DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE INGENIERÍA DE SOFTWARE

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SOFTWARE

TEMA:

SISTEMA WEB PARA EL RECONOCIMIENTO Y LA NORMALIZACIÓN DE ENTIDADES BIOMÉDICAS MEDIANTE TÉCNICAS DE CROSS-LINGUAL.

AUTORES:

PALLO TASIGUANO, BRANDON EDUARDO Y
SALAZAR RIVERA, ADRIAN ALEXANDER

DIRECTOR:
ING. UYAGUARI UYAGUARI, ALVARO DANILO

LATACUNGA FEBRERO, 2023

Orden del día

- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones

Orden del día

- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones

Problema

- El reconocimiento y la normalización de entidades biomédicas en notas clínicas, es un recurso fundamental para ayudar a los profesionales de la salud mediante encontrar rasgos en el texto a determinar unas posibles predicciones de diagnóstico y tratamientos médicos de un paciente.

- Existen varios enfoques y métodos para realizar este proceso de predicción, dependiendo de la naturaleza de los datos y de los recursos disponibles en el idioma en el que se redacten las notas médicas. (Quevedo-Marcos, 2020)

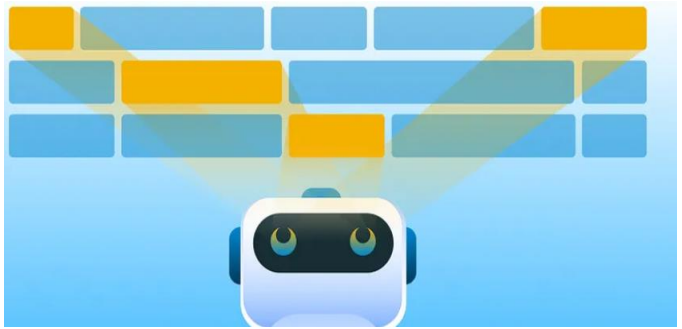
Problema

- La escasez de información estructurada en el campo de la medicina, especialmente en el español a lo largo de los años, imposibilita la aplicación en esta área de nuevas tecnologías de Inteligencia Artificial relacionadas con el análisis de datos. Nuevas aplicaciones de PLN han sido creadas con el objetivo de procesar textos médicos de manera automática y aumentar así la cantidad de datos estructurados.

Planteamiento de la solución

- Para solventar las necesidades previamente descritas se decide realizar una investigación para desarrollar un sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.

- Construye el sistema de reconocimiento y normalización de entidades médicas que integre tantos modelos supervisados como etiquetadores automáticos, para lograr un mejor desempeño.



Objetivo General

Desarrollar un sistema web que permita el reconocimiento de entidades biomédicas mediante técnicas de cross-lingual utilizando modelos supervisados de inteligencia artificial .

Objetivos Específicos

Explorar nuevos métodos para el reconocimiento y normalización de conceptos biomédicos.

Desarrollar y aplicar nuevos métodos de reconocimiento de entidades biomédicas mediante el uso de técnicas de procesamiento del lenguaje natural con cross-lingual.

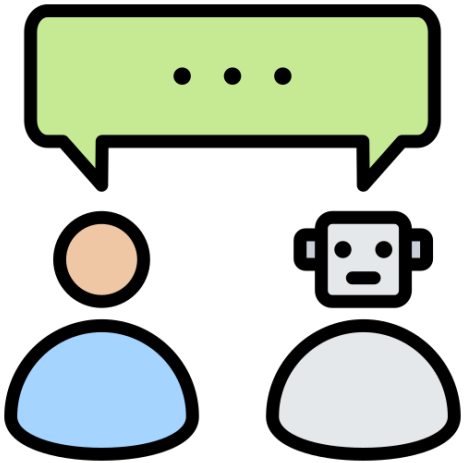
Aplicar buenas prácticas en el ciclo de desarrollo e implementar el sistema con frameworks y arquitecturas actuales.

