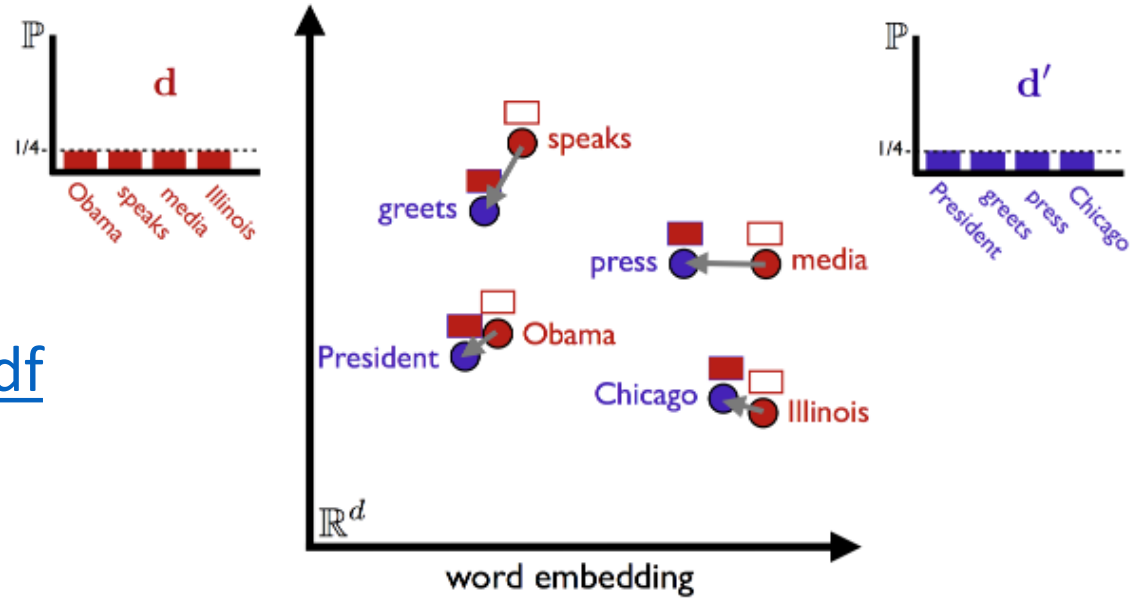
- El LSA tiene su origen en la LSI (del inglés Latent Semantic Indexing), método automático de recuperación de información (e.g. los buscadores de Internet) que incorpora la descomposición de valores singulares (SVD) con el propósito de superar las dificultades semánticas, generadas por la **sinonimia** y la **polisemia**, en la correlación de las palabras que las personas emplean para las búsquedas y los documentos (contextos verbales) contenidos en las bases de datos.
- LSI y LSA difieren principalmente en lo que respecta a su definición de contexto. Para LSI es un documento, mientras que para LSA es más flexible, aunque a menudo es un párrafo de texto. Si la unidad de contexto en LSA es un documento, LSA y LSI se convierten esencialmente en la misma técnica.
- https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-09342005000300003
- http://cs.uns.edu.ar/~agm/mineriaweb/downloads/Slides/clase18-slides-marcelo-amaolo.pdf

WMD

- **Word Mover's Distance** es un método que nos permite evaluar la "distancia" entre dos documentos de manera significativa, incluso cuando ***no tienen palabras en común***. Utiliza las incorporaciones de palabras del vector word2vec.
- Se basa en resultados recientes en **incrustaciones de palabras** que aprenden representaciones semánticamente significativas de palabras de coincidencias locales en oraciones.
- Esta técnica de incrustación demuestra que las relaciones semánticas a menudo se conservan en operaciones de vectores en vectores de palabras.
- La distancia entre dos documentos de texto A y B se calcula por la distancia acumulativa mínima que las palabras del documento de texto A deben viajar para coincidir exactamente con la nube de puntos del documento de texto B

