Herramienta de apoyo al diagnóstico basada en el análisis de historias clínicas

Facultad de Informática
Universidad Complutense de Madrid

Departamento de Ingeniería del Software e Inteligencia Artificial
Curso 2016/2017

Fernando Miñambres González
Zhihao Zheng

Director:
Alberto Díaz Esteban

Fernando Miñambres González y Zhihao Zheng, autores del presente documento y del proyecto “Herramienta de apoyo al diagnóstico basado en el análisis de historias clínicas” autorizan a la Universidad Complutense de Madrid a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a los autores, tanto la propia memoria, los códigos, el prototipo desarrollado y las imágenes utilizadas.

Madrid, 16 de junio de 2017.

Fernando Miñambres González y Zhihao Zheng.

Queremos agradecer a un gran número de personas, sin los que este proyecto no se habría llevado a cabo.

En primer lugar, a nuestro tutor, Alberto Díaz Esteban, por confiar en todo momento en nuestras posibilidades así como por su colaboración a lo largo de todo el año académico.

También queremos agradecer a todos los profesores que nos han impartido clase a lo largo de estos años y nos han intentado formar para poder alcanzar nuestras metas.

Finalmente, a todos nuestros familiares, amigos y compañeros que han mostrado su interés así como brindado su ayuda en los momentos que la hemos necesitado.

A todos, muchas gracias.

RESUMEN

Actualmente una de las tareas más importantes a las que se enfrenta un médico es encontrar el diagnóstico de un paciente en el menor tiempo posible para poder hacer frente a la situación. Este proyecto tiene como objetivo ayudar al personal médico ofreciéndoles la posibilidad de comparar un gran número de informes en busca de aquellos más similares respecto a la información extraída de cada informe.
Para ello, hay que conseguir representar todos los informes de una manera similar, tal y como se describe en este proyecto, de forma que la comparación sea lo más sencilla y precisa posible.

Esta aplicación se divide en tres partes: las dos primeras llevan a cabo la construcción de las representaciones extrayendo la mayor información posible de cada informe médico y la tercera parte se encarga de realizar la búsqueda por similitud en dichas representaciones.

Palabras clave: informe médico, ontología, WordNet, SNOMED-CT, FreeLing, ElasticSearch.

ABSTRACT

At the moment one of the most important tasks that a doctor faces is to find the diagnosis of a patient in the shortest time possible to be able to face the situation. This project aims to help medical staff by offering the possibility of comparing a large number of reports in search of those more similar to the symptoms presented by the patient.
To do this, you must be able to represent all reports in a similar way, as described in this project, so that the comparison is as simple and accurate as possible.

This application is divided into three parts: the first two carry out the construction of the representations by extracting as much information as possible from each medical report and the third part is responsible for the search for similarity in said representations.

Keywords: medical report, ontology, WordNet, SNOMED-CT, FreeLing, ElasticSearch.

Índice general

1 Introducción

1.1 Motivación

En los últimos años, gracias a los avances de las nuevas tecnologías, se ha producido un gran auge en el uso y aplicación de las Tecnologías de la Información y las Comunicaciones (TIC) en los centros especializados de la salud como hospitales o policlínicos ayudando, entre otras cosas, en la gestión de la información. Pero este gran almacenamiento de datos puede resultar difícil de tratar para un personal médico que quiera repasar el historial clínico de un paciente o simplemente compararlo con el de otros con tal de encontrar un diagnóstico lo más rápido y preciso posible.

Es este último caso, la base a partir de la cual surge el tema principal de este trabajo, la creación de una herramienta de ayuda al diagnóstico basada en el análisis de historias clínicas. Se tratará de una herramienta que sirva de apoyo al personal médico a la hora de realizar el diagnóstico de un paciente. Dicha herramienta se basará en la comparación entre diversos historiales médicos, buscando aquellos más similares en cuanto a la información contenida en cada informe en cuestión.

Este trabajo surge como continuación de otro titulado "Herramienta para búsqueda de casos médicos semejantes" realizado por Agustín Pastore Burgos bajo la dirección de Alberto Díaz Esteban. Este proyecto tenía como objetivo principal analizar un gran número de historiales médicos y extraer aquellos cuyo índice de similitud era mayor. Pero, este trabajo, tenía una serie de limitaciones a partir de las cuales surge nuestra motivación para este proyecto. Dicho trabajo necesitaba del uso de un traductor externo ya que la herramienta utilizada para analizar los historiales sólo aceptaba textos en inglés. Este hecho, dotaba a la aplicación final de una gran complejidad debido a esta dependencia de factores externos. Es por ello que nuestra motivación ha sido encontrar una forma de poder hacerlo en textos tanto en español como en inglés.

El proyecto del año pasado contaba con la colaboración del Hospital Público Comarcal de Jarrio en Asturias, el cual proporcionó una gran cantidad de historiales médicos para llevar a cabo el trabajo. Todos estos historiales son anónimos y su explotación en este ámbito académico fue aprobada por un Comité de Ética, el cual dio luz verde para que se nos fueran proporcionados dichos historiales.

Para este proyecto seguiremos contando con esta colaboración y con todos los informes que fueron utilizados en el proyecto del año pasado que nos servirán de base para llevar a cabo el objetivo de este trabajo.

También han sido de gran ayuda para la consecución de este proyecto, un artículo publicado en 2013, el cual trataba sobre el uso de la herramienta FreeLing con la ontología SNOMED-CT. Este artículo titulado *"Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names"* publicado por Maite Oronoz, Arantza Casillas, Koldo Gojenola y Alicia Pérez, nos facilitó la idea de cómo enriquecer el diccionario de FreeLing de tal forma que pudiera analizar textos médicos identificando conceptos SNOMED-CT.

1.2 Objetivos

El objetivo principal de este trabajo es la creación de una herramienta útil e intuitiva que sirva de apoyo a todo el personal médico a la hora de realizar el diagnóstico de un paciente. Esta herramienta ayudará al médico a encontrar de una manera rápida aquellos historiales médicos similares basándose en los síntomas presentados por el paciente.

Para cumplir con el objetivo, lo más importante es extraer la mayor cantidad de información de todos los archivos proporcionados, independientemente del formato que tenga cada archivo.

Otro aspecto importante será resolver el problema del idioma. Actualmente casi todas las herramientas para el análisis de documentos médicos utilizan el inglés como idioma principal. Este hecho obliga a la utilización de un traductor para pasar el texto en español que nos ha sido facilitado al inglés para que pueda ser utilizado por las diferentes herramientas. Esto conlleva una dependencia total del traductor para un uso adecuado de las herramientas. Además, siempre existe la posibilidad de que el traductor haga una mala traducción y esta traducción genere errores que se propaguen al resto de la aplicación. Para hacer frente a este problema del traductor, hemos recurrido a una herramienta llamada FreeLing la cual es capaz de analizar textos escritos tanto en español como en inglés.

Para llegar a cumplir con el objetivo de este trabajo será necesario conseguir un documento con una estructura similar para todos los informes que nos han sido facilitados de manera que la comparación sea lo más efectiva posible. Además, habrá que presentar el resultado en una interfaz intuitiva y familiar que ayude al usuario final de la aplicación a un manejo correcto y eficiente de la misma, puesto que éste puede carecer de conocimientos informáticos avanzados y podría resultarle demasiado difícil utilizar algo que no entiende.

La herramienta se dividirá en tres fases: las dos primeras estarán relacionadas con todo lo referente al procesado de los historiales médicos y la tercera se encargará de buscar aquellos más similares mostrando los resultados en una interfaz web:

1. **Pre-procesamiento** de los historiales médicos proporcionados para obtener una representación estructurada por distintas secciones médicas.

2. **Procesamiento** de las secciones médicas de las representaciones obtenidas en el pre-procesamiento en busca de conceptos médicos.

3. **Búsqueda de similitud** en las representaciones procesadas, para encontrar aquellos más similares en base a los conceptos médicos extraídos en el procesamiento.

1.3 Visión general del documento

El presente documento está dividido en varios capítulos, los cuales son descritos a continuación:

1. **Introducción**: describe brevemente el punto de partida de este proyecto, así como los objetivos que se desean lograr con él.

2. **Estado de la Cuestión**: describe el estado actual de desarrollo en las áreas a las que afecta el proyecto, así como las tecnologías utilizadas para la realización del mismo.

3. **Procesado de los informes**: en este capítulo se explica en profundidad todo lo relacionado con el procesamiento de los historiales médicos.

4. **Búsqueda de similitud en informes procesados**: en este capítulo se explica de manera detallada cómo se realiza la búsqueda de historiales médicos similares en base a unos conceptos médicos.

5. **Conclusiones y trabajo futuro**: en este capítulo se incluyen las conclusiones sacadas tras la realización del proyecto, así como una serie de nuevas funcionalidades que podría prestar nuestra aplicación en un futuro.

6. **Bibliografía**: en este capítulo se mostrarán una lista de enlaces y referencias bibliográficas que nos han servido de ayuda para la realización del proyecto.

7. **Apéndices**: este capítulo servirá para proporcionar información extra sobre nuestro proyecto tal como los manuales de uso de la aplicación.

2 Estado de la cuestión

2.1 Introducción

Este capítulo tiene como objetivo mostrar el estado actual de las tecnologías relacionadas con el ámbito de este proyecto, es decir, aquellas que hemos decidido utilizar ya fuera reciclándolas del trabajo realizado previamente, así como las nuevas tecnologías introducidas con el fin de alcanzar el objetivo de este proyecto.

Para hablar sobre estas tecnologías, procederemos a realizar una división entre aquellas relacionadas con el procesamiento de los datos de los historiales médicos, las encargadas de realizar la búsqueda de historiales similares respecto a la terminología médica y la encargada de mostrar los resultados de una manera visual de tal forma que sea mucho más sencillo entenderlos.

- **Procesamiento de datos:**

  Ontología general: En este proyecto hemos realizado pruebas previas con la ontología **WordNet** con el objetivo de entender el funcionamiento de la herramienta FreeLing, para más tarde poder llevar a cabo el propósito de este proyecto.

  Ontología médica: tiene el propósito de expresar en términos precisos los complejos conceptos e ideas del mundo de la medicina. Cada término debe poseer un significado único aceptado por la comunidad científica, facilitando, así, el intercambio de información a nivel internacional. Las terminologías que se han tenido en cuenta, por ser las más completas, son SNOMED-Clinical Terms (**SNOMED-CT**) y Unified Medical Language System (**UMLS**).

  ApachePOI: se trata de una librería de Java destinada a la lectura y escritura de documentos con formatos de Microsoft Office como .doc, .xml, etc. En nuestro caso, su utilización es primordial ya que nos ayuda a la lectura de todos los informes que nos han sido facilitados con el fin de extraer la mayor cantidad de datos existentes y estructurarlos de tal forma que podamos realizar el procesamiento oportuno.

  FreeLing: es una biblioteca de C++ que proporciona funcionalidades de análisis del lenguaje (análisis morfológico, análisis sintáctico, detección de entidades nombradas, etc.). Elegimos esta herramienta porque es capaz de trabajar con distintos idiomas, entre ellos el castellano, eliminando la dependencia previamente mencionada del traductor. Para que esta herramienta tuviera éxito en nuestro proyecto, tuvimos que configurarlo de tal manera que permitiera analizar conceptos SNOMED-CT.

- **Búsqueda por similitud:**

  Lucene: una vez procesados los documentos, necesitábamos de una herramienta que nos permitiera comparar estos documentos en busca de similitudes. Decidimos seguir utilizando Lucene visto el buen resultado obtenido en el proyecto del año anterior.

- **Exposición de los resultados:**

  Javascript: es un lenguaje de programación interpretado, orientado a objetos, basado en prototipos, débilmente tipado y dinámico. Decidimos utilizarlo en nuestro proyecto para crear una interfaz web sencilla e intuitiva que permitiera al usuario acceder a ella desde cualquier sitio además de ver los resultados de una manera visual.

2.2 Ontologías

Una *Ontología* es una definición formal de tipos, propiedades y relaciones entre entidades existentes para un dominio de discusión en particular. Por ejemplo, existen una gran cantidad de idiomas en el mundo, cada uno de ellos con un vocabulario amplio. Estos idiomas necesitan del concepto de **Ontología general** para realizar la clasificación de cada término del vocabulario, así como relacionarlo con otros términos similares (sinónimos) o diferentes (antónimos).

Por otro lado, gracias al avance tecnológico, son descubiertas nuevas terminologías médicas que necesitan ser clasificadas. Estas terminologías designan un conjunto de términos propios que pertenecen a un campo de extensión variable, debidamente diferenciado de otros campos.

En el ámbito informático es utilizado el concepto denominado como **Ontología médica** para llevar a cabo la clasificación o agrupación de esa nueva información en clases que contienen una información similar, es decir, agrupar cada término médico con aquellos que tienen un significado parecido o con los que existe una relación directa.

Esta ontología se puede aplicar también en Web Semántica e Inteligencia Artificial con el fin de asimilar y codificar el conocimiento, definiendo relaciones existentes entre los conceptos de un determinado dominio, en nuestro caso la Medicina.

2.2.1 Ontología general

2.2.1.1 WordNet

WordNet es una base de datos léxica que agrupa palabras en conjuntos de sinónimos llamados synsets, proporcionando definiciones cortas y generales, así como almacenando las relaciones semánticas entre dichos conjuntos. *WordNet* es el lexicón computacional más usado para desambiguar el significado de las palabras, una tarea que consiste en asignar el concepto más apropiado a los términos en contexto.