Orden del día

- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones

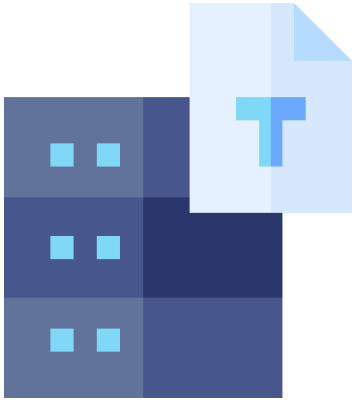
Procesamiento de Lenguaje Natural (NLP)

- El procesamiento de lenguaje natural es aquel que abarca distintas metodologías y conceptos, con el objetivo de poder ser un medio de comunicación efectivo entre personas y ordenadores, de tal manera que el ordenador tenga un lenguaje mediante el cual pueda entender las órdenes que recibe de un humano (Kang et al., 2020).
- El procesamiento de lenguaje natural se subdivide en dos ramas importantes que son Comprensión de Lenguaje Natural (NLU), que se encarga de interpretar los información para obtener los datos más importantes y la segunda rama es la Generación de Lenguaje Humano (NLG) (Kang et al., 2020).



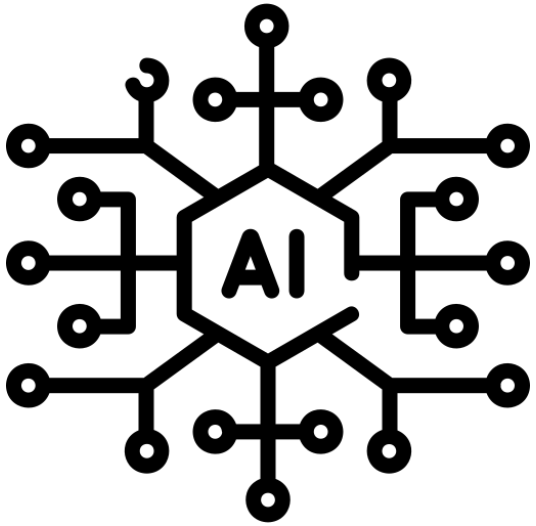
Minería de texto

- La minería de textos tiene un proceso definido que arranca con un conjunto de documentos que vienen de múltiples fuentes, para procesarlos mediante un tipo de tratamiento poniendo énfasis en sus caracteres.
- Para proceder con el análisis del texto mediante el que se determina información de elevada precisión, es preciso nombrar que este proceso puede ser repetitivo dando como resultado una mejor calidad en la información (Gaikwad et al., 2014).



Modelos basados en redes tipo Transformer

- **BERT:** es un modelo que se entrena de manera bidireccional con el propósito de extraer un sentido más recóndito de un contexto (Auquilla Vicuña & Mora Alvarez, 2022).
- **RoBERTa:** El modelo Transformer roBERTa es una variante del modelo Transformer original el mismo que fué entrenado en un corpus mucho más grande y con una serie de técnicas de optimización adicionales (Liu et al., 2019). Fue desarrollado por Facebook AI y se ha demostrado que supera a otros modelos de lenguaje populares en una serie de tareas de PLN (Liu et al., 2019).