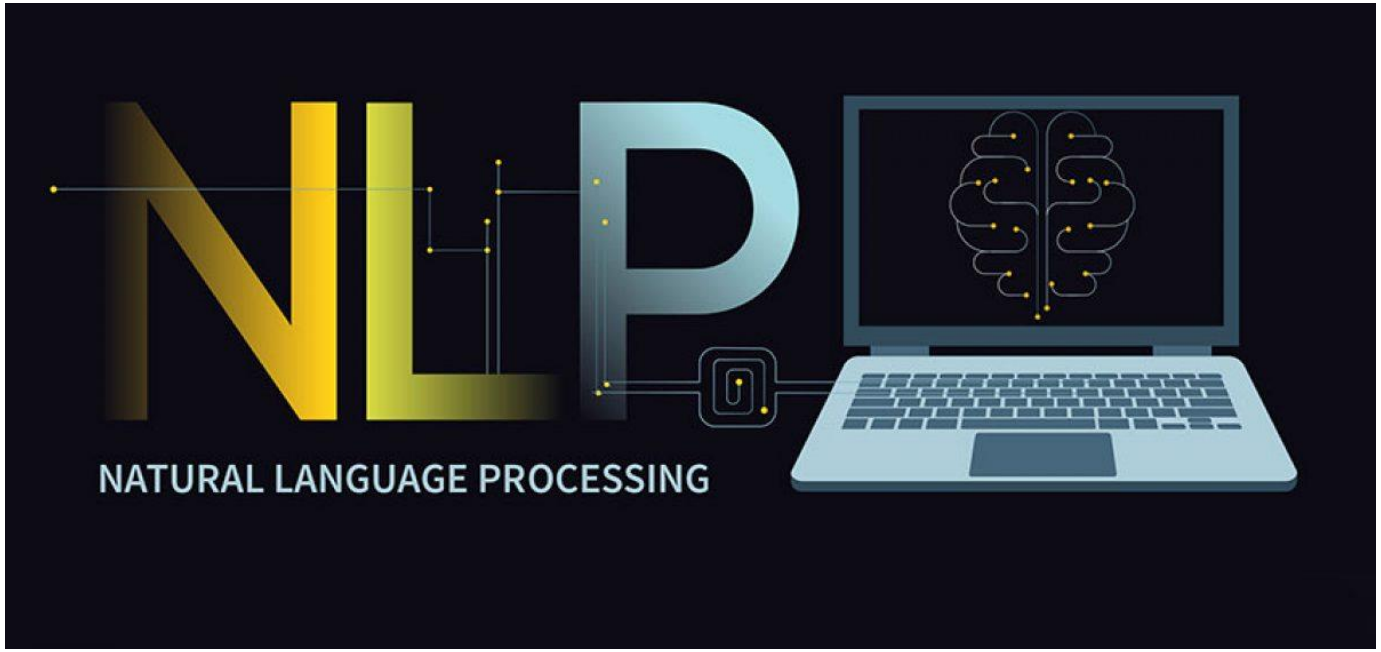
WMD

- Las oraciones no tienen palabras en común, pero al unir las palabras relevantes, WMD es capaz de medir con precisión la (des) similitud entre las dos oraciones. El método también utiliza la representación de bolsa de palabras de los documentos (simplemente, las frecuencias de las palabras en los documentos), anotadas como **d** en la figura. La intuición detrás del método es que encontramos la "distancia de viaje" mínima entre los documentos, en otras palabras, la forma más eficiente de "mover" la distribución del documento 1 a la distribución del documento 2.



- http://proceedings.mlr.press/v37/kusnerb15.pdf

Introducción
a PLN

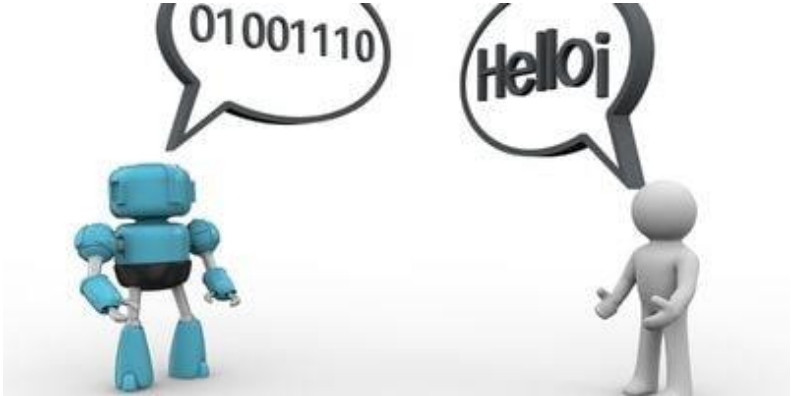


¿Qué es el Procesamiento del Lenguaje Natural (NLP)?

