



Universidad Politécnica de Madrid
Facultad de Informática



Trabajo de Fin de Máster
Máster Universitario en Inteligencia Artificial

**USO DE TÉCNICAS DE MINERÍA DE TEXTO PARA LA
IDENTIFICACIÓN DE ENSAYOS CLÍNICOS EN
NANOMEDICINA**

Autor: **Charles Pérez Espinoza**

Tutor:
Miguel García Remesal

Madrid, Julio 2015

**Universidad Politécnica de Madrid
Facultad de Informática**

**Trabajo de Fin de Máster
Máster Universitario en Inteligencia Artificial**

**USO DE TÉCNICAS DE MINERÍA DE TEXTO PARA LA
IDENTIFICACIÓN DE ENSAYOS CLÍNICOS EN
NANOMEDICINA**

Autor: Charles Pérez Espinoza

**Tutor:
Miguel García Remesal**

Madrid, Julio 2015

A mi familia

Ser original, no es aquella persona que
no copia a los demás, ser original es aquella
persona que aunque otros lo intenten
jamás podrán ser como el

Charles Perez
2004

AGRADECIMIENTOS

Estoy muy agradecido principalmente con Dios Todopoderoso, quien me ha dado la inteligencia, la fuerza y la constancia para lograr cada uno de mis retos, yo como extranjero tuve inconvenientes que me ayudaron a mejorar cada día, y sé que los supere porque Dios estuvo conmigo en todo momento.

Agradezco a mi tutor el Phd. Miguel García por creer en que haría un gran trabajo, por su paciencia, y su constancia cuando necesitaba aclarar una duda o cuando tenía un problema con el tema propuesto, realmente es una excelente persona y un excelente profesor.

Además no hubiera sido posible llegar hasta esta meta, si no fuera por las increíbles personas que tengo a mi lado, ayudándome día a día a ser mejor, mi estabilidad emocional, sentimental e intelectual se los debo a ellos.

Como es mi padre Carlos, a la cual le debo mi parte fuerte y mi constancia de seguir trabajando para mejorar, a mi madre Carmen, a la cual le debo mi parte sentimental y mi acercamiento a Dios, a mi hermana por sus frases de no rendirme y su total atención cuando yo estaba en problemas, a mi hermano por ser mi consejero y que fue uno de los factores importantes por el cual elegí venir a España a estudiar mi máster, a mi amada Paola por ser siempre mi fuente de apoyo, por ayudarme a descargar los ensayos para este trabajo, por ser mi paz en todo momento, que aunque hayamos estado tan lejos ha sido mi estabilidad para seguir luchando. Gracias a todos ellos por confiar en mis decisiones, y lo más importante gracias por sus constantes oraciones.

Agradezco también a mis sobrinos porque cuando los veía por cámara, cuando los saludaba me daban una razón más para terminar esta meta, ya que se que ellos se iban a sentir orgullosos de mi y así convertirme en un buen ejemplo para ellos.

A mis mejores amigos del grupo Nessit que están en mi país, ya que siempre estuvieron apoyándome así sea con un mensaje haciéndome reír o con unas palabras de aliento para seguir adelante. A mis nuevos amigos que conocí durante mi estadía en España como son Carlos y Lilian que me ayudaron a mantenerme tranquilo, y en toda circunstancia jamás me dejaron solo, gracias por la unión y la increíble amistad que me brindaron, eso no lo olvidare jamás.

Madrid, Julio 2015

Charles Perez

RESUMEN

La nanotecnología es el estudio que la mayoría de veces es tomada como una meta tecnológica que nos ayuda en el área de investigación para tratar con la manipulación y el control en forma precisa de la materia con dimensiones comprendidas entre 1 y 100 nanómetros. Recordando que el prefijo nano proviene del griego *nano* que significa enano y corresponde a un factor de 10^{-9} , que aplicada a las unidades de longitud corresponde a una mil millonésima parte de un metro. Ahora sabemos que esta ciencia permite trabajar con estructuras moleculares y sus átomos, obteniendo materiales que exhiben fenómenos físicos, químicos y biológicos, muy distintos a los que manifiestan los materiales usados con una longitud mayor. Por ejemplo en medicina, los compuestos nanométricos y los materiales nano estructurados muchas veces ofrecen una mayor eficacia con respecto a las formulaciones químicas tradicionales, ya que muchas veces llegan a combinar los antiguos compuestos con estos nuevos para crear nuevas terapias e inclusive han llegado a reemplazarlos, revelando así nuevas propiedades diagnósticas y terapéuticas. A su vez, la complejidad de la información a nivel nano es mucho mayor que en los niveles biológicos convencionales y, por tanto, cualquier flujo de trabajo en nano medicina requiere, de forma inherente, estrategias de gestión de información avanzadas.

Muchos investigadores en la nanotecnología están buscando la manera de obtener información acerca de estos materiales nanométricos, para mejorar sus estudios que muchas veces lleva a probar estos métodos o crear nuevos compuestos para ayudar a la medicina actual, contra las enfermedades más poderosas como el cáncer. Pero en estos días es muy difícil encontrar una herramienta que les brinde la información específica que buscan en los miles de ensayos clínicos que se suben diariamente en la web. Actualmente, la informática biomédica trata de proporcionar el marco de trabajo que permita lidiar con estos retos de la información a nivel nano, en este contexto, la nueva área de la nano informática pretende detectar y establecer los vínculos existentes entre la medicina, la nanotecnología y la informática, fomentando así la aplicación de métodos computacionales para resolver las cuestiones y problemas que surgen con la información en la amplia intersección entre la biomedicina y la nanotecnología.

Otro caso en la actualidad es que muchos investigadores de biomedicina desean saber y comparar la información dentro de los ensayos clínicos que contiene temas de nanotecnología en las diferentes paginas en la web por todo el mundo, obteniendo en si ensayos clínicos que se han creado en Norte América, y ensayos clínicos que se han creado en Europa, y saber si en este tiempo este campo realmente está siendo explotado en los dos continentes. El problema es que no se ha creado una herramienta

que estime un valor aproximado para saber los porcentajes del total de ensayos clínicos que se han creado en estas páginas web.

En esta tesis de fin de máster, el autor utiliza un mejorado pre-procesamiento de texto y un algoritmo que fue determinado como el mejor procesamiento de texto en una tesis doctoral, que incluyo algunas pruebas con muchos de estos para obtener una estimación cercana que ayudaba a diferenciar cuando un ensayo clínico contiene información sobre nanotecnología y cuando no. En otras palabras aplicar un análisis de la literatura científica y de los registros de ensayos clínicos disponibles en los dos continentes para extraer información relevante sobre experimentos y resultados en nano medicina (patrones textuales, vocabulario en común, descriptores de experimentos, parámetros de caracterización, etc.), seguido el mecanismo de procesamiento para estructurar y analizar dicha información automáticamente. Este análisis concluye con la estimación antes mencionada necesaria para comparar la cantidad de estudios sobre nanotecnología en estos dos continentes. Obviamente usamos un modelo de datos de referencia (gold standard) —un conjunto de datos de entrenamiento anotados manualmente—, y el conjunto de datos para el test es toda la base de datos de estos registros de ensayos clínicos, permitiendo distinguir automáticamente los estudios centrados en nano drogas, nano dispositivos y nano métodos de aquellos enfocados a testear productos farmacéuticos tradicionales.

ABSTRACT

Nanotechnology is the scientific study that usually is seen as a technological goal that helps us in the investigation field to deal with the manipulation and precise control of the matter with dimensions that range from 1 to 100 nanometers. Remembering that the prefix nano comes from the Greek word *νάνοσ*, meaning dwarf and denotes a factor of 10^{-9} , that applied the longitude units is equal to a billionth of a meter. Now we know that this science allows us to work with molecular structures and their atoms, obtaining material that exhibit physical, chemical and biological phenomena very different to those manifesting in materials with a bigger longitude. As an example in medicine, the nanometric compounds and the materials in nano structures are often offered with more effectiveness regarding to the traditional chemical formulas. This is due to the fact that many occasions combining these old compounds with the new ones, creates new therapies and even replaced them, revealing new diagnostic and therapeutic properties. Even though the complexity of the information at nano level is greater than that in conventional biologic level and, thus, any work flow in nano medicine requires, in an inherent way, advance information management strategies.

Many researchers in nanotechnology are looking for a way to obtain information about these nanometric materials to improve their studies that leads in many occasions to prove these methods or to create a new compound that helps modern medicine against powerful diseases, such as cancer. But in these days it is difficult to find a tool that searches and provides a specific information in the thousands of clinic essays that are uploaded daily on the web. Currently, the bio medic informatics tries to provide the work frame that will allow to deal with these information challenge in nano level. In this context, the new area of nano informatics pretends to detect and establish the existing links between medicine, nanotechnology and informatics, encouraging the usage of computational methods to resolve questions and problems that surge with the wide information intersection that is between biomedicine and nanotechnology.

Another present case, is that many biomedicine researchers want to know and be able to compare the information inside those clinic essays that contains subjects of nanotechnology on the different webpages across the world, obtaining the clinic essays that has been done in North America and the essays done in Europe, and thus knowing if in this time, this field is really being exploited in both continents.

In this master thesis, the author will use an enhanced text pre-processor with an algorithm that was defined as the best text processor in a doctoral thesis, that included many of these tests to obtain a close estimation that helps to differentiate when a clinic essay contains information about nanotechnology and when it does not. In other words, applying an analysis to the scientific literature and clinic essay available in both continents, in order to extract relevant information about experiments and the results in nano-medicine (textual patterns, common vocabulary, experiments descriptors, characterization parameters, etc.), followed by the mechanism process to structure and analyze said information automatically. This analysis concludes with the estimation, mentioned before, needed to compare the quantity of studies about nanotechnology in these two continents. Obviously we use a data reference model

(Gold standard) – a set of training data manually annotated –, and the set of data for the test conforms the entire database of these clinic essay registers, allowing to distinguish automatically the studies centered on nano drugs, nano devices and nano methods of those focus on testing traditional pharmaceutical products.

TABLA DE CONTENIDO

1 INTRODUCCION	1
2 OBJETIVOS Y ALCANCE	6
2.1 Objetivos.....	7
2.2 Alcance	8
2.3 Organización del trabajo realizado	9
3 ESTADO DEL ARTE.....	11
3.1 Nanotecnología	11
3.1.1 Nanoinformática	18
3.2 Minería de texto	22
3.2.1 Algoritmos clasificadores usados en aprendizaje automático.....	24
3.2.2 Herramienta WEKA.....	28
3.2.3 Librería LibLINEAR	30
3.3 Minería de texto en nanomedicina.....	32
3.4 Ensayos clínicos.....	35
3.4.1 Ensayos clínicos sobre nanopartículas.....	36
3.4.2 Ensayos clínicos sobre nanomateriales.....	40
3.4.3 Ensayos clínicos sobre nanodispositivos	42
3.4.4 Ensayos clínicos no nanos	43
3.5 Repositorios de registros de ensayos clínicos	44
3.5.1 Clinical Trials (Estados Unidos)	46
3.5.2 Clinical Trials Register (Unión Europea).....	50
4 METODOLOGÍA	54
4.1 Obtención de los recursos	54
4.1.1 Obtención del repositorio para el entrenamiento del algoritmo	54
4.1.2 Obtención de los repositorios para el ejecutar el algoritmo	57
4.2 Ensayos clínicos.....	61
4.3 Pre-procesamiento de texto.....	61
4.3.1 Tokenización	64
4.3.2 Reemplazo de dígitos	65

4.3.3 Texto en minúsculas.....	67
4.3.4 Eliminación de Stopwords.....	67
4.3.5 Eliminación de características textuales cortas.....	68
4.3.6 Obtención de frecuencias por documento.....	68
4.3.7 Obtención de la suma de frecuencias.....	69
4.3.8 Primera regla para la eliminación de palabras.....	70
4.3.9 Obtención de los IDF de los repositorios.....	71
4.3.10 Segunda regla para la eliminación de palabras.....	72
4.4 Herramienta creada para hacer el Pre-procesamiento.....	74
4.5 Desarrollo del modelo.....	78
4.6 Aplicación del modelo y el algoritmo establecido.....	80
5 RESULTADOS.....	82
5.1 Resultados para la aplicación del modelo de entrenamiento.....	82
5.2 Resultados de la base de ClinicalTrials.gov (Norte América).....	88
5.3 Resultados de la base de ClinicalTrialsRegister.eu (Europa).....	89
6 DISCUSIÓN.....	90
7. CONCLUSIÓN Y TRABAJOS FUTUROS.....	94
7.1 Conclusiones.....	94
7.2 Trabajo futuro.....	95
ANEXO I: LISTA DE ENSAYOS CLÍNICOS SEPARADOS MANUALMENTE EXTRAIDOS DE CLINICALTRIALS.GOV.....	97
ANEXO II: LISTA DE LAS STOPWORDS DE PUBMED HASTA EL 2015.....	103
ANEXO III: LISTA DE UNIGRAMS USADOS PARA CLASIFICAR LAS ETIQUETAS NANO Y NO NANO.....	105
REFERENCIAS.....	123

INDICE DE FIGURAS

Figura 1.1. Comparación de las longitudes en nanómetros.....	1
Figura 1.2. Explicación gráfica del concepto de nanoinformática.....	3
Figura 1.3. Aplicación de nanoinformática en ensayos clínicos de diferentes continentes.....	4
Figura 2.1. Folleto del Grupo de Biomedicina Informática de la UPM.....	6
Figura 2.2. El alcance de esta TFM: Clasificación de los ensayos clínicos.....	10
Figura 3.1. Logo de la National Nanotechnology Initiative.....	11
Figura 3.2. Comparación de tamaño, 3 átomos $\frac{1}{2}$ de oro representa 1 nanómetro.....	12
Figura 3.3. Los cambios de tamaño de los transistores por año diseñados por Intel....	13
Figura 3.4. Invenciones que ha hecho la nanotecnología en estos últimos años, a) el Nano BugBot, b)Nanopartículas de Hierro, c) Nanopartículas de Oro.....	15
Figura 3.5. Número de publicaciones aproximadas por países según la pagina de PubMed en el 2011.....	16
Figura 3.6. Explicación gráfica del uso de la nanoinformática.....	18
Figura 3.7. Portada principal de la Web del proyecto ACTION-Grid.....	20
Figura 3.8. Conceptos que encierra el Text Mining.....	22
Figura 3.9. Parte de la página Web de Oracle que habla acerca de los algoritmos de aprendizaje automático más usado para clasificaciones.....	25
Figura 3.10. Logo de la Página Web de WEKA.....	27
Figura 3.11. Principales funciones de la herramienta Weka.....	28
Figura 3.12. Código demostrando que la opción “6” define el uso del algoritmo Regresión Logística.....	30
Figura 3.13. Explicación de los pasos de la herramienta RedNano, sacado de su publicación.....	33
Figura 3.14. Farmaco Abraxane® para el consumo humano.....	36

Figura 3.15. Nanovacuna Inflexal® para el consumo humano.....	37
Figura 3.16. Ejemplo de un ensayo clínico que habla sobre la nanopartícula Abraxane® en el titulo.....	37
Figura 3.17. Ejemplo de un ensayo clínico que habla sobre la nanopartícula Abraxane® en el cuerpo.....	38
Figura 3.18. Nanomateriales mas usados en el campo medico, a) Fullerenos, b) Nanotubos, c) Dióxido de Titanio, d) Dendrímeros, e) Nanoarcilla y f) Nanopartículas de Oro.....	40
Figura 3.19. Ejemplo de un ensayo clínico que habla sobre el nanomaterial Nanotubo en el cuerpo.....	41
Figura 3.20. NA-NOSE, nanodispositivo usado para la respiración artificial.....	42
Figura 3.21. Ejemplo de un ensayo clínico que habla sobre el nanodispositivo Na-NOSE en el cuerpo.....	43
Figura 3.22. Logos de los diferentes repositorios sobre medicina en el mundo: a) BioMedical Central, b) DialNet, c) DOAJ, d) HighWire, e) PLoS, f) PubMed y g) Recolecta.....	45
Figura 3.23. Página Web de Clinical Trials.gov.....	46
Figura 3.24. Curva que demuestra el crecimiento de los registros de ensayos médicos en este repositorio, información tomada de la página de ClinicalTrials.gov.....	47
Figura 3.25. Ejemplo de cómo se ve un ensayo clínico y sus tipos de vistas: a) Full Text View y b) Tabular View.....	49
Figura 3.26. Página Web de ClinicalTrialsRegister.eu.....	50
Figura 3.27. Ejemplo de cómo se ve un ensayo clínico de forma resumida.....	51
Figura 3.28. Ejemplo de cómo se ve un ensayo clínico al elegirlo para leerlo.....	52
Figura 4.1. Procedimiento de la metodología a seguir.....	54
Figura 4.2. Partes de la página web de NCImetathesaurus, información tomada de la tesis doctoral.....	55
Figura 4.3. Explicación gráfica de como descargar los ensayos clínicos en la página Web ClinicalTrials.gov.....	58

Figura 4.4. Explicación gráfica de como descargar los ensayos clínicos en la página Web ClinicalTrialsRegister.eu.....	59
Figura 4.5. Los tipos de ensayos clínicos que necesitamos buscar en los dos repositorios.....	60
Figura 4.6. Pantalla principal de la Herramienta creada.....	73
Figura 4.7. Opciones de la herramienta: a) Nombre y elección de las categorías, b) se debe elegir la Categoría principal c) Debe estar activada la Lista Inteligente.....	74
Figura 4.8. Resultado de los archivos revisados, mostrando su título, numero de unigrams y la fecha de publicación.....	75
Figura 4.9. Pregunta principal para empezar el pre-procesamiento.....	75
Figura 4.10. Tabla de Resultados, mostrando los unigrams y los valores correspondientes de cada uno, la cantidad mínima de repetición del total de los documentos, el máximo de repetición, la media, la desviación estándar, el valor IDF, y el total de documentos en el cual no aparecieron.....	75
Figura 4.11. Cuadro que muestra si se desea seguir con el procesamiento.....	76
Figura 4.12. Captura total de la pantalla principal, con sus resultados correspondientes.....	76
Figura 4.13. Ultimo mensaje para poder empezar a la creación del modelo para la predicción.....	77
Figura 4.14. Creación del modelo y predicción del repositorio de entrenamiento.....	79
Figura 4.15. Parte de la pantalla en donde se escribe el nombre de la carpeta en donde se encuentran los ensayos clínicos, el nombre del archivo para predecir, y el nombre de un archivo para escribir los nombres de los ensayos.....	80
Figura 4.16. Pedazo de código, en el cual se muestra como se agrega un archivo para predecirlo, el modelo a usar y la línea para predecir.....	80
Figura 5.1.- Evolución de los modelos de entrenamiento.....	84
Figura 5.2.- Accu-Check Aviva Nano (dispositivo para medir la presión), dispositivo que en su nombre tiene la etiqueta “nano”, pero no es un nanodispositivo.....	85
Figura 6.1.- Liposoma.....	90

Figura 7.1.- En la herramienta solo está marcada la lista inteligente pero en el cuadro de alado está listo para poder modificarse para la lista de palabras..... 94

INDICE DE TABLAS

Tabla 3.1. Tabla explicativa de la Regresión Logarítmica_____	26
Tabla 3.2. Tabla explicativa del algoritmo de Naïve Bayes_____	26
Tabla 3.3. Tabla explicativa del SVM_____	26
Tabla 3.4. Tabla explicativa del algoritmo Árbol de Decisión_____	27
Tabla 3.5. Principales Lenguajes en los que está desarrollada la librería LibLINEAR y sus respectivas versiones_____	30
Tabla 3.6. Total exacto de publicaciones por año de la página ClinicalTrials.gov, información tomada en la misma página_____	47
Tabla 5.1.- Tabla que demuestra los tipos de modelos y sus resultados porcentuales, desde el primero hasta el que se uso_____	83
Tabla 5.2.- Resultado porcentual de resultados de los 250 ensayos clínicos_____	84
Tabla 5.3.- Tiempo total para crear la bolsa de unigrams_____	86
Tabla 5.4.- Resultado del repositorio Norteamericano. Contiene todas las 9 partes que se dividió el repositorio para hacer la comparación y el número de ensayos clínicos con etiqueta “nano”_____	85
Tabla 5.5.- Valores que se obtuvieron probando el modelo 10 y el modelo final para la clasificación para el repositorio Norteamericano_____	87
Tabla 5.6.- Resultados del repositorio Europeo. En este repositorio no fue necesaria la división de sus ensayos por lo cual se tomo una sola parte del mismo_____	87
Tabla 5.7.- Valores que se obtuvieron probando el modelo 10 y el modelo final para la clasificación para el repositorio Europeo_____	88
Tabla 5.8.- Comparación de los dos repositorios y sus respectivos números de ensayos clínicos y sus porcentajes respectivos_____	88
Tabla 6.1.- Comparación de los dos repositorios y sus respectivos números de ensayos clínicos y sus porcentajes respectivos_____	91

1 INTRODUCCION

Cuando hablamos de nanotecnología se nos vienen muchas definiciones a la cabeza, una definición común dentro de la opinión mundial, es que al escuchar la palabra "nano", la relacionemos con robots sumamente diminutos que hacen ciertas funciones dentro del lugar en donde se los ponga a trabajar, realmente esta definición es solo una parte de la gran gama de divisiones que tiene esta ciencia que se está estudiando. Para obtener una mejor definición acudimos a una de las más importantes organizaciones de nanotecnología que es "U.S. National Nanotechnology Initiative" (La Iniciativa Nacional de Nanotecnología de Estados Unidos) llamada por sus siglas en inglés NNI, y este organismo define a la nanotecnología como "la ciencia, la ingeniería y la tecnología realizada en la nano escala" o, en otras palabras, "la manipulación y el control de la materia con al menos una dimensión de tamaño de 1 a 100 nanómetros". Muchos investigadores hoy en día tienen una disputa de cuáles deben ser las medidas para que esta ciencia los estudie, pero la mayoría se han puesto de acuerdo en dejarlo entre 1 a 100 nanómetros. Si nos vamos un poco a la historia de la nanotecnología recordaremos la famosa charla de uno de los precursores de esta ciencia, el físico Richard Feynman en Caltech (1959), que nos ilustra con su famosa charla llamada "There's Plenty of Room at the Bottom" (Hay suficiente espacio en el fondo), la cual Feynman consideraba la posibilidad de la manipulación de los átomos individuales con materiales muy pequeños en este caso los que llamaremos nanomateriales. Y así después de muchos años podemos afirmar que la nanotecnología ha evolucionado a través de varias generaciones de nanomateriales y cada nueva generación ha ofrecido los avances en la eficacia y la seguridad para muchas ciencias.

Una comparación de la longitud de lo que es un nanómetro se muestra en la Figura 1.1.

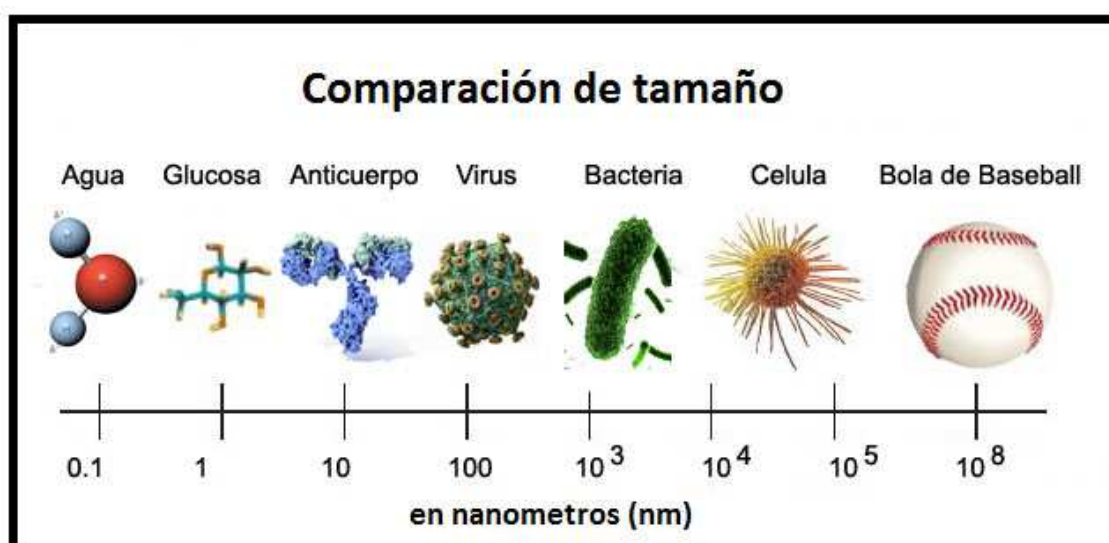


Figura 1.1. Comparación de las longitudes en nanómetros

Actualmente, la nanotecnología representa una de las más grandes promesas de muchas ciencias y representa muchas nuevas oportunidades para la creación de nuevas soluciones que llegan a ser significativas a través de múltiples disciplinas científicas y, en concreto, la denominada biomedicina.

El término biomedicina, en sí fue adquirido entre 1945 y 1975, ya que justamente coincidió con la aparición de un nuevo sistema de innovación médica que en ese tiempo tenía la biología y la política sanitaria la cual estaban relacionándose. Pero aun así el significado de biomedicina fue profundamente influenciado por las diferentes culturas científicas y nacionales que han dado forma a la medicina occidental desde finales del siglo XIX. Y así obtenemos que la palabra biomedicina apareciera por primera vez en 1923 en el diccionario médico de Dorland, y este decía que "es una medicina clínica basada en los principios de la fisiología y la bioquímica" [Quirke y Gaudillie're, 2008]. Pero poco a poco se ha ido mejorando y tomando forma y obteniendo la definición de biomedicina, en concreto, la biomedicina es considerada La biomedicina, como una ciencia que aplica todos los principios de las ciencias naturales en la práctica clínica mediante el estudio e investigación de los procesos fisiopatológicos considerando desde las interacciones moleculares hasta el funcionamiento dinámico del organismo a través de las metodologías aplicadas en la biología, química y física, tomando en cuenta que no se relaciona con la práctica de la medicina. Ahora tomando en cuenta las propiedades especiales de la nanotecnología, sabemos que las nano partículas son adecuadas para el diagnóstico y la terapia moleculares a nivel celular, ofreciendo numerosas aplicaciones médicas, tales como el desarrollo de dispositivos biomédicos, la reparación de tejidos, sistemas de suministro de fármacos, biosensores y un sin número de numerosas funciones que al unir estas dos ciencias nos dan como resultado.

Otra característica de la nanotecnología es que la investigación de esta ciencia tiene conceptos fundamentados sobre la caracterización de nanopartículas y nanomateriales que son usados para el análisis y el tratamiento de datos experimentales, y estos buscan correlacionar las nanopartículas, sus funcionalidades y sus efectos secundarios. Y cuando estos son aplicados a la medicina, se le denomina nanomedicina, que según el International Journal of Nanomedicine (IJN) la define como -una rama de la nanotecnología, la cual es tomada como una ciencia y una tecnología para el diagnóstico, tratamiento, prevención de enfermedades y lesiones traumáticas, para preservar y mejorar la salud humana, con el uso de herramientas moleculares [Webster, 2006]-.

Gracias a la nanotecnología se pudo obtener conocimiento molecular del cuerpo humano, que contribuyen a estudios efectuados actualmente en nanomedicina, la cual están produciendo grandes cantidades de datos estructurales, destacando el papel de los enfoques computacionales en gestión de la información y el análisis masivo de datos complejos. Además, ayuda a las mejoras en la precisión de los métodos computacionales, como es la denominada minería de datos, la cual nos ayuda al descubrimiento de nuevos conocimientos, de crear un modelado gracias al sin número de simulaciones nos han permitido crear herramientas eficaces para automatizar la extracción, gestión, análisis y almacenamiento de estos vastos volúmenes de datos. Un

enfoque en las estrategias para la gestión de la información obtenida de la información relacionada a nanomedicina ha hecho hincapié en la necesidad de mejorar métodos de minería de datos según el NNN [National Nanomanufacturing Network, 2007]. Y se ha creado una nueva disciplina llamada nanoinformática la cual nos ayuda para proporcionar la visión y los enfoques de la informática necesaria para tales fines [Figura 1.2].



Figura 1.2. Explicación gráfica del concepto de nanoinformática

Esta rama de la nanotecnología implica el desarrollo de herramientas eficaces, tecnologías y métodos para recopilar, estandarizar, integrar, analizar y visualizar la información pertinente a la nanomedicina, incluidos los datos que pueden ser relevantes, tales como las propiedades físico-químicas, efectos biológicos, clínicos y toxicológicos, y así sucesivamente [Maojo et al., 2010a, de la Iglesia et al. 2011a]. Y justamente sobre trabajo de fin de máster se trata de nanoinformática.

Debemos considerar que la nanoinformática tiene el potencial de introducir nuevos objetivos y conceptos transformadores a las investigaciones científicas, como son la investigación fundamental, la innovación y la fabricación sostenible de productos, y ahora actualmente que la parte ecológica esta en todo su auge también podemos decir que ayuda la seguridad de las personas y el medio ambiente. En si la nanoinformática es la ciencia y práctica de conocer y determinar qué información es relevante para los investigadores científicos y la ingeniería a nanoescala, para luego desarrollar e implementar mecanismos o herramientas eficaces para recopilar, compartir, validar, almacenar, modelar, analizar, y la aplicación de esa información con los datos involucrados en estos procesos. Es además considerada como una tecnología de amplio impacto que puede contribuir a la mejora de productos y procesos de fabricación a través de diversos sectores de comercio incluyendo la asistencia sanitaria; agua y energía; el transporte; defensa y seguridad; la remediación ambiental y fabricación ecológica; y la seguridad alimentaria con la producción y envasado. Según el NNN [National Nanomanufacturing Network, 2011] En las últimas dos décadas, las minerías de datos a gran escala han comenzado a combinar la ciencia experimental y computacional, con métodos informáticos que utiliza redes de computación masiva, herramientas de ciencias de la información y algoritmos de aprendizaje automático y las tecnologías en las redes sociales. La nanoinformática, incluye el desarrollo y aplicación

de las herramientas críticas necesarias para simulaciones, cálculos, y el modelado predictivo de los nanomateriales, nanodispositivos y nanosistemas.

Dentro de la nanoinformática los investigadores de todas las partes del mundo y la industria gestionan sus datos obtenidos en los repositorios, así como la forma en que descubren datos generados por otros datos. Este tema abarca temas como la gestión de bases de datos, instrumentación, la recopilación de datos de alto rendimiento, nanometrología, etc. Por ejemplo la industria basada en la nanotecnología, cuando se trata de creación de nanomateriales o nanodispositivos requiere una instrumentación que sea tan exacta, barata y fiable como sea posible y que se asocie con las normas internacionalmente aceptadas, en estos temas tiene que ver la nanometrología. Por ende los fabricantes, investigadores, las autoridades u organizaciones públicas y no gubernamentales demandan métodos de prueba y muestreo, mediciones e instrumentaciones, regulaciones y normas para evitar que los seres humanos padezcan por este nuevo desarrollo nanotecnológico, etc. Otro ejemplo es la gestión de las bases de datos para obtener información relevante, este tema preocupa por la seguridad de los datos así sea que sean solo un conjunto mínimo de información sobre un repositorio o texto. La discusión engloba que si esa cantidad mínima de datos está ligado a los requisitos mínimos de estos conjuntos de datos y si esto es un factor de calidad para los datos obtenidos futuros. Ya que algunos requisitos dependen de los tipos de datos y de su cantidad de datos para cuantificar, entender o clasificar el material dado. Y muchos investigadores se preguntan sobre cuanta información sería capaz de predecir los riesgos de materiales basados en las similitudes en los datos aplicando clasificadores con aprendizaje automático.



Figura 1.3. Aplicación de nanoinformática en ensayos clínicos de diferentes continentes

Ahora si hablamos un poco más allá sobre la nanoinformática, existen muchas otros desafíos con la gestión de bases de datos o repositorios de información, como es la clasificación automática de estos, para que los diferentes investigadores puedan escoger solo los repositorios o registros dentro de los mismo para su uso, y no estén buscando uno a uno para ver si realmente son o no de su tema a tratar. Así como saber en qué parte del mundo se están haciendo más ensayos clínicos sobre la nanomedicina o ciencias afines. Todo esto también es un desafío para las personas que estamos desarrollando herramientas para ayudar estas necesidades, y este trabajo de tesis justamente desarrolla una técnica o método para poder clasificar de forma automática estos ensayos, además deja una herramienta para solo agregar estos repositorios y comenzar la clasificación de los mismos. Este trabajo también nos ayudara a conocer el porcentaje de trabajos de nanomedicina –nanopartículas, nanodispositivos, nanomateriales- de dos repositorios grandes de registros de ensayos clínicos que contienen muchos tipos de temas como la biomedicina, medicina en general, nanomedicina, los usos de dispositivos médicos, etc. Y estos son el repositorio de ClinicalTrials.gov que su base está en Norteamérica y del repositorio ClinialTrialsRegister.eu que su base está en Inglaterra y tiene ensayos relacionados de toda la unión europea [Figura 1.3].

2 OBJETIVOS Y ALCANCE

En este capítulo se mostraran los objetivos y el alcance del trabajo fin de máster, tomando en cuenta lo escrito en la introducción se responderán ciertas cuestiones que los investigadores necesitan resolver, para este estudio se uso un trabajo anterior (de la Iglesia, 2013) la cual se investigo durante varios años en los campos de la biomedicina y nanomedicina, aplicando enfoques computacionales, métodos y técnicas de la información extraída de ambos dominios. Esta autora estuvo dentro del Grupo de Informática Biomédica (GIB, Figura 2.1) que se estableció formalmente en 1993 bajo la dirección y coordinación del profesor Víctor Maojo, un excelente miembro de la facultad del Departamento de Inteligencia Artificial de la Facultad de la Universidad Politécnica de Madrid. Este grupo llamado GIB ha establecido la colaboración en los últimos 15 años con el programa de Ciencias de la Salud y Tecnología de Harvard-MIT, con el que se desarrolló un programa de intercambio intenso, con seis estudiantes, la Universidad de Rutgers, la Universidad de Utah y el Instituto Nacional del Cáncer de EE. El GIB ha trabajado en la integración de bases de datos clínica-genómica, una investigación que fue pionera en el proyecto financiado por la Comisión Primera Europea en este ámbito (INFOGENMED). En 2008-2011 Prof. Maojo dirigió el proyecto de Acción de cuadrícula, que fue el primer proyecto financiado por la CE en la nueva zona de Nanoinformática. Varios trabajos seminales escritos en el GIB han contribuido a consolidar el campo, y así un sin número de nuevas herramientas y proyectos que ha ayudado a mejorar muchas ciencias de la investigación [1]. El autor es estudiante de la Universidad Politécnica de Madrid en la facultad de Informática, cursando el Máster Universitario de Inteligencia Artificial, y por esto ha podido usar los anexos del trabajo doctoral anteriormente mencionado [De la Iglesia, 2013].



Figura 2.1. Folleto del Grupo de Biomedicina Informática de la UPM

2.1 Objetivos

El objetivo general del presente trabajo es la comparación de dos repositorios de registros de ensayos clínicos tomando en cuenta que contienen información relacionada con el dominio de la nanotecnología dentro de la medicina, y así poder proporcionar un número aproximado sobre la cantidad de registros de ensayos clínicos que estos repositorios contienen acerca de temas relacionados sobre nanomedicina hoy en día. Se tomara en cuenta si tienen una proporción igual, menor o mayor, ya que las cantidades de registros no son las mismas, ni siquiera cercanas; por ende, solo tomaremos en cuenta el porcentaje aproximado de cada uno. Para ello, el autor proporcionara un nuevo método de pre-procesamiento de texto y un método de procesamiento usando un clasificador de aprendizaje automático, utilizándolos para poder diferenciar entre estos registros de ensayos clínicos, etiquetándolos como "NANO" si es que en el contexto del ensayo clínico trata sobre nanomedicina o la etiqueta "NO-NANO" si es que dentro de su contexto no trata sobre nanomedicina.

En resumen, este trabajo nos proporcionara, además de un nuevo método de pre-procesamiento, una discusión sobre como los trabajos de los investigadores de la nanomedicina están trabajando alrededor del mundo, y cuantos registros están creados actualmente en cada repositorio, tomando en cuenta el modelo creado por la herramienta de clasificación a usar -Adaptando el vocabulario, la complejidad, la desorganización y singularidades de la información dentro de cada ensayo clínico - . Se usara como base un estudio reciente por una tesis doctoral que demostró que cierta herramienta fue la mejor para clasificar estos registros la cual hizo un análisis exhaustivo de la diferencia y las características específicas de la nanomedicina en registros separados manualmente. En este contexto, la hipótesis central es que es posible inferir y descubrir nuevos registros valiosos que contengan información de estos temas, relevante para la nanomedicina, tanto de la literatura científica y los resúmenes de ensayos clínicos disponibles en la Web. La metodología propuesta, por lo tanto, tendrá como objetivo esa demostración.

Como fue indicado en el capítulo anterior los repositorios que son más usados para subir los registros de ensayos clínicos son dos grandes, el primero ClinicalTrial.gov que es usado frecuentemente por investigadores que se encuentran en Norteamérica – Estados Unidos, Canadá y otros países fuera de Norteamérica-, y el ClinicalTrialsRegister.eu que es usado por países de la unión europea -Austria, Bélgica, Bulgaria, Croacia, Ciprés, Republica Checa, Dinamarca, Estonia, Finlandia, Francia, Alemania, Grecia, Hungría, Islandia, Irlanda, Italia, Letonia, Liechtenstein, Lituania, Luxemburgo, Malta, Holanda, Noruega, Polonia, Portugal, Rumania, Eslovaquia, Eslovenia, España, Suiza, Reino Unido y las afuera de EU/EEA- para subir ensayos clínicos de medicina, la cual nos ayuda a separara en estos dos continentes grandes que y cuanto está trabajando cada uno en estos temas.

Aunque este es el objetivo general, este método que se va a usar también puede ayudar a categorizar otros tipos de ensayos clínicos o cualquier otro título que se desea, pueden

ser a textos planos, como a archivos xml, siempre y cuando se hagan unos cambios referentes según sea el caso.

2.2 Alcance

Para tomar en cuenta el alcance de este trabajo, se debe puntualizar que este trabajo es una continuación de lo que la tesis doctoral usada quiso lograr y uno de sus futuros trabajos que mencionada es el objetivo mencionado anterior, por ende el alcance de este trabajo es muy cercano a lo que la autora quiso referirse ya que se tuvo que analizar diversos escenarios por las diferentes actividades que los investigadores podrían emprender mientras realizan sus investigaciones en este caso sobre nanotecnología:

1. buscar dentro de la Web, fuentes de datos y recursos computacionales para apoyar su investigación,
2. buscar dentro de la archivos bibliográficos los resultados experimentales actuales y publicaciones relacionadas con su investigación y,
- 3. *buscar registros de ensayos clínicos para conocer los resultados clínicos relacionados con su investigación.***

El alcance de este trabajo solo es para desarrollar la tarea 3, o sea, buscar registro de ensayos clínicos para conocer los resultados clínicos relacionados con la investigación, la cual el desarrollo de esta actividad de investigación dependerá de la utilización de la herramienta informática que se usa en este trabajo, y los repositorios que se encuentren en la web ya antes mencionado, recordando que son bases de datos de registros de ensayos clínicos basados en nanotecnología.

Una vez más dando a entender que este método que se usara no solo es válido para los ensayos clínicos, se puede usar con cualquiera de las 2 tareas antes mencionadas, para ayudar a los investigadores en sus procesos.

Como vamos a usar la herramienta solo en ensayos clínicos debemos destacar que estos son esenciales para la aplicación clínica ya que tratan de una investigación experimental sobre nuevos medicamentos o terapias. Al igual que los ensayos clínicos de medicamentos y productos biológicos tradicionales nos han ayudado a acelerar la mejora de los hallazgos biomédicos en la práctica médica, los ensayos clínicos de nanofármacos y nanodispositivos podría generar nuevos nanomateriales, nanodispositivos e inclusive nanopartículas como agentes esenciales para el diagnóstico y las terapias médicas.

Debemos centrarnos que este trabajo se trata de minería de datos y dentro de la nanomedicina y la biomedicina no hay ninguna propuesta metodológica conocida por los investigadores acerca de la minería de datos, ya que se necesita alguno que diferencie de las metodologías tradicionales en alguna forma significativa. Hace unos

años cuando se hablaba del tema de minería de datos en medicina [Cios y Kacprzyk, 2001], se propusieron que las aplicaciones en el área son diferentes de otras debido a dos temas principales:

- i) Heterogeneidad de los datos médicos, ya que estos datos están compuestos por varios tipos de vocabularios y especialidades, es muy probable que cuando se hable de una enfermedad como el cáncer, existan muchas variables de distintas especialidades como si es un cáncer al hígado, tenemos desde temas hepáticos hasta temas de gastroenterológicos, y se hace muy complicado hacer estas diferencias.
- ii) Asuntos éticos, legales y sociales, otro de los problemas es que no todos los médicos dejan sus trabajos (ensayos clínicos) en la Web para que los demás investigadores puedan tener acceso fácilmente, y no es muy ético robar esta información si es que no la obtenemos por los mismo autores o si ellos han dejado estos datos en algún lugar libre.

