Universidad Nacional Autónoma de México

Facultad de Ciencias

PATENTES UTILIZADAS COMO INDICADORES DE INNOVACIÓN
TESIS

QUE PARA OBTENER EL TÍTULO DE:

FÍSICO

PRESENTA:

LUIS DAVID RODRÍGUEZ DÍAZ

TUTOR:

DR. ALEJANDRO PÉREZ RIASCOS

CIUDAD UNIVERSITARIA, CD. MX., ABRIL 2023

Resumen

En este trabajo de investigación se analizaron las patentes producidas a lo largo de la historia de las empresas de la industria tecnológica así como de la industria farmacéutica más influyentes del mundo en el periodo del año 1970 al 2019, partiendo de las 100 empresas más relevantes alrededor del mundo de cada una de estas industrias según la revista Forbes del año 2019, tomando las empresas con más de 50 patentes en total o en su defecto menos de una patente por año de existencia, llegando a tener 78 empresas de la industria tecnológica y 81 de la industria farmacéutica a partir de las cuales se realizará todo el estudio. Las patentes dan información acerca de las características de las empresas y son suministradas de manera abierta por Google Patentes. En la primera parte, se analiza la entropía de Shannon aplicada a los títulos de las patentes para mostrar si una mayor diversidad de palabras dentro de las patentes se encuentra acompañada de una mayor entropía, lo cual podría estar ligado a una mayor innovación a través del tiempo, así como la creación de redes y detección de comunidades como complemento del comportamiento de dichas industrias utilizando conceptos físicos y computacionales.

Posteriormente, se implementaron medidas de similitud entre la evolución del número de patentes y palabras. El conjunto de valores obtenido permite organizar en un arreglo matricial las distancias entre cada una de las empresas, para obtener una red equivalente que describe similitudes entre empresas a la que se le aplicó un algoritmo para detectar la formación de comunidades en una red que describe el comportamiento en cuanto a patentes para las empresas. Los resultados obtenidos mediante la detección de comunidades se interpretaron en tablas donde se identifican grupos con características similares. Los resultados muestran que las empresas con mayor diversidad de palabras y patentes dentro de la industria de la tecnología son Siemens, Amazon, Texas Instruments y Hp, mientras que para la industria farmacéutica son Abbot, Roche, Novartis y Bayer ligado a una mayor entropía lo que podría indicar que estas empresas posiblemente tengan una mayor innovación a lo largo de su historia. Las comunidades encontradas para la industria de la tecnología fueron 3, comunidades que esta compuestas de empresas que patentan

poco, pero innovan mucho, empresas que innovan poco pero patentan mucho y empresas de telecomunicaciones, mientras que para las empresas de farmacéutica igualmente son 3 comunidades, empresas que patentan poco pero innovan mucho, empresas que innovan poco pero patentan mucho y empresas que se dedican a la creación de medicamentos para trastornos neurológicos.

Índice general

Índice de figuras

Índice de tablas

Introducción

Los sistemas complejos se encuentran en diversas situaciones de la naturaleza, su estudio está presente en las ciencias naturales como lo son la física, la biología al estudiar el ADN, la medicina en donde se puede estudiar el comportamiento de una epidemia [1]; así como en sectores sociales, por ejemplo, en el estudio del crecimiento de una ciudad. Las diferentes escalas en las que encontramos sistemas complejos son muy amplias y pueden ir desde el interior de una célula hasta fenómenos de organización colectiva, cómo la innovación a nivel global, el avance de la innovación mediante el registro de patentes de ciertos productos [2], o el comportamiento de las personas en el transporte público.

Vivimos en un mundo en donde cada cosa se encuentra relacionada con otra. De esta manera, se forman estructuras de relaciones que pueden ser modeladas por medio de redes. Existen redes de diversos tipos desde las conocidas redes sociales, las redes de transporte, hasta las redes creadas por la interacción de entidades financieras y nuestras propias células. Estas estructuras están definidas por las entidades que las componen y sus relaciones, lo cual nos permite estudiarlas como un conjunto. De esta manera, nos aporta más información el estudio de grupos de entidades (comunidades) asociadas a fenómenos colectivos. Al estudiar estas entidades por separado, la información proporcionada será mucho menor.

Por otra parte, muchos fenómenos colectivos en sistemas complejos pueden ser descritos por medio de una red. En particular, el análisis de las redes y procesos dinámicos en estas estructuras nos permite estudiar la propagación de enfermedades, crisis financieras, identificación de puntos críticos y demás fenómenos [3].

En esta investigación se hace uso de distintas técnicas de la física estadística como la entropía de Shannon [4] así como teoría de redes [5], herramientas matemáticas y computacionales propias de la ciencia de datos aplicándolas al estudio del comportamiento de las empresas y el cómo podrían estar relacionadas con la innovación a través de la diversidad de patentes sin tomar en cuenta el rubro de los derechos de autor y partiendo de todas

las patentes independientemente de la nacionalidad de la empresa y las diferencias en cuanto a reglas de patentado en cada país [6]. Se estudia la relación entre la cantidad de información obtenida de patentes y la innovación en las empresas, tomando a dos de las industrias más importantes las cuales son la de tecnología y la farmacéutica del año 1970 al 2019.

La investigación desarrollada se reporta en capítulos de los cuales, el capítulo 1 contiene los conceptos principales para el entendimiento del contexto en el que se realiza la investigación y el procedimiento que se seguirá, dividido en dos principales conceptos, patentes e innovación, qué son y cómo son vistas en las dos principales industrias que abordaremos tecnología y farmacéutica, centrándonos en estos dos últimos como objetivo principal de la investigación. El capítulo 2 aborda con detenimiento cuales son las herramientas usadas a lo largo de la investigación tanto para el tratamiento de los datos como para la visualización de ellos, así como el proceso de obtención de los datos utilizados, la estructura de cada una de las bases de datos de las empresas que se utilizarán tanto las pertenecientes a la industria tecnológica como las de la industria farmacéutica, el tratamiento de cada una de ellas y la visualización de sus patentes a lo largo de la vida de las empresas. En el capítulo 3 se encuentra uno de los conceptos más importantes de la física estadística, la entropía, en este caso la entropía de Shannon también conocida como entropía de la información. Se presenta la definición de esta cantidad y como se encuentra relacionada con las patentes analizadas. En el capítulo 4, se implementan conceptos de teoría de redes, así como el proceso de detección de comunidades donde se observan grupos de empresas que guardan características entre sí, mientras que al final del texto se presentan las conclusiones así como algunas propuestas para trabajos posteriores.

Capítulo 1

Patentes e innovación

1.1. Introducción

A menudo se considera que el estudio de las patentes y la innovación es un campo que solo compete a quienes se forman en las ciencias sociales, tales como la sociología o la administración pública; la fama de los aportes que hacen la física o las matemáticas para ampliar nuestro entendimiento es menor. El capítulo inicial de este trabajo tiene como finalidad mostrar al lector lo que son las patentes y la innovación, en lo particular tomado como contexto para el análisis en posteriores capítulos y así poder saber cómo se encuentran relacionados estos dos conceptos. En la primera parte se muestra lo que son las patentes, así como su historia a lo largo del tiempo lo cual ayuda a tener un mayor entendimiento no solo de lo que es una patente, sino de todos los tipos de patentes que puede haber, así como los objetivos de una patente donde se destacan principalmente el incentivar la innovación. En la segunda parte, se presenta lo que es la innovación, así como distintas áreas en las que se presenta, sus principales tipos y el cómo la innovación implica patentar.

1.2. Patentes

Una patente es un derecho otorgado por el estado (u otra autoridad pertinente) a un inventor de modo que este sea el único autorizado para explotar y obtener ganancias de su invención durante un período de tiempo limitado, una patente es un derecho exclusivo

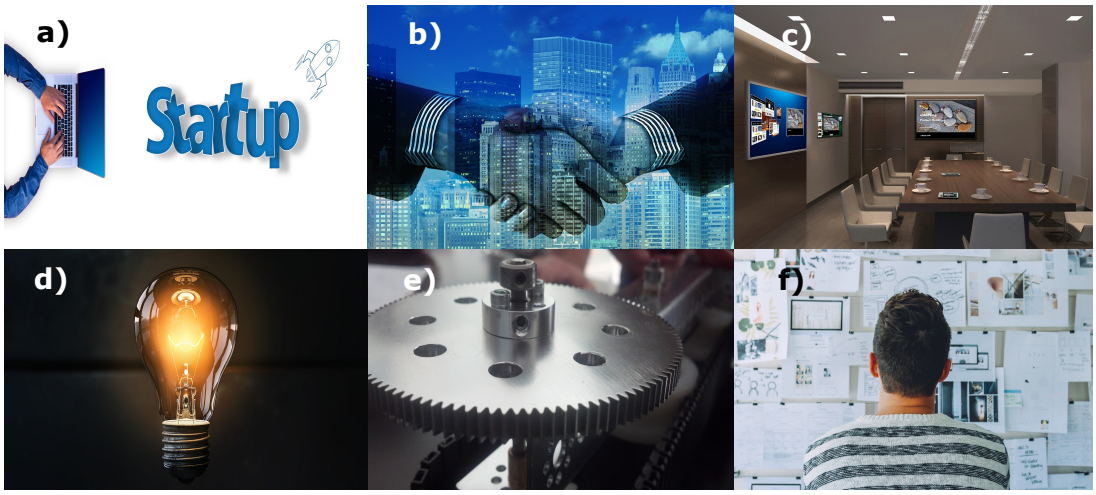
que se otorga al creador o inventor de un cierto producto o servicio. Este derecho le permite fabricar y comercializar su invento de manera exclusiva, pudiendo demandar a cualquier otra persona u organización que intente copiar su invento. La duración de la patente es limitada y a su vencimiento el inventor debe revelar los secretos de su invención (composición, estructura, etc.) de modo que todas las personas que estén interesadas puedan copiarlo y comercializarlo [7].

1.2.1. Historia de las patentes

El concepto de otorgar monopolios sobre las nuevas invenciones para fomentar la innovación y ofrecer una ventaja competitiva al inventor es algo relativamente reciente en términos históricos. Si realizamos un rastreo documental sobre la historia de las patentes para poder descubrir los orígenes de este asunto encontramos que la concesión de este tipo de privilegios se formalizó por primera vez en Italia mediante el Estatuto de Venecia de 1474 por el cual las nuevas invenciones una vez puestas en práctica tenían que ser comunicadas a la República para obtener protección jurídica contra los potenciales infractores, este privilegio se concedía por un período de 10 años. Se exigía que las invenciones debiesen de ser nuevas y útiles, cumplido esto se otorgaban los derechos, mientras que si se encontraban infractores a estos estatutos se les juzgaba y terminaban destruyendo sus dispositivos, gracias a esto se sentaron las bases de lo que hoy conocemos como patentes [8].

Los ingleses adoptaron este sistema por el que la corona concedía privilegios especiales a los empresarios de modo que solo ellos pudieran utilizar una invención importada, en 1623 se promulgó la Ley de Patentes de Invenciones (Statute of Monopolies) [9]. Por otra parte, la llegada de la revolución industrial a Inglaterra supuso un hito importante en la historia de las patentes, pues sirvió como catalizador en materia de patentes que aceleró la adopción legislativa en los diferentes países industrializados. Francia creó su primera Ley de patentes en 1791, y Alemania lo hizo en 1877. En 1883 los sistemas de patentes se internacionalizaron a través de la firma del Convenio de París [10].

En América, las primeras patentes fueron expedidas en 1641 por los gobiernos coloniales y las primeras leyes de patentes de los EE. UU. fueron establecidas por el Congreso en 1790, bajo la autoridad del Artículo 1 Sección 8 de la Constitución [11]. En México el Instituto Mexicano de la Propiedad Industrial (IMPI), es un Organismo público descentralizado encargado de la recepción, estudio y otorgamiento de registro de marcas y patentes. La primera patente en México se otorgó en el año 1843, durante la época en que México



A pesar de que

Figura 1.1: Tipos de patentes. Hay diversos tipos de patentes entre ellos: (a) patentes industriales utilizadas en los comercios o negocios que trabajan en la elaboración de distintos tipos de productos, (b) patentes comerciales utilizadas en los negocios de compra y venta de productos, (c) patentes profesionales son las que se le confieren a profesionistas a la hora de abrir sus oficinas por ejemplo los bufetes de abogados, (d) patentes de invención dadas a aquellas invenciones que logran cumplir con las normas legales, (e) patentes de utilidad son las dadas a la persona que inventa algo nuevo o mejorar una ya existente, (f) patentes de diseño logra inventar o descubrir una idea de diseño original o nueva para algún producto.

era todavía una república federalista. La patente fue otorgada por el presidente Antonio López de Santa Anna a un inventor mexicano llamado Juan Nepomuceno Almonte por su invención de una máquina para procesar la paja de maíz y convertirla en papel. Actualmente el costo de una patente varía dependiendo del país mientras que en México el costo de una patente según el IMPI va desde los $6000 hasta los $60000 dependiendo del tipo de patente, mientras que a nivel internacional suele costar hasta 4000 dólares, lo cual suele ser un valor menor dado los beneficios de una patente [12].

De hecho, las patentes únicamente brindan protección en el país en el que se otorgan, por lo que los inventores deben solicitar patentes en todos los países en los que deseen proteger su invención. La protección de patentes se rige por las leyes nacionales de propiedad intelectual de cada país, y las normas pueden variar de un país a otro. Sin embargo, algunos países no tienen leyes de propiedad intelectual y no cumplen completamente con los estándares internacionales de protección de patentes.

1.2.2. Características y tipos de patentes

Las patentes se pueden distinguir en función del objeto patentable o según su ámbito de aplicación, dentro de ellas existen diversos tipos que son [13]:

- Patentes comerciales: Son aquellas patentes utilizadas en los negocios y tiendas de compra y venta de productos.
- Patentes profesionales: Son aquellas que se les confieren a todos los profesionales cuando deciden abrir sus oficinas en una zona determinada de la comunidad. Por ejemplo, las oficinas para bufetes de abogados, locales para ingenieros, médicos, odontólogos, etc.
- Patentes industriales: Son utilizadas en los comercios o negocios que trabajan con la elaboración de diferentes tipos de productos y alimentos manufacturados.
- Patentes de utilidad: Se le otorga a la persona que inventa o descubre algo nuevo, una máquina nueva, mejora algún producto que ya existe, inicia un nuevo proceso de algo, etc.
- Patentes de diseño: Logra inventar o descubrir una idea de diseño original o nuevo para algún producto que se va a elaborar.
- Patentes de invención: Es la que se les permite a aquellas invenciones que logran cumplir con las normas legales que establecen las leyes de patente.
- Patentes internacionales: Estas patentes tienen validez en más de 140 países, dependen de factores como la inversión directa extranjera, del tamaño de los mercados de determinado país, las diferentes corrientes de tipo comercial, entre otras.

Estos tipos podemos observar a su vez en la figura 1.1.

Las patentes tratan acerca de un derecho exclusivo, tienen duración limitada es decir su duración depende del producto o servicio patentado. Usualmente no superan los 20 años, a su vez estas las otorga el Estado, generalmente a través de un organismo público especialmente dedicado al registro de patentes, marcas y protección de la propiedad intelectual e industrial. La patente genera un monopolio que favorece al inventor o creador del producto o servicio mientras que la copia o explotación de un producto o servicio sin contar con su patente es ilegal y puede dar lugar a procedimientos sancionatorios (multas, cárcel, etc.). Esto no sólo se aplica a productos o procesos complejos, sino que también a inventos sencillos o ideas originales (por ejemplo, un sujetapapeles, filtro, etc.) [14].

Figura 1.2: Principales áreas de innovación. La innovación puede darse en distintas áreas, las más destacadas a la hora de innovar son: (a) innovar en procesos y productos se refiere a la mejora de procesos o productos ya existentes, (b) innovar en materia de organización responde al hecho de mejorar la organización o planificación de una empresa no necesariamente un producto, (c) innovar en el área comercial a la hora de poner nuevos productos en el mercado con mejores opciones de supervivencia, (d) innovar en tecnología a través del mejoramiento de técnicas para la creación de productos.

1.2.3. Objetivos de una patente

El objetivo principal de la patente es incentivar la creación e innovación permitiendo que el inventor, que ha gastado energía y recursos en crear algo nuevo, pueda obtener una retribución. Cuando no existen patentes, los incentivos para invertir en desarrollar nuevos productos, servicios o tecnologías se reducen ya que inmediatamente otros podrán copiarlos, apropiándose de gran parte de los beneficios. En otras palabras, si el inventor no puede obtener ganancias de sus creaciones, no le será rentable invertir en crear algo nuevo lo que terminará perjudicando al conjunto de la sociedad [15].

Si bien la patente genera un monopolio, este será de duración limitada ya que cuando finalice el período de protección, el inventor compartirá su creación permitiendo que otras empresas compitan con él. La mayor oferta y competencia a su vez permitirán que más personas puedan acceder a los bienes que se encontraban patentados [16].

1.3. Innovación

1.3.1. ¿Qué es la innovación?

La innovación es un proceso que modifica elementos, ideas o protocolos ya existentes, mejorándolos o creando nuevos que impacten de manera favorable, es un concepto muy ligado al ámbito empresarial. Innovar es mejorar lo que existe, aportando nuevas opciones que suplan las necesidades de los consumidores, o incluso crear nuevos productos con el fin de que tengan éxito, a través del conocimiento de los productos a lo largo del tiempo una empresa puede tener una idea de cómo mejorar sus productos y de esta manera tener un mayor impacto en el mercado esto es un ejemplo de lo que es la innovación [17].

1.3.2. Innovación en distintas áreas

La innovación puede darse en diferentes áreas: sociales, empresariales, de organización, tecnológicas, entre otras, las más destacadas a la hora de innovar son el innovar en el área de procesos y productos. En este caso la innovación se dirige a mejorar los productos existentes, y permitir que el área de procesos sea igualmente innovadora para obtener los resultados deseados. Por ejemplo: fabricar productos con envoltorios que mejoren su durabilidad. En materia de organización, la innovación no solo puede responder al hecho de mejorar o crear un producto que revolucione el mercado, sino que también se puede aplicar a la organización de la propia empresa [18].