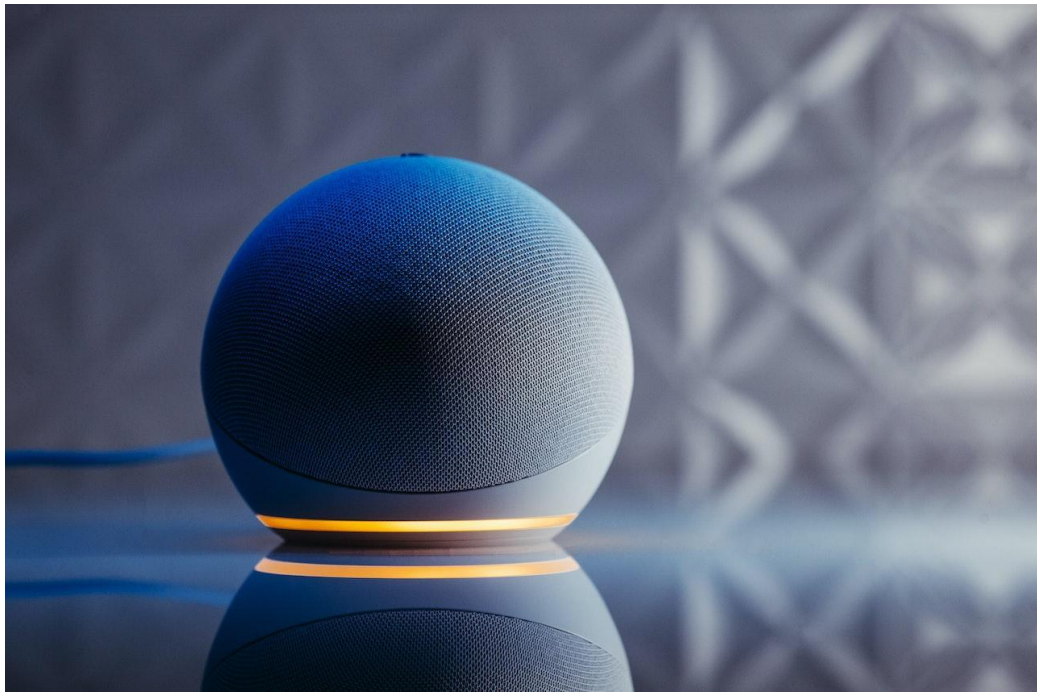
- El término Procesamiento del Lenguaje Natural (*Natural Language Processing*) abarca un amplio conjunto de técnicas para la generación, manipulación y análisis automatizados del lenguaje natural.
- Aunque la mayoría de las técnicas de PLN provienen en gran medida de la Lingüística e Inteligencia Artificial, también están influenciados por tecnologías relativamente más nuevas como Machine Learning, Estadística Computacional y Ciencia Cognitiva.
- El PLN se ocupa de la ***interacción*** entre las ***computadoras*** y el ***lenguaje humano***. Su objetivo es permitir que las máquinas comprendan, interpreten y generen lenguaje humano de manera natural

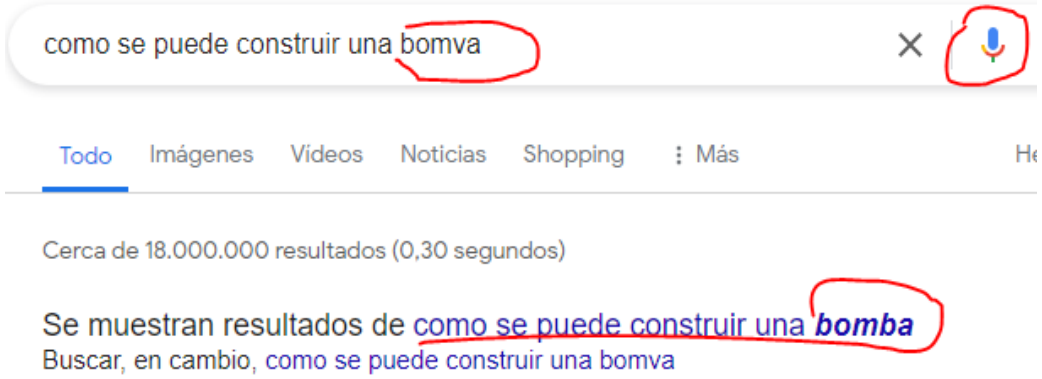
¿Por qué es necesario?

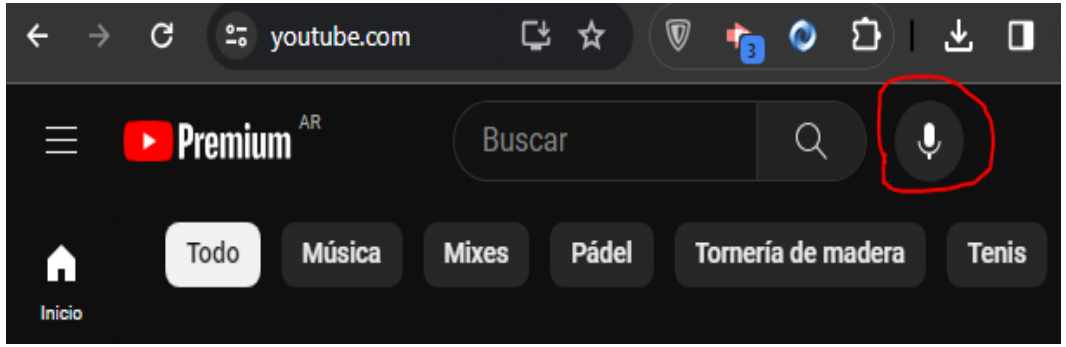
- El lenguaje natural es el lenguaje humano, la lengua que un grupo o comunidad de gente ha establecido de manera espontánea para comunicarse entre ellos, transmitir pensamientos, ideas y conceptos así como para referenciar el mundo que les rodea.
- **El lenguaje humano** es una de las características más relevante, si no la principal, que **nos distingue a los humanos del resto de seres vivos**.
- La ambigüedad del lenguaje es la característica más compleja y, por ello, el entendimiento del lenguaje es difícil de modelar.