El problema para buscar este tipo de información y aplicar la minería de texto es de que los investigadores no tienen un repositorio en la Web que contenga material de nanomedicina públicamente y que solo hablen de este tema, si no que existen archivos en línea con esta clase de información que tienen que ver con criterios que se diferencian entre los estudios con los nanomateriales o procesos basados en la nanotecnología (nano) y los ensayos clínicos que no implican la nanotecnología (no nano). Solo de pensar que se debe averiguar si nanofármacos y nanodispositivos participaron en un estudio de ensayos clínicos es una tarea difícil, ya que se tendría que leer muchas líneas y abrir cada registro.

Para finalizar con este capítulo resumimos que el alcance para este trabajo de fin de máster se basa en descubrir por medio de la minería de texto con ayuda de algoritmos de clasificación de aprendizaje automático la cantidad de ensayos clínicos que existen en los dos repositorios antes mencionados que hablen o dentro de su contexto usen algún tipo de nanotecnología, para que los investigadores al momento de querer obtener resultados de estos ensayos vayan encaminados automáticamente a aquellos registros de ensayos clínicos que se tratan sobre este tema, además conociendo en que parte del mundo se habla o se crean más experimentos sobre una nanopartícula, un nanomaterial o un nanodispositivo. [Figura 2.2]

2.3 Organización del trabajo realizado

El trabajo está organizado de la siguiente manera. El tercer capítulo analiza el estado del arte acerca de la intersección de la meta antes mencionada de áreas científicas: la biomedicina, medicina molecular, la medicina personalizada y la nanomedicina, tomando en cuenta la detección de los recientes avances en nanomedicina, así como los desafíos presentados en la gestión de información en el nivel nano, también realizaremos un trabajo acerca de los métodos más usados de la minería de texto en

estos últimos años. El cuarto capítulo presenta la metodología seguida en este trabajo de fin de máster, al presentar la descripción del experimento específico realizado mostrando como fue el pre-procesamiento paso a paso y el procesamiento de texto. El capítulo cinco evalúa los resultados obtenidos por el experimento obtenidos en el capítulo anterior, destacando los diferentes problemas que enfrentan durante la ejecución, así como la comparación de los resultados obtenidos en los dos repositorios antes mencionados que tienen que ver con el campo de la nanomedicina. En el sexto capítulo se muestra un análisis, discusión y comentarios por los resultados obtenidos, con los respectivos cuadros obtenidos por los resultados. En el séptimo capítulo, se proporciona observaciones finales, las conclusiones de este trabajo y una lista de posibles líneas de investigación para el trabajo futuro de la zona. Por último, este documento incluye tres anexos que contienen información complementaria con todo lo referido en esta tesis.

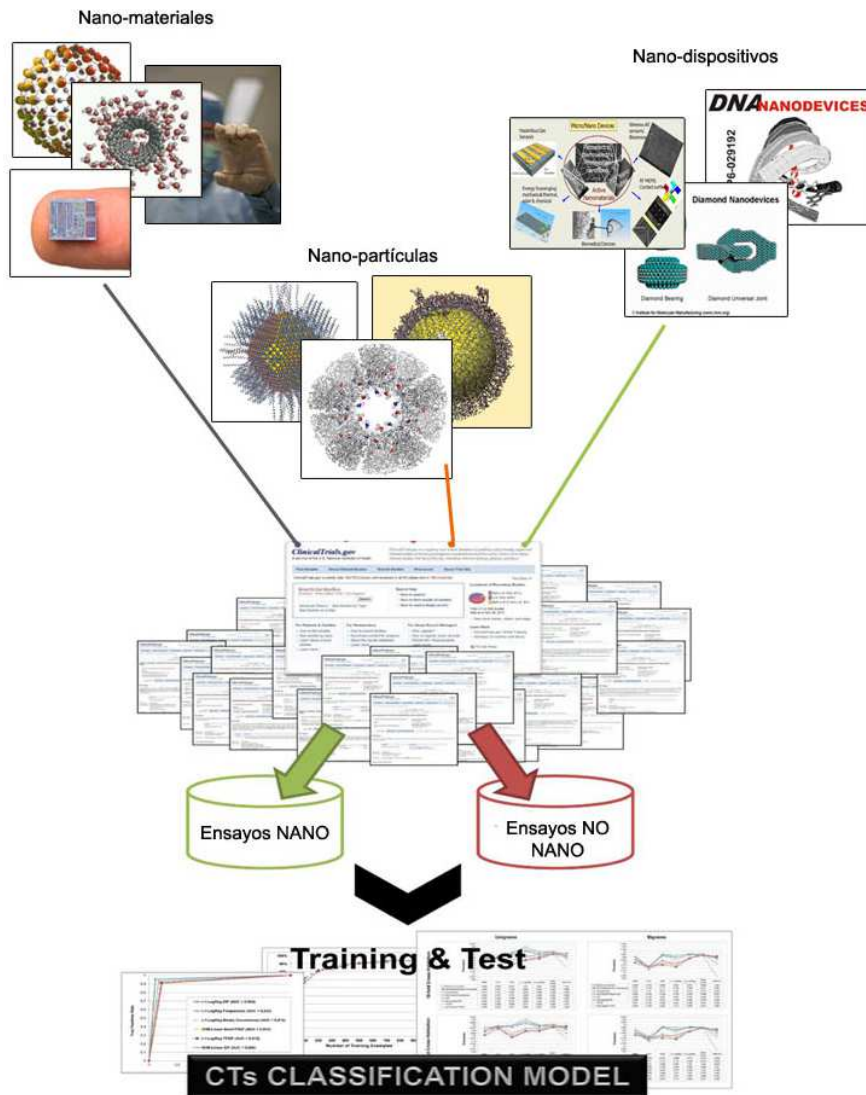


Figura 2.2. El alcance de esta TFM: Clasificación de los ensayos clínicos

3 ESTADO DEL ARTE

En este capítulo denominado estado del arte, vamos a revisar ciertos términos que se van a usar en este trabajo de fin de máster, tomando en cuenta desde la ciencia la cual estudia este trabajo que es la nanotecnología hasta la historia de los repositorios que vamos a usar, de que están conformados cada uno y en qué se diferencia del uno con el otro. También vamos a revisar los últimos trabajos de minería de texto que se han realizado en estos años, para comparar el trabajo realizado con algunos anteriores, aunque no existen muchos trabajos que tengan que ver con la minería de texto en registros de ensayos clínicos, se lo comparara con minería de texto aplicado en otras aéreas. Revisaremos la herramienta Weka que es aquella que hace el modelo para la clasificación al final del pre-procesamiento, y por último es aquella que muestra los resultados tomando en cuenta el algoritmo a usar.

3.1 Nanotecnología

La nanotecnología en si es considerada solo una ciencia, también es considerada como la ingeniería y la tecnología realizada en la nano escala, que actualmente su escala su escala va alrededor de 1 a 100 nanómetros, y está considerada por el instituto más importante que estudia esta ciencia en el mundo, llamado National Nanotechnology Initiative [Figura 3.1]. Ellos toman muy en cuenta las ideas y los conceptos detrás de la nanociencia y la nanotecnología, y conmemoran que todo esta idea comenzó con una charla llamada “There’s Plenty of Room at the Bottom” (Hay mucho sitio al fondo, en español) por un físico muy importante en ese tiempo llamado Richard Feynman en una reunión de la Sociedad Americana de Física que se realizo en el Instituto de Tecnología de California (Caltech), el 29 de diciembre de 1959. Feynman en esta charla, nos describió un proceso futurístico en ese entonces, en el cual los científicos serían capaces de manipular y controlar los átomos y moléculas individuales de cada cuerpo que nos rodea. Y al pasar más de una década, el profesor Norio Taniguchi en sus muchas exploraciones e investigaciones con sus mecanismos de ultra precisión, probó y definió el término nanotecnología. No fue hasta 1981, los físicos Gerd Binnig y Heinrich Rohrer de IBM que crearon el microscopio de efecto túnel, la cual sirve para tomar imágenes de superficies de los átomos individuales, que la nanotecnología moderna comenzó a desarrollarse.



Figura 3.1. Logo de la National Nanotechnology Initiative

Hablar sobre nanotecnología en estos tiempos es hablar del futuro no solo de la ciencia, sino también de la tecnología, el futuro del mundo y todo lo que nos rodea. El mundo y el pensamiento de cada persona está cambiando gracias este tema, pero la pregunta antes de comenzar a explicar todo lo que este tema se relaciona es explicarnos él porque debemos usar nuevas tecnologías con esta medida, el porqué debemos usar cosas tan diminutas.

Si hablamos de comparar a que es igual un nanómetro la Comisión Europea en el libro llamado “Nanotecnologías: Principios, Aplicaciones, Implicaciones y Manualidades” publicado en el año de 2013, demostró que el tamaño de un nanómetro es igual a 3 y medio átomos de oro en fila, asumiendo que cada átomo de oro tiene una longitud de 0.144 nm. [Figura 3.2]

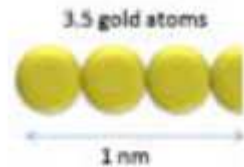


Figura 3.2. Comparación de tamaño, 3 átomos $1/2$ de oro representa 1 nanómetro

Un estudio realizado por el NNIN por sus siglas en ingles (National Nanotechnology Infrastructure Network), explicó porque debemos trabajar a esta medida, cual serian sus ventajas en esta nano escala, y tomo en consideración que si trabajamos a esa longitud los dispositivos creados iban a ser mucho más ligeros, ocuparían menos espacio para trabajar, serian muchas baratos crearlos, usarían menos energía que los materiales de grandes longitudes, se usarían menos productos lo que produciría menos gastos en los materiales que componen al nuevo dispositivo, y así podríamos llegar a pensar en una ayuda en el ámbito ecológico en donde no hubiera tanto desgaste al momento de desechar los residuos que no son usados en la creación de estos, además se afirmó que en el escala nanométrica, las propiedades de la materia, como la energía se comportan de manera diferente. Esta es una consecuencia directa del tamaño de los nanomateriales, explicó físicamente hablando por los efectos cuánticos que se producen. La consecuencia por ejemplo es que un material como el metal plata (Au), la cual a granel no es tóxica, cuando hablamos dentro de su nanoescala, o sea, las nanopartículas de este metal plata son capaces de matar a los virus al entrar con un simple contacto. Y si hablamos de otras propiedades podemos definir algunas como la conductividad eléctrica, el color, la fuerza y el cambio de peso cuando se encuentra a este nivel tan diminuto, así que este metal puede convertirse en un semiconductor o un aislante según sea las propiedades y el proceso que es llevado a cabo para convertirlo en una nanopartícula.

Esta ciencia ha motivado mucho a las empresas tecnológicas como un ejemplo tenemos a la empresa Intel en la producción de los transistores [Figura 3.3], las cuales muestra cuadro a cuadro el tamaño con el respectivo año de fabricación que esta empresa ha

creado para ocupar menos espacio en las microcomponentes que este fabrica pero aun así aunque el transistor es más pequeño la función de este componente mejora.

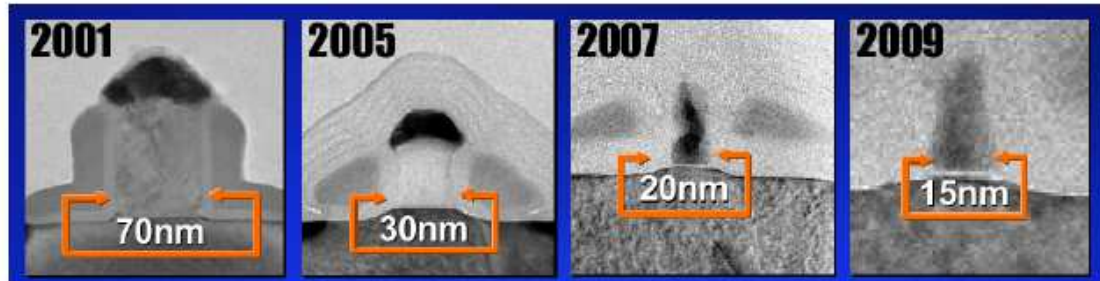


Figura 3.3. Los cambios de tamaño de los transistores por año diseñados por Intel

Y como este tenemos varios ejemplos de aplicaciones de nanotecnología, como son los motores, bombas, giroscopios, acelerómetros, sensores, batería solares, etc. Un sin número de creaciones que nos han ayudado día a día pero para este trabajo de fin de máster realmente nos interesan los trabajos que fueron creados para la medicina, biomedicina y la nanomedicina, que serán definidas más adelante.

Pero no solo los robots o los dispositivos electrónicos a nivel nano deben llevarse toda la atención, existen también otros temas que son de gran importancia como las actividades relacionadas con la educación y la formación de nuevos investigadores de otras ramas, la medicina, la informática. Ya que por otra parte, la realización de una verdadera investigación interdisciplinaria en el área de la nanotecnología requiere muchos nuevos planteamientos de educación y formación aplicables tanto a la investigación como a la industria, y para esto muchos tipos de disciplinas deben trabajar para mejorar todos estos puntos.

Realmente, todas las diferentes disciplinas científicas, incluyendo cada cosa “nano” (como los nanomateriales, nanodispositivos, nano-máquinas, nanoelectrónica), tienen su propia imagen del mundo, o sea todas estas disciplinas describen una forma distinta de uso. Esta es la razón por la cual todas ellas crean innovaciones y desarrollos industriales que son profundamente diferentes. Sin embargo, estos campos están fuertemente interrelacionados, por eso es necesario que para hacer un estudio del mundo nanotecnológico tenemos dos caminos, elegir una rama nanotecnológica o dedicarnos a algo más general como el estudio de un dispositivo, es decir, si deseamos estudiar en si la nanotecnología es mejor hacer nuestros estudios más interdisciplinario con el fin de nos permitirá entender el nanomundo.

Tomando en cuenta la idea del párrafo anterior como nuestra idea base para nuestro trabajo de fin de máster, se ha introducido a la nanotecnología en una escala específica y así podemos dar a ese tema una gama más amplia posible mostrando su base científica común, así como sus múltiples interconexiones, en este caso hablo de la nanomedicina y la nanoinformática. La nanomedicina es la práctica de medicina pero usando materiales que están dentro de la nano escala (1 a 100 nm, aunque a veces llegan a 200 nm). Y la nanoinformática es una ayuda para la investigación dentro de publicaciones, ensayos

clínicos que usaron a la nanomedicina como objetivo para crear sus diferentes herramientas, algoritmos o funciones.

3.1.1 Nanomedicina

Lo primero para poder hablar de este tema, necesitamos definir que es en si la nanomedicina, y así tenemos que la nanomedicina es el proceso para diagnosticar, crear o hacer un tratamiento, además ayuda a la prevención de enfermedades y lesiones traumáticas, para aliviar los dolores, y lo más importante es que ayuda a preservar y mejorar la salud humana, con el uso de herramientas moleculares y conocimiento molecular o atómico del cuerpo humano. En un tiempo no muy lejano debemos saber que la nanomedicina puede aclarar muchos problemas o enigmas médicos importantes mediante el uso de materiales a nanoescala-estructurada y nanodispositivos simples que pueden ser fabricado hoy sin ningún inconveniente, incluyendo la interacción de materiales nano estructurados con los sistemas biológicos. Y es así como un Journal de Computación y nanociencia de los Estados Unidos en el 2005 nos dice que a mediano plazo, la biotecnología y la nanotecnología hará posibles mejores avances moleculares que serán más notables en el campo de la medicina y biobotics - desarrollo de robots biológicos, incluyendo biorobots microbiológicos o organismos manipulados por el ser humano. Y unos años mas tarde o sea una tarea a largo plazo, tal vez 10 a 20 años, los sistemas de máquinas moleculares y los nanorobots pueden unirse al arsenal médico, dando finalmente a los médicos herramientas más potentes e imaginables para conquistar y vencer a las enfermedades, a la mala salud y el envejecimiento. [Freitas R., 2005]

El objetivo de la nanomedicina puede ser ampliamente definido como el monitoreo integral, control, construcción, reparación, defensa y mejora de todos los sistemas biológicos humanos, que trabajan desde el nivel molecular utilizando dispositivos de ingeniería a nivel nano y nanoestructuras fabricadas por esa misma ingeniería, y en última instancia, para lograr beneficio médico. En este contexto, esas nanoestructuras pueden ser incluidas en un micro-dispositivo (que podría tener un macro-interface) o un entorno biológico. El enfoque, sin embargo, está siempre en nanointeracciones para el funcionamiento de un dispositivo más grande, y si hablamos biológicamente dentro de un sistema celular.

Las creaciones y desarrollos de la nanomedicina para la investigación y la práctica biomédica son impresionantes, que van desde la mejora de los productos farmacéuticos -hacer más efectivos estos productos con el objetivo de reducir su contraindicaciones- a la creación de nuevos dispositivos, que ayudan a crear nuevos procedimientos de diagnóstico o para el desarrollo de nuevas técnicas y materiales para sustitución de tejidos, que ayudan a la reparación en el área de la medicina regenerativa [Figura 3.4]. Otra de los desarrollos más importantes en este campo son las nanopartículas que recientemente comenzaron a ser ampliamente consideradas para la práctica médica como herramientas de diagnóstico y terapéuticos a fin de comprender, detectar y lo más importante tratar enfermedades. Pero, para este tipo de nuevas aplicaciones clínicas, los efectos secundarios -como la toxicidad en los animales y las repercusiones del medio

ambiente- deben ser analizados cuidadosamente por los investigadores antes de que estas nanopartículas estén aprobados para su uso en la rutina clínica, o sea para el uso humano. Como ejemplo para comprender la toxicidad de las nanopartículas es importante considerarlas como partículas de muy pequeño tamaño, la cual son habilitadas para circular a través de la sangre, la linfa u otros caminos que son importantes para todo ser biológico. El mal uso de estas tecnologías y desarrollos podrían tener consecuencias no deseadas, formando algún efecto adverso, algún resultado tóxico temporal o permanente muy dañino para el ser humano.

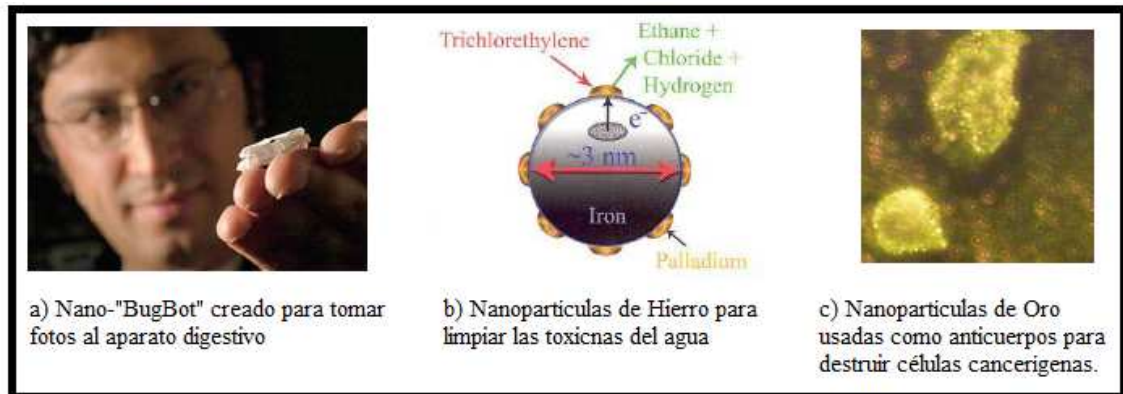


Figura 3.4. Invencciones que ha hecho la nanotecnología en estos últimos años, a) el Nano BugBot, b) Nanopartículas de Hierro, c) Nanopartículas de Oro

Otras de las tecnologías usadas en la nanomedicina son el análisis de las herramientas nano y las nanoimágenes que de igual forma son aspectos que han sido vistos como complementarios. La formación de nanoimágenes se la tomo como una oportunidad para refinar nanotécnicas existentes que permiten el análisis de los tejidos normales y patológicos dentro de un ser biológico, en si para traer rápidamente una mejor comprensión de la iniciación y el progreso de una enfermedad en especifica. Existe un gran interés y necesidad de desarrollar estas técnicas nano, ya que nos traen ideas novedosas para el monitoreo en tiempo real de procesos celulares y moleculares, con mejoras que anteriormente no podíamos tener como la resolución y que todos estos procesos lo vemos en tiempo real. Si pensamos más allá, podemos llegar a la conclusión que la investigación debe desarrollar un enfoque multimodal para las tecnologías para la nanoimagen, y ayudar al diseño de nano herramientas analíticas con alta reproducibilidad, sensibilidad y fiabilidad. Tales herramientas permitirían cambiar el paradigma en la vigilancia de los pre-síntoma de una enfermedad, y permitir el uso de medicamentos preventivos en una etapa temprana de esta.

El desarrollo de esta tecnología científica y nanotecnológica permitió la fabricación de nuevas partículas y dispositivos a nanoescala y además el estudio de sus efectos. La aplicación de estos nuevos materiales y dispositivos para la aplicación en el ser humano ha evolucionado rápidamente en los últimos años, gracias a la creación de la nanomedicina. Si revisamos en la web el uso de estas herramientas tenemos que existen casi 110.000 documentos que aparecen indexados en Medline (PubMed) si hacemos una búsqueda con el término "nanopartícula" y también encontramos más de 10.000

documentos bajo el término "nanomedicina", obviamente estos estudios fueron creados por todas partes del mundo como lo muestra la Figura 3.5. La publicación más antigua acerca de nanomedicina fue publicada en 1999 [Weber T., 1999]. Sin embargo, muchos otros ejemplos se pueden encontrar en este repositorio de publicaciones, aunque no han sido indexados bajo el término "nanomedicina". Esta gran cantidad de publicaciones en este campo que es todavía en desconocida en gran medida para muchos investigadores biomédicos indica su el increíble y tremendo impacto que ha proporcionado este tema en la ciencia y tecnología biomédica moderna. Al mismo tiempo, también señala una nueva tendencia científica, en donde se tendrá una enorme cantidad de información que se ha manejado de manera eficiente por medio de un conjunto de herramientas informáticas, además de nuevos enfoques para la gestión de estos datos y el uso inteligente de la información y el conocimiento. Muchas de estas investigaciones se destacan por los desarrollos necesarios de nuevos estándares y herramientas informáticas -como bases de datos, modelos de simulación, vocabularios, ontologías, etc.- para gestionar y estructurar el conocimiento biomédico a varios niveles biológicos. En este sentido, la ampliación de esta información en el nivel nano podría seguramente será un nuevo reto para los informáticos biomédicos. Y aquí es donde encontramos a una herramienta que ayuda a encontrar y gestionar esta información a la cual llamamos nanoinformática, la cual veremos más adelante.

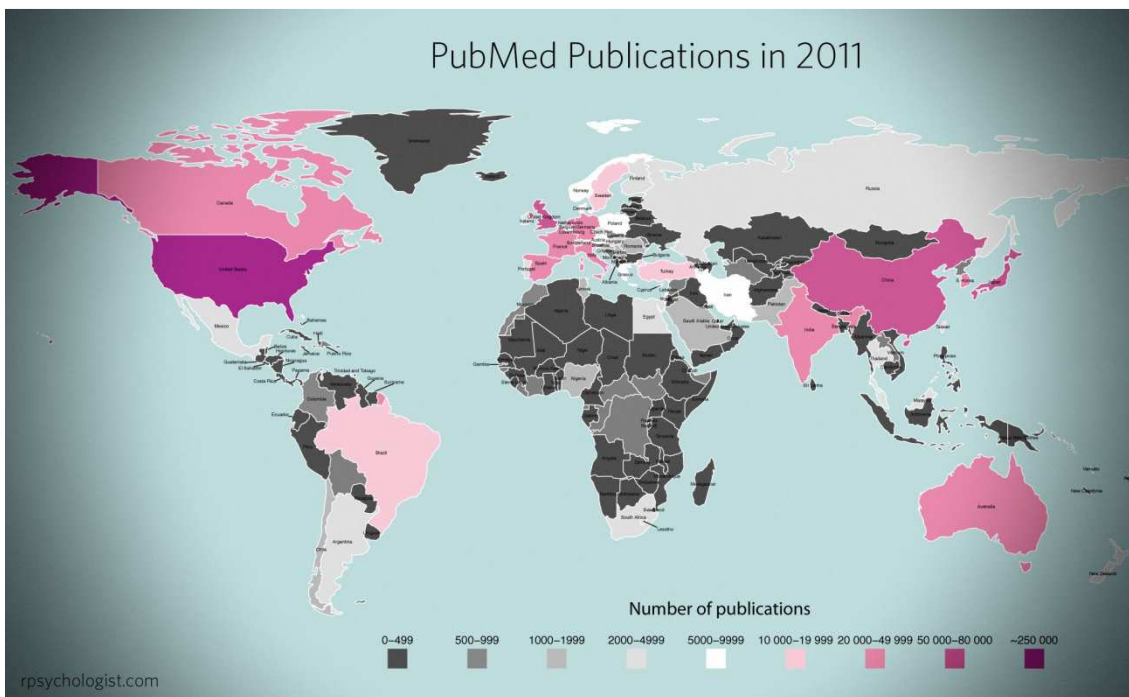


Figura 3.5. Número de publicaciones aproximadas por países según la pagina de PubMed en el 2011

Existe una necesidad urgente de acortar los plazos entre la adquisición de datos y conocimientos fundamentales sobre estos nanomateriales, sobre las publicaciones de estudios y ensayos relacionados en los medios científicos, sobre las partes negativas de

su uso, los métodos que no fueron buenos para obtener lo que se quiso mostrar, o los métodos que fueron exitosos y principalmente su aplicación clínica. Esto podría solucionarse mediante el apoyo al desarrollo de herramientas para obtener una nanotecnología "responsable", que el suministro de información sea de alta calidad y pueda obtenerse a través de un sistema de mejores intercambio de información y compartir estas novedades como los datos fisicoquímicos para determinar de manera eficiente la asociación entre la exposición a los nanomateriales y los efectos secundarios que estos pueden causar. Actualmente, en la mayoría de los casos, los ensayos clínicos en nanopartículas no revelan nuevos efectos tóxicos que una nanopartícula suministra en sí o cualquiera de sus componentes y, por lo tanto, debe ser necesario diferenciar entre la toxicidad del recipiente que lo contiene y el de su interior, que son las piedras angulares en la determinación y eficacia para conocer los efectos adversos contra la salud. De hecho, existe un estudio que nos indica que en el caso de la nanomedicina, la elaboración y aprobación de nuevos fármacos no debe tomar tanto tiempo y los ensayos clínicos no son necesariamente vitales en algunos casos, sobre todo si se considera que los nanomateriales se derivan generalmente de los medicamentos previamente autorizados por la Administración Norteamericana de Comidas y Drogas (FDA, U.S. Food and Drug Administration) [Etheridge et al. 2013] o si hablamos del continente Europeo, este papel lo tiene la Agencia Europea de Medicamentos (EMA, European Medicines Agency) [Karalis y Macheras, 2012].

Ahora dada la gran cantidad de registros, investigaciones, ensayos clínicos, publicaciones sobre nanomedicina que día a día se incrementan y que cada vez son generados más rápidamente por la investigación nanomédica actualmente, es evidente que se necesitaba el desarrollo de nuevas herramientas para la gestión de toda esta información, ya que se ha convertido en un paso crucial para avanzar en el campo de esta investigación. Los informáticos se propusieron a crear ciertas herramientas -algunas todavía están en desarrollo- para mejorar la eficiencia de esta investigación en las áreas relacionadas con la nanomedicina y así se construyeron tomando en consideración todos los puntos anteriores, creando la nanoinformática.

3.1.2 Nanoinformática

La informática es la aplicación de métodos de información y ciencias de la computación para recopilar, analizar y aplicar la información. La "X-informática" (X, cualquier termino de alguna ciencia a estudiar) se ha convertido en el descriptor predeterminado para la aplicación de estos métodos a un conjunto de problemas dentro de un campo o disciplina específica, como la bioinformática en la biología o ahora la nanoinformática en nanotecnología. Así podemos decir que esta ciencia representa un sin número de herramientas que tienen como objetivo ayudar a estudios, investigaciones, repositorios, búsquedas, clasificadores que acerca de nanomedicina. En otras palabras, la nanoinformática es la ciencia y la práctica de determinar qué información es relevante para la comunidad de la ciencia, sean estos doctores, practicantes, científicos,

investigadores, ingenieros que estudien el tema de nanotecnología, para luego desarrollar e implementar mecanismos eficaces y eficientes para recopilar, validar, almacenar, compartir, analizar, modelar, y aplicar esa información para un bien en el campo a nanoescala, obteniendo así la información relevante de cada tema. [Figura 3.6]

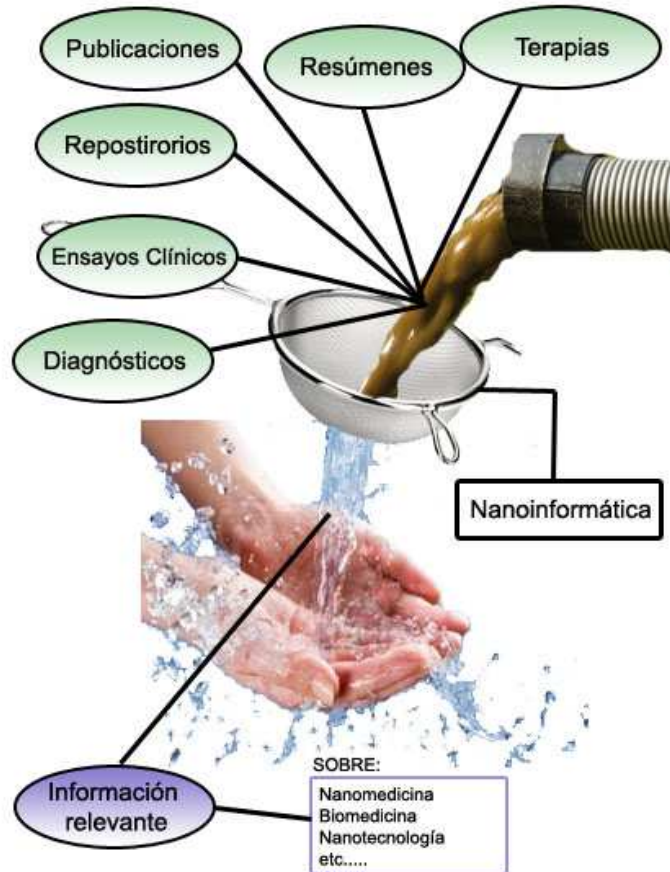


Figura 3.6. Explicación gráfica del uso de la nanoinformática

La nanotecnología ha surgido como una pieza clave en el campo de la nanomedicina, ya que el uso de nanopartículas dentro de la ingeniería está aumentando rápidamente, pero como dijimos en el punto anterior la evaluación de la seguridad también es importante para su uso beneficioso, teniendo en cuenta que la evaluación experimental de estos nuevos nanomateriales es costosa y laboriosa, y por esta razón se deben de tener datos que nos ayudan a mostrar todas las ventajas y desventajas de estas nanopartículas, claro está que todo esto está en los datos relevantes de las publicaciones o ensayos clínicos del uso de estas. Además la nanoinformática no solo es la ingeniería de nanopartículas también es necesaria para el desarrollo inteligente, la caracterización comparativa de los materiales y dispositivos a nanoescala, para el diseño y uso de nanosistemas optimizados, para el desarrollo de los procesos de instrumentación y fabricación avanzada, y para la garantía de la seguridad y salud ocupacional y ambiental. Esta ciencia también implica la utilización de las herramientas de comunicación en redes

para lanzar y apoyar a las comunidades eficientes en este ámbito y ayuda a fomentar el descubrimiento científico eficiente a través de la minería de datos y aprendizaje automático.

Esta ciencia nació en Junio de 2007, en una conferencia celebrada en Arlington, Virginia con el apoyo de la Fundación Nacional de Ciencias de Estados Unidos (NSF, US National Science Foundation) la cual el actual Co-Director del NSF y del MassNanoTech y director del National Nanomanufacturing Network, el físico Mark Tuominen dio a conocer los primeros conceptos de nanoinformática [National Nanomanufacturing Network & National Nanotechnology Initiative, 2007]. En esta conferencia, la nanoinformática se definió por como "el desarrollo de mecanismos eficaces para la recopilación, el intercambio, la visualización, modelado y análisis de datos e información de interés para la comunidad científica y la ingeniería a nanoescala". Esta definición no ha cambiado mucho desde ese entonces, lo único que se ha hecho es aumentar las capacidades de la misma, esto incluye mostrar los datos que pueden ser relevantes, como la literatura, las propiedades físico-químicas, biológicas, clínicas y los efectos toxicológicos de un nanomaterial.

De esta conferencia los investigadores pudieron sacar conocimientos sobre extraer información a partir de conjuntos de datos en bruto, la cual implica varios pasos:

1. La recolección de datos;
2. El análisis de datos y anotación;
3. La extracción de información; y, finalmente,
4. La validación de datos por expertos en el campo.

Pero nos damos cuenta que estos son los típicos pasos para clasificar todos los tipos de informaciones de cualquier tipo de ciencia, pero lo que diferencia es que la nanoinformática se enfrenta a muchos desafíos y limitaciones dentro de estos datos, como las definiciones de la semántica en el campo; la existencia de numerosos repositorios de datos, la integración de datos heterogéneos que todos los sistemas de salud reconozcan como general, la falta de normas para lograr esta heterogeneidad de datos y recursos y por ultimo y no menos importante la integración con herramientas de análisis de datos en estos repositorios de información. Por ejemplo si hablamos de la recolección de datos, actualmente enfrentamos serias restricciones para acceder a los datos de nanomateriales, es mas se pidió en este trabajo de fin de máster la ayuda de un repositorio para poder obtener los datos, pero jamás fue contestado y así concluimos que la mayoría tienen la falta de voluntad para compartir experiencias y resultados, o la existencia de datos de propiedad que no está disponible públicamente. Además, en algunos casos, la mayoría de investigadores o personas que se dedican a producir a estos resultados experimentales reportados y que son explicados en la literatura científica contienen errores o no se describen en términos de protocolos exactos con parámetros experimentales o técnicas, por lo que no es posible reproducir el experimento, como normalmente debe de ser. Por lo tanto, con el fin de extraer información de alta calidad a partir de conjuntos de datos en bruto, se necesita un gran esfuerzo para validar y analizar repositorios existentes, eliminar basura -datos no válidos o redundantes.

En el 2008 con la ayuda de la Universidad Politécnica de Madrid como coordinador de este proyecto presento un proyecto innovador que fue presentado a la Comisión Europea, la cual consiste en establecer bases de un diálogo entre los investigadores en Informática Médica, Bioinformática y Nanoinformática para proporcionar información sobre los recursos de información utilizadas y creadas por investigadores profesionales en las disciplinas mencionadas. El proyecto es llamado: ACTION-Grid [Figura 3.7], y así fue la primera iniciativa en nanoinformática a nivel europeo, destinado a ampliar la cooperación internacional en el campo entre los socios y expertos que anteriormente eran cuatro: Europa, los Balcanes Occidentales, África, América Latina y actualmente se unió los Estados Unidos. Al compartir esta información, la comunidad científica y de investigadores colaborando entre sí pueden desarrollar mejores repositorios de datos, bases de datos, herramientas o servicios que pueden ser utilizados por otros investigadores en estas áreas antes mencionadas y así identificar partes más relevantes en una investigación de este ámbito, y así obtener resultados que serán de gran ayuda para los futuros proyectos. Ahora ya no solo está la UPM dentro de este proyecto, tienen ahora muchos más socios como son: el ISCIII (Instituto de Salud Carlos III, FORTH (Foundation for Research and Technology), HIBA (Hospital Italiano de Buenos Aires), UTALCA (Universidad de Talca), la asociación de HEALTHGRID y UNIZG (Universidad de Medicina de Zagreb). [De la Iglesia, 2009]



Figura 3.7. Portada principal de la Web del proyecto ACTION-Grid

En los últimos años, las exploraciones de datos a gran escala han comenzado a combinar la ciencia experimental y la ciencia computacional usando métodos informáticos que utilizan redes de computación masiva, herramientas de ciencias de la información y las tecnologías de redes sociales; como son las base de datos tales como ISA-TAB-Nano -la cual ofrece un marco común para registrar e integrar descripciones de nanomateriales y usos, esta base define cuatro formatos de archivo para la distribución de los datos: archivos de investigación, archivos de casos de estudio, archivos de ensayos y archivos sobre materiales- [Thomas D.G., 2013] , caNanoLab -la cual investigadores la utilizan para la presentación y recuperación de información sobre las nanopartículas, incluyendo su composición, función, características físicas, y en vitro caracterizaciones experimentales- [Sharon G., 2013], Nanomaterial Registry -esta base de datos creada en la web ofrece datos relevantes caracterizados por cada nanomaterial que existe. Esta herramienta tiene doble funcionalidad: la introducción manual de datos y la carga de datos automática- [Ostraat, 2013] y Toxicology Data Network (TOXNET) -que es una base de datos que proporciona acceso a los datos de toxicología para unos 5.000 aproximadamente productos químicos, incluidos los nanomateriales- que ayudan en el intercambio de datos y el desarrollo de estándares de datos, es decir tratando de solucionar la heterogeneidad de términos, como la cantidad de datos de nanomateriales crece, estas bases de datos proporcionaran una manera de desarrollar métodos y herramientas específicas para la nanomedicina, o una ciencia nano afín. [Panneerselvam y Choi, 2014]

En este trabajo de fin de máster, hablaremos de una de las metodologías que ayudan a estos científicos, investigadores o informáticos para encontrar este tipo de información que es relevante para las diferentes investigaciones, justamente hablaremos de uno de los puntos mencionados en los objetivos que es el categorizar registros de ensayos clínicos. Y todo esto haremos mediante el uso de las herramientas de minería de texto a estos registros, pero aun así existen muchos tipos de herramientas para esto, en la actualidad no existen muchos clasificadores para la nanoinformática. Pero existen muchos clasificadores para poder categorizar muchas otras áreas, como la biología, medicina, física, química, etc. Y todas estas usan diferentes herramientas y diferentes clasificadores, y así justamente en una publicación se experimento cuál de estos clasificadores era mejor para dividir estos ensayos y se obtuvo que el mejor de todos fue el algoritmo de regresión logarítmica LASSO, con variables de IDF que fueron tomadas de los unigrams de los registros. Eso no significa que solo este clasificador podemos usar, ya que existen muchos, esto significa que fue el que mejor tuvo rendimiento y porcentaje positivo clasificador para los registro de nanomedicina. [De la Iglesia, 2013]

Y en el siguiente punto podemos observar los diferentes algoritmos que muchos investigadores actualmente usan para clasificar registros o repositorio en las categorías que ellos quieren dividir. Desde las más básicas pero que nos brindan buenos resultados a aquellas que tienen muchas variables, haciendo que sean más difíciles usarlos pero que tienen mejores resultados que los anteriores.

de datos puede ser más plenamente caracterizada como la extracción de implícita, previamente información desconocida, y potencialmente útil a partir de los datos [Witten y Frank, 2000]. Con la minería de texto, sin embargo, la información que se extrae está clara y explícitamente indicado en el texto, no está oculta dentro de los datos. El problema, por supuesto, es que la información no está expresada en una forma que es susceptible de procesamiento automático. La minería de texto se esfuerza por llevar de forma adecuada para que este texto vaya directamente a los ordenadores, sin necesidad de un intermediario humano.

Actualmente, la aplicación de la minería de texto sobre la nanotecnología está en una etapa temprana, algunos estudios recientes se han publicado como son: ayudar a los sistemas del conocimiento comparando sistemas terapéuticos y toxicológicos en sistemas de nanoescala [Gulke C., 2012], o un estudio de esta universidad que fue creado con mi tutor identificando entidades nanotoxicológicas en la literatura de nanomedicina [García Remesal, 2012] e incluso estudios para encontrar información relevante en el uso de nanopartículas [Lui R, 2014].

Entre los objetivos de la minería de texto incluyen la categorización de texto, el agrupamiento de texto, extracción de entidades y conceptos, la producción de taxonomías, análisis de la información, reconocimiento de patrones, distribución de frecuencia de palabras y el etiquetado de información. Y en este trabajo hablaremos sobre el etiquetado de información. Para ayudar a la categorización de esta información la minería de texto trabaja en conjunto con los sistemas de aprendizaje automático basado en la construcción de clasificadores que pueden operar con cualquier tipo de información.