Por otra parte, también se puede innovar en el área comercial. En este caso, un elemento indispensable es poder introducir al mercado productos que tengan éxito y supongan la supervivencia de las marcas. En cuestiones comerciales se puede trabajar la innovación en el packaging y diseños de productos de tal forma que causen un impacto destacado y positivo en los consumidores, en el modo de colocarlos en los puntos de venta, donde la creatividad y el estudio de la acción de los consumidores tiene mucho que ver para ayudar e innovar, o a la hora de llevar a cabo promociones innovadoras que se salgan de lo cotidiano y que llamen la atención de los posibles consumidores. Por ejemplo: crear escaparates llamativos, combinar colores, aromas, iluminación para captar la atención del público. Sí se puede innovar en el aspecto social, comercial, de organización, también se puede mencionar el hecho de hacerlo en el aspecto tecnológico a través de la utilización de técnicas de fabricación de producto, de maquinaria o herramientas que aporten valor al producto y se obtengan novedosos resultados. Por ejemplo: utilizar la inteligencia artificial

para desarrollar y mejorar productos ya existentes [19].

1.3.3. Innovaciones revolucionarias

En términos de innovación existen muchas variables que intervienen en el área de la innovación, dentro de los cuales destacan las innovaciones revolucionarias. Se trata de tecnologías que transforman la sociedad y la actividad comercial, alteran las prácticas establecidas y pueden generar nuevas industrias. Algunos ejemplos de este tipo de innovación son: el motor de combustión interna, los antibióticos y, más recientemente, el teléfono móvil, en cambio, el segundo tipo, denominado innovación incremental, abarca mejoras secundarias a la tecnología existente. Dichas innovaciones no generan grandes adelantos, sino pequeños avances. Si bien en ocasiones se considera que las innovaciones incrementales son irrelevantes, en realidad, la mayoría de las innovaciones son de ese tipo y la acumulación de avances graduales puede generar cambios importantes mientras que el tercer tipo, la innovación frugal, describe un enfoque de innovación que consiste en crear un producto de mayor valor social mediante la utilización de escasos recursos. Este tipo de innovación suele producirse en entornos donde los recursos son limitados, para satisfacer las necesidades de comunidades de bajos y medianos ingresos [20].

Capítulo 2

Bases de datos de patentes

2.1. Introducción

En este capítulo se describen las distintas herramientas implementadas en la investigación, donde destaca el lenguaje de programación Python y sus correspondientes librerías para manipulación de datos como lo son Pandas, NLTK, Re, Numpy entre otras más especializadas para la visualización de los datos como lo es Matplotlib. Estas técnicas fueron utilizadas tanto para la extracción y manipulación de los datos. También se describe el proceso de la descarga de las bases de datos o datasets de Google Patentes que contienen las patentes y características de las 100 empresas de tecnología y farmacéutica con mayor influencia en la actualidad, al ser las que cuentan con una mayor relevancia en el mundo. También se abordará la estructura de cada una de las bases de datos, sus campos y características seguido del cómo se realizó el tratamiento y la limpieza de estos datos implementado con el propósito de facilitar la manipulación y el posterior análisis de resultados. En el apéndice 4.4 se presenta una tabla con la nación en la que cada una de las empresas analizadas se encuentra registrada.

2.2. Software utilizado para el análisis

A continuación, se describen las herramientas que fueron utilizadas para el tratamiento de los datos así como para el análisis y las visualizaciones de las patentes. **Python:** Python es un lenguaje de programación desarrollado en 1991 por Guido van Rossum en

el Centro para las Matemáticas y la Informática. Python el cual es un lenguaje de programación interpretado, multiparadigma y multiplataforma. En este contexto, interpretado significa que Python "interpreta" el código del programador, es decir, lo traduce y lo ejecuta a la vez. Por otra parte, se dice que es multiparadigma porque es un lenguaje de programación que admite el uso de varios paradigmas de programación (modelos de desarrollo), por lo que no exige a los programadores un estilo único para programar, los paradigmas de programación que permite Python son programación orientada a objetos, programación imperativa y programación funcional. También es multiplataforma porque el lenguaje Python puede ejecutarse en diferentes sistemas operativos como Unix, Linux, macOS y Windows [21].

Se utilizó Python a través del software de Jupyter el cual es una aplicación web que aporta una interfaz para el desarrollo de diferentes tipos de códigos de programación (su nombre es resultado de la fusión de los lenguajes de programación Julia, Python y R) y permite su interpretación, así como la visualización de gráficas, mapas, entre otros. Haciendo uso de algunas librerías especializadas de Python, una librería o biblioteca es el conjunto de implementos funcionales que te ayudan a codificar lenguajes de programación para crear una interfaz independiente. Las librerías tienen la libertad de ser utilizadas por otros programas independientes y simultáneamente [22].

Matplotlib: Matplotlib es una librería de Python especializada en la creación de gráficos. Permite crear y personalizar los tipos de gráficos más comunes, entre ellos: Diagramas de barras, histogramas, diagramas de sectores, diagramas de líneas, diagramas de áreas, mapas de calor [23].

Re: Esta librería de Python proporciona operaciones de coincidencia de expresiones regulares, estas expresiones regulares son patrones de coincidencia de texto descritos con una sintaxis formal. Los patrones se interpretan como un conjunto de instrucciones, que luego se ejecutan con una cadena como entrada para producir un subconjunto de coincidencia o una versión modificada del original. El término "expresiones regulares" con frecuencia se acorta en conversación a "regex" o "regexp". Las expresiones pueden incluir correspondencia de texto literal, repetición, composición de patrones, ramificación y otras reglas sofisticadas. Una gran cantidad de problemas de análisis son más fáciles de resolver con una expresión regular que mediante la creación de un analizador léxico de propósito especial y un analizador sintáctico [24].

Numpy: Es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos, incorpora una clase de objetos llamados arrays que permite representar colecciones de datos de un mismo tipo en varias

dimensiones, y funciones muy eficientes para su manipulación, un array es una estructura de datos de un mismo tipo organizada en forma de tabla o cuadrícula de distintas dimensiones.[25]

Datetime: Para manejar fechas en Python se suele utilizar la librería datetime que incorpora los tipos de datos date, time y datetime para representar fechas y funciones para manejarlas. Algunas de las operaciones más habituales que permite son acceder a los distintos componentes de una fecha (año, mes, día, hora, minutos, segundos y microsegundos), convertir cadenas con formato de fecha en los tipos date, time o datetime, convertir fechas de los tipos date, time o datetime en cadenas formateadas de acuerdo a diferentes formatos de fechas y hacer aritmética de fechas (sumar o restar fechas) [26].

NLTK: El kit de herramientas de lenguaje natural, o más comúnmente llamado NLTK es un módulo de Python que contiene muchas funciones diseñadas para su uso en el análisis lingüístico de documentos y en el procesamiento de lenguaje natural. Este tipo de herramientas son la base de nuestra investigación al momento de hacer el análisis de las patentes de las empresas [27].

Pandas: Es una librería especializada en el manejo y análisis de estructuras de datos. La principal característica de esta librería es: el análisis de datos toma datos (CVS, TSV, SQL, etc.) y los convierte en un objeto de Python con columnas y filas llamadas DataFrames [28].

2.3. Obtención de los datos

Para el proceso de la obtención de datos se recurrió a la búsqueda de bases de datos en Google Patentes, el cual es un servicio de búsqueda de patentes y solicitudes de patentes que ofrece Google, fue creado en el 2006 y desde entonces ha acumulado 120 millones de patentes pertenecientes a 100 oficinas de patentes en todo el mundo [29]. Ofrece además la posibilidad de descargarlas y guardarlas en distintos formatos entre ellos CSV. Se pueden encontrar traducciones de millones de patentes y solicitudes registradas, también es posible encontrar documentos técnicos organizados de forma automática los cuales podemos filtrar a través de la misma plataforma.

Para la selección de las empresas a analizar se tomó como referencia las 100 empresas más influyentes de tecnología y las 100 más influyentes en la industria farmacéutica listadas en el último ranking de la revista Forbes (mayo, 2020), la cual es una revista especializada en el mundo de los negocios y las finanzas [30, 31]. Utilizando Google Patentes se obtuvieron

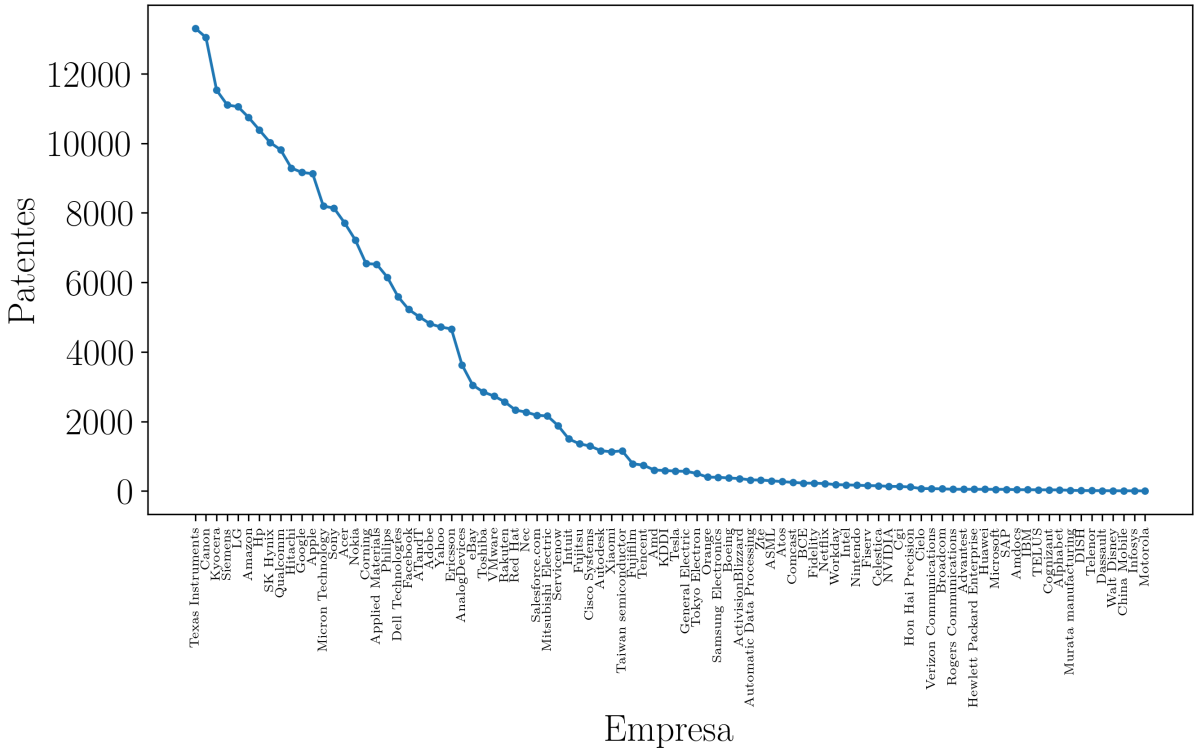


Figura 2.1: Patentes de empresas de tecnología. Se muestra el número total de patentes por empresa de las mejores empresas de tecnología donde se observa que las empresas de tecnología que cuentan con un mayor número de patentes son Texas Instruments y Canon contando con más de 12000 patentes cada una de ellas, es posible observar a su vez que alrededor de la mitad de las empresas cuentan con un número de patentes menor a 1000 donde destacan grandes empresas como Intel, Nintendo e IBM.

las bases de datos de patentes para cada una de las empresas, estas fueron descargadas en idioma inglés y con formato CSV para que puedan ser analizadas en Python tomando como intervalo de tiempo del año 1970 al año 2019 ya que son los años con los que actualmente se encuentran los datos de las patentes. Finalmente, se descartaron las empresas que tuvieran menos de 50 patentes, o en su defecto menos de una patente por año de existencia llegando a tener 78 empresas de la industria tecnológica y 81 de la industria farmacéutica.

Para cada una de las empresas se observa que su respectiva base de datos que a su vez está compuesta por Título de la patente, Cesionario, Inventor o autor, Fecha de creación, Fecha de publicación, Fecha de prioridad, Fecha de concesión.

Para el análisis de las patentes solamente requerimos de dos de estos campos, Titulo de la patente y Fecha de concesión por lo que los demás serán omitidos, cada base de datos cuenta con un número de patentes el cual se muestra en la tabla 2.1. Análogamente para las empresas farmacéuticas se tiene un tabla que muestra las empresas y su número de patentes, tabla 2.2. Se obtiene a su vez de las tablas 2.1 y 2.2 el comportamiento

Tabla 2.1: Empresas de tecnología. Se muestran las empresas de tecnología con las que se trabajará, así como su número de patentes a lo largo de su historia.

Empresas de Tecnología					
Empresa	Patentes	Empresa	Patentes	Empresa	Patentes
Acer	7708	eBay	3046	Nokia	7219
Activision B.	360	Ericsson	4662	NVIDIA	133
Adobe	4809	Facebook	5221	Orange	399
Advantest	51	Fidelity	226	Philips	6147
Amazon	10743	Fiserv	156	Qualcomm	9812
Amd	605	Fujifilm	784	Rakuten	2564
Amdocs	59	Fujitsu	1357	Red Hat	2330
Analog D.	3626	General E.	567	Rogers C	52
Apple	9126	Google	9162	Salesforce.com	2179
Applied M.	6522	Hewlett P. E.	50	Samsung E.	396
ASML	295	Hitachi	9289	Servicenow	1181
Atos	271	Hon Hai P.	116	Siemens	11101
AT&T	5009	Hp	10385	SK Hynix	10025
Autodesk	1160	Huawei	50	Sony	8136
Automatic D. P.	321	IBM	58	Taiwan S.	1156
BCE	228	Intel	177	Tencent	744
Boeing	375	Intuit	1499	Tesla	574
Broadcom	63	KDDI	591	Texas I.	13310
Canon	13044	Kyocera	11527	Tokyo E.	506
Celestica	153	LG	11053	Toshiba	2846
Cgi	132	Micron T.	8195	Verizon C.	70
Cielo	72	Microsoft	59	VMware	2732
Cisco S.	1295	Mitsubishi E.	2159	Workday	184
Comcast	249	Nec	2269	Xiaomi	1130
Corning	6541	Netflix	216	Yahoo	4117
Dell T.	5590	Nintendo	168	Zte	316

del número de patentes correspondientes a cada una de las empresas tanto de tecnología como de farmacéutica en donde se observa acerca de las empresas de tecnología tienen un máximo de patentes por empresa de 13310, más de la mitad de empresas cuentan con menos de 2000 patentes a lo largo de su historia como podemos observar en la figura 2.1 mientras que las empresas de farmacéutica tienen un máximo de 25000 patentes dentro de las cuales más de mitad cuentan con menos de 5000 patentes como podemos observar en la figura 2.2.

Tabla 2.2: Empresas de farmacéutica. Se muestran las empresas de Farmacéutica con las que se trabajara, así como su número de patentes a lo largo de su historia.

Empresas de Farmacéutica					
Empresa	Patentes	Empresa	Patentes	Empresa	Patentes
Abbott	14617	GlaxoSmithKline	5427	Octapharma	133
AbbVie	2310	Grifols	377	Ono P.	1534
Alexion	429	Grunenthal	678	Otsuka	11045
Allergan	3447	Hikma P.	50	Perrigo	269
Amgen	2950	Hisamitsu P.	1326	Pfizer	14045
Angelini	560	Humanwell H.	124	Pierre fabre	1156
Astellas P.	1392	Incyte	1883	Purdue Pharma	460
AstraZeneca	6757	Intas P.	63	Recordati	295
Aurobindo P.	171	IPsen	1151	Regeneron	999
Bausch H.	200	Jazz P.	62	Richter gedeon	1857
Baxter I.	5338	Jiangsu H.	779	Roche	20616
Bayer	24842	J&J	7166	Sanofi	11794
Biogen	1422	Kowa	5262	Santen P.	602
Biomarin P.	169	Krka	419	Sawai P.	125
Boehringer I.	8069	Kyowa H.	684	Servier	907
Bracco	936	LEO Pharma	501	Shanghai P.	9644
Celgene	796	Livzon P.	333	Sichuan K.	1743
Cipla	579	Lundbeck	1212	Stada A.	80
CSL	933	Lupin	535	Sumitomo D.	1024
CSPC	72	Menarini	408	Sun P.	508
Daiichi S.	1354	Merc&Co	85	Taisho P.	1374
Dr. Reddys	98	Merck	25000	Takeda	19332
Eisai	3901	Merz Pharma	348	Teijin	24999
Eli Lilly	8683	Mitsubishi T. P.	631	Teva	2347
Ferring	503	Mylan	531	UCB	152
Fresenius SE & Co	573	Novartis	21521	Vertex	1882
Gilead	1626	Novo Nordisk	5123	Yuhan	1489

2.4. Tratamiento de los datos

Para el tratamiento de los datos se comenzó por la hacer la configuración del ambiente de trabajo Jupyter así como la instalación de Python y la librerías que serán utilizadas, a partir de las tablas 2.1 y 2.2 creadas con el nombre de las empresas así como la cantidad de patentes que han tenido a lo largo de su historia, se optó por escoger a las empresas con más de 30 patentes ya que nos aportan una cantidad de información mayor.