Actualmente la base de datos contiene 155.287 palabras organizadas en 117.659 synsets entre los que se distinguen sustantivos, verbos, adjetivos y adverbios.

2.2.2 Ontología médica

2.2.2.1 SNOMED-CT

SNOMED-CT, o *Systematized Nomenclature of Medicine – Clinical Terms*, es la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia a nivel mundial. Con SNOMED-CT, la información clínica se registra haciendo uso de identificadores que se refieren a los conceptos que se definen formalmente como parte de la terminología. Asimismo, SNOMED-CT soporta el almacenamiento de la información clínica en los niveles apropiados de detalle utilizando los conceptos clínicos relevantes.

SNOMED-CT está estructurado de tal forma que, al introducir información, se hace utilizando sinónimos que se adapten a las preferencias locales mientras se almacena la información de una forma coherente y comparable. Además, gracias a su jerarquía, permite que la información se almacene con diferentes niveles de detalles para adaptarse a cualquier uso (por ejemplo, |neumonía|, |neumonía bacteriana| o |neumonía neumocócica|). También permite añadir detalles adicionales mediante la combinación de conceptos para dotar al concepto de más precisión (por ejemplo, |neumonía neumocócica| con |sitio de búsqueda| con |lóbulo superior derecho del pulmón|).

SNOMED-CT permite una serie de diferentes opciones para la obtención inmediata y posterior reutilización para atender tanto los requisitos clínicos inmediatos como de largo plazo, así como los requisitos de otros usuarios. La jerarquía de SNOMED-CT permite que esa información que se desea obtener pueda satisfacer diferentes necesidades en diferentes niveles de generalización (por ejemplo, obtener subtipos de |trastorno pulmonar| o |infección bacteriana| incluirían |neumonía bacteriana|).

Un concepto SNOMED-CT se representa usando tres tipos de componentes:

- **CONCEPTOS:** representan significados clínicos que se organizan en jerarquías. Cada concepto tiene un único identificador numérico de concepto. Dentro de cada jerarquía, los conceptos se organizan desde lo más general a lo más detallado. Esto permite a los datos clínicos detallados ser almacenados y más tarde agregados a un nivel más general.

- **DESCRIPCIONES:** vinculan términos legibles por el ser humano a los conceptos. Un concepto puede tener asociado varias descripciones, cada una representando un sinónimo que describe el mismo concepto clínico. Cada traducción de SNOMED-CT incluye un conjunto adicional de las descripciones, que vinculan los términos en otro idioma a los mismos conceptos SNOMED-CT. Cada descripción tiene un identificador numérico de descripción.

- **RELACIONES:** unen cada concepto a otros conceptos con los que esté relacionado. Estas relaciones proporcionan definiciones formales y otras propiedades del concepto. Por ejemplo, la relación |es una| relaciona un concepto con unos conceptos más generales. Cada relación tiene un identificador numérico de relación.

Estos componentes se complementan con unos conjuntos de referencia, los cuales proporcionan características flexibles adicionales y habilitan la configuración de la terminología para hacer frente a diversos requisitos.

- **CONJUNTOS DE REFERENCIA**: son un enfoque estándar flexible utilizado por SNOMED-CT para soportar la personalización y enriquecimiento de una variedad de requisitos. Por ejemplo, la representación de subconjuntos, las preferencias del idioma para el uso de determinados términos y el mapeo desde o hacia otros sistemas codificados. Cada conjunto de referencia tiene un identificador único de conjunto.

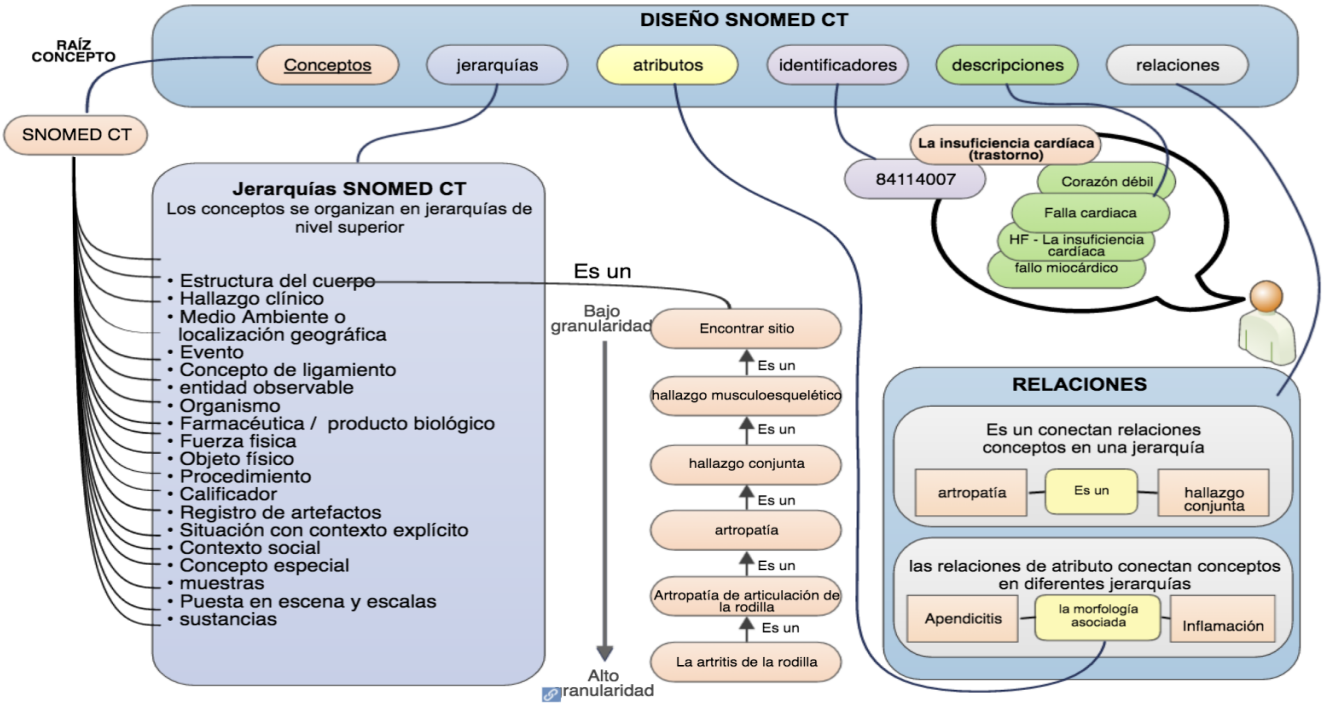


Figura 1. Ejemplo de la estructura de SNOMED-CT.
Fuente: https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics

2.2.2.2 UMLS

UMLS, o *Unified Medical Language System*, es un conjunto de archivos y software que reúne una gran cantidad de terminología médica y de salud, así como distintas normas para permitir la interoperabilidad entre sistemas informáticos.

Uno de los principales usos de UMLS es la vinculación de información de salud, terminología médica, nombres de fármacos y los códigos de facturación a través de diferentes sistemas informáticos.

UMLS cuenta con tres herramientas denominadas como Fuentes de conocimiento:

- **Metatesauro**: forma la base del UMLS y comprende más de un millón de conceptos biomédicos, así como cinco millones de nombres de conceptos derivados de los más de cien vocabularios incorporados. El Metatesauro está organizado por conceptos, y cada concepto tiene atributos específicos que definen su significado y está vinculado a esos mismos conceptos en los distintos vocabularios de origen.

- **Red Semántica**: cada concepto del Metatesauro tiene uno o más tipos semánticos, que están unidos entre sí a través de relaciones semánticas. La red semántica es una colección de estos tipos semánticos y relaciones, siendo bastante amplia ya que cuenta con 135 tipos semánticos y 54 relaciones. Cada tipo semántico viene identificado por un identificador único, una definición, unos ejemplos, su información jerárquica y sus relaciones asociativas.

- **Lexicón ESPECIALISTA y Herramientas léxicas**: el Lexicón ESPECIALISTA contiene información sobre el vocabulario común en inglés, términos biomédicos, términos encontrados en MEDLINE y términos que se encuentran en el Metatesauro. Cada entrada contiene información sintáctica, morfológica y ortográfica el concepto en cuestión.

2.3 ApachePOI

ApachePOI es una API para Java, destinada a la obtención de la información contenida en los ficheros de Microsoft (1997-2008). Su uso en este proyecto es de gran importancia pues nos ayuda a extraer la máxima cantidad de información posible de los documentos Word que nos han sido facilitados.

2.4 FreeLing

2.4.1 Introducción

FreeLing es una biblioteca de C++ que proporciona servicios de análisis del lenguaje. Actualmente proporciona servicios de identificación de lenguajes, análisis léxico, análisis morfológico, detección y clasificación de negaciones, reconocimiento de fechas, magnitudes, monedas, etc., etiquetado PoS, análisis sintáctico superficial, anotación de conceptos WordNet y desambiguación semántica, entre muchas otras cosas.

Esta herramienta está diseñada para ser utilizada como una biblioteca externa desde cualquier aplicación que necesite estos servicios. Además, existen diferentes APIs compatibles con aplicaciones desarrolladas en lenguajes como Java, Perl o Phyton.

2.4.2 Estructura

FreeLing procesa textos y crea una estructura de datos que representan los objetos lingüísticos que aparecen en esos textos. Se entiende por objeto lingüístico los elementos como una palabra, una etiqueta gramatical, una oración, etc. Esto se realiza gracias a unos módulos de procesamiento, los cuales reciben alguno de estos objetos (por ejemplo, una frase), a la cual añaden información adicional como puede ser la adición de etiquetas gramaticales a las palabras de la frase.

Los módulos de procesamiento más importantes de la herramienta FreeLing son:

- **Módulo para identificar el idioma:** se encarga de comparar el texto dado con los diferentes modelos de idiomas disponibles, devolviendo el idioma al cual se asemeja más el texto escrito.

- **Módulo tokenizer:** se trata del primer módulo en la cadena de procesamiento. Su función reside en convertir un texto plano a una lista de palabras en base a un conjunto de reglas de tokenización. Estas reglas son expresiones regulares que se comparan con el comienzo de la primera línea del texto que se está procesando. Una vez se encuentra un token, la subcadena coincidente se elimina de la línea y este proceso se repite hasta que la línea está vacía.

- **Módulo splitter:** este módulo recibe la lista de palabras previamente obtenida y la va procesando con el objetivo de devolver una lista de frases. El buffer del splitter puede retener parte de los tokens si la lista no termina de una forma clara. Además, cada frase obtenida tendrá su propio identificador asignados secuencialmente (comenzando en el 1).

- **Módulo de análisis morfológico:** se trata de un meta-módulo que no realiza ningún procesamiento por sí mismo. Se encarga de crear las instancias y realizar las llamadas a los distintos submódulos que lo componen como el de detección de puntuación, detección de números, detección de fechas, etc.

2.4.3 Referencias

La decisión de utilizar esta herramienta reside principalmente en el hecho de que es una herramienta multilingüe, es decir, es capaz de analizar textos en varios idiomas (entre ellos el castellano) por lo que resulta mucho más beneficioso ya que no depende de factores externos como un traductor, disminuyendo el esfuerzo a realizar, así como la complejidad de la aplicación final.

Por otra parte, también nos ha sido de gran ayuda un artículo publicado en 2013 en el cual se intenta desarrollar una herramienta denominada FreeLing-Med, la cual tiene como propósito crear un analizador de textos médicos en español haciendo uso de la herramienta FreeLing, un analizador morfosintáctico multilingüe.

Este artículo titulado *"Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names"* y publicado por Maite Oronoz, Arantza Casillas, Koldo Gojenola y Alicia Pérez cuenta cómo modificar el diccionario de búsqueda de la herramienta FreeLing de tal manera que sea capaz de reconocer terminología médica. Para ello, se comprobó que es necesario añadir los conceptos de la ontología SNOMED-CT al diccionario de búsqueda de FreeLing, siempre y cuando estos no pertenezcan ya a este diccionario. Los resultados de este estudio muestran cómo en un principio el diccionario de FreeLing albergaba 9.302 conceptos SNOMED-CT, al cual fueron añadidos 23.399 nuevos conceptos con lo que la aplicación final sería capaz de identificar un total de 32.701 conceptos SNOMED-CT.

2.4.4 FreeLing API

FreeLing cuenta con una API desarrollada para varios lenguajes de programación como Java, Perl, PHP, Python y Ruby. En nuestro proyecto hemos decidido utilizar la API para Java, ya que nuestra aplicación está desarrollada en dicho lenguaje. Para poder hacer uso de esta API en el sistema operativo Windows es necesario realizar una configuración de archivos descrita en el punto 1 de los Apéndices de este documento.

2.5 Lucene

Toda la representación que hemos conseguido tras realizar el procesamiento de los informes médicos, está construida de tal manera que pueda ser utilizada por un motor de búsqueda que permita realizar la búsqueda por similitud de una forma eficaz y rápida. Para ello, hemos decidido utilizar *Lucene* visto el buen rendimiento obtenido en el proyecto del año anterior. La única novedad que cabe destacar es el uso de un servidor de búsqueda basado en Lucene y llamado ElasticSearch.

Su uso se debe a que tanto la búsqueda de similitudes como la muestra de los datos resultantes de dicha búsqueda se realizan a través de una interfaz web que hemos desarrollado también.