Herramientas de etiquetado automático de entidades médicas

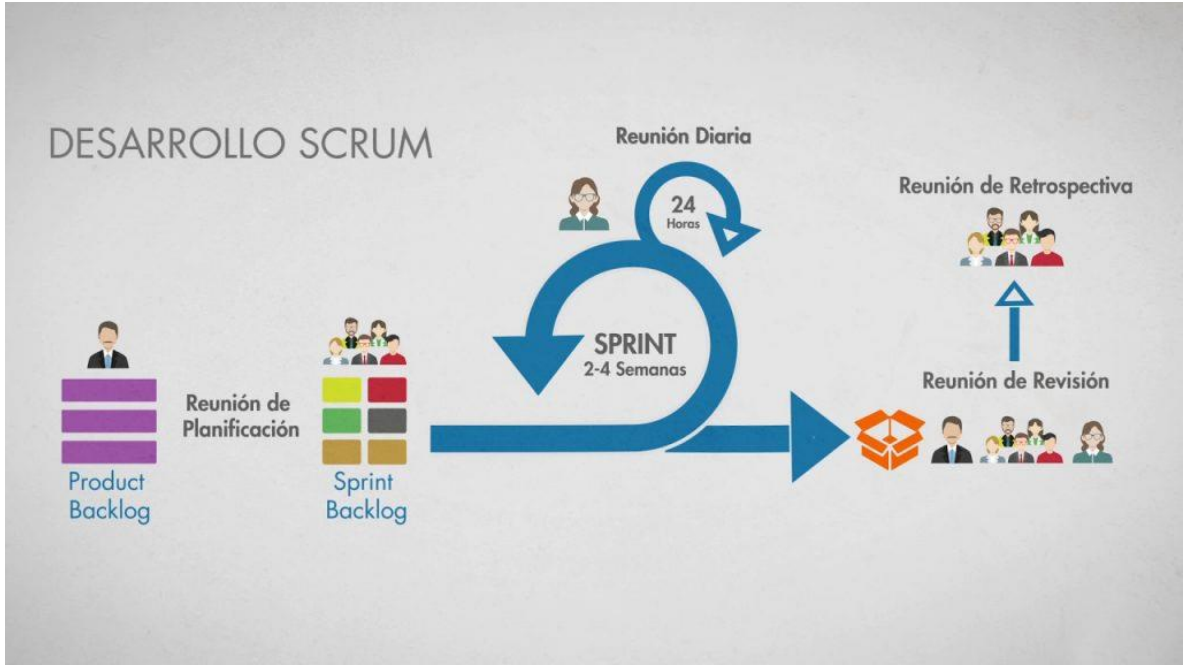
- **Metamap:** es una herramienta de reconocimiento y extracción de información biomédica, fue desarrollada en 2001 por la Biblioteca Nacional de Medicina (Peng et al., 2020). Esta biblioteca de información biomédica nace como una necesidad de recuperar información, las fuentes que está herramienta utiliza es el sistema unificado de lenguaje médico (UMLS) (Aronson & Lang, 2010b).

- **Google NLP:** El API de Cloud Healthcare de Google nos sirve para poder realizar el reconocimiento y extracción de información biomédica, mediante el uso de técnicas para poder guardar, trasladar, interrelacionar datos del ámbito médico (*Descripción general de la API de Cloud Healthcare | API de Cloud Healthcare | Google Cloud*, s. f.).

- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones

Metodología de desarrollo

- Esquema de la metodología Scrum



Recuperado de (HDC, 2020)

Análisis del sistema

Sprint 01: ***Utilizar herramienta de metamap para etiquetado automático de entidades médicas***

Historia de usuario 01
Como programador.
Quiero utilizar herramientas como Metamap
Para el etiquetado automático de entidades médicas

Sprint 02: ***Utilizar herramienta de Google para etiquetado automático de entidades médicas***

Historia de usuario 02
Como programador.
Quiero quiero utilizar herramientas como Google NLP
Para Para el etiquetado automático de entidades médicas

Análisis del sistema

Sprint 03: ***Crear un algoritmo que me permita tokenizar entidades para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus.***

Historia de usuario 03
Como programador.
Quiero crear un algoritmo que me permita tokenizar entidades.
Para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus

Sprint 04: ***Aplicar un modelo pre entrenado para reconocer entidades biomédicas a partir de un texto ingresado***

Historia de usuario 04
Como programador.
Quiero aplicar un modelo pre entrenado
Para reconocer entidades biomédicas a partir de un texto