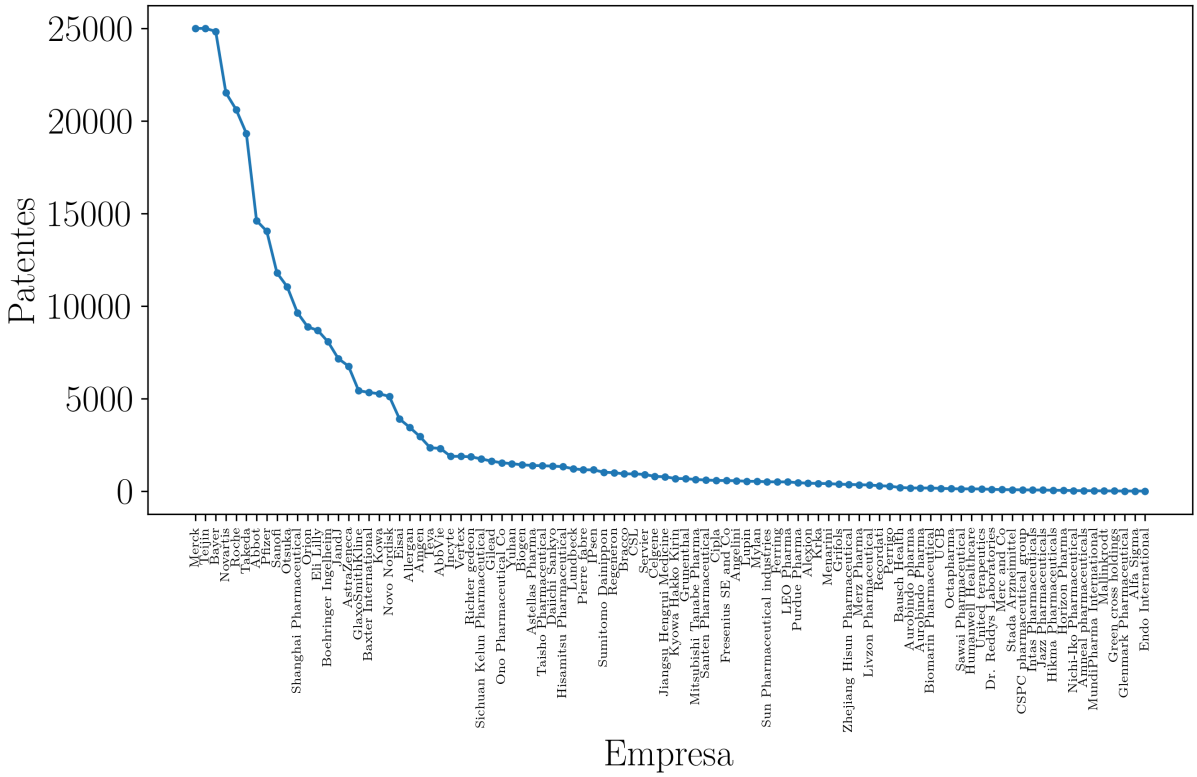


Figura 2.2: Patentes de empresas farmacéuticas. Se muestra el número total de patentes por empresa de las mejores empresas de la industria farmacéutica donde se observa que las empresas con más patentes son Merck, Teijin, y Bayer con más de 24000 patentes cada una, a su vez se observa que cerca de tres cuartas partes de las empresas tienen menos de 2500 patentes donde destacan empresas como Incyte y Biogen.

Con la ayuda de la librería Matplotlib se realizó a su vez una gráfica con la cantidad de patentes por empresa dentro de las cuales es posible observar el comportamiento de estas dos industrias, ya que observando la figura 2.1 es posible percatarse de que en alrededor del 50 % de las empresas es en donde se encuentra la mayor concentración de patentes, mientras que para la industria farmacéutica observada en la figura 2.2 la mayor concentración de patentes se encuentra en alrededor del 25 % de las empresas. Para iniciar el tratamiento de los datos se comenzó por importar cada una de las librerías que se utilizarán y en el cual para cada una se cargaron las bases de datos en formato csv creando un DataFrame haciendo uso de Python y de la librería Pandas, para el análisis se utilizaron dos campos en específico de las bases de datos la fecha de prioridad y el título de la patente.

Con ayuda de la librería NLTK cargamos la lista de stopwords (palabras sin significado como artículos, pronombres, preposiciones, etc.) con la que cuenta, estas son palabras que no aportan ningún tipo de información, a la cual se le agrego también símbolos y números

ya que tampoco nos aportan información de utilidad para el análisis.

Los títulos de las patentes se ordenaron de acuerdo con la fecha de concesión y con ayuda de la lista de stopwords creada anteriormente filtramos solamente las palabras que nos aportarán una mayor información acerca de la patente a la que hace referencia con esto se termina el filtrado de las palabras teniendo como resultado un DataFrame de la fecha de concesión y las palabras clave correspondientes a cada una de 5las patentes.

Para las mejores empresas en la industria farmacéutica se sigue el mismo proceso pero en este caso antes de hacer el filtrado con la lista de las stopwords creada para las empresas de tecnología, igualmente con la ayuda de la librería NLTK se creó una lista especial de simbología química (la simbología química es un conjunto de símbolos y convenciones utilizados para representar los elementos químicos y las moléculas en la química), la cual sirve para hacer un primer filtro a los títulos de las patentes, posteriormente se hace el filtrado análogamente como a las empresas de tecnología para culminar con un DataFrame de la fecha de concesión y las palabras decir las palabras que nos otorgan la información acerca de la patente.

Gracias a los procesos de filtrado y limpieza de datos realizados anteriormente para cada una de las empresas de tecnología y farmacéutica fue posible obtener los DataFrames correspondientes a cada una de ellas los cuales serán el punto de partida para el posterior análisis que se abordará los siguientes capítulos. En particular, en el capítulo 3 se utilizará la entropía de Shannon como medida de diversidad, partiendo de los títulos de las patentes se calculará esta entropía y mediante el estudio de su comportamiento será posible observar que a un mayor crecimiento de la entropía hay una mayor diversidad de palabras dentro de las patentes de cada una de las empresas, lo cual podría estar ligado a una mayor innovación. Por otra parte, en el capítulo 4 se generarán redes con las cuales se detectarán comunidades de empresas con las mismas características.

Capítulo 3

Evolución de la innovación en empresas

3.1. Introducción

En este capítulo se presenta el análisis del comportamiento de las patentes, así como de las palabras que comprenden dichas patentes a través del tiempo. Se exploran las diferencias del comportamiento que existen entre estas dos, también el concepto de la entropía, más específicamente la entropía de Shannon. Se abordará su definición y su aplicación dentro de esta investigación para el análisis de las patentes de las empresas de farmacéutica y tecnología. Los resultados muestran que las empresas que posiblemente tienen mayor innovación, es decir que a una mayor diversidad de palabras encontradas, presentan una mayor entropía. Al relacionar la entropía con el número de patentes se puede dar evidencia de que a mayor número de patentes podría existir mayor innovación. Finalmente se da a conocer las empresas con una posible mayor innovación y sus características.

3.2. Palabras y patentes acumuladas en el tiempo

En esta sección se explora el comportamiento de las palabras y patentes acumuladas a lo largo del tiempo para cada una de las empresas descritas en el capítulo 2 normalizadas de acuerdo al número de patentes. Donde las patentes acumuladas normalizadas se encuentran definidas como la cantidad de patentes a un tiempo T donde cada T equivale a un año

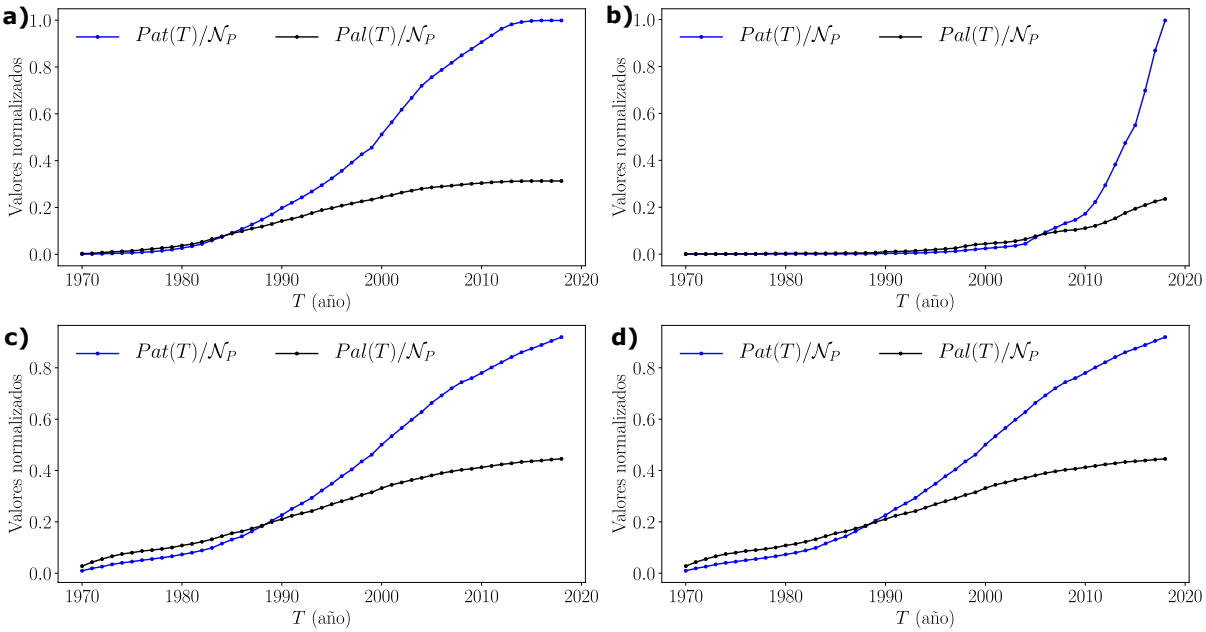


Figura 3.1: Patentes y palabras acumuladas de empresas de tecnología. Se muestra las gráficas correspondientes al número de patentes $Pat(T)$ y palabras acumuladas $Pal(T)$ a través del tiempo T en años desde 1970 hasta el 2019 para las 4 empresas de tecnología con más patentes a lo largo de su historia las cuales son a) Cannon, b) Kyocera, c) Texas Instruments, y d) Siemens. En cada caso, los valores se encuentran normalizados dividiendo entre el número total de patentes $\mathcal{N}_p$.

dividido por la cantidad total de patentes N_p de la empresa $Pat(T)/N_p$, mientras que las palabras acumuladas normalizadas se encuentran definidas como la cantidad de palabras a un tiempo T equivalente a un año dividido por la cantidad de patentes $Pal(T)/N_p$. De esta manera, en la primera parte se crea una lista con las patentes y una lista con las palabras que nos dan más información acerca de la patente de cada una de las empresas las cuales se encuentran agrupadas por año, este valor se normaliza con el número total de patentes.

En la figura 3.1 se presentan los resultados obtenidos para 4 empresas de tecnología con más patentes a lo largo de su historia. Las diferentes curvas muestran el comportamiento de estas, así como el comportamiento de sus palabras, estas 4 empresas son Texas Instruments, Canon, Kyocera y Siemens. Para Canon "a)" tienen un crecimiento con una intensidad media hasta el año 2000, mientras que del año 2000 al año 2010 tienen un gran crecimiento, posterior a este año se mantienen constantes en sus patentes, por otro lado las palabras siguen el mismo ritmo de crecimiento en el tiempo, mientras que Kyocera "b)" inició patentando muy poco y alrededor del año 2004 comenzó a tener grandes cantidades de patentes por año siguiendo un crecimiento exponencial, mientras que sus palabras solo tuvieron un crecimiento bajo. De manera complementaria a esta información, el comportamiento de las palabras nos indica que tanto varían las cosas que patentan. Se observa

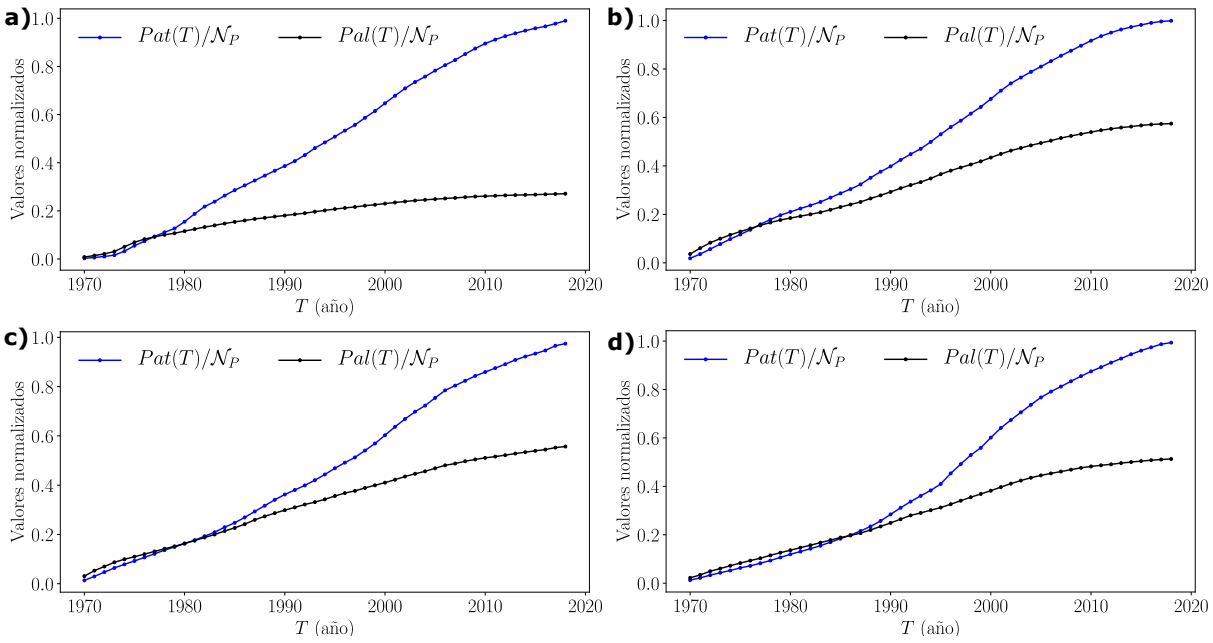


Figura 3.2: Patentes y palabras acumuladas de empresas farmacéuticas, Se muestran las gráficas correspondientes al número de patentes $Pat(T)$ y palabras acumuladas $Pal(T)$ a través del tiempo T en años desde 1970 hasta el 2019 de las 4 empresas farmacéuticas con más patentes a lo largo de su historia las cuales son a) Teijin, b) Bayer, c) Novartis y d) Merck n cada caso, los valores se encuentran normalizados dividiendo entre el número total de patentes N_p.

que las cuatro empresas mantienen poca variación entre las cosas que patentan. Texas Instruments "c)" a medida que pasan los años patentan con una intensidad media desde el año 1970 hasta el año 1987 a partir de este año patentan con una intensidad mayor donde su curva tiende a una recta, igualmente con las palabras de dichas patentes su crecimiento es sigue el mismo ritmo lo que indica que las palabras cambian constantemente por el tiempo. Por otra parte, se observa que Siemens "d)" tuvo un gran crecimiento desde el año 1990 hasta el 2007 donde su crecimiento disminuyó mientras que sus palabras igualmente tuvieron un gran crecimiento hasta el año 2008, posteriormente se mantuvieron con un crecimiento extremadamente bajo lo cual indica que no hubo gran variedad de palabras desde el año 2008.

Aplicando el mismo tipo de análisis, en la figura 3.2 se presentan los resultados obtenidos para las 4 empresas farmacéuticas con más patentes a lo largo de su historia. Igualmente se observa el comportamiento de las patentes y de las palabras, en este caso las empresas son: Bayer Merck, Teijin y Novartis. Se observa que para la empresa Teijin "a)" las patentes a lo largo de su historia han mantenido un crecimiento constante desde el año 1970 hasta la actualidad mientras que la cantidad de palabras se ha mantenido casi sin cambios desde el año 1980, se observan pequeñas variaciones en cuanto al número de palabras. Por otra parte, para el comportamiento de Bayer "b)" sus patentes siguen un

crecimiento alto constante desde el año 1970 y hasta el año 2012 en donde su crecimiento se vuelve bajo mientras que sus palabras siguen igualmente una tendencia de crecimiento medio constante lo cual indica que la variedad de la palabra ha cambiado con el tiempo. Para Novartis "c)" al igual que para Teijin se sigue la misma tendencia en cuanto a sus patentes ya que del año 1970 al año 1995 tiene un crecimiento alto y a partir de ese año su crecimiento aumenta aún más mientras que sus palabras tienen un crecimiento constante desde los años 1970 hasta 2005 donde su crecimiento se ve mermado lo que indica que desde el año 2005 sus palabras no han variado. Finalmente, la empresa Merck "d)" se encuentra que sus patentes tienen un crecimiento alto y constante durante toda su historia, al igual que sus palabras lo que arroja que tienen palabras distintas a través del tiempo. De manera complementaria es posible decir que el comportamiento de estas empresas es muy parecido ya que sus patentes tienen un crecimiento alto a lo largo de su historia y a su vez sus palabras también cambian con el pasar de los años.

3.3. Entropía para observar la innovación

La entropía de Shannon nos ayuda a medir la incertidumbre de una fuente de información, también se puede considerar como la cantidad promedio de información que contienen los objetos de estudio usados; es decir, los objetos con menor probabilidad son los que aportan mayor información [4]. Por ejemplo, si se considera como sistema al conjunto de palabras que conforman el título de una patente palabras frecuentes como: "que", "el", "a", aportan poca información, mientras que palabras menos frecuentes como "fotografía", "transistor", "lente" aportan más información, por lo que si de un texto borramos la palabra "el" no afectará en la comprensión del texto mientras que si borramos la palabra "lente" no ocurrirá lo mismo, si todos los objetos son igualmente probables todos ellos aportan información relevante y la entropía es máxima. La entropía de Shannon se encuentra estrechamente relacionada con la entropía en termodinámica.

En la termodinámica se estudia un sistema de partículas cuyos estados X (usualmente posición y velocidad) tienen una cierta distribución de probabilidad, pudiendo ocupar varios microestados posibles. De esta manera, la entropía termodinámica es igual a la entropía de la teoría de la información de esa distribución (medida usando el logaritmo natural) multiplicada por la constante de Boltzmann k (constante física que relaciona temperatura absoluta y energía).

Con el fin de presentar una definición formal de la entropía de información, supongamos

que un evento (variable aleatoria) tiene un grado de indeterminación inicial igual a k (i.e. existen k estados posibles) y supongamos todos los estados equiprobables. Entonces la probabilidad de que se dé una de esas combinaciones será $p = 1/k$. Luego es posible representar la expresión c_i como [32]

$$c_i = \log_2(k) = \log_2([1/(1/k)]) = \log_2(1/p) = \log_2(1) - \log_2(p) = -\log_2(p). \tag{3.1}$$

Si ahora cada uno de los k estados tiene una probabilidad p_i entonces la entropía vendrá dada por la suma ponderada de la cantidad de información [32]

$$H = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - .... - p_k \log_2(p_k) = -\sum_{i=1}^{k} p_i \log_2(p_i). \tag{3.2}$$

Por lo tanto, la entropía de un mensaje X, denotado por $H(X)$ es el valor medio ponderado de la cantidad de información de los diversos estados del mensaje [32]. Así, utilizando propiedades de los logaritmos

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i) = \sum_i p(x_i) \log_2(1/p(x_i)). \tag{3.3}$$

Relación que representa una medida de la incertidumbre media acerca de una variable aleatoria y por tanto de la cantidad de información [32].