2.5.1 ElasticSearch

ElasticSearch es un servidor de búsqueda basado en Lucene. Provee de un motor de búsqueda de texto completo, distribuido y con capacidad de dar servicio a múltiples clientes a través de una interfaz web RESTful y utilizando como entrada archivos JSON. Está desarrollado en Java y actualmente está publicado como código abierto bajo las condiciones de la licencia Apache.

Esta herramienta nos permite indexar un gran volumen de datos (en nuestro caso, todos los informes médicos) para posteriormente hacer consultas sobre ellos soportando entre muchas otras cosas búsquedas aproximadas, facetas y resaltado.
Al estar todos los datos indexados, los resultados se obtienen de formar muy rápida. Lo único que hay que hacer es añadir los ficheros JSON con sus propiedades y ElasticSearch se encarga de indexarlo y asignarle un identificador para que más tarde la búsqueda se ejecute rápidamente.

La función más destacada de ElasticSearch, la cual es imprescindible en nuestro proyecto es la función *More Like This Query*. Esta función permite al usuario, realizar una búsqueda de similitud entre un conjunto de documentos. Para ello, la función More Like This Query selecciona un conjunto de términos representativos del documento de entrada, en nuestro caso, el archivo procesado. Con este conjunto forma una consulta, la ejecuta y devuelve los resultados. El usuario es capaz de controlar los campos (*fields*) del documento sobre los que se desea realizar la búsqueda, el número mínimo de veces que debe aparecer un término para que se contabilice como similar (*min_term_freq*), así como el número máximo de resultados que se desea obtener (*max_query_terms*).

2.6 JavaScript

JavaScript es un lenguaje de programación interpretado, dialecto del estándar ECMAScript. Se define como orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico. Su uso se centra principalmente en el lado del cliente, implementado como parte de un navegador web permitiendo mejoras en la interfaz de usuario y páginas web dinámicas. Hemos decidido usarlo para desarrollar una pequeña interfaz web que nos ayude a reflejar todos los resultados obtenidos en este proyecto.

3 Procesado de los informes

3.1 Introducción

Este capítulo tiene como objetivo explicar el funcionamiento, la estructura, así como los problemas que han ido apareciendo en el núcleo central de este proyecto, el procesado de los informes médicos.

Para llevar a cabo dicho procesado, la aplicación final cuenta con dos partes diferenciadas y dependientes entre sí:

- **Pre-procesado:** se encarga de leer los informes médicos facilitados en formato .doc o .txt con el fin de obtener representaciones con la extensión .mxml, es decir, un fichero .xml estructurado por secciones médicas tal y como el usuario lo indique en los archivos de configuración. Se trata de la primera toma de contacto con los nuevos informes.

- **Procesado:** se encarga de analizar las representaciones .mxml pre-procesadas anteriormente utilizando la herramienta FreeLing con el objetivo de identificar conceptos respecto a la ontología SNOMED-CT. A cada concepto se le asignará un identificador de tal manera que posteriormente sea más sencillo realizar una búsqueda por similitud entre las distintas representaciones.

3.2 Informes médicos

Como se ha indicado en el capítulo anterior, el objetivo es extraer la mayor información posible de los informes médicos que nos han sido facilitados. Es en este punto donde nos hemos encontrado los primeros problemas:

- Los informes presentan una estructura diferente y vienen dados en diferentes formatos (Word y Excel).

- Los informes contienen distinta información médica (existen documentos de Cirugía, Informe de alta, Registro de Enfermería, etc.).

Esto nos llevó a dedicar una parte del tiempo del desarrollo del proyecto a analizar los diferentes informes médicos para crear una estructura unificada para todos ellos, organizados en las siguientes secciones médicas: Motivo ingreso, Alergias, Medicación actual, Antecedentes familiares, Antecedentes personales, Anamnesis, Exploración, Pruebas complementarias, Evolución, Intervención Quirúrgica, Juicio clínico, Tratamiento y Plan terapéutico.

3.3 Pre-procesado

La primera actividad que realiza nuestra aplicación es el Pre-procesado de los informes médicos. El objetivo de esta actividad es obtener la mayor cantidad de información posible de todos los documentos facilitados, en formato .doc, y almacenarlos como representaciones con formato .xml estructurados y organizados en las secciones médicas descritas en el punto anterior.

Estas secciones médicas se pueden configurar en los archivos de configuración *conf/secciones.properties* antes de arrancar la aplicación, de tal manera que el usuario tenga la posibilidad de extraer información de todos los campos o de sólo aquellos que considere oportunos para su objetivo.

Por otro lado, disponemos de un archivo de configuración *conf/AbreviaturasSiglas.properties* en el cual hemos registrado todas las posibles abreviaturas de conceptos que pueden aparecer en los informes. Una vez se ejecuta, el pre-procesado las abreviaturas que aparecen en los informes se intercambian por su nombre completo.

Además, la aplicación necesita que el usuario aporte dos datos desde la interfaz para que esta actividad pueda llevarse a cabo, la ruta donde se encuentran los ficheros a pre-procesar y la ruta donde quiere que se almacene el resultado.

Otra cuestión que es importante destacar es que esta actividad pre-procesa independientemente cada archivo, es decir, que no es necesario pre-procesar todos los archivos cuando lo único que se quiere es extraer la información de uno sólo.

Por último, cabe mencionar que, para llevar a cabo esta actividad, lo único que hemos necesitado usar ha sido la librería Apache POI para ir leyendo los documentos en formato .doc que nos han sido facilitados, ya que se trataba del formato predominante entre todos ellos.

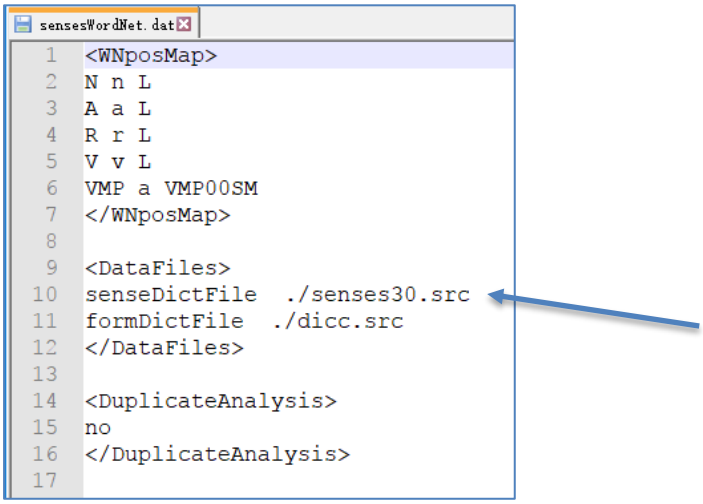
3.4 Procesado

La segunda actividad de la aplicación consiste en el procesado de las representaciones generadas en el punto anterior. El objetivo es analizar dichas representaciones con el fin de identificar conceptos de la ontología SNOMED-CT.

En esta actividad es en la que más problemas hemos encontrado puesto que hemos utilizado gran parte del tiempo en entender cómo funciona la herramienta FreeLing. Esto se debe a que no existe gran información respecto a ella ya que se trata de una herramienta que aún en la actualidad sigue desarrollándose.

3.4.1 Uso de FreeLing con WordNet

Por esta razón, comenzamos procesando los documentos .xml identificando conceptos de la ontología WordNet. Esto lo conseguimos configurando el archivo de configuración *..\freeling-3.1-win64\data\es\sensesWordNet.dat* indicándole que realizará el procesado en base a dicha ontología añadiendo ***./senses30.src*** en el apartado *senseDictFile* de la sección *DataFiles*.



Una vez ejecutado el procesado de los informes, se observa un resultado como el de la figura 3. En él se aprecia los conceptos extraídos, clasificados por secciones médicas y a cada concepto se le asocia una etiqueta gramatical extraída del diccionario propio de FreeLing, así como el identificador de concepto extraído de la base de datos de WordNet. Por ejemplo, al término *tratamiento* se le asocia la etiqueta **NCMS000** que indica que es un Nombre Común Masculino Singular y por otro lado se le asocia el identificador **00658082-n** que confirma que este término es un Nombre por la terminación –n del identificador.

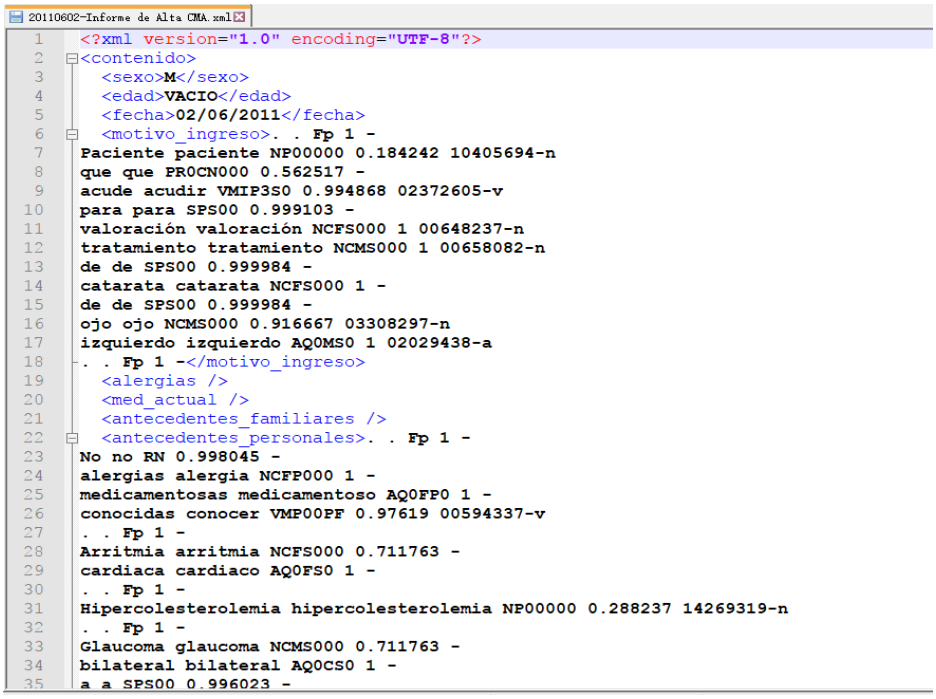


Figura 3. *Resultado de la ejecución del procesado mediante WordNet.*

3.4.2 Adaptación SNOMED-CT a FreeLing

Una vez visto el resultado obtenido, creímos que lo único que nos quedaba por hacer sería cambiar el archivo de configuración anteriormente mencionado de tal manera que en vez de procesar los documentos buscando conceptos de la ontología WordNet lo hiciera respecto a conceptos SNOMED-CT. Pero es aquí donde encontramos el principal inconveniente.

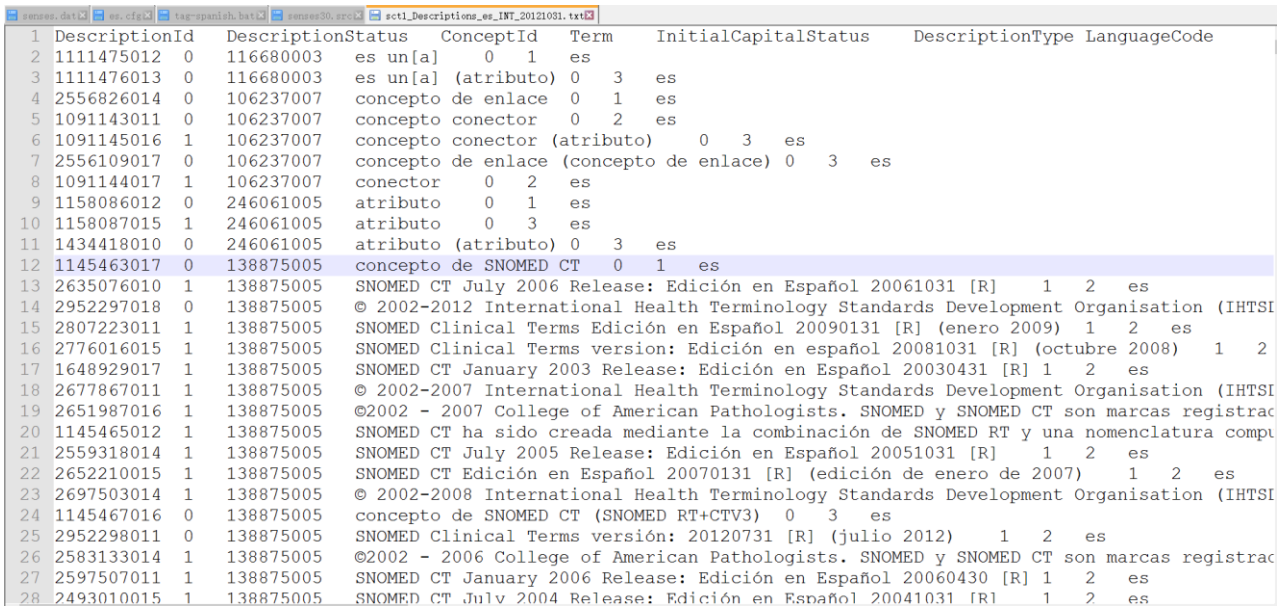
El archivo donde se encuentra la base de datos con los conceptos de la ontología WordNet cuenta con dos columnas referentes al identificador y al concepto. Por otro lado, el de la ontología SNOMED-CT contiene siete columnas y, por tanto, la herramienta FreeLing no realizaba la búsqueda correctamente ya que era incapaz de asociar los conceptos con sus identificadores.

Es por ello, que el siguiente paso a dar fue reestructurar el archivo donde se encuentra la base de datos con los conceptos de la ontología SNOMED-CT, de tal manera que se asemejara al de WordNet con el objetivo de hacer efectivo el uso de FreeLing.