Esta categorización de texto es la asignación de etiquetas a los documentos en lenguaje natural de acuerdo a su contenido. Este procedimiento es una técnica tradicional para el Big Data que se usa para la recuperación de información en bibliotecas, o lugares con grandes cantidades de datos, formando un "vocabulario estandarizado" o lista de palabras que es único en cada categoría. El objetivo es proporcionar este vocabulario estandarizado para todas las categorías de conocimiento, descendiendo a un nivel bastante específico, por lo que los libros-sobre cualquier tema, en cualquier lenguaje puede ser descrito de una manera que ayuda a los investigadores en el área puedan obtener todos los libros sobre un tema determinado [Witten I., 1999].

Actualmente para los informáticos la categorización de texto debe ser automática ya que este tipo de clasificación tiene muchas aplicaciones prácticas, incluyendo la indexación de documentos la recuperación, la extracción automática de metadatos, la detección de temas necesarios para la persona que está buscando, la organización de información y el mantenimiento de grandes catálogos de recursos Web. Y hace algunas décadas este tema siempre ha sido importante por su alta contribución y sobre todo por el enfoque dominante de la comunidad de investigación que la ha dado uso por medio de técnicas de aprendizaje automático para inferir categorías automáticamente de un entrenamiento conjunto de documentos previamente clasificados, como este trabajo de fin de máster.

3.2.1 Algoritmos clasificadores usados en aprendizaje automático

En esta parte del trabajo hablaremos de algunos algoritmos que son usados por los informáticos para encontrar estos datos relevantes. Existen un sin número de algoritmos de aprendizaje automático usado en minería de datos, así como tenemos publicaciones de muchos años atrás que empezaron a calificar estos algoritmos, con cuales serian los más efectivos en este tema, en una publicación realizada en el 2010 [Khan A., 2010] se dijo que los algoritmos que más se usan o más efectivos para este tipo de clasificación eran: Algoritmo de Rocchio [Rocchio, 1965] -es un método de espacio vectorial, este construye el prototipo de vectores para cada clase utilizando un conjunto de entrenamiento de los documentos-; el algoritmo de k-NN [Tam V., 2002] - el cual se utiliza para probar el grado de similitud entre documentos y k-datos de entrenamiento-; el árbol de decisiones -que reconstruye la categorización manual de documentos mediante la construcción de una estructura en forma de árbol decidiendo si el documento es verdadero / falso-; el algoritmo de reglas de decisión [Apte C., 1994] - método de clasificación utiliza el basado en normas o inferencia para clasificar documentos en categorías-; el algoritmo de Naïve Bayes [Brücher H., 2002] - que es un simple clasificador probabilístico basado en la aplicación de teorema de Bayes con una fuerte independencia a futuros supuestos-; las redes neuronales artificiales [Ruiz M., 1998] -que se construyen a partir de un gran número de elementos con una orden de las magnitudes mayores, estos elementos, son considerados como neuronas artificiales que están interconectados en grupo usando un modelo matemático-; la correlación Fuzzy o Difusa [Que H., 2000] -el cual puede hacer frente a la información incompleta, y convertir el valor de la propiedad en conjuntos difusos para la clasificación de documentos múltiples-; los algoritmos genéticos [Wang X., 2002] -que tiene como objetivo encontrar una característica óptima utilizando los mecanismos de la evolución genética y la supervivencia de los más aptos en la selección natural-; y los Support Vector Machines (SVMs) [Thorsten J., 1998] -que son uno de los métodos de clasificación mas discriminativo, ya que se basa en la minimización de riesgos estructurales mediante la teoría del aprendizaje computacional.

Otro estudio en años recientes [Korde V. y Mahender, 2012] también hablan de otras técnicas de las antes mencionadas, pero aumentan unas más : el LLSF, Linear Least Squares Fit [Yang, 1994] que da -un enfoque de mapeo de datos de entrenamiento y se representan en forma de pares de vectores de entrada / salida donde el vector de entrada es una documento en el modelo de espacio vectorial convencional (que consta de palabras con sus respectivos pesos), y el vector salida consta de categorías (con pesos binarios) del documento correspondiente-; y otros algoritmos que son en sí una mezcla de los anteriores mencionados, ya que la mayoría son métodos mezclados para una mejor precisión usando solo las cosas positivas de algunos.

Pero en este resumen nos enfocaremos en un estudio que realizó una gran organización de la computación, que es la empresa ORACLE, ya que su investigación se basa en un año más cercano al actual, y es en el año 2014 [Figura 3.9], que muestran que para la tarea de clasificación -que es lo que estamos realizando en este trabajo de fin de máster-

los algoritmos de aprendizaje automático más usados son las Regresiones Logísticas, Naïve Bayes, Support Vector Machine y por último los Árboles de Decisión. Estos algoritmos se usan para predecir si un documento o un repositorio de datos se etiqueten con una clasificación que se da por la persona o grupo de personas que están categorizando estos documentos.

Oracle Advanced Analytics Data Mining Algorithms and Functions SQL API

Oracle Advanced Analytics' Oracle Data Mining (ODM) component provides a broad range of in-database implementations of powerful workhorse data mining techniques and algorithms to solve many types of business problems:

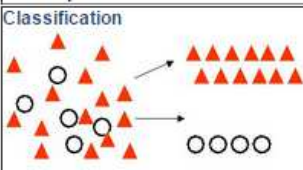
Technique	Applicability	Algorithms
<p>Classification</p> 	<p>Most commonly used technique for predicting a specific outcome such as response / no-response, high / medium / low-value customer, likely to buy / not buy.</p>	<p>Logistic Regression — classic statistical technique but now available inside the Oracle Database and supports text and transactional data</p> <p>Naive Bayes —Fast, simple, commonly applicable</p> <p>Support Vector Machine— Next generation, supports text and wide data</p> <p>Decision Tree —Popular, provides human-readable rules</p>

Figura 3.9. Parte de la página Web de Oracle que habla acerca de los algoritmos de aprendizaje automático más usado para clasificaciones

Todos estos algoritmos antes de ser usados, se debe tener en cuenta que primero se debe crear un modelo con datos manualmente clasificados para que el algoritmo vaya aprendiendo y se cree una bolsa de palabras para que sepa que etiqueta colocar al documento revisado, a estos datos revisados manualmente se le denomina documentos para el TESTING del algoritmo.

Hablaremos solo de los 4 algoritmos que se nombro anteriormente ya que estos también se encuentran dentro de la Herramienta llamada Weka que utilizamos para predecir y clasificar los repositorios que tenemos.

Realizaremos un estudio con su definición, su forma de clasificar, sus ventajas y desventajas de cada uno dentro de una tabla para su comparación:

- Regresión Logística LASSO [Tabla 3.1]
- Naïve Bayes [Tabla 3.2]
- Support Vector Machines (SVM) [Tabla 3.3]
- Árboles de decisión [Tabla 3.4]

Regresión Logarítmica			
Definición	Ventajas	Desventajas	Método de Clasificación
Es un algoritmo que se usa para predecir las categorías por medio de regresión, usando la función LOGIT ya que puede borrar en el resultado los datos que no se usan.	<ul style="list-style-type: none"> * Usadas en ciencias medicas por su precisión y calculo de probabilidad por su función LOGIT * Se usa para grandes bases de datos * Existen diferentes tipos pero la mas usada es LASSO * Es facil incorporar datos para el entrenamiento 	* El tiempo para obtener los valores para el testing es alto computacionalmente	Este método primero debe obtener un conjunto de datos del testing, con la suma de los valores absolutos de los paretros como limite superior para su comparación. Aplica una restricción por la función LOGIT y aplica una penalización de Lagrange para que las probabilidades sean mas facil de estimar.

Tabla 3.1. Tabla explicativa de la Regresión Logarítmica

Naïve Bayes			
Definición	Ventajas	Desventajas	Método de Clasificación
Es un clasificador probabilistico fundamentado en el teorema de Bayes, que contiene variables predictoras que ayudan a la clasificación	<ul style="list-style-type: none"> * Es fácil de implementar y hacer una comparación computacional con otros algoritmos * Trabaja muy bien con datos numericos y datos textuales. 	<ul style="list-style-type: none"> * La independencia condicional de las variables predictoras a veces están fuera del rango de los datos reales * Y no funciona tan bien cuando los criterios están muy correlacionados 	Se obtiene un modelo tomado en un training, para darle valor a criterios (probabilidad) para la clasificación, y en el testing se aplica el teorema de Bayes en donde la probabilidad de la palabra dada debe estar dentro del cluster seleccionado por los criterior del training

Tabla 3.2. Tabla explicativa del algoritmo de Naïve Bayes

Support Vector Machine			
Definición	Ventajas	Desventajas	Método de Clasificación
Este algoritmo es un modelo que representa a los puntos de muestra en el espacio, separando las clases dadas por una distancia que ha sido calculada como la mas lejana posible.	<ul style="list-style-type: none"> * Al igual que el Naïve Bayes trabaja muy bien con datos numericos y con datos textuales * Es muy fácil de implementar y su tiempo computacional es bajo * Son facilmente escalables a grandes bases de datos 	<ul style="list-style-type: none"> * La independencia condicional de las variables predictoras a veces están fuera del rango de los datos reales * Y no funciona tan bien cuando los criterios están muy correlacionados 	Se crea un conjunto de datos en el entrenamiento y se crean vectores con pesos, y se aplica la diferencia en el testing de cada vector para clasificar el documento dado.

Tabla 3.3. Tabla explicativa del SVM

Árboles de Decisión			
Definición	Ventajas	Desventajas	Método de Clasificación
Es un modelo de predicción que se construye con nodos en forma de árbol, dándole a cada nodo un valor de criterio para poder clasificarlos.	<ul style="list-style-type: none"> * Las generación de reglas para la creación son faciles * Reduce el problema de complejidad * Es muy fácil enterderlo 	<ul style="list-style-type: none"> * El tiempo de entrenamiento es muy largo * Si existe un error en una parte superior del arbol, todo el arbol posterior esta incorrecto * Un documento esta conectado a una sola rama * Puede sufrir un overfitting * No puede mantener variables continuas 	Se entrena una parte del repositorio y se crea la lista asociada de clasificación con atributos y otra lista de atributos probables, Después de esto a cada documento a clasificar se le asigna un atributo para determinar el cirterio que lo clasifica en la mejor particion en una clase individual , llevando los mejores valores desde el primer nodo hasta el ultimo.

Tabla 3.4. Tabla explicativa del algoritmo Árbol de Decisión

3.2.2 Herramienta WEKA

Esta herramienta es usada en este trabajo de fin de máster para crear el modelo de clasificación, para el training del repositorio de entrenamiento y para el testing de los repositorios a usarse en el este estudio. La herramienta llamada Weka (Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») [Figura 3.10] es una plataforma de software que contiene un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software ha sido desarrollado en la universidad de Waikato (Nueva Zelanda) bajo la Licencia Pública General de GNU-GPL. Además Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos como: las transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. La palabra Weka proviene de las islas de Nueva Zelanda, y se llama así a un tipo de ave robusta de color marrón, del tamaño de un pollo e incapaces de volar. Generalmente miden alrededor de 50 cm de alto y pesan en torno a 1 kg.



Figura 3.10. Logo de la Página Web de WEKA

Históricamente la versión original de Weka es un front-end en TCL/TK para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para pre-procesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura, pero la versión más reciente y estable basada en Java (WEKA 3.6), que es nombrada en el libro de data mining (Prácticas herramientas y técnicas de aprendizaje automático), la cual empezó a desarrollarse en 1997, se utiliza en muchas áreas, en particular con finalidades docentes y de investigación.

Weka en su primera pantalla muestra en qué tipo de aplicación se necesita usar esta herramienta [Figura 3.11]:

1. *Explorer*.- esta opción permite llevar a cabo las ejecuciones de los algoritmos de análisis de Weka sobre ficheros de entrada, tomando una ejecución independiente por cada prueba que se realiza.
2. *Experimenter*.- esta opción permite hacer experimentos más complejos, con el objeto de ejecutar más de un algoritmo sobre los ficheros de entrada, y al final hacer la comparación estadística de cada uno.
3. *Knowledge Flow*.- esta opción permite hacer la misma función que explicamos con la opción Explorer pero sus opciones se las hace gráficamente, muy parecido a la herramienta "data-flow" para seleccionar lo necesario y conectarlos a los ficheros de entrada y sus procesos.
4. *Simple CLI*.- esta opción es para expertos en esta herramienta, ya que es un ventana de comandos, y se utiliza para ejecutar por medio de comandos las opciones para ejecutar las funciones que se desea dar.4



Figura 3.11. Principales funciones de la herramienta Weka

Tomando los datos de la página de Weka esta herramienta se ha descargado aproximadamente unas 680.000 veces desde que fue puesto en SourceForge en abril 2000, y actualmente investigadores, docentes, empresas u organizaciones descargan un promedio de 10.000 / mes. La lista de correo Weka cuenta con más de 1.100 suscriptores en 50 países, existen 15 proyectos sustanciales bien documentados que incorporan usos de esta herramienta, y sin duda muchos más que no han sido reportados en Sourceforge. Los creadores de Weka Ian H. Witten y Eibe Frank también han escrito un libro muy popular "Minería de Datos: Aprendizaje Automático Práctico Herramientas y Técnicas". Este libro se convirtió en uno de los libros de texto más populares para la minería de datos y aprendizaje automático, y es muy frecuentemente citado en las publicaciones científicas.

Weka usa muchas librerías para solucionar estas clasificaciones, pero la que vamos a usar en este trabajo de fin de máster es la llamada LibSVM, librería creada en el año 2008 en la Universidad de Taiwan y ahora ya está en la versión 1.96 que fue creada en el 2014.

3.2.3 Librería LibLINEAR

LIBLINEAR es una librería de código abierto para la clasificación lineal a gran escala. Es compatible con la regresión logística y los algoritmos de máquinas de vectores de soporte lineales. Esta librería proporciona herramientas de líneas de comandos muy fáciles de usar. Y por la cantidad de publicaciones que usan esta librería, la mayoría de experimentos por no decir todos, demuestran que LIBLINEAR es muy eficiente y en su gran mayoría los experimentos son aplicados en grandes conjuntos de datos dispersos. [Fan y Chan, 2008].

Este conjunto de algoritmos con ayuda de muchos investigadores están creados en diferentes lenguajes de programación para su implementación, aunque no en todos los lenguajes esta actualizado con la última versión. [Tabla 3.5]

Dentro de la última versión de LibLINEAR podemos usar los siguientes algoritmos de clasificación:

- Clasificador de Regresión Logística L2 regularizada
- Clasificador de Regresión Logística L1 regularizada (LASSO)
- Clasificador de SVM (Support Vector Machine)
- Clasificador de SVR (Support Vector Regression)
- Clasificador de SVMR (Support Vector Machine Regularized)
- Clasificador de SVRR (Support Vector Regression Regularized)

Lenguaje	Creadores y Asistentes de Mantenimiento	Versiones de la librería
MATLAB	Universidad Nacional de Taiwan	1.96
Octave	Universidad Nacional de Taiwan	1.96
Java	Benedikt Waldvogel	1.92
Python	Universidad Nacional de Taiwan	1.96
Python	Uwe Schmitt	1.32
Ruby	Kei Tsuchiya	1.93
Perl	Koichi Satoh	1.93
Weka	Benedikt Waldvogel	1.5
R	Thibault Helleputte	1.94
Common LISP	Gabor Melis	1.92
Scilab	Universidad Técnica de Berlin	1.8

Tabla 3.5. Principales Lenguajes en los que está desarrollada la librería LibLINEAR y sus respectivas versiones

El diseño de esta librería está inspirado en el paquete LIBSVM, y contiene herramientas que son llamados por medio de líneas de comando para aplicar una tarea específica. Esta librería LibLINEAR comparten con LibSVM un uso similar, así como interfaces de programación de aplicaciones (API), para que los usuarios/desarrolladores pueden utilizar fácilmente estos dos paquetes. Sin embargo, sus modelos de clasificación después del entrenamiento son muy diferentes, debido a estas diferencias, es mejor no combinar estos dos paquetes juntos.

Todo el procedimiento; es decir, creación del modelo y las pruebas, tarda menos de 15 segundos en un equipo medio, aunque esto depende del número de registros que se usa para hacer este modelo y el número de documentos para las pruebas. El tiempo de entrenamiento sin incluir disco I/O es de menos de un segundo. Por esta razón se puede decir que es muy sencillo ejecutar LibLINEAR, aunque para informáticos expertos que usan estas herramientas tienen varios parámetros que están disponibles para el uso avanzado. Por ejemplo, se puede especificar un parámetro para obtener un conjunto de salidas de probabilidad aplicando la regresión logística L1, con solo una línea de código que especifica este número de opción, que es este caso es el número 6. [Figura 3.12]

```

__author__ = 'charlles'
from liblinearutil import *

y,x = svm_read_problem('train_2')
m= train(y, x, '-s 6')
save_model('tesis7.model', m)
#m = load_model('tesis6.model')
p_label, p_acc, p_val = predict(y, x, m)

```

Figura 3.12. Código demostrando que la opción “6” define el uso del algoritmo Regresión Logística

3.3 Minería de texto en nanomedicina

En esta parte del capítulo de estado del arte vamos a revisar las últimas actualizaciones en nanoinformática, aunque la mayoría de estudios han sido publicados por mi tutor de tesis, mi profesor de Bioinformática Médica y por la autora de la tesis doctoral que he usado para hacer este trabajo de fin de máster, Phd. Miguel García, Phd. Víctor Maojo y Phd. Diana de la Iglesias respectivamente. Esto no quiere decir que no existan otros autores pero en sí la mayor parte de publicaciones actuales sobre nanoinformática son de estos 3 autores.

Una parte importante de estos tipos de publicaciones es que buscan que la aplicación de cualquier clasificador pueda reducir drásticamente el esfuerzo humano en la identificación de los informes pertinentes de la gran piscina de ensayos que se tienen para una mayor investigación del tema que se desea obtener. Pero siempre es bueno considerar que el clasificador siempre se debe basar en una base de datos que contenga registros reales y así se pueda lograr un alto rendimiento en el filtrado de los informes pertinentes.

En el 2009 un Doctor publicó una publicación muy peculiar especificando el estado del arte hasta ese año, definiendo y enumerando diferentes experimentos que iban desarrollando poco antes de llegar a la nanoinformática, describiendo en sí la nanotecnología y dando a entender como poco a poco se ha llegado tan lejos, y diciendo que en ese año recién se empezó a usar la nanoinformática y que solo estamos comenzando, dejando en sí en claro que la nanoinformática tiene una tarea difícil de seguir. [McGonigle D., 2009]

En los últimos años se han mejorado técnicas para la minería de texto en muchas áreas, inclusive en la medicina, revisando el material en la Web, existen muchas publicaciones sobre cómo encontrar, separar y clasificar los registros de ensayos médicos, pero no de nanomedicina, si no de otra rama de medicina. En el año 2014 tenemos un ejemplo que se usa con los ensayos radiológicos que de igual forma que este trabajo de fin de máster se leyeron los ensayos clínicos para poder crear un sistema automatizado con algoritmos de aprendizaje activo supervisado para la clasificación de los exámenes, con el objetivo de separar los informes que describían una enfermedad cancerosa y los que no. Además en esta publicación de clasificación de estos ensayos se usó otros métodos de aprendizaje supervisado tradicionales como campos aleatorios condicionales y máquinas de vectores soporte, pero al final el aprendizaje activo (AL) fue el que mejor éxito tuvo para optimizar la producción de la formación y mejorar aún más el rendimiento de clasificación. En estas pruebas el rendimiento del clasificador logra un 98,25% de sensibilidad y especificidad 96,14% en el set de prueba celebrada fuera del registro de cáncer. [Dung H., 2014]

Otra publicación realizada en el año de 2014 trata de solucionar la clasificación de diferentes etiquetas para documentos que estén asociados con un subconjunto de categorías, y se escogió esta publicación porque se encuentra publicada en una revista

de Nanotecnología llamada "Journal of Nanoscience and Nanotechnology". Y se centra en la selección de características para reducir la dimensionalidad de los vectores de palabras que existen en una publicación, además, se uso el clasificador por K-vecinos más cercanos (KNN) para predecir la categoría en el cual debe estar siguiendo el esquema de este clasificador de aprendizaje automático. [Gayathri K., 2014]

Una publicación del departamento de Biomedicina Informática de la UPM sacado en el año 2008 propuso la creación de nuevos métodos para organizar automáticamente los recursos nanomédicos en función de sus funcionalidades, este trabajo es una parte del proyecto ACTION-Grid, ya antes mencionado. Esta publicación presenta una breve reseña sobre la disciplina nanomédica y sus tecnologías relacionadas e introduce un método para la creación automatizada de un índice de recursos nanomédicos, la cual se basa en un enfoque existente para construir un índice de recursos biomédicos de trabajos de investigación utilizando la minería de texto. [Chiesa S., 2008]

Otra publicación que habla de este ámbito es una hecho por mi tutor en el año 2013, la cual, se centro a dar soluciones para facilitar la investigación nanotoxicología dentro del texto medico. Para esto, adopto un enfoque computacional para reconocer automáticamente y extraer entidades relacionadas a la nanotoxicología dentro de la literatura científica, y clasifico estas entidades en cuatro categorías diferentes: nanopartículas, las vías de exposición, efectos tóxicos, y objetivos. Obteniendo buenos resultados que demostraron la eficacia de utilizar métodos computacionales para realizar de forma fiable este trabajo y llamo así este proceso como Reconocimiento de Entidad (NER). [García-Remesal M, 2013]

Un estudio muy parecido al anterior, fue uno publicado en el 2013, por la Universidad de Tel Aviv en Israel que fue desarrollado para predecir la nanotoxicología que existe en las nanopartículas de ferrita de cobalto, ya que generalmente estas nanopartículas son muy usadas para las terapias que usan nanotecnología. La conclusión que este estudio saco fue que es muy importante conocer la correlación observada entre el “estrés oxidativo”, que es provocado por la presencia de las nanopartículas de Ferrita de Cobalto con las diferentes células. El algoritmo que se uso para poder aplicar el modelo de clasificación para observar esta toxicidad fue un tipo de clasificador supervisado llamada Árbol de Decisión J48 que se encuentra en la herramienta Weka, y además para saber que era un buen clasificador se lo comparo con el algoritmo de Naives Bayes. [Horev-Azaria L., 2013]

En el 2013, en el workshop de la IEEE de Nanoinformática para Biomedicina, algunos investigadores sacaron una publicación demostrando diferentes aspectos de la bioinformática para apoyar la investigación biomédica y para avanzar en el conocimiento sobre las interacciones de los nanomateriales. Ellos estudiaron diferentes temas que incluían curación de datos, estándares de datos, minería de datos y el modelo predictivo por medio algoritmos de aprendizaje automático. Uno de ellos creo modelos predictivos de los efectos biológicos de los nanomateriales -por ejemplo, la mortalidad, la malformación, inhibición del crecimiento- utilizando algoritmos árboles de regresión; otro del grupo de investigadores se centraron en la predicción de la supervivencia en

pacientes con cáncer utilizando sus perfiles y se modelaron con algoritmos K-NN y regresión; otro grupo que participo aplicaron conjuntos aproximados para clasificar los tipos de cáncer que existen en el ser humano -por ejemplo, tumor de pulmón, tumor de hígado, etc.- por medio del algoritmo de Support Vector Machine (SVM). [Liu, Xiong y Webster, 2013]

Otros estudios en el año 2015, de diferentes países fueron citando muchos trabajos del grupo de Biomedicina Informática, citando publicaciones de nanoinformática, como por ejemplo en India se escribió una publicación definiendo algunos términos de la nanotecnología, nanociencia, nanoinformática, etc. Este trabajo es muy interesante ya que define conceptos puntuales y sus futuros retos en el campo de la nanoescala. Por último se concluye con la explicación de nanoinformática, dando a entender que es una ciencia emergente que podría ser ampliamente utilizada para evolucionar con nuevos fármacos y puede ser utilizado para conservar la información “nano” en todos los aspectos. [Manonmani V, 2015]

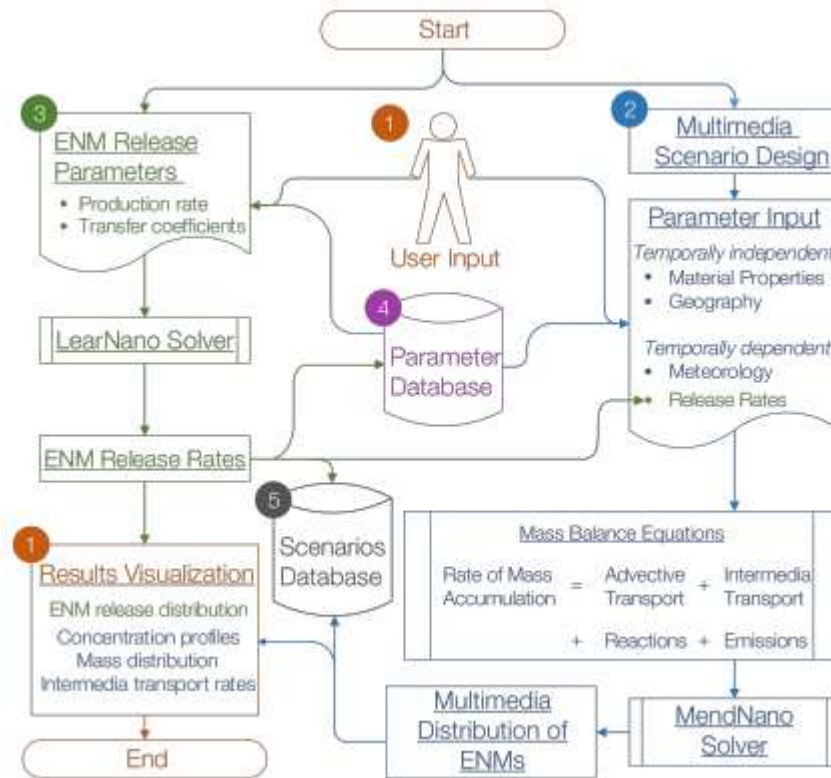


Figura 3.13. Explicación de los pasos de la herramienta RedNano, sacado de su publicación

Otra publicación en el año 2015 muestra el uso de una nueva herramienta de simulación integrada, que ha sido desarrollada para evaluar la posible liberación y distribución ambiental de los nanomateriales, llamada RedNano. La herramienta de simulación Rednano y su aplicación de software basada en la web permiten hacer predicciones de

análisis de escenarios, a fin de evaluar la respuesta de un sistema ambiental a diversos escenarios de liberación de los nanomateriales artificiales. Esta herramienta se divide en 5 partes principales [Figura 4.17], la primera es el interfaz del usuario para la visualización de resultados y el diseño del problema, la segunda parte es llamada MedNano que es un modelo ya predeterminado para los sistemas ambientales, la tercera parte es la aplicación del algoritmo de predicción llamado LearNano, la cuarta parte es la base de datos de los parámetros que se estiman y por último es un repositorio para construir escenarios y casos de simulación. Además, la herramienta de simulación Rednano según esta publicación se ha usado para la investigación, académicas, y propósitos regulatorios. Y para concluir se puede decir que esta herramienta puede servir como un apoyo a las decisiones de forma rápida y críticamente evaluar las posibles repercusiones ambientales de la nanomateriales y así asegurar que la nanotecnología se desarrolla en un productivo y de manera responsable con el medio ambiente. [Liu H., 2015]

3.4 Ensayos clínicos

Un ensayo clínico es una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, en su aplicación a seres humanos, pretende valorar su eficacia y seguridad. Los estudios de prometedores tratamientos nuevos o experimentales en pacientes se conocen como ensayos clínicos. Un ensayo clínico se realiza sólo cuando hay razones para creer que el tratamiento que se está estudiando puede ser beneficioso para el paciente. Los tratamientos usados en los ensayos clínicos con frecuencia demuestran tener beneficios reales. Los investigadores realizan estudios sobre nuevos tratamientos para conocer la utilidad del nuevo tratamiento, el mecanismo de acción del nuevo tratamiento, si la efectividad es mayor que otros tratamientos ya disponibles, los efectos secundarios del nuevo tratamiento y si son mayores o menores que el tratamiento convencional, si supera los beneficios a los efectos secundarios y en qué pacientes el nuevo tratamiento es más útil.

Según la página MedLine Plus, la cual está encargada la NIH (Biblioteca Nacional de medicina de EEUU), define a los ensayos clínicos como estudios de investigación que se prueban que tan efectivo pueden aplicarse los descubrimientos médicos, biomédicos o biológicos a los pacientes. Por lo cual cada ensayo clínico se establece con una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, en su aplicación a seres humanos, pretende valorar su eficacia y seguridad. Cada estudio que se realiza con estos ensayos clínicos responden preguntas científicas e intentan encontrar mejores formas de prevenir, examinar, diagnosticar o tratar una enfermedad, inclusive los más importantes ensayos clínicos para los investigadores es aquellos que permiten la comparación de nuevos productos o tratamientos contra los que ya existen, para poder sacar nuevas conclusiones.

Todo ensayo clínico se inicia cuando surge una hipótesis que suele darse a partir de estudios no controlados observacionales, descriptivos o retrospectivos, o de estudios preclínicos. Frecuentemente se realizan actividades médicas cuya utilidad no ha sido demostrada mediante un ensayo clínico, sin embargo llevarlo a la práctica es difícil, sobre todo por el costo económico y además el tiempo que un ensayo clínico lleva puede ser de meses o años, este factor hace que a veces pocos investigadores dejen estos ensayos a medias ya que no cuentan con el dinero suficiente para poder seguir. Además los investigadores corren el riesgo de que se descubran en investigaciones preclínicas posibilidades terapéuticas que no tienen ningún beneficio en un ensayo clínico, y esto les hace perder más tiempo del que tienen. Se debe aclarar que todo ensayo clínico tiene un protocolo a seguir. Este plan después de ser diseñado debe ser aprobado por un comité de bioética, ya que los pacientes que forman parte deben conocer los objetivos del estudio, sus riesgos y beneficios y firmar el consentimiento informado y por consiguiente podrán abandonar el estudio cuando ellos deseen. El ensayo clínico finaliza cuando acaban los plazos de tiempo definidos en el protocolo, o cuando de forma prematura son manifiestamente perjudiciales o beneficiosos los efectos en el brazo experimental.

En este trabajo de fin de máster ensayos clínicos solo me referiré a aquellos que nos interesan, como son los que dentro de su estructura usan términos nano médicos, como son las nanopartículas o los nanomateriales o algún referente a este tema. A continuación expondré unos ejemplos de estos dos tipos de ensayos clínicos.

3.4.1 Ensayos clínicos sobre nanopartículas

Para diferenciar un ensayo clínicos sobre nanopartículas se deben tomar en cuenta una cosa en particular y ese es el tipo de droga que se está utilizando. Muchos medicamentos nuevos ya tienen nombre comercial de cada nanopartícula que se ha utilizado, para lo cual con ayuda de estos nombres es más fácil encontrar un ensayo clínico que tenga que ver con nanomedicina.

La mayoría de estos nombres se encuentran en la publicación llamada “Vacunas de Nanopartículas” [Zhao and Seth, 2014], el cual muestra un sin número de productos ya con nombres comerciales como lo había mencionado anteriormente, así si uno de estos términos es usado en un ensayo clínico, es más fácil saber si se trata o no un registro con etiqueta “nano”. La herramienta que estamos creando hará que automáticamente la computadora llegue a designar esta etiqueta a uno de los ensayos clínicos de ejemplo que vamos a poner para que el algoritmo funcione y lo etiquete correctamente.

Un ejemplo de los productos más usados con nanopartículas es el medicamento llamado Abraxane® [Figura 3.14], que es un medicamento inyectable de paclitaxel, un fármaco inhibidor mitótico utilizado en el tratamiento de cáncer de mama, cáncer de pulmón y cáncer de páncreas. Este fármaco, el paclitaxel se une a la albúmina creando un medio

de suministro en contra de estos males. También es conocido como nab-paclitaxel (que significa que está constituido de nanopartículas de paclitaxel).



Figura 3.14. Fármaco Abraxane® para el consumo humano

Otra de las funciones importante que hace que la nanotecnología cada día incremente su papel significativo en el desarrollo de medicamentos, es el desarrollo de vacunas que conlleva a crear composiciones menos inmunogénicas "minimalistas", estas formulas aumentan la eficacia del antígeno que hace combatir a la enfermedad la cual está atacando al cuerpo. El uso de nanopartículas en las creaciones de nuevas vacunas no sólo permite mejorar la estabilidad del antígeno y la inmunogenicidad, sino que también hace que este actúe más lento y haga lo que debe hacer. Muchas de las vacunas creadas con nanopartículas las cuales varían en su composición, tamaño, forma y propiedades físicas han sido aprobadas para el uso humano y cada día el número de estas nanovacunas está aumentando.

Sin embargo, sigue habiendo problemas debido a la falta de conocimientos fundamentales sobre el comportamiento directo de estas nanopartículas, por lo cual se está trabajando en muchos ensayos clínicos, ya que estas pueden operar como un sistema de suministro para mejorar el procesamiento de antígenos y / o como un agente inmunoestimulante para activar o mejorar la inmunidad. Esta revisión proporciona una visión general de los avances recientes en el estudio de estas nanovacunas. Muchas de las nanovacunas conocidas son las siguientes: INFLEXAL® [Figura 3.15] que esta constituido por un liposoma llamado por sus siglas DOTAP®, que es una fórmula de carbonos, otra llamada EPAXAL® que también es otra vacuna conocida por estar creada por nanopartículas. Existen muchos otros medicamentos ya aprobados o en condiciones de uso comercial, como los de la terapia contra el cáncer -Daunoxome® [Petre y Dittmer, 2007], Doxil® [Pearce, 2012] y Caelyx® [Li, 2012], entre otros -, además sustitutos óseos tales como CarriGen®, compuestos dentales tales como FilTek® [Moosavi, 2012], o plataformas como el Sistema Verigene® [Buxton, 2009], que se utiliza ampliamente en los hospitales para fines de diagnóstico.



Figura 3.15. Nanovacuna Inflexal® para el consumo humano

Y así obtenemos muchos términos nuevos para estos ensayos clínicos. Generalmente estos ensayos ponen el nombre de la droga que van a usar en el título de su ensayo [Figura 3.16].

Gemcitabine, Cisplatin, and Abraxane in Advanced Biliary Cancers

This study is currently recruiting participants (See Contacts and Locations)
 Verified April 2015 by M.D. Anderson Cancer Center

Sponsor:
 M.D. Anderson Cancer Center

Collaborator:
 Celgene Corporation

Information provided by (Responsible Party):
 M.D. Anderson Cancer Center

ClinicalTrials.gov Identifier:
 NCT02392637

First received: March 13, 2015
 Last updated: April 6, 2015
 Last verified: April 2015
 History of Changes

Full Text View Tabular View No Study Results Posted Disclaimer How to Read a Study Record

Tracking Information	
First Received Date ICMJE	March 13, 2015
Last Updated Date	April 6, 2015
Start Date ICMJE	April 2015
Estimated Primary Completion Date	April 2018 (final data collection date for primary outcome measure)
Current Primary Outcome Measures ICMJE	Progression-Free Survival (PFS) of Gemcitabine, Cisplatin, and Abraxane in Advanced Biliary Cancers [Time Frame: 63 days] [Designated as safety issue: Yes]

Figura 3.16. Ejemplo de un ensayo clínico que habla sobre la nanopartícula Abraxane® en el título

Pero otros autores de ensayos clínicos no lo ponen ya que solo lo comparan la nanopartícula dentro del estudio que lo están haciendo con otra nueva partícula o

nanopartícula que ayude a mejorar el anterior [Figura 3.17], en el ejemplo propuesto, nos damos cuenta que esta el nombre de *Paclitaxel*, que es el compuesto del que está formado el Abraxane®, pero este compuesto no siempre es el de nanopartículas, el nombre de la formula que compone el Abraxane® es el nab-paclitaxel o llamado (ABI-007), esto implica que cuando vemos un la palabra “Paclitaxel” no siempre indica que este hablando del compuesto que contiene nanopartículas llamado Abraxane®, y por ende puede causar cierta complejidad al momento de buscar este compuesto en los diferentes ensayos.

Gemcitabine/Irinotecan/ZD1839 vs Paclitaxel/Carboplatin/Etoposide/ZD1839 in Carcinoma of Unknown Primary Site

This study has been completed.

Sponsor: SCRI Development Innovations, LLC

ClinicalTrials.gov Identifier: NCT00193596

First received: September 12, 2005

Last updated: March 22, 2013

Last verified: March 2013

History of Changes

Collaborators: AstraZeneca, Pharmacia and Upjohn, Eli Lilly and Company

Information provided by (Responsible Party): SCRI Development Innovations, LLC

Full Text View | Tabular View | Study Results | Disclaimer | How to Read a Study Record

Tracking Information	
First Received Date <small>ICMJE</small>	September 12, 2005
Last Updated Date	March 22, 2013
Start Date <small>ICMJE</small>	September 2003
Primary Completion Date	September 2008 (final data collection date for primary outcome measure)
Current Primary Outcome Measures <small>ICMJE</small> (submitted: March 22, 2013)	Overall Survival (OS), the Length of Time, in Months, That Patients Were Alive From Their First Date of Protocol Treatment Until Death [Time Frame: 24 months] [Designated as safety issue: No] Length of time, in months, that patients were alive from their first date of protocol treatment until death.
Study Design <small>ICMJE</small>	Allocation: Randomized Endpoint Classification: Safety/Efficacy Study Intervention Model: Parallel Assignment Masking: Open Label Primary Purpose: Treatment
Condition <small>ICMJE</small>	Neoplasms, Unknown Primary
Intervention <small>ICMJE</small>	<ul style="list-style-type: none"> Drug: Etoposide 50 mg alternating with 100 mg PO, days 1 and 10 in regimen A Other Names: <ul style="list-style-type: none"> Etopophos Toposar Drug: Gemcitabine 1000 mg/m² IV, days 1 and 8, in regimen B Other Name: Gemzar Drug: Irinotecan 1000 mg/m² IV days 1 and 8 in regimen B Other Name: Camptosar Drug: Paclitaxel 200 mg/m² day 1-hour IV infusion, day 1, regimen A Other Name: Abraxane Drug: Carboplatin Area under the curve (AUC) 6.0 IV, day 1, regimen A Other Name: Paraplatin
Study Arm (s)	Experimental: Regimen A

Figura 3.17. Ejemplo de un ensayo clínico que habla sobre la nanopartícula Abraxane® en el cuerpo

Es por esta razón la cual mejor confiamos en el algoritmo que en la búsqueda visual, ya que este lee todo el artículo y por medio del valor que aparezca define lo que es un ensayo medico sobre nanomedicina o no lo es.

3.4.2 Ensayos clínicos sobre nanomateriales

Primero debemos saber que los nanomateriales pueden ser divididos en nanopartículas, nanocapas y nanocompuestos, obviamente son considerados nanomateriales si su dimensión está entre 1 a 100 nm. El enfoque de los nanomateriales es una aproximación desde abajo hacia arriba a las estructuras y efectos funcionales de forma que la construcción de bloques de materiales que son diseñados y ensamblados de forma controlada, recordando que la mayoría son de uso médico, pero esto no significa que solo para esto son usados. En el futuro se dice que de estos nanomateriales estarán hechos los nanorobots que también lucharan contra las células malignas dentro del cuerpo humano.

En la actualidad cientos de productos que contienen nanomateriales ya están en uso en cientos de lugares, como por ejemplo dentro de las baterías, revestimientos, ropa antibacterianas, etc. Los analistas esperan que los mercados crezcan a cientos de miles de millones de euros, ya que la nanoinnovación se verá en muchos sectores, como en la salud pública, en el empleo, la informática, la industria, la innovación, el medio ambiente, la energía, el transporte, la seguridad y el espacio. Los nanomateriales tienen el potencial de mejorar la calidad de vida y contribuir a la competitividad industrial en el mundo entero. Sin embargo, los nuevos materiales también pueden presentar riesgos para el medio ambiente como lo hablamos en los capítulos anteriores por la existencia de la toxicidad y crea preocupaciones en la salud y seguridad. Estos riesgos, siempre han sido objeto de varios dictámenes del Comité Científico de los Riesgos Sanitarios Emergentes y Recientemente Identificados (CCRSERI). La conclusión general es que hasta ahora, a pesar de que los nanomateriales no son de por sí peligrosos, todavía hay incertidumbre científica sobre la seguridad de los nanomateriales en muchos aspectos y por lo tanto la evaluación de la seguridad de las sustancias debe hacerse sobre una base de caso por caso.