3.4. Entropía de Shannon para títulos de patentes

A partir de la información de palabras correspondientes a las patentes por año de las empresas de tecnología y farmacéutica se obtuvo la entropía acumulada de cada una de las empresas a lo largo de toda su historia dada por:

$$H(X) = \sum_i p(x_i) \log_2(1/p(x_i)). \tag{3.4}$$

La entropía de Shannon aplicada a las palabras en este contexto podría funcionar como un indicador de mayor o menor innovación ya que entre mayor sea la entropía de una empresa esto significa que sus respectivos conjuntos de palabras son independientes entre sí es decir existe una mayor diversidad de palabras lo que nos muestra que la empresa correspondiente a ese conjunto de palabras tiene una diversidad de elementos dentro de sus patentes muy grande y esto estar ligado a una variación en la innovación.

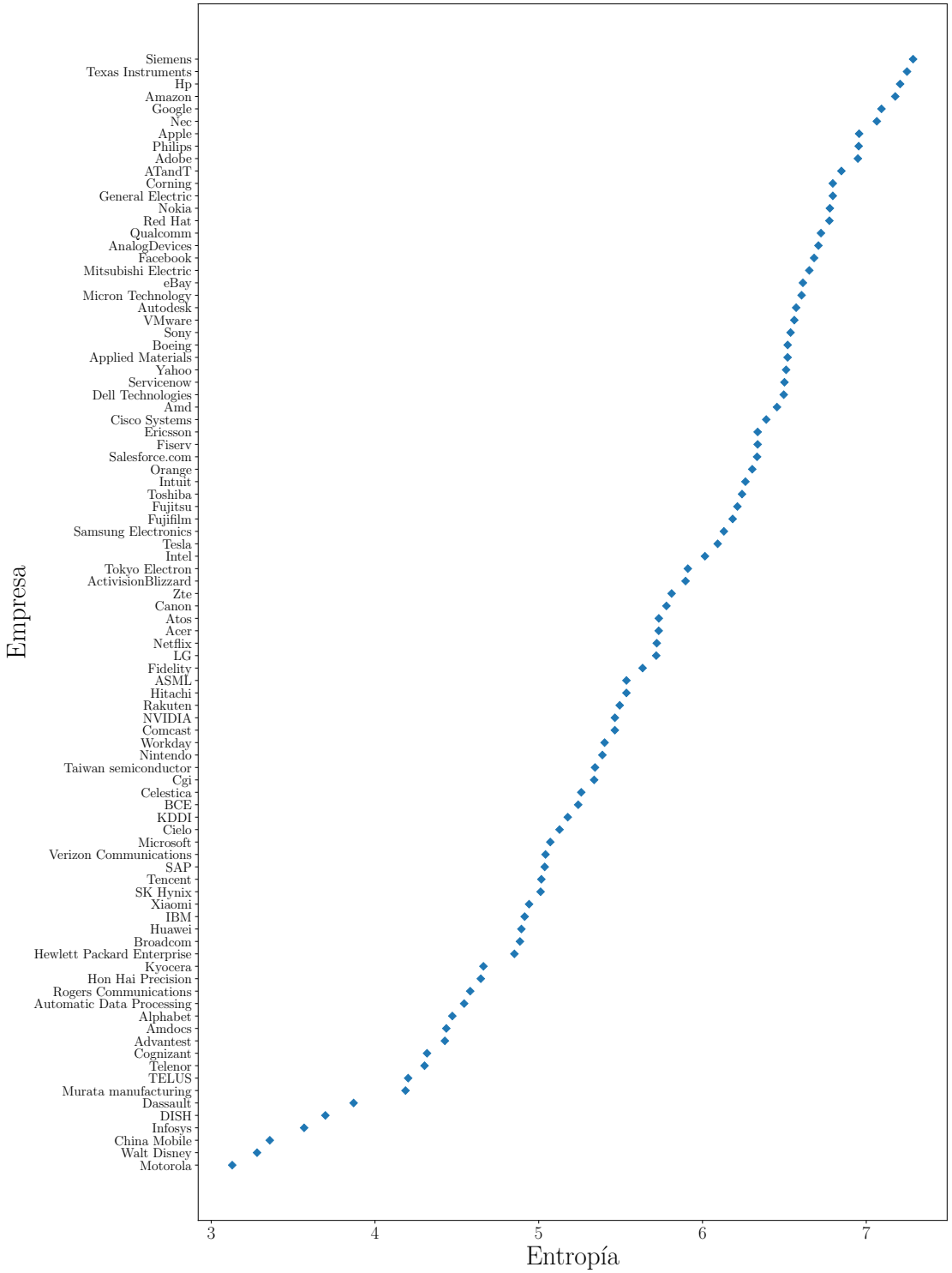


Figura 3.3: Entropía acumulada de empresas de tecnología. Se muestra la entropía acumulada de las mejores empresas de tecnología del mundo, donde es posible percatarse de las empresas con una mayor entropía las cuales corresponden a las empresas “Siemens”, “Google”, “Texas Instruments”, “Hp” y “Amazon”.

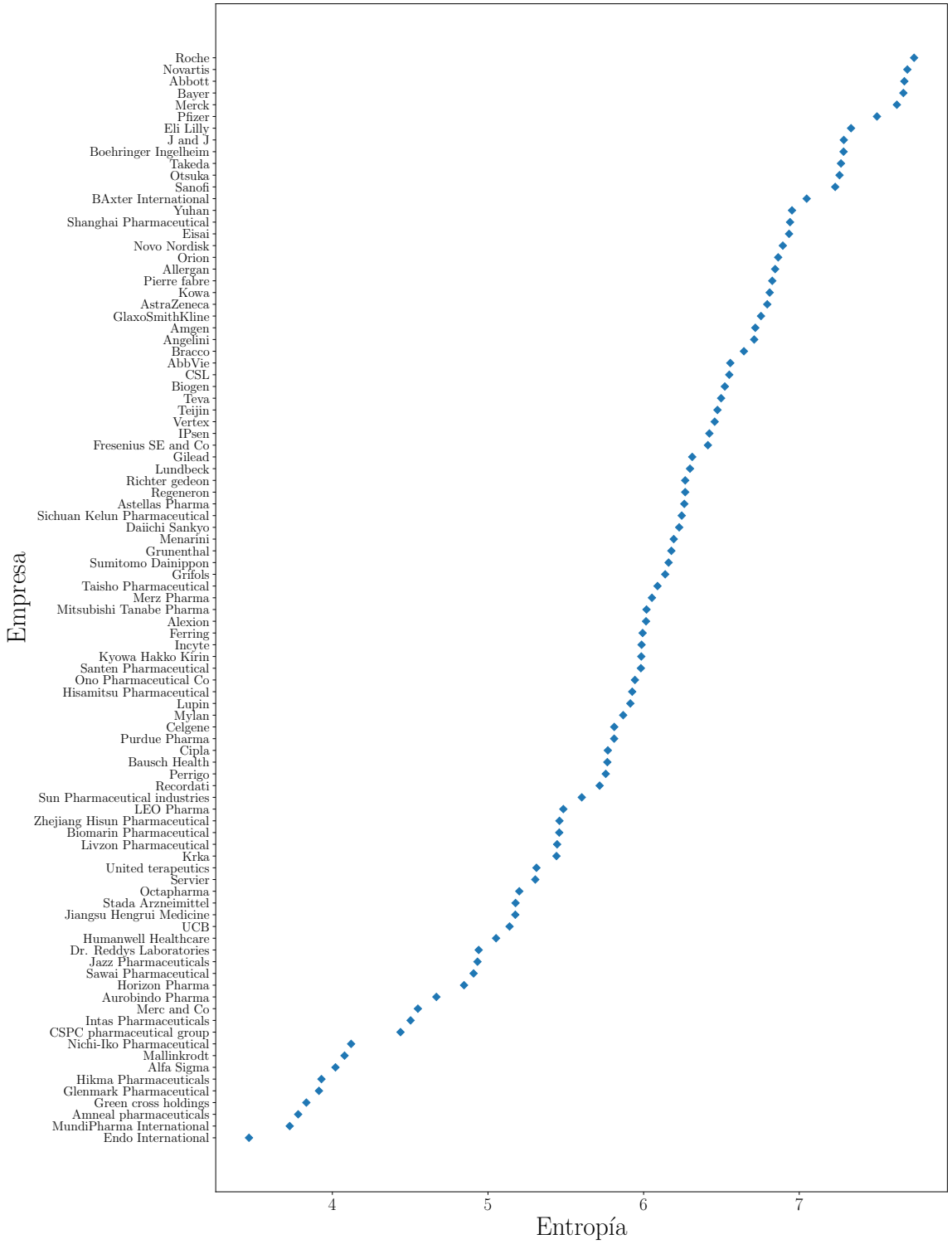


Figura 3.4: Se muestra la entropía acumulada de las mejores empresas de farmacéutica del mundo, donde es posible percatarse de las empresas con una mayor entropía las cuales corresponden a las empresas, "Roche", "Novartis", "Abbott", "Bayer", y "Merck".

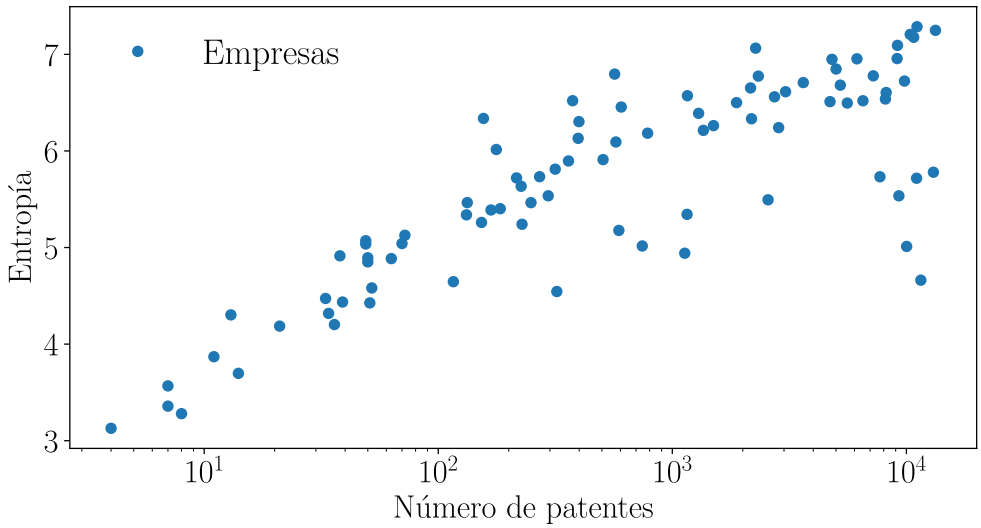


Figura 3.5: Comportamiento de las empresas de tecnología con respecto del número de patentes y la entropía para cada una de ellas el cual se observa con una escala logarítmica, y realizando un ajuste lineal se encuentra que no sigue una tendencia lineal, con resultados de un $R^2 = 0.26$

Para tener un panorama general del comportamiento de la entropía para las empresas de tecnología se muestran en la figura 3.3 las empresas de tecnología con su correspondiente entropía donde es posible percatarse de las empresas con una mayor entropía las cuales corresponden a las empresas "Siemens"," Amazon", "Texas Instruments", "Hp" y "Google", gracias a esto se observa que no necesariamente el tener más patentes corresponde a una mayor innovación como podría llegar a creerse, ya que de las las empresas con mayor número de patentes tenemos el caso de Canon empresa que cuenta con 13044 patentes y una entropía de 5.77 muy por debajo de las empresas con mayor innovación.

Análogamente para las empresas de farmacéutica se muestra en la figura 3.4 el comportamiento de su entropía donde se muestra que las empresas que tiene una entropía mayor son "Roche", "Novartis", "Abbott", "Bayer", y "Merck", igualmente que con las empresas de tecnología no necesariamente al tener muchas patentes le corresponde una entropía mayor, tal es el caso de la empresa Teijin que cuenta con 24999 patentes pero una entropía de 6.47, abajo de las empresas con mayor innovación.

Lo antes mencionado puede ser observado con mayor detalle en las figuras 3.5 y 3.6 donde se muestra la relación patentes-entropía donde tanto las empresas de tecnología como las de farmacéutica se encuentran casos con grandes cantidades de patentes y una entropía chica o menos patentes con una entropía grande, ya que la mayoría de las empresas siguen un comportamiento en el que a mayor número de patentes su entropía también es mayor

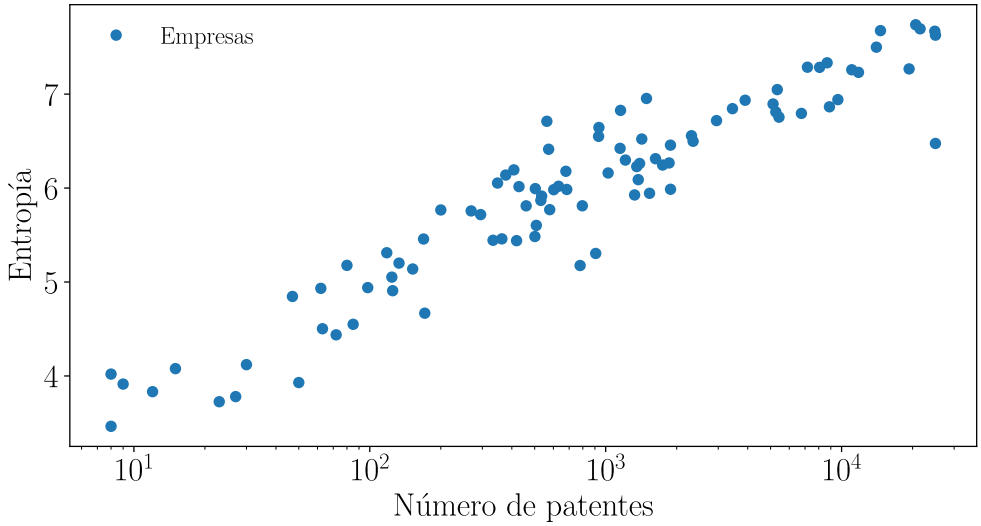


Figura 3.6: Comportamiento de las empresas de farmacéutica con respecto del número de patentes y la entropía para cada una de ellas el cual se observa con una escala logarítmica, y realizando un ajuste lineal se encuentra que no sigue una tendencia lineal, con resultados de un $R^2 = 0.40$

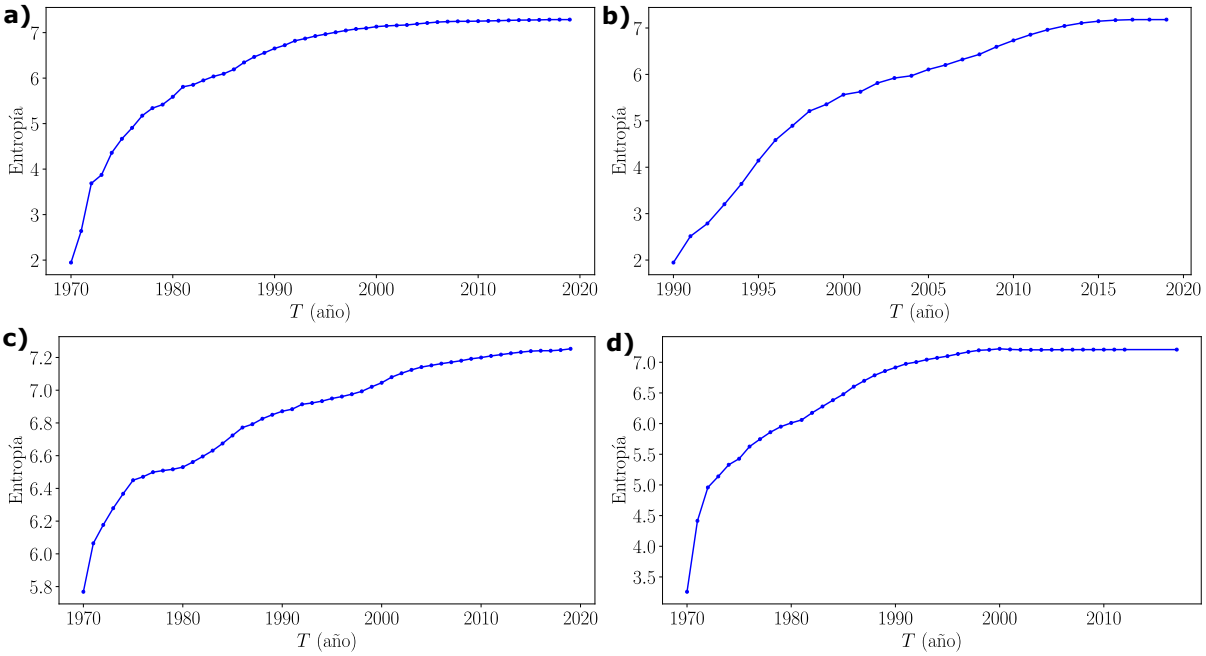


Figura 3.7: Se observa el comportamiento a lo largo de su historia de las empresas de tecnología con mayor entropía es decir con un mayor diversidad de palabras, lo que podría indicar una mayor innovación, a)Siemens, b)Amazon, c)Texas Instruments, d)Hp.

y por ende innovan más. Sin embargo, también existen casos en donde no se cumple esto, empresas con grandes cantidades de patentes y una entropía chica o por el contrario no muchas patentes con una entropía grande como los vistos anteriormente de Canon para las empresas de tecnología y Teijin para las empresas de farmacéutica.

