

**“Algoritmo de clasificación de noticias mediante técnicas de procesamiento de lenguajes naturales”**

**Autor**: Estudiante. Rios, Leandro Agustin.

**Institución**: Colegio Universitario IES 21

**Carrera**: Tecnicatura superior en ciencia de datos e inteligencia artificial.

**Materia**: Procesamiento de lenguajes naturales II

**Docente**: Lic. Sufotinsky, Alejandro Julio

Córdoba, Argentina.

2022

# 1.Introducción

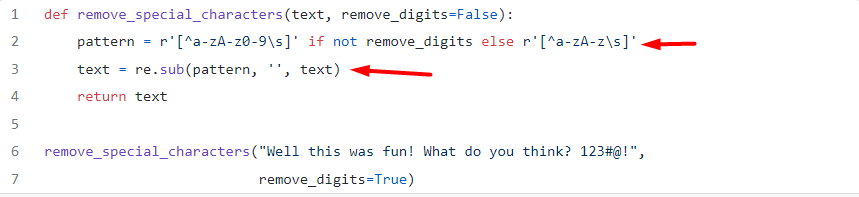
El procesamiento de lenguajes naturales con la programación está teniendo un gran crecimiento en distintas industrias. De esta forma las computadoras están pudiendo entender mejor la forma de comunicarse de los humanos. Pudiendo descifrar lo que están comunicando. Mediante noticias, documentos, tuits, etc.

Para resolver este problema nos vamos a basar en un sistema de scoring que cuente con técnicas de NLP para la limpieza de los documentos. Al finalizar con estos pasos tendremos un puntaje con el cual clasificaremos.

# 2. Marco teórico

Como Sarkar nombre en su publicación llamada “Guía práctica para el procesamiento del lenguaje natural (Parte I) — Procesamiento y comprensión del texto” (2018). Citando, “Los datos no estructurados, especialmente los textos, las imágenes y los videos, contienen una gran cantidad de información”. A su vez, Sarkar, expone una preocupación respecto de la falta de profesionales que se atrevan a adentrarse en el mundo del procesamiento de lenguajes naturales, debido a su complejidad en el procesamiento y análisis de datos (2018).

Sarkar en esta publicación dice lo siguiente sobre el pre-procesado del texto “...por lo general hay múltiples pasos involucrados en la limpieza y el preprocesamiento de los datos textuales“. Como primera medida nombra que debemos eliminar los caracteres acentuados. También comenta que debemos eliminar los caracteres especiales usando alguna expresión regular. Como por ejemplo la que propone Sarkar es:



Aclarando que cada expresión debe de adecuarse al contexto y a la estrategia que se está siguiendo para llevar ese proyecto a cabo.

El próximo paso que nombra es la eliminación de palabras claves, el dice “Las palabras que tienen poco o ningún significado, especialmente cuando se construyen características significativas a partir del texto, se conocen como stopwords o stop words. Por lo general, se trata de palabras que terminan teniendo la frecuencia máxima si se hace un término simple o una frecuencia de palabra en un corpus. Por lo general, estos pueden ser artículos, conjunciones, preposiciones, etc.”

# 3. Marco metodológico

Hoy en día poder clasificar noticias puede ser una ventaja para las empresas. Poder entender cuales son los temas más hablados y con eso poder orientarse para generar nuevas noticias. Por esto mismo, nuestro problema a resolver sería poder armar un algoritmo que nos clasifique las noticias en distintas categorías.

En primera instancia para poder realizar el clasificador de noticias. Se debe de armar el conjunto de datos. Para hacer esto estuvimos viendo noticias de diversas páginas y guardarlas en archivos .txt.

Una vez realizado el armado de datos, pasamos a aplicar técnicas de procesamiento de lenguajes naturales, como dice Sarkar “... aventurarse en estos datos puede ser una mina de oro potencial” (2018).

Nos planteamos la siguiente estrategia. Conseguir que el texto esté todo tokenizado y obtener la frecuencia relativa de las palabras.

El objetivo principal sería tener el modelo armado para presentar como nota del primer parcial de esta materia.

Ya teniendo los archivos en txt, lo que hicimos fue generar funciones para poder trabajarlos. Una de esas funciones lo que haces es recorrer una carpeta y va juntando todo el contenido de la misma y almacenando en una variable. Por lo cual nos quedaría una por cada categoría que tenemos armada.

La otra función es la encargada de aplicar el procesamiento de lenguajes naturales. En la misma lo que hace es pasar todo el texto a minúsculas, luego tokenizar todo el contenido y por último eliminamos las stop words.

Al tener esto realizado, lo que hacemos es un conteo de la frecuencia absoluta de las palabras pasándolas en un diccionario con su key que es la palabra y el value que es la cantidad de veces que dicha palabra aparece en el archivo.

Teniendo esto armado, al mismo diccionario le aplicamos la frecuencia relativa de dichas palabras dividiendo el contenido total de palabras sobre el contenido total de tokens.

Con estas funciones listas, ya estamos en condiciones de agregar un texto nuevo que sea para testear lo que hemos armado. Hacemos el mismo procedimiento para el preprocesado de la información. En cambio ahora, vamos a comparar los tokens del nuevo diccionario, noticia nueva, con los token de cada diccionario de categorías.

Cada vez que encuentre una palabra que se repite vamos a ir agregando su value a una variable acumulativa.

Llegamos a la parte del resultado. cuando ya tengamos las distintas comparaciones realizadas. Vamos a poder notar un puntaje asignado para cada tipo de categoría. La que tenga su puntaje más alto significa que esa noticia que se cargó pertenece a ese tipo de noticias.

# 4. Conclusiones

Las conclusiones que pudimos notar es que el algoritmo que realizamos aproximadamente tiene una certeza de un 66.66 % de decir que tipo de categoría es. Es decir, que cada tres noticias que se cargan, dos de ellas están bien y una está mal.

El modelo se puede mejorar más todavía. Lo que podemos hacer para mejorar sus resultados es entrenarlo con más datos. Es decir, ampliar las noticias que tienen cada categoría. Gracias a esto el modelo va a contar con más palabras y su frecuencia cambiaría para cada palabras. A su vez, podemos agregar distintos pasos de procesamiento de lenguajes naturales para mejorarlo. Como por ejemplo ver cómo funciona aplicando stemming a los tokens o lematizandolos también. Estas técnicas no fueron completadas en este modelo.

# 5. Bibliografía

**Dipanjan (DJ) Sarkar - Guía práctica para el procesamiento del lenguaje natural (Parte I) — Procesamiento y comprensión del texto (2018)** <https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72>

Sarkar, D. 2018. *Guía práctica para el procesamiento del lenguaje natural (Parte 1) - Procesamiento y comprensión del texto.* Obtenido de <https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72>