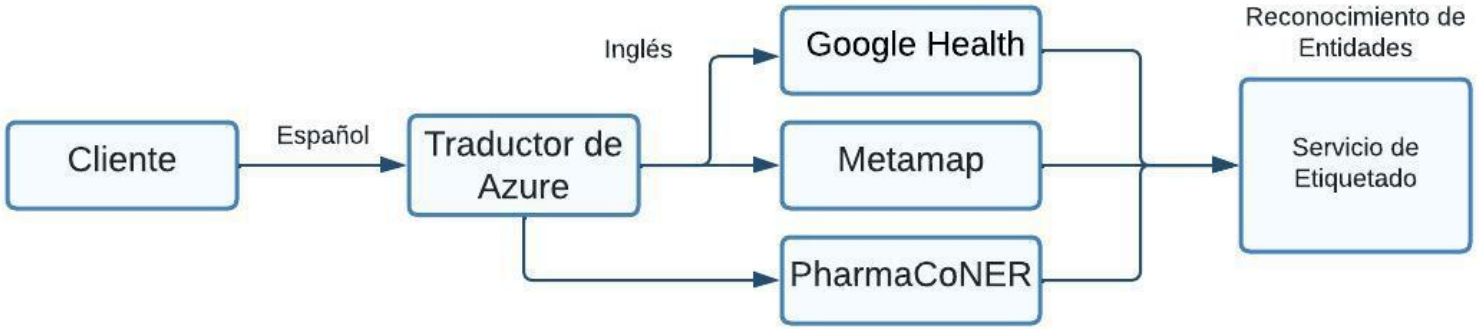
Análisis del sistema

- **Sprint 05:** ***Unificar el algoritmo y el modelo pre entrenado para obtener un mejor reconocimiento de entidades biomédicas***

Historia de usuario 05
Como programador.
Quiero unificar mi algoritmo a las herramientas de metamap y Google.
Para obtener un mejor reconocimiento de entidades biomédicas.

Diseño del sistema

- Arquitectura Lógica



Selección de las herramientas y recursos

Selección de corpus

PharmaCoNER: Un corpus lingüístico es un conjunto de documentos lingüísticos seleccionados y ordenados según criterios lingüísticos explícitos con el objetivo de ser usados como muestra del lenguaje («Corpus lingüístico», 2022). El corpus puede consistir en diferentes tipos de textos, como discursos, narrativas, diálogos, entrevistas, artículos, etc. Los textos del corpus pueden ser escritos o hablados, y se usan para estudiar el uso del lenguaje, las estructuras gramaticales, las relaciones entre palabras y frases, etc («Corpus lingüístico», 2022).

Selección de las herramientas y recursos

Selección de corpus

Medline (Del Clef Gold Corpus): Medline es una fuente de información creada por la Librería Nacional de Medicina (NLM), correspondiente al Clef Gold Corpus, es una colección de varios temas entre ellos biomedicina, ciencias de la vida, salud, ciencias químicas y bioingeniería entre otros. Temáticas imprescindibles en el conocimiento de profesionales de salud, científicos en investigación de ciencias biomédicas. (Costas et al., 2008).