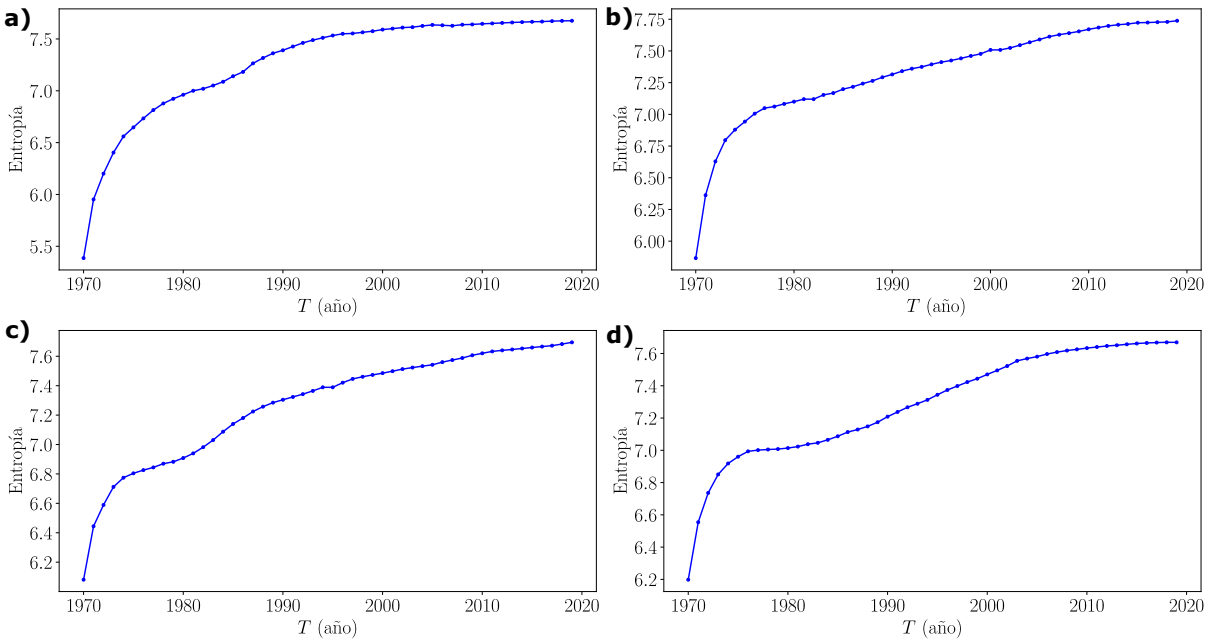


Figura 3.8: Se observa el comportamiento a lo largo de su historia de las empresas de Farmacéutica con mayor entropía es decir con un mayor diversidad de palabras, lo que podría indicar una mayor innovación, a)Abbott , b)Roche , c)Novartis , d)Bayer

En la figura 3.7 se muestra la entropía a lo largo de toda su historia de las empresas con una mayor entropía, comenzando con Siemens "a)" donde se muestra un crecimiento de entropía desde el año 1970 hasta los años 2000 esto muestra que en este intervalo de tiempo esta empresa contó con la innovación más grande a lo largo de su historia, seguida por una de las empresas más jóvenes Amazon "b)" a pesar de contar con menos tiempo que las anteriores empresas cuenta con un comportamiento ascendente en la mayoría de su historia mayor innovación dentro de la empresa, se observa que Texas instruments "c)" la cual tiene un comportamiento de mayor innovación en los años 70's aunque a diferencia de Siemens está aún sigue con un comportamiento ascendente. Finalmente Hp "d)" sigue un comportamiento similar a Siemens ya que tiene un intervalo de crecimiento y mayor innovación que va de los años 70's a los 2000, mientras que posteriormente le sigue un intervalo más estable lo cual indica que la empresa dejó de innovar.

Así como para las empresas de tecnología, en la figura 3.8 se muestra el acumulado de la entropía de las 4 empresas de farmacéutica con mayor entropía a lo largo de su historia, comenzando por Abbott "a)" la cual sigue un comportamiento en el cual tiene un intervalo de mayor innovación entre los años 70's y 2000 con una entropía de 7.5, seguido por un intervalo de poco crecimiento es decir menor innovación, a diferencia de la empresa Roche "b)" la cual tiene un crecimiento muy grande en cuanto a su entropía desde el año 1970 y hasta 1977 teniendo una entropía de 7.0 en donde después de este año su entropía

sigue aumentando pero a un menor ritmo lo que nos arroja que en el primer periodo su innovación fue mayor igualmente Novartis "c)" tiene un periodo de mayor innovación el cual comprende desde el año 1970 hasta 1975 donde su entropía se encontraba alrededor de 7.0 aunque a diferencia de Roche su entropía después de este periodo crece pero a un menor ritmo llegando al punto de que en el periodo de entre los años 2005 y 2006 a la actualidad no tienen una variación en su entropía manteniéndose en alrededor de 7.6 y por ende se observa una menor innovación. Finalmente, para Bayer en "d)" se observa un comportamiento similar de crecimiento durante toda su historia con un intervalo de mayor crecimiento en los años 70's, este crecimiento nos indica que a lo largo de toda su historia ha estado innovando pero manteniendo su periodo de mayor innovación en sus primeros 3 años pasando de una entropía de 0 hasta 6.8.

Capítulo 4

Detección de patrones en grupos de patentes

4.1. Introducción

En este capítulo se explora la detección de patrones en bases de datos utilizando redes de similitud. En la primera parte se abordan conceptos de la ciencia de redes, los cuales son aplicados en el análisis de detección de patrones en comunidades de empresas con similares características creando redes de empresas, esto permite detectar grupos con actividad similar. En la primera parte se define el concepto de red y la historia acerca de ellas, así como las características y propiedades que comprenden a una red compleja. La descripción se limita al concepto de redes simples.

En la segunda parte se aborda la detección de patrones utilizando redes, comenzando por la creación de las correspondientes matrices de distancias entre curvas de la evolución de palabras en las patentes registradas por empresas de tecnología y farmacéuticas. A partir del análisis probabilístico de estas distancias, tanto para las empresas como para las patentes, se obtienen valores de referencia para obtener indicadores de similitud entre ellas llamados parámetro umbral de similitud para palabras y parámetro umbral de similitud para patentes. Finalmente, aplicamos la detección de comunidades para detectar patrones en bases de datos. Un enfoque similar se ha implementado recientemente para el minado de textos en los resúmenes de patentes [33], o en el análisis de datos con registros del movimiento de vehículos en el sistema metrobús de la Ciudad de México [34]. En el caso estudiado para el texto de títulos de patentes se utilizan dos parámetros umbrales para

definir una red y aplicando algoritmos de detección de comunidades se definen grupos de empresas con características similares entre sí.

4.2. Redes

Las redes o grafos son objetos matemáticos constituidos por un conjunto de nodos llamados también vértices y conectados mediante uniones o bien aristas formando una estructura. La teoría de grafos es una rama de las matemáticas la cual se comenzó a desarrollarse a mediados del siglo XVIII por el matemático y físico Leonhard Euler (1707-1783) [35], aunque más concretamente, el campo de las redes complejas aparece en la última década del siglo pasado con dos artículos de Watts & Strogatz sobre redes y mundo pequeño [36] y el de Barabási & Albert sobre la aparición de redes complejas invariantes de escala [37].

Algunos ejemplos de lo que son las redes son, por ejemplo, el internet, infraestructuras complejas como redes de interacción de proteínas dentro de las células, las redes de carreteras o vuelos, redes sociales y el cerebro por mencionar algunas [37]. El concepto de red permite el estudio de sistemas complejos describiendo las interacciones presentes de un sistema [38].

Una red compleja puede ser representada formalmente como un grafo. Un grafo está conformado por dos conjuntos $\mathcal{V} = \{n_1, ..., n_N\}$ cuyos elementos son los nodos, vértices, siendo $N = |\mathcal{V}|$ el número total de nodos. Por otra parte $\mathcal{E} = \left\{l_1, ..., l_{|\mathcal{E}|}\right\}$ es el conjunto de conecciones cuyos elementos son las uniones o aristas. Por lo tanto un grafo tiene $\mathcal{V}$ nodos y $\mathcal{E}$ aristas y se denota como $G = (\mathcal{V}, \mathcal{E})$.

Las aristas se definen por los órdenes de los nodos que unen, es decir, la unión entre los vértices n_i y n_j se denota por $l_k = (i, j) = (n_i, n_j) = l_{ij}$. En este caso, si hay una unión o arista entre dos nodos, los nodos se llaman vecinos o adyacentes [39].

Existen diversos tipos de grafos, entre ellos están [38]:

- Grafo no-dirigido: aquel en el que el orden de los índices en las uniones es el mismo: $l_{ij} = lji$.
- Grafo dirigido: se tiene que el orden de los índices en las uniones es importante de forma que $l_{ij} \neq l_{ji}$.

- Multigrafos: son grafos que tienen auto uniones o lazos, por ejemplo l_{ii}, o múltiples uniones entre los mismos dos nodos. Los grafos no dirigidos que no tienen lazos ni aristas paralelas se llaman grafos simples.
- Grafos pesados: Son grafos en los que a cada arista se le asigna un peso o valor numérico que mide la intensidad de la unión. En caso contrario la red o grafo se llama no pesado.
- Grafo vacío: grafo sin aristas, sólo con nodos.
- Grafo nulo: el que no tiene vértices (y por lo tanto no tiene aristas).

Las redes, representadas matemáticamente por medio de grafos, son estructuras en las que los nodos se describen por puntos y las uniones mediante líneas que conectan puntos adyacentes. Para caracterizar a dichas redes se usan dos tipos de representaciones: la lista de aristas y la matriz de adyacencia. Ambas caracterizaciones dependen del tipo de grafo que estemos estudiando[38].

En el caso de un grafo simple la lista de aristas es el conjunto $\mathcal{E}$ pares de nodos (n_j, n_i), o sus etiquetas (j, i) indicando que el nodo j está unido al i por una arista [40]. Por otra parte, en la representación utilizando una matriz de adyacencia A de un grafo simple de N nodos, toda la conectividad de la red está dada por una una matriz cuadrada $N \times N$ cuyos elementos son:

$$A_{ij} = \begin{cases} 1 & \text{si } (i,j) \in \mathcal{E}, \\ 0 & \text{cualquier otra situación.} \end{cases} \tag{4.1}$$

El grado de un nodo i se define como el número total de aristas incidentes en dicho nodo y se denota por k_i, el grado de un nodo se puede calcular directamente a partir de la matriz de adyacencia

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ji}. \tag{4.2}$$

También existe una medida del grado en el que los nodos de una red tienen a agruparse entre ellos. Dado un grafo $G(N, \mathcal{E})$ no dirigido se define el coeficiente de agrupamiento de un nodo i de dicho grafo y se denota como C_i al cociente el cual se llama coeficiente de agrupamiento (clustering coefficient) [41]:

$$C_i = \frac{\text{Pares } (l,m) \text{ de nodos vecinos de } i \text{ conectados por aristas}}{\text{Total de pares que pueden existir con los nodos vecinos de } i}. \tag{4.3}$$

4.3. Comunidades

La idea intuitiva utilizada en la gran mayoría de los trabajos para establecer qué es una comunidad, es que los vértices de la misma deben estar más relacionados entre sí que con el resto de los vértices de la red. En función de ésta idea general se han propuesto numerosos criterios cuantitativos para definir que es una comunidad. Estos criterios pueden clasificarse en:

- Definiciones locales: se analiza la estructura interna de la comunidad, sin tener en cuenta el resto de la red.
- Definiciones globales: se analiza el papel de la comunidad en la estructura global de la red.

Como ya se definió, una comunidad en una red es un subgrupo de ésta, el cual está más densamente conectado entre sí que con el resto de la red. Cuando un nodo de esta red pertenece a más de un subgrupo a la vez se dice que la red está solapada. La definición de solapamiento es válida y es precisamente lo que vemos en la realidad. Por ejemplo, vemos en una red social como Facebook en donde una persona cualquiera (la cual representaría un nodo) puede pertenecer a varias comunidades o grupos si tomamos en cuenta sus intereses [42]. De forma intuitiva podemos deducir que las personas podrían llegar a tener más de un interés, pero en este caso no podríamos asignar cada vértice (persona) a una sola comunidad, es acá donde aparece el término de comunidad solapada, en donde estamos en presencia de comunidades que comparten nodos entre sí [43], es decir, vértices que pertenecen a dos o más comunidades. Algunos métodos de detección de comunidades son:

- Métodos divisivos que consiste en eliminar las aristas que conectan vértices pertenecientes a distintas comunidades, de esta manera quedan aisladas unas de otras.
- Métodos de clustering que buscan determinar cómo separar adecuadamente un conjunto de nodos respecto a un número predefinido de clusters o grupos. Para poder utilizar este tipo de métodos es necesario asociar al grafo una métrica espacial, de tal manera de que cada nodo se encuentre a una distancia respecto a cada uno del resto de nodos que componen el grafo. Esta distancia puede ser una medida de similitud o de disimilitud [44].
- Métodos Jerárquicos los cuales tiene por objetivo buscar las divisiones naturales de la red en grupos, basados en la idea de que el grafo tiene una estructura jerárquica,

es decir, pequeños grupos de nodos que son parte de grupos medianos de nodos y que a su vez éstos pertenezcan a grupos más grandes y así sucesivamente.

- Métodos basados en la Modularidad la cual representa uno de los primeros intentos por lograr entender los principios del problema de clustering, integrando en su función de calidad todos los elementos esenciales, desde la definición de comunidad, pasando por la elección de un modelo nulo de comparación, hasta la expresión de solidez o fortaleza de las comunidades y particiones encontradas [45].

Una red, o un grafo, se denomina completa si el número de sus conexiones, o vértices, es igual $L_{max} = \frac{N(N-1)}{2}$ por lo que su correspondiente matriz adjunta estará llena de valores no nulos, salvo aquellos de la diagonal, siempre que los elementos no interactúen consigo mismos. En caso de que no sea completa, puede suceder que haya subconjuntos de elementos que interactúan mucho entre sí con respecto a aquellos elementos que no están en tales subconjuntos en este caso, se dirá que se ha formado una comunidad. Considerando un subgrafo C de N_C nodos dentro de una red. El grado externo, K_i^{ext} es el número de conexiones de ese nodo i con el resto de la red, mientras que grado interno, K_i^{int} , es el número de conexiones de i con otros nodos contenidos en C. Si $K_i^{ext} = 0$ entonces todos los vecinos de i están dentro de C y la comunidad es adecuada para i. Si $K_i^{int} = 0$ quiere decir que todos los vecinos de i están fuera de C, y de esta manera i debe ser asignado a otra comunidad. Lo anterior, ofrece un parámetro para determinar si una comunidad es fuerte o débil. Así, se tendrá que una comunidad es fuerte sí [41]

$$\forall i \in K_i^{int}(C) > K_i^{ext}(C) \tag{4.4}$$

y una comunidad débil si:

$$\sum_{i \in C} K_i^{int}(C) > \sum_{i \in C} K_i^{ext}(C) \tag{4.5}$$

esto es, si el grado interno total es mayor que el grado externo total. Con el fin de buscar de manera eficiente comunidades dentro de una red, hay que definir los conceptos de bisección del grafo y de tamaño mínimo de corte [41]. El primer término se refiere, simplemente, a partir en dos comunidades a un grafo de manera que no tengan elementos en común. En cuanto al tamaño mínimo de corte, la bisección debe hacerse de manera tal que el número de conexiones entre las dos comunidades sea mínimo. Ahora bien, el número de

combinaciones posibles que cumplen esta condición es [41]:

$$B \simeq \frac{N^{N+\frac{1}{2}}}{N_1^{N+\frac{1}{2}} N_2^{N+\frac{1}{2}}} \tag{4.6}$$

donde N es el número de nodos de la red y N_1 y N_2 son los nodos en la comunidad 1 y 2, respectivamente. Si consideramos el caso en que $N_1 = N_2$, entonces.

$$B \simeq \frac{2^N + 1}{\sqrt{N}} = \exp\left\{(N+1)\log 2 - \frac{1}{2}\log N\right\} \tag{4.7}$$

con lo que se ve que el número de combinaciones crece de forma exponencial con cada aumento de nodos. Así, para una red de 10 nodos en las que las comunidades 1 y 2 tengan el mismo tamaño y con un tiempo de cálculo de 1 milisegundo por inspección, tardaríamos 10^{16} años en encontrar las comunidades correspondientes. Gracias a este sencillo ejemplo nos podemos dar cuenta lo importantes que son los algoritmos de detección de comunidades. A continuación se describen algunos de los algoritmos usados para detección de comunidades[41].

- Algoritmo de agrupamiento jerárquico. Este procedimiento ayuda a conseguir tiempos de ejecución algorítmica que crecen polinomialmente con cada nodo N. El método inicia con una matriz de similaridad X cuyos elementos x_{ij} representan la similitud entre los nodos i y j. Esta distancia puede calcularse por la similitud coseno, el índice de Jaccard y la distancia Hamming [46]. Una vez que las distancias para cada par de nodos ha sido calculada, el algoritmo agrupa aquellos nodos con alta similitud; esté último paso puede hacerse de diferentes maneras: por medio de algoritmos aglomerativos o con algoritmos divisivos. Entre los algoritmos aglomerativos está el Algoritmo de Ravasz y entre los divisivos, el Algoritmo de Girvan-Newman[47].

- Algoritmo de Ravasz. Se lleva a cabo en cuatro pasos. En el primero se calcula la matriz de traslapamiento topológico, cuyos coeficientes X_{ij}^o se calcula a partir de

$$x_{ij}^o = \frac{J(i,j)}{\text{mín}(k_i, k_j) + 1 - \theta(A_{ij})} \tag{4.8}$$

  donde $\theta(x)$ es la función escalón de Heaviside, $J(i,j)$ es el número de vecinos en común de los nodos i y j; a este valor se le suma un uno si existe una conexión entre los nodos; $\text{mín}(k_i, k_j)$ es el más pequeño de los grados de dichos nodos. Entonces, se tendrá que: $x_{ij}^o = 1$ si los nodos están conectados y tienen los mismos vecinos y

$x^o_{ij} = 0$ en caso contrario. El siguiente paso es determinar qué tan similares son las comunidades que se han obtenido, ello se hace calculando el promedio de x_{ij} sobre todos los pares de nodos que pertenecen a comunidades distintas. El tercer paso es aplicar el agrupamiento jerárquico. Por último, se usa un dendograma para extraer la organización de comunidad que subyace en este conjunto [48].