La Organización para el Desarrollo Económicos Corporativo, Dirección de Medio Ambiente realizó una lista actualizada de los nanomateriales manufacturados hasta ese momento, esta tuvo discusión que tuvo lugar en la séptima reunión del Grupo de Trabajo de la OCDE sobre Nanomateriales Manufacturados (WPMN) en julio de 2010. La lista es la que está a continuación [Figura 3.18]:

- Los fullerenos (nanotubos de carbono)
- Nanotubos de carbono de pared simple
- Nanotubos de carbono de pared múltiple
- Las nanopartículas de plata
- Nanopartículas de hierro
- El dióxido de titanio
- El óxido de aluminio
- El óxido de cerio
- El óxido de zinc
- El dióxido de silicio

- Los dendrímeros
- Nanoarcillas
- Las nanopartículas de oro

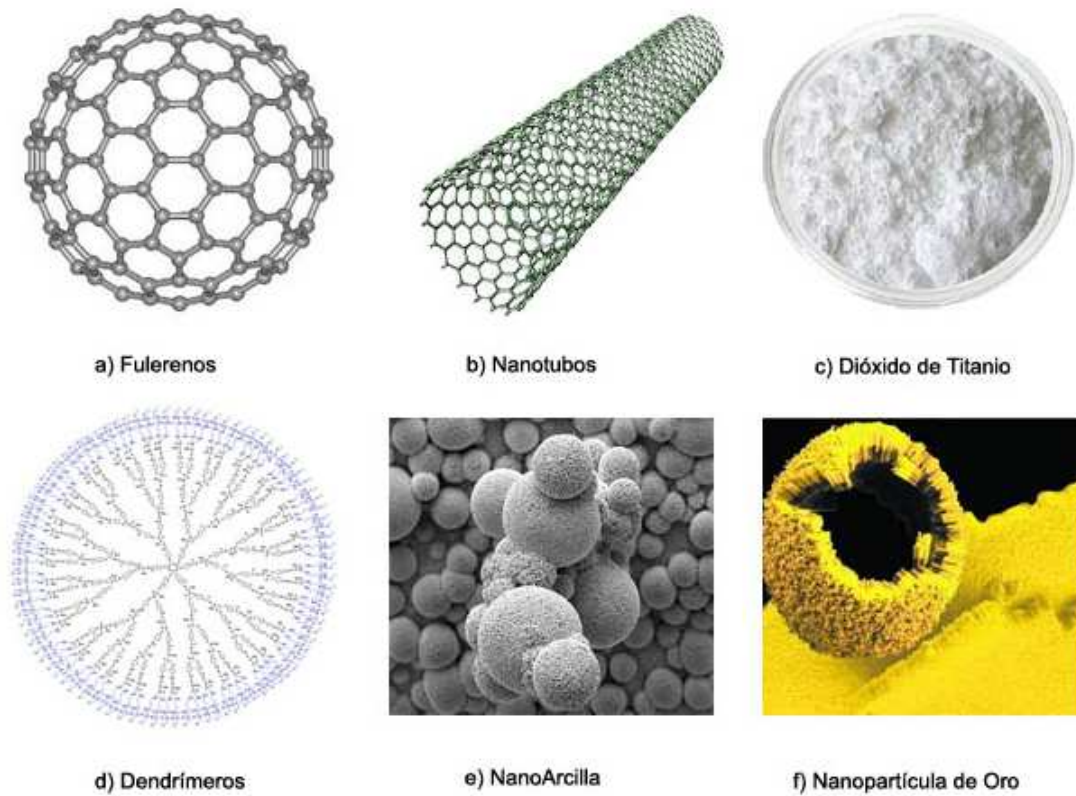


Figura 3.18. Nanomateriales mas usados en el campo medico, a) Fullerenos, b) Nanotubos, c) Dióxido de Titanio, d) Dendrímeros, e) Nanoarcilla y f) Nanopartículas de Oro

Una cosa importante es que los nanomateriales no son simplemente otro paso en la miniaturización de los materiales, estos a menudo requieren enfoques muy diferentes en su producción. Existen varios procesos para crear nanomateriales, aunque muchos nanomateriales están actualmente en fase de de fabricación dentro de los laboratorios, y otros nanomateriales se están comercializando. Otra cosa importante es que estos materiales en nano escala representan desarrollos incrementales gracias a sus propiedades mejoradas y la mayoría son usados en fabricar productos en los sectores de energía y en fabricar productos químicos, en este caso los farmacéuticos.

En los registros de ensayos clínicos, tenemos algo parecido al igual que la nanopartícula, pero en este caso el nombre de estos nanomateriales casi nunca están dentro del título del ensayo clínico siempre están en el cuerpo, por ende, para saber si están usando estos nanomateriales es necesario leer el cuerpo y así poder clasificarlo o

etiquetarlos, ya que la mayoría da uso a la fabricación de nanodispositivos o solo se dan a conocer como una nanopartícula, como los nanotubos que son usados en muchas fabricaciones de medicamentos y como en estos ensayos solo muestran el nombre del medicamento total y no su composición es muy difícil saber si estos fueron creados con estos materiales de nano escala, así como se muestra en el ejemplo en la Figura 3.19, en donde en este ensayo si se habla de la composición del medicamento a usar y se puede apreciar que habla sobre los nanotubos y las nanopartículas de oro.

Diagnosis of Gastric Lesions With Na-nose

The recruitment status of this study is unknown because the information has not been verified recently.
 Verified June 2013 by Anhui Medical University.
 Recruitment status was Recruiting

Sponsor:
 Anhui Medical University

Collaborators:
 Technion, Israel Institute of Technology
 University of Latvia

Information provided by (Responsible Party):
 Hu Liu, Anhui Medical University

ClinicalTrials.gov Identifier:
 NCT01420588

First received: August 18, 2011
 Last updated: June 23, 2013
 Last verified: June 2013
[History of Changes](#)

Full Text View Tabular View No Study Results Posted Disclaimer How to Read a Study Record

Descriptive Information

Brief Title ICMJE	Diagnosis of Gastric Lesions With Na-nose
Official Title ICMJE	Study of the Exhaled Breath of Patients With Malignant and Benign Lesions With Na-Nose
Brief Summary	<p>The investigators study the feasibility of a novel method in oncology based on breath analysis with nanosensors array for identifying gastric diseases. Alveolar exhaled breath samples collected from volunteers referred for upper endoscopy or surgery are analyzed using a custom-designed array of chemical nanosensors based on organically functionalized gold nanoparticles and carbon nanotubes. Predictive models are built employing discriminant factor analysis (DFA) pattern recognition method. Classification accuracy, sensitivity and specificity are determined using leave-one-out cross-validation or an independent blind test set. The chemical composition of the breath samples is studied using gas chromatography coupled with mass spectrometry (GC-MS).</p> <p>A pilot study is conducted first (enlistment of 160 subjects at the Department of Oncology, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China.)</p> <p>The pilot study is followed by a large-scale clinical trial to confirm the preliminary results of the Chinese pilot study (enlistment of 800 subjects at the Digestive Diseases Centre GASTRO, Riga East University Hospital, 6 Linezera iela, LV1006 Riga, Latvia). 25% of the samples are used as independent blind test set. The samples are blinded by the medical team and are not disclosed until prediction of blind sample identity is complete.</p>

Figura 3.19. Ejemplo de un ensayo clínico que habla sobre el nanomaterial Nanotubo en el cuerpo

Por esta razón es necesario un algoritmo clasificador que lea estos registros de ensayos clínicos para poder saber que etiqueta ponerle, compararlos con sus vectores de palabras o con sus bolsas de palabras, dependiendo del clasificador que se esté usando.

3.4.3 Ensayos clínicos sobre nanodispositivos

Los avances tecnológicos sobre nanotecnología que están relacionados con la medicina y todo aquellas maneras que tiene que ver con ayudar a los demás, está incrementando año a año y es que una de las últimas investigaciones por ejemplo en la Universidad de Harvard se centra en una cadena de ADN de un nanómetro de ancho [Kleiner R, 2011]. En si comenzando con estas investigaciones se pueden crear cuerpos diminutos que hagan funciones de este tipo, es decir, la creación de nanodispositivos auto ensamblados para facilitar el tratamiento desde la parte interior del cuerpo biológico, administrando

fármacos o incluso reprogramando las células madre para poder regenerar tejido óseo o neuronal. O ayudando a ciertos órganos a funcionar mejor.

Se debe conocer que estos nanodispositivos son creados por medio de nanomateriales o nanopartículas que fuimos mencionando en los puntos anteriores y así crean un sin número de estos dispositivos de nano escala que se usan en los registros de ensayo clínico en las diferentes ramas de medicina.



Figura 3.20. NA-NOSE, nanodispositivo usado para la respiración artificial

De igual forma que en el de nanopartículas, existen ciertos ensayos clínicos que esconden en el cuerpo el uso de estos nanodispositivos y otras que en su título lo colocan sin problema, siempre y cuando sepamos que el nombre del nanomaterial que estamos usando. [Figura 3.21].

En este caso habla de NA-NOSE [Figura 3.20] que es una nariz artificial hecho a nano escala y que fue inventada para los pacientes que tienen problema en los pulmones, para ser más específico el cáncer de los pulmones.

3.4.4 Ensayos clínicos no nanos

Son aquellos ensayos clínicos que no tienen nada que ver con la nanomedicina, y en todo su contexto usan cosas que no están dentro de este rango.

Después de conocer los tipos de ensayos clínicos, ya puedo exponerles el método que se creó para saber si un ensayo clínico es sobre nanomedicina o no.

Electronic Nose for Diagnosis of Neurodegenerative Diseases Via Breath Samples

The recruitment status of this study is unknown because the information has not been verified recently.

Verified September 2010 by Rambam Health Care Campus.
Recruitment status was Recruiting

Sponsor:
Rambam Health Care Campus

Collaborator:
Technion, Israel Institute of Technology

Information provided by:
Rambam Health Care Campus

ClinicalTrials.gov Identifier:
NCT01291550

First received: February 7, 2011
Last updated: NA
Last verified: September 2010
History: No changes posted

Full Text View Tabular View No Study Results Posted Disclaimer How to Read a Study Record

Descriptive Information	
Brief Title ICMJE	Electronic Nose for Diagnosis of Neurodegenerative Diseases Via Breath Samples
Official Title ICMJE	Not Provided
Brief Summary	<p>The diagnosis of neurodegenerative conditions and ADHD still mostly relies on clinical symptoms as there are no validated, inexpensive, and simple biomarkers available yet. The purpose of this study is to deliver a proof-of-concept for novel biomarkers to identify neurodegenerative conditions and ADHD based on breath testing.</p> <p>Alveolar breath will be collected from healthy volunteers, patients with extrapyramidal conditions, patients diagnosed with dementia and from ADHD subjects. The discriminative power of a tailor-made Nanoscale Artificial Nose (™NA-NOSE) containing an array of six nanomaterial-based sensors will be tested. Discriminant factor analysis will be applied to the NA-NOSE signals in order to detect statistically significant differences between the sub-populations, and classification success will be estimated using leave-one-out cross-validation. The identification of NA-NOSE patterns will be supported by analyzing the chemical composition of the breath using gas-chromatography in conjunction with mass-spectrometry (GC-MS).</p>
Detailed Description	Not Provided

Figura 3.21. Ejemplo de un ensayo clínico que habla sobre el nanodispositivo Na-NOSE en el cuerpo

3.5 Repositorios de registros de ensayos clínicos

En la web existen algunos repositorios de registros que contienen a los ensayos clínicos, en esta parte del capítulo hablaremos de los dos que usaremos en este trabajo de fin de máster, no sin antes hablar de los que actualmente son más usados. Un repositorio es un depósito o un archivo, es un sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos de cualquier tipo áreas.

En este trabajo sobre nanomedicina nos centraremos solo en repositorios que tengas registros médicos o publicaciones sobre medicina, en la página BVS (Biblioteca Virtual en Salud) de España enseña una lista de repositorios de publicaciones médicas:

- **BIOMED CENTRAL**

Es un repositorio de publicaciones científicas creada en Inglaterra especializada en publicaciones de acceso abierto. Todos los artículos publicados por este repositorio son de carácter investigativo, los artículos van quedando inmediata y permanentemente accesibles en línea, sin ningún tipo de cargo o barrera para ser leídos. Y es completamente seguro con su fidelidad de ser originales. (<http://www.biomedcentral.com/>) [Figura 3.22 a]

- **DIALNET**

Es un portal de publicaciones científicas de habla hispana que inició su funcionamiento en el año 2001, en la Universidad de Rioja España y estas publicaciones médicas están especializadas en ciencias humanas y sociales. Pero la diferencia en este repositorio es que no solo ofrece publicaciones científicas, también ofrece información como artículos de revistas, artículos de obras colectivas, documentos de trabajo-pre publicaciones, libros, etc. (<http://dialnet.unirioja.es/revistas>) [Figura 3.22 b]

- **DOAJ**

Es un directorio en línea, los cuales sus índices proporcionan acceso a revistas sobre ciencias de la salud, farmacia, lengua y literatura, arte y arquitectura, derecho y ciencias políticas, ciencias sociales, etc que son de alta calidad investigativa, y el acceso es abierto. (<https://doaj.org/>) [Figura 3.22 c]

- **HIGHWIRE PRESS**

Actualmente es una empresa estadounidense que ofrece el desarrollo de contenidos digitales, servicios de hosting y soluciones. La cual empezó como una división de las bibliotecas de la Universidad de Stanford, y se centra en editoriales universitarias y editoriales independientes que producen revistas de alto impacto, libros y otras publicaciones académicas. (<http://home.highwire.org/>) [Figura 3.22 d]

- **PLOS**

Es un repositorio de acceso libre que ofrece el contenido a revistas científicas y divide su contenido en diferentes páginas: ciencias biológicas, ciencias médicas, bioinformática, genética, ensayos clínicos, medicina tropical, etc. Sus siglas significan Public Library of Science (Librería Pública de Ciencia) y tiene su sede en la ciudad de San Francisco. (<https://www.plos.org/publications/journals/>) [Figura 3.22 e]

- **PUBMED CENTRAL**

Que más que en simple repositorio de documentos científicos, es considerada como el más grande repositorio digital gratuito que archiva artículos académicos de texto completo de acceso público que se han publicado en la literatura de revistas biomédicas y de ciencias de la vida. Este portal es actualizada y administrada por el National Institutes of Health (NIH), el National Center for Biotechnology Information (NCBI) y la National Library of Medicine (NLM) de EE.UU. (<http://www.ncbi.nlm.nih.gov/pmc/>) [Figura 3.22 f]

- **RECOLECTA**

Es una plataforma que agrupa a todos los repositorios científicos nacionales y que provee de servicios a los gestores de repositorios, a los investigadores y a los agentes implicados en la elaboración de políticas (decisiones públicas).(<http://recolecta.fecyt.es/>) [Figura 3.22 g]

Y así como esta página de España, mostrando listas de estos repositorios existen muchas en toda la Web. Pero generalmente estos repositorios son de publicaciones, revistas, y solo uno de repositorios de ensayos clínicos que lo revisaremos en otro punto de esta parte del capítulo, por ahora nos centraremos en los que se utilizó para el desarrollo de este trabajo de fin de máster, y estos son los repositorio de ensayos clínicos que muestran los trabajos actuales o pasados que se han hecho probando diferentes medicamentos o dispositivos para ayudar a los científicos o investigadores a saber sus fallas o sus ventajas que estos desean saber.

En estos siguientes puntos revisaremos los dos repositorios de ensayos clínicos más grandes en cada continente, en Norteamérica esta el Clinical Trials y en el continente europeo tenemos al Clinal Trials Register.



Figura 3.22. Logos de los diferentes repositorios sobre medicina en el mundo: a) BioMedical Central, b) DialNet, c) DOAJ, d) HighWire, e) PLoS, f) PubMed y g) Recolecta

3.5.1 Clinical Trials (Estados Unidos)

Esta página contiene uno de los más grandes repositorios de ensayos clínicos que existe en el mundo, actualmente contiene más 190.000 registros que contienen los resultados de los estudios clínicos mediante el apoyo público y privado de los participantes humanos, en este caso, científicos, médico o investigadores que fueron realizados en

todo el mundo, [Figura 3.22] este sitio web es mantenido y actualizado por la Biblioteca Nacional de Medicina (NLM) y los Institutos Nacionales de Salud (NIH).

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. Learn more [about clinical studies](#) and [about this site](#), including relevant [history](#), [policies](#), and [laws](#).

Find Studies | About Clinical Studies | Submit Studies | Resources | About This Site

ClinicalTrials.gov currently lists 192,809 studies with locations in all 50 States and in 189 countries. Text Size ▾

Search for Studies
Example: "Heart attack" AND "Los Angeles"
[Search Box] Search

Advanced Search | See Studies by Topic
See Studies on a Map

Search Help

- How to search
- How to find results of studies
- How to read a study record

Locations of Recruiting Studies

- Non-U.S. Only (53%)
- U.S. Only (41%)
- Both U.S. and Non-U.S. (6%)

Total N = 35,747 studies
(Data as of June 18, 2015)

- See more trends, charts, and maps

Learn More

- Tutorials for using ClinicalTrials.gov
- Glossary of common site terms
- For the Press
- Using our RSS Feeds

For Patients and Families

- How to find studies
- See studies by topic
- Learn about clinical studies
- Learn more...

For Researchers

- How to submit studies
- Download content for analysis
- About the results database
- Learn more...

For Study Record Managers

- Why register?
- How to register your study
- FDAAA 801 requirements
- Learn more...

HOME | RSS FEEDS | SITE MAP | TERMS AND CONDITIONS | DISCLAIMER | CONTACT NLM HELP DESK

Copyright | Privacy | Accessibility | Viewers & Players | Freedom of Information Act | USA.gov
U.S. National Library of Medicine | U.S. National Institutes of Health | U.S. Department of Health and Human Services

Figura 3.23. Página Web de Clinical Trials.gov

Históricamente la Web de repositorios de ensayos clínicos ClinicalTrials.gov fue creado en el año de 1997 como resultado de la Ley de Modernización de Alimentos y Medicamentos (FDAMA, Food and Drug Administration Modernization Act). FDAMA pidió al Departamento de Salud y Servicios Humanos de Estados Unidos, a través de los Institutos Nacional de Salud (NIH, National Institutes of Health), poder establecer un repositorio de todos registros de información de ensayos clínicos para los aquellos estudios financiados por el gobierno federal y estudios privados que fueron realizados bajo las condiciones de la aplicación IND (Investigational New Drug) para probar la efectividad de los fármacos experimentales para conocer las ventajas y desventajas que pueden ocasionar a la vida, y para saber los mecanismos y resultados contra enfermedades. El NIH y la Administración de Alimentos y Medicamentos (FDA) trabajaron juntos para que en febrero del 2000 este portal este completo y sea acceso público en la web.

Este repositorio contiene información sobre estudios de medicina que contienen como punto principal voluntarios humanos. Un ensayo clínico es un estudio de investigación en el cual a voluntarios humanos se le aplican intervenciones como bien puede ser, un producto médico, tratamiento o procedimiento sobre la base de un plan médico y se van evaluando los efectos sobre los resultados biomédicos por fases hasta llegar al final. ClinicalTrials.gov también incluye registros que describen los estudios y programas de acceso a medicamentos en investigación fuera de los ensayos clínicos de observación. Los estudios que figuran en la base de datos se llevan a cabo en los 50 estados y en 189

países. Y el número de estudios registrados cada año se ha ido incrementado con el tiempo [Figura 3.24].

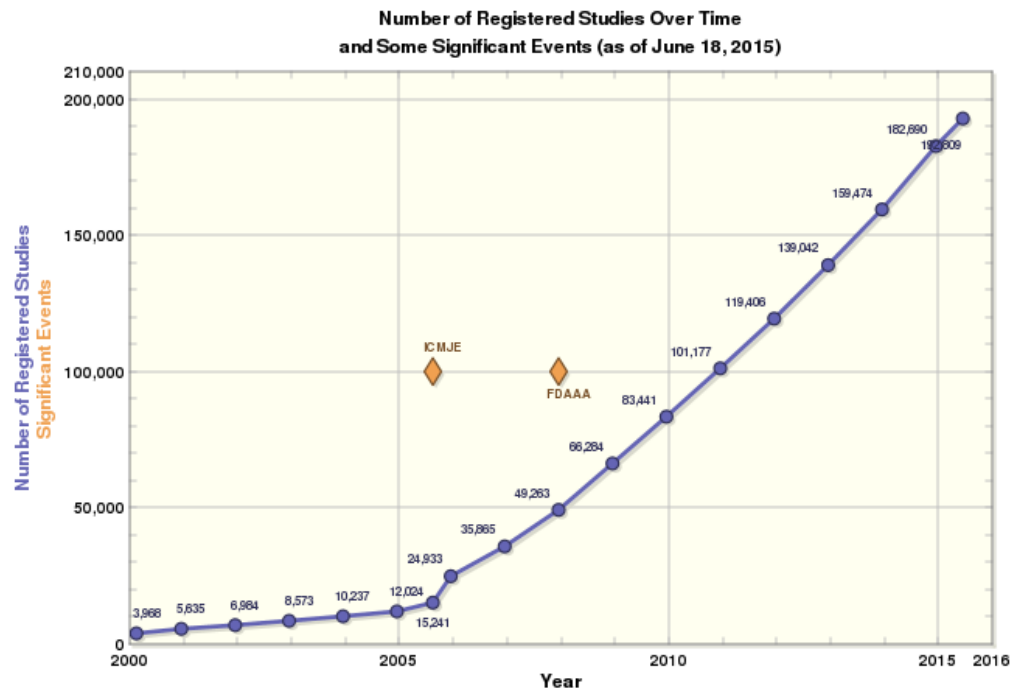


Figura 3.24. Curva que demuestra el crecimiento de los registros de ensayos médicos en este repositorio, información tomada de la página de *ClinicalTrials.gov*

Esta información fue tomada en este mismo portal y esta actualizada hasta Junio del 2015, y así también mostramos un número exacto de ensayos clínicos de los últimos 5 años [Tabla 4.9], y si observamos la figura anterior nos daremos cuenta que a partir del 2008 hubo un incremento considerable de ensayos clínicos, y esto se debe a que en el año 2007 el congreso de este país aprobó una ley de enmiendas de la FDA, que quito algunas restricciones para que un estudio sea considerado ensayo clínico y además en una parte de esta ley obligo a que el uso de fármacos y tratamientos tenga muchos ensayos clínicos probados para ser usado al público en general. Los investigadores tratan de determinar la seguridad y eficacia de la intervención mediante la medición de ciertos resultados en los participantes.

Año	Publicaciones
2010	101,177
2011	119,406
2012	139,042
2013	159,474
2014	182,690
2015	192,809

Tabla 3.6. Total exacto de publicaciones por año de la página *ClinicalTrials.gov*, información tomada en la misma página.

Para que este portal elija que el ensayo clínico sea aceptado pasa por un estudio para comprobar su autenticidad y leyes como el que cada estudio clínico debe estar dirigido por un investigador principal, que suele ser un doctor en medicina y es muy casual que también contengan un equipo de investigación que puede incluir médicos, enfermeras, trabajadores sociales y otros profesionales de la salud.

Estos estudios clínicos pueden ser patrocinados o financiados por empresas farmacéuticas, centros médicos académicos, grupos de voluntarios y otras organizaciones, además de las agencias federales, tales como los Institutos Nacionales de Salud, el Departamento de Defensa de Estados Unidos, el Departamento de Asuntos de Veteranos de EE.UU o médicos, profesionales de la salud y otras personas también pueden patrocinar la investigación clínica. Pero debido a que ClinicalTrials.gov es un sitio web del gobierno, no está permitido recibir financiamiento de la publicidad sobre este portal o la visualización de contenido comercial.

En este portal un estudio tiene dos vistas distintas, el primero es para leer el estudio en forma de texto plano, para leer lo que el creador del ensayo clínico ha escrito desde el principio hasta el final del ensayo, y el otro es una vista tabular que solo muestra unos resúmenes de cada parte del ensayo, y así están descritos todos los ensayos de la página, ya que así son ordenados y si un investigador solo quiere saber una parte del ensayo, por medio de esta vista solo la busca y no sería necesario leer todo el ensayo. [Figura 3.25]

Estos ensayos clínicos también se los pueden descargar como archivos de texto o como archivo XML, pero esto hablaremos en el capítulo de metodología en el cual se muestra paso a paso lo que se hizo para descargar todos estos ensayos clínicos.

Otro punto importante a considerar es que la mayoría de los ensayos clínicos en estos momentos no solo representan a Estados Unidos, de todo el repositorio este representa solo el 39% total, y el 4% más representa estudios en Estados Unidos y otros países al mismo tiempo pero el 46% de estos ensayos clínicos son de otro lugar que no es Estados Unidos, y por último tenemos un 10% que no se especifica el lugar y así hacemos un 100%.

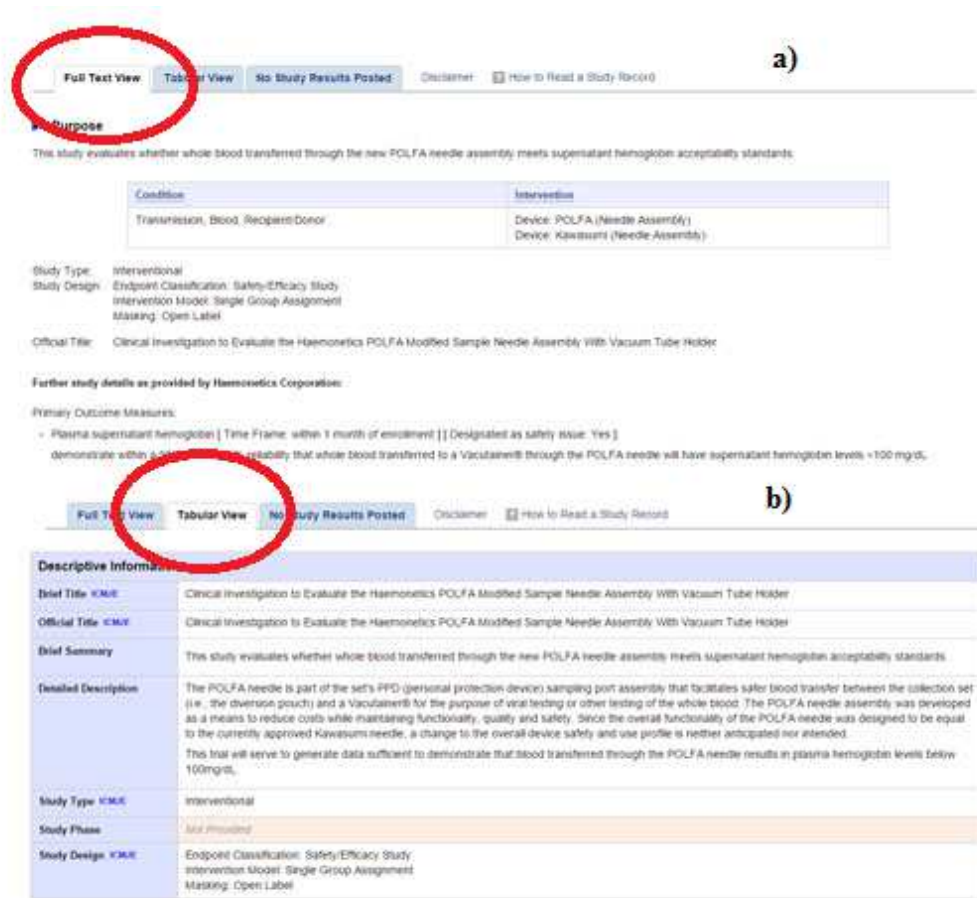


Figura 3.25. Ejemplo de cómo se ve un ensayo clínico y sus tipos de vistas: a) Full Text View y b) Tabular View

3.5.2 Clinical Trials Register (Unión Europea)

Al igual que el anterior repositorio, este contiene registro de ensayos clínicos sobre medicina, actualmente contiene aproximadamente más de 25500 registros que contienen información sobre ensayos clínicos de intervención sobre los medicamentos llevados a cabo en la Unión Europea (UE) o del Espacio Económico Europeo (EEE), [Figura 3.26] a diferencia del anterior en este repositorio no se sabe si su información es 100% original, ya que la Agencia Europea de Medicamentos (EMA) no se responsabiliza de la integridad o exactitud de esta información. Además, la EMA y las autoridades nacionales competentes no asumen ninguna responsabilidad por el uso de cualquier partido, o los resultados de dicho uso, de cualquier parte de la base de datos.

The European Union Clinical Trials Register allows you to search for protocol and results information on:

- interventional clinical trials that are conducted in the European Union (EU) and the European Economic Area (EEA);
- clinical trials conducted outside the EU / EEA that are linked to European paediatric-medicine development.

Learn [more about the EU Clinical Trials Register](#) including the source of the information and the legal basis.

The EU Clinical Trials Register currently displays **25767** clinical trials with a EudraCT protocol, of which **3745** are clinical trials conducted with subjects less than 18 years old.

The register also displays information on **18612** older paediatric trials (in scope of Article 45 of the Paediatric Regulation (EC) No 1901/2006).

Please enter search term...

Examples: Cancer AND drug name, Pneumonia AND sponsor name.
[How to search \[pdf\]](#)

Advanced Search: [Search tools](#)

EU Clinical Trials Register Service Desk: euctr@ema.europa.eu
 European Medicines Agency © 1995-2015 | 30 Churchill Place, Canary Wharf, London E14 5EU

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

HMA
Health Medicines Agency

Figura 3.26. Página Web de ClinicalTrialsRegister.eu

Históricamente este portal fue creado el 1 de mayo de 2004, siete años más tarde que el anterior repositorio, y es parte de la comunidad de bases de datos médicas EudraPharm. Este repositorio se va actualizando por medio de versiones, en este momento está en la versión 2.1 que fue establecida el 26 de Enero del 2007, en su nueva versión contienen información sobre ensayos clínicos pediátricos que anteriormente no tenía. Al igual que en Estados Unidos la Unión Europea también creó una ley en la cual la Directiva de Ensayos Clínicos de la Unión Europea 2001/20 / CE establece un marco que estableció cómo deberían llevarse a cabo ensayos clínicos que investigan la seguridad o eficacia de un medicamento en humanos. Clinical Trials Register le permite buscar información en la base de datos URL EudraCT (European Union Drug Regulating Authorities Clinical Trials), que ya está por la versión 10. Esta es la base de datos utilizada por los reguladores nacionales de medicamentos (European Medicines Agency) que son los que regulan los datos relacionados con los protocolos de ensayos clínicos. Y la validación no es tan rigurosa como en el anterior repositorio aunque los datos sobre los resultados de estos ensayos se introducen en la base de datos por medio de los patrocinadores y se publican en este Registro final una vez que los mismos patrocinadores han validado los datos.

Este portal tiene reglamentos en la publicación de sus ensayos clínicos como:

- El ensayo clínico debe proporcionar información sobre los ensayos clínicos observacionales de medicamentos autorizados;

- Debe proporcionar información sobre los ensayos clínicos, como son los procedimientos quirúrgicos, dispositivos médicos o procedimientos psicoterapéuticos;
- Debe proporcionar información sobre los ensayos clínicos en los que todos los sitios de investigación están fuera de la UE / EEE
- Y algo distinto del anterior repositorio fue proporcionar la navegación y el contenido web en idiomas distintos del inglés.

Es importante saber que la información que aparece en los Registros de Ensayos Clínicos de la Unión Europea está originalmente suministrada por la empresa u organización responsable del ensayo clínico. La información de protocolo relacionado es un componente de la aplicación del patrocinador a una autoridad nacional competente para que esta la pueda autorizar. Y por ende la autoridad nacional competente que autorizo este trabajo se suma a la información suministrada del ensayo clínico y la opinión del comité de ética pertinente. Si es que el ensayo clínico fue creado fuera de la Unión Europea y del Espacio Económico Europeo, esta autorización es suministrada por la empresa u organización responsable del mismo.

EudraCT Number: 2010-023457-11 Sponsor Protocol Number: S52798 Start Date *: 2011-02-15					
Sponsor Name: UZ Leuven					
Full Title: Immune regulation and timing of chemotherapy in advanced/recurrent ovarian and endometrial cancer					
Medical condition: advanced/recurrent ovarian and endometrial cancer					
Disease:	Version	SOC Term	Classification Code	Term	Level
	12.1		10057529	Ovarian cancer metastatic	LLT
	12.1		10066697	Ovarian cancer recurrent	LLT
	12.1		10057529	Ovarian cancer metastatic	PT
	12.1		10066697	Ovarian cancer recurrent	PT
	12.1		10014734	Endometrial cancer metastatic	LLT
	12.1		10014736	Endometrial cancer recurrent	LLT
	12.1		10014734	Endometrial cancer metastatic	PT
	12.1		10014736	Endometrial cancer recurrent	PT
Population Age: Adults, Elderly			Gender: Female		
Trial protocol: BE (Ongoing)					
Trial results: (No results available)					

Figura 3.27. Ejemplo de cómo se ve un ensayo clínico de forma resumida

También la presentación del ensayo clínico es muy diferente del anterior mencionado, es más simple, dependiendo de lo que se busque, y de igual forma se presenta de dos formas una resumida [Figura 3.27] y la otra en la cual presenta todo lo que tiene el ensayo clínico [Figura 3.28]. Y aquí se dividen según los países de donde fue proporcionado el estudio. La lista de los países que aportan información y registros de ensayos clínicos a este repositorio son: Austria, Bélgica, Bulgaria, Croacia, Ciprés, Republica Checa, Dinamarca, Estonia, Finlandia, Francia, Alemania, Grecia, Hungría, Islandia, Irlanda, Italia, Letonia, Liechtenstein, Lituania, Luxemburgo, Malta, Holanda, Noruega, Polonia, Portugal, Rumania, Eslovaquia, Eslovenia, España, Suiza, Reino Unido y las afuera de la EU/EEA.

Index		
A. PROTOCOL INFORMATION		
B. SPONSOR INFORMATION		
C. APPLICANT IDENTIFICATION		
D. IMP IDENTIFICATION		
D.8 INFORMATION ON PLACEBO		
E. GENERAL INFORMATION ON THE TRIAL		
F. POPULATION OF TRIAL SUBJECTS		
G. INVESTIGATOR NETWORKS TO BE INVOLVED IN THE TRIAL		
N. REVIEW BY THE COMPETENT AUTHORITY OR ETHICS COMMITTEE IN THE COUNTRY CONCERNED		
P. END OF TRIAL		

[Expand All](#) [Collapse All](#)

A. Protocol Information		
A.1	Member State Concerned	Slovakia - SIDC (Slovak)
A.2	EudraCT number	2004-000534-36
A.3	Full title of the trial	Efficacy and Safety of Piasclidine 300 versus Chondroitin Sulfate in a 6 Months Treatment plus 2 Months Observation in Patients with Osteoarthritis of the Knee
A.4.1	Sponsor's protocol code number	PR 1903
A.7	Trial is part of a Paediatric Investigation Plan	Information not present in EudraCT
A.8	EMA Decision number of Paediatric Investigation Plan	

B. Sponsor Information		
------------------------	--	--

Figura 3.28. Ejemplo de cómo se ve un ensayo clínico al elegirlo para leerlo

Actualmente para que la información que contienen estos registros de ensayos clínicos sea confiable, las autoridades nacionales competentes y la Agencia están trabajando para desarrollar validaciones más exactas, mejorando la calidad de los nuevos registros mediante una mayor comprobación automatizada, control de calidad y a través del mayor uso de datos estandarizados.

Al igual que el anterior repositorio tiene la opción de poder descargar estos ensayos clínicos, pero en el capítulo siguiente se dará los pasos correspondientes para poder bajar esta información de manera más precisa.

4 METODOLOGÍA

El método usado en este trabajo de fin de máster fue la unión de tres fases importantes que ayudaron a este método de minería de texto que se hizo en los registros de ensayos clínicos, el primero es la recolección de la base de datos para el entrenamiento del algoritmo y la base de datos para la prueba del algoritmo; segundo es el pre-procesamiento de los ensayos clínicos para obtener la bolsa de unigrams (se puede definir como un token extraído del texto) y para transformarlos en archivos que el algoritmo clasificador lo pueda leer y así construir el modelo para la comparación con el repositorio total que incluye todos los ensayos clínicos obtenidos por las páginas de ClinicalTrials.gov y ClinicalTrialsRegister.eu. Y tercero es ya la comparación del modelo con las bases de datos y obtener los resultados para la correcta discusión del tema de este trabajo de fin de máster. Un resumen de la metodología se muestra en la Figura 4.1.

4.1 Obtención de los recursos

En esta parte del capítulo conoceremos como se obtuvo la primera fase de la metodología, que es la obtención de la base de datos para hacer modelo, en este caso estoy hablando de los ensayos clínicos que se usaron para el entrenamiento del algoritmo y la obtención de la base de datos para aplicar el algoritmo y obtener los resultados deseados

4.1.1 Obtención del repositorio para el entrenamiento del algoritmo

En la mayoría de casos los investigadores toman nuevas bases de datos para formular algoritmos, herramientas, medicinas, tratamientos o terapias para obtener buenos resultados, en mi caso como hemos visto, se ha utilizado una publicación [De la Iglesia, 2013] para poner a prueba el algoritmo y obtener mejores resultados. En su publicación explica que los investigadores han desarrollado directorios de recursos que funcionan como "páginas amarillas" para la comunidad científica biomédica. Páginas como las que he descrito en la historia del arte en el Capítulo 3, como es ClinicalTrials.gov, ClinicalTrialsRegister.eu, y un sin número más. Dentro de la comunidad nano médica, como nuestra unidad de bioinformática médica se están desarrollando un enfoque similar, para promover la difusión y el intercambio de los recursos disponibles en la actualidad en el campo.

Para el entrenamiento del modelo para la comparación con la bases de datos de los registros de ensayos clínicos, se ha usado el que la autora de la publicación [De la Iglesia, 2013] ha usado dentro de su Anexo (Anexo I), ya que ella separó manualmente los registros de ensayos clínicos en las dos categorías que deseo separar la base de datos completa, los ensayos clínicos sobre nano medicina y los ensayos clínicos que no son sobre nano medicina. A continuación se explicará como la autora separó los ensayos

clínicos, como primer paso se escogió un conjunto representativo de los resúmenes de ensayos clínicos que se encuentran en el repositorio *ClinicalTrials.gov*, en base a los criterios y procedimientos que se describen a continuación:

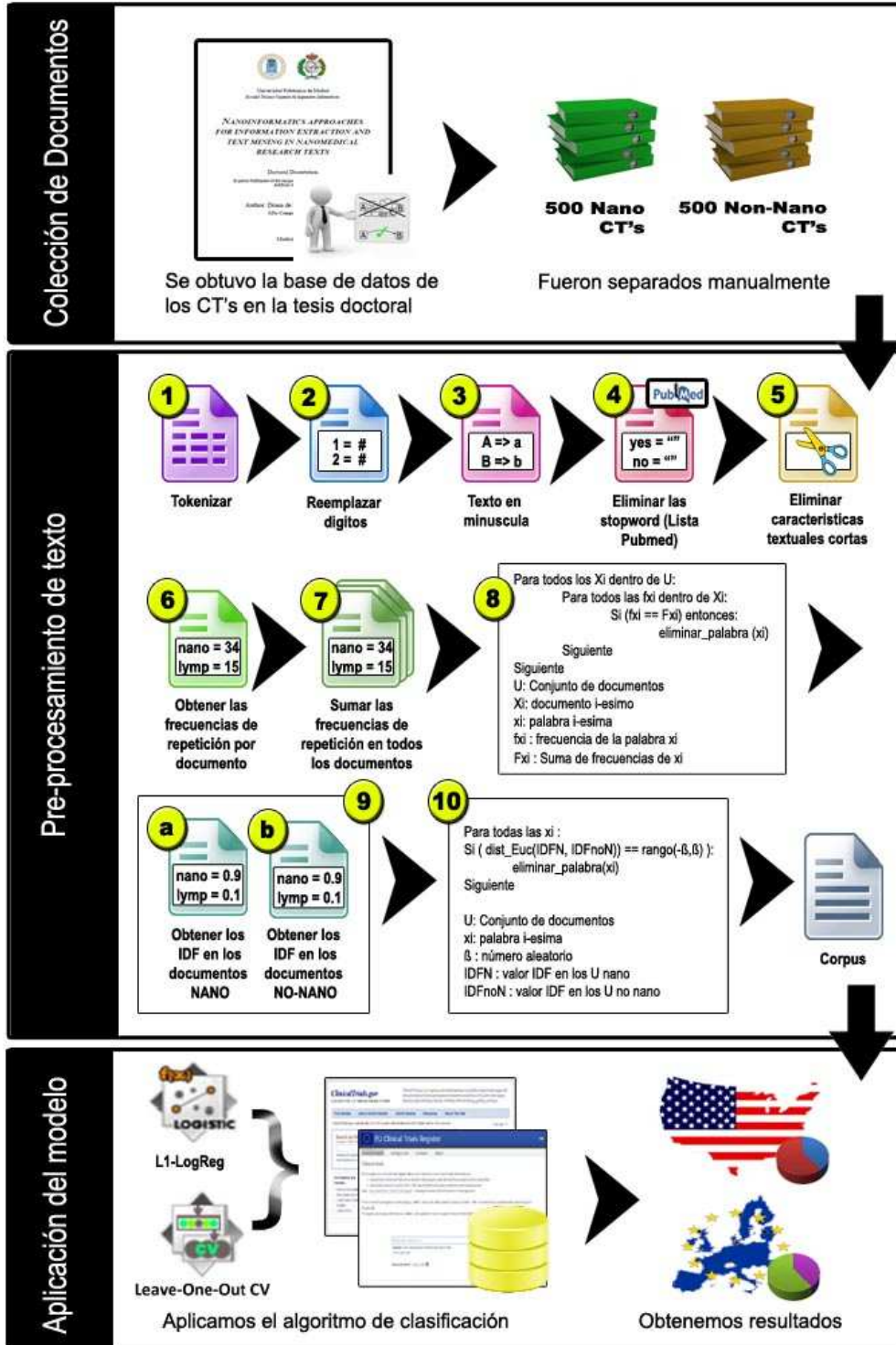


Figura 4.1. Procedimiento de la metodología a seguir

1. Simples búsquedas dependiendo de la palabra clave:
 - i) Los ensayos clínicos que contenían el término "nano"
 - ii) Los ensayos clínicos que contienen el término "nanopartículas"

En este pequeño grupo de ensayos se obtuvieron 188 registros validados y etiquetados manualmente como nano, teniendo en cuenta los numerosos productos nanomédicos que están actualmente presentes, lo cual contenían el uso de nanofármacos y nanodispositivos y así se podía deducir que el ensayo clínico era el correcto.