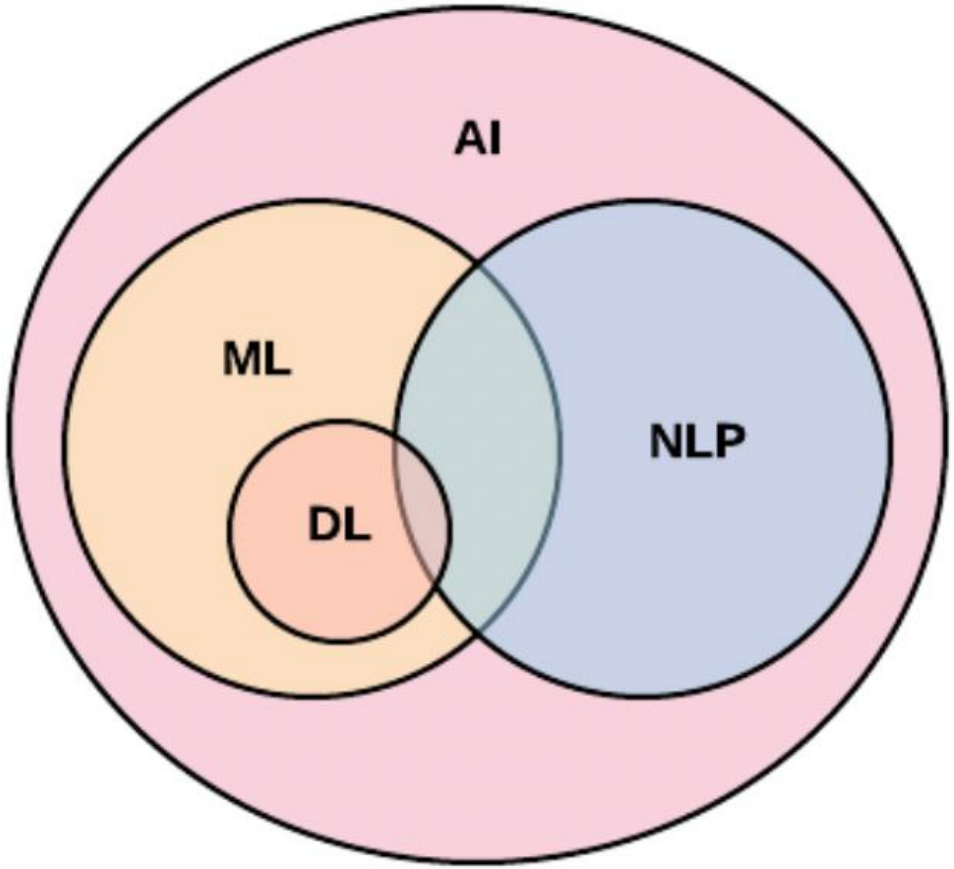
Está por todos lados…







¿Dónde se ubica el Procesamiento del Lenguaje Natural (NLP)?



