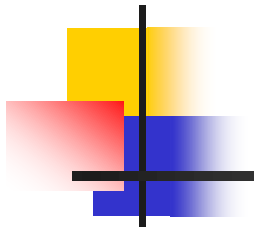


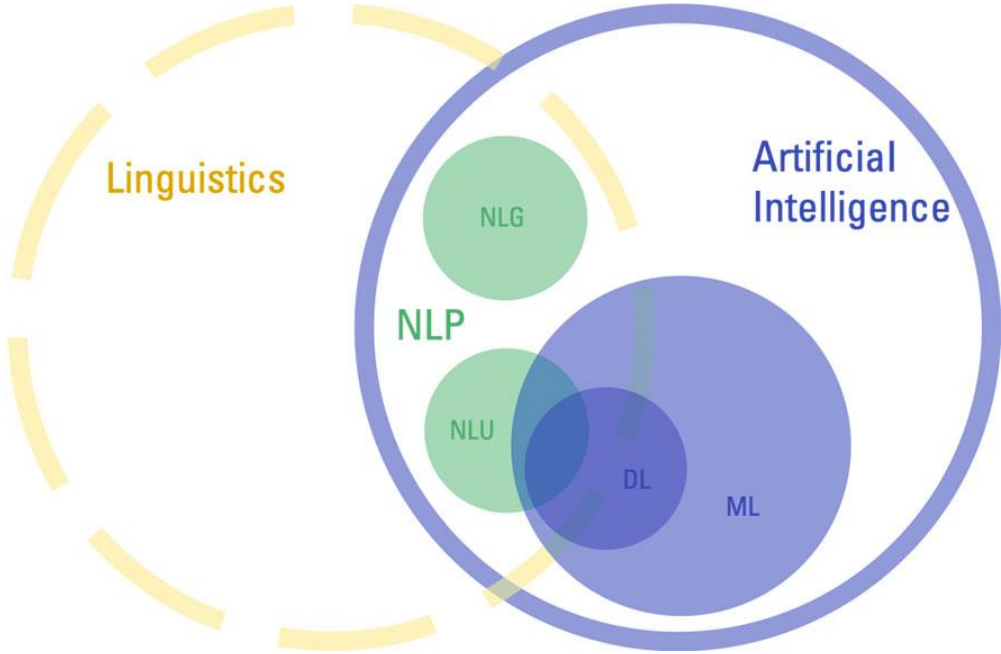
Procesamiento de lenguaje natural

CURSO GRUPO BANCOLOMBIA

Universidad Nacional de Colombia

Procesamiento de lenguaje natural

- El procesamiento de lenguaje natural (NLP) estudia la interacción humano-computador, a través de una mezcla de IA y lingüística.
- Se entiende como “natural” a los idiomas de uso común (español, inglés, etc.) y no a los lenguajes creados artificialmente para las máquinas.
- NLP tiene diferentes fases, desde identificar las palabras, contarlas o encontrar su relevancia, hasta “entender” semánticamente el contexto de las mismas, o analizar sentimientos.
- La mayoría de información del mundo se encuentra en datos no estructurados como documentos, manuales, informes, artículos, etc., los cuales no son fácilmente entendibles para las máquinas.



ML=Machine learning, DL=Deep Learning, NLU=Natural Language Understanding (lectura de lenguaje), NLG=Natural Language Generation (generación de lenguaje)

Procesamiento de lenguaje natural

- Aplicaciones del NLP incluyen conversión de textos a formatos estructurados, chatbots, traducción, filtros de spam, corrección de ortografía, generación automática de informes y reportes, conversión de voz a texto, etc.

- Una de las herramientas más utilizadas para el modelamiento de problemas de NLP son las redes neuronales, gracias a su capacidad para adaptarse fácilmente a problemas complejos.

- Keras y scklearn incluyen diferentes herramientas en Python que facilitan el procesamiento de lenguaje natural. Existen además muchas librerías adicionales para realizar operaciones básicas como separar palabras, encontrar sinónimos, lematizaciones, etc.

Stopwords

- El primer paso para aplicar NLP sobre un texto es eliminar las palabras vacías o *stopwords.* Estas palabras depende de cada idioma y son aquellas que no aportan un significado importante en un texto, además de que son muy comunes.
- Eliminar las *stopwords* es un paso importante ya que permite a la máquina centrar su atención en el análisis de las palabras con mayor semántica.
- En este grupo se incluyen pronombres, artículos, conjunciones, preposiciones, etc. Algunos ejemplos para el idioma español pueden ser: a, al, con, de, el, es, esta, ha, me, mi, que, tu, y, etc.
- Supongamos que tenemos dos oraciones: “yo vivo feliz” y “ser feliz es vivir pleno”. Después de realizar la eliminación de las *stopwords*, tendríamos los grupos “vivo feliz” y “feliz vivir pleno”.

Lematización

- Otro paso importante en el procesamiento de lenguaje natural es la lematización. Este proceso consiste en encontrar el lema a partir de una palabra cuya forma puede estar flexionada (gerundio, plural, femenino, conjugación, etc.).

- El lema es la palabra que se encuentra en el diccionario o enciclopedia. Por ejemplo, el lema de “hablamos” es “hablar”, el lema de “peones” es “peón”, etc.

- Para los conjuntos de palabras “vivo feliz” y “feliz vivir pleno” se obtendrían los siguientes lemas: “vivir feliz” y “feliz vivir pleno”

Vectorización

- Las palabras se representan como vectores de números con el objetivo de utilizar métodos de machine learning como las redes neuronales.

- Se puede indicar con 0 y 1 si una palabra está o no en una oración.

- Para el ejemplo en cuestión tendríamos:

Oración	Vivir	Feliz	pleno
vivir feliz	1	1	0
feliz vivir pleno	1	1	1

Tf-idf

- Una medida importante utilizada en NLP es el Tf-idf (*term frequency – inverse document frequency*), la cual indica la relevancia de un palabra para un documento en una colección.
- Un documento puede ser una oración, un párrafo, un texto, una página de libro, un libro, etc. Una colección sería un conjunto de los documentos que hayamos elegido.
- Si una palabra aparece varias veces en un documento, el tf-idf aumenta; pero si aparece muchas veces en la colección, su tf-idf disminuye. Esta idea trata de compensar el hecho de que hay palabras mucho más comunes que otras y aparecerán en muchos documentos

Oración	Vivir	Feliz	pleno
vivir feliz	0.707	0.707	0.0
feliz vivir pleno	0.501	0.501	0.707

Deep learning

- Una vez se tenga la representación numérica de los documentos, se puede proceder a aplicar aprendizaje profundo para la clasificación de los mismos.
- La capa de entrada de la red neuronal tendrá tantas variables como palabras diferentes existan. Normalmente se suele limitar a un máximo de caracteres y se deben llenar todos los documentos con ceros para las posiciones restantes. El ejemplo en cuestión se podría representar como: [[1, 1, 0], [1, 1, 1]]
- La capa de salida dependerá de las variable de clasificación utilizada. Por ejemplo, si el texto se va a clasificar en tres posibles géneros (ciencia ficción, novela, historia), entonces la capa tendrá tres neuronas de salida.
- Existen diversas configuraciones posibles de la red neuronal, utilizando redes convolucionales, redes recurrentes, memoria a largo plazo, etc. Puede ser un trabajo difícil encontrar la mejor configuración posible. Se recomienda iniciar con una configuración sencilla y ver los resultados del modelo; luego realizar ajustes de parámetros o variaciones en las capas y observar si hay mejoras

¡A codificar!