Para llevar a cabo este paso, decidimos quedarnos con las columnas ***ConceptId*** y ***Term*** del archivo con la base de datos de los conceptos SNOMED-CT a parte de añadir al ***ConceptId*** la terminación "-x", pudiendo tomar la x los valores "n", "v", "r", "a" si el concepto es un nombre, un verbo, adverbio o un adjetivo.
Para ello, hemos decidido transformar el archivo original con la base de datos de SNOMED-CT a un archivo reconocible por FreeLing para poder llevar a cabo el procesamiento.

Para ello, hemos extraído del archivo original la información referente al ConceptId y al Term almacenándola en un nuevo archivo. Además de esto, hemos añadido la terminación "-n" al ConceptId para hacer posible que FreeLing lo reconozca como Nombre y así pueda ejecutar el procesamiento.

Hemos decidido utilizar sólo la terminación "-n" debido a que el gran tamaño de la base de datos nos ha hecho imposible analizar si cada concepto se refería a un nombre, un verbo, un adverbio o un adjetivo.



	DescriptionId		DescriptionStatus	ConceptId	Term	InitialCapitalStatus	DescriptionType	LanguageCode
2	1111475012	0	116680003	es un[a]	0	1	es	
3	1111476013	0	116680003	es un[a] (atributo)	0	3	es	
4	2556826014	0	106237007	concepto de enlace	0	1	es	
5	1091143011	0	106237007	concepto conector	0	2	es	
6	1091145016	1	106237007	concepto conector (atributo)	0	3	es	
7	2556109017	0	106237007	concepto de enlace (concepto de enlace)	0	3	es	
8	1091144017	1	106237007	conector	0	2	es	
9	1158086012	0	246061005	atributo	0	1	es	
10	1158087015	1	246061005	atributo	0	3	es	
11	1434418010	0	246061005	atributo (atributo)	0	3	es	
12	1145463017	0	138875005	concepto de SNOMED CT	0	1	es	
13	2635076010	1	138875005	SNOMED CT July 2006 Release: Edición en Español 20061031 [R]	1	2	es	
14	2952297018	0	138875005	© 2002-2012 International Health Terminology Standards Development Organisation (IHTSI				
15	2807223011	1	138875005	SNOMED Clinical Terms Edición en Español 20090131 [R] (enero 2009)	1	2	es	
16	2776016015	1	138875005	SNOMED Clinical Terms version: Edición en español 20081031 [R] (octubre 2008)	1	2		
17	1648929017	1	138875005	SNOMED CT January 2003 Release: Edición en Español 20030431 [R]	1	2	es	
18	2677867011	1	138875005	© 2002-2007 International Health Terminology Standards Development Organisation (IHTSI				
19	2651987016	1	138875005	©2002 - 2007 College of American Pathologists. SNOMED y SNOMED CT son marcas registrac				
20	1145465012	1	138875005	SNOMED CT ha sido creada mediante la combinación de SNOMED RT y una nomenclatura compu				
21	2559318014	1	138875005	SNOMED CT July 2005 Release: Edición en Español 20051031 [R]	1	2	es	
22	2652210015	1	138875005	SNOMED CT Edición en Español 20070131 [R] (edición de enero de 2007)	1	2	es	
23	2697503014	1	138875005	© 2002-2008 International Health Terminology Standards Development Organisation (IHTSI				
24	1145467016	0	138875005	concepto de SNOMED CT (SNOMED RT+CTV3)	0	3	es	
25	2952298011	0	138875005	SNOMED Clinical Terms versión: 20120731 [R] (julio 2012)	1	2	es	
26	2583133014	1	138875005	©2002 - 2006 College of American Pathologists. SNOMED y SNOMED CT son marcas registrac				
27	2597507011	1	138875005	SNOMED CT January 2006 Release: Edición en Español 20060430 [R]	1	2	es	
28	2493010015	1	138875005	SNOMED CT July 2004 Release: Edición en Español 20041031 [R]	1	2	es	

Figura 4. Archivo original con la base de datos de SNOMED-CT.

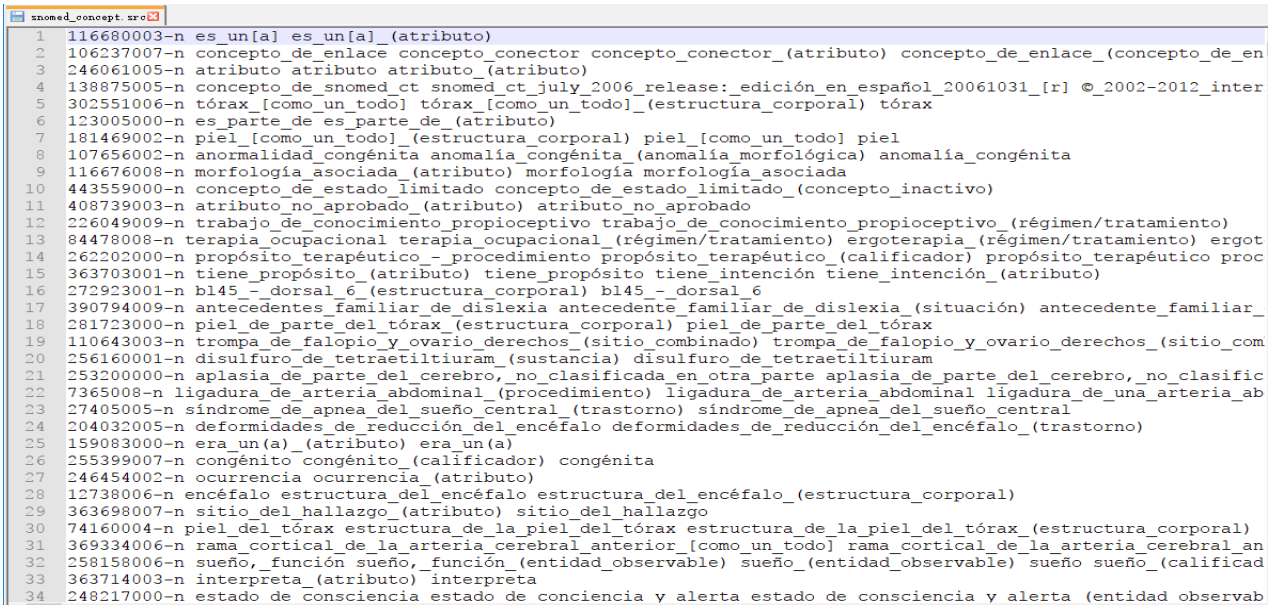
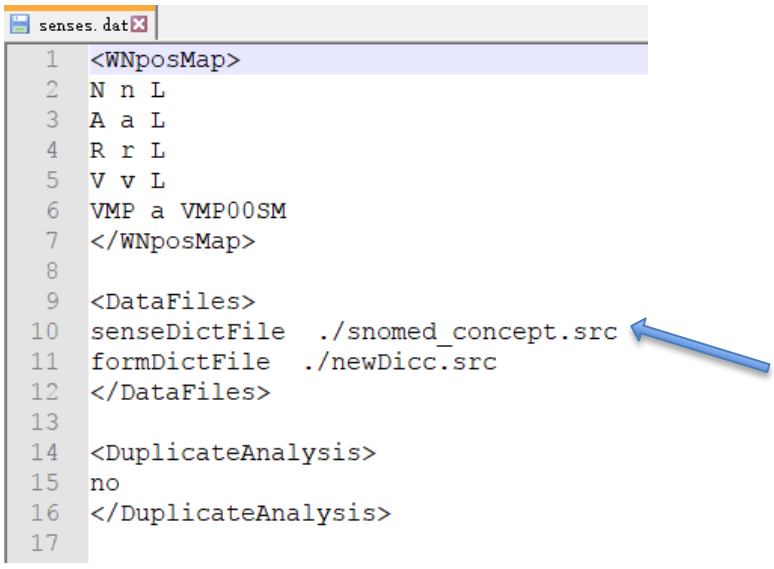


Figura 5. Archivo adaptado para poder ejecutar FreeLing.

3.4.3 Uso de FreeLing con SNOMED-CT

Una vez conseguimos un archivo con una base de datos reconocible por FreeLing, tuvimos que configurar el archivo de configuración *..\freeling-3.1-win64\data\es\senses.dat* de tal forma que ahora el procesamiento se realice utilizando el nuevo archivo obtenido. Para ello hay que indicarle que realice el procesamiento en base a la ontología SNOMED-CT añadiendo ***./snomed_concept.src*** en el apartado *senseDictFile* de la sección *DataFiles*.



Una vez se procesan los informes, se genera una representación en formato XML. Esta representación debe ser transformada al formato JSON puesto que es el formato que se necesita para realizar la búsqueda de similitud mediante ElasticSearch.

En la representación se aprecian los conceptos SNOMED-CT identificados por la etiqueta gramatical del diccionario propio de FreeLing y por el ConceptId al cual le hemos añadido anteriormente la terminación "-n". Por ejemplo, el término *tratamiento* está identificado por la etiqueta gramatical **NCMS000** que indica que es un Nombre Común Masculino Singular y por el ConceptId **116154003-n** extraído del archivo previamente obtenido.

```
<?xml version="1.0" encoding="UTF-8"?>
<contenido>
  <sexo>M</sexo>
  <edad>VACIO</edad>
  <fecha>02/06/2011</fecha>
  <motivo_ingreso>Paciente paciente NP00000 0.184242 116154003-n
tratamiento tratamiento NCMS000 1 276239002-n
catarata catarata NCFS000 1 62795009-n</motivo_ingreso>
  <alergias />
  <med_actual />
  <antecedentes_familiares />
  <antecedentes_personales>alergias alergia NCFP000 1 106190000-n
Arritmia arritmia NP00000 0.288237 53488008-n
Hipercolesterolemia hipercolesterolemia NP00000 0.288237 13644009-n
Glaucoma glaucoma NP00000 0.288237 155120009-n
tratamiento tratamiento NCMS000 1 276239002-n
ojos ojo NCMP000 1 81745001-n</antecedentes_personales>
  <anamnesis />
  <exploracion />
  <pruebas_complementarias />
  <evolucion />
  <intervencion_quirurgica>estudios estudio NCMP000 1 224699009-n
previos previo NCMP000 0.605221 9130008-n
ausencia ausencia NCFS000 1 16757004-n
facoemulsificación facoemulsificación NCFS000 1 84149000-n
implante implante NCMS000 0.918697 385235008-n
cápsula cápsula NCFS000 1 428641000-n
complicaciones complicación NCFP000 1 263718001-n</intervencion_quirurgica>
```

Figura 6. Representación, en formato XML, generada tras el procesamiento.

Comparando las salidas obtenidas procesando los informes respecto a la base de datos WordNet y a la base de datos de SNOMED-CT, hemos llegado a la conclusión de que ésta última es la más adecuada para nuestro proyecto. Aunque es cierto que ambos procesamientos extraen una información similar, el hecho de que el procesamiento en base a la ontología SNOMED-CT extraiga únicamente conceptos propiamente médicos, resulta más beneficioso para nuestra aplicación ya que a la hora de realizar la comparación por similitud la hará sobre términos puramente médicos.

Actualmente el procesado de cada representación tarda en torno a 20-40 segundos. Aunque se puede considerar como lento, el hecho de que la base de datos tenga un gran tamaño, hace que consideremos este tiempo como adecuado para este tipo de aplicación. Este hecho, unido a que la búsqueda se realiza de forma casi inmediata, dota a la aplicación final de una rapidez tal que pudiera ejecutarse casi en tiempo real.

4 Búsqueda de similitud en informes procesados

4.1 Introducción

Una vez tuvimos los informes procesados, pudimos pasar a la última parte de la aplicación, la búsqueda de similitud entre dichos informes.

Para ello, lo primero que hicimos fue convertir las representaciones en formato XML al formato JSON ya que es el formato que utiliza ElasticSearch para realizar la búsqueda. Esta conversión la conseguimos utilizando el paquete *xml2js*, el cual se encarga de transformar el formato XML al formato JSON.

Una vez tuvimos todas las representaciones en formato JSON, las almacenamos en el servidor de ElasticSearch. A cada una de ellas se les asigna, por defecto, un identificador incremental mediante el cual quedan registradas en el servidor. Pero esto supone un problema ya que, si subimos varias veces una misma representación, cada una de ellas tendrá un identificador distinto y podría llevar a engaño a la hora de realizar la comparación. Por ejemplo, si subimos una representación dos veces, por defecto se les asignará el identificador 1 y 2 y a la hora de mostrar el resultado, se muestra la similitud obtenida para ambos identificadores aun tratándose de la misma representación. Por este motivo, decidimos asignar directamente nosotros el identificador a cada representación. Este identificador que nosotros asignamos corresponde a la posición que ocupa cada informe médico en el conjunto de carpetas que nos fueron facilitadas. Por otro lado, almacenamos también la ruta de cada informe, de cara a mostrar en el resultado final a qué grupo y carpeta pertenecen.

Además, inicialmente la herramienta ElasticSearch indexaba todos los documentos subidos mediante un mismo índice. Con el fin de aumentar la precisión de búsqueda por similitud, decidimos dividir cada representación por campos. Estos campos se corresponden con las secciones médicas en las que fueron estructuradas las representaciones obtenidas tras el pre-procesado de los informes médicos. Este hecho permite además modificar la búsqueda de tal forma que solamente se busquen similitudes en determinados campos.

