Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación

Proyecto #2

Guido Padilla 19200
Oscar Paredez 19109
Diego Alvarez 19498
Julio Herrera 19402

Introducción

En este documento se ha innovado en la tecnología del procesamiento del lenguaje a través de programación dando solución al reto LitCoin NLP Challenge: Part 1. Este consiste en identificar la posición y clasificar la entidad biomédica dentro de artículos de investigación, dentro de su título y/o en el área de “abstract”. Como solución se ha diseñado un sistema el cual puede reconocer efectivamente conceptos científicos asociados a afirmaciones de conocimiento y determinar si es un hallazgo novedoso o únicamente información de relleno. A su vez este sistema es una aplicación en la cual un usuario puede cargar sus propios datos y pueda alterarlos de manera interactiva dentro de esta. Cabe aclarar que se realizó únicamente la parte 1 del reto y que una entidad biomédica es un concepto científico como: Organismo, Taxon, Célula, entre otros.

Si bien podemos obtener las distintas formas de las que se habla sobre una entidad biológica en los textos, ¿Cómo sabremos de cuál se está hablando y clasificarla a un *output* utilizando un modelo computacional? Para entender las relaciones entre las entidades biológicas mencionadas en cada uno de los artículos se usará el modelo de datos de alto nivel llamado BioLink, que proporciona asociaciones entre estas entidades biológicas, sus propiedades y relaciones entre sí de una manera estructurada. (Unni DR, et al., 2022)

Objetivos

- General
    - Identificar las entidades biológicas a las que se hacen referencia - de distintas formas - en los artículos biomédicos.
- Específicos
    - Procesamiento básico de palabras y símbolos a partir de diferentes *inputs* de texto (limpieza de datos).
    - Comprobar la eficiencia del modelo utilizando los *datasets* de validación.
    - Permitir el uso de diversos conjuntos de datos subidos a la aplicación así como también visualizar los resultados de forma interactiva.

Marco Teórico

Las publicaciones biomédicas exploran temas cada vez más profundos e interconectados con otros artículos, si bien las búsquedas directas a través de las *keywords* proporcionadas por los investigadores son de ayuda, es necesario crear métodos de procesamiento de lenguaje para descifrar todas las entidades que se abordan en los artículos biomédicos de forma masiva. Estas entidades se pueden expresar de distintas maneras entre diferentes artículos, ya sea usando abreviaturas, con sinónimos o explicaciones más largas usando diferentes palabras, símbolos y puntuaciones. Esto complica cada vez más la tarea de detectar estas entidades biológicas, a la cuál se le conoce como NER (Named Entity Recognition). (Cho, H., et al., 2019)

BioLink clasifica las entidades utilizando ontologías biomédicas de nivel superior a las que se puede llegar a partir de otras ontologías de nivel inferior. Las ontologías biomédicas describen una entidad a partir del significado de los datos y por lo tanto podemos determinar varias de ellas a través de asociaciones (Bodenreider O., et al., 2005). BioLink tiene una estructura donde considera las entidades, los predicados que las componen (slots), las asociaciones entre entidades y los slots para las asociaciones.

Este modelo de datos nos abre una vía para entrenar modelos de NLP (Natural Language Processing). Mediante asociaciones de palabras ya conocidas (abundancia que afecta a, actividad que involucra a, degradación afectada por, etc) decir qué entidades biológicas se están haciendo referencia en los artículos científicos. Así mismo se hizo uso de otros modelos pre entrenados para diferentes objetivos como entrenamiento, evaluación y procesamiento todos estos están basados en el algoritmo de NER y serán explicados más adelante. Cabe recalcar que el procesamiento de datos recayó en estos modelos, especialmente en el BIONLP13CG el cual es utilizado para clasificar conceptos científicos, no fue necesario limpiar datos ya que los archivos proporcionados por el reto ya están listos para su uso.

Metodología

Primero se parte por el problema que se plantea basado en el reconocimiento y análisis de texto dentro de artículos, el cual propone el uso de NLP para interpretar y comprender los textos propuestos, de modo que la tarea final sea la clasificación de las entidades propuestas por los artículos.

Los grupos de entrenamiento y prueba fueron previamente seleccionados por el challenge, por lo cual estaban preparados para uso de los algoritmos y modelos propuestos, utilizamos python por la familiarización con el lenguaje además de que las librerías conocidas están disponibles en este. Se utilizó el algoritmo de NER debido a que este puede analizar texto de una manera más eficiente que los bigramas y/o trigramas, scispacy proveé modelos pre-entrenados basados en NER, spaCy para la visualización de resultados e interacción con usuarios.

A su vez se utilizaron librerías de apoyo para el entrenamiento y visualización de los datos. Estos son pandas, pandas, pandas_profiling, seaborn, matplotlib, warnings y streamlit. Pandas se utilizó para el almacenamiento de datos y limpieza, pandas_profiling para realizar un reporte del análisis exploratorio, seaborn y matplotlib para la creación de gráficas de los datos, warnings sirvió para ignorar alertas innecesarias al momento de procesar los datos y streamlit como una ayuda para scispacy para la interacción con el usuario.

Cada integrante utilizó su computadora personal para realizar este reto, se usaron 2 computadoras HP con windows, dos razer con linux y una Mac con ios. La cantidad de datos no es extensa, por lo que no se necesitó de una computadora que tenga altas especificaciones para reducir el tiempo de procesamiento de los datos.

Los modelos utilizados fueron los siguientes:

- Craft Corpus: Entrenamiento y evaluación de métodos automatizados para la identificación de conceptos científicos.
- JNLPBA Corpus: Es un conjunto de datos de GENIA que contiene conceptos científicos, utilizado para validación.
- BC5CDR Corpus: Es un conjunto de datos que contiene múltiples artículos científicos, publicaciones médicas, químicas, entre otros. Utilizado para validación.
- BIONLP13CG Model: Es un modelo utilizado para el procesamiento de conceptos científicos y previamente entrenado por la librería de scispacy (NLP: Natural Language Processing).

Resultados

Se poseen en el conjunto de datos de entrenamiento un total de 400 observaciones para las propiedades de los artículos científicos dentro de abstracts_train que son comprendidas por 3 variables:

- abstract_id: El id único de cada artículo de tipo cuantitativo discreto
- title: El título de los artículos de tipo cualitativa
- abstract: el área de abstract de cada artículo de tipo cualitativa

Luego para las propiedades de las menciones de las entidades donde se encuentra y tipo de entidad dentro del artículo perteneciente al set de entities train, hay 13636 observaciones comprendidas por 7 variables:

- id: El id único de cada dato registrado en la tabla tipo cuantitativo discreto
- abstract_id: El id único de cada artículo al que pertenece de tipo cuantitativo discreto
- offset_start: La posición de inicio de la aparición de la entidad de tipo cuantitativo discreto
- offset_finish:La posición final de la aparición de la entidad de tipo cuantitativo discreto
- type: La categoría de la entidad mencionada de tipo cualitativa nominal
- mention: La mención en forma de texto a la entidad biomédica de tipo cualitativa nominal
- entity_ids: El id único de cada entidad de tipo cuantitativo discreto

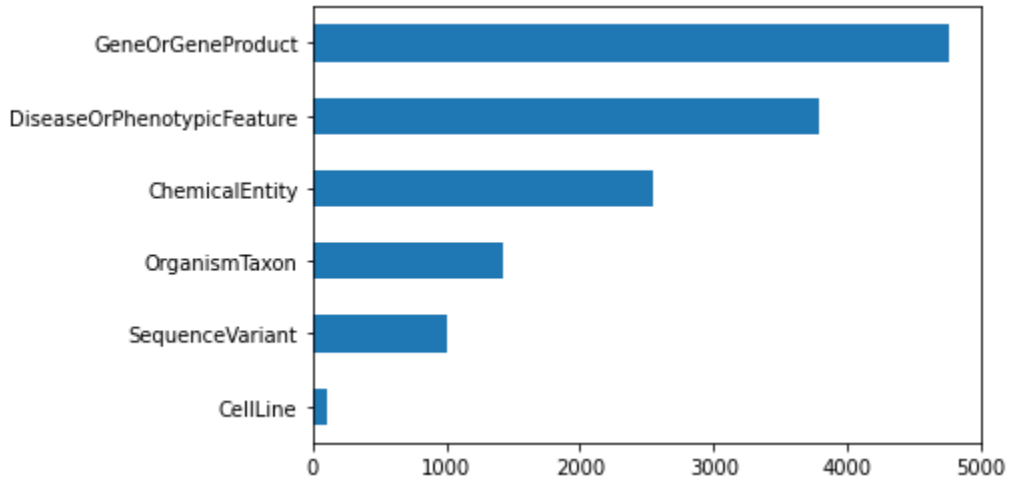
Realmente no se realizaron ningún tipo de limpiezas a los datos debido a que estos ya vienen para su uso. Los dos procesamientos que se realizaron fueron al momento de cargar los datos del archivo .csv y delimitarlo por comas para su correcta manipulación en los modelos. Luego al proveer de datos a los modelos ya mencionados, estos procesaron lo proporcionado y nos dieron los resultados. Por lo que no se necesito de algún ajuste a los parámetros del modelo solo se evaluó qué tan efectivo es con respecto a el input del usuario sabiendo la cantidad de entidades que hay.

Lo que sí se realizó debido al modelo utilizado es la clasificación, ya que muchas entidades no aparecen como GeneOrGeneProduct, sin embargo si pertenece a esta. Por lo que la siguiente clasificación por tipo de entidades fue necesaria de realizar.

Imagen 1: Clasificación de entidades

```
{
    "GENE_OR_GENE_PRODUCT": "GeneOrGeneProduct",
    "GGP": "GeneOrGeneProduct",
    "ORGANISM": "OrganismTaxon",
    "CANCER": "DiseaseOrPhenotypicFeature",
    "DISEASE": "DiseaseOrPhenotypicFeature",
    "CHEBI": "ChemicalEntity",
    "CHEMICAL": "ChemicalEntity",
    "PROTEIN": "SequenceVariant",
    "AMINO_ACID": "SequenceVariant",
    "ORGANISM_SUBDIVISION": "OrganismTaxon",
    "ORGANISM_SUBSTANCE": "OrganismTaxon",
    "TAXON": "OrganismTaxon",
    "CL": "CellLine",
    "CELL_LINE": "CellLine",
    "CELL": "CellLine",
    "SIMPLE_CHEMICAL": "ChemicalEntity",
}
```

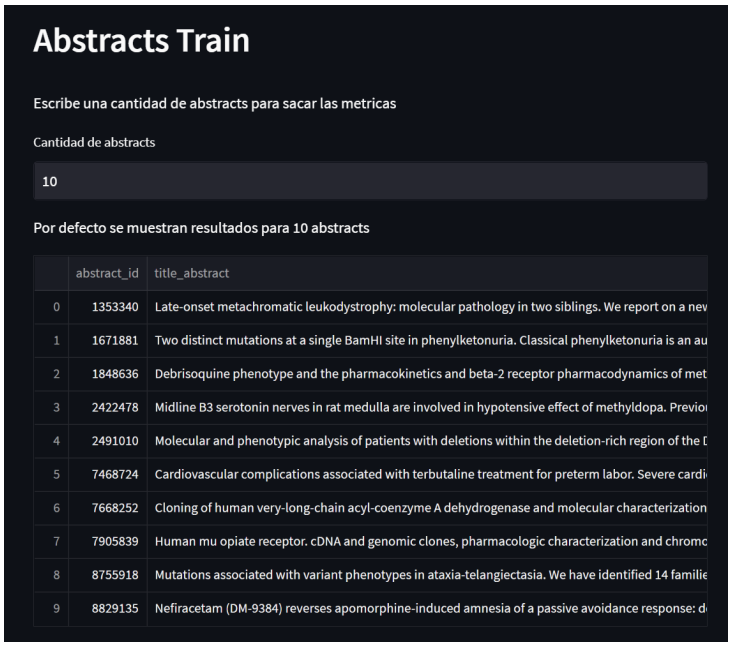
Gráfica 1: Frecuencia de las categorías contenidas en múltiples textos abstracts.



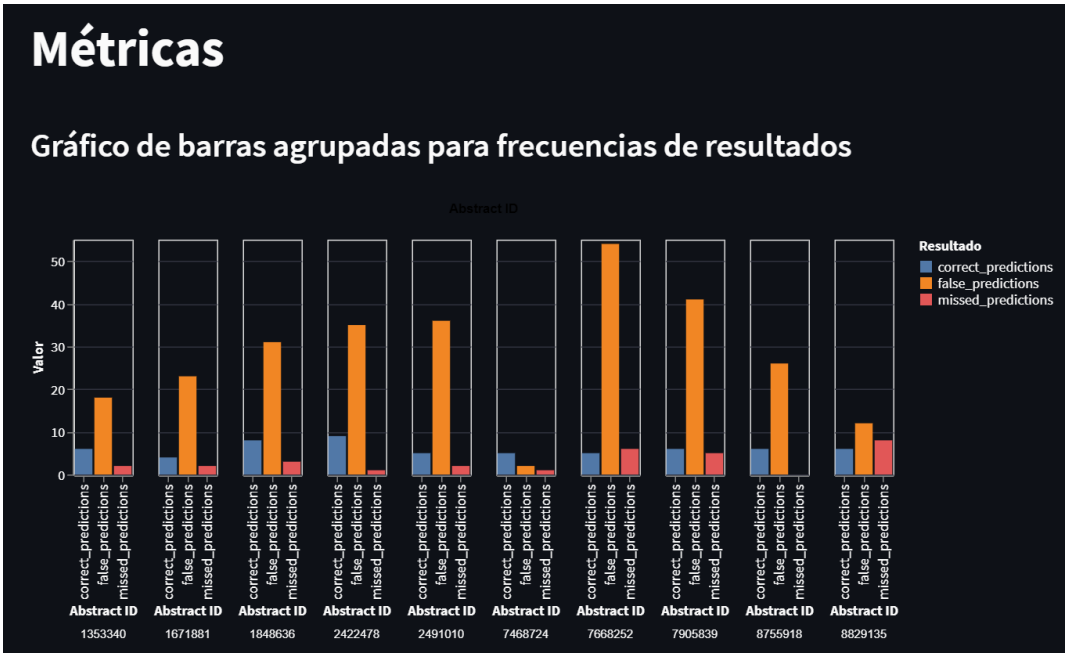
El reto menciona que se localicen ciertos tipos de entidades estas son: DiseaseOrPhenotypicFeature, ChemicalEntity, OrganismTaxon, GeneOrGeneProduct, SequenceVariant y CellLine. Sin embargo, el modelo puede encontrar otras entidades que también se muestran en las gráficas de resultados.

Aplicación

Esta aplicación muestra inicialmente resultados para las métricas de precisión, recall y f1 sobre cada abstract dentro de un rango de abstracts elegido por el usuario, por defecto se saca sobre 10 abstracts ya que ejecutar el modelo suele demorar un tiempo considerable.



Estas métricas se obtienen sobre los datasets de entrenamiento, usando los modelos para predecir las entidades en el texto (title + abstract) de cada abstract ID y comparar con las entidades reales que se mencionan según el dataset de entidades de entrenamiento para ese abstract ID. Se compara cuáles entidades predichas sí están en el dataset de entrenamiento (correct_predictions), cuales están en el de entrenamiento pero no en las predichas (missed_predictions) y cuales se predijeron pero no están en el dataset de entrenamiento (false_predictions).



Habiendo obtenido estos resultados podemos calcular las métricas de precisión, recall y f1 igualmente para cada abstract. Estas fueron sacadas utilizando las siguientes fórmulas:

$TP = True\ Positive$

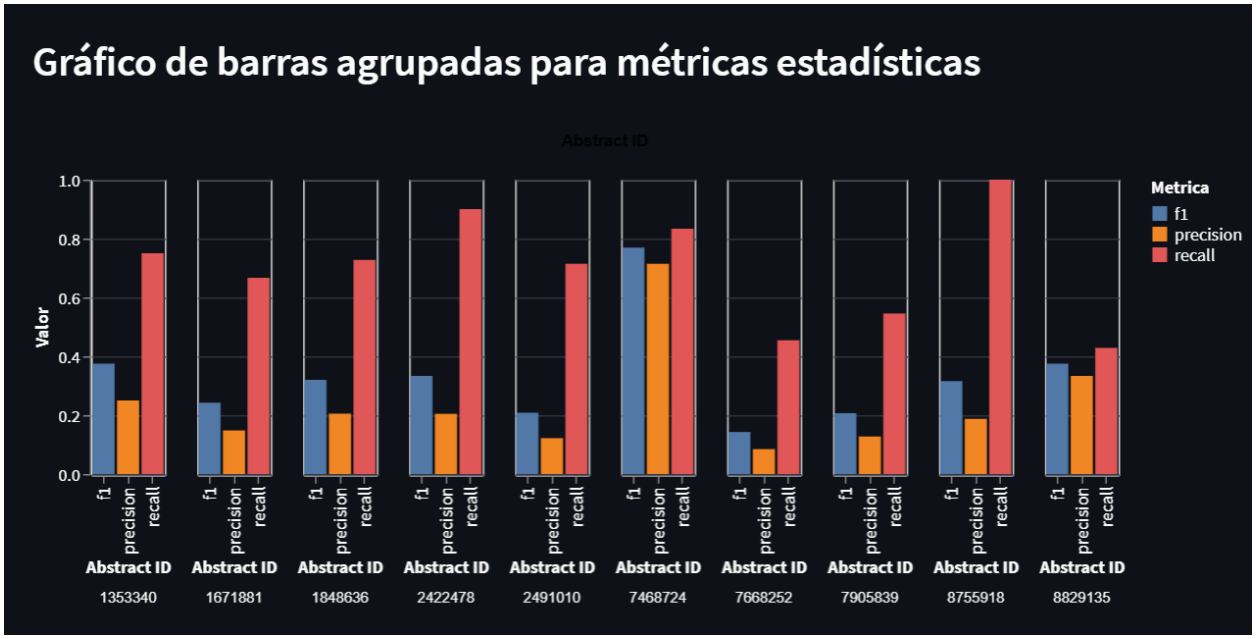
$FP = False\ Positive$

$FN = False\ Negative$

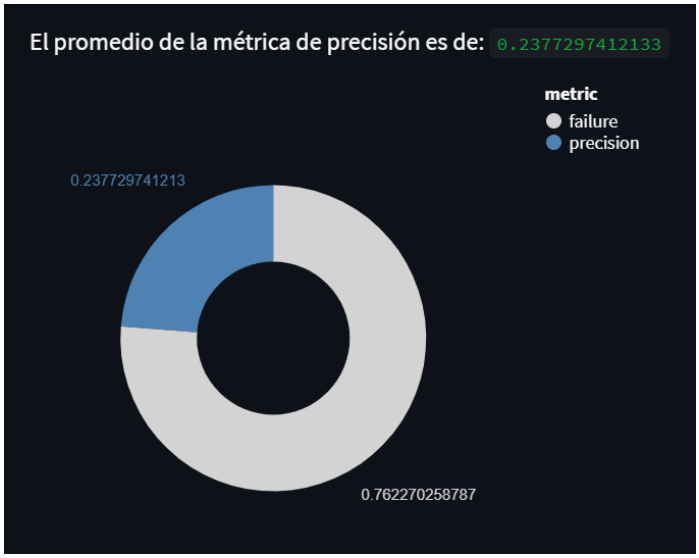
$precision = \frac{TP}{TP+FP}$

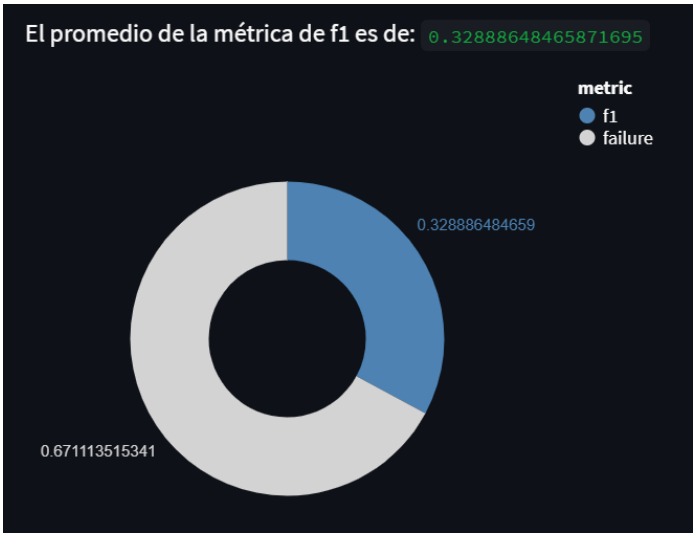
$recall = \frac{TP}{TP+FN}$

$F1 = \frac{2 \cdot precision \cdot recall}{precision+recall}$



Y para tener una vista general sobre el modelo en sí, se obtiene el promedio de cada una de estas métricas y se saca una gráfica de dona para saber el rendimiento de cada una sobre un 100%.





Dentro de los errores podemos mencionar que hay entidades que son encontradas pero no coinciden exactamente con lo que menciona el grupo de datos de entrenamiento, por lo que da como resultado una precisión menor a las esperadas. Sin embargo, analizándolo detenidamente si es una predicción válida. Acá un ejemplo de ello:

El modelo detecta la entidad de tipo DiseaseOrPhenotypicFeature en la mención “Late-onset metachromatic leukodystrophy” y no la detecta como correcta solo porque la mención en el dataset de entidades de entrenamiento en realidad es “metachromatic leukodystrophy” y por ende tampoco coincide en las posiciones donde se encuentra la mención y donde finaliza (offset_start y offset_finish).

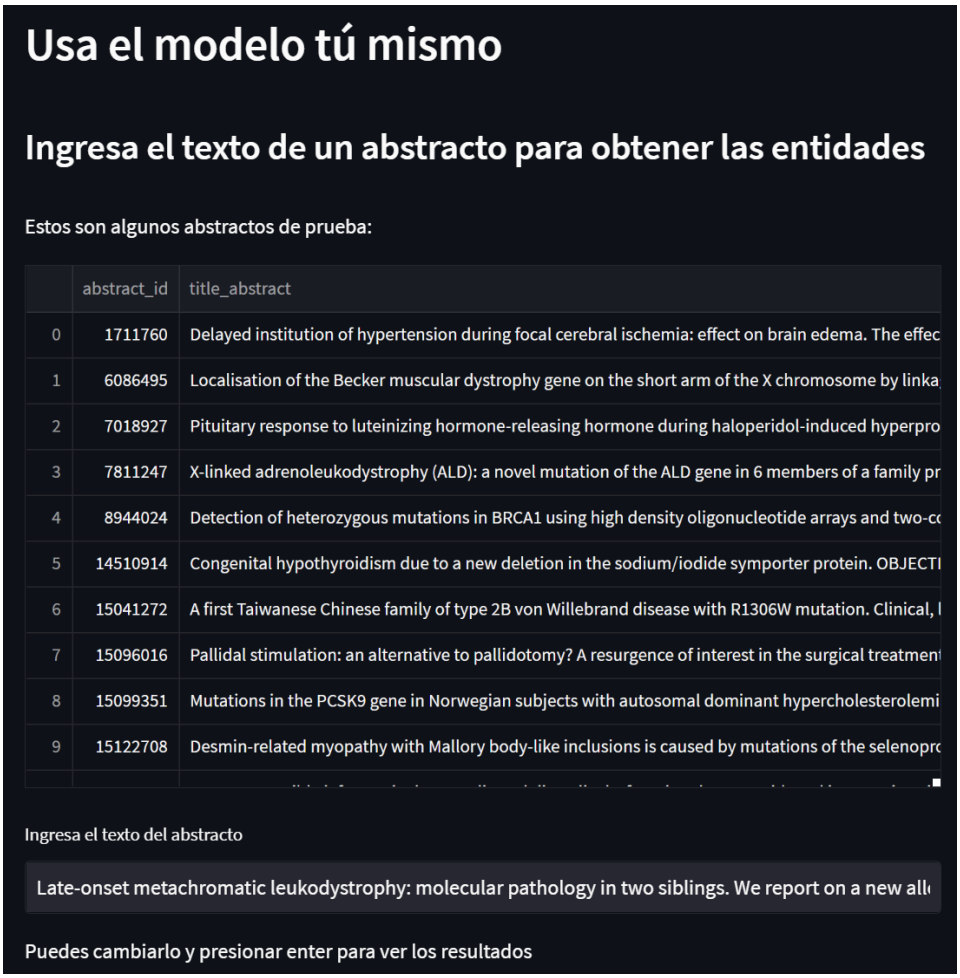
Predicción:

	text	label_	start_char	end_char
31	Late-onset metachrom	DiseaseOrPhenotypicF	1	39

Real:

id	abstract_id	offset_start	offset_finish	type	mention	entity_ids
0	1353340	11	39	DiseaseOrPhenotypicFeature	metachromatic leukodystrophy	D007966

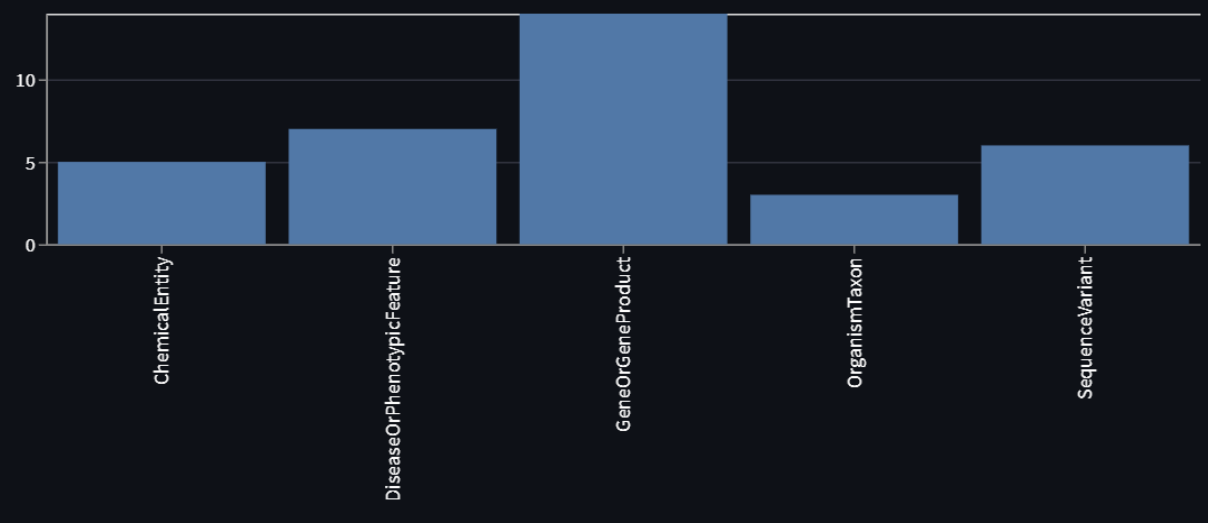
Por último en la aplicación se incluye una sección de “Usa el modelo tú mismo” donde se puede ingresar cualquier texto para sacar las entidades de las que se hablan dentro de él, es decir las predicciones y obtener una lista de ellas así como una gráfica de la frecuencia de cantidad sobre cada tipo de entidad.



Entidades extraídas

	text	label_	start_char	end_char
31	Late-onset metachrom	DiseaseOrPhenotypicF	1	39
32	late-onset metachroma	DiseaseOrPhenotypicF	149	187
33	MLD	DiseaseOrPhenotypicF	190	192
34	glutamine	ChemicalEntity	307	315
35	MLD	DiseaseOrPhenotypicF	364	366
36	glutamine	ChemicalEntity	387	395
37	MLD	DiseaseOrPhenotypicF	577	579
38	MLD	DiseaseOrPhenotypicF	713	715
39	allele	<NA>	98	103
40	arylsulfatase A (ARSA lo	<NA>	112	139

Frecuencia de entidades por tipo



Referencias

Bodenreider O, Mitchell JA, McCray AT. Biomedical ontologies. Pac Symp Biocomput. 2005:76-8. doi: 10.1142/9789812704856_0016. PMID: 15759615; PMCID: PMC4300097

Cho, H., Lee, H. Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics 20, 735 (2019). https://doi.org/10.1186/s12859-019-3321-4

Naseem, U. (2022, 21 abril). Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - BMC Bioinformatics. BioMed Central. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04688-w

Unni DR, Moxon SAT, Bada M, Brush M, Bruskiewich R, Caufield JH, Clemons PA, Dancik V, Dumontier M, Fecho K, Glusman G, Hadlock JJ, Harris NL, Joshi A, Putman T, Qin G, Ramsey SA, Shefchek KA, Solbrig H, Soman K, Thessen AE, Haendel MA, Bizon C, Mungall CJ, The Biomedical Data Translator Consortium (2022). Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clin Transl Sci. Wiley; 2022 Jun 6; https://onlinelibrary.wiley.com/doi/10.1111/cts.13302