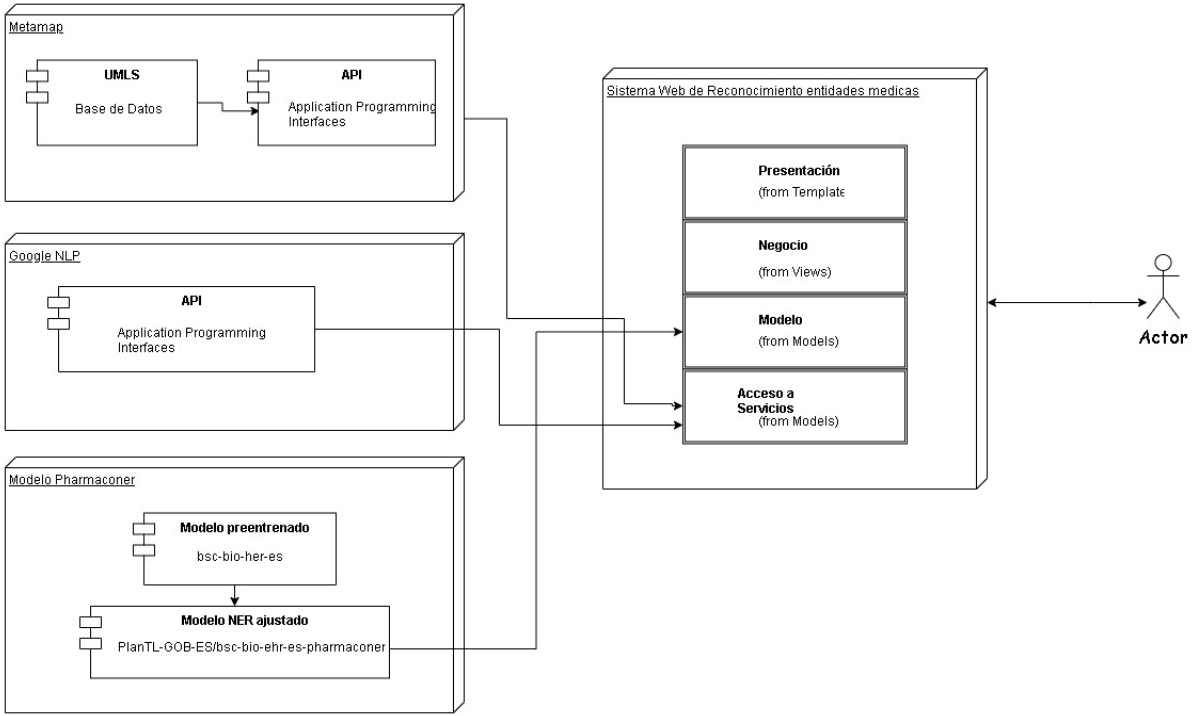
Selección de las herramientas y recursos

Selección de corpus

EMEA (Del Clef Gold Corpus): EMEA es un corpus de información biomédica correspondiente al Clef Gold Corpus el cual fue creado por la Agencia Europea de Medicina (EMEA). En este corpus se anexa información de medicamentos específicamente traducidos a varios idiomas de toda la Unión Europea. (CORRALES et al., s. f.)