Una vez tuvimos cada informe indexado, dividido por campos e identificado unívocamente, pudimos comenzar la búsqueda.

4.2 Búsqueda

Para llevar a cabo la búsqueda, utilizamos la interfaz web desarrollada en la que el usuario selecciona el archivo procesado y comienza a realizar la búsqueda. Esta búsqueda se ejecuta mediante la función More Like This Query de ElasticSearch, la cual se encarga de buscar todas las coincidencias presentes por cada campo en los que dividimos cada representación.

Como resultado de la búsqueda, se crea un valor numérico por cada caso médico. Este valor conocido como *score*, es la media de los resultados de todas las búsquedas por cada campo en todos los documentos y representa la similitud entre un caso y el que ha sido seleccionado por el usuario.

Este valor score se ordena de mayor a menor por lo que, a la hora de mostrar el resultado, el primer resultado será la representación más similar al archivo seleccionado. También cabe destacar que para evitar confundir al usuario no se mostrará en el resultado el propio archivo seleccionado ya que éste tendría un porcentaje de similitud del 100%.

El resultado a mostrar se estructura de la siguiente manera:

- Nombre del archivo coincidente.
- Grupo y carpeta al que pertenece el archivo coincidente.
- Porcentaje de similitud obtenido.

4.2.1 Búsqueda de similitud con FreeLing

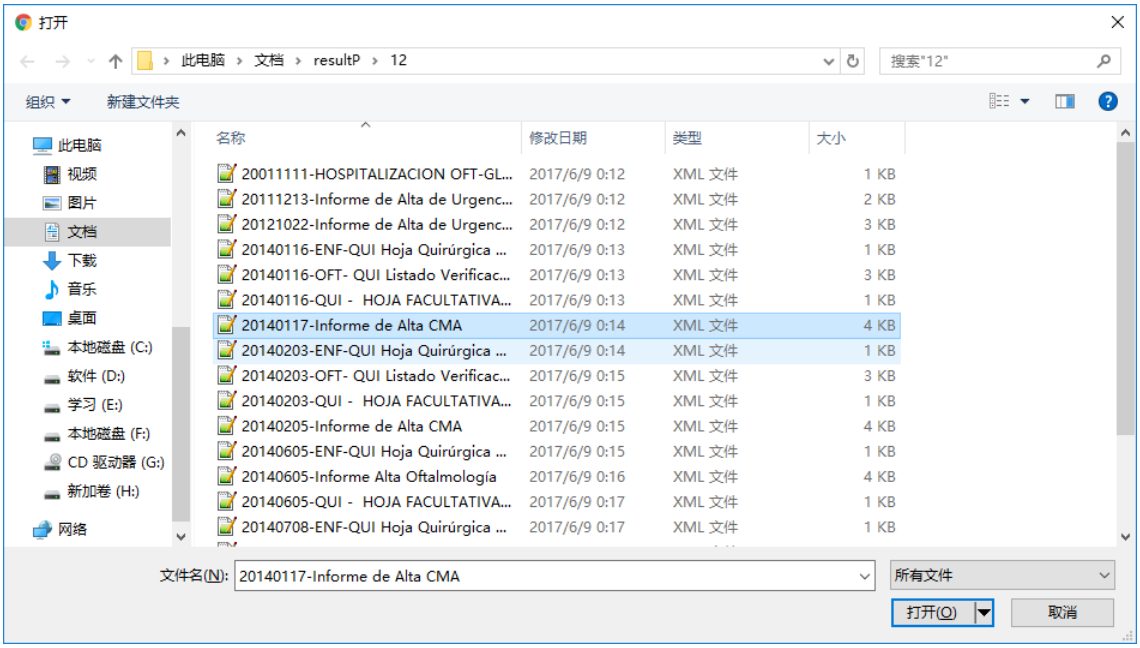


Figura 7. Representación en XML seleccionada para realizar la búsqueda.

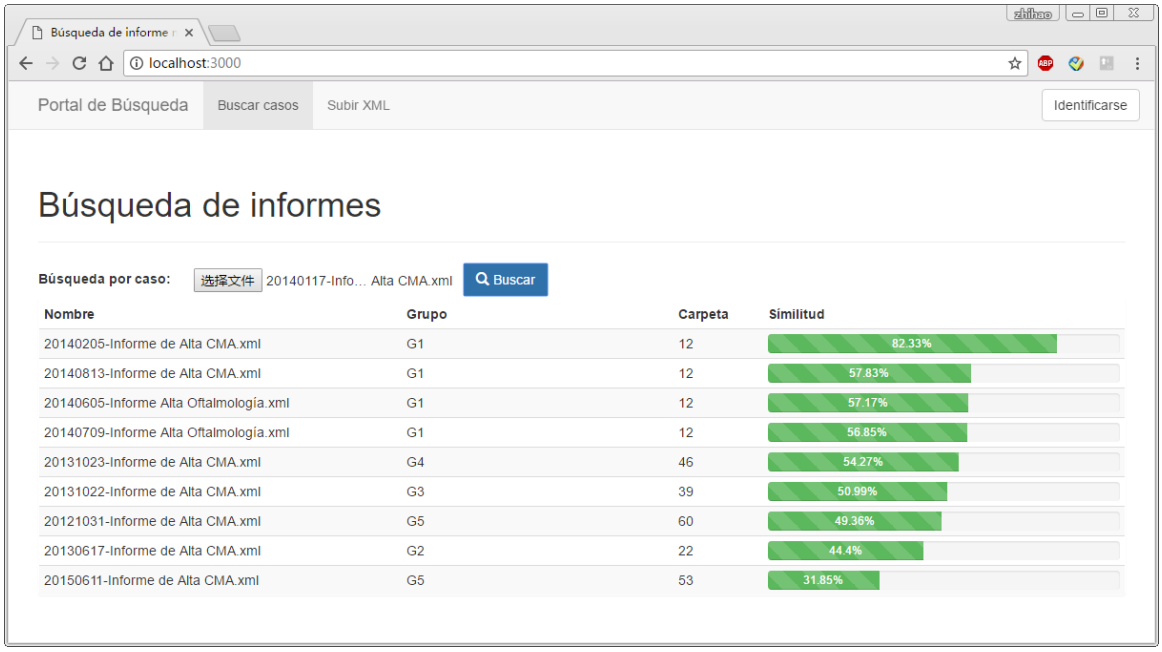


Figura 8. Resultado obtenido tras la realización de la búsqueda en base a la ontología SNOMED-CT.

4.2.2 Búsqueda de similitud con WordNet

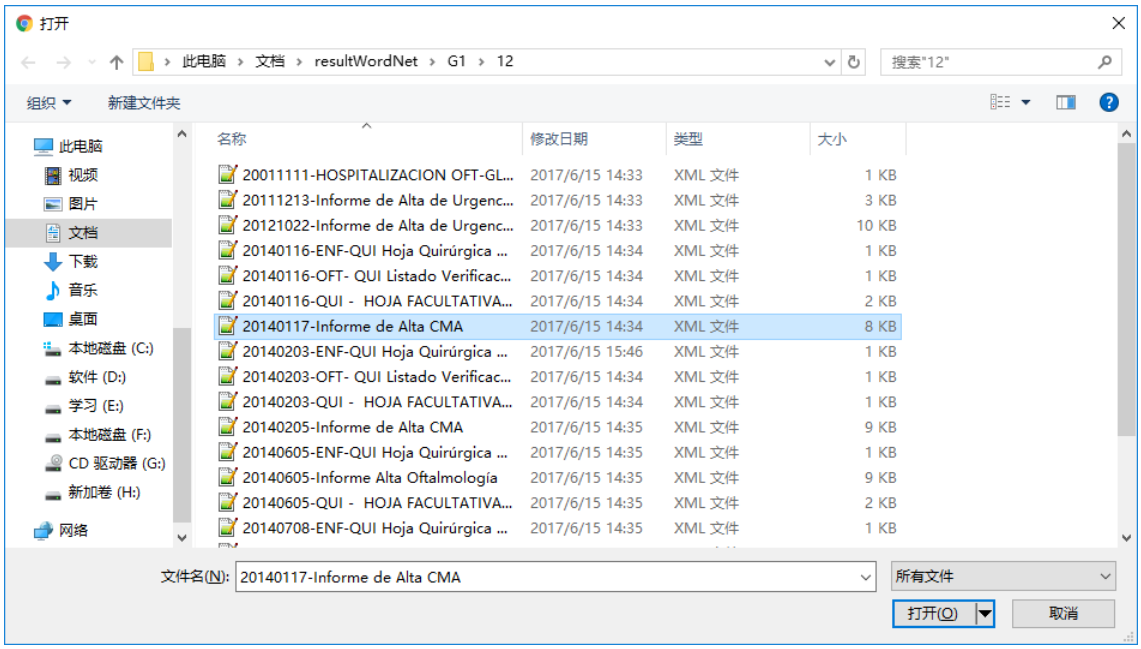


Figura 9. Representación en formato XML seleccionada para realizar la búsqueda.

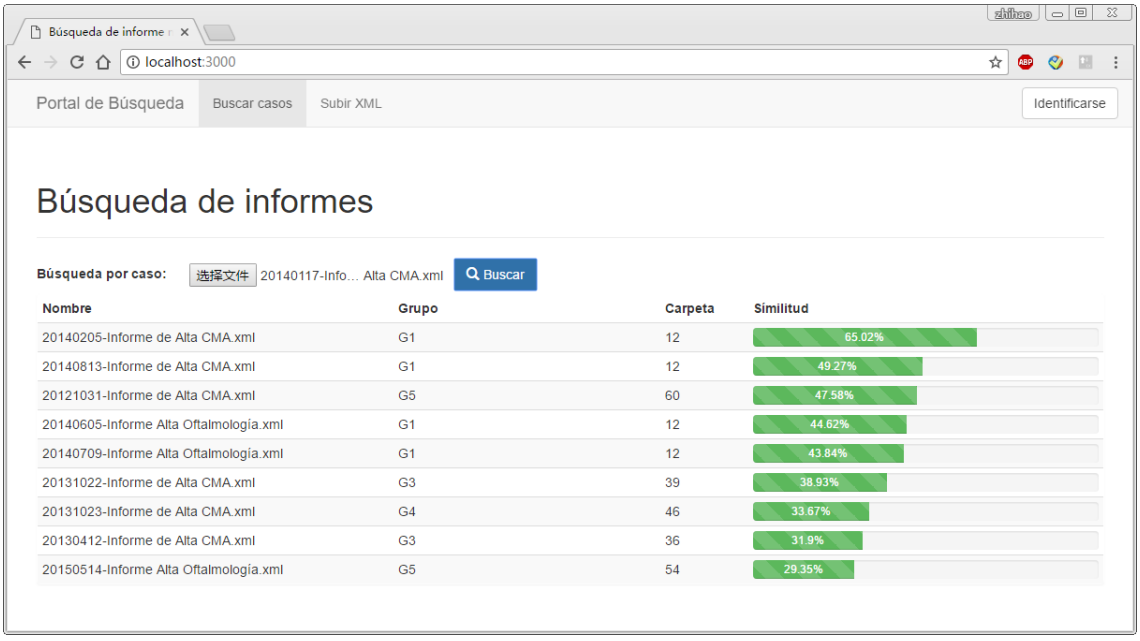


Figura 10. Resultado obtenido tras la realización de la búsqueda en base a la ontología WordNet.

4.2.3 Comparación de los resultados obtenidos

A la vista de los resultados obtenidos tras realizar la búsqueda de similitud en base a ambas ontologías, podemos concluir que se tratan de resultados completamente distintos.

Si observamos el resultado obtenido en la búsqueda de similitud de conceptos SNOMED-CT, se puede apreciar como los informes con un índice de mayor similitud corresponden a aquellos que pertenecen a un mismo grupo de enfermedad. Este hecho tiene sentido ya que significa que todos esos informes tienen un gran número de conceptos SNOMED-CT en común y es lógico puesto que están contenidos en el mismo grupo de enfermedad.

Por otro lado, si nos fijamos en el resultado conseguido tras la búsqueda de similitud de conceptos WordNet, podemos observar como los informes más similares corresponden a aquellos que comparten una estructura similar. Como se observa en el ejemplo, el archivo seleccionado es un informe de Alta y todos los resultados obtenidos se refieren a informes de Alta también. Esto indica que la similitud se basa más en la estructura de los informes, pues todos son informes de Alta, que en la información que realmente contiene cada informe.

Por tanto, podemos concluir que, para conseguir llevar a cabo el objetivo de nuestro proyecto, es más recomendable utilizar la búsqueda de similitud basada en la ontología SNOMED-CT, ya que realmente se basa en el contenido de cada informe, es decir, realiza la búsqueda en base a la terminología médica desechando el resto de términos que aparecen.

5 Conclusiones y trabajos futuros

5.1 Conclusiones

Una vez hemos concluido el desarrollo de este proyecto académico llevado a cabo durante todo el curso, podemos concluir que hemos conseguido alcanzar los objetivos. En primer lugar, hemos sido capaces de corregir aquellas limitaciones que tenía el proyecto que nos servía de base.

Hemos construido un proyecto dividido en dos aplicaciones. La primera de ella es capaz de leer los informes médicos facilitados, dividirlos por secciones médicas y procesarlos para obtener la mayor información posible en base a diferentes ontologías, como WordNet y SNOMED-CT. Esta aplicación puede resultar muy interesante ya que simplemente cambiando la configuración de la aplicación puede realizar procesados en base a múltiples ontologías. Además, el uso de la herramienta FreeLing es de gran ayuda ya que evita el uso de un traductor externo, disminuyendo la complejidad final de la aplicación.