- Algoritmo de Girvan-Newmann. Inventado en 2001 por M. Girvan y M. E. J. Newman, el algoritmo de Girvan-Newmann se lleva a cabo en cuatro pasos[38]:
  - 1.- Calcular la betweenness(medida utilizada en análisis de redes para determinar la importancia relativa de un nodo o una arista en una red) para todos los vértices en la red; este es el término con el que se le denomina a los elementos x_{ij}.
  - 2.- Eliminar el vértice con la betweenness más pequeña.
  - 3.- Recalcular la betweenness para todos los vértices afectados.
  - 4.- Repetir desde el paso 2 hasta que no sobren vértices.

La maximización de la modularidad es uno de los métodos más usados. En este método, la modularidad se refiere a qué tan buena es la división de una red en comunidades. Para ello, se toma una red de N nodos y L enlaces, en la que cada comunidad tiene N_c nodos. Entonces, se mide la diferencia M_c entre el diagrama real de enlaces de la red y el número esperado de enlaces entre los nodos i y j si los enlaces de la red se construyen aleatoriamente. Así, la modularidad se calcula como [49]:

$$M_c = \frac{1}{2L} \sum_{(i,j)\in C_c} (A_{ij} - p_{ij}) \tag{4.9}$$

donde p_{ij} suele expresarse como $p_{ij} = \frac{k_i k_j}{2L}$ los valores de la modularidad llevan a 3 posibles casos: 1. si $M_c > 0$ entonces el subgrafo C_c tiene más conexiones que las esperadas por azar; 2. para $M_c = 0$ la conectividad entre los N_c nodos es aleatoria; 3. si $M_c < 0$ entonces los nodos de C_c no forman una comunidad. El método de Louvain, llamado así por el lugar en el que se creó, fue ideado por Vincent Blondel y sus estudiantes en 2007. Este método se usa en redes pesadas y propone la optimización de la modularidad, que se calcula como [49]:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{4.10}$$

se reemplaza a L por m, que es la suma de todos los pesos de las conexiones de la red y se calcula como $m = \frac{1}{2}\sum_{ij} A_{ij}$ y k_i es el grado de i pero considerando los pesos de las conexiones. El método en sí, consta de dos fases que se repiten de manera iterativa. Primero, se asigna una comunidad diferente a cada nodo de la red, por lo que en este paso hay tantas comunidades como nodos en la red. Entonces, se consideran los vecinos j del nodo i y se evalúa la ganancia en modularidad ΔQ que habría si i se extrajera de su comunidad y se colocará en la de j. De esta manera el nodo se queda en la comunidad donde la ganancia de modularidad sea mayor siempre que esta ganancia sea positiva. Si no existe ganancia positiva i se queda en su comunidad, el cálculo se hace como sigue:

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]. \tag{4.11}$$

En la segunda fase del método, se construye una nueva red, donde los nodos son las comunidades obtenidas en el paso anterior. Las conexiones entre nodos de la misma comunidad se convierten en auto-conexiones y las conexiones a otras comunidades se condensan en una sola conexión cuyo peso se determina con el número de conexiones anteriores. En esta fase se crea una nueva red, y el proceso se repite tantas veces como se considere necesario[49].

4.4. Detección de patrones en bases de datos

Para comenzar con la detección de patrones en la base de datos de patentes, una forma de realizar este análisis es por medio de la creación de vectores $\vec{p_i} = (p_i(t_0), p_i(t_1), \ldots, p_i(t_M))$ y $\vec{w_i} = (w_i(t_0),\ w_i(t_1), \ldots, w_i(t_M))$ los cuales se encuentran definidos como $p_i(t)$ correspondiente a la fracción de patentes determinada en cierto tiempo t, mientras que $w_i(t)$ corresponde a la fracción de la palabras, representando las curvas de palabras y patentes mostradas en el capítulo anterior en las figuras 3.1 y 3.2 las cuales muestran el comportamiento de las palabras y patentes a lo largo de la historia de las empresas.

De esta manera, para todas las empresas de tecnología y farmacéutica que se analizarán se tienen sus correspondientes vectores de palabras y patentes y con la información de estos vectores se construye una matriz de distancias a partir de la cual es posible comparar las curvas de cada una de las empresas con respecto de las demás tanto por palabras como por patentes. Para ello se tomó cada uno de los vectores de la empresas y se compararon a partir de la norma Euclidiana donde $D_{patentes}(i, j)$ es la distancia entre la empresa i con

respecto de la empresa j para sus patentes en determinados períodos de tiempo. En forma similar, $D_{palabras}(i,j)$ corresponde a la distancia entre curvas asociadas a palabras. La norma Euclidiana en dos dimensiones para un vector $\vec{A}-\vec{B}$ es definida como la distancia (en línea recta) entre dos puntos $\vec{A}=(a_1,a_2,\ldots)$ y $\vec{B}=(b_1,b_2,\ldots)$ que delimitan dicho vector [50]:

$$\|\vec{A}-\vec{B}\| = \sqrt{(b_1-a_1)^2+(b_2-a_2)^2+....}$$

Gracias a esto se obtienen las distancias entre ellos la cual nos arroja que tanto se separa una curva con respecto de las demás como es mostrado a continuación.

Por otra parte, para cada fracción p_i, p_j de patentes y las cantidades w_i, w_j asociadas a palabras, correspondientes a las empresas i, j que se encuentran definidas como:

$$p_i = (p_i(t_0), p_i(t_1), p_i(t_2), ..., p_i(t_f)),$$

$$w_i = (w_i(t_0), w_i(t_1), w_i(t_2), ..., w_i(t_f)).$$

A partir de estas definiciones, se puede establecer la distancia entre listas de patentes:

$$\begin{aligned} D_{patentes}(i,j) = \|\vec{p_i}-\vec{p_j}\| &= \sqrt{(p_i(t_0)-p_j(t_0))^2+(p_i(t_1)-p_j(t_1))^2+...} \\ &= \sqrt{\sum_{n=1}^{M}(p_i(t_n)-p_j(t_n))^2} \end{aligned} \tag{4.12}$$

y entre listas de palabras:

$$\begin{aligned} D_{palabras}(i,j) = \|\vec{w_i}-\vec{w_j}\| &= \sqrt{(w_i(t_0)-w_j(t_0))^2+(w_i(t_1)-w_j(t_1))^2+...} \\ &= \sqrt{\sum_{n=1}^{M}(w_i(t_n)-w_j(t_n))^2}. \end{aligned} \tag{4.13}$$

En la figura 4.1 se presentan las matrices de distancias para todas las empresas. Donde $D_{palabras}(i,j)$ y $D_{patentes}(i,j)$ corresponde a la norma euclidiana tanto para palabras como para patentes de las distintas empresas. En esta representación se puede observar tanto la distancia entre patentes, así como la distancia entre las curvas de palabras para las empresas de farmacéutica y tecnología. De esta manera, se cuantifica qué tan alejado o cercano está el comportamiento de las curvas que describen a cada empresa. A partir de

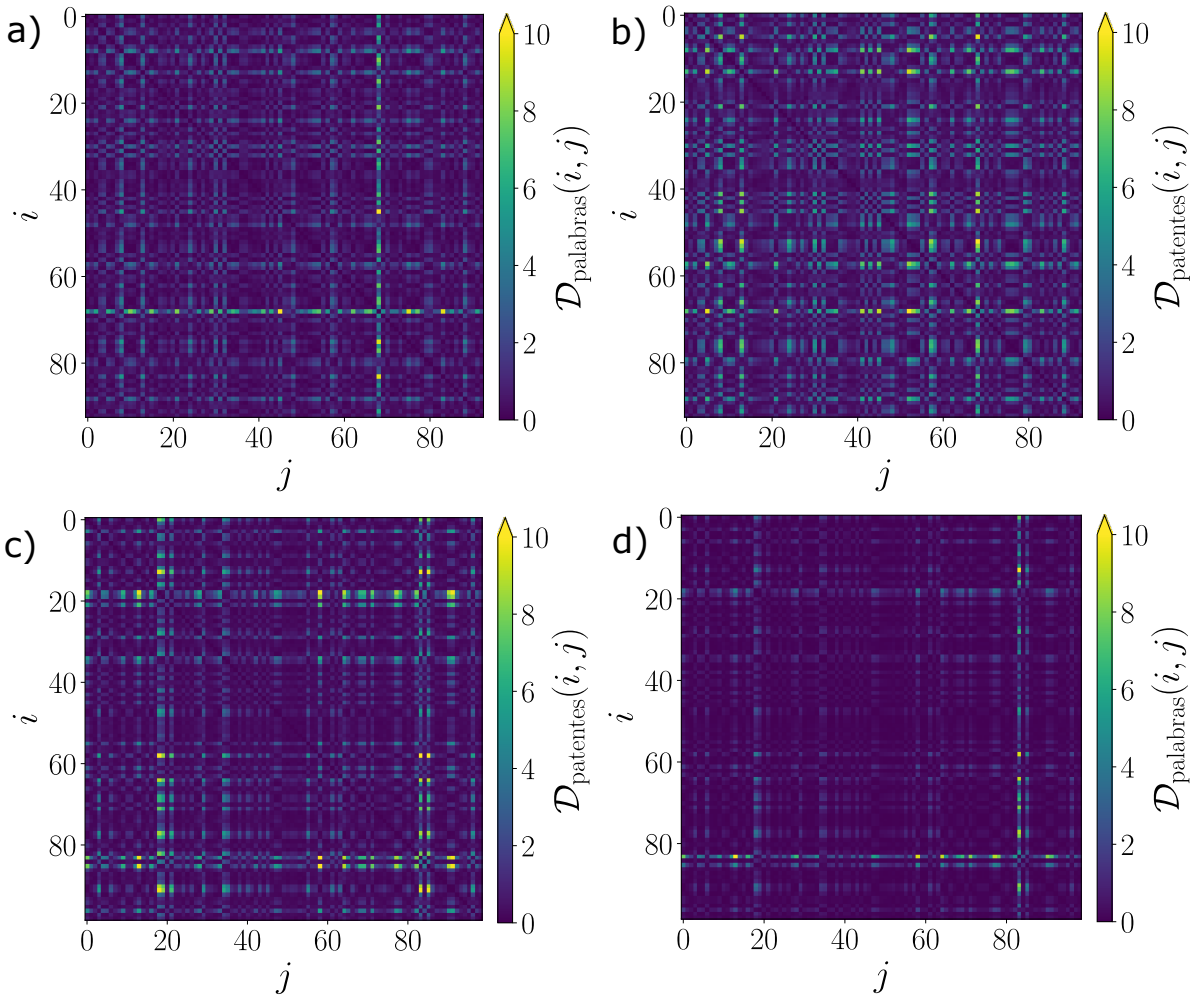


Figura 4.1: Matrices de distancias entre empresas de tecnología y farmacéutica. Se observan las matrices de las distancias obtenidas entre las empresas, matriz a) matriz de distancias entre las palabras de las empresas de tecnología, con los valores de distancias más grandes en $D_{palabras}$ = (69,46), (69, 76) y (69,84) correspondientes a la empresa Recruit Holdings con las empresas Hp, Siemens y Texas Instruments respectivamente b) matriz de distancias entre patentes de las empresas de tecnología donde los valores $D_{patentes}$ se encuentran muy dispersos al igual que en las patentes de las de farmacéutica de la matriz c), finalmente la matriz d) muestra las distancias entre palabras de las empresas de farmacéutica donde las distancias más grandes se encuentran en $D_{palabras}$ = (83, 14) y (83,59) que corresponden a la empresa Shanghai Pharmaceutical con las empresas Bayer y Merck.

la escala cromática a un costado de la matriz, se puede apreciar que gran parte de las distancias son menores a 2, dada la predominancia visual de los tonos oscuros. También se puede advertir la formación de bloques de tonos oscuros y líneas de tonos brillantes.

Se observa también que para la primera matriz a) correspondiente a las distancias entre las palabras de las empresas de tecnología donde los valores mostrados más grandes son $D_{palabras}$ = (69,46), (69, 76) y (69,84) correspondientes a la empresa Recruit Holdings con las empresas Hp, Siemens y Texas Instruments respectivamente. Esto se observa ya que

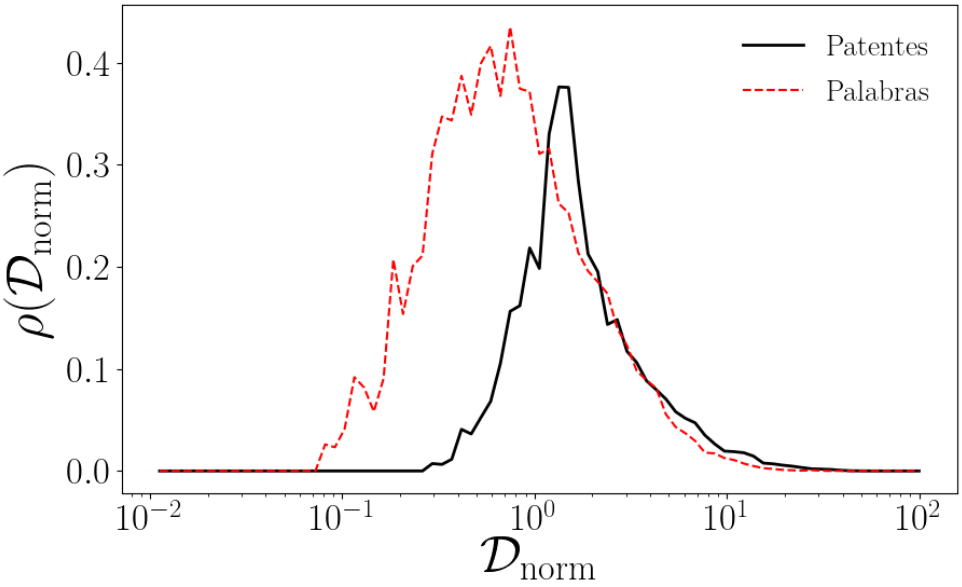


Figura 4.2: Densidad de probabilidad para las distancias en empresas de tecnología. Donde D_{norm} representa las distancias entre palabras $D_{palabras}$ (línea roja punteada) y distancias entre patentes para $D_{patentes}$ (línea negra). Se observa la densidad de probabilidad en las empresas de tecnología tanto para las palabras como para las patentes, en donde encontramos la mayoría de los valores en torno a 1.7 para las patentes y 1.0. Estos valores nos dan información sobre el intervalo en el que se encuentra la mayor cantidad de datos $D_{palabras}$, $D_{patentes}$ para definir valores umbrales.

estas últimas son empresas que cuentan con una gran variedad de palabras por lo que hay una gran distancia entre ellas con respecto de Recruit Holdings. A diferencia de la matriz de distancia entre palabras, se observa que la matriz de distancia entre patentes b) mantiene una proporción de distancias grandes entre empresas lo cual ocurre también para la matriz las patentes de empresas farmacéuticas c).

Para la matriz que corresponde las distancias entre palabras de empresas de farmacéutica d) es posible percatarse que los valores más altos que se observan son $D_{palabras} = (83, 14)$ y (83,59) que corresponden a la empresa Shanghai Pharmaceutical con las empresas Bayer y Merck respectivamente lo que al igual que en las empresas de tecnología nos muestra la distancia tan grande que hay entre ellas de acuerdo con la gran variedad de palabras con que cuentan cada una de estas últimas empresas.

En la figura 4.2 se muestra la estadística de las distancias creadas a partir de las matrices de distancias, tanto para las palabras y patentes de las empresas de tecnología las cual nos permite conocer sus correspondientes valores característicos. Estas cantidades nos son de ayuda para determinar los intervalos en los que se encuentra la mayor cantidad de datos. A partir de $D_{patentes}$, se puede apreciar que el máximo para las distancias de patentes

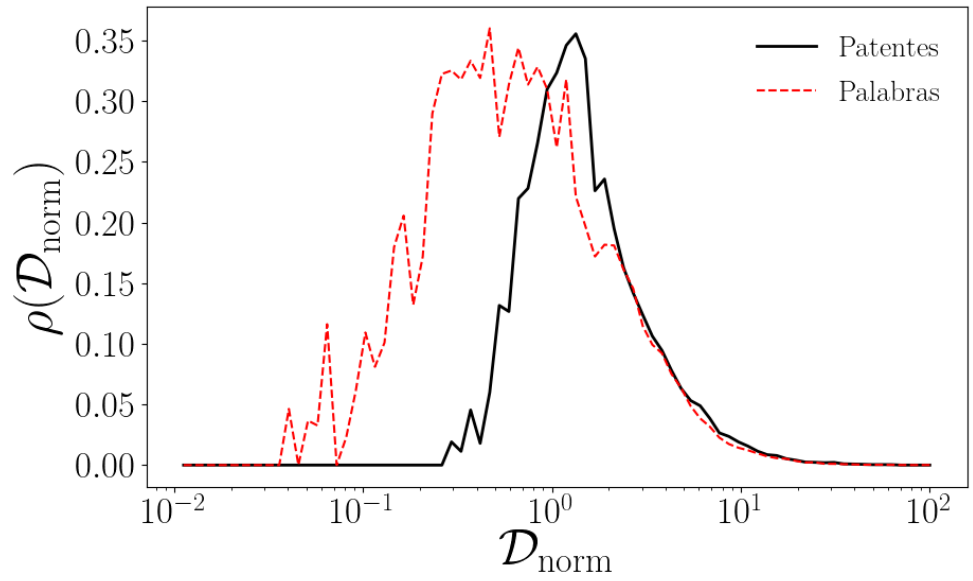


Figura 4.3: Densidad de probabilidad para las distancias en empresas farmacéuticas. Donde D_{norm} representa las distancias entre palabras $D_{palabras}$ (línea roja punteada) y a las distancias entre patentes $D_{patentes}$ (línea negra). Se observa la densidad de probabilidad en las empresas de farmacéutica tanto para las palabras como para las patentes, en donde encontramos la mayoría de valores en torno a 1.8 para las patentes, y 0.8 para las palabras los cuales nos aportan información acerca de valores característicos asociados a $D_{palabras}$ y $D_{patentes}$.

entre empresas se sitúa en 1.7 para las patentes mientras que su media es de 5.1 con una desviación estándar de 6.0 lo que nos muestra que hay una gran variabilidad en las distancias de patentes. En forma similar, mediante los valores $D_{palabras}$, se observa que el máximo para las distancias entre palabras se encuentra en 1.0 con una media de 2.8 y una desviación estándar de 3.4 que nos indican que los datos tienen una tendencia central de alrededor de 2.8 y una dispersión relativamente grande alrededor de esa media, teniendo en cuenta que el 40 % de todos los valores de las distancias entre palabras se encuentran entre 0.3 y 5.0 mientras que las patentes el 37 % de todos los valores se encuentran entre 0.8 y 8.0. A su vez es posible percatarse que los valores medios son distintos por lo que deben tratarse de una manera distinta ya que no son comparables.