Diseño del sistema

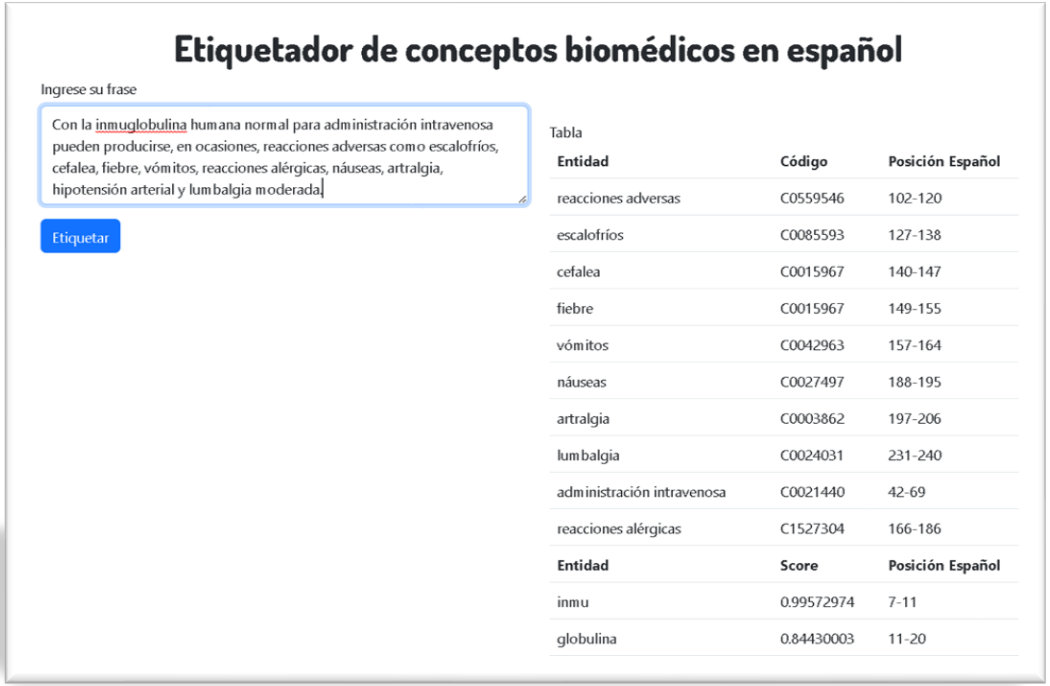
Esquema funcional



oración propia

Diseño del sistema

Interfaz de usuario del etiquetador de conceptos biomédicos



- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones

Validación del Sistema

Ejemplo de combinación de herramientas en el corpus Medline

Frase	Entidad	Tipo	Google NLP	Metamap	PharmaCoNER
Modificaciones de los valores de K + en la inducción con pentotal y succinilcolina.	K	CHEM	0	0	1
	succinilcolina	CHEM	1	1	0
	pentotal	CHEM	0	0	0

Validación del Sistema

Ejemplo de combinación de herramientas en el corpus EMEA

Frase	Entidad	Tipo	Google NLP	Metamap	PharmaCoNER
Retacrit fue tan eficaz como EPREX/ ERYPO para corregir y mantener los recuentos de glóbulos rojos.	EPREX	CHEM	0	0	1
	recuentos de glóbulos rojos	PHEN	0	1	0
	recuentos de glóbulos rojos	PROC	1	1	0
	Retracrit	CHEM	0	0	1
	ERYPO	CHEM	0	0	1

Validación del Sistema

Resultados de efectividad de las herramientas de etiquetado (Google NLP y Metamap)

Corpus	Etiquetado con combinación de herramientas (Google NLP y Metamap)	% de efectividad
Medline	166/316	52,53%
EMEA	237/430	55,11%

Validación del Sistema

Reconocimiento de nuevas entidades con PharmaCoNER

Corpus	Etiquetado con combinación de herramientas (Google NLP y Metamap)	Nuevas entidades identificadas con PharmaCoNER
Medline	166/323	7/323
EMEA	237/449	17/449

Validación del Sistema

Resultados de efectividad de las herramientas de etiquetado (Google NLP, Metamap y PharmaCoNER)

Corpus	Etiquetado con combinación de herramientas (Google NLP, Metamap y PharmaCoNER)	% de efectividad	% que incrementa al agregar PharmaCoNER con respecto a las anteriores dos herramientas
Medline	173/323	53,56%	1,03%
EMEA	254/449	56,57%	1,46%

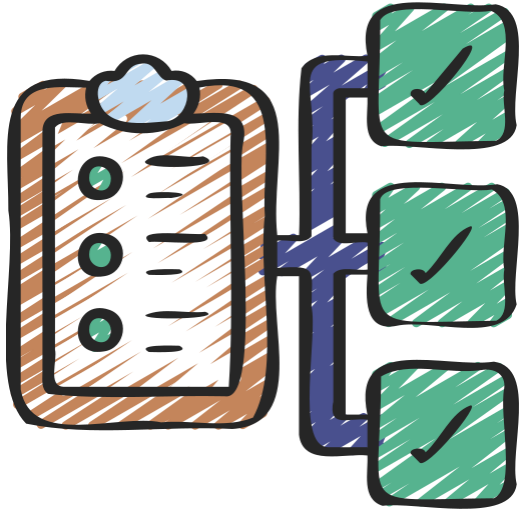
Análisis de resultados

- Resultados obtenidos al evaluar las herramientas de Google NLP, Metamap y PharmaCoNER en base a los corpus de Medline y EMEA, donde se evidencia que evaluando el corpus de Medline reconoce 173 entidades de un total de 323, obteniendo un resultado del 53,56%. Cabe mencionar que gracias a la integración de esta última herramienta el porcentaje de efectividad aumentó en un 1,03%. Por otra parte en el corpus EMEA se reconocieron 254 entidades de 449, obteniendo un resultado de efectividad del 56,57%, destacando que la efectividad aumentó en un 1,46% en comparación a lo anterior.