La otra parte del proyecto, consiste en una interfaz web que es capaz de realizar búsqueda entre todos los informes procesados de una forma muy rápida y precisa gracias a que dichos informes contienen la información puramente necesaria en cuanto a la ontología médica SNOMED-CT. El hecho de que sea una aplicación web, permite a cualquier usuario hacer uso de ella desde cualquier sitio y realizar una localización de informes médicos similares en un escaso plazo de tiempo.

5.2 Trabajos futuros

Este apartado está dedicado a hablar de posibles mejoras de este proyecto, corrigiendo aquellos errores que han surgiendo durante el mismo, así como comentar algunas aplicaciones que pueden surgir teniendo de base este proyecto.

5.2.1 Mejora en la adaptación de SNOMED-CT a FreeLing

Como se ha comentado en el apartado 3.4.2 que a la hora de adaptar el archivo con la base de datos de SNOMED-CT, se han considerado todos los conceptos como nombres, añadiendo la terminación “-n” al ConceptId. Consideramos que una mejora importante podría ser conseguir añadir la terminación adecuada a cada concepto, de manera que se diferenciara si se trata de un nombre, verbo, adjetivo o adverbio. Este hecho dotaría a la aplicación de mayor precisión a la hora de realizar las búsquedas por similitud.

5.2.2 Mejora en el tiempo de procesamiento

Como se ha comentado en el apartado 3.4.3 el tiempo actual de procesamiento de cada representación gira en torno a los 20-40 segundos. Aunque puede parecer un tiempo rápido, este tipo de aplicaciones necesita que las funcionalidades se ejecuten en el menor tiempo posible. Es por ello, que consideramos que un trabajo futuro podría ser mejorar dicho tiempo de procesamiento de tal forma que se pueda conseguir realizar el procesamiento en pocos segundos, consiguiendo dotar a la aplicación de aún más rapidez.

5.2.3 Aplicación para dispositivos móviles

Hoy en día, el mundo de las aplicaciones para dispositivos móviles crece exponencialmente año tras año. Es por ello, que se podría aprovechar lo hecho en este proyecto y trasladarlo al mundo de las aplicaciones, creando una aplicación para Android/iOS y ponerla a prueba en algún centro médico para comprobar su efectividad.

6 Bibliografía

1. Beyad, A. *ElasticSearch. Open Source, Distributed, RESTful Search Engine* (GitHub). Disponible en: https://github.com/elastic/elasticsearch
2. Casillas, A; Díaz de Ilarraza, A; Gojenola, K; Mendarte, L; Oronoz, M; Peral, J; Pérez, A. 2016. "Deteami research-transference Project: natural language processing technologies to the aid of pharmacy and pharmacosurveillance" en *Procesamiento del Lenguaje Natural,* nº57, pp. 155-158. Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5351
3. Elastic. *ElasticSearch Reference [5.4].* Disponible en: https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html
4. Ministerio de Sanidad, Servicios Sociales e Igualdad (Gobierno de España). *SNOMED-CT*. Disponible en: http://www.msssi.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/home.htm
5. Oronoz, M; Casillas, A; Gojenola, K; Perez, A. 2013. "Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names" en *Research Gate.* Disponible en: https://www.researchgate.net/publication/263651479_Automatic_Annotation_of_Medical_Records_in_Spanish_with_Disease_Drug_and_Substance_Names
6. Oronoz, M; Díaz de Ilarraza, A; Torices, O. 2010. "First Steps in The Manual and Automatic Annotation of Clinical Notes in Spanish" en *Procesamiento del Lenguaje Natural*, nº 45, pp. 259-262. Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/815
7. Padró, L. 2011. "Analizadores Multilingües en FreeLing" en *Freeling Home Page.* Disponible en: http://nlp.cs.upc.edu/freeling/
8. Padró, L. *FreeLing* (GitHub). Disponible en: https://github.com/TALP-UPC/FreeLing
9. Padró, L. *FreeLing User Manual*. Disponible en: https://talp-upc.gitbooks.io/freeling-user-manual/content/
10. Princeton University. *WordNet. A lexical database for English*. Disponible en: https://wordnet.princeton.edu/
11. SNOMED International. 2017. *SNOMED-CT Starter Guide*. Disponible en: https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide
12. SNOMED International. Disponible en: http://www.snomed.org/
13. The Apache Software Foundation. *Apache POI –the Java API for Microsoft Documents*. Disponible en: https://poi.apache.org/
14. U.S. National Library of Medicine. *SNOMED-CT Home*. Disponible en: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html
15. U.S. National Library of Medicine. *Unified Medical Language System (UMLS)*. Disponible en: https://www.nlm.nih.gov/research/umls/
16. Wikipedia. *Apache POI*. Disponible en: https://en.wikipedia.org/wiki/Apache_POI
17. Wikipedia. *JavaScript*. Disponible en: https://es.wikipedia.org/wiki/JavaScript
18. Wikipedia. *Lucene*. Disponible en: https://es.wikipedia.org/wiki/Lucene
19. Wikipedia. *WordNet*. Disponible en: https://es.wikipedia.org/wiki/WordNet

7 Apéndices

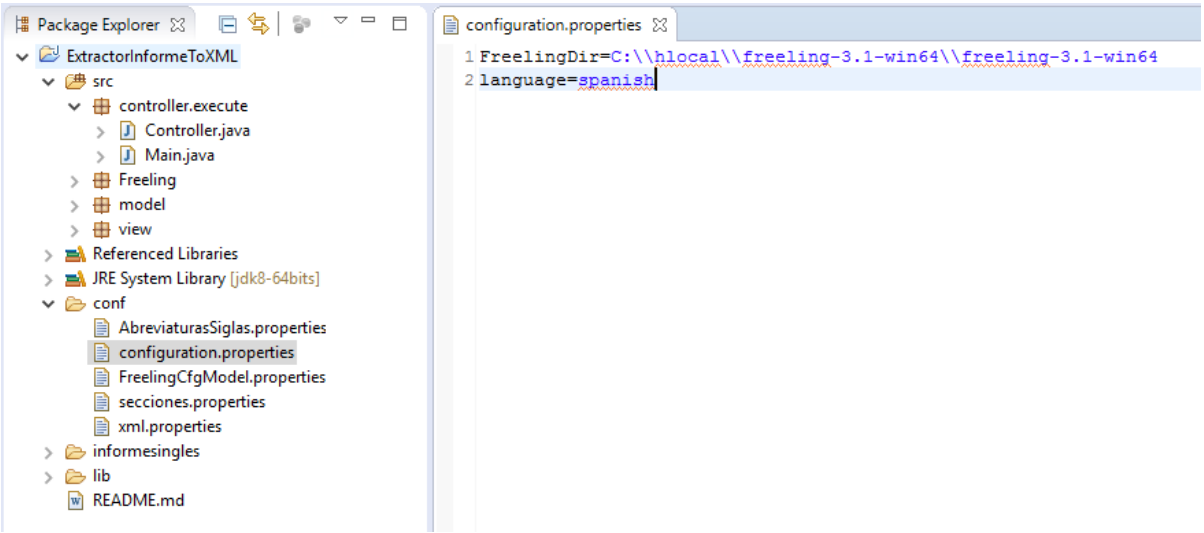
7.1 Configuración para el uso de FreeLing en Windows

Para configurar debemos ir al código de la aplicación, en la carpeta *conf*, abrir el archivo *configuration.Properties* y en el campo *FreeLingDir* añadir la ruta en la que se encuentre FreeLing.

Una vez se ejecuta la aplicación, el archivo FreelingConf.properties rellena todos sus campos con las rutas adecuadas para la ejecución de FreeLing. Esto nos ahorra cambiar la configuración del archivo es.cfg si cambiamos la ruta de ubicación de FreeLing.

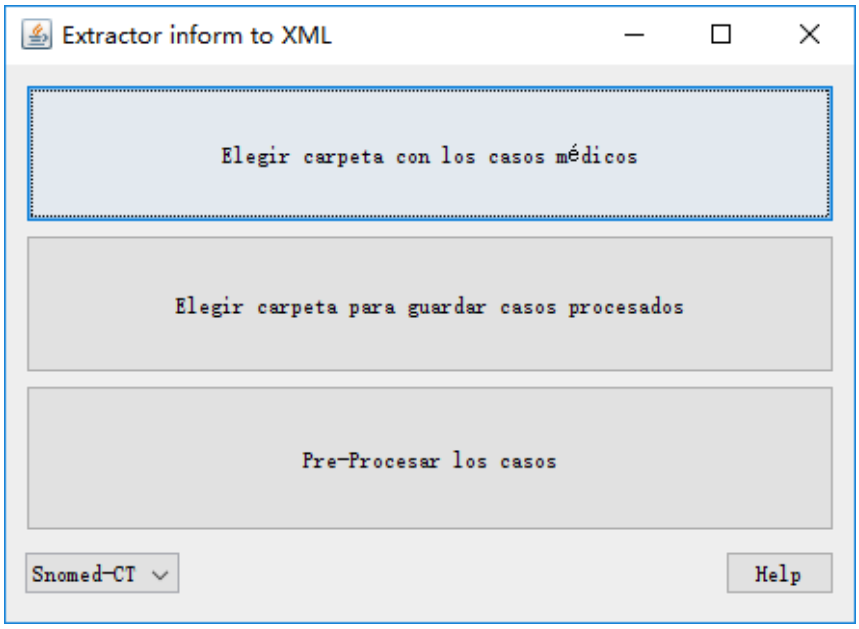
7.2 Manual para pre-procesar y procesar los informes médicos

1. Abrimos el archivo ***conf/configuration.properties*** en la ruta principal donde se encuentre la aplicación.

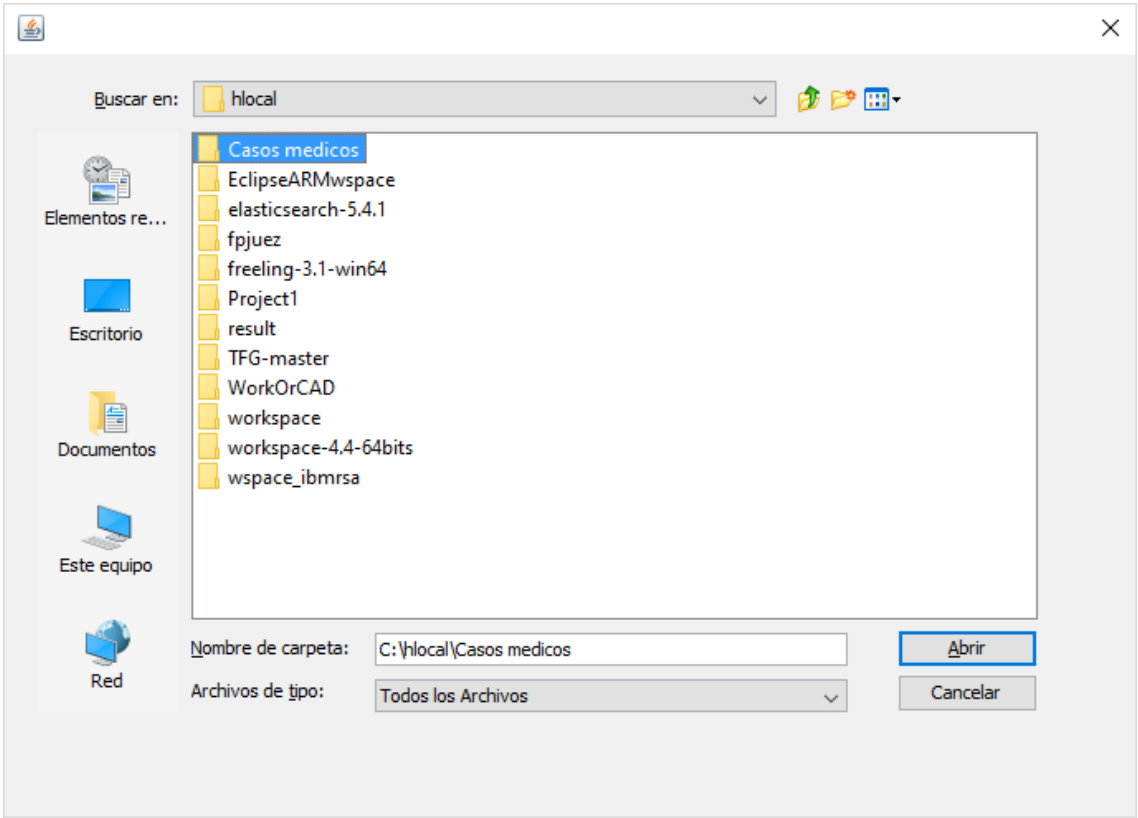


2. Cambiamos el valor del campo ***FreelingDir*** por la ruta principal donde se encuentra la herramienta Freeling, utilizando como separación entre carpetas el símbolo ”\\”.

