



**COMILLAS**

UNIVERSIDAD PONTIFICIA

ICAI

ICADE

CIHS

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN INGENIERÍA INDUSTRIAL

# VIGILANCIA TECNOLÓGICA POR BIG DATA DE PATENTES Y ESPIONAJE INDUSTRIAL

JUAN CARLOS GARCÍA LÓPEZ


TRABAJO DE FÍN DE MÁSTER

Madrid

Julio 2020



Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título  
**VIGILANCIA TECNOLÓGICA POR BIG DATA DE PATENTES Y ESPIONAJE INDUSTRIAL**  
en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el  
curso académico 2019/2020 es de mi autoría, original e inédito y  
no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni  
parcialmente y la información que ha sido tomada  
de otros documentos está debidamente referenciada.

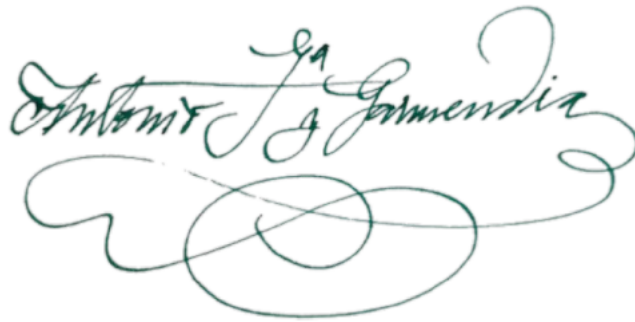


Fdo.: Juan Carlos García

Fecha: 15 / 07 / 2020

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO



Fdo.: Ph.D. Antonio García de Garmendia

Fecha: 15 / 07 / 2020

# **TECHNOLOGICAL SURVEILLANCE OF PATENTS AND INDUSTRIAL ESPIONAGE USING BIG DATA**

**AUTHOR: García López, Juan Carlos**

ADVISOR: Ph.D. Antonio García de Garmendia

## **SUMMARY OF THE PROJECT:**

### **Introduction:**

We currently live and move in highly changing VUCA environments that require constant adaptation. Therefore, innovation is already a mandatory value in organizations. To innovate, companies are not happy enough with their own experience, since they are aware that the central pillar and prior to innovation is extensive and comprehensive technological information. From it, it is about building by exploring, analyzing, identifying and mixing, since the generation of ideas is but a combination of some previous ones. This will undoubtedly facilitate strategic, operational and tactical decision-making in companies, businesses and governments, which makes technological information an essential pillar for innovation.

The sources of information that companies, organizations and governments routinely use are numerous: direct contacts with stakeholders, scientists, engineers, entrepreneurs, professionals in a specific technical sector, professional and scientific literature, participation in events, conferences, academic theses and patents. However, very often the latter and their trends are not estimated, as sources of information to prospect. At the moment, they are not common instruments in all innovative areas of research. Perhaps due to ignorance, since, without a doubt, they can be a stimulus for new ideas, or they can respond to different technical problems. This is valuable information. One of the most

complete, accessible, manageable, practical and updated, available to different users.

Among the advantages of using patent documentation and its tendencies for decision-making at different levels, is the fact that it is up-to-date information recently published. As they are not documents edited for commercial purposes, the highest percentage of information has not been published previously, so being able to have it in real time is of value for management and business strategy. Another advantage that it offers is that the documentation is mostly available on the internet, so it can be accessed in an accessible way when necessary.

Trend information is valuable information for its possible uses for decision-making in marketing, risk analysis, or strategic planning in the field of research and development. Through it, we can access knowledge and updates on technology uses, companies and patenting organizations, emerging sectors, sectors in decline, etc. All of this is, without a doubt, a facilitating element for the industry and administrations in making strategic, tactical and operational decisions. It is also for the planning of activities, as well as for monitoring the evolution of certain sectors, and not only for monitoring, but for forecasting and planning the process of technological development, as well as for conducting economic and scientific research. and technological for different purposes.

Among other uses of patent information that affect decision-making by are: legal purposes, technological purposes that include efficiency in solving problems based on previous experiences reflected, aid to R&D plans and technological development, promotion of innovation policies, control of lines of research in competition and justification before decision makers of certain financial investments.

The analysis of information on patent trends can therefore be considered a flow of knowledge, an effective tool that can undoubtedly influence the economic use of science and research for the development and empowerment of businesses and public bodies.

Knowledge of trends is undoubtedly part of the management of Competitive Intelligence tools that help the implementation of innovative processes and economic and social improvements. In the areas of management, the search and location of relevant information to define the strategy, plan and manage companies and businesses are key aspects.