2. La autora para detectar la presencia de conceptos y entidades semánticas que podrían estar relacionados con el campo de la nanomedicina en los registros de la base de datos ClinicalTrials.gov, enriquece el conjunto de términos de búsqueda mediante la inclusión de conceptos específicos de una fuente externa desarrollada por el National Cancer Institute, el NCI Metathesaurus (Covitz et. al 2003). El NCI Metathesaurus combina un gran número de vocabularios clínica y biológicamente relevante controlados y terminologías, algunos directamente relacionados con la nanotecnología y, en concreto, a la nanomedicina. En este paso se obtuvieron 38 conceptos (190 términos) que devolvieron los resultados y se adaptan a los ensayos clínicos. Con este proceso, se obtuvieron un total de 414 registros, validados y anotados con la etiqueta "nano", obviamente después de una verificación manual de cada uno.

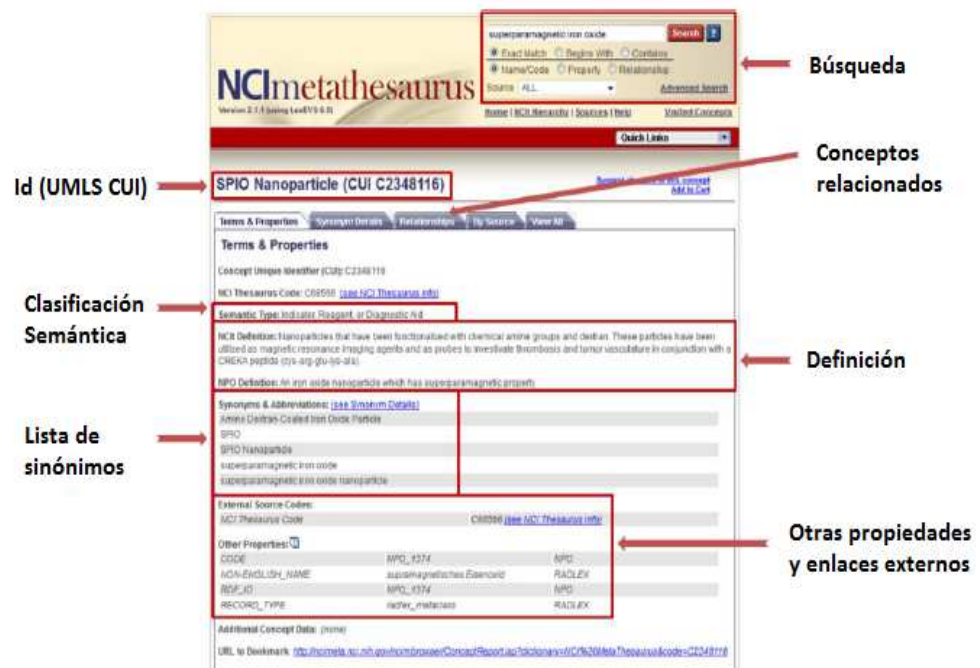


Figura 4.2. Partes de la página web de NCI Metathesaurus, información tomada de la tesis doctoral

3. Como el objetivo fue encontrar 500 ensayos clínicos relacionados con la nanomedicina en la web de *ClinicalTrials.gov*. La autora comenzó a investigar más a fondo acerca de las nanodrogas que se utilizaban y así reviso una lista de páginas que contenían información acerca de este tema. De la información analizada, la autora encontró un conjunto de términos que se utilizan ampliamente relacionados con la nanomedicina que se puede considerar como patrones de texto en los ensayos clínicos -como por ejemplo liposomal, micelas, nanomaterial, nanosuspensión, nanocoloide, cristal, nanotubos, gel, PEG, etc. Estos términos también se utilizaron para completar los criterios de búsqueda y realizar nuevas búsquedas en la base de datos *ClinicalTrials.gov* que permitió obtener los registros de ensayos clínicos adicionales para completar el cuerpo de los mismos y etiquetarlos como "nano".

4. Ahora los 500 ensayos clínicos que no hablaban sobre nanomedicina fue más sencillo ya que se escogió aleatoriamente en el vasto numero de ensayos clínicos de la pagina *ClinicalTrials.gov*, y así la autora terminó con una formación equilibrada y conjuntos de pruebas, incluyendo 500 TC nano y 500 TC no nano.

A través del procedimiento que la autora hizo y que fue descrito anteriormente, se obtuvo un conjunto de 1000 resúmenes de anuncios, dividido en las dos clases diferentes de nano y no nano, dependiendo de si estos ensayos clínicos eran o no relevante para la focalización de nanopartículas, nanomateriales, o nanotecnologías, y por lo tanto caer en el campo de la nanomedicina. Y de esta forma se obtuvo la base para el entrenamiento para la aplicación del algoritmo.

4.1.2 Obtención de los repositorios para el ejecutar el algoritmo

La obtención de los repositorios de los ensayos clínicos fue la segunda tarea después de obtener los ensayos clínicos de entrenamiento para crear el modelo. Los repositorios que usamos para poder obtener los ensayos clínicos fueron dos:

1. El que representa para nosotros Norte América: *ClinicalTrials.gov*
2. El que representa la Unión Europea: *ClinicalTrialsRegister.eu*

Para obtener todos los registros del primer repositorio no fue complicado, ya que cuenta con una forma de descargar todos los registros que uno desee de forma automática, siguiendo estos pasos:

1. Primero se debe ingresar a la página de *ClinicalTrials.gov* y en el cuadro de texto no se debe ingresar nada, sino directamente presionar "Search", para que así la búsqueda nos de todos los registros de ensayos clínicos que contiene y obtener todo el repositorio, así como se muestra en la figura 4.3 (1).

2. Segundo aparecerán todos los registros en una lista de 20 ensayos clínicos por pagina, y en la parte de abajo el número total de páginas que existen con la búsqueda realizada, esta búsqueda la realice el 20 de Mayo del 2015, y encontré 186.333 ensayos clínicos., así como se muestra en la figura 4.3 (2)

3. Tercero debajo del número total de registros, esta un enlace con la palabra "Download" al presionar aparecerá un cuadro con opciones de descarga, la cual representa diferentes opciones que nos ayudaran a obtener estos ensayos clínicos. Como necesitaba descargarme todo el repertorio escogí en el primer cuadro de combo los 186.333 ensayos clínicos y abajo de esto escogí la opción de "Download All Study and Results Fields as XML", en esta opción se descarga cada uno de los ensayos en archivos XML diferente, obteniendo en si los 186.333 ensayos clínicos separados por el número de registro correspondiente, así como se muestra en la figura 4.3 (3).

4. Y por ultimo presionamos el botón de "Download Zip File", en el cual dependiendo del navegador se descargara de forma automática, como por ejemplo en el navegador Google Chrome aparece en la parte derecha, en la esquina inferior el archivo descargándose, y al final usted ya tiene el archivo comprimido llamado (search_result.zip), así como se muestra en la figura 4.3 (4). Cabe recordar que para obtener los archivos XML, se debe descomprimir este archivo descargado.

En cambio para obtener los registros de los ensayos clínicos en el otro repositorio "ClinicalTrialsRegister.eu", no fue tan fácil como el anterior, ya que ellos no cuentan con una opción de descargar todos los archivos que fácilmente aparecen comprimidos y separados como lo hace ClinicalTrials.gov y en XML.

En este repositorio se encuentran los ensayos clínicos de diferentes países de la unión europea:

- Austria,
- Bélgica,
- Bulgaria,
- Croacia,
- Republica Checa,
- Dinamarca,
- Estonia,
- Finlandia,
- Francia,
- Alemania,
- Grecia,
- Entre otros

De igual forma como el anterior explicare como se tuvo que descargar todos estos archivos de este repositorio paso a paso.

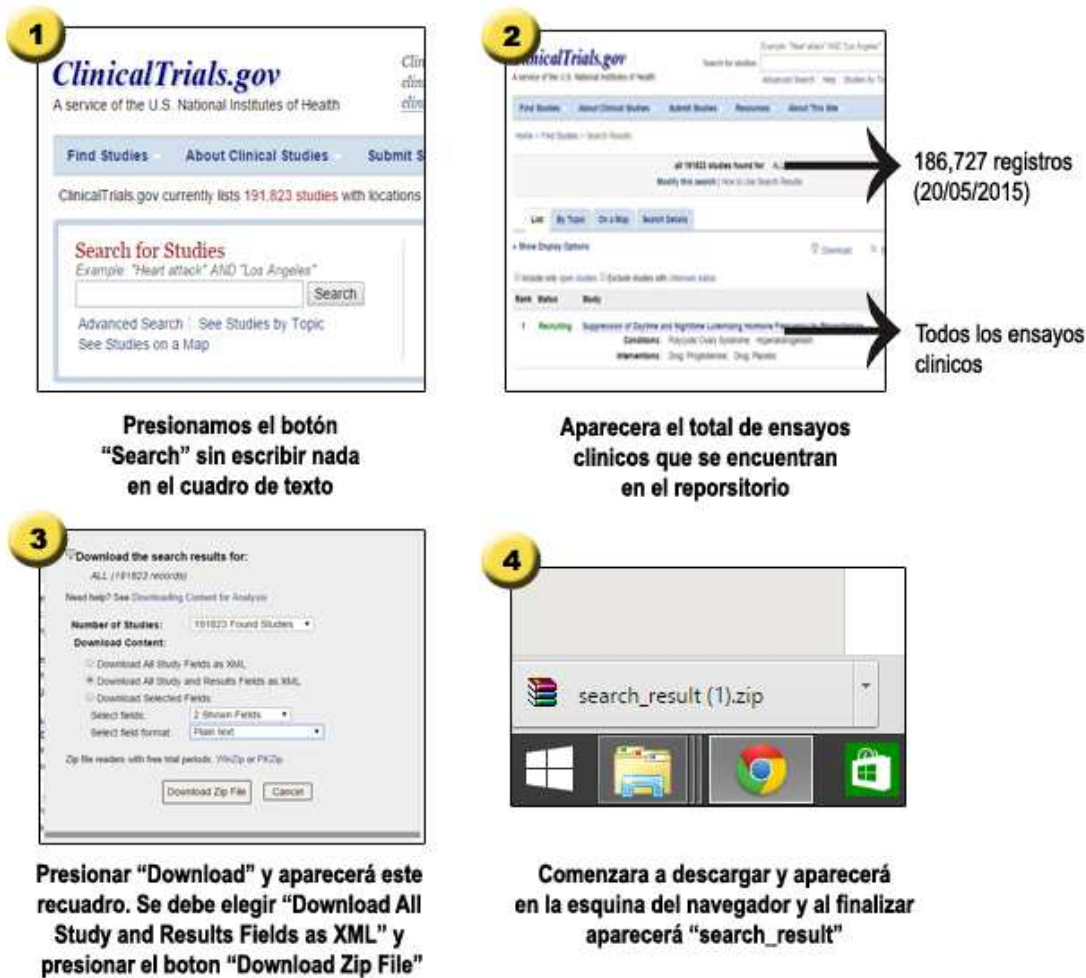


Figura 4.3. Explicación gráfica de como descargar los ensayos clínicos en la página Web ClinicalTrials.gov

Vamos a enumerar los pasos:

1. Primero se debe ingresar a la página de ClinicalTrialsRegister.eu y en el cuadro de texto no se debe ingresar nada, sino directamente presionar "Search", para que así la búsqueda nos de todos los registros de ensayos clínicos que contiene y obtener todo el repositorio, así como se muestra en la figura 4.4 (1).
2. Segundo aparecerán todos los registros en una lista de 25 ensayos clínicos por pagina, y en la parte de superior el número total de páginas que existen con la búsqueda realizada, esta búsqueda la realice el 02 de Junio del 2015, y encontré 25.657 ensayos clínicos, así como se muestra en la figura 4.4 (2)
3. Tercero en la parte izquierda del cuadro que contiene los resultados de los ensayos clínicos, esta un cuadro con opciones y el botón "Download", la cual representa diferentes opciones que nos ayudaran a obtener estos ensayos

clínicos. Se debe elegir las opciones de "Trials shown on current page", y en el siguiente combo se debe elegir "Full Trials Details" y por ultimo en el combo final seleccionar "Plain Text", así como se muestra en la figura 4.4 (3).

4. Y por ultimo presionamos el botón de "Download", en el cual dependiendo del navegador se descargara de forma automática, como por ejemplo en el navegador Goolge Chrome aparece en la parte derecha, en la esquina inferior el archivo descargándose, y al final usted ya tiene el archivo comprimido llamado (trials-full.txt), Aquí es la diferencia entre el primer repositorio, ClinicalTrialsRegister.eu no me permite bajarme todos los registros de los ensayos clínicos a la vez, si no que tuve que descargarlos pagina por pagina, y no solo eso, sino que vienen todos los ensayos clínicos dentro de un mismo archivo, es decir los 25 uno tras de otro en el mismo archivo de texto, tuve que recurrir a una herramienta creada para el experimento para poder separarlos, pero esta herramienta se la explicara más adelante. Otro inconveniente es que solo tenemos la opción de bajarlo como archivo plano, así se tuvieron que bajar 1.283 archivos de texto, así como se muestra en la figura 4.4 (4).



Figura 4.4. Explicación gráfica de como descargar los ensayos clínicos en la página Web ClinicalTrialsRegister.eu

Y así es como obtuvimos los dos repositorios de registros de ensayos clínicos y también el repositorio para poder crear el modelo y aplicar el algoritmo necesario para que este pueda obtener los resultados que queremos tener.

4.2 Ensayos clínicos

Un ensayo clínico es una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, en su aplicación a seres humanos, pretende valorar su eficacia y seguridad. Y en estos métodos solo usaremos los registros de ensayos clínicos basados en nanomateriales, nanodispositivos y nanopartículas, para ejecutar nuestro pre-procesamiento de texto para etiquetar a la categoría “NANO”. [Figura 4.5]

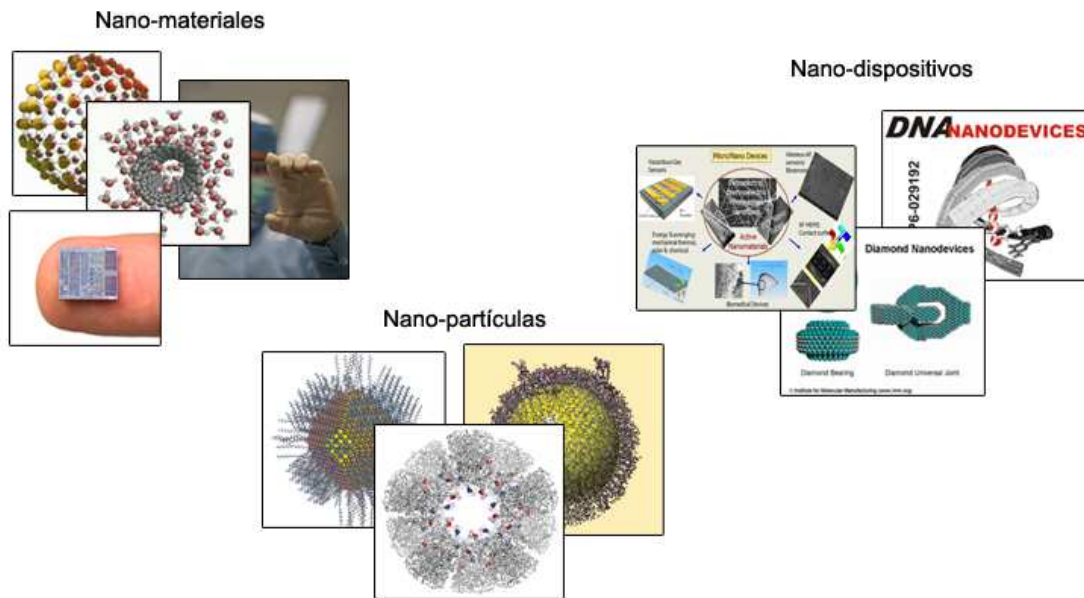


Figura 4.5. Los tipos de ensayos clínicos que necesitamos buscar en los dos repositorios

4.3 Pre-procesamiento de texto

En esta parte del capítulo conoceremos como se construyó la segunda fase de la metodología, que es la llamada pre procesamiento del repositorio de prueba que se obtuvo para crear el modelo para poder usarlo en la fase final. Tomamos en cuenta que

este fase es muy crítica para ya que si no se la hace correctamente pueden producir errores y no nos darán los resultados que deseamos. Además si mejoramos esta fase nos ayudara a mejorar el rendimiento de los resultados del clasificador que se aplicara en la tercera fase, aplicando técnicas de aprendizaje automático se obtendrá lo q deseamos. Lo primero que se hizo fue tomar muy en cuenta los pasos del pre-procesamiento a todos los ensayos clínicos que obtuvimos en la primera fase.

La función de este pre-procesamiento de datos es filtrar la basura –términos, dígitos y frases que no son necesarias para la búsqueda de los patrones del texto- que se obtiene de los ensayos clínicos y dejar lo que realmente importa, y eso es la bolsa de palabras o de unigrams ya que este paso ayuda convirtiéndolos en datos que el clasificador pueda leer y entender.

Para hacer los pasos que se desarrollaron a continuación se debió tener en cuenta que clasificador íbamos a usar y con qué términos iba a ser usado. La publicación que usó como ejemplo [De la Iglesia, 2013], demostró por medio de varios experimentos utilizando la colección de algoritmos de aprendizaje automático con tecnología de última generación para la minería de datos proporcionada es el banco de trabajo Weka [Witten y Frank 2005], y que su mejor clasificador es el llamado Regresión Logística Regularizada o también conocida como Regresión Logística Lasso que en la herramienta Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) es conocido como el L1-LogReg.

La regresión logística Lasso fue usada a principio solo para modelos lineales, la llamada Least Absolute Shrinkage and Selection Operator (con sus siglas en ingles LASSO) se introdujo para regresión logística [Lokhorst, J., 1999]. LASSO combina la penalización y selección de variables, imponiendo una penalización sobre los coeficientes de la regresión, de modo que para valores altos del parámetro de penalización algunos de estos coeficientes se fijan a 0. La regresión logística regularizada o LASSO minimiza la log-verosimilitud negativa, sujeta a una penalización en los coeficientes de regresión. Así, siendo β el vector de coeficientes de regresión y L la función log-verosimilitud negativa, el estimador lasso se define como:

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ L(\beta) + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

Formula 4.1

Donde encontramos que si $\lambda \geq 0$, se determina la cantidad de penalización. La ventaja es que puede generar modelos dispersos, donde la mayoría de los coeficientes son igual a 0. Y esta penalización es denominada penalización L1.

Ahora la autora de la publicación [De la Iglesia, 2013] utiliza los valores de los coeficientes IDF de cada termino para utilizar la penalización. Muchos investigadores usan en la minería de texto coeficientes para poder clasificar como son el TF, IDF y el TF-IDF.

1. El TF es la frecuencia del término en un documento, este coeficiente es denotado con las siglas $tf(t, d)$ donde se lee como: **tf** es el número de veces que el término t ocurre en el documento d . Existen varios tipos de tf , pero el que usaremos en este trabajo de fin de máster es este mencionado.
2. El IDF es la frecuencia inversa del término en una colección de documentos, este coeficiente es denotado con la siguiente fórmula:

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Formula 4.2

Donde:

D = es la colección de documentos

$|\{d \in D : t \in d\}|$ = es el número de documentos que contiene el termino t

Pero en este trabajo de fin de máster usaremos el IDF logarítmico regularizado, ya que si el termino no se repite en la colección de documentos se producirá una división para cero, y en la mayoría de lenguajes aparece un error o da infinito, para que esto no ocurra se le sumara el valor de 1 en la parte baja de la formula, quedando la siguiente fórmula:

$$idf(t, D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|}$$

Formula 4.3

3. El último es el TF-IDF, que es la multiplicación del TF X IDF, pero este coeficiente no vamos a usar en este trabajo de fin de máster.

Para la regresión logística LASSO usaremos como coeficiente el IDF del término, y el TF es usado para un paso del método de pre-procesamiento, los términos que usaremos son la bolsa de unigrams obtenidas en el último paso de este pre-procesamiento.

Los pasos que se siguieron fueron muy parecidos a los que uso la autora de la tesis doctoral [De la Iglesia, 2013], pero se cambio ciertos pasos ya que al hacer pruebas con los mismo pasos que ella, se obtenían resultados menos eficientes a los que se usó al

final en el pre-procesamiento; por ende estas pruebas demostraron que tenía mejores resultados, y así se obtuvo un mejor porcentaje de aceptación al aplicar el clasificador de aprendizaje automático (L1-LOGREG) con el resultado del modelo.

Estos pasos que se cambiaron fueron porque se tomó en cuenta un estudio realizado en 2013 en Iraq, la cual se observó que para hacer una clasificación más inteligente es necesario hacer una reducción de datos cuando se tiene la bolsa de palabras ya obtenida luego de todos los pasos de pre-procesamiento, para que la clasificación se mas efectiva. [Sadiq A., 2013]

Los pasos que se siguieron son 10 en resumen: la tokenización, el reemplazo de dígitos, el poner el texto en minúsculas, la eliminación de stopwords, la eliminación de características textuales, la obtención de las frecuencias de cada termino por documento, la obtención de la suma de las frecuencias de cada termino por el conjunto de documentos, la obtención de los IDF (Inverse Document Frequency) de cada termino en el conjunto de documentos, y aplique dos reglas que ayudaron a disminuir la bolsa de palabras para poder hacer más preciso la clasificación. Para poder hacer estos pasos se usó el lenguaje de programación Visual Basic 2010 Express, ya que es un lenguaje fácil de usar para las funciones con palabras, y contiene funciones que me ayudaron a separar y reemplazar textos de una forma más sencilla y rápida. A continuación se describirán los pasos que se hicieron:

4.3.1 Tokenización

La tokenización es la acción de convertir cada palabra que contiene un texto en un token (es una cadena de caracteres que tiene un significado coherente), para esto se usó una la librería de *System.IO* que se encuentra en el lenguaje Visual Basic.Net y la función de *Split*, la cual recibe una cadena y un carácter o un conjunto de caracteres que definen la separación de cada uno de los tokens, en este caso se usó el espacio como este carácter especial, como salida nos da un arreglo con todos los tokens, se repitan o no se repitan. Un Ejemplo de tokenizar lo muestro a continuación:

Si tenemos la cadena:

Cadena = “The Abraxane was used during treatment for 3 to 4 weeks after the last chemotherapy”

Arreglo = *split(Cadena, “ ”)*

print Arreglo

‘Obtendriamos:

Arreglo = [‘The’, ‘Abraxane’, ‘was’, ‘used’, ‘during’, ‘treatment’, ‘for’, ‘3’, ‘to’, ‘4’, ‘weeks’, ‘after’, ‘the’, ‘last’, ‘chemotherapy’]

Ejemplo 4.1

Pero también tenemos ciertas características textuales que están unidas por un símbolo que son reconocidas como un token de igual forma, por eso es necesario en un paso más adelante borrar estos errores. Como por ejemplo

Cadena = “*The Abraxane is used on cancer’s patients under 70-80 years?. The Abraxane is still testing*”

Arreglo = *split(Cadena, “ ”)*

print Arreglo

‘*Obtendriamos:*

Arreglo = [*‘The’, ‘Abraxane’, ‘is’, ‘used’, ‘on’, ‘cancer’s’, ‘patients’, ‘under’, ‘70-80’, ‘years?.’, ‘The’, ‘Abraxane’, ‘is’, ‘still’, ‘testing’*]

Ejemplo 4.2

Al finalizar cada documento obtenemos una bolsa de palabras que se repiten muchas veces, y también se llenan de dígitos, de stopwords que no son necesarios para el clasificador ya que hace que tengan muchas palabras innecesarias, para esto hacemos los otros pasos.

4.3.2 Reemplazo de dígitos

El reemplazo de dígitos, no solo es el reemplazo de los números, sino también de símbolos dentro del contexto que no son necesarios para el clasificador, estos son reemplazados por un mismo símbolo que en este caso es el numeral (#), y seguido a esto son borrados reemplazados por un espacio vacío para que sean desaparezcan.

Existen ciertos símbolos especiales que se fueron tratados de forma distinta como son los paréntesis – () -, los corchetes – [] -, las llaves - { } -, los signos de exclamación - ¡! -, los signos de interrogación - ¿? -, las apostrofes – ‘ - , las comas - , -, los puntos - . - y por último las comillas simples y dobles – ‘ “ - ya que la mayoría de estos están seguidos por una palabra sin espacio y al tokenizarlas se transforman en una sola palabra conjunta, lo cual no debe ocurrir, ya que más adelante se la tomaría como una palabra distinta, así que se tomó la iniciativa de agregarle un espacio antes y un espacio después de cada símbolo, y seguido a esto fueron reemplazadas por el símbolo numeral (#) y por ultimo reemplazadas por un espacio vacío. Al final se uso la función Split nuevamente por si se separaban palabras que tenia apostrofes o algún símbolo especial y se agregaba a una nueva lista.

```

Reemplazar_digitos(Arreglo){
  Para todo token en Arreglo{
    Para todo carácter en token{
      Si carácter == (lista_simbolos_especiales) {
        carácter = Concatenar(" ", carácter);
        carácter = Concatenar(carácter, " ");
      }

      Si carácter == [0,1,2,3,4,5,6,7,8,9]{
        carácter = "#";
      }

      Si carácter == (lista_simbolos){
        carácter = "#";
      }

      Si carácter == "#" {
        carácter = "";
      }
      palabra = Concatenar(palabra, carácter);
    }
    Si palabra.noEstaVacía{
      Arreglo_tmp = Split(palabra, " ");
      Nuevo_arreglo.Agregar(Trim(Arreglo_tmp));
    }
  }
  Retornar Nuevo_arreglo;
}

```

La función **Trim** hace que los espacios vacíos desaparezcan, al principio o al final de la palabra.

A continuación muestro un ejemplo (Siguiendo con el ejemplo 4.2):

```

Arreglo = ['The', 'Abraxane', 'is', 'used', 'on', 'cancer's', 'patients', 'under',
'70-80', 'years? .', 'The', 'Abraxane', 'is', 'still', 'testing']

```

```

Nuevo_Arreglo[ ] = Reemplazar_digitos(Arreglo);

```

```

print Arreglo;

```

'Obtendremos como salida:

```

Nuevo_arreglo = ['The', 'Abraxane', 'is', 'used', 'on', 'cancer', 's',
'patients', 'under', 'years', 'The', 'Abraxane', 'is', 'still', 'testing'];

```

Ejemplo 4.3

4.3.3 Texto en minúsculas

El siguiente paso fue pasar todos los token a minúsculas, eliminando las mayúsculas. En Visual Basic existe una función llamada *ToLower* de la librería String que lee toda la cadena y la transforma en minúscula.

Seguimos con el Ejemplo 4.3

```
Nuevo_arreglo = Transformar_minusculas(Nuevo_Arreglo);
print Nuevo_arreglo;
‘Obtenemos como salida
  Nuevo_arreglo = [‘the’, ‘abraxane’, ‘is’, ‘used’, ‘on’, ‘cancer’, ‘s’,
‘patients’, ‘under’, ‘years’, ‘the’, ‘abraxane’, ‘is’, ‘still’, ‘testing’];
```

Ejemplo 4.4

4.3.4 Eliminación de Stopwords

Después de obtener todos los textos en minúsculas, se buscó una lista de stopwords (el nombre que reciben las palabras sin significado en el contexto que se encuentren, no siempre son las mismas en todos los documentos), esta lista de palabras deben ser eliminadas, ya que si pensamos un poco más allá en el clasificador, estas palabras se repetirán en las categorías que vayamos a categorizar y por esto se tendrá errores en el algoritmo y no llegaremos a una buena predicción.

Ya que los documentos que estamos pre-procesando son acerca de medicina, se tomo como referencia la lista de stopwords de la página de PubMed (<http://www.oocities.org/gumby9/physicians/advanced/stopwords.pdf>). Esta lista es una de las cuales tiene la mayoría de palabras que al momento de clasificar nos dan errores, como por ejemplo tenemos las palabras –a, the, on, is, yet, about, same, etc-, la lista completa esta en el Anexo II.

Se creó una función de igual forma como la del reemplazo de dígitos, en el cual si existía en la bolsa de palabras una palabra que sea igual a uno de estas stopwords, fueran eliminadas de la lista.

```
Eliminar_stopwords(Arreglo){
  Para todo token en Arreglo{
    Si token == lista_stopwords {
      Elimar_de_Arreglo(Arreglo,token);
    }
  }
  Retornar Arreglo;
}
```

Si utilizamos el resultado del Ejemplo 4.4

```
Nuevo_arreglo = Eliminar_stopwords(Nuevo_Arreglo);
print Nuevo_arreglo;
‘Obtenemos como salida:
Nuevo_arreglo = [‘abraxane’, ‘cancer’, ‘s’, ‘patients’, ‘years’,
‘abraxane’, ‘testing’];
```

Ejemplo 4.5

4.3.5 Eliminación de características textuales cortas

La eliminación de características textuales cortas tiene como objetivo, eliminar palabras que hayan que son menores a tres caracteres y algunas palabras que no tengan un significado con el documento así como las stopwords, en el cual se aplica el ultimo filtro, dejando solo las palabras más importantes dentro de la bolsa de palabras.

Si utilizamos el resultado del Ejemplo 4.5

```
Nuevo_arreglo = Eliminar_carTextuales(Nuevo_Arreglo);
print Nuevo_arreglo;
‘Obtenemos como salida:
Nuevo_arreglo = [‘abraxane’, ‘cancer’, ‘patients’, ‘years’, ‘abraxane’,
‘testing’];
```

Ejemplo 4.6

Después de este paso en la publicación que se usó como ejemplo [De la Iglesia, 2013], hace un paso más, la autora realizó la búsqueda de las palabras derivadas utilizando el algoritmo de Porter para el idioma Inglés (Porter, 1980), que agrupa a las palabras que pertenecen a la misma familia reduciéndolas a su madre común, así desaparecen todas las palabras en plural y las deja todas en singular o si no existen en singular las deja en plural), al hacer este algoritmo la bolsa de palabras se reducía más, y no realice este paso para revisar las pruebas, y aun así tuve un buen resultado, es probable que cuando se categoricen más de dos categorías, este algoritmo tenga un mejor desempeño.

4.3.6 Obtención de frecuencias por documento

En esta etapa de pre-procesamiento, cogemos la bolsa de palabras de todo el documento y aplicamos la búsqueda de sus TF (frecuencia del término en un documento), y también eliminamos las palabras que están repetidas, dejando solo que se repitan una

vez cada una de ellas. Obteniendo un nuevo arreglo que contenía la palabra y el TF correspondiente.

Si utilizamos el resultado del Ejemplo 4.6

```
Arreglo_final = Obtener_frecuencias(Nuevo_arreglo);
print Arreglo_final;
```

‘Obtenemos como salida:

```
Arreglo_final = [(‘abraxane’,2), (‘cancer’, 1), (‘patients’,1),
(‘years’,1), (‘testing’,1)];
```

Ejemplo 4.7

4.3.7 Obtención de la suma de frecuencias

Después de aplicar la fórmula de frecuencia para todos los documentos, se hicieron dos pasos el primero fue que se sumó todas las frecuencias de los términos y se grabaron en un arreglo total, y el segundo paso fue obtener la frecuencia máxima, que es el número máximo que se obtuvo de frecuencias de cada termino que se obtuvo en todos los documentos, guardándolo en un arreglo distinto, esto nos ayudo a crear la primera regla para eliminar palabras dentro de la bolsa de palabras.

Si utilizamos el resultado del Ejemplo 4.7

```
Arreglo_final_documento_1 = [(‘abraxane’,2), (‘cancer’, 1), (‘patients’,1),
(‘years’,1), (‘testing’,1)];
```

```
Arreglo_final_documento_2 = [(‘abraxane’,3), (‘cancer’, 3), (‘patients’,1),
(‘years’,1), (‘testing’,1), (‘nanoparticle’, 4), (‘vaccine’, 2)];
```

•
•
•
•

```
Arreglo_final_documento_i = [(‘cancer’, 1), (‘patients’,1), (‘nanoparticle’,2),
(‘nose’,1), (‘lung’, 2), (‘lungs’,3), (‘close’, 10)];
```

```
Arreglo_Sumas = Aplicar_suma_frecuencias(Arreglo_final_1,
Arreglo_final_2, ..... Arreglo_final_i);
```

```
Arreglo_Maximos = Obtener_maximos(Arreglo_final_1, Arreglo_final_2,
..... Arreglo_final_i);
```

```
print Arreglo_Sumas;
print Arreglo_Maximos;
```

‘Obtenemos como salida:

```
Arreglo_Sumas = [(‘abraxane’,50), (‘cancer’, 10), (‘patients’,51), (‘years’,42),
(‘testing’,11), (‘nanoparticle’, 94), (‘vaccine’, 62), (‘nose’,25), (‘lung’, 14),
(‘lungs’,13), (‘close’, 10)];
```

```
Arreglo_Maximos = [(‘abraxane’,13), (‘cancer’, 3), (‘patients’,4), (‘years’,3),
(‘testing’,2), (‘nanoparticle’, 4), (‘vaccine’, 5), (‘close’, 10)];
```

Ejemplo 4.8

4.3.8 Primera regla para la eliminación de palabras

Obteniendo estos dos arreglos con las sumas de las frecuencias y los máximos correspondientes, aplico la primera regla de eliminación. La regla dice:

```
Primera_regla(X, Y){
  Para todos los X dentro de U{
    Para todos los fx dentro de X{
      Para todo Fx dentro de Y{
        Si (fxi = Fxi){
          Elimar_de_Arreglo(X, xi);
        }
      }
    }
  }
}
```

Teniendo en cuenta que:

- U** : conjunto de documentos
- X** : Arreglo de la suma de frecuencias
- Y** : Arreglo de los máximos de las frecuencias
- xi** : palabra i-ésima
- fxi** : frecuencia máxima de palabra i-ésima
- Fxi** : Suma de frecuencias de palabra i-ésima

Lo que trata de decir en esta regla es que si la suma del número de coeficiente TF es igual al máximo de frecuencias entonces esa palabra queda eliminada.

Si obtenemos los resultados del Ejemplo 4.8

```
Arreglo_Sumas = [(‘abraxane’,50), (‘cancer,’, 10), (‘patients’,51),
(‘years’,42), (‘testing’,11), (‘nanoparticle’, 94), (‘vaccine’, 62), (‘nose’,25),
(‘lung’, 14), (‘lungs’,13), (‘close’, 10)];
```

```
Arreglo_Maximos = [(‘abraxane’,13), (‘cancer,’, 3), (‘patients’,4), (‘years’,3),
(‘testing’,2), (‘nanoparticle’, 4), (‘vaccine’, 5), (‘close’, 10)];
```

```
Primera_regla(Arreglo_Sumas, Arreglo_Maximos);
print Arreglo_Sumas;
```

‘Como el único que es igual es la palabra ‘close (tiene la suma de frecuencia y el máximo igual) , esta se elimina y obtenemos como salida:

```
Arreglo_Sumas = [(‘abraxane’,50), (‘cancer,’, 10), (‘patients’,51),
(‘years’,42), (‘testing’,11), (‘nanoparticle’, 94), (‘vaccine’, 62), (‘nose’,25),
(‘lung’, 14), (‘lungs’,13)];
```

Ejemplo 4.9

4.3.9 Obtención de los IDF de los repositorios

Antes de aplicar la fórmula para obtener los IDF, cabe recalcar que aplicamos los pasos anteriores solo a la categoría con etiqueta “nano”, y los documentos que no tienen etiqueta no fueron usados para obtener la bolsa de palabras. Y para obtener las frecuencias en los documentos que no tenían la etiqueta “nano”, usamos la lista de palabras que encontramos en la categoría “nano”, y solo fueron calculadas las frecuencias en esa lista de documentos. Al obtener los dos arreglos totales después del paso anterior, obtenemos las bolsas de palabras con sus respectivas frecuencias de cada categoría.

Ahora que tenemos estos arreglos aplicamos la fórmula del IDF dada anteriormente para cada término y frecuencia de cada una de las categorías, obteniendo así dos arreglos nuevos con sus IDF correspondientes.

Si obtenemos el resultado del Ejemplo 4.9

```
Arreglo_Sumas_Nano = [(‘abraxane’,50), (‘cancer’, 10), (‘patients’,51),
(‘years’,42), (‘testing’,11), (‘nanoparticle’, 94), (‘vaccine’, 62), (‘nose’,25),
(‘lung’, 14), (‘lungs’,13)];
```

```
Arreglo_Sumas_NoNano = [(‘abraxane’,0), (‘cancer’, 30), (‘patients’,71),
(‘years’,40), (‘testing’,31), (‘nanoparticle’, 0), (‘vaccine’, 15), (‘nose’,16),
(‘lung’, 10), (‘lungs’,3)];
```

```

Arreglo_idf_nano = Obtener_idf(Arreglo_Sumas_Nano);
Arreglo_idf_Nonano = Obtener_idf(Arreglo_Sumas_NoNano);

print Arreglo_idf_Nano;
print Arreglo_idf_Nonano;

‘Obtendremos las salidas:
Arreglo_idf_Nano = [(‘abraxane’,4.3), (‘cancer’, 2.10), (‘patients’,2.9),
(‘years’,2.47), (‘testing’,1.1), (‘nanoparticle’, 0.94), (‘vaccine’, 0.32), (‘nose’,
0.15), (‘lung’, 0.03), (‘lungs’, 0.13)];

Arreglo_idf_NoNano = [(‘abraxane’,0), (‘cancer’, 1.0), (‘patients’,3.1),
(‘years’,2.50), (‘testing’,3.1), (‘nanoparticle’, 0), (‘vaccine’, 0.15),
(‘nose’,0.16), (‘lung’,0.01), (‘lungs’,0.3)];

```

Ejemplo 4.10

4.3.10 Segunda regla para la eliminación de palabras

Y como último paso para obtener la bolsa de palabras o unigrams final, se aplicó una segunda regla que del mismo modo que la primera: -se ejecutaron algunos modelos antes de aplicarla y cuando nos dimos cuenta que ciertas palabras se repetían tanto en el arreglo con etiqueta “nano” como en el arreglo que de etiqueta “no-nano”, decidimos aplicarla-, y al momento de aplicar el clasificador obtenía un porcentaje del 85% de aciertos. La autora de la publicación que use como ejemplo [De la Iglesia, 2013] llegó a un porcentaje de 92% de aciertos, por lo cual se tuvo que mejorar el pre-procesamiento y en este caso fue mejorando la bolsa de palabras creando una penalización con los resultados de IDF en los dos arreglos.