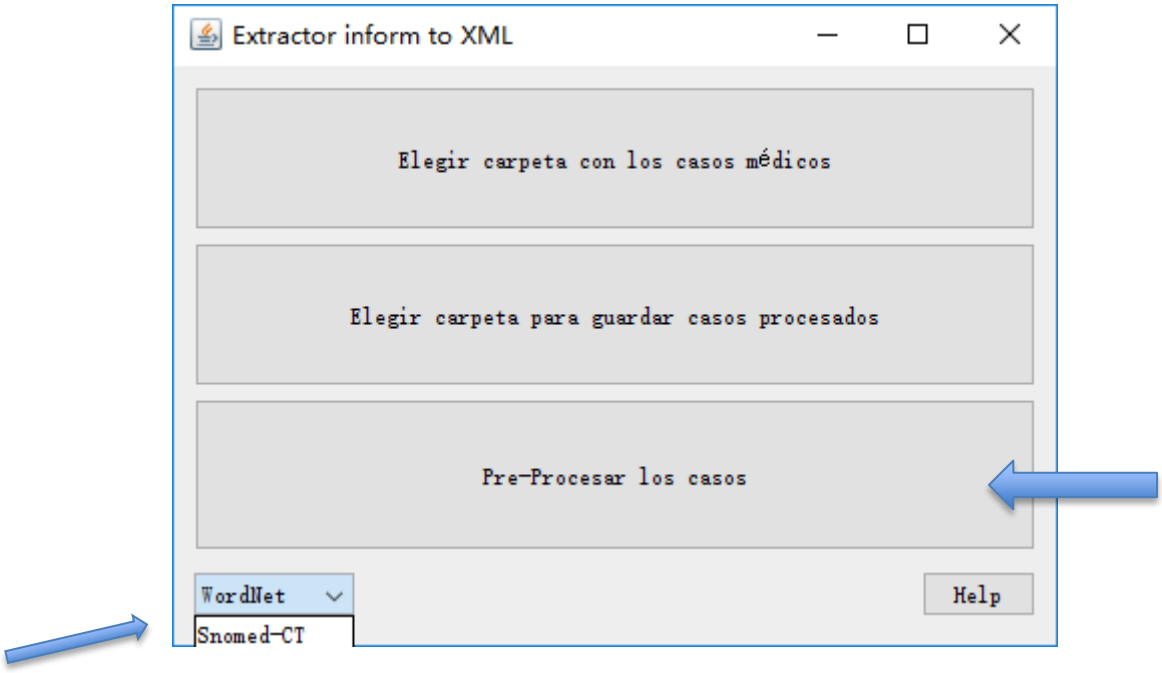
3. Ejecutamos la aplicación haciendo click derecho sobre controller.execute.Main.java y seleccionando Run as -> Java Application.



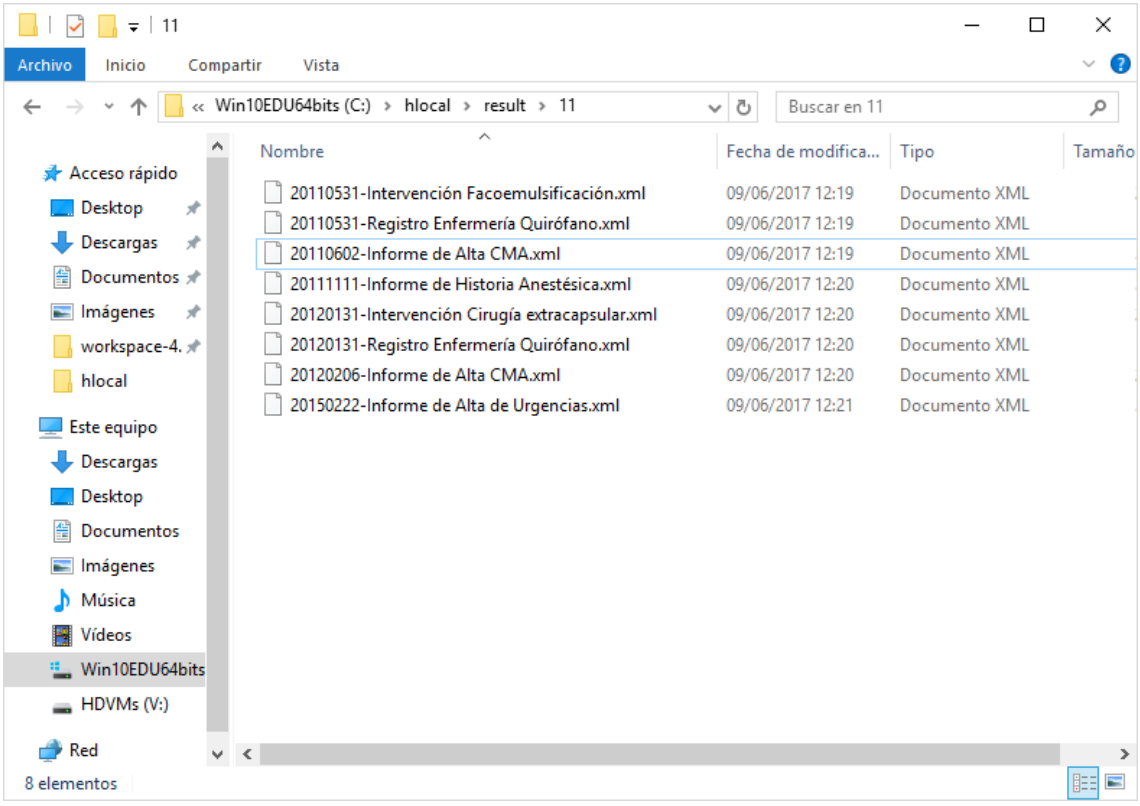
4. Seleccionar la carpeta donde se encuentra los casos médicos a procesar y la carpeta donde se desee guardar el resultado del procesamiento.



5. En el comboBox que aparece abajo a la izquierda seleccionamos la ontología que deseamos utilizar para procesar los casos y pulsamos sobre el botón **"Pre-Procesar los casos"** para comenzar con el procesamiento.

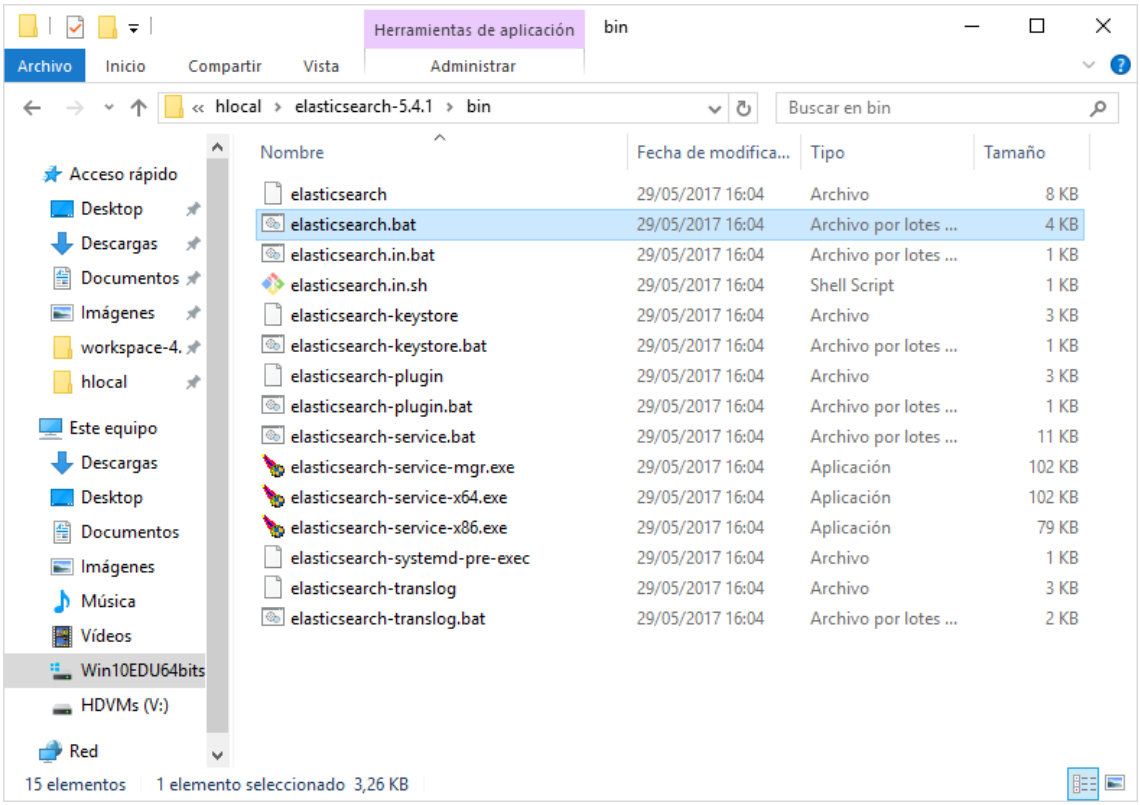


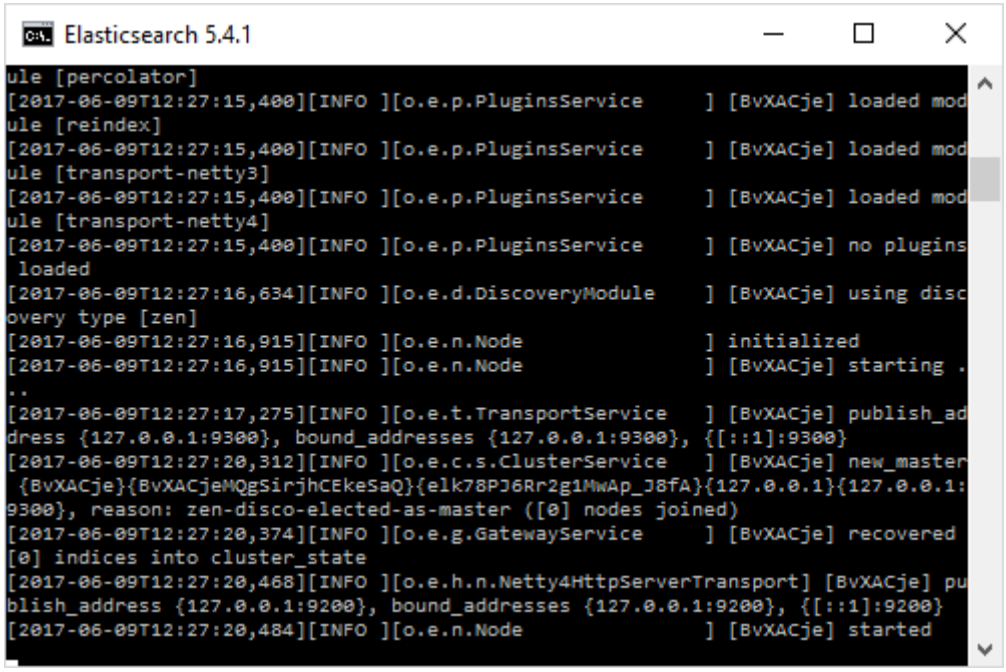
6. Las representaciones obtenidas en el procesamiento se encontrarán en la ruta previamente seleccionada.



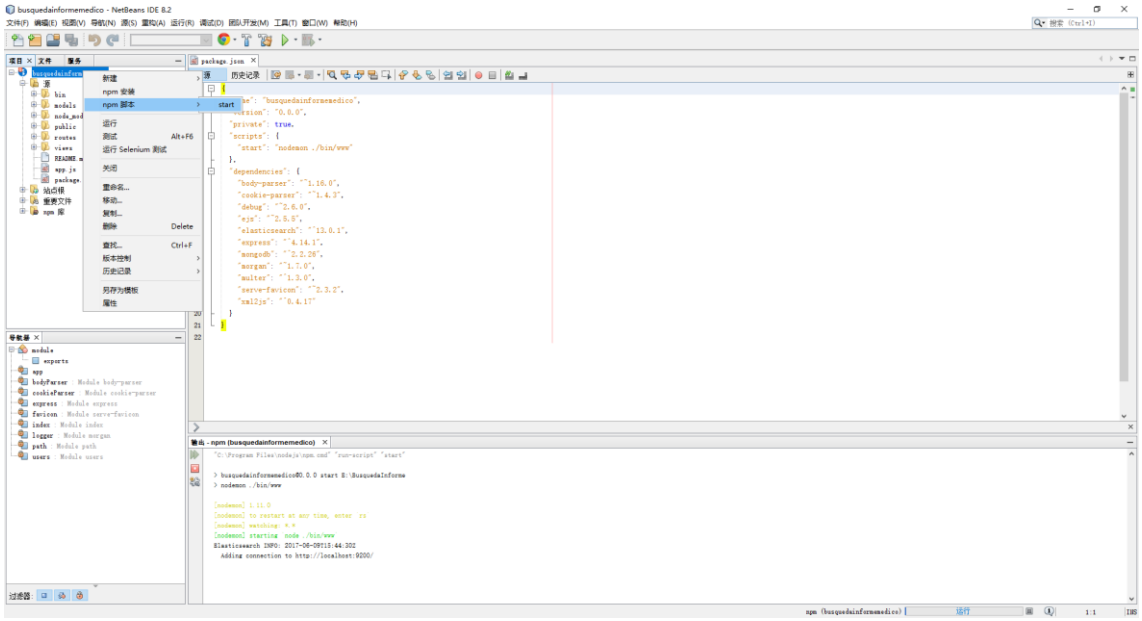
7.3 Manual para realizar la búsqueda de similitudes