Análisis de resultados

- Estos resultados se obtuvieron por la integración de estas tres herramientas de etiquetadores automáticos (Google NLP, Metamap y PharmaCoNER) dando como resultado una mejor etiquetación de entidades biomédicas, ya que cada herramienta obtiene diferentes etiquetas, como podemos evidenciar a continuación.

- Introducción
- Marco Teórico
- Implementación del Sistema
- Validación del Sistema
- Conclusiones



Conclusiones

- Se cumplió con el objetivo de desarrollar un sistema de reconocimiento y normalización de entidades biomédicas mediante técnicas de cross-lingual, y la incorporación de un modelo supervisado, logrando un incremento de entidades biomédicas reconocidas automáticamente.
- El desarrollo del marco teórico permitió la obtención de conocimientos acerca de Procesamiento de Lenguaje Natural (NLP), Minería de texto, Métodos y técnicas de reconocimiento de entidades médicas nombradas, corpus etiquetados, BIO Scheme y las redes transformer BERT y RoBERTa.

Conclusiones

- El algoritmo implementado para etiquetar las entidades biomédicas basado en el corpus Pharmacomer, dio un excelente resultado, aumentando en 1,03% en la evaluación del corpus de Medline y un 1,46% en el corpus de EMEA como parte de la base del proceso de reconocimiento y normalización de entidades biomédicas.

Conclusiones

•El uso de herramientas como paperspace fue de gran ayuda en el desarrollo del trabajo, ya que la misma poseía GPU en la nube brindándonos de la misma manera un mejor rendimiento con respecto a la GPU de nuestras máquinas.

•El modelo implementado ayudó al reconocimiento de nuevas entidades, en su mayoría del tipo CHEM (Chemicals & Drugs).

Bibliografía

- Argyriou, A., González-Fierro, M., & Zhang, L. (2020). Microsoft Recommenders: Best Practices for Production-Ready Recommendation Systems. *Companion Proceedings of the Web Conference 2020*, 50-51. https://doi.org/10.1145/3366424.3382692
- Aronson, A. R., & Lang, F.-M. (2010a). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236.

Bibliografía

- Aronson, A. R., & Lang, F.-M. (2010b). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236. https://doi.org/10.1136/jamia.2009.002733
- Auquilla Vicuña, J. F., & Mora Alvarez, J. C. (2022). *Diseño de un sistema prototipo de diálogo persona-máquina basado en la arquitectura BERT* [BachelorThesis]. http://dspace.ups.edu.ec/handle/123456789/22403

Bibliografía

- Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, *58*(4), 102569. https://doi.org/10.1016/j.ipm.2021.102569
- Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., & Dai, L. (2021). ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. *Complexity*, *2021*, 6633213. https://doi.org/10.1155/2021/6633213
- Campos, D., Matos, S., & Oliveira, J. L. (2012). Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. En S. Sakurai (Ed.), *Theory and Applications for Advanced Text Mining*. IntechOpen. https://doi.org/10.5772/51066

Bibliografía

- Castillo Molina, C. A., Gutierrez, R. E., & Solarte, O. (2015). Prototipo para el reconocimiento de entidades nombradas en el idioma Español. *2015 10th Computing Colombian Conference (10CCC)*, 364-371. https://doi.org/10.1109/ColumbianCC.2015.7333447
- Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., & Xu, H. (2015). A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, *58*, 11-18.

Bibliografía

- Corpus lingüístico. (2022). En *Wikipedia, la enciclopedia libre*.

https://es.wikipedia.org/w/index.php?title=Corpus_ling%C3%BC%C3%ADstico&oldid=147071464

- CORRALES, M., ANDRÉS, S., IZA, I., LIBELIA, J., UYAGUARI, I. U., & DANILO, A. (s. f.). *DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN CARRERA DE SOFTWARE*.

Software
DOMINA EL MUNDO

Gracias por su atención

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA