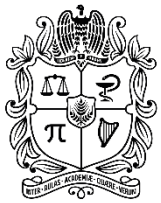


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo Computacional para el análisis de historias clínicas de pacientes con Artritis Reumatoide aplicando bioinformática traslacional y minería de textos

Alexander Del Risco Morales

Universidad Nacional de Colombia
Facultad, Departamento Ingeniería de sistemas e industrial
Bogotá, Colombia
2021

Modelo Computacional para el análisis de historias clínicas de pacientes con Artritis Reumatoide aplicando bioinformática traslacional y minería de textos

Alexander Del Risco Morales

Tesis presentada como requisito parcial para optar al título de: Magister en bioinformática

Director:
Luis Fernando Niño Vásquez, Ph.D.

Asesor:
Gerardo Quintana, M.D., Ms. Sc.

Línea de Investigación:
Tecnologías computacionales en Bioinformática

Grupo de Investigación:
LISI

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento ingeniería de Sistemas e Industrial
Bogotá, Colombia
2021

A la vida, los anhelos e ilusiones
y a la esperanza porque que me
enseñan cada día a vencer mis
temores.

Agradecimientos

A mi director de tesis, el Ph.D. Luis Fernando Niño y al Dr. Gerardo Quintana por la oportunidad y las enseñanzas brindadas.

A todos los compañeros del grupo LISI que con sus aportes y ánimo nos permiten circular por este camino, especialmente al Dr. Luis Morales y al ingeniero Andrés Cely por su apoyo.

A mi esposa por su paciencia, apoyo y todo su amor.

A Jorge Eliecer Viafara Morales, porque en uno de los momentos más críticos durante este trabajo me dio el aliento y soporte para continuar.

Resumen

Este trabajo tiene como finalidad crear un modelo computacional que permita identificar el avance de la enfermedad de artritis reumatoide (AR) con base en el análisis de historias clínicas de pacientes diagnosticados con Artritis Reumatoide. Se plantea que mediante la minería de texto, se puede extraer la información que permita a los profesionales del área identificar datos relevantes para el proceso de diagnóstico de AR y de esta forma hacer un diagnóstico temprano de la misma, así también, se pretende aplicar el concepto de bioinformática traslacional, esto implica que la información de valor y que cumpla con los objetivos propuestos de esta investigación pueda ser transferida de forma efectiva a los pacientes que sufren esta enfermedad.

Se ha desarrollado un modelo que aplica minería de textos, recuperación de la información, lingüística computacional, aprendizaje de máquina y otras áreas del conocimiento relacionadas, que permiten transformar y tratar los datos no estructurados para poder hacer el análisis correspondiente de las historias clínicas y así descubrir conocimiento implícito inmerso en las narrativas de las historias clínicas que ayude con el propósito de tener más y mejor información asociada a la artritis reumatoide y la evolución de los pacientes.

Palabras clave: Artritis reumatoide, bioinformática traslacional, minería de textos, aprendizaje de máquina, procesamiento de lenguaje natural, Snomed.

Abstract

The purpose of this work is to create a computational model that allows to identify the progression of rheumatoid arthritis (RA) disease based on the analysis of medical records of patients diagnosed with rheumatoid arthritis. It is proposed that through text mining, information can be extracted which allows professionals in the area to identify relevant data for the RA diagnosis process and thus make an early diagnosis, therefore, it is also intended to apply the concept of translational bioinformatics, which implies that valuable information that meets the proposed objectives of this research can be effectively transferred to patients suffering from this disease.

A model has been developed that applies text mining, information retrieval, computational linguistics, machine learning and other related areas of knowledge, which allow the transformation and processing of unstructured data in order to carry out the corresponding analysis of medical records and thus discover implicit knowledge immersed in the narratives of medical records that helps with the purpose of having more and better information associated with rheumatoid arthritis and the evolution of patients.

Title

Computational Model for the analysis of clinical records of patients with Rheumatoid Arthritis applying text mining and translational bioinformatics.

Keywords: Rheumatoid arthritis, translational bioinformatics, text mining, machine learning, natural language processing, Snomed.

Esta tesis de maestría se sustentó el 07 de Octubre de 2021 a las 3:00 pm,
y fue evaluada por los siguientes jurados:

Emiliano Barreto Hernandez
Ph.D. Doctorado En Ciencias - Estadística
Profesor, Instituto de Biotecnología, Universidad Nacional de Colombia

John Jaime Sprockel Diaz
Msc. Ingeniería de Sistemas y Computación
Coordinador Semiología en Fundación Universitaria de Ciencias de la Salud

Contenido

Lista de figuras

Lista de Tablas

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
NLP	natural language processing
HCE	historia clínica electrónica
CIE-9	Clasificación internacional de enfermedades
IC	insuficiencia cardíaca
GWAS	estudio de asociación de genoma completo
GHR	receptor de la hormona de crecimiento
UniProtKB	base de datos de conocimiento UniProt
COSMIC	Catálogo de mutaciones somáticas en cáncer
DISTILD	enfermedades y rasgos en el desequilibrio de ligamiento
EMR	historia clínica electrónica
CE	cáncer de esófago
LSTM	Long Short Term Memory
AI	Inteligencia artificial
ML	Aprendizaje de máquina
NHIRD	investigación de seguros de salud nacional
CRF	campos aleatorios condicionales
SIAPS	Sistema Integral para la Atención Primaria de Salud
SSE	suma de errores cuadrados
NLM	Nacional de Medicina de los Estados Unidos
NLTK	Trabajo práctico en procesamiento del lenguaje natural
CUI	Identificador único de concepto
LOINC	conjunto de identificadores, nombres y códigos
SNOMED	Nomenclatura sistematizada de la medicina
IL-1	interleucina-1
IL-6	interleucina-6
FAME	fármacos antirreumáticos modificadores de la enfermedad

1. Introducción

La artritis reumatoide (AR) es una enfermedad crónica y prevalente que genera importante morbilidad y discapacidad. La enfermedad es mucho más frecuente en mujeres y su prevalencia en la población latinoamericana es cercana al 0,91%, (Quiceno et al., 2011). La literatura reciente evidencia que el abordaje temprano y en algunos casos agresivo de la patología puede cambiar de forma significativa el curso clínico de la enfermedad (K et al., 2008). Es una artropatía no benigna, inflamatoria y destructiva que sin tratamiento efectivo lleva a discapacidad produciendo limitación parcial en el 80% de los pacientes y limitación total en el 16% después de doce años de enfermedad. La capacidad laboral de los pacientes se restringe 1/3 en un año y un 40% a los tres años (K et al., 2008).

La importancia del diagnóstico temprano es un imperativo debido a que las erosiones, que representan la destrucción del hueso yuxtaarticular, y la disminución del espacio articular (DEA), que traduce la pérdida difusa del cartílago, la pérdida de funcionalidad y la pérdida de densidad mineral ósea axial y periférica ocurren en el curso temprano de la enfermedad, (Pedrosa, 2002).

La duración de la enfermedad es factor predictor de la respuesta a medicamentos modificadores de enfermedad (DMARD) y su inicio temprano ofrece mejor pronóstico. De manera similar se constituye una ventana de oportunidad terapéutica para la mayoría de los grupos farmacológicos en cuanto a su perfil de

eficacia-efectividad y seguridad, incluyendo la terapia biológica, con mayor probabilidad de remisión en los primeros tres meses de enfermedad, (K et al., 2008).

La discapacidad puede reducirse entre más temprano se inicie el tratamiento y más agresivo sea; una demora en solo tres meses en el inicio de DMARD produce un pobre desenlace a cinco años comparado con el inicio previo de estos medicamentos.

Por estas razones es fundamental reconocer de forma temprana al paciente con enfermedad inflamatoria articular para realizar el diagnóstico expedito de (AR) y así mismo cumplir los objetivos del tratamiento: aliviar el dolor, controlar la inflamación, preservar la funcionalidad, mejorar la calidad de vida y prevenir la destrucción articular, permitiendo además la reducción de los costos en el sistema de salud y todos sus componentes: Sociedad, Gobierno, entidad promotora de salud (EPS), institución prestadora de salud (IPS) y paciente.

Como antecedente a esta investigación, se tiene la tesis de maestría de Luis Antonio Morales Muñoz, (Morales Muñoz, 2014): Modelo computacional para la identificación de endofenotipos en pacientes con artritis reumatoide utilizando información del Antígeno Leucocitario Humano (HLA) clase II, en la cual se realizó el análisis de un conjunto de datos de pacientes con artritis reumatoide aplicando métodos de inteligencia computacional con el fin de tratar de identificar el punto intermedio entre el genotipo y el fenotipo para la enfermedad. Los datos incluían información genética, clínica y serológica. También, se trató de obtener marcadores para determinar la severidad de la enfermedad en los pacientes. Se identificaron relaciones de dependencia entre algunas de las variables a través de una red bayesiana. Específicamente, se estableció que las variables del antipéptido cíclico citrulinado (anti CCP) y el factor reumatoide están implicadas

de manera significativa en la enfermedad. También, se identificaron mecanismos bioquímicos, neurofisiológicos, neuroanatómicos o neuropsicológicos relacionados con la severidad de la enfermedad en pacientes; particularmente se identificaron secuencias de aminoácidos comunes entre pacientes con el mismo desenlace, (Morales Muñoz, 2014).

1.1 Identificación del problema

La (AR) es una artropatía no benigna, inflamatoria y destructiva que sin tratamiento efectivo lleva a discapacidad produciendo limitación parcial en el 80% de los pacientes y limitación total en el 16% después de doce años de enfermedad. La importancia de asociar la información clínica y genética al desarrollo de la enfermedad está en tener la posibilidad de reconocer de forma temprana al paciente con enfermedad inflamatoria articular para realizar el diagnóstico expedito de (AR) y, así mismo, cumplir los objetivos del tratamiento, (K et al., 2008).

Con base en lo anterior se formula la siguiente pregunta de investigación: ¿es posible desarrollar un método computacional, basado en la minería de textos y la bioinformática traslacional, que permita asociar la información clínica del paciente al estado de desarrollo de la enfermedad y a información genética disponible en bases de datos biológicas y en estudios previos asociados a pacientes?

1.2 Justificación

Cada vez es más común obtener información médica en formato electrónico. Esto incluye tanto artículos científicos como revisiones de gestión clínica, así como registros médicos con datos de pacientes. Sin embargo, las herramientas

tradicionales, tanto individuales como institucionales, son menos útiles para seleccionar la información más adecuada para cada caso, independientemente del contexto clínico o de investigación, (Piedra et al., 2014).

Los avances en la tecnología de la información significan que enormes cantidades de datos relacionados con la salud ahora se pueden acceder y analizar. Esto incluye no sólo los datos personales, sino también diagnósticos, la gravedad y los resultados de pruebas de laboratorio, pruebas de función y medicamentos, y detalles sobre los contactos del paciente con el sistema hospitalario, (Guerrero Pupo et al., 2004). Hay tres ventajas importantes en el registro de esta gran cantidad de datos en formato digital:

- La calidad se mejora.
- Se reduce el tiempo que los trabajadores de la salud gastan en tareas improductivas.
- Los datos pueden ser utilizados en los sistemas de automatización, tales como minería de texto o minería de datos.

1.3 Objetivo general y objetivos específicos

1.3.1 Objetivo general

Diseñar un modelo computacional que permita asociar la información clínica del paciente al desarrollo de la enfermedad (AR) y a la información genética relacionada con (AR) disponible en bases de datos biológicas.

1.3.2 Objetivos específicos

Establecer el desarrollo de la enfermedad (AR) de un paciente mediante el análisis de la historia clínica con métodos de minería de textos.

Asociar información clínica del paciente a la información genética relacionada con AR que se encuentre disponible en bases de datos biológicas.

En este trabajo se realizó el análisis de un conjunto de historias clínicas de pacientes con artritis reumatoide aplicando métodos de minería de textos, aprendizaje de máquina, lingüística computacional y otras áreas del conocimiento relacionadas, con el fin de identificar el avance de la enfermedad a partir del corpus analizado.
Los datos tratados están en formato no estructurado y semi estructurado que representa la narrativa de las historias clínicas de pacientes diagnosticados con la enfermedad de artritis reumatoide. Además, se trató de obtener información relevante que no estuviera explícita en el conjunto de datos.

Incluir un párrafo que describa la organización del resto del documento
Los objetivos planteados se desarrollan a través de la aplicación de un flujo de trabajo en el cual se llevan a cabo diferentes etapas iniciando con la adquisición de los datos representados por las historias clínicas de pacientes diagnosticados con artritis reumatoide, luego se realiza el preprocesamiento de la información no estructurada con el propósito de transformarla en estructuras que permitan aplicar algoritmos y métodos de análisis computacional, posteriormente se realiza el análisis aplicando métodos y algoritmos computacionales dentro del contexto de la información médica y

farmacológica, por tanto se asocian palabras a la terminología médica de SNOMED lo que permite tener una interoperabilidad semántica que ayuda a la normalización de la información contenida en las historias clínicas y por otro lado se asocian los términos que hacen referencia a tratamientos farmacológicos con la base de datos ChEMBL.

2. Marco teórico

La minería de textos surge como un enfoque particular del proceso de descubrimiento de conocimiento, específicamente orientado al descubrimiento en fuentes textuales y no estructuradas. (Montes y Gómez, Manuel and Gelbukh, Alexander and López López, 2005). No es fácil acotar el significado del término "minería de textos" como disciplina, porque desde que se comenzó a hablar de minería de textos se han integrado diferentes tecnologías y principios teóricos desarrollados en otras disciplinas.

Si se recurre a la literatura sobre el tema, se encuentran distintas definiciones. (Sullivan, 2001), autor de una de las monografías dedicadas en exclusividad al tema, recoge dos de ellas: la primera define minería de textos como cualquier operación realizada para extraer y analizar textos procedentes de distintas fuentes externas con el objetivo de poder tomar decisiones. La segunda define minería de texto como el descubrimiento de información y conocimiento que anteriormente no se conocía, a partir de corpus textuales.

Esta segunda definición coincide en líneas generales con la que quizá sea la más popular y que formuló (Hearst, 1999), en su artículo titulado "*Untangling text data mining*". En ese texto, Hearst señala que ésta tiene como objetivo descubrir información y conocimiento que previamente se desconocía, y que no aparecía en ninguno de los documentos analizados. De acuerdo con esta definición, la

minería de textos sería un proceso con el que se pretende descubrir nueva información o conocimiento, y en el que la información que se descubre debe ser desconocida de antemano, incluso por los autores de los documentos que se hayan tomado como punto de partida del proceso.

De igual manera, en el trabajo sobre la permeabilidad de las disciplinas científicas, (Smalheiser & Swanson, 1994), señaló las limitaciones derivadas del desconocimiento de la literatura publicada por expertos en un determinado campo por parte de otras áreas del conocimiento y que pueden estar relacionadas con su tema de estudio. Tomando como ejemplo la migraña, (Swanson, 1987), ilustra estas limitaciones. Swanson extrajo una serie de declaraciones de diferentes artículos publicados por expertos en diferentes campos, sobre lo cual hace énfasis en que es muy poco probable que un mismo científico tenga acceso a todas estas afirmaciones. Swanson utilizó las siguientes declaraciones:

- El estrés está relacionado con las migrañas.
- El estrés puede producir pérdidas de magnesio.
- Los bloqueos de calcio previene las migrañas.
- El magnesio es un bloqueante natural del calcio.
- Los niveles altos de magnesio inhiben la muerte cardiaca súbita (SCD, por su sigla en inglés, *Sudden Cardiac Death*).

Estas afirmaciones, extraídas de una serie de elementos no interrelacionados, permiten inferir la relación entre la deficiencia de magnesio y la migraña. Vale la pena señalar que esta suposición no está documentada en ningún artículo. En otras palabras, se pueden extraer nuevos conocimientos directamente del corpus de texto inicialmente desarticulado. Promover y permitir tales inferencias a partir

de conexiones ocultas entre diferentes textos será el verdadero objetivo de la minería de textos.

2.1 La minería de textos como herramienta para el análisis

El objetivo de la minería de textos es extraer nuevo conocimiento a partir del análisis de corpus textuales. Esta capacidad permite a un usuario adquirir conocimientos útiles codificados en el texto que no es fácil de obtener para un usuario sin sintetizar y analizar una parte relativamente grande de los datos. En este caso, una herramienta de análisis de datos puede examinar una gran cantidad de datos para descubrir patrones interesantes ocultos en el texto. (Zhai & Massung, 2016)

De manera similar, (Sullivan, 2001), señaló cómo la minería de textos es el proceso de compilar, organizar y analizar una gran colección de documentos para apoyar la distribución de información a analistas y tomadores de decisiones, y para descubrir hechos relevantes que se distribuyen en diferentes campos de investigación.

Como objetivo inicial, la minería de textos debería facilitar el análisis del corpus, es decir, el conocimiento que esté contenido en una extensión de documentos y que es difícil de manejar debido a su tamaño. Así pues, los investigadores podrán analizar estos datos, identificar relaciones entre documentos y sacar conclusiones. (Hearst, 1999), aclaró el alcance de la minería de textos, indicando que no es necesario utilizar inteligencia artificial para analizar el propio texto, pero una combinación de análisis manual y análisis automático puede producir excelentes resultados. La autora incluso define la minería de texto como el descubrimiento semiautomático de patrones y tendencias en grandes conjuntos

de datos. Por tanto, la minería de textos servirá como objetivo intermedio, y antes de descubrir nuevos conocimientos, se procesará y presentará la información disponible en una gran cantidad de documentos en un formato que facilite su comprensión y análisis.

Desde un punto de vista técnico, la minería de textos recopila técnicas tradicionalmente utilizadas para la recuperación de textos y la lingüística computacional. Este hecho ha llegado a tal punto que es difícil afirmar que la minería de textos haya incorporado su propia tecnología (Hearst, 1999).

Moreno Sandoval se refiere a la lingüística computacional como la ciencia que permite el análisis sintáctico y gramatical de textos en formato electrónico y la alineación e identificación de correspondencias entre textos escritos en diferentes idiomas, entre otras. Sus principales resultados se han materializado en los sistemas de traducción automática (Moreno Sandoval, 1998).

Desde otra perspectiva, IBM, fabricante de una de las herramientas comerciales de minería de textos, define la minería de textos como el proceso de extraer automáticamente información básica del texto y detectar automáticamente los temas principales en conjunto de documentos y realizar búsqueda de texto relevante mediante consultas flexibles y de alto rendimiento (Tkach, 1998). En esta definición, sin embargo, IBM evita mencionar la posibilidad de identificar nuevos conocimientos a partir de documentos existentes, controvirtiendo así la propuesta original propuesta por (Hearst, 1999).

Esto nos conduce a una comparación entre la minería de textos y la recuperación de información, teniendo esta última la finalidad de identificar documentos relevantes para los usuarios de una colección de documentos. Sin embargo, la

recuperación de texto no es tan útil para el proceso de análisis o la extracción de nuevos conocimientos como la minería de texto. Hearst enfatizó esta diferencia y señaló cómo se difunde de manera incorrecta esta apreciación, lo que hace que la minería de texto sea equivalente a sistemas avanzados de recuperación de información o sistemas de recuperación de información adecuados para Internet, (Hearst, 1999).

En síntesis, algunas de las funciones que principalmente debería satisfacer una herramienta de minería de textos, o la salida que se espera de ella, incluiría:

- Identificar hechos y datos concretos a partir del texto de los documentos: extracción de características (*feature extraction*).
- Agrupar documentos similares (*clustering*).
- Determinar el tema o temas tratados en los documentos mediante la categorización automática de los textos (*indexing*).
- Identificar los conceptos tratados en los documentos y crear redes de conceptos (*topics*).
- Facilitar el acceso a la información repartida entre los documentos de la colección, mediante la elaboración automática de resúmenes y la visualización de las relaciones entre los conceptos tratados en la colección. (*summarise*)
- Visualización y navegación de colecciones de texto (*visualization*).

2.2 Técnicas de minería de textos.

La minería de textos incluye tareas que permiten procesar la información no estructurada con el fin de hacer que dicha información adquiera una estructura legible, que permita realizar el análisis y visualización de la misma y, finalmente, que sea un insumo relevante para la toma de decisiones. Dentro de las tareas

que se ejecutan en el proceso de minería de texto podemos citar la adquisición de conocimiento, preguntas/respuestas, resúmenes automáticos, recuperación de la información, argumentación computacional, categorización del texto, análisis de sentimientos; además, el proceso de adquisición de conocimiento está compuesto por técnicas de extracción de información, representación del conocimiento, co-referencia/resolución de anáfora, extracción de relaciones y descubrimiento de conocimiento. Por otra parte, para el proceso de recuperación de información podemos identificar las siguientes tareas: búsqueda semántica, búsqueda y descubrimiento exploratorio, filtrado recomendación y similaridad. Estas técnicas incluyen:

- Preprocesamiento de los documentos, que contendría la extracción de términos, eliminación de las palabras vacías (*stopwords*) y normalización de los términos restantes mediante *stemming*.
- Identificación de nombres propios. Análisis sintáctico y gramatical de los textos.
- Representación de los documentos mediante el modelo vectorial.
- Fórmulas para el cálculo de la similitud entre pares de documentos.
- *Clustering* o agrupación automática de documentos, que a su vez también toma como punto de partida la representación de los documentos según el modelo vectorial y el cálculo de similitudes.
- Categorización automática.
- Relaciones entre términos y conceptos.

2.3 Artritis Reumatoide

La artritis reumatoide es una enfermedad autoinmune común asociada con dolencia progresiva, complicaciones sistémicas, muerte prematura y costos socioeconómicos, (Firestein, 2003) Se desconoce el origen de la artritis

reumatoide y el pronóstico es reservado. Sin embargo, los avances en la comprensión de la patogenia de la enfermedad han impulsado el desarrollo de nuevas terapias con resultados probados. Las estrategias de tratamiento actuales que reflejan estos avances son comenzar un tratamiento agresivo inmediatamente después del diagnóstico, mejorar el tratamiento basado en una evaluación de la actividad de la enfermedad y continuar la remisión, (O'Dell et al., 2013).

La enfermedad es mucho más frecuente en mujeres y su prevalencia en la población latinoamericana es cercana al 0,91%, (Quiceno et al., 2011). La literatura reciente evidencia que el abordaje temprano y en algunos casos agresivo de la patología puede cambiar de forma significativa el curso clínico de la enfermedad, (K et al., 2008). La AR se clasifica como una artropatía no benigna, inflamatoria y destructiva que, sin tratamiento efectivo, lleva a discapacidad produciendo limitación parcial en el 80% de los pacientes y limitación total en el 16% después de doce años de enfermedad. La capacidad laboral de los pacientes se restringe 1/3 en un año y un 40% a los tres años, (K et al., 2008).

La capacidad de restaurar la remisión molecular y la tolerancia inmunológica sigue siendo difícil de alcanzar. El esclarecimiento de los mecanismos patogénicos de aparición y persistencia de la artritis reumatoide abre la puerta al progreso en cada una de estas áreas.

La artritis reumatoide se clasifica principalmente según el fenotipo clínico, (Aletaha et al., 2010), por tal razón los especialistas consideran que es importante recurrir a nuevas clasificaciones moleculares que identifiquen los diferentes subtipos de enfermedades con diferentes implicaciones pronósticas y terapéuticas, (Cantaert et al., 2006).

La importancia del diagnóstico temprano radica en:

- El daño radiológico, la pérdida de funcionalidad y la pérdida de densidad mineral ósea axial y periférica ocurren en el curso temprano de la enfermedad.
- La duración de la enfermedad es factor predictor de la respuesta a (DMARD) y su inicio temprano ofrece mejor pronóstico. De manera similar, se constituye una ventana de oportunidad terapéutica para la mayoría de los grupos farmacológicos en cuanto a su perfil de eficacia-efectividad y seguridad, incluyendo la terapia biológica, con mayor probabilidad de remisión en los primeros tres meses de enfermedad, (K et al., 2008).
- La discapacidad puede reducirse entre más temprano se inicie el tratamiento y más agresivo sea; una demora en solo tres meses en el inicio de DMARD produce un pobre desenlace a cinco años comparado con el inicio previo de uso de estos medicamentos.

Por estas razones, es fundamental reconocer de forma temprana al paciente con enfermedad inflamatoria articular para realizar el diagnóstico expedito de AR y así mismo cumplir los objetivos del tratamiento: aliviar el dolor, controlar la inflamación, preservar la funcionalidad, mejorar la calidad de vida y prevenir la destrucción articular, permitiendo además la reducción de los costos en el sistema de salud y todos sus componentes: Sociedad, Gobierno, entidad promotora de salud (EPS), institución prestadora de salud (IPS) y paciente.

2.4 Bioinformática Traslacional

La minería de textos para la bioinformática traslacional es un campo nuevo con un gran potencial de investigación. Es un subcampo del Procesamiento Biomédico del Lenguaje Natural (BioNLP), que trata directamente temas

relacionados con la investigación biomédica básica, la práctica clínica, y viceversa. La pregunta básica en la minería de textos para la bioinformática traslacional es cuáles son los casos de uso. Todavía no es obvio cómo las preguntas que la minería de textos para la bioinformática traslacional debería tratar de responder son diferentes de las que generalmente se resuelven en BioNLP. La respuesta depende, al menos en parte, de la naturaleza del tipo específico de información que la minería de textos debería intentar recopilar y del propósito de esa información. Sin embargo, esto solo puede recoger la superficie del campo de la minería de textos para la investigación en bioinformática traslacional, que aún no se ha definido claramente, (Cohen & Hunter, 2013).

La bioinformática traslacional representa la integración entre la medicina traslacional y la bioinformática. De forma similar, la medicina traslacional se describe como la traducción efectiva de la información obtenida de la investigación biomédica en los últimos cincuenta años para obtener conocimientos sustentables para favorecer el estado de salud humana y, por tanto, el desarrollo de enfermedades. De manera similar, para que esta transformación funcione de manera efectiva, se necesitan dos procesos: primero, la base básica de la biología debe aplicarse al desarrollo de la biología humana; segundo, desde la conclusión de la investigación clínica, se debe mejorar la salud de las personas.

Para obtener el impacto de la medicina traslacional, es necesario ampliar el papel y el alcance de la bioinformática y la informática clínica, así como desarrollar tecnología de almacenamiento de información, procesos de análisis y métodos de interpretación. La optimización transforma cada vez más datos biomédicos en métodos proactivos, predictivos, preventivos y participativos para la salud. La transformación de la bioinformática incluye investigar el desarrollo de nuevas tecnologías para integrar datos biológicos y clínicos, así como desarrollar una

metodología informática clínica que cubra las observaciones biológicas, (Dai et al., 2015).

Hay varios métodos para evaluar las aplicaciones de minería de texto, incluidos los corpus, los conjuntos de pruebas estructuradas y la evaluación de correlación coincidente. Los dos principios básicos de la estructura del lenguaje están relacionados con la construcción de aplicaciones de minería de texto. Una es que la estructura del lenguaje incluye múltiples niveles. La otra es que todos los niveles de la estructura del lenguaje son ambiguos. Hay dos métodos básicos de minería de texto: minería basada en reglas, también conocida como basada en el conocimiento; y otro basado en el aprendizaje automático, también conocido como métodos estadísticos. Muchos sistemas son una mezcla de dos métodos. Las tareas compartidas tienen un gran impacto en la dirección del campo. Como todo software de bioinformática traslacional, el software de minería de textos para bioinformática traslacional puede considerarse vital para la salud y debe cumplir con las pruebas de software más estrictas y los estándares de garantía de calidad. El producto final de la bioinformática traslacional es resultado del conocimiento que surge de estos esfuerzos de integración que pueden difundirse a una variedad de partes interesadas, incluidos los científicos biomédicos, clínicos y pacientes, (Cohen & Hunter, 2013).

3. Estado del arte

El análisis de datos médicos y biomédicos es un campo complejo y muy importante. Aunque se ha trabajado mucho con base en métodos estadísticos, ha habido pocos avances y los médicos admiten que todavía están haciendo medicina probatoria en lugar de hacer diagnósticos basados en hechos concretos. Todavía queda mucho por hacer para introducir métodos que puedan ayudar a los médicos a realizar sus diagnósticos basados en datos y métodos objetivos, (Perner, 2006).

Durante la última década, el resumen automático de documentos ha despertado el interés de la comunidad de investigadores de inteligencia artificial (IA). Recientemente, ese interés también se ha despertado en la comunidad de investigadores médicos, debido al rápido crecimiento de información disponible, por medio de revistas especializadas, conferencias, sitios médicos, portales web y registros médicos electrónicos, entre otros. Así pues, es importante entender el contexto general del resumen de documentos médicos, mostrando los factores asociados al proceso de resúmenes, discutiendo asuntos relacionados con la evaluación, así también, describiendo brevemente las diversas técnicas utilizadas para hacer resúmenes. Por tanto, se examinan las características del dominio médico a través de diferentes tipos de documentos médicos. Además, se presentan y discuten las diferentes técnicas utilizadas en este contexto, haciendo referencia a los sistemas correspondientes y sus características. Por último, se

analiza a fondo enfoques prometedores para la investigación futura en técnicas para resúmenes automáticos de documentos médicos. El enfoque está principalmente en grandes colecciones de documentos multimedia y en varios idiomas, cuestiones de personalización, portabilidad a nuevos subdominios e integración de tecnología de abstracción en aplicaciones prácticas, (Afantenos et al., 2005).

Dentro de este marco ha de considerarse también el razonamiento basado en casos (CBR) en las ciencias de la salud, que describe las tendencias y problemas actuales, y proyecta las direcciones futuras para el trabajo en este campo. De tal modo se representan los aportes de investigadores de dos talleres de razonamiento basado en casos (ICCBR-03) y la Séptima Conferencia Europea sobre Razonamiento Basado en Casos (ECCBR-04). La investigación actual sobre la (RBC) en las ciencias de la salud está marcada por su riqueza. Las tendencias destacadas incluyen trabajos en bioinformática, el apoyo a los ancianos y las personas con discapacidades, la formalización de la (CBR) en biomedicina y la minería de características y casos. Así las cosas, los sistemas (CBR) se están diseñando mejor para tener en cuenta la complejidad de la biomedicina, para integrarse en entornos clínicos y para comunicarse e interactuar con diversos sistemas y entornos, (Bichindaritz & Marling, 2006).

A propósito de los repositorios clínicos, es importante resaltar como característica que estos contienen grandes cantidades de datos biológicos, clínicos y administrativos que están cada vez más disponibles a medida que los sistemas de atención médica integran la información del paciente para los objetivos de investigación y utilización. Para investigar el valor potencial de buscar información novedosa en estas bases de datos, se aplican métodos innovadores mediante la minería de datos; así, por ejemplo, se utiliza la herramienta HealthMinder® en

una gran cohorte de 667.000 registros digitales de pacientes hospitalizados y ambulatorios de un sistema médico académico. Conviene subrayar que, HealthMinder® se acerca al descubrimiento de conocimientos utilizando tres métodos no supervisados: CliniMiner®, Análisis Predictivo y Descubrimiento de Patrones. Por último, los resultados iniciales encontrados sugieren que estos enfoques tienen el potencial de expandir las capacidades de investigación mediante la identificación de asociaciones de enfermedades clínicas potencialmente nuevas, (Mullins et al., 2006).

Bastante similar parece el trabajo de adquisición automatizada de conocimientos sobre la relación enfermedades-fármacos a partir de documentos biomédicos y clínicos: un estudio inicial. En este, se explora la capacidad de utilizar estadísticas y técnicas de minería de textos para identificar automáticamente el conocimiento de la literatura clínica y biomédica. Para este fin, se recogió información biomédica mediante las herramientas BioMedLEE y MedLEE y la información clínica se obtuvo a partir de las historias clínicas de los registros de los pacientes que reposan en las bases de datos que mantiene el NewYork-Presbyterian (NYP). Así, pues, se logró extraer conocimiento sobre la relación entre enfermedades y entidades farmacológicas. Los estudios se realizaron en un subconjunto de ocho enfermedades y aplicaron estadísticas de co-ocurrencia para calcular y evaluar la fuerza de asociación entre cada enfermedad y fármacos relacionados. Como resultado, se generó una lista ordenada de pares de enfermedad-fármaco y se calculó un punto de corte para identificar asociaciones más fuertes entre estos pares, para un análisis adicional. Se observaron diferencias y similitudes entre fuentes de texto (es decir, literatura biomédica y registros de pacientes) y anotaciones, es decir, conceptos del sistema de lenguaje médico unificado (UMLS por sus siglas en inglés) extraídos de títulos de temas médicos (MeSH por sus siglas en inglés) y procesamiento de lenguaje natural (NLP por sus siglas en

inglés), para el conocimiento sobre enfermedades y medicamentos, (Chen et al., 2008).

Ahora bien, (Coden et al., 2009) presentan un modelo de representación de conocimiento que puede ser ampliado y modificado para describir las características del cáncer de una manera comparable y consistente. Por lo tanto, utilizaron el software MedTAS/P que ejemplifica un modelo de representación de conocimiento a partir de informes de patología registrados en texto no estructurado. Así mismo, la herramienta utiliza componentes de lenguaje natural, aprendizaje automático y reglas de asociación. Además, los autores construyeron un corpus estándar de informes de patología de cáncer de colon anotado por humanos, lo que les permitió verificar la precisión del modelo. Por consiguiente, se utilizaron 302 informes del año 2004, pertenecientes a 222 pacientes seleccionados al azar con diagnóstico de cáncer de colon. Como resultado, se obtuvieron los siguientes: la puntuación F1 de MedTAS/P alcanza un puntaje de 0.97 y 1.0, para la categoría histología o anatomía en el modelo, mientras que la puntuación F1 es de 0.82 y 0.93 para la categoría tumor primario o ganglios linfáticos. Como resultado, para la categoría de tumor metastásico, la puntuación F1 obtenida es 0.65, que representa una puntuación más baja, principalmente, debido al pequeño número de casos en el conjunto de datos de entrenamiento y los datos de prueba. Asimismo, la conclusión a la que se llega es que elementos específicos del cáncer pueden ser reemplazados por elementos ajenos a la enfermedad, lo que permitirá al sistema adaptarse a modelos de representación del conocimiento relacionados con otras enfermedades o con entornos distintos al contexto de la medicina.

Otro ejemplo ilustrativo se encuentra en "La construcción de un corpus clínico semántico anotado"; en este trabajo, se describe la construcción de un corpus de

textos clínicos anotado semánticamente con el objetivo de desarrollar y evaluar sistemas que permitan extraer automáticamente información clínica relevante a partir de registros médicos de pacientes y que cuya estructura está en formato de texto libre o no estructurado. De esa manera, se detalla la revisión del texto a partir de una colección de 20.000 registros de pacientes con cáncer del Royal Marsden Hospital (RMH), así también, el desarrollo de un esquema de anotación semántica, la metodología de anotación para la cual se realizaron tareas de procesamiento del lenguaje natural. De manera similar, se usaron métodos que pueden anotar entidades clínicas y temporales y establecen correlaciones en el corpus. Como resultado, se obtuvo un corpus con anotaciones semánticas para el procesamiento clínico de textos; cabe subrayar que su uso en sistemas de extracción de información puede llegar a ser muy importante. Por tanto, se pudo establecer que la solución se puede aplicar a varios tipos de textos clínicos y, además, es muy probable que el método desarrollado se pueda aplicar a diferentes tipos de textos, (Roberts et al., 2009).

De manera semejante, se ilustra el tratamiento de registros de enfermería incluidos en los registros médicos electrónicos de pacientes hospitalizados en el Hospital Universitario de Miyazaki, para los cuales se aplicó minería de texto basada en el algoritmo KeyGraph con el objetivo de realizar análisis de los documentos y extracción de palabras claves. Los documentos analizados hacen parte de todos los departamentos de la facultad de medicina del Hospital Universitario de Miyazaki y los registros de enfermería datan del año 2007. La metodología desarrollada se basó en la extracción y clasificación de los términos más frecuentes y las coocurrencias de los términos en el conjunto de datos. De ahí que, los autores obtuvieron como resultado 18 grafos asociados a cada uno de los departamentos de la facultad de medicina de la Universidad de Miyazaki, para los cuales definieron las especialidades asociadas en cada departamento,

las patologías más frecuentes de los pacientes para dichas especialidades. Los términos relacionados en los tratamientos farmacológicos, efectos adversos, procedimientos quirúrgicos, entre otros. Además, un conjunto de palabras clave asociadas que constituyen las bases del grafo que se obtienen del texto con conjuntos de eventos. Las oraciones se analizaron en morfemas y las relaciones entre los vocabularios de características se analizaron mediante una técnica de minería de texto. El resultado del análisis de los registros cualitativos de enfermería hospitalaria utilizando la técnica de minería de texto logró el objetivo de obtener un registro visual de la información. Además, este resultado identificó vocabularios relacionados con los métodos adecuados de tratamiento, lo que resultó en un resumen conciso de los vocabularios extraídos de los registros de enfermería de pacientes hospitalizados. Finalmente, esta investigación sugiere la fructífera posibilidad de detectar automáticamente una enfermedad y clasificarla a partir de documentos utilizados en los tratamientos médicos. En el futuro, el uso de un enfoque de minería de texto y el procesamiento lateral de documentos médicos respalda la clasificación de enfermedades al recuperar ejemplos de síndromes similares, (Kushima et al., 2011).

Algo semejante ocurre en el enfoque de minería de textos basado en la clasificación internacional de enfermedades (CIE-9) para la clasificación de epilepsia infantil usando el algoritmo de los k-vecinos más cercanos, para un conjunto de datos reales de pacientes anónimos del distrito hospitalario de emergência de Pombal - Centro hospitalario de Leiria Portugal. Se clasificaron de manera automática diagnósticos epilépticos con el objetivo de asociar los códigos de las enfermedades al estándar CIE-9. Para llevar a cabo el proceso, se ejecutaron tareas de preprocesamiento de la información almacenada como texto libre. Como resultado, los autores encontraron que los registros de crisis focal simple no fueron suficientes para obtener un resultado confiable luego de aplicar

los algoritmos de aprendizaje automático. Debido a lo anterior, se llevaron a cabo pruebas adicionales, eliminando los registros de convulsión focal simple del conjunto de datos. Finalmente, los autores concluyen que este trabajo es una propuesta que permite brindar apoyo en las decisiones de médicos pediatras en un escenario real; no obstante, los autores recalcan que los resultados son aún preliminares, esto significa que para poder tener unos resultados satisfactorios, es necesario tener un volumen significativo de datos que permita generar un modelo de entrenamiento robusto, así también, información que permita clasificar otros tipos de convulsiones y, además, poder obtener las características más relevantes de la enfermedad que permita lograr resultados con niveles de precisión más significativos, (Pereira et al., 2013).

La prevalencia de signos y síntomas de insuficiencia cardíaca en una gran población de atención primaria identificada mediante el uso de minería de textos aplicado a la historia clínica electrónica. Como se ha mencionado en párrafos anteriores, la historia clínica electrónica (HCE) contiene una enorme cantidad de datos que, si se detectan adecuadamente, pueden conducir a una identificación más temprana de estados patológicos como la insuficiencia cardíaca (IC). Así, utilizando una novedosa herramienta de análisis de textos, se hace un reconocimiento de la HCE de más de 50.000 pacientes de atención primaria, con el fin de identificar la documentación de los signos y síntomas de la IC en los años anteriores a su diagnóstico. Bajo esta perspectiva, se realizaron análisis retrospectivos de 4,644 casos incidentes de IC y 45,981 sujetos de control emparejados por grupos. La documentación recogida de los signos y síntomas de Framingham de falla cardiaca (HF por sus siglas en inglés), se llevó a cabo con el uso de un procedimiento de procesamiento del lenguaje natural previamente validado. Como resultado de este trabajo, se documentaron un total de 892,805 criterios ratificados durante un período de observación promedio de 3.4 años.

Entre los casos eventuales de IC, el 85% tenía un criterio en el año anterior al diagnóstico de IC, al igual que el 55% de los sujetos de control. Se encontró una variabilidad sustancial en la prevalencia de signos y síntomas individuales tanto en sujetos de casos como de controles. Por consiguiente, los signos y síntomas de la IC se documentan con frecuencia en una población de atención primaria según se identifican mediante la extracción automática de textos de las HCE. En conclusión, su identificación frecuente demuestra la gran cantidad de datos disponibles dentro de los registros electrónicos que permitirán el trabajo futuro en la identificación de criterios automatizados para ayudar a desarrollar modelos predictivos para (HF) (Vijayakrishnan et al., 2014).

Las descripciones precedentes han permitido darnos cuenta de que la minería de textos es una tecnología flexible que se puede aplicar a numerosas tareas diferentes en biología y medicina. Bajo esta perspectiva, se presenta un sistema a fin de extraer asociaciones de genes-enfermedades a partir de resúmenes biomédicos. El sistema consiste en un marcador basado en un diccionario eficiente para identificar entidades nombradas en genes humanos y enfermedades. Este se combina con un esquema de puntuación que tiene en cuenta la co-ocurrencia dentro y entre oraciones. A causa de ello, se demuestra que esta metodología es capaz de extraer la mitad de todas las anotaciones curadas manualmente con una tasa de falsos positivos de sólo el 0.16%. Sin embargo, la minería de textos no debe ser independiente, sino que debe combinarse con otros tipos de evidencia. Por lo tanto, se desarrolló el recurso DISEASES, que combina los resultados de la minería de texto con asociaciones de genes-enfermedades seleccionadas manualmente, datos de mutaciones del cáncer y estudios de asociación de todo el genoma en bases de datos existentes, como por ejemplo el estudio de asociación de genoma completo (GWAS por sus siglas en inglés). En consecuencia, se ha desarrollado una herramienta de

reconocimiento de entidades nombradas (NER por sus siglas en inglés) que permite saber de la enfermedad, cuál es su naturaleza y cuáles sus síntomas, características y peculiaridades y se ha combinado con un esquema de puntuación de co-ocurrencias que permite extraer de forma eficiente y precisa asociaciones genes-enfermedades desde Medline. Así mismo, se han integrado estas herramientas con las asociaciones establecidas de forma manual de las bases de datos del receptor de la hormona de crecimiento (GHR por sus siglas en inglés) y base de datos de conocimiento UniProt (UniProtKB por sus siglas en inglés), así como datos de mutaciones somáticas almacenadas en GWAS desde el Catálogo de mutaciones somáticas en cáncer (COSMIC por sus siglas en inglés) y enfermedades y rasgos en el desequilibrio de ligamiento (DISTILD por sus siglas en inglés), respectivamente. Finalmente, la base de datos resultante está disponible como un recurso web de búsqueda en http://diseases.jensenlab.org/, donde los conjuntos de datos masivos y el software (NER) también están disponibles para ser descargados, (Pletscher-Frankild et al., 2015).

En, (Wang et al., 2018), el conjunto de datos o corpus está constituido por la historia clínica electrónica (EMR por sus siglas en inglés) de los pacientes con cáncer de esófago (CE) del departamento de oncología radioterápica, del Hospital Oncológico de Shandong, afiliado a la Universidad de Shandong, Jinan, China, para los que se incluyeron eventos médicos tales como los síntomas clínicos, los exámenes, los resultados de los exámenes y los planes de diagnóstico y tratamiento del paciente. Así, pues, la técnica utilizada para el análisis de la información fue aprendizaje profundo de word2vec; esta se utilizó para entrenar un gran corpus de registros médicos electrónicos. A diferencia de otros métodos, en esta investigación, se construyó la representación semántica de cada entidad (palabras distribuidas) y se intentó detectar grupos de entidades en la EMR

basados en la similitud semántica de las entidades. De este modo, combinaron la representación de palabras distribuidas y los métodos de agrupamiento y luego proporcionaron los detalles del algoritmo y los resultados experimentales del esquema de agrupamiento. En conclusión, al utilizar una gran cantidad de información médica basada en dominios para localizar palabras en un espacio de alta dimensión, se puede mejorar la recuperación y la riqueza de la base de conocimientos de (CE). Por ende, utilizar más vectores de palabras basados en métodos de aprendizaje profundo (como una red neuronal Long Short Term Memory (LSTM) para realizar la representación distribuida de entidades médicas y el reconocimiento de estas entidades es muy importante para encontrar las relaciones entre entidades y así lograr el propósito de construir el gráfico de conocimiento CE.

En otro trabajo, en (Vittal & Karthikeyan, 2018) se propone encontrar compuestos que inhiben el cáncer oral mediante el modelado de detección de asociación, se utiliza el algoritmo Medusa en paralelo con la clasificación binaria que es una técnica de aprendizaje automático, lo que permite descubrir compuestos potenciales que inhiben el cáncer oral. El cáncer oral afecta la cavidad oral y la faringe, y tiene una alta tasa de mortalidad debido a su detección tardía. Los métodos actuales de tratamiento del cáncer oral, como la quimio radiación y la cirugía, no pueden brindar una mejor probabilidad de supervivencia, lo que justifica la búsqueda de otro método. Generalmente, el aprendizaje automático (especialmente basado en la supervisión y algoritmos basados en gráficos) proporciona un nuevo método de análisis de datos biológicos, que no solo es más económico que las pruebas, sino que también tiene gran precisión debido a la gran cantidad de datos disponibles en servidores web y bases de datos. Se pueden ver ejemplos de tales aplicaciones en el análisis con técnicas de bosques y el aprendizaje profundo, lo que demuestra la alta precisión de la predicción de la

interacción fármaco-objetivo. Se cree que el uso de bases de datos de grafos jugará un papel en el descubrimiento de compuestos que inhiben el cáncer oral y otras enfermedades. Específicamente, los algoritmos basados en grafos son particularmente importantes para encontrar conexiones de fármacos a través de vínculos y detección de comunidades, y el uso de la teoría de grafos químicos para examinar la relación entre la estructura química de un compuesto y su impacto biológico. Específicamente para oncología, el análisis basado en la web puede lograr una oncología precisa al permitir un análisis genómico más preciso, pero aún enfrenta desafíos en la calidad de los datos, la realización de resultados y la escalabilidad. Estos desafíos prueban las limitaciones del algoritmo Medusa. Sin embargo, estos desafíos indican el potencial de mejora, ya que los datos de mayor calidad permitirán tener un mejor análisis y la mayor escalabilidad permitirá el análisis de una gama más amplia de compuestos. En conclusión, se espera, que la importancia estadística de los compuestos encontrados en este artículo se traduzca en importancia biológica; además, se espera que este trabajo sea una invitación a la aplicación del aprendizaje automático para examinar las interacciones entre medicamentos.

En (Harrer et al., 2019), se plantea que los ensayos clínicos consumen la segunda mitad del ciclo de desarrollo de 10 a 15 años, entre 1.5 y 2.0 mil millones de dólares, para llevar un único fármaco nuevo al mercado. Por lo tanto, un ensayo fallido reduce no solo la inversión en el ensayo en sí, sino también los costos de desarrollo preclínico, lo que hace que la pérdida por ensayo clínico fallido sea de 800 millones a 1,400 millones de dólares. Las técnicas de selección y reclutamiento de cohortes de pacientes sub-óptimas, junto con la incapacidad de monitorear a los pacientes de manera efectiva durante los ensayos, son dos de las principales causas de las altas tasas de fracaso de los ensayos: sólo uno de los 10 compuestos que ingresan a un ensayo clínico llega al mercado. Bien,

pareciera por todo lo anterior, que es interesante advertir, cómo los avances recientes en inteligencia artificial (IA) se pueden utilizar para remodelar los pasos clave del diseño de ensayos clínicos para aumentar las tasas de éxito de los ensayos. Los métodos de (IA) y aprendizaje de máquina (ML) también se pueden utilizar para predecir dinámicamente el riesgo de abandono de un paciente específico, en otras palabras, para detectar el inicio de la conducta del paciente que sugiere que el paciente podría estar experimentando problemas para adherirse al protocolo del estudio. Uno de estos ejemplos describió el uso de algoritmos de aprendizaje de refuerzo profundo, para determinar la menor cantidad, la menor dosis que aún podría encoger los tumores cerebrales, mientras se reduce la toxicidad asociada con los regímenes de dosificación de quimioterapia, (Yauney & Shah, 2018).

Impulsado por una técnica de aprendizaje automático, el sistema analiza los regímenes de tratamiento actualmente en uso y ajusta las dosis de forma iterativa. Finalmente, se encuentra un plan de tratamiento óptimo, con la menor potencia y frecuencia de dosis posibles que aún deberían reducir el tamaño de los tumores en un grado comparable al de los regímenes tradicionales. En ensayos simulados de 50 pacientes, el modelo de (ML) diseñó ciclos de tratamiento que redujeron la potencia a un cuarto o la mitad de casi todas las dosis, manteniendo el mismo potencial de reducción del tumor y, por lo tanto, promete mejoras en la adherencia del paciente y reducciones en los abandonos. Así, también, detectar las señales de advertencia tempranas de incumplimiento permite un compromiso proactivo con los pacientes individuales y permite abordar las causas fundamentales de la conducta problemática: por ejemplo, los efectos secundarios graves o la incompatibilidad del estudio y las rutinas personales podrían detectarse y corregirse antes de que lleven al abandono. La elección de sensores

y modelos analíticos depende en gran medida de la enfermedad y deberá formar parte del diseño del estudio clínico.

Un trabajo similar es la investigación relacionada con la detección automática de patrones de diagnóstico basados en notas clínicas a través de la minería de texto por (Ribeiro et al., 2019); aquí, se destaca la importancia del tratamiento estandarizado para los pacientes porque puede reducir el tiempo de espera, reducir los costos hospitalarios y hacer que el tratamiento sea más efectivo para los pacientes. Para analizar el texto no estructurado, se utilizó la herramienta RapidMiner con la que se logró obtener la frecuencia de palabras y el correspondiente recuento de estas, así como el análisis de grupos para crear combinaciones de palabras. La herramienta RapidMiner ayuda a transformar las anotaciones clínicas aplicando las diferentes técnicas del procesamiento de lenguaje natural, lo que permite obtener información útil para el análisis. Con la creación de modelos de análisis de texto basados en el uso de grupos (clústeres), fue factible detectar los patrones clínicos existentes de forma automática, Finalmente, se concluye que, para lograr el propósito principal de la investigación, el foco debe estar en la creación de modelos de pronóstico de diagnóstico. En la actualidad, la predicción de diagnósticos es muy útil porque puede reducir el tiempo de inactividad y reducir la grave posibilidad de errores o fallas que puedan existir en este, lo que beneficia a la organización y la sociedad en su conjunto. En definitiva, según los autores, será muy beneficioso y gratificante continuar toda la investigación médica, ya que hay una ingente cantidad de información no estructurada que se puede analizar en beneficio de la toma de decisiones en relación con los diagnósticos clínicos.

El uso de la minería de textos para extraer síntomas depresivos de registros médicos electrónicos y verificar el diagnóstico de depresión mayor, tiene como

objetivo evaluar el registro en la Base de datos de investigación de seguros de salud nacional (NHIRD) de Taiwán. Por lo tanto, se utilizó la minería de textos para extraer las características de los síntomas y el deterioro funcional a partir de la historia clínica electrónica. Desde esta perspectiva, se seleccionaron al azar un total de 500 notas de alta de la base de datos del centro médico. Así también, se revisaron las anotaciones para establecer el “estándar de oro”. De manera similar, se evaluó la precisión del código de diagnóstico para enfermedades mentales graves. Además, los métodos de minería de texto se utilizaron para extraer síntomas depresivos y características funcionales e identificar pacientes con depresión severa. Por último, los resultados de la precisión del código de diagnóstico para la depresión mayor, la esquizofrenia y la demencia fueron aceptables, pero los resultados del diagnóstico para el trastorno bipolar y la depresión leve no fueron satisfactorios. En este estudio, también, encontraron que la precisión de las instrucciones de descarga del código de diagnóstico ICD-9 es aceptable. Estos hallazgos apoyan el uso del diagnóstico psiquiátrico en el NHIRD de Taiwán. También, descubrieron que tanto los métodos basados en diccionarios de campos aleatorios condicionales (CRF por sus siglas en inglés) pueden identificar con precisión los síntomas de depresión. Sin embargo, la evaluación del nivel funcional con este método aún no es satisfactoria. Puede requerir un tamaño de muestra de investigación más grande o el uso de otros métodos de aprendizaje automático para mejorar la precisión, (Wu et al., 2020).

Lo anterior, hace parte de la revisión de literatura en el contexto Norte Americano, europeo no hispano parlante y asiático, principalmente, por lo que es importante referenciar dentro del ámbito latinoamericano y colombiano el desarrollo de investigaciones relacionadas con la minería de textos en contextos médicos y, de manera más general, en ciencias de la vida. Una vez hecha esta precisión, continuemos con el descubrimiento de conocimiento en historias clínicas

mediante minería de textos por (Carrascal et al., 2019), en pacientes de traumatología del Hospital Universitario San Vicente Fundación de Medellín, para esta investigación se aplicaron técnicas para identificar palabras más frecuentes, segmentación de los episodios para encontrar similitudes: agrupación (*clustering*) jerárquica, palabras más usadas en cada tipo de trauma, coocurrencias, predicción del trauma según la palabra ingresada. A partir de lo anterior, los autores analizaron y clasificaron los resultados de la siguiente forma, palabras más frecuentes en el área de traumatología, segmentación de los episodios para encontrar similitudes en los traumas, encontrar las palabras más utilizadas en cada tipo de trauma, analizar la ocurrencia de las palabras en los diferentes tipos de trauma, predecir el tipo de trauma según las palabras ingresadas por el médico. Por último, los resultados de los modelos desarrollados fueron frecuencia de aparición de cada palabra, cuatro grupos de traumas que usan palabras muy similares y las palabras que ocurren de forma conjunta y la predicción de tipo de trauma según las palabras del médico tratante.

En (Hernandez & Quimbaya, 2016) se desarrolló un modelo para la extracción, estructuración y visualización de eventos médicos a partir de texto narrativo en historias clínicas electrónicas presenta el modelo Health Text Line Model (HTL); este modelo permite realizar tareas de extracción de texto, estructuración y visualización de citas médicas, tratamientos, prescripción de fármacos, entre otras, a partir del texto libre consignado en las historias clínicas electrónicas. El modelo HTL fue implementado en un software que integra los procesos antes mencionados con el fin de identificar y establecer una relación de tiempo con los eventos médicos, lo que permitió compararlos con las guías de tratamiento establecidas para un conjunto específico de enfermedades. Por esta razón, es oportuno decir que el modelo y el software fueron validados tomando un caso de estudio aplicado a historias clínicas en formato de texto libre del Hospital

Universitario San Ignacio en la ciudad de Bogotá. Bajo esta perspectiva y para concluir, los autores resaltan la importancia de la herramienta debido a que permite al médico tratante contar con un soporte que fortalece la toma de decisiones, así como también, para realizar un análisis retrospectivo en el diagnóstico, seguimiento y tratamiento de una enfermedad.

En otro caso de estudio, se plantea un análisis para el Sistema Integral para la Atención Primaria de Salud (SIAPS), aplicando la técnica de agrupamiento enmarcada en el algoritmo K-Means, con el propósito de realizar observaciones de la información clínica de los pacientes; para ello se proyecta la extracción del conocimiento del almacén de datos mantenido en el repositorio de historias clínicas electrónicas. La investigación se apoya en la herramienta de libre distribución WEKA, esta funciona de forma aislada al SIAPS. En el desarrollo de la solución, los autores utilizaron el lenguaje de programación Java 1.6, el servidor de aplicaciones JBoss 4.2 y Eclipse 3.4 como plataforma de desarrollo, y el sistema de gestión de base de datos PostgreSQL 8.4 y JBoss SEAM como marco integrado; además, durante todo el proceso, se utilizó la plataforma Java Enterprise Edition 5.0. Como resultado, se plantea que es deseable obtener una visión analítica que ayude a comprender el modelo generado, apoyando así el proceso de toma de decisiones clínicas. Es importante subrayar, que se deben seguir haciendo esfuerzos en el desarrollo de sistemas informáticos que ayuden a satisfacer las necesidades actuales, dentro de este campo de conocimiento. En definitiva, se puede afirmar que la minería de datos es un proceso eficaz para responder preguntas complejas en el contexto del negocio. En suma, esta es una excelente manera de convertir datos en información y luego en conocimiento para tomar decisiones clínicas correctas, (Ochoa et al., 2013).

4. Metodología para el análisis de historias clínicas de pacientes con Artritis Reumatoide

En este capítulo se presenta la metodología por medio de la cual se implementa un flujo de trabajo que permite llevar a cabo los objetivos planteados para esta investigación. Para tal efecto, se ejecutan actividades que permiten:

- Acceder y transformar los datos no estructurados de las historias clínicas
- Llevar a cabo tareas de limpieza de los datos
- Encontrar el número óptimo de temas
- Anotar el texto
- Buscar términos relevantes
- Hacer un análisis de las secciones de la historia clínica
- Encontrar palabras claves
- Implementar la interoperabilidad semántica
- Evidenciar los tratamientos farmacológicos.

4.1 Acceso y transformación de los datos

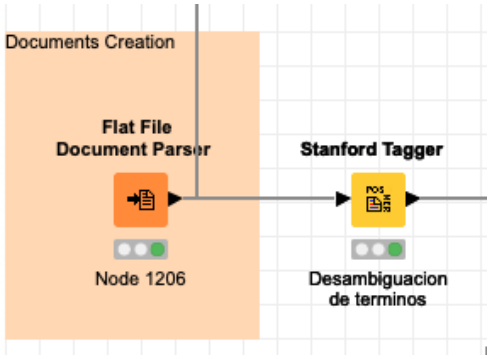
El modelo computacional que se propone para el análisis de historias clínicas de pacientes con artritis reumatoide analiza un conjunto de documentos (2788 historias clínicas) que constituye el corpus del problema y que contienen las características de la enfermedad dado que son pacientes diagnosticados con Artritis Reumatoide. El corpus está compuesto por historias clínicas de consulta (primera consulta con el especialista) e historias clínicas de control o evolución de los pacientes. Las historias clínicas son de un conjunto de pacientes de la clínica Fundación Santa Fe de Bogotá.

El desarrollo de este modelo se realizó mediante la aplicación de minería de textos, procesamiento de lenguaje natural, lingüística computacional, recuperación de la información, ciencias de datos clínicos y otras áreas relacionadas con el análisis de datos no estructurados o semi estructurados.

Teniendo en cuenta que la información contenida en las historias clínicas es de carácter personal y privado, la primera tarea que se realizó fue la de garantizar el anonimato absoluto de dicha información, tanto de pacientes como de los profesionales de la salud allí relacionados. Es importante anotar que se siguieron las mejores prácticas recomendadas en el área de ciencias de datos clínicos (*clinical data science*).

Las historias clínicas analizadas fueron entregadas por el doctor Gerardo Quintana, y como se dijo anteriormente, este conjunto de historias clínicas pertenecen a una población de pacientes diagnosticados con Artritis Reumatoide en la Fundación Santa Fe de Bogotá.

Figura 4-1 Carga de documentos y desambiguación



4.2 Limpieza de las historias clínicas

La primera tarea realizada para el análisis de las historias clínicas fue escribir un script en Python que permitió transformar el formato de las historias de PDF a formato TXT, encontrar dentro de las historias clínicas los pacientes y médicos y, así mismo, eliminar los nombres de pacientes y médicos registrados. Además, se creó un script que permitió dividir el conjunto de 2,788 historias clínicas en dos subconjuntos: historias clínicas de primera consulta e historias clínicas de control médico. Una vez realizadas estas tareas iniciales se comenzó a construir el flujo de análisis de las historias clínicas en Knime como se ilustra en la figura 4-1. Este flujo de trabajo inició con la carga de las historias clínicas, que conforman el corpus documentos para el análisis.

Knime provee artefactos denominados nodos que permiten ejecutar las actividades necesarias para llevar a cabo cada paso del proceso de minería de textos. En este primer paso del flujo de trabajo (*workflow*) en Knime se utilizó el nodo Flat File Document Parser, que permitió cargar los archivos y tratarlos directamente como documentos, sin necesidad de hacer transformaciones adicionales. En este nodo se configuró el directorio donde están almacenadas las historias clínicas, se marcó la opción de hacer una búsqueda recursiva de los

archivos y se indicó que se ignoraran los archivos ocultos. Además, se establecieron los metadatos: categoría del documento, fuente del documento, tipo de documento, segmentador de palabras (análisis léxico) y codificador de caracteres (UTF-8).

Posteriormente, se realizó la tarea de desambiguación del texto, para tal fin se utilizó el nodo Stanford Tagger; este nodo se configuró estableciendo la columna Documento como entrada y, luego, se configuró para agregar una nueva columna que permitió identificar los documentos etiquetados. Así también, se parametrizó el nodo para utilizar el modelo Stanford NLP Spanish Tokenizer, esto con el fin de poder reconocer textos en español, ya que el modelo utiliza el corpus de AnCora 3.0/DEFT Spanish Treebank.

Una característica de las historias clínicas es que contienen una gran cantidad de abreviaturas que identifican tratamientos, fármacos y otras entidades que no son inicialmente fáciles de reconocer. Por lo tanto, se creó un diccionario de abreviaturas médicas que consta de 3,406 registros y que se clasifican en dos columnas: abreviatura y definición. Este diccionario y las historias clínicas etiquetadas con el nodo Stanford Tagger son las entradas para el nodo Dictionary Replacer que permitió reemplazar las abreviaturas encontradas en las historias clínicas por la definición de estas, lo que tiene como resultado un enriquecimiento y desambiguación del texto.

En la siguiente actividad, se creó en un meta-nodo nombrado preprocesamiento; este metanodo está compuesto por ocho tareas como se puede ver en la figura 4-2:

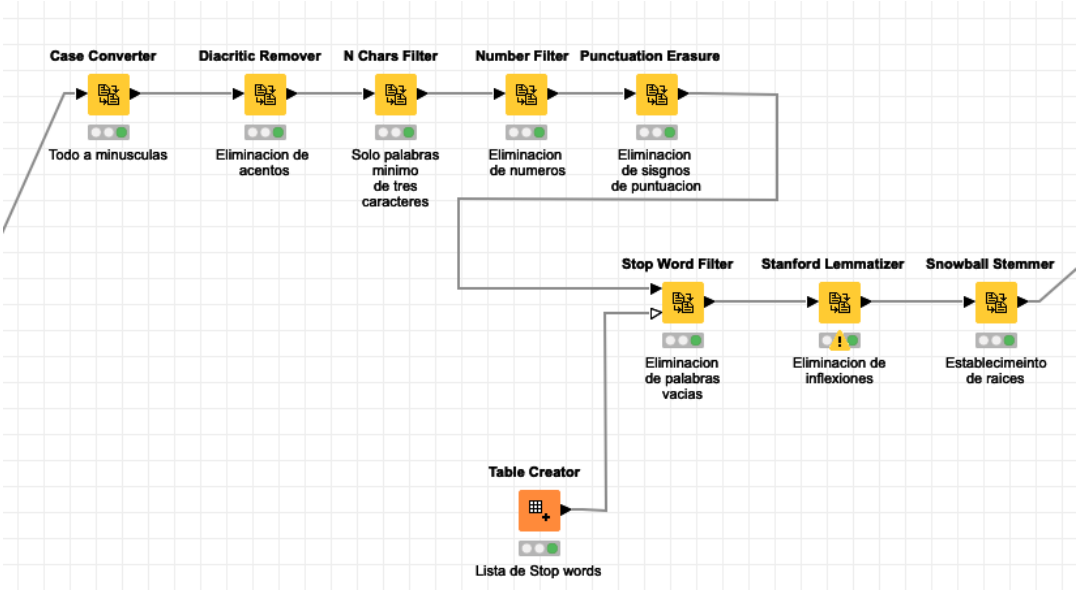
1. Case Converter: en este nodo en la opción del convertidor se configura la opción *convertir todo a minúsculas* y en la pestaña de preprocesamiento se

asigna la columna que contiene los documentos y se mantiene la opción de reemplazar columna por la que contiene los documentos etiquetados. Este nodo permite transformar los documentos llevando todas las palabras a minúsculas, el objetivo de esta actividad es normalizar el texto y asegurar que todas las palabras tienen el mismo nivel de importancia.

2. Diacritic Remover: la configuración de este nodo sólo pide que se seleccione la columna donde están los documentos de entrada, la tarea ejecutada fue eliminar los acentos de los términos para seguir normalizando el texto.
3. N Chars Filter: en las opciones de filtro en el parámetro N Chars se establece el número tres, esto quiere decir que las palabras que se encuentren en el texto que tengan menos de tres caracteres serán eliminados; este parámetro se escogió debido a que dentro de las anotaciones que contiene una historia clínica se encuentran muchas abreviaturas compuestas por tres letras que pueden aportar algún significado en el análisis realizado. Sin embargo, en un paso previo se trata de mitigar este hecho, creando un diccionario de abreviaturas y haciendo el reemplazo de estas por su definición.
4. Number Filtert: en la pestaña de opciones del filtro se seleccionó el parámetro Filter terms containing numbers; esta opción permite filtrar todos aquellos términos que contengan al menos un dígito, además de los números y signos decimales; el objetivo es eliminar todas las referencias numéricas que se encuentren dentro de los documentos, ya que estas no aportan un valor dentro del análisis de minería de textos.
5. Punctuation Erasure: con este nodo se eliminaron todos los caracteres de puntuación existentes en el texto, este nodo sólo recibe como parámetros la columna de documentos pre procesados.

6. Stanford Lemmatizer: con este nodo se realizó la lematización de las palabras; esta tarea consistió en normalizar las palabras eliminando las flexiones de estas, es decir, se relaciona una palabra flexionada o derivada con su forma canónica o lema. En lingüística el lema es una unidad autónoma constituyente del léxico de un idioma. En la pestaña de preprocesamiento se establece la columna que contiene los documentos de entrada y se mantiene la columna de los documentos pre procesados.
7. Snowball Stemmer: la tarea ejecutada con este nodo consistió en ejecutar el proceso que permite hallar las raíces de las palabras, esto significa dejar la unidad invariable de las palabras contenidas en los documentos de entrada. En las opciones de radicalización o stemmer se estableció en el parámetro Snowball Stemmer el valor Spanish, esto debido a que los documentos analizados están en idioma español. Al igual que en el nodo anterior, en la pestaña de preprocesamiento se establece la columna que contiene los documentos de entrada y se mantiene la columna de los documentos preprocesados.
8. Stop Word Filter: para la realización de esta tarea se creó una lista personalizada de palabras vacías que se configuró en un nodo Table Creator; en este nodo se cargó una lista personalizada de palabras. El nodo Stop Word Filter tiene dos interfaces de entrada, una para los documentos y la otra que es opcional para otra fuente de datos que en este caso fue la lista personalizada. En la pestaña de opciones del filtro se establece el parámetro Use custom list, este parámetro pide que se establezca la columna Stopword column, la cual tiene como valor column1. Al igual que en el nodo anterior, en la pestaña de preprocesamiento se estableció la columna que contiene los documentos de entrada y se mantuvo la columna de los documentos preprocesados.

Figura 4-2 Preprocesamiento básico de texto



4.3 Encontrar el número óptimo de temas

Una vez que los datos se han limpiado y filtrado, el nodo "Extractor de temas" se puede aplicar a los documentos. Este nodo utiliza una implementación del modelo LDA (Latent Dirichlet Allocation), que requiere que se defina la cantidad de temas que se deben extraer de antemano. Esto sería relativamente fácil si ya se supiera cuántos temas extraer de los datos. Pero a menudo, especialmente en datos no estructurados como el texto, puede ser bastante difícil estimar por adelantado cuántos temas hay.

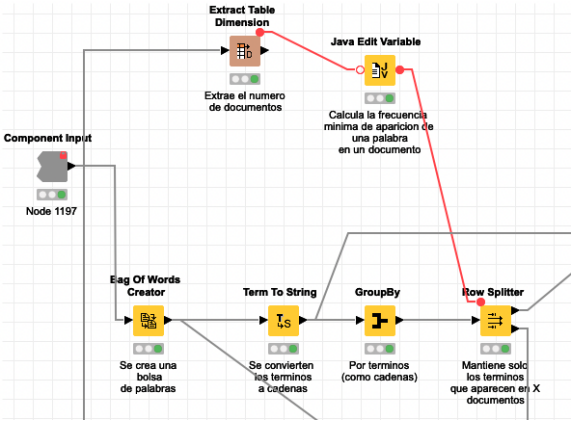
Hay algunos métodos entre los que se puede elegir para determinar cuál sería un buen número de temas. En este flujo de trabajo, se utiliza el método del "Codo" para agrupar los datos y encontrar el número óptimo de grupos y se verifica con el método silueta. Se plantea que el número óptimo de clúster se relaciona con un buen número de temas.

Para usar la agrupación en clústeres, se requiere pre procesar los datos nuevamente, esta vez extrayendo los términos que se necesitan usar como características para los vectores del documento. Todos los pasos de preprocesamiento se empaquetan en otro metanodo llamado "Preprocesamiento". Básicamente, los pasos implican crear una bolsa de palabras (BoW) de los datos de entrada. Puede resultar útil eliminar términos que ocurren muy raramente en toda la colección de documentos, ya que no tienen un gran impacto en el espacio de funciones, especialmente si la dimensión del BoW es muy grande. A continuación, se crean los vectores del documento.

Después de todo este preprocesamiento para la elaboración del texto, se llega a un punto central de este análisis: extraer los términos, que se van a utilizar como características en los vectores del documento, y así tener en cuenta para la clasificación de los documentos posteriormente. Para poder crear vectores de documento, usando el nodo "Vector de documento", primero se debe crear una tabla de datos de bolsa de palabras, usando el nodo "Bag Of Words Creator". El nodo "Vector de documento" requiere una bolsa de palabras como tabla de datos de entrada y tiene en cuenta todos los términos contenidos en esta bolsa para crear los vectores de documento.

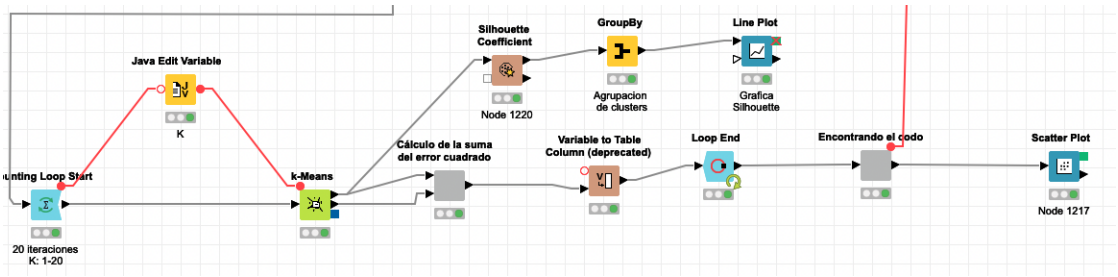
Una vez creada la bolsa de palabras, filtramos todos los términos que aparecen en menos de 54 documentos. Hacemos esto agrupando por términos, contando todos los documentos únicos que contienen estos términos, filtrando esta lista de términos y, finalmente, filtrando la bolsa de palabras con el nodo "Filtro de fila de referencia". De este modo, reducimos el espacio de características de 876,558 palabras distintas a 2,793. La extracción de características es parte del metanodo "Preprocesamiento" y en la figura 4-3 se pueden ver los nodos que hacen parte de esta actividad.

Figura 4-3 Filtrado de palabras en función de la frecuencia en el corpus:



4.4 El método del codo

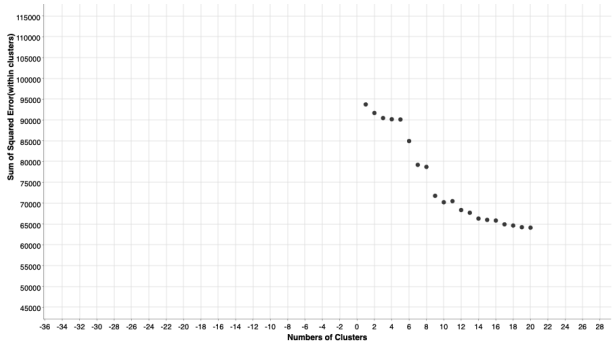
Figura 4-4 Bucle para calcular agrupaciones de k-medias basadas en diferentes valores de k:



Ahora que se han convertido las historias clínicas en vectores de documentos, se puede comenzar a agruparlos usando el nodo "k-Means". La figura 4-4 muestra el proceso para poder obtener una gráfica mediante el método del “Codo” que es básicamente ejecutar la agrupación de k-medias en los datos de entrada para un rango de valores del número de grupos k (por ejemplo, de 1 a 20), y para cada valor de k calcular posteriormente la suma de errores cuadrados dentro del grupo, que es la suma de las distancias de todos los puntos de datos a sus respectivos centros de agrupamientos. Luego, se traza en un gráfico de dispersión. El mejor

número de conglomerados es el número en el que hay una caída en el valor de SSE, lo que da un ángulo en la gráfica, como se puede observar en la figura 4-5.

Figura 4-5 Gráfica del codo:

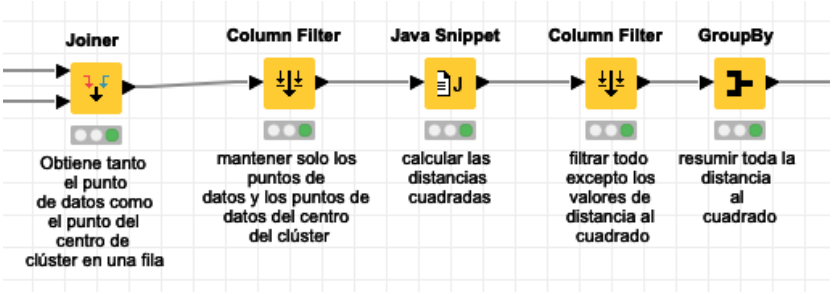


Como se ha mencionado, el nodo "k-Means" se aplica para agrupar los datos en k grupos. El nodo tiene dos interfaces de datos de salida: el primer puerto contiene una tabla de datos de todos los vectores de documentos y sus correspondientes identificadores de grupos a los que están asignados, y el segundo puerto contiene vectores de todos los centros de los clústeres. A continuación, como se observa en la figura 4-6, se utiliza el nodo "Joiner" para unir ambas tablas de datos de salida en función del ID del agrupamiento. El objetivo es obtener tanto los vectores del documento como el vector de sus respectivos centros de clústeres en cada fila para facilitar el cálculo. Después de eso, el nodo "Java Snippet" se usa para calcular la distancia al cuadrado entre un vector y su vector central de clúster en cada fila. El valor del (SSE) para este cierto número k de clústeres es la suma de todas las distancias al cuadrado, donde la suma es calculada por el nodo "GroupBy".

Para buscar el codo en un gráfico de dispersión, este cálculo se ejecuta en un rango de números de k, que en este caso es de 1 a 20. Se logró esto mediante el

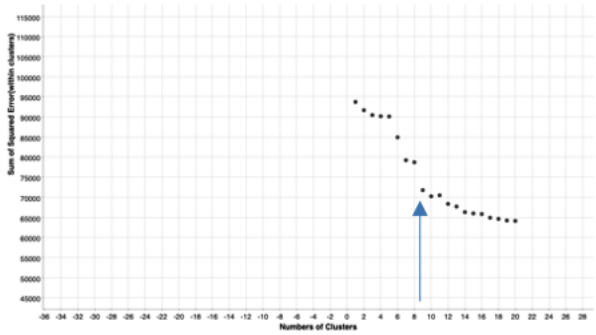
uso de nodos de bucle para realizar la tarea de manera repetitiva. El valor k de la iteración actual se proporciona como variable de flujo y también controla la configuración k del nodo k-Means.

Figura 4-6 Flujo suma errores cuadrados:



El nodo "Scatter Plot" se utiliza para generar un gráfico de dispersión figura 4-5 del número de grupos o conglomerados k frente al valor de la suma del error cuadrado (SSE). Se puede observar que el error disminuye a medida que k aumenta. La idea del método del codo es elegir el número de puntos en los que el (SSE) disminuye abruptamente. Esto produce un llamado "codo" en el gráfico. En el gráfico de la figura 4-5, se puede ver que esta inflexión está después de k = 9. Por lo tanto, una elección de 10 clústeres parecería ser el número óptimo, como se observa en la figura 4-7.

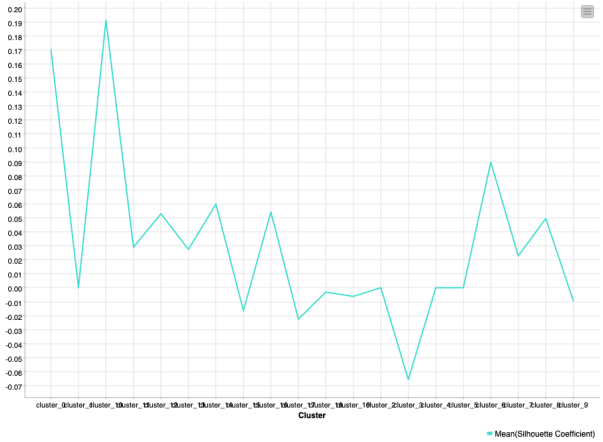
Figura 4-7 Cluster 10 donde se produce la inflexión:



El número óptimo de clústeres se determina automáticamente en el flujo de trabajo, tomando las distancias de los valores SSE posteriores y ordenando el paso con la mayor distancia en la parte superior. A continuación, se proporciona el número relacionado k como variable de flujo para el nodo “Extractor de temas”. Es importante mencionar que el método del codo es heurístico y es posible que no siempre funcione para todos los conjuntos de datos. Si no se encuentra un codo claro en el gráfico, se puede intentar utilizar un enfoque diferente, por ejemplo, el coeficiente de Silueta, que, para este caso, confirma la información obtenida del método de codo. En la figura 4-8 se puede observar que el valor más alto de la gráfica es el que pertenece al clúster 10 por lo que podemos validar la escogencia del codo mediante la gráfica del coeficiente de Silueta. El coeficiente de Silueta es una métrica usada para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering. El objetivo de Silueta es identificar cuál es el número óptimo de agrupamientos.

Figura 4-7 coeficiente de silueta para el resultado de agrupación proporcionado:

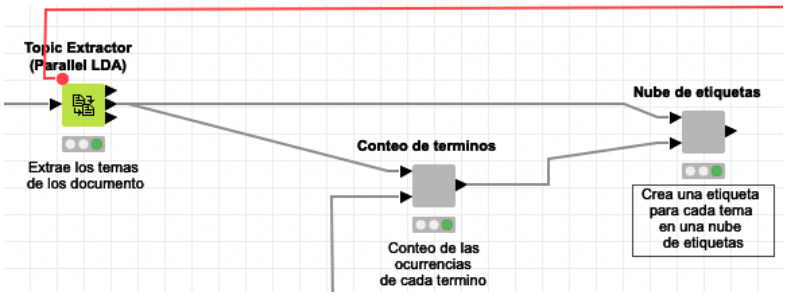
Figura 4-8 Coeficiente silueta



4.5 Extracción de temas

En la Figura 4-9 se observa el flujo que permite obtener el número óptimo de temas. Una vez que se ha determinado el posible número óptimo de temas para las historias clínicas, se puede ejecutar el nodo *Topic Extractor*. El nodo asigna un tema a cada documento y genera palabras clave para cada tema (se puede especificar cuántas palabras clave deben generarse para cada tema en el cuadro de diálogo del nodo), para este flujo ese valor se pasa por parámetro y en la figura 4-9 se ve representado por la línea roja que llega al nodo *Topic Extractor*, por lo tanto, el número de temas a extraer en este flujo se proporciona mediante una variable de flujo.

Figura 4-9 Extracción de temas:



4.6 Anotación del texto

En lo que respecta a la siguiente etapa de análisis, se realizó un proceso de anotación del texto que consistió en una tarea de extracción de información que busca localizar y clasificar el texto en categorías como, ubicación, organizaciones, expresiones químicas y biomédicas de las entidades nombradas encontradas en el corpus, este proceso se conoce como reconocimiento de entidades nombradas (NER, por su sigla en inglés), (Cely, 2018).

Para la creación del corpus anotado se utiliza como fuente de datos del dominio biomédico la ontología UMLS. A continuación, se explica la manera de disponer de la base de términos de UMLS y como se realiza la anotación de los términos.

4.7 Construcción del modelo de extracción de variables relevantes.

4.7.1 UMLS Metathesaurus

De acuerdo con el manual UMLS, “Metathesaurus es una base de datos de vocabulario muy grande, multiusos y multilingüe que contiene información sobre conceptos biomédicos y relacionados con la salud, sus diversos nombres y las relaciones entre ellos. Diseñado para ser utilizado por desarrolladores de sistemas. Metathesaurus se construye a partir de versiones electrónicas de varios tesauros, clasificaciones, conjuntos de códigos y listas de términos controlados utilizados en la atención al paciente, facturación de servicios de salud, estadísticas de salud pública, indexación y catalogación de literatura biomédica o investigación básica, clínica y de servicios de salud”, (UMLS, 2009).

El Metathesaurus hace parte de un conjunto de herramientas construidas y publicadas por la Biblioteca Nacional de Medicina de los Estados Unidos (NLM, por su sigla en inglés) como PubMed, MedLINE Plus, MeSH, etc; que apoyan la digitalización, normalización y publicación de la información del dominio biomédico. Para el uso del Metathesaurus y demás herramientas de UMLS fue necesario solicitar la licencia de uso; en este caso, para uso únicamente académico, la cual se solicita directamente en el sitio web de UMLS.

4.7.2 Red Semántica de UMLS

El propósito de la Red Semántica es proporcionar una categorización consistente de todos los conceptos representados en el MetamorphoSys UMLS y proporcionar un conjunto de relaciones útiles entre estos conceptos. Toda la información sobre conceptos específicos se encuentra en el Metathesaurus, UMLS, 2009. La red semántica ha definido 134 categorías o tipos semánticos que permiten identificar los conceptos; estos abarcan diferentes tipos: enfermedades, medicamentos, función biológica, etc., los cuales se establecen en un conjunto de categorías jerárquicas.

El proceso de creación del corpus anotado está determinado por los siguientes pasos:

1. Obtención del fragmento del corpus médico**:** de las sentencias contenidas en el corpus UFAL Medical Corpus en el idioma español se extraen las frases que corresponden al dominio biomédico. Este proceso de filtrado se realiza identificando en los metadatos el valor *medical corpus*.
2. Generación de tokens**:** una vez obtenido el corpus del dominio biomédico agrupado, se realiza la generación de los tokens, usando la librería de Trabajo práctico en procesamiento del lenguaje natural (NLTK, por su sigla en inglés) .
3. Buscador del término médico en UMLS: se realiza la búsqueda de los términos médicos y se aplica al listado de tokens. Si el registro se encuentra en Metathesaurus, entonces lo anota con la etiqueta encontrada asociada al tipo semántico que corresponde.
4. Generación de archivo con corpus anotado: se finaliza el proceso de anotación salvando la información del corpus anotado en un archivo de texto.

5. El resultado final del proceso de anotación es el archivo que contiene el corpus anotado con la tripleta definida por término, el identificador único de concepto (CUI) y etiqueta. Con el fin de ilustrar, a continuación en la figura 4-10 se muestra un fragmento de una historia clínica.

Figura 4-10 Fragmento de corpus anotado, (Cely, 2018):

```
:artritis C0003873 B-T047
reumatoide C0003873 I-T047
o C0439114 B-T170
lumbago C0949075 B-T184
crónico C0949075 I-T184
: O O

Además O O
, O O
la O O
actividad C0441655 B-T052
farmacodinámica O O
de C0332285 B-T082
clopidogrel C0070166 B-T121
no C1298908 B-T033
se C0373721 B-T059
vio O O
significativamente O O
influenciada O O
por C0678226 B-T169
la O O
administración C0001554 B-T057
concomitante C0205420 B-T079
de C0332285 B-T082
fenobarbital C0031412 B-T121
:
```

4.8 Búsqueda de términos relevantes

Las historias clínicas contienen un amplio acumulado de términos que caracterizan en conjunto información relacionada con tratamientos, antecedentes, hallazgos, evolución de la enfermedad y del paciente, efectos adversos y otra información de interés. En esta investigación, el corpus analizado está compuesto por historias clínicas de pacientes con diagnóstico de artritis reumatoide, por lo que encontrar términos como artritis o reumatoide sería evidente, por lo tanto, se

plantea la búsqueda de términos relevantes y no obvios dentro del corpus analizado. Para lograr hacer evidentes dichos términos se aplica el método de frecuencias TF-IDF; este método permite dar relevancia a aquellas palabras que aparecen pocas veces en las historias clínicas, pero que son persistentes o constantes dentro de todo el conjunto de estas.
Para llevar a cabo esta actividad, se ejecutaron las siguientes tareas que se agruparon en un metanodo nombrado Frecuencias.

El nodo *Term Frequence* (TF) permite calcular las frecuencias relativas de aparición de una palabra, para poder ejecutar esta tarea previamente se debe crear una bolsa de palabras (BoW) que es una estructura que contiene una tabla compuesta por términos y documentos; los términos obtenidos en este caso están asociados a etiquetas que indican la categoría gramatical de las palabras por lo que se decide hacer una transformación previa antes de ejecutar el nodo TF; esta transformación consiste en convertir los términos a cadenas, es decir, que se obtienen las palabras sin las categorías gramaticales asociadas. TF calcula la frecuencia de cada término de acuerdo con cada documento y agrega una columna que contiene el valor TF. El valor se calcula dividiendo la frecuencia absoluta de un término de acuerdo con un documento por el número de todos los términos de este documento. La configuración que se realizó en este nodo fue establecer en la pestaña "*Document col*", el conjunto de documentos a analizar, es decir, las historias clínicas.
En la pestaña TF options se marcó la opción *Relative frequency*; esta opción permitió calcular la frecuencia relativa de los términos. Los valores calculados para cada término representativo son normalizados por la frecuencia máxima en el documento según:

$$TF_{i,j} = \frac{f_{i,j}}{f_{i,j}} \quad (4.1)$$

Inverse Document Frequence (IDF), este nodo recibe en su interfaz de entrada al nodo TF. La tarea que se ejecutó con este nodo fue la de calcular la frecuencia inversa del documento para cada término; esto permite identificar o evidenciar qué tan importante es una palabra contenida en una historia clínica dentro una colección de documentos (historias clínicas). La configuración realizada en este nodo fue igual a la realizada en el nodo TF en la pestaña Document col. En la pestaña IDF options se establece para el parámetro Document column el valor *smooth* que es el valor por defecto.

$$IDF_{t_i} = log_{10}(\frac{Q}{n_i+1}) \quad (4.2)$$

TF-IDF, para ejecutar esta tarea se utilizó el nodo Java Snippet que permitió asignar a una variable TF-IDF el valor de la operación matemática del producto de la frecuencia relativa de los términos por la frecuencia inversa de los documentos. Posteriormente, se utilizó el nodo Sorter para organizar de forma ascendente los resultados de la columna TF-IDF.

$$w(t_i, d_j) = Tf_{t,d}\, x\, IDF_t \quad (4.3)$$

4.9 Análisis de contexto

El análisis de las palabras de forma individual dentro de un documento o conjunto de documentos es importante como se ha mostrado hasta este punto, no obstante, se encontrará con frecuencia que analizar las palabras en conjunto

tiene más sentido que cuando se analizan individualmente. En tal sentido, dentro del modelo propuesto se realizó el análisis a grupos de palabras midiendo las frecuencias con las que dos o más de estas ocurren juntas, de esta forma se puede capturar la relación entre las mismas. Esta actividad está enmarcada dentro de la estadística de concurrencia y puede calcularse en diferentes niveles: a nivel de documento, a nivel de oración, a nivel de párrafo, etc.

La ventana de palabras clave es una técnica de procesamiento de texto que extrae el texto que rodea a una palabra clave para proporcionar el contexto sobre cómo se usa esa palabra clave. En este modelo computacional se plantea la extracción de dicha ventana teniendo en cuenta las palabras claves encontradas, sin embargo se hace una búsqueda de palabras como diagnóstico, antecedentes, tratamiento, eventos, enfermedad.

Utilizando diferentes nodos de Knime se hizo la extracción del texto que rodea esas palabras clave, así también se ha utilizado la técnica de expresiones regulares y se ha utilizado AntConc para la ejecución de dicha actividad.

El objetivo es identificar a los pacientes que experimentaron alguna condición particular mientras tomaban un medicamento administrado en el tratamiento. En cada historia clínica se busca una o más palabras clave. Como se trata de identificar los efectos secundarios con los medicamentos, las palabras clave son los nombres de esos medicamentos. En cada lugar de la historia clínica en donde se encuentra la palabra clave, se extrae el texto alrededor de la palabra clave, esto se llama la ventana.

Los tamaños de las ventanas pueden variar, a veces la ventana será más pequeña, por ejemplo, cinco palabras antes y cinco palabras después de la palabra clave, o puede ser mas larga, quizas dos oraciones antes y despues de la palabra clave.

Una vez extraída la ventana de texto se comienzan a buscar otras palabras y frases que indiquen la condición que se está buscando. Este proceso sería un proceso de inclusión, si se tiene alguno de los términos buscados, se deben incluir en un grupo que se considere como casos, es decir aquellos que presenten la condición objetivo. A continuación se desarrolla una lista de palabras clave y frases basada en el conocimiento del experto clínico y en la revisión manual de las ventanas de texto.

4.10 Palabras Clave

En general, la taxonomía de los métodos actualmente disponibles para realizar la extracción automática de palabras clave se puede representar de la siguiente manera: asignación de palabras clave, extracción de palabras clave.
En la asignación de palabras clave, las palabras clave se eligen de un vocabulario controlado de términos o de una taxonomía predefinida, para este caso los registros médicos de los paciente. Luego, los documentos se clasifican en clases según su contenido de palabras. Las palabras clave no necesitan mencionarse explícitamente en el texto. Las palabras relacionadas deben ser suficientes.

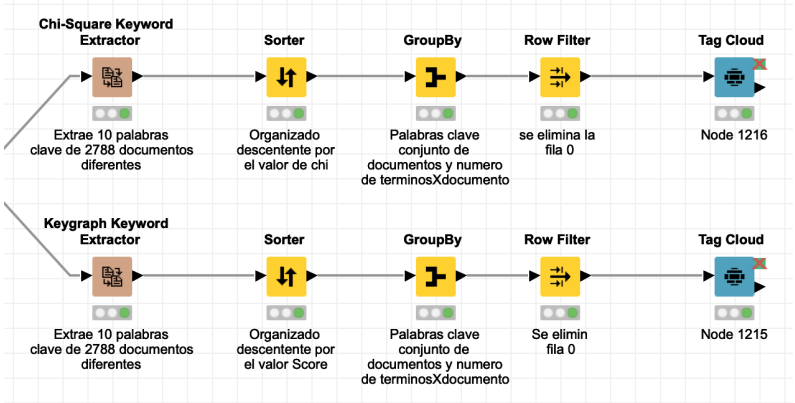
En la extracción de palabras clave, las palabras clave se detectan en el documento y forman el vocabulario de la colección. Los enfoques supervisados utilizan un clasificador para detectar palabras candidatas. Estos métodos supervisados requieren una colección etiquetada de documentos, que no es fácil de tener. Cuando las etiquetas no están disponibles para los documentos, los métodos no supervisados son útiles.

Las etapas del algoritmo, que se ilustra en la figura 4-11, se realizaron para la extracción de palabras clave y en este modelo se resumen en tres fases principales:

1. Palabras candidatas: se extraen todas las palabras posibles que pueden ser palabras clave
2. Extracción de características: para cada palabra candidata se calculan las características para medir que es una palabra clave.
3. Evaluación: se verifica la pertinencia de las características propuestas.

En el modelo propuesto, se utilizaron dos nodos provistos por Knime y que permite hacer la asignación y extracción automática de palabras clave, como se aprecia en la figura 4-11, Chi-Square Keyword Extractor y Keygraph Keyword Extractor. Chi-Square Keyword Extractor utiliza una medida chi-cuadrado para clasificar la relevancia de una palabra para un texto dado. Keygraph Keyword Extractor utiliza una representación gráfica de la colección de documentos para encontrar las palabras claves del mismo. El algoritmo Chi-Square Keyword Extractor detecta las palabras clave relevantes utilizando una medida estadística de Co-Ocurrencia de la palabra en un solo documento.

Figura 4-11 Palabras clave: 1



Para este análisis de asignación y extracción de palabras claves se decidió extraer 10 palabras clave a partir del corpus analizado, es decir 2788 historias clínicas; posteriormente, se ordenan las palabras en orden descendente y se realiza una agrupación en donde se obtienen los conjuntos de documentos y el número de palabras por documento.

4.11 Palabras asociadas al contexto médico

“Una Terminología permite capturar el significado del suceso clínico porque está referenciado a un estándar. Hay muchas terminologías, en las que podríamos agrupar taxonomías, ontologías, tesauros, vocabularios controlados, que dan abrigo a diferentes dominios como la actividad hospitalaria, la historia clínica electrónica, laboratorio, genética, etc. (por ejemplo, ICD10, conjunto de identificadores, nombres y códigos (LOINC, por su sigla en inglés) para laboratorio o Nomenclatura sistematizada de la medicina - términos clínicos (SNOMED-CT, por su sigla en inglés)).”

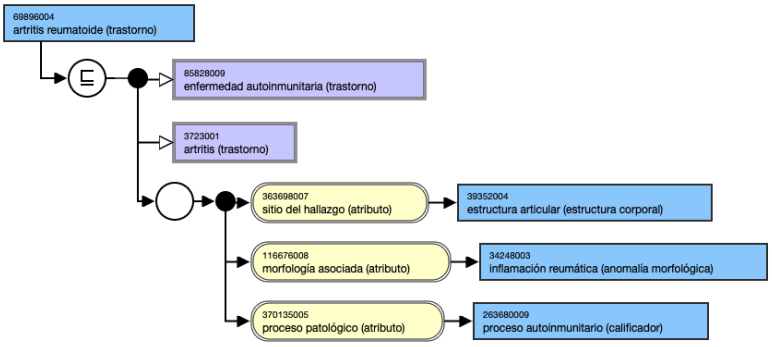
En el análisis previo, se realizaron un conjunto de tareas y actividades que permitieron identificar un agregado de términos relevantes dentro de un corpus representado por historias clínicas, estos términos constituyen entidades que dentro del contexto médico cobran gran importancia por lo que el objetivo es facilitar el reconocimiento exacto de la información médica y otros temas relacionados con la salud y la interoperabilidad semántica de los registros clínicos. Para alcanzar ese objetivo se hace uso de la terminología médica SNOMED CT en la versión 2019-04-31 para el idioma español, esta terminología proporciona una forma estandarizada de representar frases registradas por el médico y permite su interpretación automática.

Para poder realizar esta conexión entre los términos y SNOMED se utilizó Snowstorm, que es un servidor de terminología de SNOMED CT construido sobre Elasticsearch. Para la implementación de la conexión se utilizó lenguaje Python en la versión 3.9, se definieron la URL, la edición y la versión de SNOMED como variables y se definen las funciones:

- getConceptById: imprime el Nombre completamente especificado (FSN, por sus siglas en inglés)
- getDescriptionById: imprime la descripción de acuerdo con el Id.
- getConceptsByString: imprime un número de conceptos con descripciones que contienen el término de búsqueda.
- getDescriptionsByStringFromProcedure: imprime un número de descripciones que contienen el término de búsqueda con una etiqueta semántica específica.

que permiten recuperar la información asociada a: concepto, descripción, término, procedimientos; la información se recupera en formato JSON.

Figura 4-12 ejemplo estructura[1] Snomed:



[1] https://browser.ihtsdotools.org/

4.12 Medicamentos asociados a la artritis reumatoide

Los medicamentos hacen parte fundamental del tratamiento de la artritis reumatoide y, en general, están enfocados en aliviar los síntomas y mejorar la función de las articulaciones. Es posible que el especialista deba probar diferentes tratamientos y combinaciones farmacológicas antes de encontrar el adecuado para un paciente particular, de ahí la relevancia en el estudio y reconocimiento de estos procedimientos.

Existe una variedad de medicamentos para tratar la artritis y varían según el tipo de artritis. Los medicamentos más frecuentes para tratar esta enfermedad son los siguientes[2]:

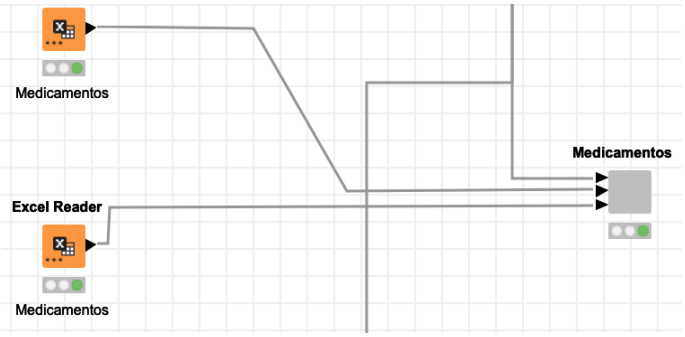
- Analgésicos: ayudan a reducir el dolor, pero no tienen ningún efecto sobre la inflamación.
- Antiinflamatorios no esteroides: reducen tanto el dolor como la inflamación. Estos medicamentos pueden provocar efectos adversos importantes como ataque cardíaco o accidente cerebrovascular.
- Contrairritantes: generalmente son ungüentos o cremas y cuyo componente principal son los pimientos.
- Medicamentos antirreumáticos modificadores de la enfermedad: estos medicamentos retardan o detienen el ataque del sistema inmunitario a las articulaciones.
- Modificadores de la respuesta biológica: son medicamentos manipulados genéticamente que captan varias moléculas de proteína que afectan la respuesta inmunitaria. Otros medicamentos se dirigen a otras sustancias que tienen un papel en la inflamación, como la interleucina-1 (IL-1), la

[2] https://www.mayoclinic.org/es-es/diseases-conditions/arthritis/diagnosis-treatment/drc-20350777

interleucina-6 (IL-6), las enzimas janocinasas y ciertos tipos de glóbulos blancos conocidos como linfocitos B y T.

- Corticosteroides: estos medicamentos reducen la inflación y suprimen el sistema inmunitario.

Figura 4-13 diccionarios para identificar medicamentos



En la figura 4-13 se muestra el flujo de trabajo realizado en este apartado, que consistió en definir diccionarios que permitieran hacer un reconocimiento de los medicamentos registrados en las historias clínicas teniendo en cuenta dos aspectos principales: el nombre del medicamento y la indicación médica principal. Por lo tanto, a partir del flujo creado se obtienen las frecuencias de aparición de los fármacos haciendo evidente el tratamiento más frecuente y, por otro lado, se puede establecer las enfermedades para las cuales se indican estos fármacos.

Al interior del meta nodo Medicamentos que se ve en la figura 4-13, se desarrolló un flujo que permite encontrar dentro de las narrativas de las historias clínicas los nombres de los fármacos y las enfermedades asociada a esos tratamientos farmacológicos, como se puede observar en la figura 4-14.

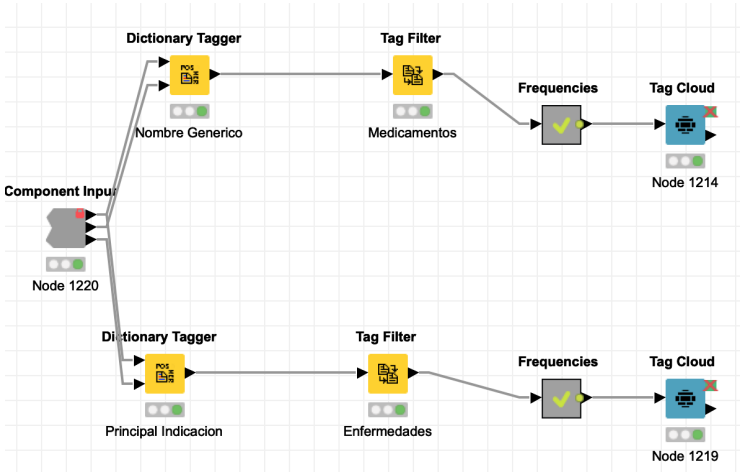
En la tabla 4-1 se puede observar un segmento del diccionario utilizado para la identificación de los tratamientos prescritos a los pacientes diagnosticados con la

enfermedad de (AR), en esta tabla se pueden observar datos asociados a: grupo terapéutico, nombre genérico, indicación principal, otras indicaciones y la descripción completa del medicamento.

Tabla 4-1 Diccionario de medicamentos:

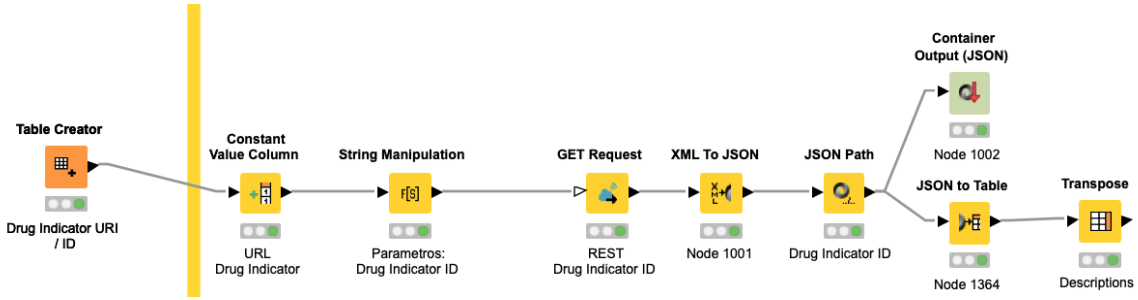
NOMBRE DE GRUPO TERAPEUTICO	NOMBRE GENERICO	PRINCIPAL INDICACION	DEMAS INDICACIONES	DESCRIPCION COMPLETA
ANALGESIA	ACIDO ACETILSALICILICO	1. ARTRITIS REUMATOIDE	2. OSTEOARTRITIS 3. ESPONDILITIS ANQUILOSANTE 4. FIEBRE REUMATICA AGUDA 5. DOLOR O FIEBRE	ACIDO ACETILSALICILICO TABLETA CADA TABLETA CONTIENE: ACIDO ACETILSALICILICO 500 MG. ENVASE CON 20 TABLETAS.
ANALGESIA	ACIDO ACETILSALICILICO	1. ARTRITIS REUMATOIDE	2. OSTEOARTRITIS 3. ESPONDILITIS ANQUILOSANTE 4. FIEBRE REUMATICA AGUDA 5. DOLOR O FIEBRE	ACIDO ACETILSALICILICO TABLETA SOLUBLE O EFERVESCENTE CADA TABLETA SOLUBLE O EFERVESCENTE CONTIENE: ACIDO ACETILSALICILICO 300 MG. ENVASE CON 20 TABLETAS SOLUBLES O EFERVESCENTES
ANALGESIA	CAPSAICINA	DOLOR DE LEVE A MODERADA INTENSIDAD EN: ARTRITIS REUMATOIDE, ARTROSIS, NEURALGIA POST-HERPETICA, NEUROPATIA DIABETICA, MIEMBRO FANTASMA.		CAPSAICINA CREMA CADA 100 GRAMOS CONTIENE: EXTRACTO DE OLEORESINA DEL CAPSICUM ANNUUNA EQUIVALENTE A 0.035 G DE CAPSAICINA. ENVASE CON 40 G.
ANALGESIA	ETOFENAMATO	1. ARTRITIS REUMATOIDE	2. ESPONDILITIS ANQUILOSANTE 3. OSTEOARTROSIS Y ESPONDILOARTROSIS 4. HOMBRO DOLOROSO 5. LUMBAGO 6. CIATICA 7.TORTICOLIS 8. TENOSINOVITIS 9. BURSITIS 10. ATAQUE AGUDO DE GOTA	ETOFENAMATO SOLUCION INYECTABLE CADA AMPOLLETA CONTIENE: ETOFENAMATO 1 G. ENVASE CON UNA AMPOLLETA DE 2 ML

Figura 4-14 flujo para identificar tratamientos farmacológicos



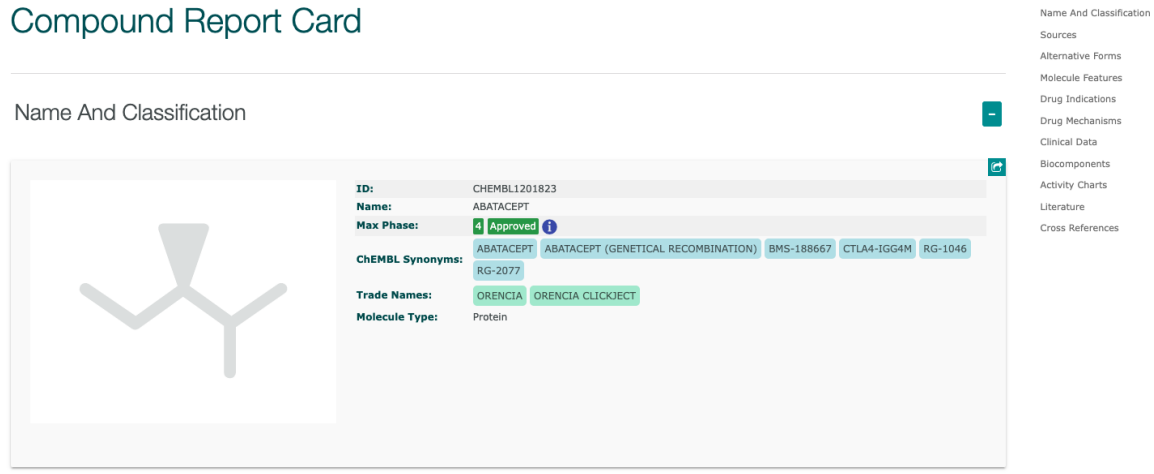
Una vez se han logrado identificar los fármacos registrados, se procede a hacer una búsqueda de los tratamientos fármaco genéticos que son Modificadores de la respuesta biológica puesto que estos son medicamentos manipulados genéticamente que captan varias moléculas de proteína que afectan la respuesta inmunitaria, para lograr este objetivo se hace una consulta a la base de datos de ChEMBL como se ilustra en la figura 4-15, en esta se muestra como se realiza la conexión a la base de datos de moléculas bioactivas con propiedades similares a las de los fármacos. Esta base de datos reúne datos químicos, de bioactividad y genómicos para ayudar a traducir la información genómica en fármacos eficaces. A la base de datos se accede mediante una conexión API REST, la cual devuelve la información en formato JSON.

Figura 4-15 flujo para extraer información fármaco genética de ChEMBL:



En dicha conexión se establece el endpoint o punto de conexión y se establece como parámetro el código del fármaco. En la imagen 4-16 se puede observar la información recuperada para el fármaco ABATACEPT, la cual incluye un identificador, nombre, la fase en la que se encuentra, sinónimos con los que puede identificar en los estudios de tratamientos, entre otros.

Figura 4-16 Información para el fármaco ABATACEPT identificado con el código 22607[3]



[3] https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL1201823/

ChEMBL es una base de datos de pequeñas moléculas similares a fármacos bioactivos que contiene datos de química medicinal, bioinformática y bioensayos integrados a partir de una amplia variedad de fuentes. Los datos se extraen manualmente del texto completo de publicaciones científicas revisadas por pares en una variedad de revistas, como Journal of Medicinal Chemistry, Bioorganic Medicinal Chemistry Letters y Journal of Natural Products.

La versión 11 de la base de datos ChEMBL contiene información extraída de más de 42.500 publicaciones, junto con varios conjuntos de datos depositados y datos extraídos de otras bases de datos. En total, hay más de 1 millón de estructuras compuestas distintas representadas en la base de datos, con 5,4 millones de valores de actividad de más de 580.000 ensayos. Estos ensayos se asignan a 8.200 dianas, incluidas 5.200 proteínas, (de las cuales 2.388 son humanas).

5. Resultados

5.1 Análisis del desarrollo de la enfermedad a partir de la minería de textos

El cumplimiento del objetivo para establecer el desarrollo de la enfermedad (AR) de un paciente mediante el análisis de la historia clínica con métodos de minería de textos se logró, mediante el reconocimiento de las palabras clave contenidas en el conjunto de documentos, ya que se identificaron términos asociados a los criterios de clasificación de la enfermedad AR que fueron actualizados en el 2010, estos términos permiten identificar desenlaces para el diagnóstico o desenlaces de la enfermedad. Los especialistas determinaron como regla de oro que un paciente particular sería diagnosticado de AR si iniciaba tratamiento con metotrexato (MTX) en los 12 meses siguientes tras la primera visita al reumatólogo, (Aletaha et al., 2010).

Esta regla de oro se cumple para las historias clínicas analizadas en esta investigación teniendo en cuenta que la evolución de la AR para los pacientes allí diagnosticados tiene en promedio más de un año y tienen indicado MTX como tratamiento.

Otros criterios para tener en cuenta son la valoración de los marcadores serológicos de la AR y de los reactantes de fase aguda. En cuanto a los

marcadores serológicos, además de valorar la presencia del factor reumatoide (FR) cuyo término se encuentra en el análisis de frecuencias (ver tabla 5-5), valora la presencia de los anticuerpos contra péptidos citrulinados (ACPA). También, valora los reactantes de fase aguda, velocidad de sedimentación globular y proteína C reactiva (PCR) como parte de los criterios de clasificación (MedlinePlus, 2021b).

Dentro de los resultados logrados en esta investigación se obtuvieron los términos PCR y VSG asociados a dichos criterios de clasificación, mediante la aplicación de un modelo de 3-gramas que permitió tener indicios de la relación entre VSG y PCR, adicionalmente, una relación con las proteínas AST y ALT; esta relación es importante debido a que el examen de PCR a menudo se realiza junto con el examen de velocidad de sedimentación globular (VSG) o de eritrosedimentación que sirve para detectar alguna inflamación. La proteína C reactiva (PCR) es producida por el hígado y cuando el nivel de PCR se eleva indica que hay inflamación en todo el cuerpo, así mismo se encuentra que la AST (aspartato aminotransferasa) es una enzima que se encuentra principalmente en el hígado y también en los músculos. Cuando el hígado está dañado, libera AST en el torrente sanguíneo. Otro indicio del desarrollo de la enfermedad fue la relación encontrada mediante el modelo de 3-gramas de los términos PCR, AST y ALT; la ALT es una enzima que se encuentra en el hígado y que ayuda a convertir las proteínas en energía para las células hepáticas. Cuando el hígado está dañado, se libera ALT al torrente sanguíneo y aumentan sus niveles (García Martín & Zurita Molina, 1998).

Así mismo, a partir de los análisis de frecuencias de palabras se puede observar que el tratamiento con la mayor frecuencia de ocurrencia en la colección de documentos es el MTX y el Rituximab tiene la mayor Frecuencia Inversa del

Documento para este término. Es importante resaltar este hecho dado que la inyección de rituximab (Rituxan) se usa con metotrexato y otros FARME para tratar los síntomas de la artritis reumatoide grave lo que es un indicio del avance de la enfermedad, (Borsari et al., 2020; MedlinePlus, 2021a).

El hecho de definir como AR a los pacientes que requirieron tratamiento con MTX en los 12 meses posteriores a la primera visita, excluyendo otros FAME como salazopirina (SZP), leflunomida (LFM) e hidroxicloroquina (HCQ), lo avalan los investigadores al tratar de evitar incluir a los pacientes con espondiloartropatías (que probablemente habrían recibido SZP) y pacientes con enfermedades del tejido conectivo, que probablemente habrían recibido HCQ (Aletaha et al., 2010).

5.2 Sobre la información genética relacionada con AR en el corpus analizado.

El cumplimiento del objetivo relacionado con asociar información clínica del paciente a la información genética relacionada con AR que se encuentre disponible en bases de datos biológicas no se logra cubrir con base en la información registrada en las historias clínicas debido a que no se halló información genética relacionada en las mismas; sin embargo, vale resaltar que existe una compleja interacción de factores genéticos y ambientales que podría llevar al desarrollo de la enfermedad (Aletaha et al., 2010).

Así, en los individuos genéticamente predispuestos, los factores ambientales como la obesidad, dieta, fumar, microbiota gastrointestinal e infecciones; que han sido términos que se identificaron en el análisis del corpus, ver tabla 5-4, podrían desencadenar la activación aberrante de la respuesta inmune innata y adaptativa, causando la pérdida de tolerancia inmunológica, presentación de autoantígenos

con activación de células T y B, y producción de citoquinas inflamatorias (Tan & Smolen, 2016; Testa et al., 2021). Esta cascada de eventos es lo que eventualmente lleva a la sinovitis, destrucción de cartílago y hueso y otros síntomas extraarticulares característicos de la enfermedad (Mease et al., 2021; Testa et al., 2021).

Con respecto a la información genética, se debe destacar el *TNF* (que codifica para el factor de necrosis tumoral alfa [TNFa]) por dos razones principales. En primer lugar, por la evidencia de que el TNFa es una citocina pro-inflamatoria crítica en la fisiopatología de la artritis reumatoide: se produce en el tejido sinovial de pacientes con artritis reumatoide por macrófagos y células T, existen niveles elevados de la citocina en el suero y líquido sinovial, y los anticuerpos anti-TNFa tienen efecto terapéutico benéfico sobre la enfermedad (Páez Leal et al., 2010). Segundo, el *TNF* es un gen altamente polimórfico, con cinco micro-satélites y numerosos SNP en su promotor, algunos de los cuales podrían regular la expresión génica (Hajeer & Hutchinson, 2001).

El TNF-α es probablemente la citocina multifuncional más importante en la AR. Esta proteína es producida por el gen TNF-α, el cual se localiza en la banda citogenética 6p21, región ligada a diversas enfermedades articulares. Esta citocina regula diversos efectos biológicos, entre los que se incluyen los siguientes: expresión de diversos genes, como IL-1, IL-6, metaloproteasas y moléculas de adhesión, proliferación, regulación de la apoptosis, activación celular e inducción de anticuerpos, que se asocian con la inflamación, la destrucción del cartílago y la erosión del hueso de los individuos con AR (Lander & Kruglyak, 1995).

A través del análisis de las historias clínicas usando minería de texto se identificaron fármacos biológicos anti-TNF-α como: abatacept, adalimumab, rituximab, tocilizumab, belimumab, infliximab, ver tabla 5-1. El **rituximab (*Rituxan*)**, es para personas con AR moderada a severa que no responden adecuadamente a la terapia con uno o más antagonistas del FNT (factor de necrosis tumoral). Ataca selectivamente los linfocitos B, involucrados en la inflamación. Ha mostrado buenos resultados en combinación con metotrexato. El **abatacept (*Orencia*)** bloquea la activación de linfocitos T en personas con AR. El **adalimumab (*Humira*)** se indica para reducir la sintomatología e inhibir el progreso del dãno articular en adultos con AR moderado o severa que tienen una respuesta inadecuada a otros medicamentos (Arthritis Foundation, n.d.).

La medicina biológica actúa sobre el sistema inmunológico y detiene la inflamación provocada por el propio sistema en las articulaciones, así como los efectos de la enfermedad a nivel general. Con estos tratamientos se consigue una mejoría de síntomas como el dolor y la rigidez; a partir del análisis de las historias clínicas usando minería de textos, fue posible identificar y anotar los términos dolor y rigidez con la etiqueta B-T184 que clasifica los términos en la categoría de signos y síntomas, ver Tabla 5-4. Vale mencionar que estos biológicos ralentizan la progresión de la enfermedad y así previenen daños a largo plazo como la deformidad de las extremidades afectadas (UNAL, 2021).

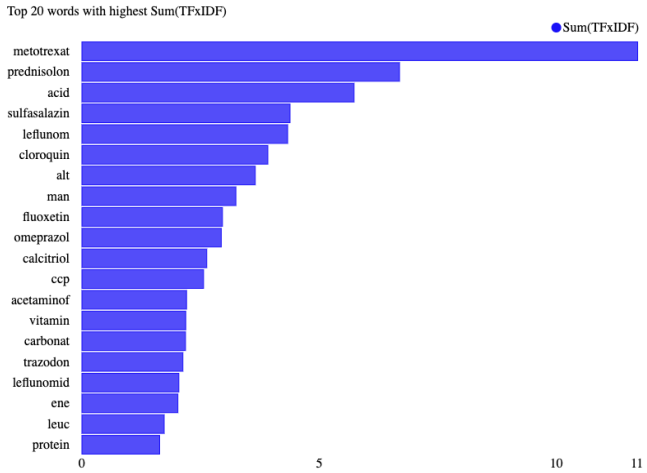
5.3 Términos relevantes a partir de una bolsa de palabras

Se obtuvo como resultado la suma de las frecuencias relativas, las cuales se representan en la figura 5-1 y que muestra los 20 primeros términos más frecuentes a partir de una bolsa de palabras que contiene todas los términos que conforman las historias clínicas, la suma de frecuencia inversa del documento,

que se ilustra en la figura 5-2 y en la cual se pueden observar los 20 primeros términos más relevantes dentro del corpus de acuerdo a la frecuencia inversa de documentos.

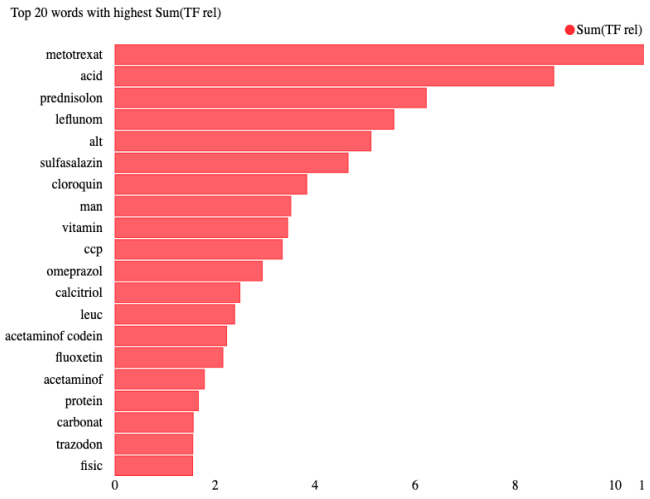
El objetivo fue poder verificar las 20 primeras palabras con mayor relevancia dentro del corpus, se pudo evidenciar que los términos con más relevancia hacen referencia a fármacos, es decir, a los tratamientos administrados a los pacientes con artritis reumatoidea siendo el metotrexato el de mayor relevancia tanto en la suma de la frecuencia relativa como en la suma de la frecuencia inversa del documento, por lo que se puede inferir su importancia en el tratamiento de la enfermedad. El metotrexato forma parte de un grupo de medicamentos llamados fármacos antirreumáticos modificadores de la enfermedad (FAME) y es el tratamiento más común para la artritis reumatoide. El metotrexato ayuda a prevenir los daños permanentes adicionales que pueden producirse si no se trata la artritis reumatoide[4].

Figura 5-1 suma de frecuencias relativas



[4] https://www.cochrane.org/es/CD000957/MUSKEL_metotrexato-para-el-tratamiento-de-la-artritis-reumatoide

Figura 5-2 suma de frecuencias inversas



5.4 Tratamientos relevantes e indicaciones principales

Es importante resaltar que el análisis de medicamentos una vez asociado el diccionario para identificar, no solo la frecuencia de aparición de estos fármacos, sino los nombres genéricos y la indicación principal arroja los siguientes resultados:

El tratamiento de Rituximab tiene la mayor frecuencia absoluta dentro del conjunto de historias clínicas seguido del tratamiento con Abatacept según se muestra en la tabla 5-1. Esta tabla contiene los primeros 20 resultados de las frecuencias de aparición de términos por documento y se observa que el fármaco Rituximab tiene una frecuencia absoluta de aparición por documentos de 15 siendo esta la mayor frecuencia.

Tabla 5-1 tabla frecuencias absolutas tratamientos:

Termino	IDF	TF rel	TF abs
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.5	15
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.517	15
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.631	12
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.423	11
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.4	10
rituximab[NCS(ANCORA) NP(ANCORA) DRUG(PHARMA)]	24.440	0.266	8
abatacept[NP(ANCORA) DRUG(PHARMA)]	17.512	0.571	8
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.75	6
omeprazol[NP(ANCORA) DRUG(PHARMA)]	0.5221	0.545	6
rituximab[NCS(ANCORA) DRUG(PHARMA)]	11.153	0.6	6
omeprazol[NP(ANCORA) DRUG(PHARMA)]	0.5221	1.0	6
		Hace parte de la tabla 5-1	
etanercept[NP(ANCORA) DRUG(PHARMA)]	15.021	0.333	6
omeprazol[NP(ANCORA) DRUG(PHARMA)]	0.5221	0.230	6
rituximab[NCS(ANCORA) NP(ANCORA) Z(ANCORA) DRUG(PHARMA)]	24.440	0.206	6

omeprazol[NP(ANCORA) DRUG(PHARMA)]	0.5221	0.24	6
rituximab[NP(ANCORA) NCS(ANCORA) DRUG(PHARMA)]	18.466	0.24	6
tocilizumab[NCS(ANCORA) DRUG(PHARMA)]	16.736	0.25	5
omeprazol[AQ(ANCORA) DRUG(PHARMA)]	10.983	1.0	5
esomeprazol[NP(ANCORA) DRUG(PHARMA)]	0.9500	0.263	5
esomeprazol[NP(ANCORA) DRUG(PHARMA)]	0.9500	0.5	5

Se revisó el estudio en (Lopez-Olivo et al., 2015) sobre los efectos del Rituximab en pacientes con artritis reumatoide. De los ocho estudios realizados en este trabajo que evaluaron a 2,720 pacientes con artritis reumatoide, se puede decir que este tratamiento:

- mejoró el dolor, la función y otros síntomas;
- redujo la actividad de la enfermedad;
- redujo el daño a las articulaciones según lo observado en la radiografía.

Cuando se tiene artritis reumatoide, el sistema inmunológico, que normalmente combate las infecciones, ataca las paredes internas de las articulaciones. Esto causa articulaciones inflamadas, rígidas y dolorosas. Actualmente, no existe cura para la artritis reumatoide, por lo que el tratamiento tiene como objetivo reducir el dolor y mejorar la capacidad de movimiento. El rituximab actúa bloqueando la actividad de los linfocitos B, un tipo de célula inmunitaria que causa inflamación y daño articular en pacientes con artritis reumatoide. El rituximab se administra por

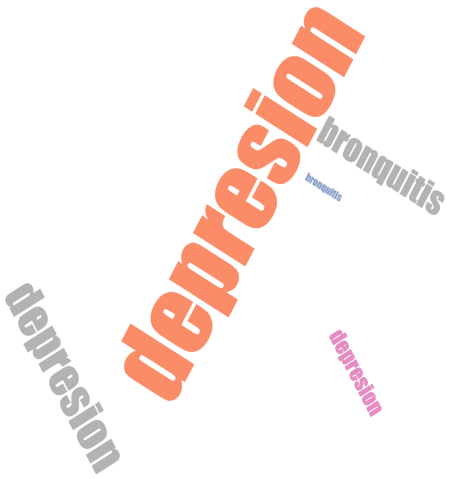
vía intravenosa. El rituximab es de gran interés para los pacientes con artritis reumatoide debido a la mejora de los síntomas y la progresión de la radioterapia, y su baja incidencia de eventos adversos a corto plazo, (Lopez-Olivo et al., 2015). La conclusión de los autores de esta investigación indica que el "tratamiento de Rituximab en combinación con el Metotrexato es significativamente más efectivo que el metotrexato solo, por lo que se evidencia una mejoría en los síntomas de AR y en la prevención de la progresión de la enfermedad".

Con relación al Abatacept, no se cuenta con información precisa acerca de los efectos secundarios y las complicaciones. Este hecho es en particular válido para los efectos secundarios poco frecuentes pero graves. Los posibles efectos secundarios pueden incluir una infección grave o una infección respiratoria alta. Entre las complicaciones poco frecuentes, se pueden incluir ciertos tipos de cáncer.

El Abatacept (Rojahn, 2011) es un medicamento que pertenece a una clase conocida como moduladores de coestimulación selectiva (moduladores inmunes). Actúa bloqueando la actividad de las células T, un tipo de célula inmunitaria que causa inflamación y daño articular en pacientes con artritis reumatoide.

Por otra parte, como se puede observar en la figura 5-2, se identifica que las indicaciones principales encontradas dentro del corpus y una vez asociado al diccionario de medicamentos y su parámetro de indicación principal, arroja como resultado que la frecuencia absoluta más relevante se obtiene para tratamientos de depresión y bronquitis.

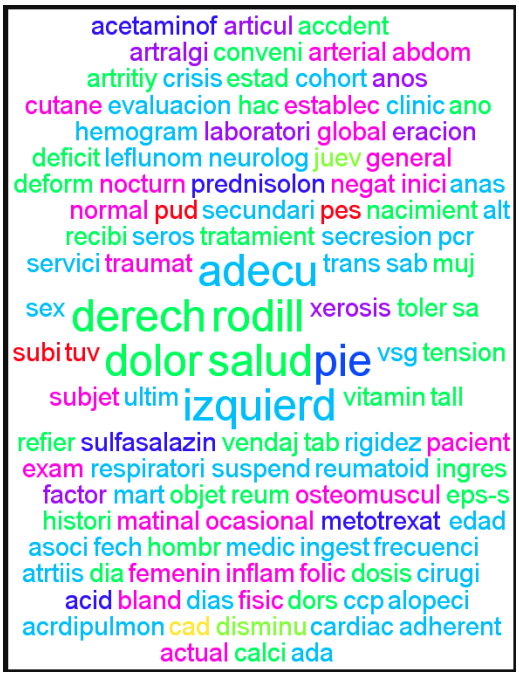
Figura 5-2 Nube de palabras de enfermedades principales tratadas:



5.5 Temas identificados a partir del corpus

Como se puede observar en la figura 5-3, los temas identificados dentro de las historias clínicas muestran términos que se pueden asociar a la enfermedad, como por ejemplo: rodilla derecha, dolor, rigidez, matinal, metotrexato, pie y un conjunto de términos más que fueron identificados automáticamente como temas mediante la aplicación del algoritmo LDA. Este resultado puede ser evidente debido a que el contenido de las historias clínicas hacen referencia a esta enfermedad, sin embargo, vale resaltar que lo importante es que el modelo sería capaz de identificar información relacionada a cualquier enfermedad en la medida que se ingrese la información pertinente.

Figura 5-3 Nube de palabras temas historias clínicas:



5.6 Hallazgos, enfermedades, síntomas, medicamentos

El listado de etiquetas generadas en este proyecto para la anotación del corpus para la identificación de términos médicos está conformado por el conjunto de términos disponibles en la red semántica de UMLS. Para este caso de estudio se han aplicado las 134 etiquetas disponibles en la red semántica para los procesos de anotación del corpus. En la siguiente tabla se muestran las etiquetas más importantes para esta investigación, (Cely, 2018).

Tabla 5-2 Categorías usadas para la anotación del corpus:

ETIQUETA	DESCRIPCIÓN
T047	Disease or Syndrome (enfermedad)
T033	Finding (hallazgos)
T184	Sign or Symptom (signos o sintomas)
T121	Pharmacologic Substance (medicamentos)

A partir de una nube de palabras (ver figura 5-4) se puede evidenciar cuales son algunas de las enfermedades que aparecen con mayor relevancia dentro de las historias clínicas siendo evidentemente la **artritis reumatoide** la que tiene una mayor frecuencia de aparición, sin embargo, se puede observar que la **diabetes**, enfermedades **coronarias**, **depresión**, **hipotiroidismo** y otras con menos frecuencias de aparición también se pueden evidenciar.

Figura 5-4 Nube de palabras enfermedades:

En la tabla 5-3 se registra la frecuencia absoluta de las enfermedades más frecuentes encontradas según las etiquetas asignadas a los términos que hacen parte de las historias clínicas. Se evidencia que el término asociado a artritis es el más frecuente.

Tabla 5-3 Diez términos más frecuentes de enfermedades:

Termino	Etiqueta	Frecuencia
artritis	B-T047	3289
enfermedades	B-T047	280
-artritis	B-T047	192
hipertension	B-T047	106
hipotiroidismo	B-T047	98
artrosis	B-T047	88
diabetes	B-T047	75
osteoporosis	B-T047	66
deshidratación	B-T047	64
dislipidemia	B-T047	57

En la tabla 5-4 se registra la frecuencia absoluta de los signos y síntomas más frecuentes encontradas, según las etiquetas asignadas a los términos que hacen parte de las historias clínicas. Se evidencia que el término asociado a dolor es el más frecuente.

Tabla 5-4 Diez términos más frecuentes de signos y síntomas:

Termino	Etiqueta	Frecuencia
dolor	B-T184	731
rigidez	B-T184	241

		Hace parte de la tabla 5-4
signos	B-T184	176
adherencia	B-T184	170
edema	B-T184	96
ictericia	B-T184	61
dificultad	B-T184	53
sintomas	B-T184	43
dispepsia	B-T184	41
diarrea	B-T184	35

En la tabla 5-5 se registra la frecuencia absoluta de las sustancias farmacológicas más frecuentes encontradas según las etiquetas asignadas a los términos que hacen parte de las historias clínicas. Se evidencia que el término asociado a control es el más frecuente, sin embargo, este término no hace referencia a un fármaco, lo cual nos indica que la anotación no fue precisa para este caso.

Tabla 5-5 Diez términos más frecuentes de sustancias farmacológicas:

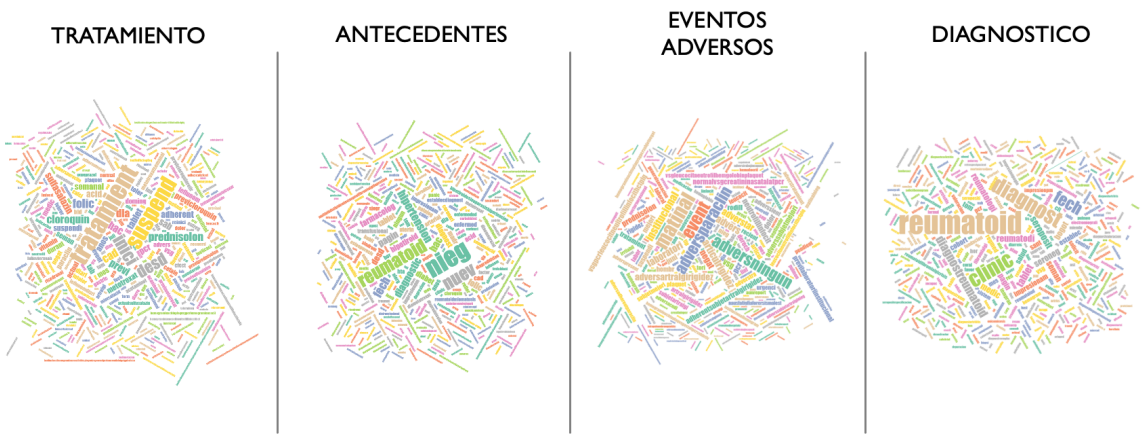
Termino	Etiqueta	Frecuencia
control	B-T121	713
acido	B-T121	333
prednisolona	B-T121	313
factor	B-T121	286
leflunomida	B-T121	224
sulfasalazina	B-T121	174
cloroquina	B-T121	152
acetaminofen	B-T121	145

omeprazol	B-T121	144
		Hace parte de la tabla 5-5
codeina	B-T121	142

5.7 Resultados análisis de antecedentes, diagnóstico, efectos adversos y tratamiento:

El resultado observado en la figura 5-5 hace referencia al análisis de frecuencia de aparición de términos asociados con tres secciones que hacen parte de la historia clínica y que se aislaron para hacer este análisis. Las secciones analizadas fueron Tratamiento, que tiene que ver con los medicamentos prescritos a los pacientes, antecedentes, en los que las narrativas incluyen aspectos familiares, condiciones médicas previas entre otros aspectos, eventos adversos, en donde se registran consecuencias debidas al tratamiento suministrado y diagnóstico, que es la condición observada por el especialista cuando evalúa al paciente.

Figura 5-5 ventanas clinical data science:



Se aplicó la metodología de "*clinical data science"*, que consiste en hacer un análisis por ventanas de textos para analizar las secciones con el objetivo de extraer información relevante; este proceso además se realizó utilizando 3 gramas, los resultados obtenidos se consignan en las tablas 5-6 para la sección de antecedentes, la tabla 5-7 para la sección de diagnóstico, la tabla 5-8 para la sección de eventos adversos y la tabla 5-9 para la sección de tratamiento, en donde se muestran los diez primeros 3-gramas y sus frecuencias de aparición por documento.

- Para la sección de antecedentes se generaron 2175 **3-gramas**
- Para la sección de diagnóstico se generaron 3370 **3-gramas**
- Para la sección de eventos adversos se generaron 740 **3-gramas**
- Para la sección tratamiento se generaron 4526 **3-gramas**

En la tabla 5-6 se tiene el registro de los 10 3-gramas más frecuentes por documento hallados en la sección de antecedentes, en este análisis podemos observar los tres términos relacionados que aparecen con más frecuencia, es importante señalar que el orden de aparición de las palabras no es conmutativo, es decir el término 1 seguido por el término 2 seguido por el término 3 no es igual al orden de aparición término 2 seguido por el término 1 seguido por el término 3, aquí el orden importa.

Tabla 5-6 Diez 3-gramas más frecuentes en antecedentes:

3-gramas	Frecuencias por documento
enfermedad factor hereditari	148
nieg nuev nieg	146

nuev nieg nuev	136
Hace parte de la tabla 5-6	
fech hor impresionpm	108
pagin fech hor	93
tablet tom tablet	76
factor hereditari nieg	65
hereditari nieg farmacolog	61
alerg nieg habit	53
clinic reumatoid salud	52

En la tabla 5-7 se tiene el registro de los 10 3-gramas más frecuentes por documento hallados en la sección de diagnóstico.

Tabla 5-7 Diez 3-gramas más frecuentes en diagnóstico:

3-gramas	Frecuencias por documento
fech hor impresionpm	200
fundacion sant bogot	161
fech impresionpagin fech	154
impres fech impresionpagin	154
impresionpagin fech hor	154
bogot impres fech	153
sant bogot impres	153
regmed fundacion sant	139
reumatoid cohort salud	123
pagin fech hor	115

En la tabla 5-8 se tiene el registro de los 10 3-gramas más frecuentes por documento hallados en la sección de eventos adversos.

Tabla 5-8 Diez 3-gramas más frecuentes en eventos adversos:

3-gramas	Frecuencias por documento
event adversparaclin institucional	8
event adversningun tratamient	4
matinalnieg dolor nocturnnieg	4
nieg rigidez matinalnieg	4
rigidez matinal negat	4
rigidez matinalnieg dolor	4
adversningun tratamient previlaboratoriinstitucional	3
ast alt pcr	3
dolor articul rodill	3
dolor nocturnnieg osteomialginieg	3

En la tabla 5-9 se tiene el registro de los 10 3-gramas más frecuentes por documento hallados en la sección de tratamiento.

Tabla 5-9 Diez 3-gramas más frecuentes en tratamiento:

3-gramas	Frecuencias por documento
tratamient previ cloroquin	37
adherent event advers	28

event advers paraclin	28
inici acid folic	21
	Hace parte de la tabla 5-9
pagin fech hor	18
pcr ast alt	18
acid folic mart	17
advers paraclin institucional	17
vsg pcr ast	17
fech hor impresionpm	16

5.8 Resultado del análisis de términos co-ocurrentes

En muchas ocasiones las palabras adquieren más sentido cuando se analizan de forma conjunta que aislada, por esto la medición de términos co-ocurrentes ayuda a identificar qué tan frecuente es que dos o más términos ocurran juntos y analizar la relación de estas palabras. Para este caso, el orden de aparición de los términos no tiene incidencia, es decir, que la ocurrencia del término 1 seguido por el término 2 se considera equivalente a la ocurrencia del término 2 seguido por el término 1.

- Para el análisis de términos co-ocurrentes para la sección de antecedentes se generó una lista de 45,874 registros.
- Para el análisis de términos co-ocurrentes para la sección de diagnóstico se generó una lista de 56,910 registros.
- Para el análisis de términos co-ocurrentes para la sección de eventos adversos se generó una lista de 14,660 registros.
- Para el análisis de términos co-ocurrentes para la sección de tratamiento se generó una lista de 45,874 registros.

En la tabla 5-10 se muestra la frecuencia de aparición de los diez primeros términos co-ocurrentes para la sección de antecedentes, siendo el conjunto de los términos factor y patología los más frecuentes con 155 apariciones.

Tabla 5-10 Diez términos co-ocurrentes más frecuentes para antecedentes:

Term1	Term2	Frecuencia
factor	patolog	155
factor	hereditari	154
hereditari	patolog	154
enfermedad	patolog	142
enfermedad	factor	141
enfermedad	hereditari	141
tablet	tom	128
hor	impresionpm	97
nieg	nuev	97
hor	tom	85

En la tabla 5-11 se muestra la frecuencia de aparición de los diez primeros términos co-ocurrentes para la sección de diagnóstico, siendo el conjunto de los términos hora e impresionpm los más frecuentes con 190 apariciones.

Tabla 5-11 Diez términos co-ocurrentes más frecuentes para diagnóstico:

Term1	Term2	Frecuencia
hor	impresionpm	190
mes	salud	166
cohort	salud	165
hor	salud	162
fundacion	hor	159
tablet	tom	153
hor	impres	149
hor	impresionpagin	149
impres	impresionpagin	149
fundacion	impres	148

En la tabla 5-12 se muestra la frecuencia de aparición de los diez primeros términos co-ocurrentes para la sección de eventos adversos, siendo el conjunto de los términos dolor y event los más frecuentes con 16 apariciones.

Tabla 5-12 Diez términos co-ocurrentes más frecuentes para eventos adversos:

Term1	Term2	Frecuencia
dolor	event	16
event	institucional	16
event	rigidez	13
adversningun	event	11
adversparaclin	event	11
dolor	rigidez	11
event	matinal	10
adversparaclin	institucional	8
articul	dolor	7
dolor	nieg	7

En la tabla 5-13 se muestra la frecuencia de aparición de los diez primeros términos co-ocurrentes para la sección de tratamiento, siendo el conjunto de los términos inici y tratamient los más frecuentes con 61 apariciones.

Tabla 5-13 Diez términos co-ocurrentes más frecuentes para tratamiento:

Term1	Term2	Frecuencia
inici	tratamient	61
metotrexat	tratamient	54
dolor	rigidez	47
inici	metotrexat	47
previ	tratamient	45
calci	metotrexat	41
dolor	tratamient	41
dolor	hor	40
cloroquin	tratamient	39

		Hace parte de la tabla 5-13
dia	metotrexat	38

5.9 clústeres de palabras clave

Se hizo un análisis de clústeres de palabras claves que permitiera ver la distribución de estas dentro de las historias clínicas (ver tabla 5-14), como resultado se obtuvieron diez clústeres de palabras claves en donde los términos con mayor frecuencia se representan en una nube de palabras en donde el término con la frecuencia mayor es el más grande.

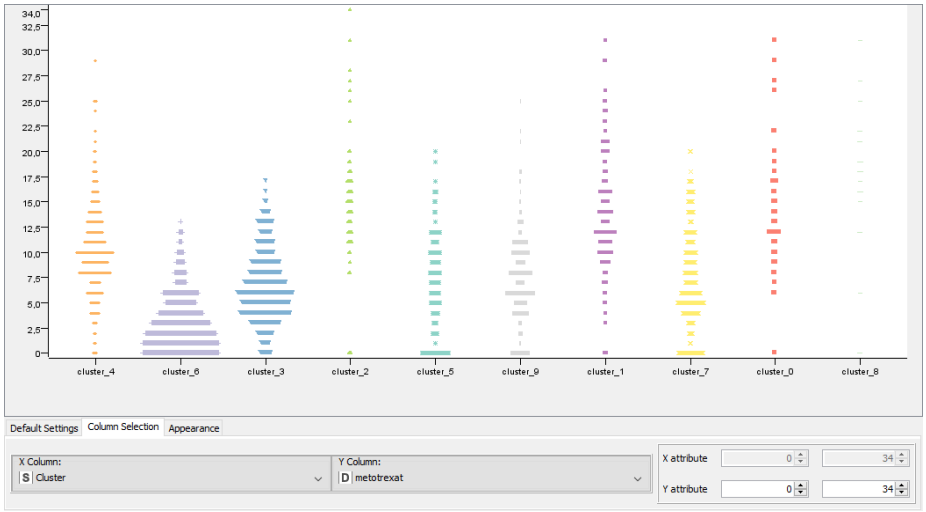
Tabla 5-14 Diez clusters de palabras clave:

CLUSTER 0	CLUSTER 1	CLUSTER 2
		
CLUSTER 3	**CLUSTER 4**	**CLUSTER 5**
		
Hace parte de la tabla 5-14		
CLUSTER 6	**CLUSTER 7**	**CLUSTER 8**

<table>
<tr><td>  </td><td>  </td><td>  </td></tr>
<tr><td>  </td><td>CLUSTER 9</td><td></td></tr>
</table>

Como ejemplo particular, se obtiene la distribución del metotrexato dentro de los conjuntos analizados (ver figura 5-6), esta elección se debe a que es el metotrexato el tratamiento farmacológico más común para tratar la artritis reumatoide.

Figura 5-6 Distribución del Metotrexato en diez clústeres:



5.10 Interoperabilidad semántica con SNOMED

A partir de la conexión establecida con la terminología médica de SNOMED se logran obtener los descriptores y las relaciones, ver figura 5-15. Dentro de las relaciones se pueden evidenciar los atributos, trastornos, estructura corporal y anomalía morfológica asociados a la artritis reumatoide.