1. Arrancamos el servidor de ElasticSearch disponible en la carpeta ***/elasticsearch-5.4.1/bin/elasticsearch.bat***.

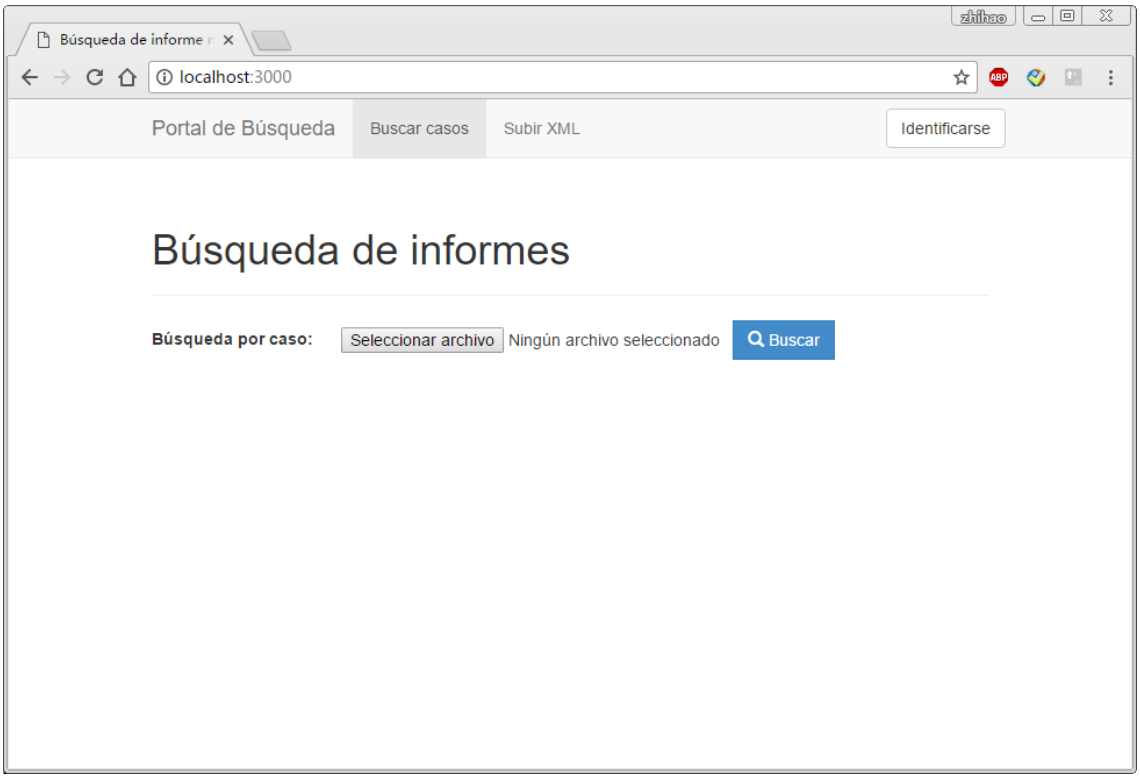




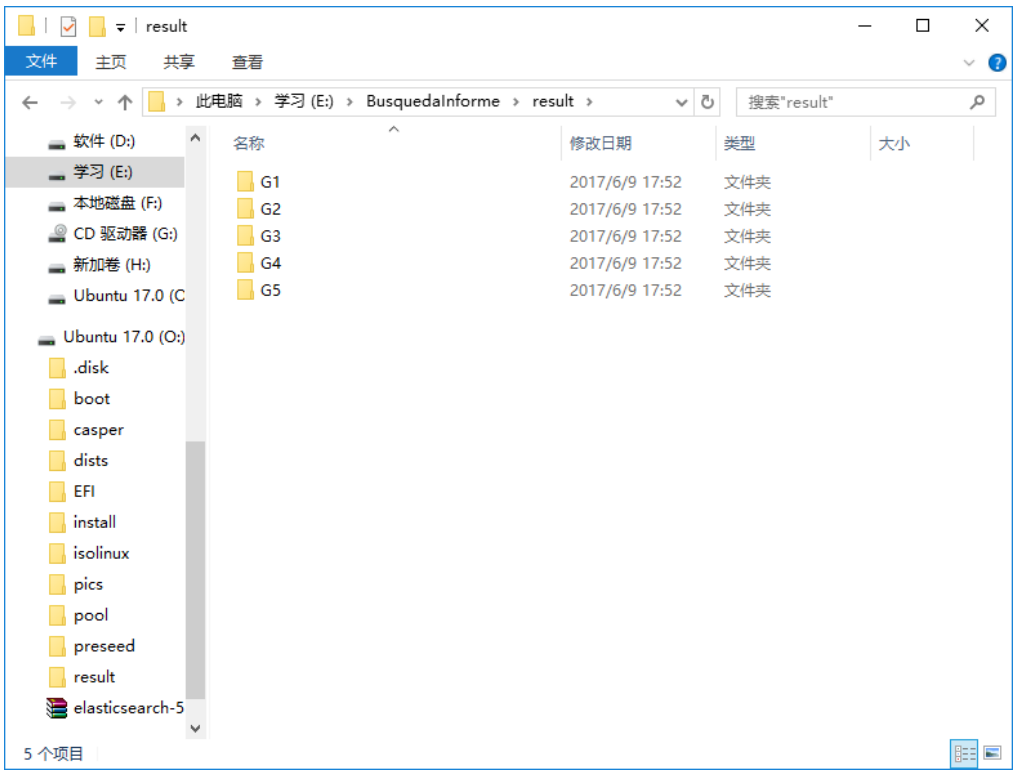
2. Ejecutamos la aplicación haciendo click derecho sobre proyecto(busquedainformemedico) y seleccionando script npm -> start.



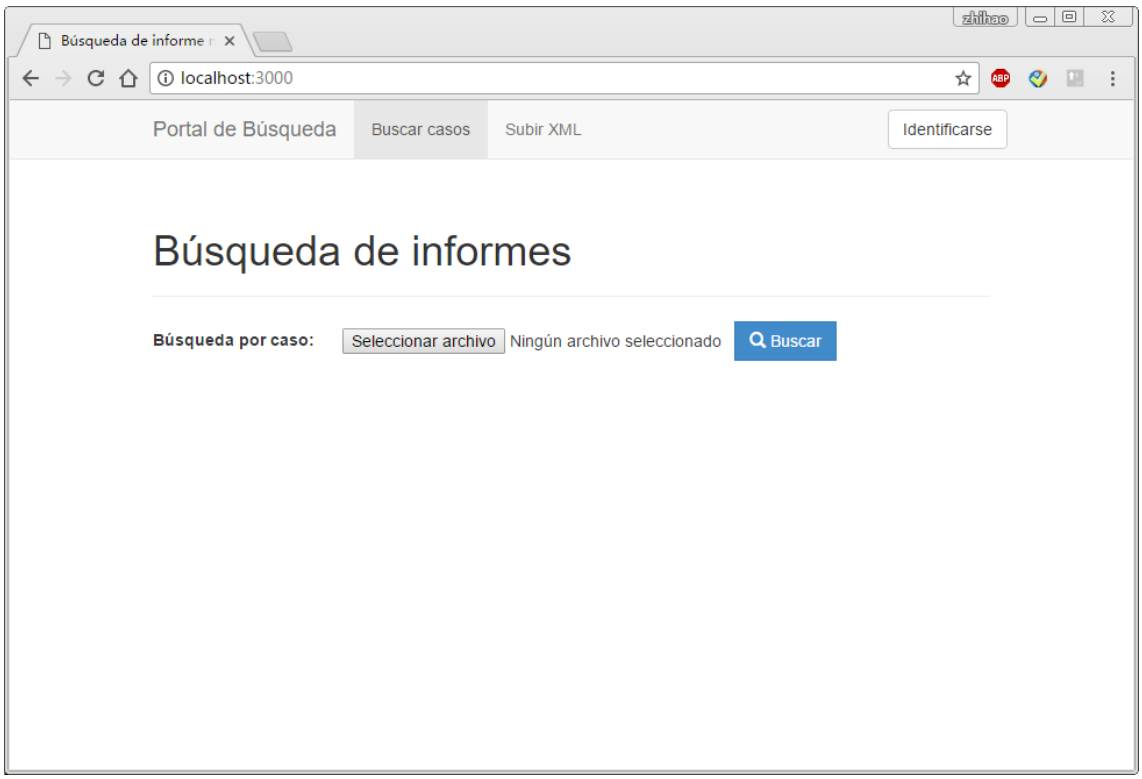
3. Abrimos la dirección **localhost:3000** en un navegador.



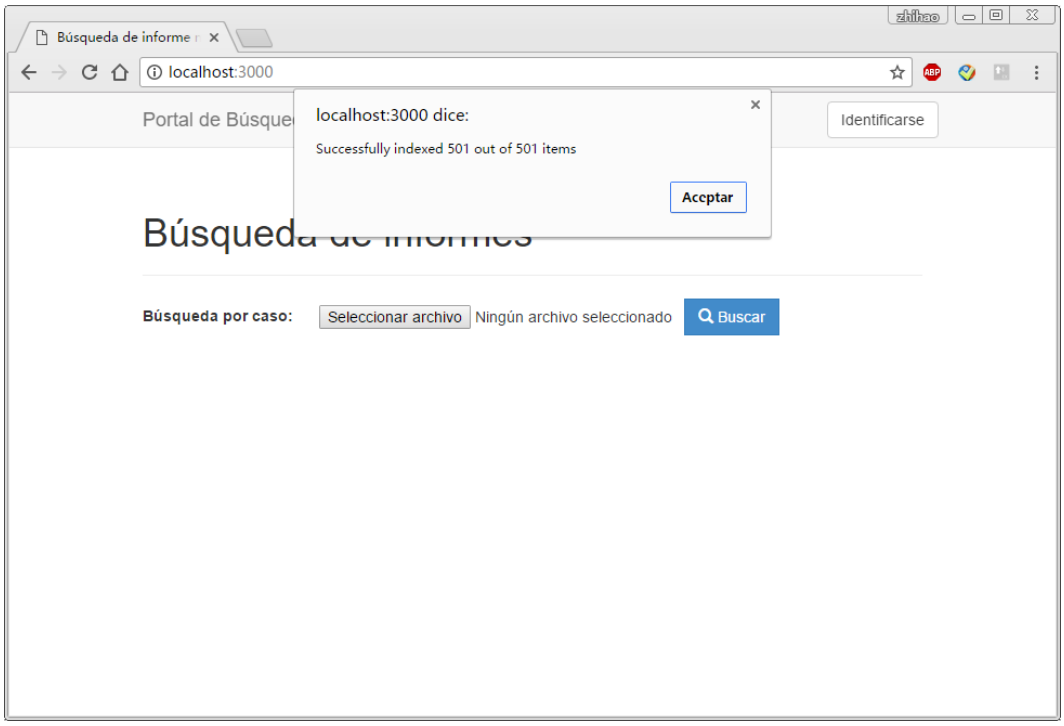
4. Movemos los casos procesados en formato xml a la carpeta result en la ruta principal de la aplicación.



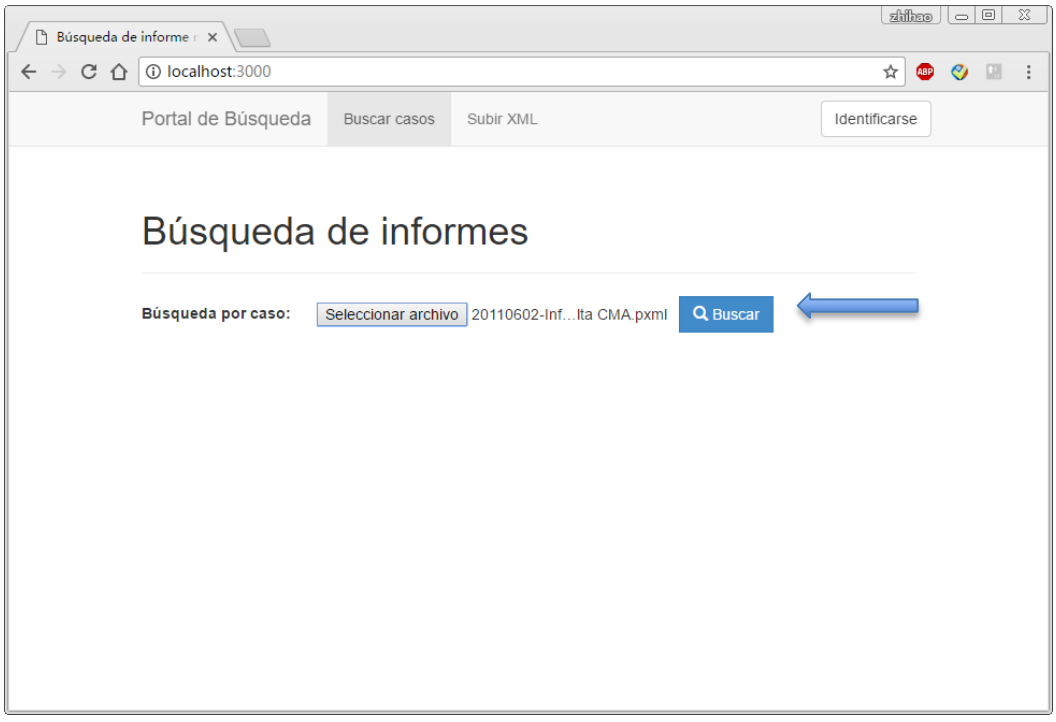
5. Hacemos click en el botón **"Subir XML"**, para subir los casos procesados al servidor.



Se muestra un mensaje confirmando que los archivos procesados han sido correctamente indexados en el servidor.



6. Hacemos click en el botón **"Seleccionar archivo"** para elegir el caso del cual queremos buscar similitudes (en formato XML) y pulsamos el botón **"Buscar"**.



7. Se muestra el resultado de la búsqueda, ordenado de mayor a menor por porcentaje de similitud.