Por otra parte, en la figura 4.3 se observa a su vez la distribución de distancias en las matrices de distancias igualmente para palabras y patentes de las empresas en la industria farmacéutica en donde se encuentra que el máximo de los valores para las distancias de las palabras se sitúa en 0.8 mientras que su media es de 4.2 con una desviación estándar de 7.7 lo que nos muestra que hay una gran variabilidad en las distancias de palabras. El máximo de las distancias entre patentes se encuentra en 1.8 con una media de 4.9 y una desviación estándar de 7.2 que nos indican que los datos tienen una tendencia central de alrededor de 4.9 y una dispersión relativamente grande alrededor de esa media,

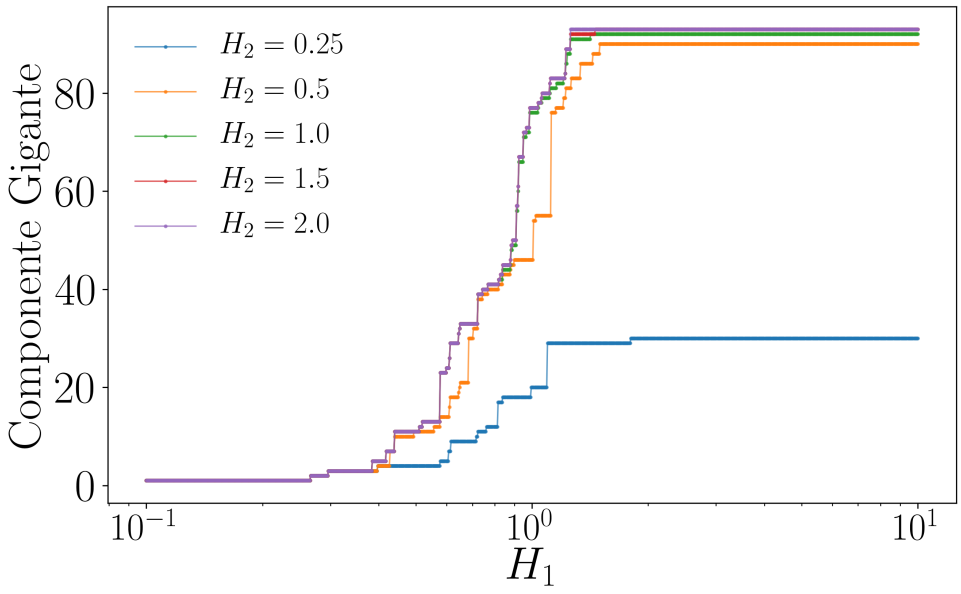


Figura 4.4: Componente gigante de acuerdo con los parámetros umbrales de palabras y patentes para las empresas de tecnología. Se presenta el comportamiento de las empresas de tecnología utilizando distintos valores para los parámetros umbrales de patentes y palabras H_1 y H_2 observando los mayores cambios entre los valores $H_2 = 0.5$ donde su componente gigante es igual a 90 hasta $H_2 = 0.25$ con componente gigante 28. La componente gigante de la red indica la presencia de un subconjunto de empresas que están altamente interconectadas y tienen una gran influencia en la red, por lo tanto, cada una de las H_1 y H_2 que se mencionan muestra diferentes niveles de detalle sobre esta componente gigante y cómo se relacionan las empresas dentro de ella.

tenemos que el 35 % de todos los valores para las distancias entre empresas a través de sus palabras están entre 0.1 y 7.0 mientras que para las patentes se encuentran entre 0.2 y 9.0. En forma similar a lo encontrado para las empresas de tecnología, en las empresas farmacéuticas los valores $D_{patentes}$ y $D_{palabras}$ deben tratarse de manera distinta ya que no son comparables.

Hasta este punto contamos con las matrices de distancias de las cuales fue posible extraer la estadística de cada una de ellas, dado que una matriz de adyacencia es equivalente a una red y ya que el propósito de esta investigación es obtener una red con la que podamos clasificar a las empresas a partir de las matrices que se obtuvieron previamente es necesario el establecer un criterio para saber qué valores serán considerados nulos y no nulos, para esto se considerará que 1 representa cercanía o similitud y 0 lejanía o no similitud. Para saber qué valor de las matrices de distancias representan cercanía o lejanía entre palabras y patentes se propone el utilizar un indicador H_1 o parámetro umbral de similitud para patentes, y H_2 o parámetro umbral de similitud para palabras. En términos de estos valores umbrales, elementos menores a H_1 y H_2 se tomarán como cercanos y se les asignará valor 1, mientras que para valores mayores se les considerará como lejanos y

su valor será 0.

En la figura 4.4 se presentan 5 curvas para las empresas de tecnología obtenidas a partir distintos valores de $0.1 \leq H_1 \leq 10$ y H_2 tomados del conjunto $\{0.25, 0.5, 1.0, 1.5, 2.0\}$. Estos valores fueron elegidos considerando las densidades de probabilidad mostradas en la figura 4.2, considerando que los máximos de distancias entre palabras y patentes se encuentran en 1.0 y 1.7 respectivamente.

En forma similar, en la figura 4.5 se presentan 5 curvas para las empresas de farmacéutica obtenidas a partir distintos valores de $0.1 \leq H_1 \leq 10$ y H_2 tomados del conjunto $\{0.25, 0.5, 1.0, 1.5, 2.0\}$. Estos valores fueron elegidos considerando las densidades de probabilidad mostradas en la figura 4.3, tomando en cuenta que los máximos de distancias entre palabras y patentes es 0.8 y 1.8 correspondientemente.

Cada una de la H mencionadas tanto para las empresas de tecnología como para las empresas de farmacéutica están relacionadas a una componente gigante la cual muestra que tanto se encuentra conectada una red es decir que tanto encuentra similitud entre las empresas, donde mientras cada componente muestra diferentes niveles de detalles sobre cómo se relacionan las empresas, mientras sea más baja corresponde a una menor conexión entre las empresas por lo que se encuentran un menor número de partes conectadas, mientras que una componente mayor corresponde a mayor conectividad global ya que hay una mayor interconexión en las empresas dentro de la red, lo que permite que al variar cada parámetro H_1, H_2 se obtienen distintos valores de componente gigante lo que permite hacer posible encontrar comunidades con caracteristicas en comun; es decir, dependiendo de qué tan profundo se quiera encontrar semejanzas entre comunidades dentro de la red nuestros parámetros umbrales H_1, H_2 deben ser ajustados de manera acorde al nivel de detalle en la información que se quiere acceder. Para las empresas de tecnología en la figura 4.4 se observa el comportamiento de las empresas para diferentes parámetros umbrales de acuerdo con su componente gigante es decir considerando como se encuentran conectados las diversas empresas lo cual nos ayudará a un posterior análisis de redes. Se observa que de acuerdo a cómo varía el parámetro umbral de las palabras, el valor de la componente gigante aumenta al aumentar H_2. En particular, se observa que para un valor de H_2 entre 0.5 y 2.0 los resultados tienen un comportamiento similar, mientras que con $H_2 = 0.5$ el tamaño de la componente gigante es de 89 y al disminuir la H_2 a 0.25, la componente gigante disminuye hasta 28 lo cual muestra una menor conexión entre las empresas que a su vez deberían de tener características similares. Usando el algoritmo modularidad en las componentes gigantes del conjunto de redes obtenidas para diferentes valores de H_1 Y H_2, se encuentran las comunidades que se forman dentro de estas redes.

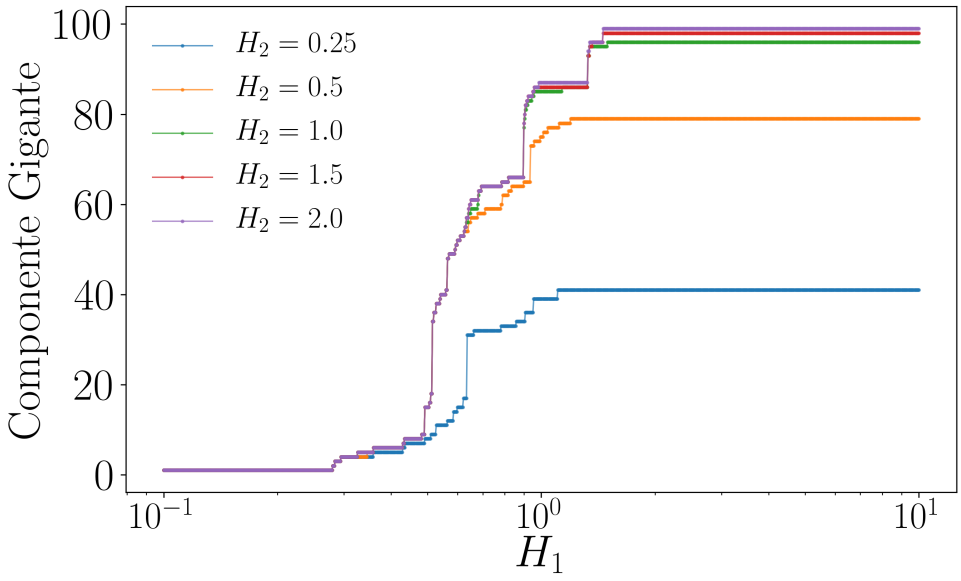


Figura 4.5: Comportamiento de las empresas de farmacéutica utilizando distintos valores para los parámetros umbrales de patentes y palabras H_1 y H_2 observando los mayores cambios entre los valores $H_2 = 1.0$ donde su componente gigante es igual a 94 $H_2 = 0.5$ con componente gigante de 81 y $H_2 = 0.25$ donde su componente gigante es igual a 42. En este caso, la componente gigante indica la presencia de un subconjunto de empresas que están altamente interconectadas y tienen una gran influencia en la red, por lo tanto, cada una de las H_1 y H_2 que se exploran podría mostrar diferentes niveles de detalle sobre esta componente gigante y cómo se relacionan las empresas dentro de ella.

Realizando un análisis similar para las empresas de farmacéutica es posible observar en la figura 4.5 que el valor del parámetro umbral de las palabras entre los valores 1.0 y 2.0 tienen un comportamiento similar, hay un cambio notable en los valores $H_2 = 0.5$ Y $H_2 = 0.25$ donde los valores de la componente gigante cambian desde 80 hasta 40 lo cual es una reducción de la mitad en este valor lo que igualmente que en las empresas de tecnología nos muestra una menor conexión entre empresas.

Mientras que para las empresas de farmacéutica siguiendo un proceso análogo a las empresas de tecnología se encontró que la comunidad 1, empresas en la tabla 4.2(a) de las empresas de farmacéutica, se encuentran compañías que cuentan con un número pequeño de patente al igual que un valor pequeño de entropía lo que nos indica que esta comunidad es de las empresas con menor innovación a diferencia de la comunidad número 2 en la tabla 4.2(b) dónde están las empresas con una mayor innovación como lo son las empresas de Merck, Roche y Bayer, continuando con la comunidad 3 reportadas en la tabla 4.2(c) dónde se encuentran empresas de procedencia japonesa como lo son Hisamitsu Pharmaceutical, Astellas Pharma, Taisho Pharmaceutical, Sumitomo Dainippon Pharma, y Daiichi Sankyo dónde también se encuentran empresas dedicadas a la creación de medicamentos contra trastornos mentales como lo es Biogen y Lundbeck.

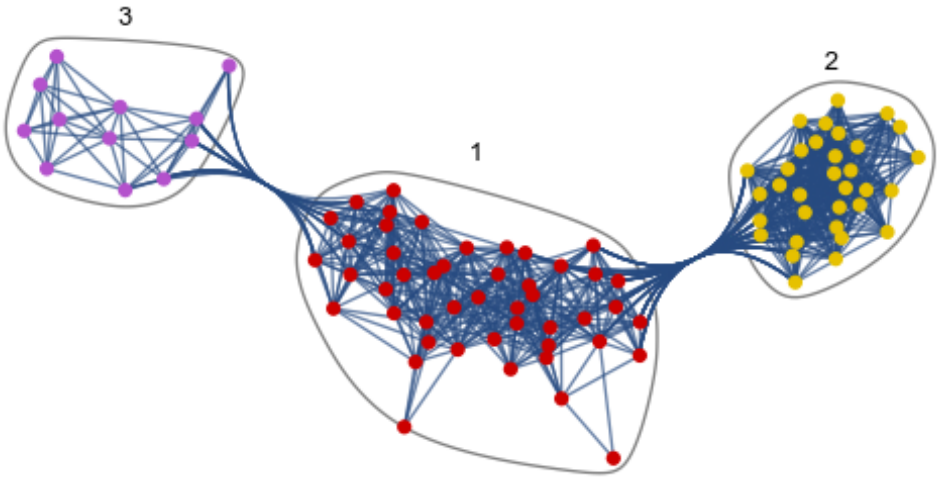


Figura 4.6: Se muestran tres comunidades detectadas a través de la creación de conexiones de acuerdo a su similitud partiendo de los parámetros umbrales de patentes y palabras obtenidas a partir del análisis de patentes de tecnología.

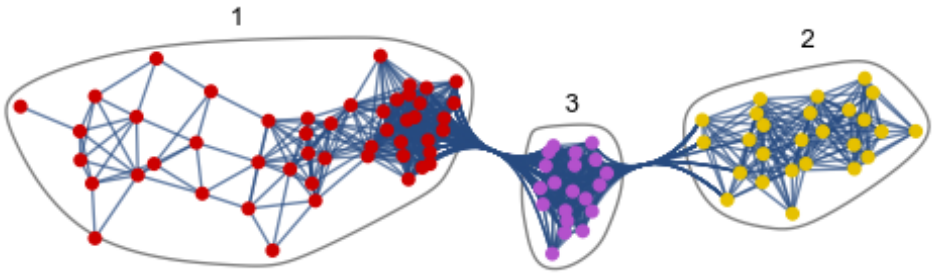


Figura 4.7: Se muestran tres comunidades detectadas a través de la creación de conexiones de acuerdo a su similitud partiendo de los parámetros umbrales de patentes y palabras

Una vez explorado el concepto de umbral y sus valores característicos, es posible utilizar como base los parámetro umbrales H_1 y H_2 de palabras y patentes para tomarlos como un punto de partida creando las matrices de adyacencia y a partir de ellas crear las redes correspondientes para las empresas de tecnología y farmacéutica con parámetros umbrales de $H_1 = 1.7$, $H_2 = 1.0$ y $H_1 = 1.8$, $H_2 = 0.8$ respectivamente. De esta manera, es sencillo convertir las matrices de adyacencia en redes. Se puede observar la red correspondiente a las empresas de tecnología en la figura 4.6 en la cual se tienen las comunidades que se forman con las empresas mediante el método de clustering, es decir la creación de grupos de empresas con características similares. El número de empresas conectadas dentro de alguna de las comunidades es menor al número total de empresas, esto se debe a que hay algunas empresas que no comparten similitudes entre ellas por lo que se encuentran fuera de las comunidades. En este caso se detectan tres comunidades las cuales se pueden ver desglosadas en la tabla 4.1(a), se puede observar la primera comunidad de las empresas de tecnología en donde se encuentran empresas las cuales cuentan con características si-

Tabla 4.1: Comunidad 1 de empresas de tecnología.

(a) Comunidad 1 de empresas de Tecnología					
Activision B.	Advantest	Amd	Amdocs	ASML	Atos
Autodesk	Automatic D. P.	BCE	Boeing	Broadcom	Celestica
Cgi	China Mobile	Cielo	Cognizant	Comcast	Fidelity
Fiserv	Fujifilm	Fujitsu	General E.	Hewlett P. E.	Hon Hai P.
Huawei	IBM	Intel	Intuit	KDDI	Microsoft
Netflix	Nintendo	NVIDIA	Orange	Rogers C.	Salesforce.com
Samsung E.	SAP	Servicenow	Taiwan S.	Tencent	Tesla
Tokyo E.	Verizon C.	Workday	Xiaomi	Zte	

(b) Comunidad 2 de empresas de Tecnología					
Acer	Adobe	Amazon	Analog D.	Apple	Applied M.
AT&T	Canon	Corning	Dell	eBay	Ericsson
Facebook	General E.	Hitachi	Hp	Kyocera	LG
Micron T.	Mitsubishi E.	Nec	Nokia	Phillips	Qualcomm
Rakuten	Red Hat	Siemens	SK Hynix	Sony	Texas I.
Toshiba	VMware	Yahoo			

(c) Comunidad 3 de empresas de Tecnología					
Alphabet	American M.	Asus	China M.	Dassault	DISH
Infosys	Motorola	Murata M.	Telenor	TELUS	Walt Disney

milares al ser empresas que patentan poco y a su vez innovan poco ya que al comparar las empresas de esta comunidad con la figura 3.3 son empresas que tienen una entropía pequeña. Por otra parte, en la comunidad 2 en las tabla 4.1(b) es posible encontrar a las empresas que además de tener a las principales empresas que patentan más como lo son Texas Instruments, Canon, Hp y Amazon , las cuales también son las empresas que innovan más, mientras que en la comunidad 3 tabla 4.1(c) la cual es una comunidad pequeña en donde gran parte de sus empresas están dedicadas a las telecomunicaciones como lo pueden ser América Móvil, China Mobile, DISH, Motorola, Telenor, Telus y Walt Disney. Cada una de estas comunidades tiene sus respectivas características en común, mientras que si se requiriera una mayor granularidad se podrían crear más comunidades variando los parámetros umbrales y por ende cada comunidad tendría mayores semejanzas entre sí.