Esta segunda regla sigue estos pasos siguientes:

- 1) Al obtener los valores IDF de los dos arreglos con diferente categoría, se obtiene la distancia euclidiana de esta tomando como x_1 = el valor del IDF de la primera categoría y x_2 = el valor del IDF de la segunda categoría,

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$y_1 = 0$$

$$y_2 = 0$$

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2}$$

$$d_E(P_1, P_2) = x_2 - x_1$$

Formula 4.4

- 2) El y_1 y el y_2 son ceros, ya que los valores son de una sola dimensión, con este resultado, se aplica una penalización β , que representa a un valor mínimo entre un rango del valor absoluto entre [0.01 a 0.05]. Cuando obtienes un β que deseas aplicar; que si la distancia euclidiana que existe entre el IDF1 y el IDF2 se encuentra dentro del rango de $[-\beta, \beta]$, entonces la palabra es eliminada de la bolsa de palabras. En este caso probé con algunos β , pero el que me dio mejor resultado fue el $\beta = 0.03$

```

Segunda_regla(X, Y){
  Para todos los IDFx dentro de X{
    Para todo IDFy dentro de Y{
      Si (Distancia_euc(IDFxi, IDFyi) == rango [-β,β]){
        Eliminar_de_Arreglo(X, xi);
        Eliminar_de_Arreglo(Y, yi);
      }
    }
  }
}

```

Teniendo en cuenta que:

X: arreglo de los valores de IDF con etiqueta “nano”
Y: arreglo de los valores de IDF con etiqueta “no-nano”
xi: palabra i-ésima
IDF_{xi}: valor del IDF palabra i-ésima en el arreglo X
IDF_{yi}: valor del IDF palabra i-ésima en el arreglo Y
β: penalización

Si obtenemos el resultado del Ejemplo 4.10

Arreglo_idf_Nano = [(‘abraxane’,4.3), (‘cancer’, 2.10), (‘patients’,2.9), (‘years’,2.47), (‘testing’,1.1), (‘nanoparticle’, 0.94), (‘vaccine’, 0.32), (‘nose’, 0.15), (‘lung’, 0.03), (‘lungs’, 0.13)];

Arreglo_idf_NoNano = [(‘abraxane’,0), (‘cancer’, 1.0), (‘patients’,3.1), (‘years’,2.50), (‘testing’,3.1), (‘nanoparticle’, 0), (‘vaccine’, 0.15), (‘nose’,0.16), (‘lung’,0.01), (‘lungs’,0.3)];

```

Segunda_regla(Arreglo_idf_Nano, Arreglo_idf_NoNano);
print Arreglo_idf_Nano;
print Arreglo_idf_NoNano;

```

‘Obtenermos las salidas:

Arreglo_idf_Nano = [(‘abraxane’,4.3), (‘cancer’, 2.10), (‘testing’,1.1), (‘nanoparticle’, 0.94), (‘vaccine’, 0.32), (‘lung’, 0.03), (‘lungs’, 0.13)];

Arreglo_idf_NoNano = [(‘abraxane’,0), (‘cancer’, 1.0), (‘testing’,3.1), (‘nanoparticle’, 0), (‘vaccine’, 0.15), (‘lungs’,0.3)];

Ejemplo 4.11

Y es así como obtenemos la bolsa de unigrams final, ya que la diferencia entre las distancias euclidianas es una longitud considerable y tenemos los términos que realmente diferencian entre una categoría y otra.

La lista de unigrams se encuentra en el Anexo III.

4.4 Herramienta creada para hacer el Pre-procesamiento

Como se mencionó en el capítulo anterior se usó el lenguaje de programación llamado Visual Basic 2010 Express, los arreglos fueron reemplazados como una base de datos llamada Microsoft Access 2010, y se desarrolló una herramienta [Figura 4.6] para seguir estos 10 pasos antes de aplicar la creación del modelo y validar el clasificador.

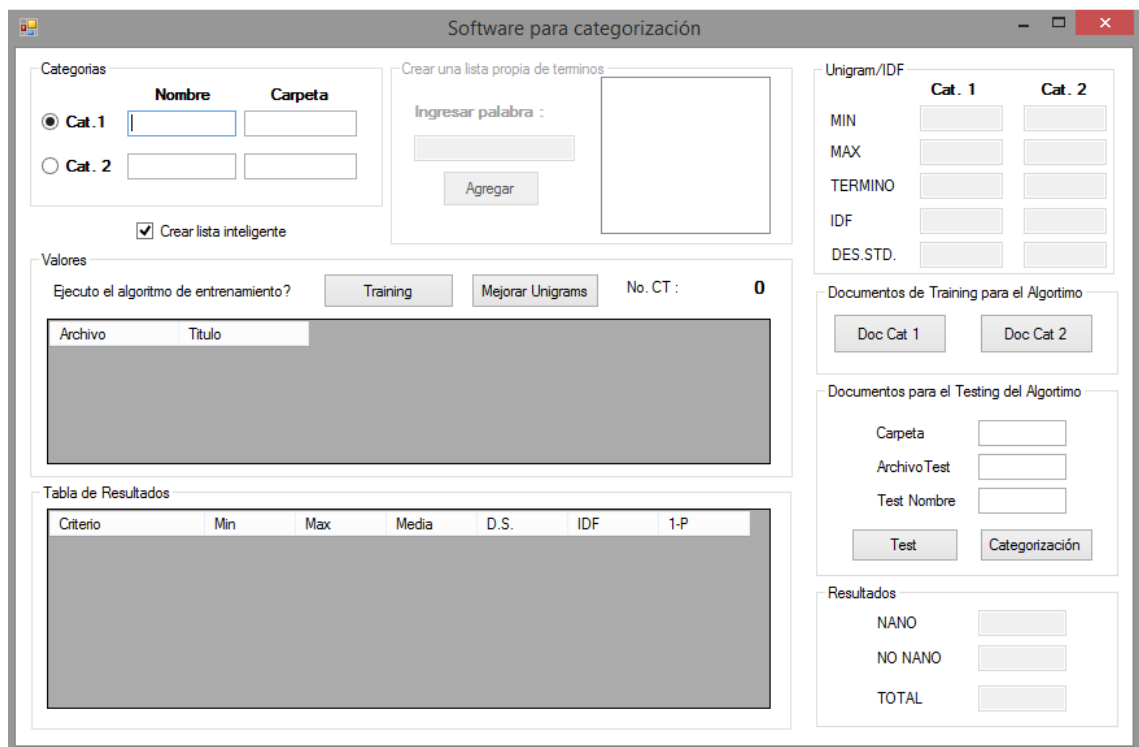


Figura 4.6. Pantalla principal de la Herramienta creada

En esta herramienta tenemos dos posibilidades para crear la bolsa de palabras o unigrams, una es para crear la lista inteligente usando los 10 pasos correspondientes y el otro es si ya tenemos una bolsa de palabras asignada y solo queremos transformar los archivos para que el clasificador pueda leerlos. En este caso se explicara cómo usar esta herramienta, y en donde están ubicados los 10 pasos de cada una.

Como primer paso es escribir la etiqueta de cada categoría que queremos separar, en este caso en el primer cuadro [Figura 4.7 a], observamos unos cuadros de texto para escribir el nombre de cada categoría, en este caso escribimos NANO y la NO_NANO, además de esto debemos elegir cuál de las dos categorías es la más importante [Figura 4.7 b], en este caso elegimos la categoría NANO como la primera en ejecutarse, esto conlleva a que la bolsa de palabras saldrán de los registros de la primera categoría. En el cuadro siguiente se escribe en que parte del ordenador están estos archivos, así no es necesario buscarlos por todo el ordenador. Y dejamos el visto en la caja llamada “Crear lista inteligente” [Figura 4.7 c], y por ultimo presionamos el botón “Training”.

<p>Categorías</p> <table border="1"> <thead> <tr> <th></th> <th>Nombre</th> <th>Carpeta</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="radio"/> Cat. 1</td> <td>NANO</td> <td>C:\Repositorio\N</td> </tr> <tr> <td><input type="radio"/> Cat. 2</td> <td>NO_NANO</td> <td>C:\Repositorio\N</td> </tr> </tbody> </table> <p>a) Elección de categorías</p>		Nombre	Carpeta	<input checked="" type="radio"/> Cat. 1	NANO	C:\Repositorio\N	<input type="radio"/> Cat. 2	NO_NANO	C:\Repositorio\N	<p>Categorías</p> <p><input checked="" type="radio"/> Cat. 1</p> <p><input type="radio"/> Cat. 2</p> <p>b) Elección de la principal</p>
	Nombre	Carpeta								
<input checked="" type="radio"/> Cat. 1	NANO	C:\Repositorio\N								
<input type="radio"/> Cat. 2	NO_NANO	C:\Repositorio\N								
<p><input checked="" type="checkbox"/> Crear lista inteligente</p> <p>c) Lista inteligente</p>										

Figura 4.7. Opciones de la herramienta: a) Nombre y elección de las categorías, b) se debe elegir la Categoría principal c) Debe estar activada la Lista Inteligente

Cuando presionamos el botón de training comenzara a crearse la lista en la fila de datos que se encuentran en la parte llamada “Valores”, lo cual mostrara el nombre del archivo que se está leyendo, el titulo del registro, la fecha en que fue publicada y el numero de unigrams encontrados en el mismo. [Figura 4.8]. Por cada paso que de la herramienta, esta le irá preguntando si desea seguir, como por ejemplo “Aplicar el algoritmo de entrenamiento” [Figura 4.9].

Valores

Ejecuto el algoritmo de entrenamiento? No. CT : **0**

Archivo	Título	Unigrams	Fecha
NCT00004705.xml	STUDY OF URI...	158	06/23/2005
NCT00005942.xml	LIPOSOMAL DA...	370	01/22/2013
NCT00039117.xml	OBLIMERSEN, C...	310	06/03/2013
NCT00046423.xml	A TRIAL OF ABI...	168	07/14/2008

Figura 4.8. Resultado de los archivos revisados, mostrando su título, número de unigrams y la fecha de publicación.

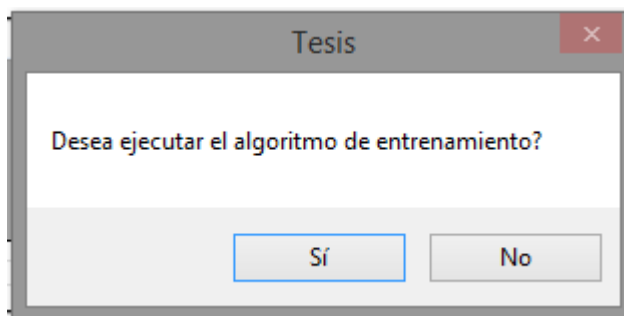


Figura 4.9. Pregunta principal para empezar el pre-procesamiento

Dentro de la función del botón de Training, se esconden los primeros pasos, hasta la obtención de los valores del IDF para la primera categoría que en este caso es la “NANO”, y se llena la tabla que es llamada TABLA DE RESULTADOS [Figura 4.10].

Tabla de Resultados

Criterio	Min	Max	Media	D.S.	IDF	1-P
study	4	16	7.9	3.81	0	0
uridine	0	4	0.4	1.2	1.7	9
triphosphate	0	3	0.3	0.9	1.7	9
utp	0	10	1	3	1.7	9
aerosol	0	3	0.3	0.9	1.7	9

Figura 4.10. Tabla de Resultados, mostrando los unigrams y los valores correspondientes de cada uno, la cantidad mínima de repetición del total de los documentos, el máximo de repetición, la media, la desviación estándar, el valor IDF, y el total de documentos en el cual no aparecieron.

Y para finalizar esta bolsa de palabras, le va a pedir que aplique la primera regla, para obtener las palabras que realmente se vaya a utilizar. [Figura 4.11], y al finalizar este proceso aparecerá un último mensaje comunicándonos que la categoría 1 está completa.

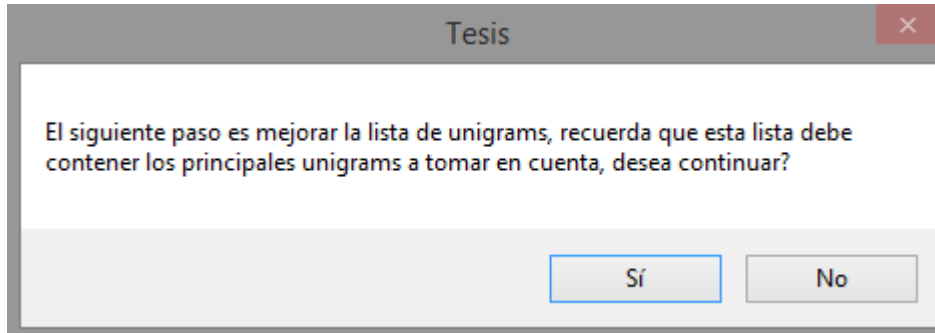


Figura 4.11. Cuadro que muestra si se desea seguir con el procesamiento

El siguiente paso es elegir la siguiente categoría que en este caso es la NO_NANO, para hacer los respectivos pasos para que termine el paso 9 del pre-procesamiento, que es la obtención de los valores de las dos categorías. De igual forma que en el anterior proceso nos mostrara un mensaje cuando ya haya finalizado el proceso con esta segunda categoría y nos mostrara el resultado en el cuadro total [Figura 4.12].

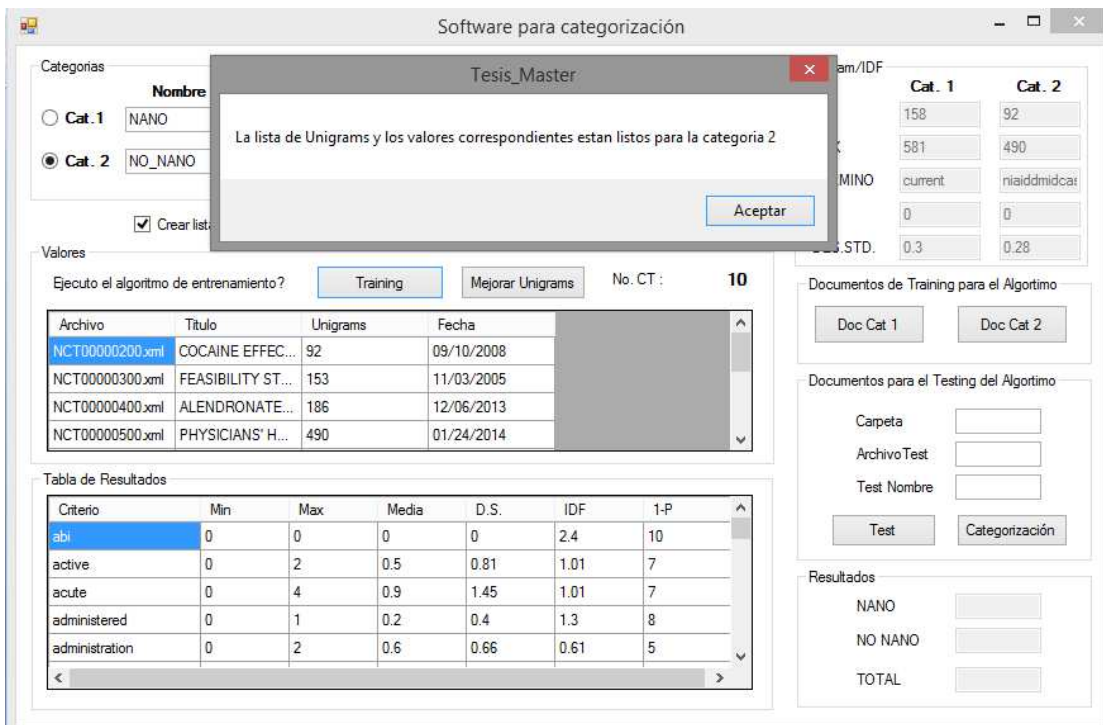


Figura 4.12. Captura total de la pantalla principal, con sus resultados correspondientes

Para completar el pre-procesamiento creamos presionamos el botón “Mejorar Unigrams”, que es la segunda regla que se creó para mejorar la bolsa de palabras, en este caso corresponde a los pasos 9 y 10 antes mencionados.

Al finalizar con este proceso, aparecerá un mensaje de finalización de las categorías y así empezar el proceso para la creación del modelo para aplicar el clasificador. [Figura 4.13]

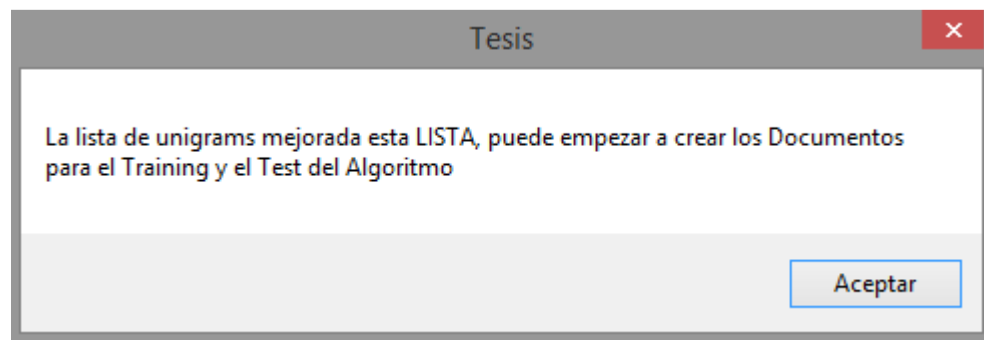


Figura 4.13. Ultimo mensaje para poder empezar a la creación del modelo para la predicción

Todo estos datos de arreglos son guardados en la base de datos correspondiente, y también se guarda la lista de palabras enumeradas para el uso del algoritmo, ahora que tenemos estos archivos listos podemos empezar con la creación del modelo y aplicar a los repositorios correspondientes.

4.5 Desarrollo del modelo

Lo primero que se debe hacer para obtener el modelo es transformar el repositorio de entrenamiento. Esta transformación es necesaria para que la herramienta Weka la pueda leer, esta transformación cumple con los siguientes pasos:

1. Escogemos el conjunto de registros de ensayos clínicos de entrenamiento con la etiqueta “NANO”.
2. Lo hacemos pasar por un proceso de transformación, la cual escoge los unigrams que tiene dentro del texto y lo transforma en un índice y seguido por el símbolo de dos puntos “:” se le agrega el valor del IDF de dicho término. Como se muestra en el ejemplo 4.12.

ID: 39 unigram: nanoparticle IDF: 2.34

Se transforma en:

[39:2.34]

Ejemplo 4.12

3. Hacemos esto con toda la lista de unigrams, y los que no existan en nuestra bolsa de palabras no aparecerán, ahora lo siguiente es al obtener todos estos valores se coloca al principio de los mismos una categoría, en este caso como vimos en la herramienta la categoría 1 es la NANO, a la cual le renombre como 1 y la categoría 2 NO_NANO la renombre como 2. Como se muestra en el ejemplo 4.13

1 39:2.34, 17:0.98, 1:0.34 1543:1.45

Ejemplo 4.13

4. El siguiente paso es hacer el mismo procedimiento con todos los documentos del repositorio y así obtener un solo documento con todos los valores. Hacemos el mismo procedimiento con la segunda categoría pero esta vez al principio colocamos la etiqueta de 2 ya no de 1, y con eso ya diferenciamos los registros de cada categoría. Esto se graba en un solo archivo. Y así obtenemos el archivo de texto listo para que sea aplicado el algoritmo y que nos muestre el modelo. Cabe recordar que usamos la herramienta Weka con un código de python para hacer este modelo, y así obtuvimos un resultado del 100% de efectividad [Figura 4.14].

Weka, mediante la librería LIBLINEAR nos ayuda con estos clasificadores y cada línea de código tiene un significado. La primera línea con la función `x,y = svm_read_problem('archivo')`, pasa los parámetros de x y y , tomando en cuenta que x es el índice de la palabra, y y es el valor del IDF de la palabra. Y en la siguiente línea de código es `m = train(y, x, '-s 6')`, esta función graba el modelo para etiquetar las categorías, y el valor de `'-s 6'`, significa que use la Regresión Logarítmica LASSO como clasificador. Y por último la línea de código `save_model('tesis9.model', m)`, graba el modelo que se obtuvo en la línea anterior, para poder usarlo en las predicciones en los repositorios nano y no nano respectivamente.

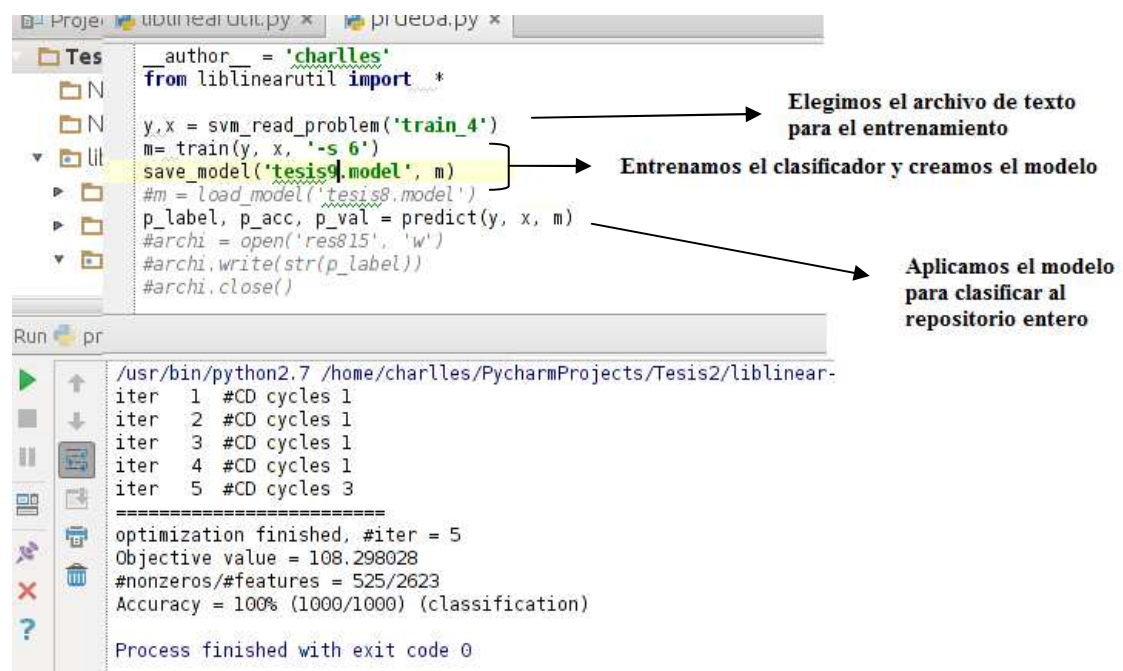


Figura 4.14. Creación del modelo y predicción del repositorio de entrenamiento

4.6 Aplicación del modelo y el algoritmo establecido

Como último paso para finalizar el procesamiento total de los repositorios, se aplicó el modelo obtenido en el paso anterior. Pero de igual forma que con los repositorios de entrenamiento se tuvo que transformar estos registros de ensayos clínicos en archivos de texto con los valores IDF. Para esto en la herramienta que se creó y se explicó anteriormente, se pueden hacer estas transformaciones [Figura 4.15] y dejar todo en un solo archivo de texto, pero como eran demasiados archivos estos fueron divididos en categorías para que su prueba pueda ser más rápida.

El único inconveniente entre estos pre-procesamientos es el tiempo que se demora en transformar estos registros en los archivos de texto, a lo que pondré como un ítem importante en las futuras investigaciones encontrar la manera de hacer estas transformaciones más rápidas.

El repositorio norte americano se dividió en ocho carpetas de 20.000 archivos cada una, una 9na carpeta con el resto que en este caso es 26.727 y una 10ma carpeta con el repositorio europeo total de 25.657 registros de ensayos clínicos.

Figura 4.15. Parte de la pantalla en donde se escribe el nombre de la carpeta en donde se encuentran los ensayos clínicos, el nombre del archivo para predecir, y el nombre de un archivo para escribir los nombres de los ensayos

La herramienta que se creó tiene 3 cuadros de texto, el primero llamado “Carpeta”, es el nombre de la carpeta que contiene todos estos registros, el segundo cuadro de texto es el llamado “Archivo Test” que es donde se pone el archivo de texto en donde se grabaran los valores de los IDF, para cada uno de estos les puse el nombre de T_n , (n representando el numero de carpeta) y el cuadro de texto final es el llamado “Test Nombre” que es como el anterior cuadro de texto, el nombre del archivo el cual contendrá el nombre de cada uno de los registros para hacer una comparación manual de cada uno de los registros que se reviso, esto se discutirá en el siguiente capítulo del trabajo de fin de máster.

Al transformar estos archivos todos aparecerán con la categoría 2, o sea con la etiqueta “NO NANO”.

Al igual que los archivos de entrenamiento estos fueron usados con la herramienta Weka, pero solo para clasificar y predecir en que categoría debe estar, utilizando el modelo ya predeterminado anteriormente. [Figura 4.16].

```
y, x = svm_read_problem('t2')
m = load_model('tesis8.model')
p_label, p_acc, p_val = predict(y, x, m)
```

Figura 4.16. Pedazo de código, en el cual se muestra como se agrega un archivo para predecirlo, el modelo a usar y la línea para predecir

En la línea de código `m = load_model('tesis8.model')`, se refiere a que va a utilizar el modelo que fue creado en el entrenamiento del repositorio, y por ultimo en la línea de código que presenta la función `predict(y, x, m)`, es donde hacen las comparaciones y presenta el uso del clasificador, los resultados se encuentran en el siguiente capítulo.

5 RESULTADOS

En este capítulo de resultados se presentara las tablas de resumen de cada parte de la metodología, y los resultados finales. Se probó con 12 modelos de entrenamiento diferentes, y se encontró que el mejor resultado fue el último ya que se agrego una tercera regla dentro de la primera regla para la disminución de la bolsa de palabras, pero esto se lo hizo para un futuro trabajo de categorizar los documentos con etiqueta “nano”, que se explicara más adelante.

Además se tomo diferentes muestras de conjunto de entrenamiento, desde 250 ensayos clínicos con etiqueta “nano” y 250 ensayos clínicos con etiquetas “no-nano”, hasta 500 ensayos clínicos con etiqueta “nano” y 26253 ensayos clínicos con etiqueta “no-nano”.

Toda esta explicación se la hará en el siguiente punto del capítulo, con las tablas correspondientes. Cabe recalcar que los ensayos clínicos con etiqueta “nano” que fueron usados para el modelo de entrenamiento fueron clasificados manualmente, siendo para los primeros modelos o para los modelos llamados “De la Iglesia” y de igual forma los que tenían la etiqueta “no-nano”.

5.1 Resultados para la aplicación del modelo de entrenamiento

En esta parte del capítulo cinco se mostraran tablas con el nombre del modelo (el tipo de repositorio de ensayos que se uso), el número de ensayos clínicos con etiquetas nano, el número de ensayos clínicos con etiqueta “no-nano”, que reglas fueron usadas, el porcentaje de aciertos que nos dio el modelo, y el numero de unigrams usados para la comparación.

En este resumen también se explicara porque existiendo modelos que me daban el 100% en el training no fueron usados para el testing. No se usaron estos modelos, porque al momento de ser probados en el testing daban porcentajes muy malos de clasificación, ya que del total de documentos del testing que fueron 212.001 y se escogía aleatoriamente 250 ensayos clínicos para ver si su clasificación era correcta. De igual forma en la tabla 5.2 se darán los resultados de esta prueba.

Para los primeros seis modelos de entrenamiento, para obtener la bolsa de palabras se buscaron aleatoriamente en el repositorio de ClinicalTrials.gov 250 ensayos clínicos que contenían las palabras “nano, nanoparticle” y 250 ensayos clínicos que no tenían nada que ver con nanotecnología y los denominamos “no nano”, y para los seis modelos más de entrenamiento; es decir, para los restantes se uso para obtener la bolsa de palabras que iba a ser usado como el modelo de clasificación los 500 ensayos clínicos con etiqueta “nano” y los 500 ensayos clínicos con etiqueta “no nano” que fueron usados en

la tesis doctoral tomada como ejemplo. En la tabla 5.1 se muestran los nombres de los modelos y estos son sus significados:

1. **UNO.-** Contiene 250 ensayos clínicos buscados en el repositorio de clinical trials y que contenían la clave “nanoparticle” y contiene 250 ensayos buscados al azar que no contenían esta clave.
2. **DOS.-** Contiene 250 ensayos clínicos buscados en el repositorio de clinical trials y que contenían la clave “nanoparticle” y contiene 20000 ensayos buscados al azar que no contenían esta clave.
3. **TRES.-** Contiene 250 ensayos clínicos buscados en el repositorio de clinical trials y que contenían la clave “nanoparticle”, 6 ensayos clínicos que contenían la palabra “nanomaterial” y 108 ensayos que contenían la palabra “nano” que en total sumados da como resultado 364 ensayos clínicos y contiene 20000 ensayos buscados al azar que no contenían esta clave.
4. **CUATRO.-** Contiene 250 ensayos clínicos buscados en el repositorio de clinical trials y que contenían la clave “nanoparticle”, 6 ensayos clínicos que contenían la palabra “nanomaterial” y 108 ensayos que contenían la palabra “nano”, y 80 ensayos clínicos con la palabra “abraxane” y que en total sumados da como resultado 444 ensayos clínicos y contiene 26253 ensayos buscados al azar que no contenían esta clave.
5. **CINCO.-** Contiene 500 ensayos clínicos con etiqueta “nano” que fueron buscados manualmente, basados en la tesis doctoral que se usó para mejorar el algoritmo, y contiene 26253 ensayos buscados al azar que no contenían las claves “nanoparticle, abraxane, nano, nanomaterial y liposomal”.
6. **SEIS.-** Contiene la unión de los 500 ensayos clínicos con etiqueta “nano” basados en la tesis doctoral mas los 444 del conjunto anterior que se uso en los modelos, dando una intersección entre ellos de 533 documentos, y contiene 26166 ensayos buscados al azar que no contenían las claves “nanoparticle, abraxane, nano, nanomaterial y liposomal, y otros términos de nanomedicina”.
7. **SIETE.-** Contiene la unión de los 500 ensayos clínicos con etiqueta “nano” basados en la tesis doctoral más 10 con nuevos términos de nanomedicina, y contiene 500 ensayos clínicos con etiqueta “no-nano” que se uso en la tesis doctoral y 10 más para igualar al conjunto de etiqueta “nano”. (Anexo I)

8. **DE LA IGLESIA.**- Contiene 500 ensayos clínicos basados con etiqueta “nano” en la tesis doctoral que se usó para mejorar el algoritmo, y contiene 500 ensayos clínicos con etiqueta “no-nano” que se uso en la tesis doctoral.

Además en la tabla 5.1 se muestran las dos reglas usadas más uno que se uso solo en el último modelo:

1. **REGLA UNO.**- se explica en el capítulo de metodología.
2. **REGLA DOS.**- se explica en el capítulo de metodología.
3. **REGLA TRES.**- se trata es que dentro de la regla uno se adjunta una condicional: si es que la palabra contiene el prefijo “nano” o “liposom”, no se debe aplicar la regla uno.

Estas reglas nos sirven para mejorar el algoritmo clasificador disminuyendo la bolsa de palabras o aumentándola. En esta tabla a continuación se muestra además de los porcentajes se muestra en la última columna el numero de unigrams que se obtuvieron sin aplicar ninguna regla o aplicando las tres reglas, y como se puede observar la cantidad de palabras, recordando que para los primeros 6 modelos se uso 250 ensayos clínicos al azar nano y no nano y para los 6 últimos los ensayos clínicos usados por la tesis doctoral.

MODELO	TIPO	ENSAYOS CLÍNICOS		REGLAS			%	UNIGRAMS
		NANO	NO NANO	PRIMERA	SEGUNDA	TERCERA		
1ER	UNO	250	250	NO	NO	NO	67	14624
2DO	UNO	250	250	SI	NO	NO	72	3565
3ERO	UNO	250	250	SI	SI	NO	99.8	2625
4TO	DOS	250	20000	SI	SI	NO	99.97	2625
5TO	TRES	364	20000	SI	SI	NO	100	2625
6TO	CUATRO	444	26253	SI	SI	NO	100	2625
7MO	DE LA IGLESIA	500	500	NO	NO	NO	83	19880
8VO	DE LA IGLESIA	500	500	SI	SI	NO	100	3556
9NO	CINCO	500	26253	SI	SI	NO	96	3556
10MO	SEIS	533	26166	SI	SI	NO	100	3556
11RO	DE LA IGLESIA	500	500	SI	SI	SI	100	3579
12DO	SIETE	510	510	SI	SI	SI	100	3579

Tabla 5.1.- Tabla que demuestra los tipos de modelos y sus resultados porcentuales, desde el primero hasta el que se uso

Para probar cada modelo de entrenamiento, se tomo una fracción del repositorio de clinicaltrials.gov y se escogieron aleatoriamente 250 ensayos clínicos y se los reviso manualmente para ver si su clasificación esta correcta. La tabla 5.2 muestra los resultados de cada uno de los modelos usados con sus nombres respectivos, y en la Figura 5.1 se muestra los resultados exactos en cada uno de los modelos establecidos.

MODELO	%
1ER	56.00%
2DO	69.60%
3ERO	76.40%
4TO	78.80%
5TO	82.00%
6TO	84.40%
7MO	61.20%
8VO	87.60%
9NO	89.20%
10MO	90.00%
11RO	94.40%
12DO	98.00%

Tabla 5.2.- Resultado porcentual de resultados de los 250 ensayos clínicos

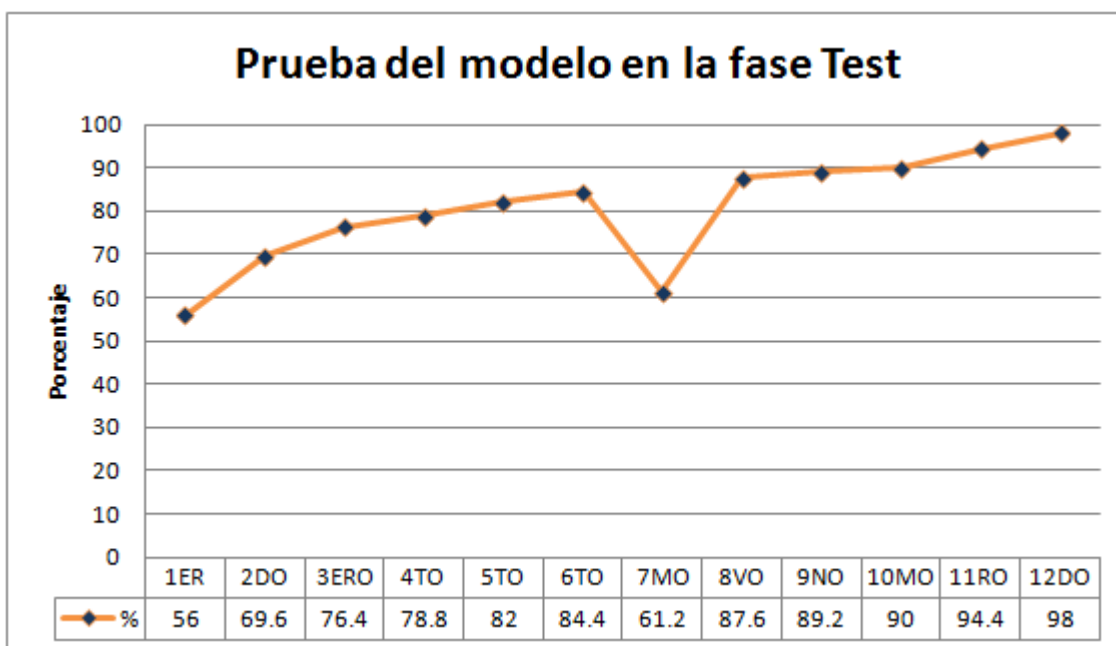


Figura 5.1.- Evolución de los modelos de entrenamiento

Con estos resultados, decidimos elegir el modelo no. 12, ya que con esto también se probó la teoría de que no importa que tan grande sea el modelo de entrenamiento en la regresión logística, lo que importa es que se diferencien entre el uno y el otro, y al probarlo con 250 ensayos clínicos al azar nos demostraban que era un mejor clasificador. Esta figura demuestra los porcentajes del modelo de entrenamiento.

Al comparar el modelo 11 al modelo 12 existe nos dimos cuenta que existe un buen porcentaje de distancia; es decir, al aumentar los 10 ensayos clínicos mas, se cubrió muchas más conceptos de nanomedicina, ya que los 500 primeros son ensayos de hasta el 2013 y los 10 nuevos son usados dentro de los años siguientes hasta el 2015. Y así como escogimos los tipo “nano”, se escogió 10 documentos con la etiqueta “no-nano”, pero para hacerlo más comparativo se buscó ensayos que contenían ciertos dispositivos que tienen en su nombre la palabra *nano* pero que realmente solo significa que son los más pequeños de su grupo de productos en comparación a los otros tamaños normales; como por ejemplo el Accu-Check Aviva Nano (Figura 5.2) que es un dispositivo para medir la presión, pero que solo es el más pequeño de toda su gama de este tipo de productos, pero que no están dentro de la escala nanométrica (1 a 100 nm) como debe ser para ser considerado como un nanomaterial; y gracias al agregar estos ensayos clínicos la comparación mejoró.

Los 250 ensayos clínicos para la prueba del modelo clasificador fueron escogidos en el repositorio de ClinicalTrials.gov, buscando ensayos con la palabra nano, y así encontramos diferentes ensayos clínicos que usaban materiales o dispositivos con nombre “nano”, pero esto no significaba que sean nanomateriales o nanodispositivos.



Figura 5.2.- Accu-Check Aviva Nano (dispositivo para medir la presión), dispositivo que en su nombre tiene la etiqueta “nano”, pero no es un nanodispositivo

Otro punto importante de analizar en esta parte del modelo fue el tiempo que se demoro en aplicar los 10 pasos para obtener la bolsa de valores. Se uso una portátil con Windows versión 8.1 con 8 Gb de RAM, y procesador Intel Core I7 con 2.45 Ghz de velocidad para la creación del modelo y para el testing se uso una máquina virtual con Linux Ubuntu versión 12 con 1 Gb de RAM.

Para esto los 10 pasos (Figura 4.1) los separe en fases:

Fase 1.- Es la extracción de todas las palabras de los documentos de registros de ensayos clínicos con etiqueta “nano” que fueron los 5 primeros pasos del pre-procesamiento

Fase 2.- Es la ejecución para obtener los diferentes valores necesarios de cada palabra dentro de la bolsa de palabras extraída de los documentos con etiqueta “nano” para poder ejecutar las reglas, estos son los pasos 6 y 7 y el 9a.

Fase 3.- Es la ejecución de la primera regla y tercera regla, este es el 8vo paso.

Fase 4.- Es la extracción de las palabras del conjunto de documentos de registros de ensayos clínicos con etiqueta “no-nano”, esto es repetir los primeros 5 pasos.

Fase 5.- Es obtener los valores necesarios de cada palabra para poder ejecutar la tercera regla, solo que esta vez es a los documentos “no-nano”, estos son los pasos 6, 7 y 9b.

Fase 6.- La aplicación de la segunda regla y es el paso final el 10.

La tabla 5.3 mostrará el tiempo requerido por cada una de las fases:

Fases	Dia	Horas	Minutos
Primera	2	6	15
Segunda	3	8	26
Tercera	0	11	21
Cuarta	1	0	12
Quinta	0	10	5
Sexta	0	0	3
TOTAL	7 dias 8 horas y 32 minutos		

Tabla 5.3.- Tiempo total para crear la bolsa de unigrams

El tiempo que tomo hacer el modelo fue de 26 segundos en la máquina virtual, el tiempo en transformar los dos repositorios en documentos que puedan ser usados por el algoritmo de regresión lineal fue de 1 día 18 horas, y por último el tiempo para darnos los porcentajes del clasificador variaba según el tamaño del repositorio, en este caso fue entre 1 minuto 26 segundos y 2 minutos 03 segundos.

Ahora si después de obtener el modelo de entrenamiento resultante, y saber el tiempo que este demoró en ser creado, como paso siguiente fue clasificar los dos repositorios en donde se obtuvieron los documentos de registros de ensayos clínicos.

5.2 Resultados de la base de ClinicalTrials.gov (Norte América)

Después de obtener el modelo a usar para la clasificación se lo uso con todo el repositorio de documentos de ensayos clínicos de medicina del portal ClinicalTrials.gov, y estos fueron los resultados:

ENSAYOS CLÍNICOS					
	NO NANO	CANT.	NANO	CANT.	TOTAL
parte 1	91.935%	18387	8.065%	1613	20000
parte 2	91.775%	18355	8.225%	1645	20000
parte 3	90.380%	18076	9.620%	1924	20000
parte 4	90.585%	18116	9.415%	1884	20000
parte 5	90.055%	18011	9.945%	1989	20000
parte 6	90.245%	18049	9.755%	1951	20000
parte 7	90.095%	18019	9.905%	1981	20000
parte 8	89.120%	17824	10.880%	2176	20000
parte 9	88.752%	23371	11.248%	2962	26333
		TOTAL	9.673%	18125	186333

Tabla 5.4.- Resultado del repositorio Norteamericano. Contiene todas las 9 partes que se dividió el repositorio para hacer la comparación y el número de ensayos clínicos con etiqueta “nano”.

	model 10	cantidad	model 12	cantidad
NANO	2.9306%	5461	9.6732%	18125
NO NANO	97.0694%	180872	90.3268%	168208
TOTAL US	186333			186333

Tabla 5.5.- Valores que se obtuvieron probando el modelo 10 y el modelo final para la clasificación para el repositorio Norteamericano.