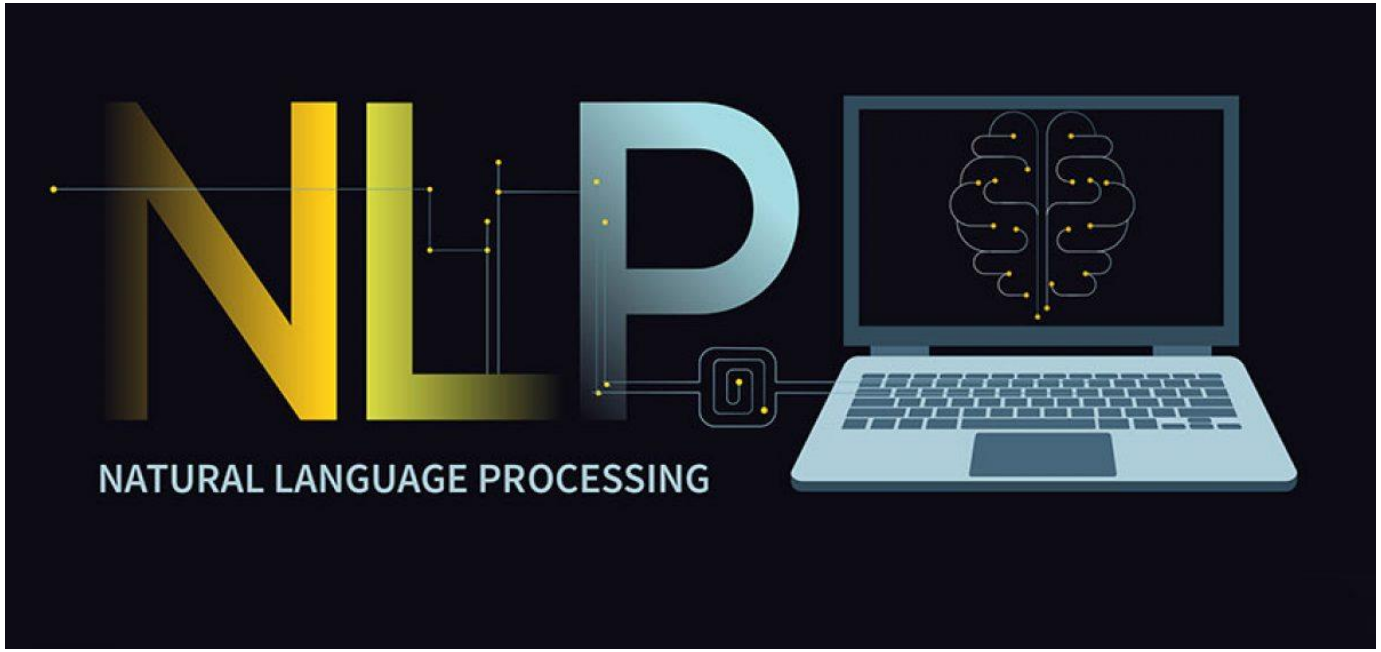
IA : Inteligencia Artificial

ML : Aprendizaje Automático

DL : Aprendizaje Profundo

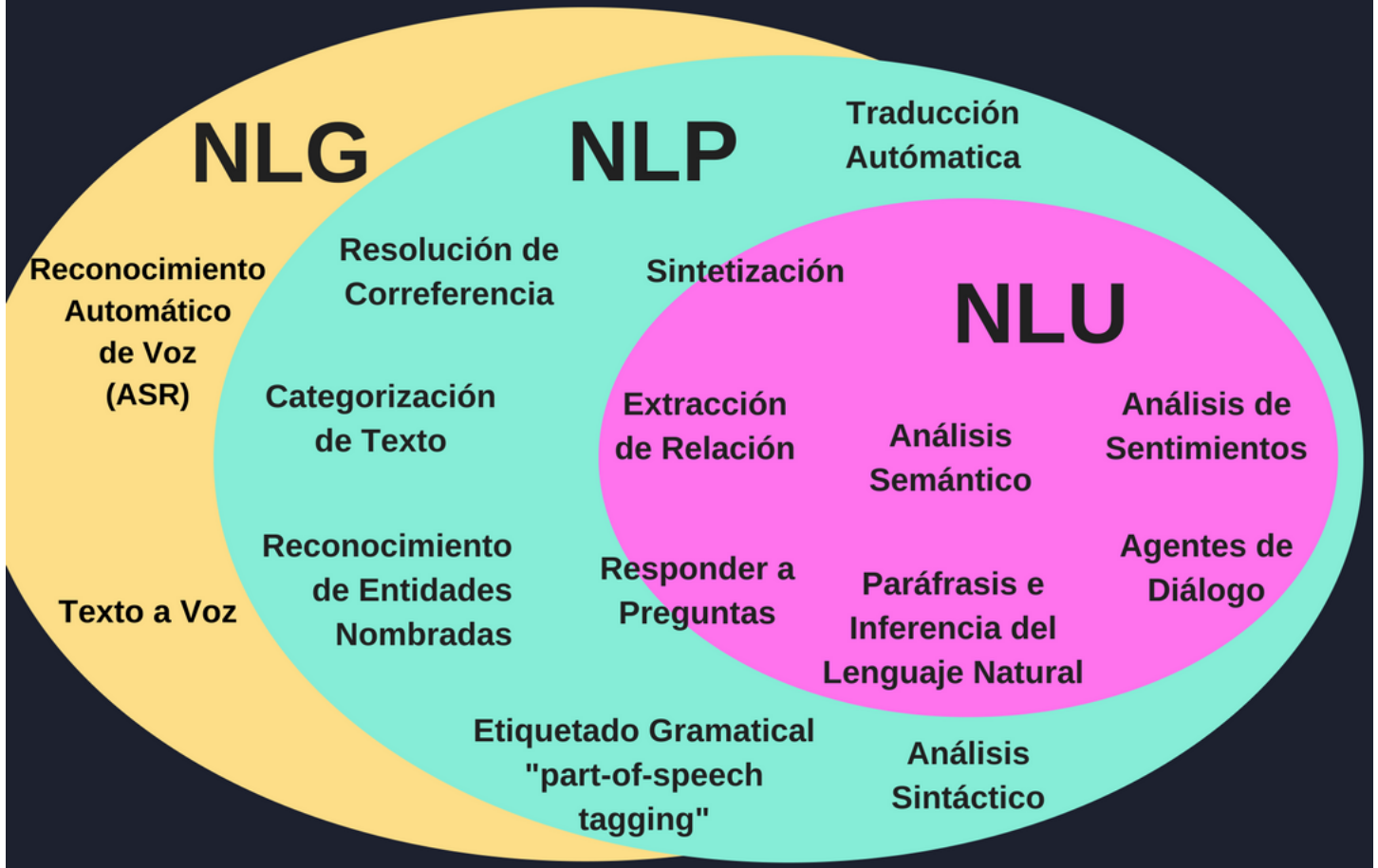
NLP : Procesamiento del Lenguaje Natural

