

Natural Language Processing

El '**NLP**' es aplicar modelos de '**Machine Learning**' a texto y lenguaje para que el '**Agente**' intente comprenderlos transformando patrones de texto a números.

El '**NLP**' en la manera en como el '**lenguaje informático**' interactúa con el '**lenguaje humano**'.

Secuencia de 8 pasos para crear y evaluar un algoritmo '**NLP**'

1. Limpieza gramatical

- Con '**Expresiones Regulares**' limpiar los '**strings**' (frases) de números, signos, exclamaciones, comas, puntos, etc. Finalmente solo quedan las palabras de la frase separadas por espacio.

2. Pasar a minúsculas

- Los caracteres de los '**strings**' (frases) son evaluados y pasados a minúsculas; esto es para una mejor manipulación de los datos por parte del algoritmo.

3. Transformar '**string**' (frase) en un array de palabras

- Cada '**string**' (frase) es evaluado para separar cada palabra cortando los espacios, y luego agregándolas a un array.

4. Quitar las '**stopwords**'

- En base a un diccionario de palabras inútiles (**stopwords**) evaluado mundialmente, se valida que el DataSet no contenga ninguna de estas palabras, se eliminan.

5. Estemizar DataSet

- Evaluar cada palabra del DataSet y reducirla a su raíz gramatical (sin tiempo, conjugaciones, etc.).

6. Reformar fila: 'array' a 'string'

- Cada fila que esta formada por un array de palabras, es re-formada a un solo '**string**'.

7. Bag of Words (bolsa de palabras)

- Después de los filtros previos hechos a las palabras del DataSet, finalmente se crea una matriz numérica (0 y 1) con filas igual a la cantidad de datos, y columnas igual a la cantidad de palabras.

8. Usar un modelo de clasificación

- Con un modelo/método de '**Classification**' evaluar el algoritmo '**NLP**'. La '**Matriz de Confusión**' funciona muy bien.