Como se explicó en un capítulo anterior, este repositorio se dividió en 9 partes por la cantidad de ensayos que eran, y en la tabla 5.4 se muestra la cantidad exacta de ensayos clínicos que se obtuvieron después de completar la fase del testing, y el porcentaje del clasificador. La última parte del repositorio norteamericano contiene los ensayos entre los años 2013 a 2015. Y para la tabla 5.5, se muestra la comparación de dos de los modelos usados para la clasificación, pero como se explicó en el punto anterior se usó el modelo 12 como clasificador y obtenemos que solo el 9.6732 % de los ensayos clínicos hablan sobre nanomedicina y sus componentes, que da un total de 18125.

5.3 Resultados de la base de ClinicalTrialsRegister.eu (Europa)

Después de obtener la clasificación del repositorio anterior, el modelo clasificador también se lo uso con todo el repositorio de documentos de ensayos que representa Europa:

ENSAYOS CLÍNICOS					
	NO NANO	CANT.	NANO	CANT.	TOTAL
parte 1	87.506%	22461	12.49%	3227	25688
		TOTAL	12.49%	3227	25688

Tabla 5.6.- Resultados del repositorio Europeo. En este repositorio no fue necesaria la división de sus ensayos por lo cual se tomo una sola parte del mismo.

	model 10	cantidad		model 12	cantidad
NANO	0.4753%	122		12.4942%	3207
NO NANO	99.5247%	25546		87.5058%	22461
TOTAL EU		25668			25668

Tabla 5.7.- Valores que se obtuvieron probando el modelo 10 y el modelo final para la clasificación para el repositorio Europeo

Este repositorio no fue necesario dividirlo, y en la tabla 5.6 se muestra la cantidad exacta de ensayos clínicos que se obtuvieron después de completar la fase del testing, y el porcentaje del clasificador.

Y para la tabla 5.7, se muestra la comparación de dos de los modelos usados para la clasificación, el modelo 10 y el modelo 12, pero como se explicó en el punto anterior se usó el modelo 12 como clasificador y así obtuvimos que solo el 12.4942 % de los ensayos clínicos hablan sobre nanomedicina y sus componentes, que da un total de 3207 de documentos.

6 DISCUSIÓN

Los recientes avances en la nanomedicina han provocado un cambio muy grande en el mundo, por sus grandes investigaciones y descubrimientos que han ayudado al ser humano para combatir las enfermedades más fuertes y difíciles como es el cancer, pero al mismo tiempo todas las ciencias que la tratan de ayudar para mejorarla como la informática han tenido que avanzar a la par, y uno de sus casos es lo que hicimos en este trabajo; que es la minería de texto, que por lo que requiere a su vez mejorar los avances relacionados con la extracción, gestión e integración de datos; es decir, avanzar en la nanoinformática. Esto depende de acceder a las fuentes de manera eficiente y eficaz. Obtener datos de nanomedicina y saber dónde encontrarlos es uno de los caminos que se debe tomar para poder incrementar estos avances y es por ahí, que cada vez se crean muchas herramientas de gestión de datos sobre nanotecnología. Y cuando un investigador necesita este tipo de información sepa dónde encontrar sin necesidad de leer cada uno de las publicaciones o ensayos clínicos que se tiene, si no directamente escoger el que necesita. En este trabajo después de mostrar los resultados se puede discutir en donde se encuentran esta mayoría de ensayos que hablan sobre nanomedicina o sobre nanotecnología en sí.

Cabe recalcar que esta tesis presenta una nueva contribución al diseño, modelado y análisis del dominio en la nanomedicina en términos de mostrar que uno puede detectar automáticamente la clasificación de un objetivo relacionado con esta área. Se ha creado un modelo de clasificación, con capacitación y conjuntos de pruebas que se pueden utilizar para desarrollar aplicaciones computacionales extendidas para el apoyo a la investigación en el campo de la nanomedicina. El enfoque presentado ha producido buenos resultados con la ayuda del subconjunto de ensayos clínicos extraídos de ClinicalTrials.gov que fueron clasificados manualmente por un trabajo anterior, y así este método puede utilizarse de manera fiable para determinar automáticamente si el ensayo clínico contiene el uso de nanofármacos, nanodispositivos o nanomateriales. Como se dijo en un capítulo anterior se usa un algoritmo (L1-regularizado regresión logística o LASSO) que fue capaz de hacer frente a un problema de tan alta dimensión, tanto en términos de rendimiento de la clasificación y el coste computacional.

Ahora ya en el capítulo anterior obtuvimos los dos resultados necesarios, podemos discutirlos. Se podrá compararlos detalladamente. Con algunas explicaciones sobre porque se uso la ultima regla –la tercera regla-, y porque se cambio los primeros ensayos clínicos por los que fueron usados en la tesis doctoral.

Primero se explicara porque se uso la tercera regla para poder obtener más unigrams en la bolsa de palabras final. En el siguiente capítulo se hablara sobre un futuro trabajo que se hará con esta bolsa de palabras, y para poder hacer que este proceso sea más fácil se hizo una exclusión al aplicar la regla uno –si su cantidad máxima de repeticiones era

igual al tf total en todos los documentos, estos documentos se eliminaban instantáneamente-, pero la tercera regla es excluir en esta eliminación a todas las palabras que contengan el prefijo “nano” y “liposom”; ya que con el primer termino nos damos cuenta que cualquier nombre que contenga “nano” como prefijo es muy probable que se trate de una nanopartícula, un nanodispositivo o un nanomaterial, lo que es muy diferente que el término “nano” se encuentre solo y no como prefijo.

Y el termino “liposom” se debe a la palabra “liposome” –liposoma- que es una pequeña esfera (vesícula), hecho con el mismo material que la membrana celular; es decir, se hacen generalmente de fosfolípidos, que tienen un grupo de cabeza y un grupo de cola, y la cabeza se siente atraído por el agua (hidrosoluble), y la cola, que está hecho de una larga cadena hidrocarbonada, es repelido por el agua (liposoluble). Pero lo que realmente importa es que cuando se lo usa en la medicina tiene un tamaño variable que va de 20 nm a 200 nm de longitud; por ende, con esta longitud podemos decir que esta dentro del concepto de nanopartícula. Y uno de sus principales objetivos es que las moléculas al interior de un liposoma pueden ser dirigidas a células específicas, siendo esto una posible respuesta de cómo llevar medicamento a células enfermas disminuyendo los efectos tóxicos en células sanas, ya que los liposomas pueden evitar el reconocimiento y destrucción del sistema inmune. (Figura 6.1)

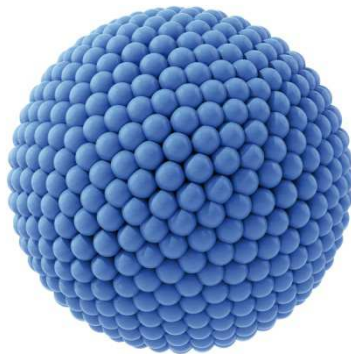


Figura 6.1.- Liposoma

Como segunda explicación tomaremos en cuenta las bolsas de palabras finales que se obtuvieron en los 12 modelos de entrenamiento. En los seis primeros modelos se observa que el número de unigrams es muy diferentes al número de unigrams en los últimos seis. Esto se debe a que en los primeros seis modelos se uso para crear la bolsa de palabras 250 ensayos obtenidos aleatoriamente con clave “nano”, “nanoparticle” o “nanomaterial”, pero no fueron revisados manualmente para observar si hablaban o no de nanomedicina o nanotecnología. De igual forma para aquellos que fueron escogidos como los “no nano”, y por este inconveniente no llegaban a un buen modelo para la clasificación. Pero en los últimos seis modelos se uso lo 500 ensayos que fueron revisados manualmente en un estudio previo que hablaban sobre nanomedicina o nanotecnología por lo que mejoro eficientemente el modelo final, y con la ayuda de las

reglas se iban reduciendo unigrams (primera y segunda regla) o aumentando unigrams (tercera regla de exclusión).

Ahora como último punto se comparó las dos tablas de los ensayos clínicos de Norteamérica y de Europa y tenemos dos teorías; ya que si comparamos sus porcentajes de los documentos clasificados como “nano” se obtiene que el repositorio europeo contiene aproximadamente un 3% más que el norteamericano, y podemos concluir que en Europa está fomentando más información sobre el tema de nanomedicina o nanotecnología. Pero si tomamos como variable de comparación el total de ensayos clínicos que tiene que ver sobre nanomedicina o nanotecnología y sus dispositivos concluimos que existen aproximadamente 6 veces más ensayos clínicos en el repositorio norteamericano que el europeo. (Tabla 6.1)

	%	cantidad
NANO	12.4942%	3207
NO NANO	87.5058%	22461
TOTAL EU		25668
NANO	9.6732%	18125
NO NANO	90.3268%	168208
TOTAL US		186333

Tabla 6.1.- Comparación de los dos repositorios y sus respectivos números de ensayos clínicos y sus porcentajes respectivos.

Otro punto que debemos tomar en cuenta es el tiempo en que estos repositorios han estado funcionando y así comparar realmente por los años de publicaciones. El primer repositorio *ClinicalTrials.gov* comenzó a trabajar en 1997 pero de forma privada solo para el gobierno, pero a comienzos del 2000 comenzó a publicar todos sus trabajos e invitó a que nuevos investigadores suban sus publicaciones o estudios; entonces decimos que este repositorio comenzó desde el año 2000. Y el segundo repositorio *ClinicalTrialsRegister.eu*; el repositorio europeo, fue creado a mediados del 2004 y comenzó su funcionamiento en ese mismo día; entonces decimos que el repositorio comenzó del 2004. Ahora según la pagina del primer repositorio en el año 2004 contenía 10,237 documentos de registros de ensayos clínicos así que aunque restemos este total de ensayos al total que tiene hasta ahora el segundo repositorio no alcanzaría ni la tercera parte del primer repositorio.

Otra observación es acerca de las estadísticas de la evolución de las publicaciones del primer repositorio y nos damos cuenta que para poder llegar a los casi 25000 documentos se demoró 6 años y medio (Figura 3.24), y el segundo repositorio ya lleva más de 10 años para poder alcanzar los 25000 ensayos clínicos. Ahora si comparamos el número de los ensayos clínicos norteamericanos de los últimos años (parte 9) con el total de ensayos europeos, podemos decir que el total de estos es muy parecido al número de todos los registros de repositorio europeo.

Estos resultados nos indican que Europa debe usar mas su repositorio para así tener más documentos de ensayos clínicos y no todos introducirlos en el repositorio norteamericano u otro internacional, porque de igual forma en el capítulo de estado del arte en donde se explica acerca estos repositorios, se da a entender que en *ClinicalTrials.gov* el total de los ensayos clínicos que son propios de Estados Unidos solo son el 41% en total y que los otros son de otra parte del mundo incluyendo Europa.

7. CONCLUSIÓN Y TRABAJOS FUTUROS

En este capítulo hablaremos de las conclusiones del trabajo realizado. En este trabajo a fin de máster se ha demostrado, a través de una revisión detallada por medio del estado del arte las técnicas más usadas en minería de texto, de cómo los investigadores de todo el mundo están tratando de superar los desafíos actuales de la nanomedicina, que son las barreras existentes para acceder y gestionar los datos generados en investigación nanomédica. Y por medio del análisis de los diferentes modelos de entrenamiento, a través de la exploración y gestión de datos en ensayos reales se pudo observar que esta herramienta puede ayudar a algunas tareas para la investigación en nanomedicina, ilustrando claves y procesamientos convenientes sin limitaciones en la cantidad de documentos en las grandes bases de datos que pertenecen al dominio nanomédico. Basados en los objetivos propuestos podemos concluir que se ha alcanzado todo lo antes mencionado, pero iremos explicando los puntos que se tomaron en consideración.

7.1 Conclusiones

Como objetivo general de este trabajo de fin de máster fue la comparación de dos repositorios de registros de ensayos clínicos -ClinicalTrials.gov y ClinicalTrialsRegister.eu- tomando en cuenta que contienen información relacionada sobre la nanotecnología dentro de la medicina (nanomedicina), y así poder proporcionar un número aproximado sobre la cantidad de registros de ensayos clínicos que estos repositorios. Por lo cual se llegó a obtener este número aproximado de documentos y fueron etiquetados como aquellos que ensayos clínicos "nano" y aquellos que no tenían definiciones o términos sobre nanotecnología como "no-nano". Con esto nos dimos cuenta que el primer repositorio (ClinicalTrials.gov) tiene muchos más documentos de nanomedicina o sobre nanotecnología que el repositorio europeo (ClinicalTrialsRegister.eu), y que si tomamos en cuenta los porcentajes el repositorio europeo tiene mayor proporción que el repositorio norteamericano, con un 12.49% y 9.673% respectivamente.

También se demostró que se pudo proporcionar un nuevo método de pre-procesamiento de texto que se basa en 10 pasos especificados en el capítulo de metodología, utilizándolos para poder diferenciar entre estos registros de ensayos clínicos, y para poder hacer más fácil este entrenamiento se creó un herramienta que se dividía en 6 fases para crear el modelo de entrenamiento, de igual forma especificadas en el capítulo de metodología. Gracias a este método aplicado en los repositorios se pudo descubrir nuevos registros valiosos que contienen información de nanomedicina y nanotecnología, que es muy relevante para investigadores que desean buscar nuevos métodos para sus estudios tanto en la literatura científica como en los resúmenes de ensayos clínicos disponibles en la Web. Además este método no solo se puede usar a

ensayos clínicos que estén en formato xml, también puede ser usado si están en texto plano o en cualquier formato de lectura normal.

Hemos aportado al análisis para la gestión de datos dentro de la nanomedicina, obteniendo resultados de los ensayos clínicos de dos diferentes y grandes bases de datos disponibles en la web que con llevan casi el 70% del mundo investigativo, pudimos hacer la extracción de los términos potenciales y nuevos patrones en los datos de áreas nanomédicas como la genotoxicidad y la terapia dirigida de fármacos, donde ciertos patrones y tendencias subyacentes podrían conducir a inferencias que informan a las futuras investigaciones en nanomedicina, otra parte importante es que pudimos diferenciar entre un nanodispositivo a un dispositivo que contenga el nombre "nano" y que solo signifique que es pequeño en comparación al grupo de sus productos. Así concluimos que este trabajo podría facilitar investigadores en el descubrimiento de nuevos conocimientos de forma automática a partir de los ensayos clínicos médicos.

7.2 Trabajo futuro

Dentro de esta parte del capítulo se hablara acerca de los futuros usos que se le puede dar a esta herramienta y del futuro uso de los resultados de los datos obtenidos en los repositorios.

Primero con los datos obtenidos en estos dos repositorios se pueden usar para poder clasificar a nuevas categorías futuras, como poder dividir estos registros de ensayos clínicos en nanomedicinas, nanotoxicidad, nanodispositivos, nanomateriales, propiedades físicas, biológicas, químicas de las nanopartículas, diagnósticos, terapias, métodos y algunas categorías que tengan que ver con la nanotecnología. Y por esta razón fue que se creó la regla número tres en los resultados, ya que con mas palabras que tengan relación con la nanotecnología o la nanomedicina, esta clasificación podría hacerse más fácil. Justamente por eso se dejo en el Anexo III la lista de estas palabras para que se pueda clasificar a futuro y crear un nuevo método o usar el mismo para esa clasificación.

Segundo la herramienta que se creó se la puede usar para clasificar cualquier otra categoría, no solo sobre nanotecnología, lo único que se tendría que hacer es modificar la herramienta para que pueda actualizar una lista de prefijos o sufijos que hagan las exclusiones de la regla numero uno.

Figura 7.1. En la herramienta solo está marcada la lista inteligente pero en el cuadro de alado está listo para poder modificarse para la lista de palabras

Tercero en la herramienta que se creó existe una opción en la cual se puede agregar una lista que sea la bolsa de palabras y no necesariamente hacer la lista inteligente que recorre los 10 pasos o las 6 fases de entrenamiento para crear el modelo de entrenamiento, ya que en esta herramienta solo se puede crear la lista inteligente y no la otra. (Figura 7.1)

Otro trabajo futuro que se puede hacer con esta herramienta es mejorar el tiempo del modelamiento, ya que como se mostro en el capítulo de los resultados, poder transformar los 1010 ensayos clínicos en una bolsa de palabras que nos sirva para poder hacer la clasificación. Recordando que casi se tomo 7 días y medio para poder crearlos.

Y con esto terminaríamos los posibles trabajos futuros con los resultados obtenidos en este trabajo de fin de máster.

ANEXO I: LISTA DE ENSAYOS CLÍNICOS SEPARADOS MANUALMENTE EXTRAIDOS DE CLINICALTRIALS.GOV

Primero se pondrá la tabla de los ensayos con etiqueta “NO NANO”

ENSAYOS NO NANO				
NCT00000200	NCT00187109	NCT00389792	NCT00596700	NCT01004770
NCT00000300	NCT00187161	NCT00390000	NCT00600600	NCT01009892
NCT00000400	NCT00189800	NCT00391300	NCT00604500	NCT01047176
NCT00000500	NCT00192400	NCT00395200	NCT00605111	NCT01047644
NCT00000600	NCT00193700	NCT00396500	NCT00608400	NCT01071005
NCT00000700	NCT00195000	NCT00396838	NCT00609986	NCT01075971
NCT00000800	NCT00200551	NCT00399100	NCT00610584	NCT01088529
NCT00000900	NCT00204100	NCT00402961	NCT00611000	NCT01094990
NCT00001000	NCT00205400	NCT00404300	NCT00613600	NCT01113502
NCT00001100	NCT00206700	NCT00409500	NCT00617500	NCT01125579
NCT00001126	NCT00208000	NCT00412100	NCT00618007	NCT01135147
NCT00001400	NCT00209300	NCT00413400	NCT00621400	NCT01135212
NCT00001500	NCT00210600	NCT00415090	NCT00622700	NCT01147588
NCT00001560	NCT00211900	NCT00418600	NCT00622934	NCT01150721
NCT00001600	NCT00213200	NCT00421200	NCT00624000	NCT01154894
NCT00002000	NCT00215800	NCT00426400	NCT00625300	NCT01172730
NCT00002100	NCT00219700	NCT00427700	NCT00630500	NCT01182961
NCT00002200	NCT00221000	NCT00429000	NCT00631800	NCT01201785
NCT00002400	NCT00221884	NCT00430300	NCT00633321	NCT01211626
NCT00002554	NCT00226200	NCT00431600	NCT00633347	NCT01221532
NCT00003000	NCT00227500	NCT00434200	NCT00634400	NCT01222845
NCT00003600	NCT00232700	NCT00435500	NCT00634660	NCT01237535
NCT00004200	NCT00235300	NCT00436449	NCT00635700	NCT01239823
NCT00004300	NCT00235677	NCT00439400	NCT00637000	NCT01251380
NCT00004331	NCT00236600	NCT00440024	NCT00638690	NCT01255124
NCT00004400	NCT00240500	NCT00442000	NCT00639730	NCT01257893
NCT00004500	NCT00248300	NCT00444600	NCT00642200	NCT01265732
NCT00004800	NCT00252200	NCT00447200	NCT00643500	NCT01270919
NCT00005225	NCT00253552	NCT00449800	NCT00644800	NCT01305811
NCT00005600	NCT00256100	NCT00451100	NCT00648700	NCT01316796
NCT00006100	NCT00258700	NCT00452400	NCT00650000	NCT01317394

NCT00006118	NCT00260000	NCT00453544	NCT00651300	NCT01319994
NCT00006138	NCT00261300	NCT00453700	NCT00652600	NCT01327443
NCT00006200	NCT00263900	NCT00455000	NCT00655200	NCT01331590
NCT00006400	NCT00264004	NCT00456300	NCT00655343	NCT01332994
NCT00014300	NCT00265200	NCT00456547	NCT00660400	NCT01343303
NCT00018200	NCT00270400	NCT00457600	NCT00661700	NCT01348490
NCT00020605	NCT00270998	NCT00458900	NCT00665600	NCT01352507
NCT00023400	NCT00273000	NCT00460200	NCT00666900	NCT01362023
NCT00032500	NCT00275600	NCT00461500	NCT00667095	NCT01371617
NCT00033566	NCT00276900	NCT00469300	NCT00670800	NCT01374360
NCT00036400	NCT00278018	NCT00470600	NCT00672100	NCT01376128
NCT00037700	NCT00278200	NCT00471900	NCT00674076	NCT01396876
NCT00039000	NCT00279500	NCT00473200	NCT00674700	NCT01412827
NCT00047762	NCT00280969	NCT00475644	NCT00678158	NCT01413178
NCT00048100	NCT00286000	NCT00475800	NCT00679900	NCT01418261
NCT00048464	NCT00287300	NCT00479700	NCT00681590	NCT01418846
NCT00048516	NCT00288600	NCT00483600	NCT00682500	NCT01435031
NCT00049985	NCT00292500	NCT00484900	NCT00685100	NCT01443585
NCT00054600	NCT00293800	NCT00486200	NCT00689000	NCT01445275
NCT00059800	NCT00296114	NCT00487110	NCT00691600	NCT01460446
NCT00061100	NCT00299000	NCT00487500	NCT00692562	NCT01461772
NCT00075400	NCT00301093	NCT00488800	NCT00695500	NCT01472055
NCT00079300	NCT00301600	NCT00490100	NCT00696800	NCT01477918
NCT00079677	NCT00305500	NCT00492466	NCT00698100	NCT01480011
NCT00081354	NCT00309634	NCT00492700	NCT00705900	NCT01481311
NCT00081900	NCT00310700	NCT00496600	NCT00709800	NCT01504763
NCT00083863	NCT00313300	NCT00504400	NCT00710632	NCT01508156
NCT00087100	NCT00314912	NCT00508300	NCT00712400	NCT01513018
NCT00088400	NCT00315900	NCT00512213	NCT00712803	NCT01529749
NCT00092300	NCT00315939	NCT00513500	NCT00713700	NCT01529996
NCT00094900	NCT00318500	NCT00518700	NCT00717600	NCT01533259
NCT00097500	NCT00322309	NCT00520104	NCT00718770	NCT01545245
NCT00098800	NCT00325000	NCT00521300	NCT00720200	NCT01574001
NCT00100100	NCT00328900	NCT00527800	NCT00721461	NCT01575847
NCT00101400	NCT00330200	NCT00530400	NCT00721500	NCT01582178
NCT00107900	NCT00331500	NCT00533000	NCT00729300	NCT01582477
NCT00109200	NCT00334867	NCT00534300	NCT00734500	NCT01586780
NCT00110500	NCT00335400	NCT00543374	NCT00735501	NCT01591317
NCT00111800	NCT00338000	NCT00543400	NCT00737100	NCT01591876
NCT00114140	NCT00342433	NCT00545961	NCT00738400	NCT01606085
NCT00114400	NCT00343200	NCT00546000	NCT00742300	NCT01614613

NCT00117000	NCT00345800	NCT00547300	NCT00747981	NCT01620619
NCT00119600	NCT00346697	NCT00549900	NCT00750100	NCT01621126
NCT00120900	NCT00347100	NCT00551200	NCT00757900	NCT01624428
NCT00121524	NCT00349700	NCT00552500	NCT00760500	NCT01625104
NCT00125463	NCT00351000	NCT00553670	NCT00765700	NCT01635504
NCT00126100	NCT00354679	NCT00555100	NCT00786006	NCT01639833
NCT00130000	NCT00356200	NCT00556400	NCT00789477	NCT01644968
NCT00131300	NCT00358800	NCT00556465	NCT00795535	NCT01647958
NCT00133900	NCT00360100	NCT00557700	NCT00800956	NCT01649752
NCT00140400	NCT00362700	NCT00557765	NCT00805038	NCT01650103
NCT00144300	NCT00362726	NCT00558012	NCT00814372	NCT01650948
NCT00145600	NCT00364000	NCT00559000	NCT00828906	NCT01650961
NCT00146731	NCT00365300	NCT00560300	NCT00831129	NCT01652690
NCT00150800	NCT00366600	NCT00561600	NCT00831610	NCT01655784
NCT00155441	NCT00369200	NCT00562900	NCT00832078	NCT01665417
NCT00156000	NCT00370500	NCT00568100	NCT00835159	NCT01666340
NCT00157300	NCT00371800	NCT00569400	NCT00843167	NCT01667055
NCT00158600	NCT00373100	NCT00569608	NCT00860314	NCT01670578
NCT00159900	NCT00375154	NCT00570700	NCT00862784	NCT01675687
NCT00161200	NCT00378300	NCT00572000	NCT00890981	NCT01676935
NCT00165100	NCT00379392	NCT00573300	NCT00895635	NCT01679171
NCT00166400	NCT00379600	NCT00574600	NCT00898313	NCT01682980
NCT00170300	NCT00380900	NCT00577200	NCT00919425	NCT01685268
NCT00171600	NCT00380926	NCT00578500	NCT00920127	NCT01687959
NCT00174200	NCT00383500	NCT00579943	NCT00928122	NCT01699828
NCT00178100	NCT00384813	NCT00581100	NCT00928759	NCT02022176
NCT00179400	NCT00386100	NCT00584883	NCT00937885	NCT02053428
NCT00182000	NCT00386152	NCT00589602	NCT00993863	NCT02078908
NCT00184600	NCT00387673	NCT00594100	NCT00997828	NCT02205385
NCT00185900	NCT00388050	NCT00596635	NCT01002833	NCT02283320

La siguiente es la tabla de los ensayos con etiqueta “NANO”.

ENSAYOS NANO				
NCT00004705	NCT00428662	NCT00731380	NCT01125215	NCT01494415
NCT00005942	NCT00432562	NCT00732836	NCT01127113	NCT01495884
NCT00039117	NCT00436410	NCT00733408	NCT01147016	NCT01506973
NCT00046423	NCT00436709	NCT00734682	NCT01148745	NCT01521520
NCT00046514	NCT00441376	NCT00736619	NCT01149798	NCT01523314
NCT00046527	NCT00442260	NCT00736944	NCT01151592	NCT01525966

NCT00046930	NCT00442910	NCT00738361	NCT01155375	NCT01550848
NCT00059605	NCT00444314	NCT00740584	NCT01155388	NCT01552135
NCT00073723	NCT00447889	NCT00741741	NCT01158079	NCT01555853
NCT00077090	NCT00453817	NCT00748163	NCT01161186	NCT01565421
NCT00077246	NCT00454324	NCT00748553	NCT01163071	NCT01565499
NCT00081042	NCT00456846	NCT00753740	NCT01167985	NCT01566357
NCT00085124	NCT00459810	NCT00754039	NCT01169870	NCT01566435
NCT00087347	NCT00461149	NCT00764621	NCT01169935	NCT01567553
NCT00093119	NCT00462423	NCT00769093	NCT01188226	NCT01572038
NCT00093145	NCT00466960	NCT00775359	NCT01188408	NCT01577238
NCT00093223	NCT00466986	NCT00777673	NCT01190982	NCT01577537
NCT00095914	NCT00470548	NCT00781872	NCT01192165	NCT01578603
NCT00103038	NCT00472693	NCT00785291	NCT01193855	NCT01583426
NCT00103506	NCT00473720	NCT00788892	NCT01201057	NCT01589562
NCT00107094	NCT00477529	NCT00794092	NCT01204437	NCT01598480
NCT00107484	NCT00479674	NCT00816829	NCT01206023	NCT01598493
NCT00110084	NCT00479856	NCT00818259	NCT01207102	NCT01599078
NCT00110695	NCT00481455	NCT00819351	NCT01224587	NCT01599845
NCT00111904	NCT00483301	NCT00820768	NCT01227616	NCT01604109
NCT00113607	NCT00484120	NCT00821964	NCT01227941	NCT01612546
NCT00121745	NCT00490152	NCT00822094	NCT01229839	NCT01616303
NCT00124943	NCT00499252	NCT00825201	NCT01234012	NCT01620190
NCT00129896	NCT00499291	NCT00826085	NCT01236716	NCT01620918
NCT00140140	NCT00500422	NCT00833807	NCT01238939	NCT01625429
NCT00147238	NCT00502606	NCT00837200	NCT01238952	NCT01625936
NCT00148148	NCT00503750	NCT00844649	NCT01242163	NCT01627041
NCT00177684	NCT00503906	NCT00848042	NCT01243320	NCT01640847
NCT00180037	NCT00504998	NCT00851721	NCT01244282	NCT01641783
NCT00184041	NCT00505271	NCT00851877	NCT01251809	NCT01644890
NCT00185029	NCT00505713	NCT00856492	NCT01258192	NCT01646762
NCT00188695	NCT00518284	NCT00864253	NCT01259570	NCT01647672
NCT00193206	NCT00520000	NCT00866307	NCT01264679	NCT01652079
NCT00195013	NCT00521781	NCT00875693	NCT01270139	NCT01655693
NCT00233597	NCT00531271	NCT00876070	NCT01278095	NCT01661335
NCT00243594	NCT00537056	NCT00876486	NCT01291342	NCT01663090
NCT00249002	NCT00540514	NCT00876824	NCT01291550	NCT01665846
NCT00250874	NCT00541931	NCT00877253	NCT01292369	NCT01667211
NCT00251472	NCT00542958	NCT00882180	NCT01296139	NCT01669239
NCT00254592	NCT00544648	NCT00882973	NCT01298011	NCT01674257
NCT00255424	NCT00549848	NCT00886717	NCT01300533	NCT01677559
NCT00255437	NCT00553462	NCT00894023	NCT01304303	NCT01679470

NCT00255450	NCT00570674	NCT00903942	NCT01304797	NCT01689194
NCT00274443	NCT00572130	NCT00905034	NCT01305512	NCT01689610
NCT00274456	NCT00583349	NCT00910741	NCT01307891	NCT01690702
NCT00281528	NCT00585689	NCT00912639	NCT01313078	NCT01693276
NCT00281697	NCT00585936	NCT00915369	NCT01323296	NCT01695187
NCT00284752	NCT00589082	NCT00920023	NCT01323894	NCT01696084
NCT00291473	NCT00594529	NCT00926484	NCT01324180	NCT01698281
NCT00291603	NCT00602316	NCT00934895	NCT01326078	NCT01698710
NCT00294996	NCT00607438	NCT00942292	NCT01336062	NCT01702129
NCT00295893	NCT00609791	NCT00944047	NCT01336803	NCT01704365
NCT00299598	NCT00616967	NCT00951054	NCT01344473	NCT01705158
NCT00306618	NCT00617942	NCT00951613	NCT01369108	NCT01707758
NCT00307255	NCT00617981	NCT00957320	NCT01374750	NCT01709019
NCT00308178	NCT00618657	NCT00960687	NCT01374919	NCT01715168
NCT00309959	NCT00621101	NCT00961116	NCT01376310	NCT01726894
NCT00313599	NCT00622973	NCT00964080	NCT01380769	NCT01730833
NCT00319839	NCT00625573	NCT00968604	NCT01386203	NCT01731210
NCT00320541	NCT00626405	NCT00970996	NCT01387997	NCT01732458
NCT00328497	NCT00629499	NCT00972946	NCT01403415	NCT01734746
NCT00331032	NCT00635284	NCT00978562	NCT01405794	NCT01736943
NCT00331630	NCT00637403	NCT00983424	NCT01406756	NCT01746225
NCT00333502	NCT00637572	NCT00989131	NCT01411904	NCT01748903
NCT00337376	NCT00637728	NCT00995488	NCT01412229	NCT01754948
NCT00337714	NCT00637806	NCT01003808	NCT01416558	NCT01763710
NCT00346229	NCT00637897	NCT01007240	NCT01420211	NCT01770795
NCT00355888	NCT00638079	NCT01007890	NCT01420588	NCT01773850
NCT00356980	NCT00642759	NCT01008969	NCT01420718	NCT01777230
NCT00360425	NCT00643461	NCT01009970	NCT01426126	NCT01792479
NCT00361842	NCT00654836	NCT01010035	NCT01432223	NCT01793818
NCT00363610	NCT00659126	NCT01010945	NCT01432275	NCT01802333
NCT00368589	NCT00659178	NCT01020578	NCT01433068	NCT01804101
NCT00370357	NCT00659204	NCT01023347	NCT01435720	NCT01804530
NCT00372281	NCT00659334	NCT01035658	NCT01436123	NCT01806571
NCT00377559	NCT00659776	NCT01036087	NCT01436604	NCT01807845
NCT00377780	NCT00660543	NCT01041235	NCT01437007	NCT01812746
NCT00389428	NCT00661167	NCT01050777	NCT01437020	NCT01815294
NCT00389714	NCT00662129	NCT01052779	NCT01437722	NCT01815333
NCT00392392	NCT00666991	NCT01054313	NCT01442974	NCT01825382
NCT00394082	NCT00667147	NCT01057264	NCT01446016	NCT01830244
NCT00394251	NCT00670618	NCT01057342	NCT01455389	NCT01848652
NCT00394810	NCT00670670	NCT01059058	NCT01461915	NCT01858935

NCT00397761	NCT00671034	NCT01062191	NCT01463072	NCT01878708
NCT00398086	NCT00672165	NCT01074255	NCT01463709	NCT01885013
NCT00400348	NCT00673257	NCT01088815	NCT01464294	NCT01892540
NCT00404235	NCT00675259	NCT01089335	NCT01464593	NCT01906385
NCT00404404	NCT00683059	NCT01090128	NCT01464996	NCT02104752
NCT00407407	NCT00689065	NCT01101529	NCT01465087	NCT02123030
NCT00407563	NCT00691054	NCT01102504	NCT01467986	NCT02247726
NCT00407888	NCT00691912	NCT01103791	NCT01470417	NCT02340858
NCT00409383	NCT00707876	NCT01113476	NCT01478685	NCT02367547
NCT00411788	NCT00709761	NCT01114139	NCT01488552	NCT02370589
NCT00416455	NCT00712881	NCT01114204	NCT01490047	NCT02400957
NCT00418951	NCT00721747	NCT01114217	NCT01490203	NCT02407977
NCT00421187	NCT00723125	NCT01120171	NCT01492101	NCT02408874
NCT00427414	NCT00729612	NCT01122771	NCT01493310	NCT02414750

ANEXO II: LISTA DE LAS STOPWORDS DE PUBMED HASTA EL 2015

Esta información fue sacada de las páginas propias de PubMED:

STOPWORDS PUBMED						
about	beyond	gone	letter	other	since	toward
above	both	got	like	others	slightly	towards
abs	but	gov	ltd	otherwise	so	try
accordingly	by	had	made	ought	some	type
across	came	has	mainly	our	somehow	ug
after	can	have	make	ours	someone	under
afterwards	cannot	having	many	ourselves	something	unless
again	cc	he	may	out	sometime	until
against	cm	hence	me	over	sometimes	up
all	come	her	meanwhile	overall	somewhat	upon
almost	compare	here	mg	owing	somewhere	us
alone	could	hereafter	might	own	soon	use
along	de	hereby	ml	oz	specifically	used
already	dealing	herein	mm	particularly	still	usefully
also	department	hereupon	mo	per	strongly	usefulness
although	depend	hers	more	perhaps	studied	using
always	did	herself	moreover	pm	sub	usually
am	discover	him	most	precede	substantially	various
among	dl	himself	mostly	predominantly	such	very
amongst	do	his	mr	present	sufficiently	via
an	does	how	much	presently	take	was
analyze	done	however	mug	previously	tell	we
and	due	hr	must	primarily	th	were
another	during	i	my	promptly	than	what
any	each	ie	myself	pt	that	whatever
anyhow	ec	if	namely	quickly	the	when
anyone	ed	ii	nearly	quite	their	whence
anything	effected	iii	necessarily	rather	theirs	whenever
anywhere	eg	immediately	neither	readily	them	where
applicable	either	importance	never	really	themselves	whereafter
apply	else	important	nevertheless	recently	then	whereas
are	elsewhere	in	next	refs	thence	whereby
arise	enough	inc	no	regarding	there	wherein
around	especially	incl	nobody	relate	thereafter	whereupon

as	et	indeed	noone	said	thereby	wherever
assume	etc	into	nor	same	therefore	whether
at	ever	investigate	normally	seem	therein	which
be	every	is	nos	seemed	thereupon	while
became	everyone	it	not	seeming	these	whither
because	everything	its	noted	seems	they	who
become	everywhere	itself	nothing	seen	this	whoever
becomes	except	just	now	seriously	thorough	whom
becoming	find	keep	nowhere	several	those	whose
been	for	kept	obtained	shall	though	why
before	found	kg	of	she	through	will
beforehand	from	km	off	should	throughout	with
being	further	last	often	show	thru	within
below	gave	latter	on	showed	thus	without
beside	get	latterly	only	shown	to	wk
besides	give	lb	onto	shows	together	would
between	go	ld	or	significantly	too	wt
yet	your	yourself	you	yours	yourselves	yr

ANEXO III: LISTA DE UNIGRAMS USADOS PARA CLASIFICAR LAS ETIQUETAS NANO Y NO NANO

En este anexo se muestra las 3579 palabras que fueron usadas para clasificar.