Tabla 5-15 Descriptores trastorno Artritis Reumatoide:

ID	Descripción
1014302013	artritis reumatoide
1014303015	artritis reumática crónica
1014304014	artritis atrófica
1014305010	gota reumática
1014306011	artritis reumatoide (trastorno)
116082011	Rheumatoid arthritis
116083018	Atrophic arthritis
116084012	Chronic rheumatic arthritis
116085013	Rheumatic gout
809891012	Rheumatoid arthritis (disorder)
1233202011	RA - Rheumatoid arthritis

Hace parte de la tabla 5-15

1233203018	Rheumatoid disease
1233204012	RhA - Rheumatoid arthritis
1788127016	Proliferative arthritis

En la tabla 5-17 se registran las relaciones asociadas a los hallazgos patológicos en tejidos blandos y óseos afectados por la artritis reumatoide. Estos hallazgos generalmente se hacen evidentes en estudios radiológicos que se les practica a los pacientes siendo las coyunturas las principales afectadas.

Tabla 5-17 relaciones atributos sitios de hallazgo artritis reumatoide:

ID	Relaciones
116680003	es un[a] (atributo)
30701005	poliartropatía inflamatoria (trastorno)
363698007	sitio del hallazgo (atributo)
39352004	estructura articular (estructura corporal)
363698007	sitio del hallazgo (atributo)
306720004	estructura sinovial (estructura corporal)
363698007	sitio del hallazgo (atributo)
64331004	estructura de la membrana sinovial (estructura corporal)
116676008	morfología asociada (atributo)
23583003	inflamación (anomalía morfológica)
116680003	es un[a] (atributo)
3723001	artritis (trastorno)
116680003	es un[a] (atributo)
426760008	trastorno de hipersensibilidad tardía (trastorno)
116680003	es un[a] (atributo)
53338001	artropatía asociada con una reacción de hipersensibilidad (trastorno)
42752001	debido a (atributo)
418925002	reacción por hipersensibilidad inmunológica (trastorno)
116680003	es un[a] (atributo)
85828009	enfermedad autoinmunitaria (trastorno)
363705008	tiene manifestación definitoria (atributo)
106182000	hallazgo del sistema inmunológico (hallazgo)
370135005	proceso patológico (atributo)
263680009	proceso autoinmunitario (calificador)
116676008	morfología asociada (atributo)
23583003	inflamación (anomalía morfológica)
363698007	sitio del hallazgo (atributo)

39352004	estructura articular (estructura corporal)
116680003	es un[a] (atributo)
3723001	artritis (trastorno)
116676008	morfología asociada (atributo)
34248003	inflamación reumática (anomalía morfológica)
363698007	sitio del hallazgo (atributo)
39352004	estructura articular (estructura corporal)
116676008	morfología asociada (atributo)
34248003	inflamación reumática (anomalía morfológica)
116680003	es un[a] (atributo)
85828009	enfermedad autoinmunitaria (trastorno)
116676008	morfología asociada (atributo)
23583003	inflamación (anomalía morfológica)
363698007	sitio del hallazgo (atributo)
39352004	estructura articular (estructura corporal)
246075003	agente causal (atributo)
77089006	factor reumatoide (sustancia)
370135005	proceso patológico (atributo)
263680009	proceso autoinmunitario (calificador)

En la tabla 5-18 se registran las relaciones morfológicas y procesos patológicos provocados por la enfermedad, aquí se establece que la artritis reumatoide es un proceso autoinmunitario y una de las relaciones morfológicas es que provoca inflamación reumática en las estructuras articulares, tiene como característica que es un trastorno de hipersensibilidad tardía y el agente causal es el factor reumatoide.

Tabla 5-18 Relaciones morfologías y procesos patológicos

ID	Relación
116680003	es un[a] (atributo)
3723001	artritis (trastorno)
370135005	proceso patológico (atributo)
263680009	proceso autoinmunitario (calificador)
	Hace parte de la tabla 5-18
116676008	morfología asociada (atributo)
34248003	inflamación reumática (anomalía morfológica)

363698007	sitio del hallazgo (atributo)
39352004	estructura articular (estructura corporal)
116680003	es un[a] (atributo)
426760008	trastorno de hipersensibilidad tardía (trastorno)
116680003	es un[a] (atributo)
85828009	enfermedad autoinmunitaria (trastorno)
116680003	es un[a] (atributo)
64572001	enfermedad (trastorno)
116676008	morfología asociada (atributo)
23583003	inflamación (anomalía morfológica)
246075003	agente causal (atributo)
77089006	factor reumatoide (sustancia)

6. Conclusiones y recomendaciones

6.1 Conclusiones

En este trabajo se desarrolló un método para el análisis de un conjunto de historias clínicas de pacientes con artritis reumatoide aplicando métodos de minería de textos, aprendizaje de máquina, lingüística computacional y otras áreas del conocimiento relacionadas, con el fin de identificar el avance de la enfermedad a partir del corpus analizado.

Este trabajo busca identificar las palabras o conjunto de palabras que permitan dar indicios importantes ya sea para el diagnóstico o desenlace de la enfermedad. Es importante mencionar que no hay términos asociados a aspectos genéticos debido a que no están registrados en las historias clínicas, pero cabe aclarar que sí existiera esta información, el modelo está en la capacidad de identificarlos.

El modelo propuesto y desarrollado está en la capacidad de recuperar información asociada a cualquier enfermedad que se registre en un conjunto de historias clínicas, en este caso particular el corpus hace parte de pacientes diagnosticados con Artritis Reumatoide, sin embargo en el momento de ingresar historias clínicas que estén relacionadas con otras enfermedades por ejemplo cáncer o diabetes, el modelo será capaz de hacer el análisis.

Hay una gran complejidad en el análisis de texto consignado en las historias clínicas, debido a que no tienen la riqueza argumentativa de un texto literario, que

permita recuperar en un espectro amplio y variado todo el detalle asociado al ejercicio médico.

Este tipo de iniciativas de automatización para el descubrimiento de información, aún requiere la intervención humana, es decir un experto en el contexto del trabajo realizado que valide los hallazgos.

Se analiza la información asociada a los tratamientos farmacológicos registrados en las historias clínicas asociandolos a la base de datos CheMBL, que es una base de datos curada manualmente de moléculas bioactivas con propiedades similares a las de los fármacos y que son utilizados en ensayos clínicos que permiten establecer presencia de polimorfismos genéticos y la efectividad clínica de los tratamiento.

Los datos analizados están en formato no estructurado y semi estructurado que representa la narrativa de las historias clínicas de pacientes diagnosticados con la enfermedad de artritis reumatoide, lo que ha sido desafiante, debido al poco desarrollo de investigaciones asociadas al análisis de textos médicos mediante la aplicación de minería de texto, lingüística computacional y demás áreas del conocimiento relacionadas dentro del contexto hispano y en contraste con los desarrollos en esta área para la lengua inglesa.

Se logra establecer una interoperabilidad semántica a partir de la relación establecida entre los términos o palabras consignadas en las historias clínicas y la terminología médica de SNOMED para el idioma español.

El desarrollo de trabajos en análisis de la información y en particular el análisis de texto aplicando métodos de minería de textos, tiene una alta carga de consumo

de recursos computacionales, por lo que puede llegar a ser un factor limitante para la ejecución de este tipo de proyectos.

6.2 Recomendaciones y trabajo futuro

Es importante reconocer que este tipo de investigaciones aún tienen un camino por recorrer, en especial en el contexto latinoamericano y para la lengua española. Se evidencia que se han venido generando trabajos en temas relacionados con la minería de texto para el ámbito médico y en español, pero es evidente que en comparación con los trabajos realizados en el contexto anglosajón aún hay mucho trabajo por hacer.

Sería importante incorporar otras herramientas y fuentes de conocimiento externas como FreeLing, GATE y cTakes para el análisis de información médica no estructurada que permita hallar información no explícita y pueda ayudar a la toma de decisiones a partir del descubrimiento de nuevo conocimiento.

Es importante poder crear un modelo de minería de textos que permita asociar la información numérica relacionada con los tratamientos, exámenes de laboratorio, rutinas de medicación y demás información que sin duda, puede enriquecer el análisis de los datos y así lograr obtener un mapa más completo de todo el ejercicio clínico registrado en la historia clínica del paciente.

Finalmente, el análisis de texto no estructurado se ha realizado sobre técnicas de representación de textos basada en el modelo de bolsa de palabras, TF-IDF y LDA principalmente. Sería interesante abrir una nueva vía de investigación

incorporando nuevos enfoques como el basado en el modelo Word2Vec o Doc2Vec.

Estos modelos pueden llegar a ser más maduros teniendo más herramientas de análisis de información clínica de acceso libre para la lengua española y más bases de datos bioinformáticas.

7. BIBLIOGRAFIA

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: A survey. *Artificial Intelligence in Medicine*, *33*(2), 157–177. https://doi.org/10.1016/j.artmed.2004.07.017

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O., Birnbaum, N. S., Burmester, G. R., Bykerk, V. P., Cohen, M. D., Combe, B., Costenbader, K. H., Dougados, M., Emery, P., Ferraccioli, G., Hazes, J. M. W., Hobbs, K., Huizinga, T. W. J., Kavanaugh, A., … Hawker, G. (2010). 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis and Rheumatism*, *62*(9), 2569–2581. https://doi.org/10.1002/art.27584

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O., Birnbaum, N. S., Burmester, G. R., Bykerk, V. P., Cohen, M. D., Combe, B., Costenbader, K. H., Dougados, M., Emery, P., Ferraccioli, G., Hazes, J. M. W., Hobbs, K., Huizinga, T. W. J., Kavanaugh, A., … Hawker, G. (2010). 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis and Rheumatism*, *62*(9), 2569–2581. https://doi.org/10.1002/art.27584

Arthritis Foundation. (n.d.). *Biológicos/Modificadores de la Respuesta Biológica (MRB)*. https://doi.org/https://espanol.arthritis.org/espanol/tratamientos/medicamentos/medicamentos-por-enfermedad/medicamentos-mrb/

Bichindaritz, I., & Marling, C. (2006). Case-based reasoning in the health sciences: What's next? *Artificial Intelligence in Medicine*, *36*(2), 127–135. https://doi.org/10.1016/j.artmed.2005.10.008

Borsari, C., Trader, D. J., Tait, A., & Costi, M. P. (2020). Designing Chimeric Molecules for Drug Discovery by Leveraging Chemical Biology. *Journal of Medicinal Chemistry*, *63*(5), 1908–1928. https://doi.org/10.1021/acs.jmedchem.9b01456

Cantaert, T., de Rycke, L., Bongartz, T., Matteson, E. L., Tak, P. P., Nicholas, A. P., & Baeten, D. (2006). Citrullinated proteins in rheumatoid arthritis: Crucial... but not sufficient! *Arthritis and Rheumatism*, *54*(11), 3381–3389. https://doi.org/10.1002/art.22206

Carrascal, A. I. O., Cotte, D. S., Arango, N. A. R., & Vélez, A. F. P. (2019). Descubrimiento de Conocimiento en Historias Clínicas mediante Minería de Texto. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, *34*, 29–43. https://doi.org/10.17013/risti.34.29-43

Cely, A. (2018). *Construcción de un modelo de procesamiento de historias clínicas electrónicas de pacientes con artritis reumatoide para la obtención de variables relevantes.* http://bdigital.unal.edu.co/71670/

Chen, E. S., Hripcsak, G., Xu, H., Markatou, M., & Friedman, C. (2008). Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association*, *15*(1), 87–98. https://doi.org/10.1197/jamia.M2401

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., & de Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, *42*(5), 937–949. https://doi.org/10.1016/j.jbi.2008.12.005

Cohen, K. B., & Hunter, L. E. (2013). Chapter 16: Text Mining for Translational Bioinformatics. *PLoS Computational Biology*, *9*(4). https://doi.org/10.1371/journal.pcbi.1003044

Dai, H. J., Wei, C. H., Kao, H. Y., Liu, R. L., Tsai, R. T. H., & Lu, Z. (2015). Text mining for translational bioinformatics. *BioMed Research International*, *2015*, 2–4. https://doi.org/10.1155/2015/368264

Firestein, G. S. (2003). Evolving concepts of rheumatoid arthritis. *Nature*, *423*(6937), 356–361. https://doi.org/10.1038/nature01661

García Martín, M., & Zurita Molina, A. (1998). Transaminasas: Valoración y significación clínica. *Hospital Universitario Virgen Macarena*, 267–275. https://www.aeped.es/sites/default/files/documentos/transaminasas.pdf

Guerrero Pupo, J. C., Amell Muñoz, I., & Cañedo Andalia, R. (2004). Tecnología, tecnología médica y tecnología de la salud: Algunas consideraciones básicas. *Acimed*, *12*(4), 1–16.

Hajeer, A. H., & Hutchinson, I. v. (2001). Influence of TNFα gene polymorphisms on TNFα production and disease. *Human Immunology*, *62*(11), 1191–1199. https://doi.org/10.1016/S0198-8859(01)00322-6

Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial Intelligence for Clinical Trial Design. *Trends in Pharmacological Sciences*, *40*(8), 577–591. https://doi.org/10.1016/j.tips.2019.05.005

Hearst, M. A. (1999). *Untangling text data mining*. 3–10. https://doi.org/10.3115/1034678.1034679

Hernandez, E. P. H., & Quimbaya, A. P. (2016). *HTL: Modelo para la extracción, estructuración y visualización de eventos médicos a partir de texto narrativo en historias clínicas electrónicas*. 1–9. https://doi.org/10.1109/columbiancc.2016.7750768

K, C. M., González, A., & L, G. Q. (2008). Guía de tratamiento de la artritis reumatoide temprana en un Hospital Universitario de Colombia Guideline on treatment of early rheumatoid arthritis in a University Hospital of Colombia. *Revista Colombiana de Reumatología*, *15*(2), 79–91.

Kushima, M., Araki, K., Suzuki, M., Araki, S., & Nikama, T. (2011). Text data mining of in-patient nursing records within electronic medical records using keygraph. *IAENG International Journal of Computer Science*, *38*(3), 215–224.

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*(3), 241–247. https://doi.org/10.1038/ng1195-241

Mease, P. J., Stryker, S., Liu, M., Salim, B., Rebello, S., Gharaibeh, M., & Collier, D. H. (2021). Treatment patterns in rheumatoid arthritis patients newly initiated on biologic and conventional synthetic disease-modifying antirheumatic drug therapy and enrolled in a North American clinical registry. *Arthritis Research and Therapy*, *23*(1), 1–13. https://doi.org/10.1186/s13075-021-02599-4

MedlinePlus. (2021a). *Inyección de rituximab*. https://medlineplus.gov/spanish/druginfo/meds/a607038-es.html

MedlinePlus. (2021b). *Proteína C reactiva*. Proteína C Reactiva. https://doi.org/https://medlineplus.gov/spanish/ency/article/003356.htm

Montes y Gómez, Manuel and Gelbukh, Alexander and López López, A. (2005). RESUMEN DE TESIS DOCTORAL Minería de Texto empleando la Semejanza entre Estructuras Semánticas. *Computacion y Sistemas*, *9*, 63–81.

Morales Muñoz, L. A. (2014). *Modelo computacional para la identificacion de endofenotipos en pacientes con artritis reumatoide utilizando informacion del antigeno leucocitario humano HLA clase II*. http://www.bdigital.unal.edu.co/44357/

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Greg Miller, W., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., & Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine*, *36*(12), 1351–1377. https://doi.org/10.1016/j.compbiomed.2005.08.003

O'Dell, J. R., Mikuls, T. R., Taylor, T. H., Ahluwalia, V., Brophy, M., Warren, S. R., Lew, R. A., Cannella, A. C., Kunkel, G., Phibbs, C. S., Anis, A. H., Leatherman, S., & Keystone, E. (2013). Therapies for Active Rheumatoid Arthritis after Methotrexate Failure. *New England Journal of Medicine*, *369*(4), 307–318. https://doi.org/10.1056/nejmoa1303006

Ochoa, I. J. A., Orellana García, I. A., Sánchez Corales, I. Y., & Hernández, F. D. (2013). *Datos Agrupamiento Sld070 Web Component for the Analysis of Clinical Infor- Mation Using the Technique of Clustering Data Mining*. 8.

Páez Leal, M., Gómez, L., & Anaya, J. (2010). Implicaciones funcionales de los linfocitos B en el desarrollo de la artritis reumatoide. *MedUNAB*, *9*(1).

Pedrosa, M. M. F. (2002). Evaluación del daño radiográfico en la artritis reum atoide. *Revista Española de Reumatología*, *1*(S1), 22–26.

Pereira, L., Rijo, R., Silva, C., & Agostinho, M. (2013). ICD9-based Text Mining Approach to Children Epilepsy Classification. *Procedia Technology*, *9*, 1351–1360. https://doi.org/10.1016/j.protcy.2013.12.152

Perner, P. (2006). Intelligent data analysis in medicine-Recent advances. *Artificial Intelligence in Medicine*, *37*(1), 1–5. https://doi.org/10.1016/j.artmed.2005.10.003

Piedra, D., Ferrer, A., & Gea, J. (2014). Minería de textos y medicina: Utilidad en las enfermedades respiratorias. *Archivos de Bronconeumologia*, *50*(3), 113–119. https://doi.org/10.1016/j.arbr.2014.02.008

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). DISEASES: Text mining and data integration of disease-gene associations. *Methods*, *74*, 83–89. https://doi.org/10.1016/j.ymeth.2014.11.020

Quiceno, J. M., Vinaccia, S., & Remor, E. (2011). Empowerment program of resilience for rheumatoid arthritis patients. *Revista de Psicopatologia y Psicologia Clinica*, *16*(1), 27–47. https://doi.org/10.5944/rppc.vol.16.num.1.2011.10349

Ribeiro, J., Duarte, J., Portela, F., & Santos, M. F. (2019). Automatically detect diagnostic patterns based on clinical notes through text mining. *Procedia Computer Science*, *160*, 684–689. https://doi.org/10.1016/j.procs.2019.11.027

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, *42*(5), 950–966. https://doi.org/10.1016/j.jbi.2008.12.013

Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. In *Neuroscience Research Communications*.

Sullivan, D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. John Wiley & Sons, Inc.605 Third Ave. New York, NYUnited States. http://dl.acm.org/citation.cfm?id=516935

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science*. https://doi.org/10.1002/(SICI)1097-4571(198707)38:4<228::AID-ASI2>3.0.CO;2-G

Tan, E. M., & Smolen, J. S. (2016). Historical observations contributing insights on etiopathogenesis of rheumatoid arthritis and role of rheumatoid factor. *Journal of Experimental Medicine*, *213*(10), 1937–1950. https://doi.org/10.1084/jem.20160792

Testa, D., Calvacchi, S., Petrelli, F., Giannini, D., Bilia, S., Alunno, A., & Puxeddu, I. (2021). One year in review 2021: Pathogenesis of rheumatoid arthritis. *Clinical and Experimental Rheumatology*, *39*(3), 445–452.

UMLS, R. M. (2009). *UMLS ® Reference Manual* (Issue September).

UNAL. (2021). *Medicamentos biológicos, caros para tratar la artritis*. https://doi.org/https://unperiodico.unal.edu.co/pages/detail/medicamentos-biologicos-caros-para-tratar-la-artritis/

Vijayakrishnan, R., Steinhubl, S. R., Ng, K., Sun, J., Byrd, R. J., Daar, Z., Williams, B. A., Defilippi, C., Ebadollahi, S., & Stewart, W. F. (2014). Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of Cardiac Failure*, *20*(7), 459–464. https://doi.org/10.1016/j.cardfail.2014.03.008

Wang, R., Zhao, J., Peng, L., Yang, B., Wang, L., & Li, B. (2018). Medical entity recognition of Esophageal Carcinoma based on word clustering. *2018 International Conference on*

Security, Pattern Analysis, and Cybernetics, SPAC 2018, *i*, 348–353. https://doi.org/10.1109/SPAC46244.2018.8965515

Wu, C. S., Kuo, C. J., Su, C. H., Wang, S. H., & Dai, H. J. (2020). Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. *Journal of Affective Disorders*, *260*(415), 617–623. https://doi.org/10.1016/j.jad.2019.09.044

Yauney, G., & Shah, P. (2018). Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection. *Proceedings of the 3rd Machine Learning for Healthcare Conference*, *85*, 161–226. http://web.media.mit.edu/~pratiks/mlhc_2018/reinforcement_learning_with_action_derived_rewards_for_chemotherapy_and_clinical_trial_dosing_regimen_selection.pdf

Zhai, C., & Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. In *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. https://doi.org/10.1145/2915031