Es posible crear comunidades más pequeñas en las cuales las similitudes se vean mayormente resaltadas como es el caso de las comunidades número 3 de las empresas de tecnología y farmacéutica tomando como consideración el cambio de los parámetros um-

Tabla 4.2: Comunidades encontradas en empresas de farmacéutica.

(a) Comunidad 1 de empresas en la industria Farmacéutica					
Alexion	Alfa Sigma	Amneal P.	Aurobindo P.	Bausch H.	Biomarin P.
Cadila H.	Celgene	Chiesi P.	Cipla	CSPC P.	Dr. Reddys L.
Endo I.	Ferring	Glenmark P.	Green C. H.	Grifols	Grunenthal
Hikma P.	Horizon P.	Humanwell H.	Intas P.	Jazz P.	Jiangsu H.
Krka	Kyowa H. K.	LEO P.	Livzon P.	Lupin	Mallinkrodt
Menarini	Merc&Co	Merz P.	Mitsubishi T. P.	MundiPharma I.	Mylan
Nichi-Iko P.	Octapharma	Perrigo	Purdue P.	Recordati	Regeneron
Santen P.	Sawai P.	Servier	Stada A.	Sun P. I.	UCB
United T.	Vifor P.	Zhejiang H. P.			

(b) Comunidad 2 de empresas en la industria Farmacéutica					
Abbott	AbbVie	Alexion	Amgen	AstraZeneca	Baxter I.
Bayer	Boehringer I.	Bracco	Eisai	Eli Lilly	GlaxoSmithKline
Incyte	IPsen	J&J	Kowa	Merck	Novartis
Novo Nordisk	Ono P.	Orion	Otsuka	Pfizer	Pierre Fabre
Richter G.	Roche	Sanofi	Shanghai P.	Sichuan K. P.	Takeda
Teijin	Teva	Vertex	Yuhan		

(c) Comunidad 3 de empresas en la industria Farmacéutica					
Angelini	Astellas P.	Biogen	CSL	Daiichi S.	Fresenius SE & Co
Gilead	Hisamitsu P.	Lundbeck	Sumitomo D.	Taisho P.	

brales tanto para patentes como para palabras.

Los resultados obtenidos por medio de la detección de comunidades en redes de similitud permiten identificar patrones a partir de los títulos de las patentes de las empresas más importantes de tecnología y farmacéutica en el mundo según la revista Forbes en el año 2019, utilizando técnicas propias de la ciencia de datos como lo son algoritmos de detección de comunidades como el algoritmo de agrupamiento jerárquico, algoritmo de Ravasz y algoritmo de Girvan-Newmann, así como conceptos de física como lo es la entropía de Shannon.

Este tipo de técnicas nos muestran que a través de ellas es posible observar incrementos o decrementos en la innovación de las empresas en ciertos lapsos de tiempo, así como la detección de grupos de empresas con índices de innovación similares.

Conclusiones

Los análisis desarrollados en este trabajo de investigación muestran que, con la ayuda de la entropía de Shannon aplicada a la distribución de palabras en los títulos de las patentes, es posible identificar las empresas con una mayor diversidad de patentes a lo largo de su historia lo que podría verse reflejado en la innovación de dichas empresas.

En el caso de empresas de tecnología, se observa mayor diversidad de palabras y patentes en las empresas Siemens, Texas Instruments, Hewlett-Packard y Amazon, mientras que para las empresas de farmacéutica son Roche, Novartis, Abbott y Bayer, lo que puede indicar una mayor innovación dentro de las mismas, igualmente es posible identificar que no necesariamente el tener una gran cantidad patentes implica que la empresa innove mucho ya que existen casos que muestran lo contrario.

Por otra parte, la aplicación de la distancia Euclidiana en las curvas que describen a las patentes y las diferentes palabras en el tiempo, es el punto de partida con el cual se puede encontrar semejanzas o diferencias entre empresas con lo cual al generar redes se reporta la aparición de comunidades las cuales cada una guarda ciertas características entre las empresas que la comprenden. El número de comunidades varía con respecto a las componentes gigantes definidas por los parámetros umbrales de patentes y palabras, las características de las comunidades se encuentran relacionadas con las palabras que comprenden las empresas dentro de estas comunidades.

Como conclusión general, fue posible observar que gracias al comportamiento que siguen las patentes de las empresas de tecnología y farmacéutica, con ayuda del método utilizado (la entropía de Shannon y el procesamiento de lenguaje natural) se encontraron relaciones entre las patentes y la innovación. Por consiguiente el objetivo se cumple totalmente ya que es posible observar el comportamiento de la distribución de palabras y patentes dando características de la innovación de las mejores empresas de tecnología y farmacéutica del mundo.

Perspectivas a futuro

Como se mencionó en el presente trabajo, la intención primordial fue hacer un estudio de las patentes de las mejores empresas de tecnología y farmacéutica del mundo y cómo influyen estas en la innovación de dichas empresas utilizando métodos de la ciencia de redes. A futuro este método de análisis podría ser implementado en otras industrias, lo cual ayudaría a conocer cuáles son las industrias y empresas más innovadoras. Por otro lado, podría implementarse a su vez al análisis del comportamiento de las empresas en la banca pudiendo visualizarlas por su valor a lo largo de su historia.

Dentro de posteriores trabajos se podría profundizar en el tema con la ayuda de algunos otros conceptos referentes al tema como lo serían la entropía mutua, la utilización de redes pesadas o el trabajar con citas de las patentes.

También es importante recalcar que los desarrollos de esta investigación pueden ser aplicables en otros ámbitos como lo puede ser la música, es decir el análisis de las letras de las canciones y a través de ellas observar las repercusiones que pueden tener en cuanto a su éxito partiendo de la información obtenida por diversas canciones con estructuras similares entre sí.

Apéndice

Tabla A1: Empresas de tecnología. Se muestran las empresas de tecnología con las que se trabajará, así como su la nación en la que se encuentran registradas.

Empresas de tecnología					
Empresa	Nacionalidad	Empresa	Nacionalidad	Empresa	Nacionalidad
Acer	Taiwanesa	eBay	Estadounidense	Nokia	Finlandesa
Activision B.	Estadounidense	Ericsson	Sueca	NVIDIA	Estadounidense
Adobe	Estadounidense	Facebook	Estadounidense	Orange	Francesa
Advantest	Japonesa	Fidelity	Estadounidense	Philips	Holandesa
Amazon	Estadounidense	Fiserv	Estadounidense	Qualcomm	Estadounidense
Amd	Estadounidense	Fujifilm	Japonesa	Rakuten	Japonesa
Amdocs	Israelí	Fujitsu	Japonesa	Red Hat	Estadounidense
Analog D.	Estadounidense	General E.	Estadounidense	Rogers C	Canadiense
Apple	Estadounidense	Google	Estadounidense	Salesforce.com	Estadounidense
Applied M.	Estadounidense	Hewlett P. E.	Estadounidense	Samsung E.	Surcoreana
ASML	Holandesa	Hitachi	Japonesa	Servicenow	Estadounidense
Atos	Francés	Hon Hai P.	Taiwanesa	Siemens	Alemana
AT&T	Estadounidense	Hp	Estadounidense	SK Hynix	Surcoreana
Autodesk	Estadounidense	Huawei	China	Sony	Japonesa
Automatic D. P.	Estadounidense	IBM	Estadounidense	Taiwan S.	Taiwanesa
BCE	Estadounidense	Intel	Estadounidense	Tencent	China
Boeing	Estadounidense	Intuit	Estadounidense	Tesla	Estadounidense
Broadcom	Estadounidense	KDDI	Japonesa	Texas I.	Estadounidense
Canon	Japonesa	Kyocera	Japonesa	Tokyo E.	Japonesa
Celestica	Canadiense	LG	Surcoreana	Toshiba	Japonesa
Cgi	Canadiense	Micron T.	Estadounidense	Verizon C.	Estadounidense
Cielo	Brasileña	Microsoft	Estadounidense	VMware	Estadounidense
Cisco S.	Estadounidense	Mitsubishi E.	Japonesa	Workday	Estadounidense
Comcast	Estadounidense	Nec	Japonesa	Xiaomi	China
Corning	Estadounidense	Netflix	Estadounidense	Yahoo	Estadounidense
Dell T.	Estadounidense	Nintendo	Japonesa	Zte	China

Tabla A2: Empresas de farmacéutica. Se muestran las empresas de Farmacéutica con las que se trabajará, así como su la nación en la que se encuentran registradas.

Empresas farmacéuticas					
Empresa	Nacionalidad	Empresa	Nacionalidad	Empresa	Nacionalidad
Abbott	Estadounidense	GlaxoSmithKline	Británica	Octapharma	Suiza
AbbVie	Estadounidense	Grifols	Española	Ono P.	Japonesa
Alexion	Estadounidense	Grunenthal	Alemana	Otsuka	Japonesa
Allergan	Irlandesa	Hikma P.	Jordana	Perrigo	Irlandesa
Amgen	Estadounidense	Hisamitsu P.	Japonesa	Pfizer	Estadounidense
Angelini	Italiana	Humanwell H.	China	Pierre fabre	Francesa
Astellas P.	Japonesa	Incyte	Estadounidense	Purdue Pharma	Estadounidense
AstraZeneca	Británica	Intas P.	India	Recordati	Italiana
Aurobindo P.	India	IPsen	Francesa	Regeneron	Estadounidense
Bausch H.	Canadiense	Jazz P.	Irlandesa	Richter gedeon	Húngara
Baxter I.	Estadounidense	Jiangsu H.	China	Roche	Suiza
Bayer	Alemana	J&J	Estadounidense	Sanofi	Francesa
Biogen	Estadounidense	Kowa	Japonesa	Santen P.	Japonesa
Biomarin P.	Estadounidense	Krka	Eslovena	Sawai P.	Japonesa
Boehringer I.	Alemana	Kyowa H.	Japonesa	Servier	Francesa
Bracco	Italiana	LEO Pharma	Danesa	Shanghai P.	China
Celgene	Estadounidense	Livzon P.	India	Sichuan K.	China
Cipla	India	Lundbeck	Danesa	Stada A.	Alemana
CSL	Australiana	Lupin	India	Sumitomo D.	Japonesa
CSPC	China	Menarini	Italiana	Sun P.	India
Daiichi S.	Japonesa	Merc&Co	Estadounidense	Taisho P.	Japonesa
Dr. Reddys	India	Merck	Alemana	Takeda	Japonesa
Eisai	Japonesa	Merz Pharma	Alemana	Teijin	Japonesa
Eli Lilly	Estadounidense	Mitsubishi T. P.	Japonesa	Teva	Israelí
Ferring	Suiza	Mylan	Estadounidense	UCB	Belga
Fresenius SE & Co	Alemana	Novartis	Suiza	Vertex	Estadounidense
Gilead	Estadounidense	Novo Nordisk	Danesa	Yuhan	Surcoreana

Bibliografía

[1] N. James and M. Menzies. Cluster-based dual evolution for multivariate time series: Analyzing covid-19. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):061108, 2020.

[2] C. G. Pereira, V. Picanco-Castro, D. T. Covas, and G. S. Porto. Patent mining and landscaping of emerging recombinant factor viii through network analysis. *Nature Biotechnology*, 36(7):585–590, August 2018.

[3] A. Barrat, M. Barth elemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, Cambridge, Octubre 2008.

[4] V. Cunha, M. do, C. C. Ribeiro Santos, M. A. Moret, and H. B. de Barros Pereira. Shannon entropy in time-varying semantic networks of titles of scientific paper. *Applied Network Science*, 5(1):53, August 2020.

[5] D. d. l. C. R. H. Manuel Alexander Molina Espinosa. Redes complejas. teoría y práctica. *Revista Académica de Investigación Tlatemoani*, 11:14, 2012.

[6] A. Bergeaud, Y. Potiron, and J. Raimbault. Classifying patents based on their semantic content. *PLOS ONE*, 12(4):e0176310, 04 2017.

[7] J. P. Sáiz. Investigación y desarrollo: patentes. *UAM*, 2005.

[8] L. C. Quevedo Cerpa. *Acuerdo sobre los aspectos de los derechos de propiedad intelectual relacionados con el comercio y control aduanero de los derechos de propiedad intelectual en la frontera : el escenario de control aduanero respecto a la violación de marcas y patentes en Colombia.* Bogotá : Universidad Externado de Colombia, 2020., 2020.

[9] J. P. S. Gonzáles. *Propiedad industrial y revolución liberal*. Oficina Española de Patentes y Marcas, 1995.

[10] P. P. y marcas. Historia de las patentes. https://www.protectia.eu/2013/11/historia-de-las-patentes/, 2010. Accedido en Octubre de 2021.

[11] M. D. y José Ángel. Una breve historia del origen de las patentes. *(INCAR)*, 2018.

[12] IMPI. ¿qué es el impi? https://www.gob.mx/impi/acciones-y-programas/conoce-el-impi-que-es-el-impi, 2021. Accedido en Octubre de 2021.

[13] pixabay. pixabay. https://pixabay.com/es/, 2021. Accedido en Enero de 2022.

[14] E. R. Garcia. Estudio sobre patentes y dominio público. *Revista la propiedad inmaterial*, (n°15):127–142, 2011.

[15] Economipedia. Patentes. https://economipedia.com/definiciones/patente.html, 2021. Accedido en Octubre de 2021.

[16] O. O. mundial de la propiedad intelectual. Patentes conceptos básicos. https://www.wipo.int/patents/es/faq_patents.html, 2021. Accedido en Octubre de 2021.

[17] Economipedia. Innovación. https://economipedia.com/definiciones/innovacion-2.html, 2021. Accedido en Octubre de 2021.

[18] D. M. Tirado. *Fundamentos de marketing*. Universitat Jaume, Universitat Jaume, España, 2013.

[19] Economipedia. La innovación y sus ámbitos de implementación. https://economipedia.com/definiciones/innovacion-2.html, 2021. Accedido en Octubre de 2021.

[20] C. I. Rodriguez. Innovación incremental e innovación radical o disruptiva y sus ejemplos. https://www.eoi.es/blogs/carollirenerodriguez/2012/03/08/innovacion-incremental-e-innovacion-radical-o-disruptiva-y-sus-ejemplos/, 2021. Accedido en Octubre de 2021.

[21] G. van Rossum and the Python development team. *The Python Language Reference*. Python Software Foundation, Python Software Foundation, March 05, 2017.

[22] J. VanderPlas, B. E. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Interactive statistical visualizations for python. *The Journal of Open Source Software*, 3(32):1057, October 2018.

[23] S. Tosi. *Matplotlib for Python Developers*. Packt Publishing Ltd 32 Lincoln Road Olton Birmingham B27 6PA UK, Matplotlib for Python Developers, March 05, 2017.

[24] F. López and V. Romero. *Mastering Python Regular Expressions*. Packt Publishing Ltd, Mastering Python Regular Expressions, 2014.

[25] F. Nelli. *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Apress, Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2018.

[26] A. Martelli. *Python in a Nutshell: A Desktop Quick Reference*. O'Reilly, Python in a Nutshell: A Desktop Quick Reference, 2017.

[27] J. Perkins. *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd, Python 3 Text Processing with NLTK 3 Cookbook, 2014.

[28] S. Molin. *Data Analysis with Pandas*. Packt Publishing Ltd, Data Analysis with Pandas, 2019.

[29] G. V.-O. R. López-Carreño. Google patents versus lens: citaciones de literatura científica en patentes. *Revista General de Información y Documentació*, Abril 2021.

[30] ayudaley. Google patents: La herramienta para buscar patentes de google. `https://ayudaleyprotecciondatos.es/patentes-marcas/google/`, 2021. Accedido en Octubre de 2021.

[31] F. Staff. Las compañías tecnológicas más grandes del mundo 2020. may https://www.forbes.com.mx/listas-companias-tecnologicas-grandes-mundo-2020-apple-mantiene-cima/ 2020.

[32] J. A. T. Thomas M. Cover. *"Elements of Information Theory"*. John Wiley & Sons, Second Edition 2006.

[33] J. Rincón-López, Y. C. Almanza-Arjona, A. P. Riascos, and Y. Rojas-Aguirre. When cyclodextrins met data science: Unveiling their pharmaceutical applications through network science and text-mining. *Pharmaceutics*, 13(8):1297, 2021.

[34] J. U. Martínez-González and A. P. Riascos. Activity of vehicles in the bus rapid transit system Metrobús in Mexico City. *Scientific Reports*, 12(1):98, 2022.

[35] R. E. B. C. E. Sandifer. *Leonhard Euler: Life, Work and Legacy*. Elsevier, Studies in the History and Philosophy of Mathematics, 2007.

[36] Collective dynamics of 'small-world' networks. department of theoretical and applied mechanics, watts, d.j.; strogatz, cornell university, ithaca, new york 14853, usa. 1998.

[37] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, Cambridge, 2008.

[38] M. Newman. *Networks : an introduction*. Oxford, 2010.

[39] M. Newman. *Random graphs as models of networks*. Bornholdt & H. G. Schuster, 2003. pp. 35-68.

[40] J. C. J. Marro, J.J. Torres and S. de Franciscis. *Complex networks with time-dependent connections and silent nodes*. Pello and Romance Edts, 2007.

[41] A.-L. Barabási. *Network science*. Cambridge University Press, Cambridge, 2016.

[42] M. van Steen. *Graph Theory and Complex Networks*. Maarten Van Steen, 2010.

[43] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[44] M. E. Newman. Fast algorithm for detecting community structure in networks. *Department of Physics and Center for the Study of Complex Systems*, Septiembre 2003.

[45] G. C. Nuwan Ganganath and C. Cheng. *Detecting Hierarchical and Overlapping Community Structures in Networks*. IEEE, 2020.

[46] C. Donnat and S. Holmes. Tracking network dynamics: A survey using graph distances. *Ann. Appl. Stat.*, 12(2):971 – 1012, 2018.

[47] A. J. Alvarez, C. E. Sanz-Rodríguez, and J. L. Cabrera. Weighting dissimilarities to detect communities in networks. *Phil. Trans. R. Soc.*, 373(2056):20150108, 2015.

[48] F. B. Palacio Niño, Julio Omar Ancízar. *Deteccion de comunidades*. Departamento de Ciencias de la Computacion e I.A., 2013.

[49] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[50] V. Boss. *Lecciones de matemática tomo 1 Análisis*. Editorial URSS, URSS, Moscú (2007).