BOLSA DE UNIGRAMS				
aall	contraceptives	health	murphy	right
abbott	contraindication	healthcare	muscle	rising
abdomen	contraindications	healthy	mutans	risk
abdominal	contralateral	heart	mutations	risks
abi	contrast	heat	myelin	ritonavir
ability	contribute	hec	myelodysplastic	rituximab
ablation	control	help	myeloid	robert
able	controlled	hematologic	myeloma	rochester
abnormal	controls	hematological	myelomonocytic	rock
abnormalities	conventional	hematology	myelosuppression	role
abnormality	cooper	hematopoietic	myocardial	roma
above	cooperative	hemochromatosis	myocet	romania
abramson	cord	hemodialysis	nab	room
abraxane	core	hemoglobin	nadir	root
abraxis	cornell	heparin	naïve	roswell
abscess	corporation	hepatic	nancy	rouge
absence	corrected	hepatitis	nano	route
absolute	correlate	hepatocellular	nanoburning	routine
absorption	correlated	herbal	nanocolloid	roy
abstain	correlation	herceptin	nanocrystal	royal
abstinence	correlative	herneu	nanodispersion	rule
abx	corresponding	herpes	nanoemulsion	run
academy	corticosteroid	hershey	nanoliposomal	russia
acceptability	corticosteroids	hgb	nanoliposomes	russian
acceptable	cost	higher	nanom	ryan
accepted	coumadin	highest	nanomaterial	sacramento
accepts	count	highly	nanomaterials	sacred
access	counts	hill	nanomedicine	safe
accessible	county	hills	nanometer	safely
accident	covered	hisher	nanometers	safety
accordance	cyp	histologic	nanoparticle	safetyefficacy
according	cpt	histological	nanoparticles	saginaw
account	cpx	histologically	nanopharmaceutical	saint
accrual	cranial	histology	nanoscale	saalem

accrued	crcl	historical	nanose	salina
accuracy	cream	history	nanosensors	saline
accurate	creatinine	hiv	nanoshells	saliva
accurately	creek	hnsc	nanospectra	salivary
acetate	crel	hodgkin	nanosuspension	salt
achieve	cremophor	hodgkins	nanotechnology	salvage
achieved	crf	hoff	narrowing	samaritan
acid	cri	holy	nasal	same
acidic	crisis	homa	nashville	samples
acids	criterion	home	nasopharyngeal	sampling
acp	critical	hong	national	samyang
acquired	crlx	honolulu	native	san
acquisition	cross	hope	natural	sanford
across	crossover	hopital	nature	sanitaire
act	cse	hôpital	nausea	sant
acting	cses	hopkins	ncd	santa
activated	csf	hormonal	nci	sao
activation	ctc	hormone	near	são
active	ctcae	hormones	nebraska	sarah
activities	ctr	hospitalier	necessary	sarcoma
activity	culture	hospitalization	neck	sarcomas
actual	cumulative	hospitals	necrosis	satisfy
acute	curative	hot	necrotic	scale
adaptive	curatively	houston	needle	scan
adc	cure	howard	needs	scanner
added	cured	hvp	negative	scanning
addition	currently	hrs	nelfinavir	scans
additional	curve	hum	neo	schedule
additionally	cut	human	neoadjuvant	schedules
adenocarcinoma	cvc	humans	neoplasia	school
adequate	cycle	hungary	neoplasm	sciences
adequately	cycles	hurley	neoplasms	scl
adhesive	cyclin	hutchinson	neoplastic	sclerosis
adjacent	cyclodextrin	hybrid	nervous	score
adjusted	cyclophosphamide	hybridization	netherlands	scores
adjuvant	cyclosporine	hydrochloride	network	scott
administer	cypa	hydrocortisone	neu	scottsdale
administered	cysteine	hydroxyurea	neurodegeneration	screened
administering	cystic	hygiene	neurodegenerative	scri
administration	cysts	hypersensitivity	neurologic	seasonal
admission	cyt	hypertension	neurological	seattle

adnce	cytarabine	hypertensive	neuronal	second
adolescents	cytokine	hyperthermia	neuropathy	secondary
adriamycin	cytokines	hypothesis	neurosurgery	seconds
adult	cytological	hypothesize	neurotoxicity	secreted
adults	cytologically	hysterectomy	neutropenia	section
advanced	cytology	iabn	neutrophil	sécurité
advantage	cytosine	ida	neutrophils	sedation
advantages	cytotoxic	idaho	neuwelt	see
adverse	cytoxan	idarubicin	nevada	segment
aes	czech	identification	new	seizure
affairs	dacarbazine	identified	newark	seizures
affect	daily	identify	newly	select
affecting	dakhil	identifying	newton	selected
affects	dakota	igg	next	selective
affiliated	dallas	ihc	ngml	selectively
africa	damage	iaa	nhl	self
afssaps	dana	iib	nhs	sensitive
after	daniel	iii	niaid	sensitivity
age	danville	iiia	nih	sensor
aged	dasatinib	iiib	nijmegen	sensors
agence	database	iiic	nilotinib	sensory
agency	date	iliac	nippon	sent
agent	dated	illinois	nitro seas	sentinel
agents	daunomycin	illness	nodal	seoul
ages	daunorubicin	illnesses	node	sep
aggressive	daunoxome	illnesssocial	nodes	separate
ago	david	image	non	separately
agonist	davidi	images	nondialysis	september
agree	davis	imaging	none	sequence
agreed	day	imf	noninvasive	sequences
agreement	days	immediate	nonmelanoma	sequential
agrees	dayton	immediately	normal	ser
aids	dce	immune	normalization	serbia
aim	dcis	immunodeficiency	normalized	serial
aims	dds	immunogenicity	north	serious
air	deaconess	immunohistochemical	northern	serous
akron	dealing	immunohistochemistry	northwest	serum
alabama	death	immunologic	northwestern	service
alanine	debulking	immunosuppressive	norway	services
alaska	dec	immunotherapy	nose	set
alberta	decadron	impact	not	setting

albumin	decatur	impair	note	seven
albuquerque	decision	impaired	noted	severe
alcohol	decitabine	impairment	nov	severity
alexander	decline	imperfect	novavax	sexual
algorithm	decrease	implantation	novel	sexually
ali	decreased	implanted	november	sgot
alive	deemed	implants	novo	sgpt
alkaline	deep	important	nps	shanghai
allergic	deficiency	impregnated	nsabp	she
allergies	define	improve	nsc	shock
allergy	defined	improved	nsclc	shore
allocation	definite	improvement	nuclear	short
allogeneic	definition	improves	nugent	shorter
allow	definitions	improving	number	showing
allowed	definitive	imrt	nursing	shrinkage
allowing	degradation	inability	nutrition	side
allows	del	inactive	nyha	sided
alone	delaware	inadequate	oakland	sidney
along	delay	inadequately	oakwood	sign
alopecia	delayed	inappropriate	oatpb	signal
alpha	delays	incidence	obese	signed
already	deli	include	objective	significance
alter	deliver	included	objectives	significant
altered	delivered	includes	oblimersen	signing
alternative	demand	including	observation	signs
amag	dementia	inclusionexclusion	observational	silica
ambisome	demonstrate	inclusive	observed	silver
amenable	demonstrated	incompatible	obstruction	similar
american	demonstrating	incorporated	obstructive	simon
aminotransferase	dendritic	increase	obtain	simultaneous
amir	denis	increased	obtaining	sinai
aml	dense	increases	occlusion	sinerem
amorphous	density	increasing	occur	singapore
amount	dent	incurable	occurred	single
amphotericin	dental	independence	occurrence	sinus
amplification	dentin	independent	occurs	sioux
amy	dentistry	index	oconto	sirna
analog	denver	india	oct	sirolimus
analysed	department	indiana	october	site
analyses	dependent	indianapolis	octreotide	siteman
analysis	depending	indicate	offer	sites

analyze	dept	indicated	offered	situ
analyzed	der	indicates	office	situation
analyzing	derived	indicating	ohio	situations
anaphylactic	dermal	indication	ohsu	size
anc	des	indications	oil	sized
anchorage	describe	indicator	oklahoma	skimmilk
anderson	described	indices	old	skin
andor	description	indinavir	omaha	sleep
androgen	descriptive	individual	onc	sloan
anemia	designed	individuals	oncaspar	slow
anesthesia	despite	induce	once	small
aneurysm	detail	inducers	oncol	smaller
angeles	detailed	inducing	oncologia	smallest
angina	detect	induction	oncología	smith
angiogenesis	detectable	inductionre	oncological	smoking
angiographic	detected	industry	oncologico	sns
angiography	detecting	ineligible	oncologist	social
angioplasty	detection	inf	oncology	society
animal	determination	infants	ondansetron	soft
ann	determine	infarction	one	sole
anne	determined	infected	ongoing	solid
annual	determining	infection	only	solution
annually	detroit	infectious	ontario	solvent
anorexia	develop	inflammation	oophorectomy	sorafenib
antagonist	developed	inflammatory	open	south
anthony	developing	influence	operable	southeastern
anthracycline	development	inform	operation	southern
anthracyclines	develops	information	operative	southwest
anti	device	informed	opinion	southwestern
antibacterial	devices	infusion	optimal	spain
antibiotic	dex	infusions	option	spanish
antibiotics	dexamethasone	ingram	optional	sparc
antibodies	dextran	inhaled	options	spartanburg
antibody	dextrose	inhibit	orally	speak
anticancer	dfs	inhibition	orange	special
anticipated	diabetes	inhibitor	order	specialists
anticipation	diabetic	inhibitors	oregon	species
anticoagulation	diagnosed	initial	organ	specific
antiemetic	diagnosis	initially	organic	specifically
antifungal	diagnostic	initiate	organization	specificity
antigen	dialysis	initiated	organs	specified

antigens	diameter	initiating	origin	specimen
antihypertensive	diameters	initiation	original	specimens
antineoplastic	diaphragm	injectable	orlando	spect
antiretroviral	diarrhea	injected	orleans	spectrometry
antitumor	diary	injection	orr	spectrum
antonio	diastolic	injections	orthodontic	speed
aortic	diathesis	injury	osaka	spermicide
apart	die	innovations	osf	spinal
apoptosis	died	inoperable	ospedale	spine
apparent	diego	inr	ospedaliera	spio
appearance	diet	inside	osteosarcoma	spiral
appears	dietary	insoluble	other	spl
appendix	difference	instead	others	spleen
appetite	differences	institut	otherwise	spokane
appliance	different	institute	ottawa	sponsor
appliances	differential	institutes	out	spot
applicable	differentiated	institution	outcome	spread
application	differentiation	institutional	outcomes	spring
applications	difficult	institutions	outline	springfield
applied	difficulty	instituto	outlined	springs
apply	diffusion	instructions	outpatient	sputum
approach	digital	instrument	outside	squamous
approaches	diluent	insufficiency	ovarian	stabilized
appropriate	diluted	insufficient	ovary	stable
approval	dimension	insulin	over	staff
approved	dimensions	int	overexpressing	stage
approx	dipstick	intake	overexpression	stages
approximately	direct	intend	overland	staging
apr	directed	intended	overload	staining
aprepitant	directly	intensification	overnight	standard
ara	director	intensity	oxaliplatin	standardized
arabinoside	disability	intensive	oxaloacetic	stanford
arbor	disappearance	intent	oxide	starpharma
area	discharge	inter	oxygen	start
areas	discontinuation	interaction	pacemaker	started
argentina	discontinued	interactions	pacific	starting
arizona	discovered	intercse	paclitaxel	states
arkansas	discretion	intercurrent	pain	statin
arm	discrimination	interest	painful	statistical
armed	discuss	interfere	paired	statistically
arms	disease	interferon	palliative	statistics

army	diseases	interim	palpable	status
around	disorder	interleukin	pancreas	stay
array	disorders	intermediate	pancreatic	stem
arrest	dispensary	intermittent	pancreatitis	stenosis
arrhythmia	dispersion	internal	panzem	stent
arrhythmias	dissection	international	papillary	stenting
arterial	distant	interpretation	par	stents
arteries	distribution	interruptions	paraffin	step
artery	divided	interstitial	parallel	stephen
artificial	dividing	interval	parameter	steps
asa	division	intervals	parameters	sterile
asan	dlt	intervention	paraplatin	sterilization
ascites	dlts	interventional	parent	sterilized
asheville	dmd	interventions	parenteral	steroid
asked	dna	intra	paris	steroids
asp	docetaxel	intracoronary	park	steven
asparaginase	doctor	intracranial	paromomycin	sti
asparagine	doctors	intramuscular	parsons	stimulating
aspartate	document	intraperitoneal	part	stimulation
aspects	documentation	intrathecal	partial	stoddard
aspirate	documented	intrauterine	participant	stop
aspirin	dodge	intravascular	participants	stopped
assay	doi	intravenous	participate	stopping
assess	donation	intravenously	participating	strasbg
assessed	donor	inv	participation	strategies
assessing	dorado	invasion	particle	strategy
assessment	dosage	invasive	particles	stratified
assessments	dose	investigate	particular	stress
assessor	dosed	investigated	partner	stroke
assigned	doses	investigation	parts	stroma
assignment	dosing	investigational	past	stromal
assistant	dotap	investigative	paste	strong
associate	double	investigator	patch	structure
associates	down	investigators	patency	studied
association	doxil	involved	patent	studies
assume	doxilcaelyx	involvement	pathologic	studying
assuming	doxorubicin	involves	pathological	surgeon
ast	drawn	involving	pathologically	sub
astalt	dresden	iowa	pathology	subcutaneous
asthma	dressing	ipsilateral	pathway	subcutaneously
asymptomatic	drug	irb	patient	subgroup

atheroma	drugs	ireland	patients	subject
atherosclerosis	dsc	irinotecan	patrick	subjects
ati	dtpa	iron	pattern	subscale
ation	duarte	irradiated	patterns	subsequent
atlanta	dublin	irradiation	paul	subsequently
atorvastatin	ductal	irrespective	pazopanib	subset
atrial	duke	irreversible	pca	substance
attack	duluth	irvine	pcr	substances
attacks	duration	ischemia	pdq	substituted
attempt	durham	ischemic	peak	subtype
attributed	dye	island	pectoris	subtypes
auc	dynamic	islet	pediatric	success
aug	dysfunction	islets	peg	successful
augusta	dyspnea	isolated	pegaspargase	successfully
aurora	earlier	israel	pegfilgrastim	sucrose
auroshell	early	istituto	pegylated	suffering
austin	east	italy	pekin	sufficient
australia	eastern	itraconazole	pelvic	sugar
austria	echo	itt	pelvis	suggest
authorization	echocardiogram	iud	pemetrexed	suggestive
authorized	echocardiography	ivpb	pennsylvania	suitable
autoimmune	ecog	ivus	people	sulfate
autologous	economic	ixabepilone	peoria	sum
autophagy	edema	jackson	peptic	summa
availability	edinburgh	jacksonville	percentage	summarized
available	edu	jacobs	perez	summary
avastin	edward	james	perforation	summit
average	effect	jan	perform	sun
avoid	effective	january	performance	sunitinib
axilla	effectiveness	japan	performed	superficial
axillary	effects	jay	perfusion	superior
azacitidine	efficient	jean	period	superparamagnetic
azar	effusion	jeffrey	periodically	supplement
azienda	effusions	jeopardize	periods	supplementation
back	efs	jersey	peripheral	supplements
background	egfr	jerusalem	peritoneal	support
bacterial	ejection	jessica	permanent	suppressive
balloon	ekg	jewish	permanente	surface
baltimore	elapsed	jnal	permeability	surgery
bangladesh	elderly	john	permit	surgical

baptist	electrocardiogram	johns	permitted	surgically
barbara	electronic	joint	persistence	surrogate
barcelona	elevated	joints	persistent	surrounding
barrier	elevation	jonathan	person	survival
basal	eligibility	joseph	personal	susan
base	eligible	josephs	persons	suspect
based	elimination	juan	perspective	suspected
basel	elisa	judged	perth	suspension
baseline	ellis	judgment	pertuzumab	suspicion
basic	eluting	juice	peru	suspicious
basis	embedded	jul	pet	suv
baton	embolism	jun	peter	swallow
bay	embolization	kaiser	petersburg	sweden
bayer	emend	kala	pfs	swedish
bcl	emergent	kalispell	pharma	switzerland
bcs	emission	kansas	pharmaceutical	sylvester
bct	emory	kantonsspital	pharmaceuticals	symptom
bearing	employ	kaplan	pharmacodynamic	symptomatic
become	emulsion	karmanos	pharmacodynamics	symptoms
becoming	enamel	karnofsky	pharmacokinetic	syndromes
beech	encapsulated	kayaku	pharmacokinetics	system
began	encephalopathy	kelly	pharmacological	systemic
begin	end	ketoconazole	phase	systems
beginning	endocrine	kettering	phases	systolic
begins	endometrial	kewanee	phd	table
beijing	endorem	key	phenobarbital	tablespoons
belgium	endoscopic	kgm ²	phenytoin	tablet
below	endothelial	kill	philadelphia	tablets
bend	endpoint	killing	phoenix	tace
beneficial	endpoints	kim	phone	tacoma
benefit	energy	kimmel	phosphatase	tafter
benefits	engl	kinase	phosphate	taipei
benign	enhance	kinetics	phosphopeptide	taiwan
benjamin	enhanced	kingdom	photographs	take
bern	enhancement	kingman	photothermal	taken
best	enhancing	kit	physician	takes
beta	enriched	klinik	physicians	taking
beth	enroll	klinikum	picn	tamoxifen
bethesda	enrolled	knight	pills	tampa
better	enrolling	known	pilot	target
bevacizumab	enrollment	knoxville	piper	targeted

beyond	ensure	kong	pittsburgh	targeting
bid	entered	korea	place	tat
big	entering	lab	placebo	tav
bilateral	entire	labc	placed	taxane
bilirubin	entry	label	placement	taxanes
billings	enzyme	labeled	plan	taxol
bind	enzymes	labelled	planned	taxotere
binding	eob	laboratory	planning	team
bio	eos	lack	plans	teaspoon
bioavailability	epeius	lactating	plaque	teaspoons
biochemical	epidermal	lactation	plaques	technical
bioequivalence	epirubicin	lake	plasma	technion
biologic	episode	laparoscopic	plasmonic	technique
biological	episodes	lapatinib	platelet	techniques
biomarker	epithelial	large	platelets	technology
biomarkers	epub	larger	platinum	teeth
biopharmaceuticals	equal	largest	pleural	temozolomide
biopsies	equivalence	las	pllc	temperature
biopsy	equivalent	laser	plus	temsirolimus
bioscience	erbb	last	point	ten
biotechnologies	erbitux	lasting	points	tennessee
biotics	eric	late	poland	teratogenic
birmingham	erlotinib	later	polyethylenimine	term
birth	erosion	lateral	polymer	terminal
bismarck	erythromycin	latvia	polymerase	terminated
bisphosphonate	escalating	lawrence	polymeric	termination
bisphosphonates	escalation	ldl	polymorphisms	terminology
bladder	escanaba	lead	poor	terms
blasts	eso	leads	poorly	tesla
bleeding	esophageal	learn	population	test
blind	essen	least	populations	tested
blinded	essential	lee	port	testicular
block	establish	left	portal	testing
blocking	established	legacy	portion	testosterone
blood	estimate	legal	portland	tests
bloodstream	estimated	legally	portugal	teva
bloomington	estimates	leishmaniasis	positive	texas
blue	estradiol	length	positivity	tfore
bmi	estrogen	leptomeningeal	positron	tgen
board	ethanol	lerable	possible	therapeutic
body	ethics	lerance	possibly	therapeutics

bolus	ethnic	lesion	post	therapies
bone	etiology	lesions	posterior	therapy
borderline	etoposide	less	postoperative	thermal
bortezomib	eureka	lethal	potassium	thermodox
boston	european	leucovorin	potent	thickness
bound	evaluabe	leukemia	potential	thinks
bowel	evaluate	leukemic	potentially	thioguanine
bozeman	evaluated	level	ppm	third
brain	evaluating	levels	pptt	thomas
branch	evaluation	lexington	practice	thoracic
brazil	evaluations	lhrh	practicing	threatening
breast	even	liberal	pratt	three
breastfeeding	event	life	pre	thrombocytopenia
breath	events	ligation	preceding	thromboembolic
brest	ever	light	preclinical	thromboplastin
british	everolimus	like	preclude	thrombosis
bronx	evidence	likely	precursor	thyroid
brown	evidenced	lima	predict	tianjin
bsa	evident	limit	predicted	time
buffalo	exam	limited	predicting	times
bulgaria	examination	limiting	prediction	tissue
burden	examinations	limits	predictive	tissues
burn	examine	linda	predictors	titer
burning	examined	line	prednisone	tmax
bypass	examiners	linear	predose	tnf
caassessed	example	lineberger	pregnancy	together
caelyx	exceed	lines	pregnant	toledo
caever	exceeding	lipid	preliminary	tolerability
calcium	except	liposomal	premedication	tolerable
calculated	exception	liposomally	preoperative	tolerance
calculation	exceptions	liposome	preparation	tolerate
california	excess	liposomes	preparations	tolerated
call	excessive	lisa	presbyterian	tomography
called	excipients	listed	prescribed	tool
camptothecin	excised	little	prescription	tooth
campus	exclude	liu	presence	toothpaste
canada	excluded	live	present	topical
canal	excluding	liver	presentation	topoisomerase
cancer	exhaled	lives	presented	toronto
cancercare	exhibit	living	presenting	total
cancerous	exist	llc	pressure	tower

cancers	existing	load	pretreatment	toxic
candidate	exists	loaded	prevalence	toxicities
candidates	expansion	loading	prevent	toxicity
cannon	expectancy	lobular	preventing	tracking
cannot	expected	local	prevention	tract
canton	experience	localization	previous	trademark
capable	experienced	localized	previously	trametinib
capacity	experiences	locally	primarily	transaminase
capecitabine	experiencing	located	primovist	transaminases
capillary	experimental	location	princess	transarterial
carbamazepine	explain	loma	princeton	transcatheter
carbon	exploratory	lombardi	principal	transfer
carboplatin	explore	london	principle	transferrin
carboplatinnab	exposed	long	prior	transformation
carcinoma	exposure	longer	probability	transformed
carcinomas	expressed	longest	probable	transfusion
cardiac	expressing	los	probably	transfusions
cardiology	expression	loss	probe	transient
cardiotoxicity	extended	lost	problem	transitional
cardiovascular	extension	louis	problems	translational
caregiver	extensive	louisville	procedure	transplant
caries	extent	low	procedures	transplantation
carious	external	lower	proceed	transplanted
carle	extra	lowering	process	transport
carlos	extracellular	lowest	processes	transporter
carmel	extract	ltd	produce	trastuzumab
carolina	extraction	lubbock	producing	traumatic
carotid	extremity	lukes	product	treat
carried	eye	lumen	products	treated
carry	ezetimibe	luminal	prof	treating
carthage	fact	lumpectomy	professor	treatment
case	factor	lung	profile	treatments
casein	factorial	lutheran	profiles	trial
cases	factors	lvef	profiling	tricolor
casi	failed	lymph	progesterone	triple
caspofungin	failure	lymphadenectomy	prognosis	true
castor	fairfax	lymphangiography	prognostic	trust
castrate	fairview	lymphatic	program	tsat
castro	fallopian	lymphoblastic	progress	ttp
category	false	lymphoid	progressed	tubal
catheter	familial	lymphoma	progressing	tube

causes	family	lymphomas	progression	tulsa
caution	farber	macomb	progressive	tum
cavity	fargo	macrophage	prohibited	tumor
cbc	fashion	macular	project	tumori
ccop	fast	mad	projected	tumors
ccr	fasting	madison	prolongation	tums
cdr	fat	madrid	prolonged	turkey
cedar	fatal	magnesium	promising	tween
cedex	fatty	magnetic	promyelocytic	twice
celator	favorable	main	prone	two
celgene	fda	maine	proof	tyler
cell	fdg	maintain	properties	type
cells	feasibility	maintained	prophylactic	types
cellsl	feasible	maintenance	prophylaxis	typical
cellsmm	features	major	propofol	typically
cellular	feb	majority	proportion	tyrosine
celsion	february	making	propose	uca
cement	fec	malabsorption	proposed	ucla
censored	fed	malate	prospective	ucsf
cent	federal	male	prostate	ugta
center	federation	males	prostatectomy	ukraine
centers	feeding	malignancies	prostatic	ulcer
centeruniversity	fekg	malignancy	protein	ulcerative
central	female	malignant	proteins	ulm
centro	females	mammary	proteinuria	uln
cerebral	femoral	mammogram	prothrombin	ultra
cerebrovascular	fenofibrate	mammography	protocol	ultrasound
certain	fenofibric	man	protocols	unable
cerulean	fer	management	proven	unacceptable
cervical	feraheme	mandatory	provided	unc
cervix	ferritin	maplewood	providence	unclear
cetuximab	ferrosoferric	mapping	provides	uncontrolled
cfu	fertile	mar	provisional	under
chain	ferumoxtran	margaret	proximal	undergo
chair	ferumoxytol	margin	psa	undergoing
chance	fetus	margins	psychiatric	undergone
chang	fever	maria	pts	underlying
change	few	marietta	ptt	understand
changed	fewer	mark	pty	understanding
changes	fiber	marked	published	undetactable
chanute	fibrinor	marker	puerto	unequivocal

chao	fibrillation	markers	pugh	unique
chapel	fibrosis	marketed	pulmonary	unit
characteristic	fibrous	marrow	pulp	united
characteristics	field	marseille	pulse	unitsm
characterize	fields	marshfield	pulsed	unity
characterized	filgrastim	martha	purposes	universitario
charles	filled	martin	qtc	universität
charleston	filling	maryland	qualify	universitätsklinikum
charlotte	filtration	marys	qualitative	university
chattanooga	final	masking	quality	unknown
check	find	mason	quantify	unl
chek	finding	mass	quantitative	unless
chemical	findings	mastectomy	quebec	unlikely
chemically	fine	material	questionnaire	unrelated
chemistry	finland	materials	questionnaires	unresectable
chemo	first	matrix	questions	unresolved
chemoembolization	fish	matthew	radiation	unsafe
chemoradiation	fistula	maturation	radical	unspecified
chemoradiotherapy	five	max	radiofrequency	unstable
chemotherapeutic	fixed	maximal	radiographic	until
chemotherapies	flint	maximum	radiographically	untreated
chemotherapy	florida	mbc	radiologic	unwilling
cheng	flow	mbp	radiological	unwillingness
chest	flowable	mcl	radiologist	upc
chewing	fludeoxyglucose	mdanderson	radiology	upper
chf	fluid	mids	radiotherapy	upstate
chicago	fluoracil	mean	raeb	uptake
chief	fluorescence	means	rambam	ural
child	fluoridated	measurable	random	urbana
childbearing	fluoride	measure	randomised	urinalysis
childhood	flynn	measured	randomization	urinary
children	foc	measurement	randomized	urine
childrens	focal	measurements	randomly	urothelial
china	folfirinox	measures	range	usa
choice	follow	measuring	ranges	useful
chol	followed	mechanical	ranging	users
cholesterol	following	mechanism	rapamune	uses
chorionic	follows	mechanisms	rapamycin	uspio
chosen	food	med	rapid	usual
chp	forest	media	rapidly	usually

christi	form	median	rapids	utah
christopher	formation	mediated	rate	uterine
chromosome	forms	medical	rates	utility
chronic	formula	medically	ratio	vaccination
chu	formulated	medication	rationale	vaccine
cimetidine	formulation	medications	ray	vaccines
cin	formulations	medicinal	rays	vaginal
cincinnati	fort	medicines	rcb	vaginally
circulating	fosaprepitant	medium	reaction	vaginosis
cirrhosis	fraction	medizinsche	reactions	valencia
cis	fractions	meet	reactive	validated
cisplatin	fracture	meeting	read	valley
citrate	frame	meets	real	value
city	française	megace	reason	values
ckd	france	megestrol	reasonable	valvular
clarithromycin	francis	meier	reasons	van
class	francisco	melanoma	receipt	vancouver
classification	frankfurt	melbne	receive	vanderbilt
classified	franklin	mellitus	received	vandetanib
claustrophobia	fred	member	receiving	vapor
clear	free	members	recent	variability
clearance	frequency	membrane	receptor	variables
clearly	frequent	memorial	receptors	variants
cleveland	fri	men	recist	variety
clin	front	menopausal	reco	varnish
clinic	fudan	menstrual	recombinant	vas
clinical	full	mental	recommend	vascular
clinically	fully	mercaptopurine	recommended	vcr
clinics	function	mercy	recorded	vector
clofarabine	functional	mesenchymal	recover	vegas
close	functioning	mesh	recovered	vegf
closed	functions	met	recruited	vehicle
cloud	fungus	metabolic	recruiting	vein
cmax	fus	metabolism	recruitment	velcade
cns	fusion	metal	rectal	venous
coagulation	future	metallic	recurrence	ventricular
coagulopathy	gadolinium	metastases	recurrent	vermont
coast	gain	metastasis	red	versus
coated	gainesville	metastatic	reduce	vessel
cockcroft	galesburg	meter	reduced	vessels
code	gamma	metformin	reducing	via

coenzyme	gastric	method	reduction	victoria
cognitive	gastro	methodist	reductions	view
cohort	gated	methods	ref	viii
cohorts	gault	methotrexate	reference	vincent
coil	gbg	methylation	referred	vincristine
coils	gdc	metro	referring	vinorelbine
collect	gdl	mexico	refractory	viral
collected	geisinger	mgday	regardless	virginia
collection	gel	mgdl	regimen	virus
college	gemcitabine	mgkg	regimens	visceral
colloidal	gemzar	mgkgday	regina	visible
colon	genasense	mgm	region	visit
colony	gender	mgm^	regional	visits
colorado	gene	mgm ²	regions	visual
colorectal	genentech	mgmday	registered	vital
columbus	general	mgmdose	registration	vitamin
combidex	generally	mgml	registry	vitro
combination	genesis	miami	regression	vivagel
combinations	genesys	micelle	regular	vivo
combined	genetic	micelles	relapse	vocs
combining	geneva	michael	relapsed	volatile
come	genexol	michigan	relapses	volume
commercial	genital	micro	related	voluntarily
commercially	genomic	microbicide	relation	volunteers
committee	genotype	microscopic	relationship	vomiting
common	geographical	microscopy	relative	von
commonly	george	mid	relatively	voriconazole
community	georgia	middletown	release	vorinostat
company	germ	midwest	releasing	vsp
comparable	german	migration	relevant	wade
comparative	germany	milano	reliable	wales
comparator	get	mild	remain	wall
compared	gets	milk	remaining	wang
comparing	gfr	miller	remains	want
complete	ghent	miltefosine	remission	warfarin
completed	gilbert	milwaukee	remitting	warren
completely	gilberts	min	removal	washington
completing	gingivitis	minimal	remove	washout
completion	giovanni	minimum	removed	water
complex	give	ministry	renal	wave
compliance	given	minneapolis	render	waves

complication	giving	minor	repair	way
complications	glaxosmithkline	minute	repeat	wayne
comply	gleason	minutes	repeated	ways
component	glioblastoma	missi	repeats	wbc
components	glioma	mississippi	replaced	wear
composite	global	missoula	replacement	website
composites	glomerular	mitomycin	report	week
composition	glucocorticoids	mitoxantrone	reports	weekly
compound	glucose	mixed	representative	weeks
compounds	glutamic	mll	reproductive	weight
comprehensive	glutamine	mlmin	republic	weighted
compression	glycol	mln	require	weill
comprising	gmbh	mm^	required	weiss
compromise	gmdl	mm ³	requirement	well
computed	goal	mmhg	requirements	wellington
con	gog	mmoll	requires	west
concentrate	gold	mobile	requiring	western
concentrated	gonadotropin	modalities	res	weston
concentration	good	model	resce	whether
concentrations	government	models	rescue	white
concept	grade	moderate	research	who
conclusion	graded	modified	researchers	whole
concomitant	grades	modulated	resectable	whom
concurrent	graft	moffitt	resected	whose
concurrently	grand	moines	resection	wichita
condition	grant	molar	reserve	wide
condom	granulocyte	molecular	residual	widely
condoms	granulocytes	molecule	resin	will
conduct	grapefruit	molecules	resins	william
conducted	great	mon	resistant	willing
conduction	greater	monday	resolution	willingness
confidence	greece	monitor	resolved	win
confirm	green	monitored	resonance	window
confirmation	greenville	monitoring	respect	winfield
confirmed	gregory	monmouth	respectively	winston
congenital	gross	monoclonal	respiratory	wisconsin
congestive	group	monotherapy	respond	withdrawal
conjugated	groups	monroe	responded	withdrawn
conjunction	grove	montana	responders	wks
conlin	grow	month	responding	woman
connecticut	growing	monthly	response	women

consecutive	growth	months	responses	womens
consent	gsk	montreal	rest	wood
conservation	guardian	morbidity	restenosis	woodbury
considered	guided	more	restoration	work
considering	guidelines	mortality	restorations	working
consist	guy	moscow	restorative	works
consistent	gynecologic	mother	restrictions	world
consisting	hackensack	motion	resulting	worse
consists	hadassah	motor	retain	worst
consolidation	hai	mount	retainer	wort
consortium	haick	mountain	retention	worth
constant	haifa	mousse	retroviral	wound
consultants	half	mouth	return	wounds
consumption	hampshire	mra	revascularization	written
contact	hand	mrd	reverse	wustl
contain	hands	mri	review	wyoming
containing	hangzhou	msec	reviewed	yale
contains	hard	mtc	rexin	year
continuation	havana	mtd	rfa	yes
continue	hawaii	mtor	rhode	york
continued	hawkins	mtx	rhtnf	you
continues	hcc	much	rhumab	younger
continuing	hcg	mucinous	rich	yrs
continuous	hcl	muga	richard	zhang
continuously	hcv	multi	richmond	zhejiang
contra	head	multicenter	rico	ziekenhuis
contraception	healed	multicentre	ridge	zubrod
contraceptive	healing	multiple	rifampin	

REFERENCIAS

Apte and Weiss, 1997 C. Apte and S. Weiss. Data Mining with Decision Trees and Decision Rules. *Future Generation Computer Systems*, 13:197-210, 1997

Brücher Heide, Gerhard Knolmayer, Marc-André Mittermayer; “Document Classification Methods for Organizing Explicit Knowledge”, Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland, 2002.

Buxton DB. Current status of nanotechnology approaches for cardiovascular disease: a personal perspective. *Wiley Interdiscip Rev Nanomed Nanobiotechnol.*;1(2):149-55. 2009

Chidanand Apte, Fred Damerau, Sholom M. Weiss; “Automated Learning of Decision Rules for Text Categorization”, *ACM Transactions on Information Systems (TOIS)*, Vol. 12, Issue 3, pp. 233 – 251, 1994.

Chiesa S, García-Remesal M, de la Calle G, de la Iglesia D, Bankauskaite V, Maojo V. Building an Index of Nanomedical Resources: An Automatic Approach Based on Text Mining. *Proc. of the KES2008*; pag 50-57, 2008

Cios, K. and Kacprzyk, J. (Eds.) *Medical Data Mining and Knowledge Discovery*, Springer Verlag, New York, USA. 2001.

De la Iglesia D, Chiesa S, Kern J, Maojo V, Martin-Sanchez F, Potamias G, Moustakis V, Mitchell JA. Nanoinformatics: new challenges for biomedical informatics at the nano level. *Stud Health Technol Inform*; pag 150:987-91, 2009.

De la Iglesia D, Maojo V, Chiesa S, Martin-Sanchez F, Kern J, Potamias G, Crespo J, Garcia-Remesal M, Keuchkerian S, Kulikowski C, Mitchell JA. International efforts in nanoinformatics research applied to nanomedicine *Methods Inf Med.* 50:84-95, 2011

De la Iglesia Jiménez, Diana de la; García Remesal, Miguel; Anguita Sanchez, Alberto; Muñoz Mármol, Miguel; Kulikowski, Casimir y Maojo García, Víctor Manuel. A machine learning approach to identify clinical trials involving nanodrugs and nanodevices from ClinicalTrials.gov. "Plos One", v. 9 (n. 10), 2013

De la Iglesia D, “Nanoinformatics Knowledge Infrastructures: Bringing Efficient Information Management to Nanomedical Research.” *Computational science & discovery* 6.1 (2013): 014011–. PMC. Web. 17 June 2015.

- Dung H. M. Nguyen, Jon D. Patrick: Research and applications: Supervised machine learning and active learning in classification of radiology reports. *JAMIA* 21(5): 893-901 (2014)
- Etheridge ML, Campbell SA, Erdman AG, Haynes CL, Wolf SM, McCullough J. The big picture on nanomedicine: the state of investigational and approved nanomedicine products. *Nanomedicine*. 9(1):1-14. 2013
- EU Framework Project 7. ACTION-Grid, <http://www.action-grid.eu>
- Fan R.E, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification *Journal of Machine Learning Research*, Vol. 9, 1871-1874., 2008
- Freitas RA. *Nanomedicine, Volume IIA: Biocompatibility*. Georgetown, Texas: Landes Bioscience; Available from: <http://www.nanomedicine.com/NMIIA.htm>., 2005
- García-Remesal M, García-Ruiz A, Pérez-Rey D, de la Iglesia D, Maojo V. Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature. *Biomed Res Int*. 2012
- Gayathri K., A. Marimutha, Multi-Class Text Classification with KNN Machine Learning Techniques, Vol 2, (6), pag 645-647, Febrero 2014
- Genkin, Alexander, David D. Lewis, and David Madigan. "Large-scale Bayesian logistic regression for text categorization." *Technometrics* 49, no. 3, 291-304, 2007.
- Grulke, C.M.; Goldsmith, M.R.; Vallerio, D.A. Toward a blended ontology: Applying knowledge systems to compare therapeutic and toxicological nanoscale domains. *J. Biomed. Biotechnol*. 2012.
- Horev-Azaria L., Baldi G., Beno D., et al. Predictive Toxicology of cobalt ferrite nanoparticles: comparative in-vitro study of different cellular models using methods of knowledge discovery from data. *Particle and Fibre Toxicology*.; 10(1, article 32) doi: 10.1186/1743-8977-10-32, 2013
- Karalis V, Macheras P. Current regulatory approaches of bioequivalence testing. *Expert Opin Drug Metab Toxicol*. (8):929-42. 2012
- Khan A., B. Baharudin, L. H. Lee, K. Khan, "A Review of Machine Learning Algorithms for Text Documents Classification", *Journal of Advances Information Technology*, vol. 1, 2010.
- Kleiner, R.E., Dumelin, C.E. & Liu, D.R. Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev*. 40, 5707–5717 (2011).

- Korde, V.& Mahender, C. Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications (IJAIA)*, Vol. 3, No. 2, pp. 85-99, 2012
- Li Y, Zheng X, Cao Z, Xu W, Zhang J, Gong M. Self-assembled peptide (CADY-1) improved the clinical application of doxorubicin. *Int J Pharm.*; 434(1-2):209-14, 2012
- Liu H., M Bilal, A Lazareva, A Keller, Y Cohen, *Beilstein journal of nanotechnology* 6 (1), 938-951, 2015
- Liu, R.; France, B.; George, S.; Rallo, R.; Zhang, H.; Xia, T.; Nel, A.E.; Bradley, K.; Cohen, Y. Association rule mining of cellular responses induced by metal and metal oxide nanoparticles. *Analyst*, 139, 943–953, 2014
- Liu, X., and Thomas J Webster. “Nanoinformatics for Biomedicine: Emerging Approaches and Applications.” *International Journal of Nanomedicine* 8.Suppl 1 (2013): 1–5. PMC. 2013
- Maojo V, Martin-Sanchez F, Kulikowski C, Rodriguez-Paton A, Fritts M. Nanoinformatics and DNA-based computing: catalyzing nanomedicine. *Pediatr Res.* 67(5):481-9, 2010
- Manonmani V., and Anitha U., Nanoinformatics: A Conjecture and a Review, *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, Vol. 6, (3), pag: 631-635, 2015
- McGonigle, D. Editorial: From Wearing Nanotechnology To Treating Patients To Nanoinformatics - We Are Just Beginning... *Online Journal of Nursing Informatics (OJNI)*, 13, (1). February, 2009
- Moosavi H, Zeynali M, Pour ZH. Fracture resistance of premolars restored by various types and placement techniques of resin composites. *Int J Dent.*;2012:973641. 2012
- National Nanotechnology Initiative. What is nanotechnology? 2007. Available: <http://www.nano.gov/nanotech-101/what/definition>.
- National Science Foundation, National Nanomanufacturing Network. Workshop on Nanoinformatics Strategies. Westin Gateway Hotel, Arlington Virginia. June 12-13, 2011.
- Ostraat, M.L.; Mills, K.C.; Guzan, K.A.; Murry, D. The Nanomaterial Registry: Facilitating the sharing and analysis of data in the diverse nanomaterial community. *Int. J. Nanomed.*, 8 (Suppl. 1), 7–13, 2013.

- Panneerselvam, Suresh, and Sangdun Choi. "Nanoinformatics: Emerging Databases and Available Tools." *International Journal of Molecular Sciences* 15.5 (2014): 7158–7182. PMC. Web. 17 June 2015.
- Pearce TR, Shroff K, Kokkoli E. Peptide targeted lipid nanoparticles for anticancer drug delivery. *Adv Mater.*;24(28):3803-22, 3710, 2012
- Petre CE, Dittmer DP. Liposomal daunorubicin as treatment for Kaposi's sarcoma. *Int J Nanomedicine.*; 2(3):277-88. 2007
- Que, H. -E. "Applications of fuzzy correlation on multiple document classification. Unpublished master thesis", Information Engineering department, Tamkang University, Taipei, Taiwan-2000.
- Quirke, V. and Gaudilliere, J.-P "The era of biomedicine: Science, Medicine and Health in Britain and France, ca 1945-65", in V. Quirke and J.-P. Gaudilliere (eds), special issue of *Medical History*, 52; 441-52, 2008
- Robert, A. F. J., "Current Status of Nanomedicine and Medical Nanorobotics", *Journal of Computational and Theoretical Nanoscience* 2: 1-25, 2005
- Rocchio, J; "Relevance Feedback in Information Retrieval", In G. Salton (ed.). *The SMART System*: pp.67-88, 1965
- Ruiz Miguel, Padmini Srinivasan; "Automatic Text Categorization Using Neural Network", In *Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research*, pp. 59-72. 1998.
- Sadiq Ahmed T. , Sura Mahmood Abdullah, "Hybrid Intelligent Techniques for Text Categorization." *International Journal of Advanced Computer Science and Information Technology(IJACSIT)* Vol. 2, No. 2, Page: 23-40, April 2013.
- Sharon, G.; George, W.H.; Stephanie, A.M.; Michal, L.; Mervi, H.; Juli, D.K. *Cananolab: Data sharing to expedite the use of nanotechnology in biomedicine. Comput. Sci. Discov.* 2013.
- Sriram, Bharath, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. "Short text classification in twitter to improve information filtering." In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 841-842. ACM, 2010.
- Tam, V., Santoso, A., & Setiono, R., "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization",

Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238, 2002.

Thomas, D.G.; Gaheen, S.; Harper, S.L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G.A.; Freund, E.T.; et al. ISA-TAB-Nano: A specification for sharing nanomaterial research data in spreadsheet-based format. *BMC Biotechnol.*, Vol. 13 (2), 2013.

Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features” ECML-98, 10th European Conference on Machine Learning, pp. 137-142. 1998.

<http://toxnet.nlm.nih.gov/>

Wang Xiaoping, Li-Ming Cao. Genetic Algorithm Theory, Application and Software [M]. XI'AN: Xi'an Jiaotong University Press, 2002

Webster, T.J., Nanomedicine: what's in a definition? *int J. Nanomedicine* 1(2) ,115-116, 2006

Witten, I.H. and Frank, E., *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, CA, 2000.

Witten, I.H., Bray, Z., Mahoui, M. and Teahan, W.J., “Text mining: a new frontier for lossless compression.” *Proc Data Compression Conference*, edited by J.A.S. Cohn, pp. 198-207, Snowbird, Utah, March. IEEE Press, Los Alamitos, CA, 1999

Weber DO. Nanomedicine. *Health Forum J.* 1999; 42(4):32, 36-7.

Yan, Su, Xiaoqian Jiang, and Ying Chen. “Text Mining Driven Drug-Drug Interaction Detection.” *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine (2013)*: 349–355. PMC. Web. 17 June 2015.

Yang Yiming, And Christopher G. Chute Mayo Clinic “An Example-Based Mapping Method For Text Categorization And Retrieval” *ACM Transactions On Information Systems*, Vol. 12, No 3, Pages 252-277, July 1994

Zhang, Harry. "The optimality of naive Bayes." *AA* 1, no. 2, 2004

Zhao L, Seth A. Wibowo N, Zhao CX, Mitter N, Yu CZ, "Nanoparticle vaccines", *Vaccine* 32 pag. 327-337, 2014