

Exploración y análisis de datos de texto masivos relativos a la crisis del COVID-19

Semillero de Investigación en Inteligencia Artificial K-IA

Luis E. Baquero-Rey – lrey@itc.edu.co
Joseph Rhomandt (jdrhomandtbretolau@itc.edu.co) - Elkin Vargas (eavargasc@itc.edu.co)

Introducción

La pandemia del COVID-19 trajo consigo grandes implicaciones para la humanidad, desde poner el mundo en jaque frente a las restricciones de movilidad, hasta llegar a implicar nuevos retos para los científicos; con el fin de tratar y hacerle frente a la situación. Esto conllevó, entre otras cosas, al desarrollo de nuevas vacunas en tiempo récord, apoyados desde luego, también en áreas de las ciencias de la computación, como lo es la Inteligencia Artificial.

El aprendizaje automático o *machine learning* es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Como apoyo a esta investigación se plantea el uso de algunos de estos algoritmos para examinar texto real de los resúmenes de publicaciones científicas para identificar comportamientos, patrones y superposiciones, especialmente en el contexto de la exploración de datos que implica la temática abordada.

Planteamiento del Problema y justificación

La propagación del Covid 19 cambió profundamente el día a día de los ciudadanos. Si bien la relevancia de las nuevas tecnologías como palancas estratégicas para la resolución de crisis ha sido ampliamente debatida antes de la pandemia, especialmente en el contexto de la exploración de datos,

La exploración y análisis de datos de textos masivos relativos al COVID-19 implica la extracción, tratamiento y carga de grandes volúmenes de datos basados en corpus de noticias y de literatura científica respectiva. Dado que es un proyecto en curso, las fuentes ya han permitido descubrir clusters (agrupaciones) de palabras relacionadas que denoten virus y tratamientos; así como deberá permitir inferir relaciones implícitas que nos permitan saber que el virus X causa la enfermedad Y y qué enfermedad causa el virus Z.

Para el desarrollo de esta investigación se están siguiendo las siguientes fases: Exploración, Descarga, Tokenización, Generación de Word Embedding, Análisis y Representación de resultados.

Metodología

Para el desarrollo de esta investigación se están siguiendo las siguientes fases:

- π Exploración
- π Descarga
- π Tokenización
- π Generación de Word Embedding
- π Análisis y
- π Representación de resultados.

Objetivos

π **Objetivo General**

Realizar un estudio exploratorio y de analítica de datos de texto masivos relativos a la crisis del COVID-19

π **Objetivos Específicos**

Diseñar e implementar una solución de análisis de datos sobre texto masivos relativos al Covid-19.

Descubrir clusters (de palabras relacionadas que denoten virus, tratamientos, para inferir relaciones implícitas basada en corpus de noticias y de literatura científica.

Marco teórico

La literatura emergente sobre inteligencia artificial (IA) y otras tecnologías disruptivas, ha reconocido la importancia de la asimilación tecnológica con beneficios organizacionales operativos y estratégicos y el crecimiento económico, dada la capacidad para transformar y apoyar operaciones en los diferentes sectores productivos [1] así como en el campo de la salud La investigación que se ha producido hasta la fecha aborda la temática del Covid-19 en este sentido el uso de la inteligencia artificial (IA y los algoritmos de aprendizaje automático pueden examinar texto real de los resúmenes de publicaciones para identificar la superposición de [2] Las herramientas de IA se basan en varios enfoques para simular la inteligencia humana, incluido el aprendizaje automático supervisado, las redes neuronales y el aprendizaje profundo Así mismo la integración progresivamente de las tecnologías de IA en las actividades rutinarias de las organizaciones, apoyan a las entidades en la continuidad del negocio [3].

Actividades

Para continuar con el desarrollo de esta propuesta investigativa, se propone realizar las siguientes actividades:

- π Revisión y actualización de la literatura científica.
- π Descarga y actualización de nuevos dataset de trabajo
- π Alistamiento de herramientas
- π Análisis de datos
- π Documentación respectiva



Figura 1. National Lib. Medicine.



Figura 2. Análisis de datos.

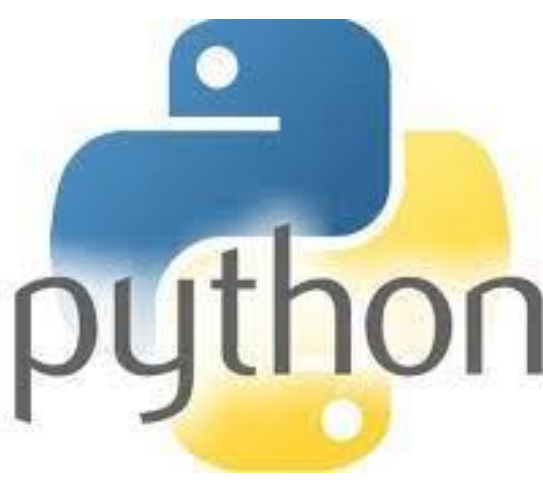


Figura 3. Lenguaje Python

Resultados

Se ha logrado establecer algunas relaciones mediante el análisis cluster con la información y dataset trabajados hasta el momento.

Dando continuidad, se espera involucrar nuevos datasets y nuevas herramientas de Inteligencia Artificial para poder establecer un análisis más cercano a la realidad y de esta manera, establecer conclusiones más concretas y fehacientes.

Referencias

- [1] O. Troisi , G. Fenza , M. Grimaldi y F. Loia , «Covid 19 sentiments in smart cities : The role of technology anxiety before and during the pandemic ,» Computers in Human Behavior , vol. 126, p. 106986, 2022.
- [2] A. Doanvo , X. Qian y D. Ramjee , «Machine Learning Maps Research Needs in COVID 19 Literature ,» Patterns , vol. 1, p. 100123,
- [3] V. Vincent, «Integrating intuition and artificial intelligence in organizational decision making,» Business Horizons, vol. 64, pp. 425 438, 2021.



**Escuela Tecnológica
Instituto Técnico Central**
Establecimiento Público de Educación Superior