Aplicacionesd el PLN



NLP v.s. NLU v.s. NLG

NLP - Procesamiento de Lenguaje Natural
NLU - Entendimiento de Lenguaje Natural
NLG - Generación de Lenguaje Natural



Fuente: https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf

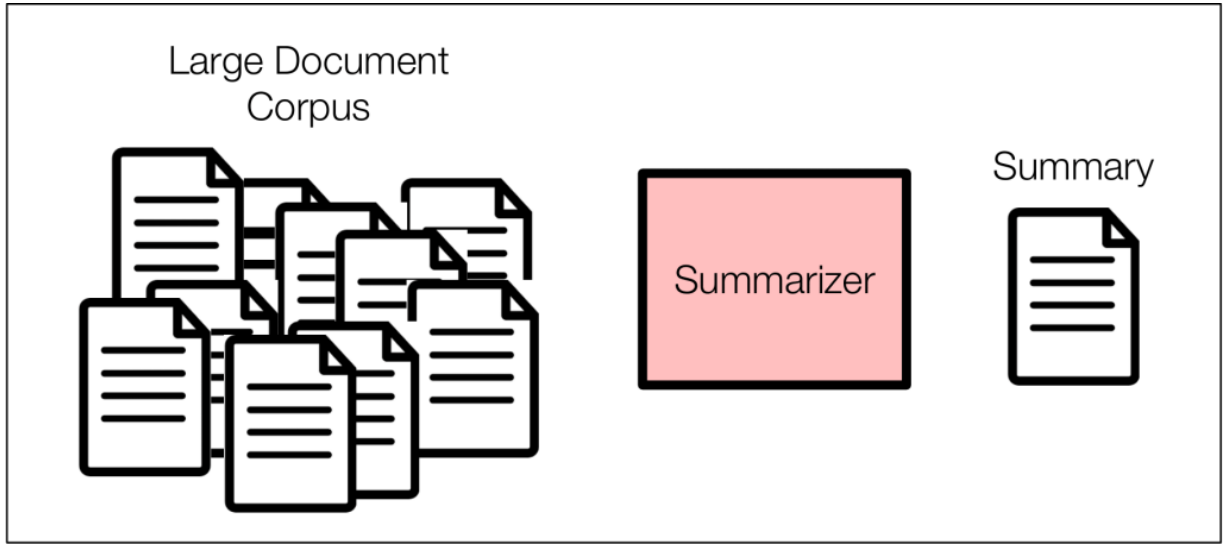
Aplicaciones Procesamiento del Lenguaje Natural

- **Análisis morfosintáctico** (*PoS tagging* en inglés) que etiqueta las palabras según su papel sintáctico (adverbio, sustantivo, verbo)
- ***Chunking* o parseado poco profundo** (*shallow parsing*) que asigna una etiqueta según los grupos relacionados sintácticamente en los que se divide el texto (sujeto, verbo, predicado).
- **Reconocimiento de entidades** (*Named Entity Recogition o NER*) que identifica los elementos de las frases en clases de objetos (“persona”, “sitio”, “animal”, “compañía”).
- **Etiquetado semántico** (*semantic parsing*)
- **Desambiguación semántica** (*word sense disambiguation*)
- **Palabras relacionadas semánticamente o sintácticamente** (similaridad): predice si dos palabras están relacionadas semánticamente o sintácticamente .
- **Traducción automática** (*machine translation*). Este, junto con los modelos de lenguaje natural, es una de las tareas insignia en el desarrollo del PLN.
- **Búsqueda de respuestas** (*question understanding*)
- **Reconocimiento del habla** (*automatic speech reconition*)

Aplicaciones Procesamiento del Lenguaje Natural

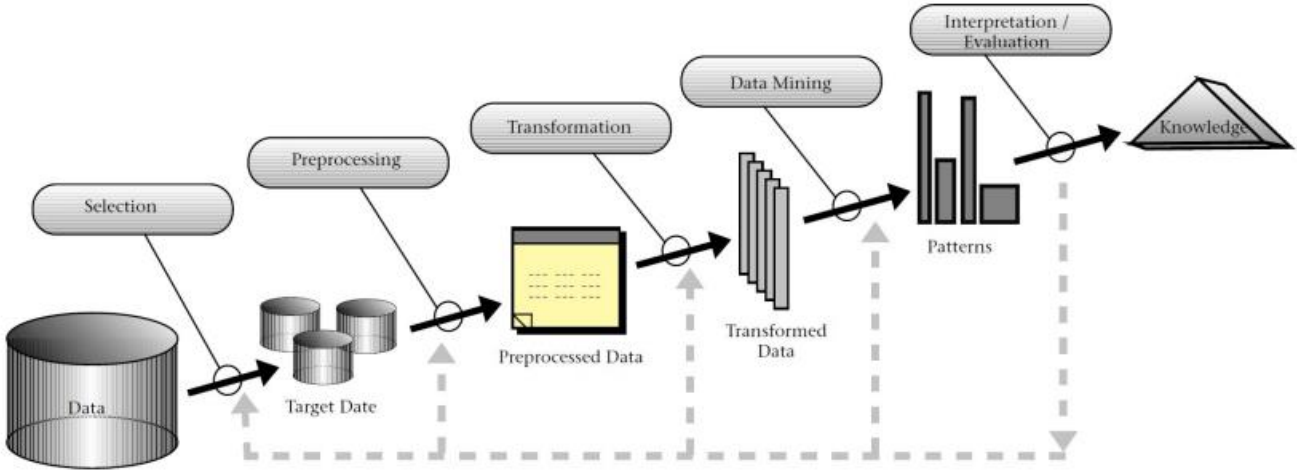
- **Análisis de opiniones/sentimientos** (*sentiment analysis*) Son los sistemas que determinan si un texto tiene connotación negativa o positiva. En las aplicaciones más avanzadas se puede asignar incluso la emoción predominante en un texto
- **Extracción de información** (*information extraction*)
- **Resumen de texto** (*text summarization*)
- **Clasificación de textos** (*text clasification*)
- **Recuperación y extracción de información**: Son sistemas que buscan información en colecciones de documentos de acuerdo con una consulta que puede estar estructurada en lenguaje natural o no. Ejemplo de esto son los buscadores de internet.
- **Sistemas/Agentes de diálogo:** Son las aplicaciones que buscan entablar conversaciones naturales con las personas. Ejemplos son los ***chatbots*** que podemos encontrar en sitios web.
- **Corrección automática de ortografía y gramática**

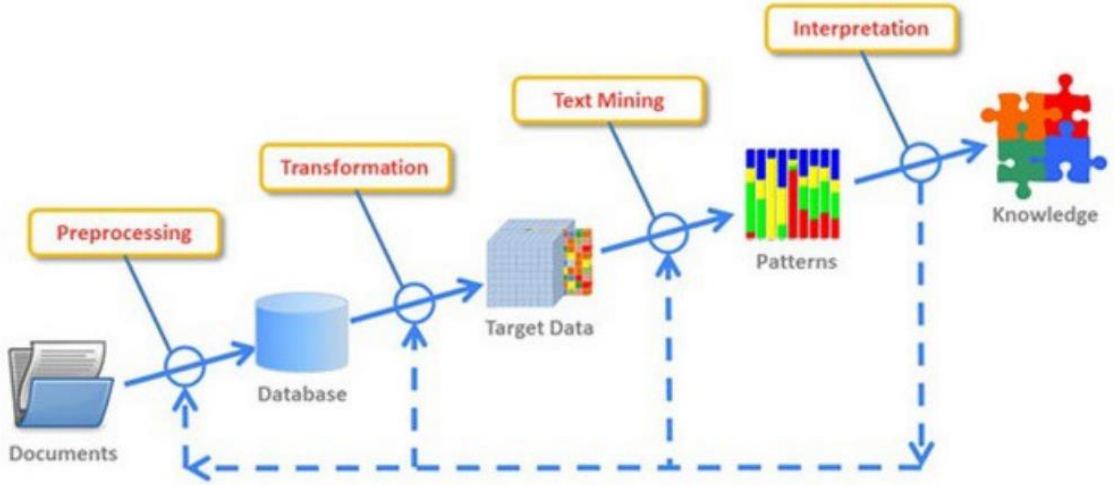
Resúmenes Automáticos



Introducción

- A diferencia del proceso de KDD (Knowledge Discovery in Databases) tradicional, el proceso de descubrimiento de conocimiento en textos (KDT Knowledge Discovery in Text, por sus siglas del inglés) parte únicamente de texto en forma de documento y se lo preprocesa utilizando un conjunto de tareas que se aplican exclusivamente a dicho tipo de dato





Introducción

- El interés en la **minería de textos** ha crecido enormemente en los últimos años, debido a la creciente cantidad de documentos disponibles en forma digital y la también creciente necesidad de organizarlos y aprovechar el conocimiento contenido en ellos.
- Un **resumen** es una transformación **reductiva** de un texto fuente a un texto resumen por reducción de su contenido mediante selección y/o generalización de lo que es importante en el texto fuente.
- La generación automática de resúmenes de texto es el proceso mediante el cual se crea utilizando una computadora una “versión reducida” de uno o más documentos con el contenido relevante. Omitir oraciones o incluso párrafos completos sin sufrir pérdida de información es algo difícil de conseguir en forma automática.

Introducción

- Entonces, un documento de texto es una unidad de datos textual que usualmente, aunque no necesariamente, se corresponde con algún documento del mundo real.
- Pueden resumirse informes, monografías, reportes, e-mails, libros, comunicados, historias clínicas, trabajos científicos, cartas, expedientes, entre otros.
- En todos los casos, **la lectura del resumen facilita la lectura del documento original**. Su extensión jamás iguala ni mucho menos supera la del documento original.

Tipos de resúmenes

- Los **resúmenes extractivos** están formados por “partes” del documento que fueron seleccionadas apropiadamente. Los **resúmenes abstractivos**, por otro lado, están formados por las “ideas” desarrolladas en el documento, sin hacer uso de las frases exactamente como aparecenen el documento original.
- En la literatura, la mayor parte de la atención se ha puesto en seleccionar del documento, mientras que generar un nuevo texto ha recibido menos atención de alguna manera.
- El extractivo es el enfoque más utilizado, ya que no tiene que reescribir el texto ni tampoco tiene que garantizar la coherencia narrativa de lo seleccionado.
- Este tipo de resumen ofrece tres ventajas: (1) el tamaño del resumen puede controlarse, (2) el contenido del resumen se obtiene con precisión, y (3) puede ubicarse fácilmente en el documento fuente

Tipos de resúmenes

- Un resumen **extractivo** está formado por un conjunto de porciones de texto (desde palabras sueltas hasta párrafos enteros) literalmente copiadas de la entrada, a las que llamaremos a partir de ahora “sentencias”, haciendo referencia a oraciones únicamente. Este enfoque extrae partes de un documento sin requerir un análisis semántico complejo.
- En cambio, para construir un resumen por **abstracción** se necesitan recursos de conocimiento lingüístico específicos tales como *ontologías*, *tesauros* y *diccionarios* para comprender el contenido. Este enfoque resulta más complejo de llevar a cabo. Generalmente, utiliza algunas pocas tareas de preprocesamiento que no “destruyen” (en el sentido de transformar demasiado) el documento, ya que las palabras y su contexto resultan fundamentales.
- Una vez preprocesado el documento, se crea una representación intermedia que lejos está de ser el documento original.

Mecanismo

- En general, la tarea de resumir automáticamente se puede dividir en dos grandes fases:
  - (a) Construcción de la representación del texto;
  - (b) Generación del resumen, que puede incluir la extracción de oraciones existentes o la construcción de nuevas oraciones.

Mecanismo

- Las estrategias poco profundas en general no analizan el texto más allá del nivel sintáctico y utilizan características superficiales del mismo tales como :
  - **frecuencia de los términos**: tales medidas estadísticas pueden capturar el tema del texto, asumiendo que las frases importantes son las que contienen palabras que ocurren frecuentemente en el documento
  - **ubicación**: la intuición aquí es que las frases importantes están situadas en ubicaciones particulares, que dependen del género del texto, aunque hay algunas reglas que podrían ser generales como tomar las primeras frases del texto o los encabezados
  - **sesgo**: la relevancia de ciertas frases puede depender de que incluyan términos que aparecen en el título o en encabezados del documento, o hasta en la consulta del usuario que requiere el resumen.
  - **palabras clave** como “en resumen”, “en conclusión”, o este trabajo describe” u otras dependientes del dominio pueden señalar la relevancia (o irrelevancia) de una cierta frase en el texto.

Mecanismo

- Otras estrategias exploran el texto a mayor profundidad, modelando las entidades que aparecen en el texto y sus relaciones, o hasta al nivel de discurso, modelando la estructura global del texto y su relación con metas de comunicación.
- En general, las estrategias poco profundas generan resúmenes que son extractos, mientras que las más profundas son necesarias para generar abstractos