We can include its analysis as part of the Competitive Intelligence of the different entities, as it focuses on the process of obtaining, analyzing, interpreting and disseminating information of strategic value about the industry, competitors, innovative sectors, etc. Each organization must

identify the relevant and key information for it, obtained from the analysis of trends that will have different priority according to each case. Examples include identifying emerging technology that may pose a threat or an opportunity to react in a timely manner if appropriate, knowing what the competition is doing, and having information about a potential market.

There is no doubt, therefore, that developing a model for handling this information and having this relevant information is part of the broad fabric of competitive intelligence in today's business world by answering questions such as: What companies are patenting? In which sectors? What technology is emerging? In which countries? In addition, it will allow identifying possible strategies of the competitors detecting opportunities, facilitating the anticipation of future market and customer needs, as well as knowing the trends of certain sectors and being updated on technological innovations that may cause a change in the market, in customers and in providers.

## **Objectives**

The main objective of this thesis is to propose a methodology, associated with the current techniques of extraction, transformation and data loading through Machine Learning, in order to analyze trends in patent applications and have advanced analytics to help us in the Knowledge for decision-making on Competitive Surveillance. This objective can be broken down in depth into the following:

1. Review the State of the Art regarding the analysis of industrial property data
2. Provide a tailored, scalable and cost-effective solution to the main concerns of companies and governments in the field of Technology Surveillance
3. Optimize the use of public access data from the EPO
4. Apply data cleaning and analysis processes in a practical way

## **Methodology**

The methodology applied in this work is based on five different steps:

1. Work Plan: At this point the different stages of the work were established, with their clear objectives (deliverables) for each stage.

2. Research on the “State of the Art” in the area of Patent Analytics, developing the current processes related to patentability and patent applications. Texts, articles, books, web pages and data sources were included fulfilling the following criteria: Belonging, Updating and Internationality.

3. Approach to the problems to be studied in the Work: Present theoretical framework and cases and explanatory examples of the current situation.

4. Choice of algorithm: to predict and describe output data, using Artificial Intelligence or more specifically Machine Learning. It was necessary to train in ML techniques, attending face-to-face and "on-line" at different courses and seminars and in particular those taught by Stanford University. Priority was given to the intention of “clustering”, as far as possible, patent information and determining trends based on this “clustering”. Thus, work had to be done in the field of “Unsupervised Learning” algorithms in their different variants, each with a specific applicability.

5. Obtaining data: an analysis was carried out of the main sources of patent data currently available and an exhaustive comparison was made of them to define from which to obtain the information that would later feed the algorithm.

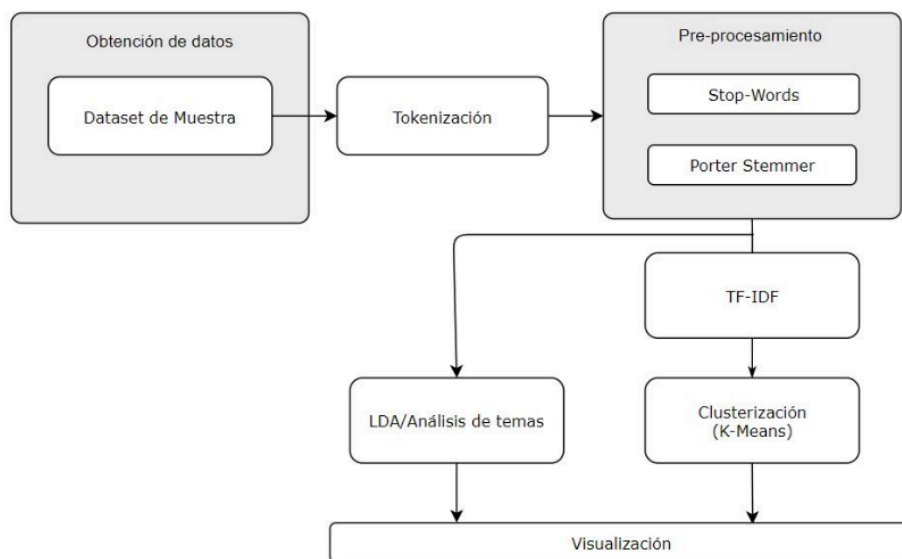


Figura 1. Flowchart of string type data in Patent Radar

## Results

Clusters of three terms were generated with the sole objective of providing added value and not generating useless clusters for the investigation, therefore avoiding that the clustering was limited

to generic terms (i.e. invention, cure ...). In Figure 1, the 15 trinomial clusters generated by the K-Means algorithm are shown. From these, as is evident, generic concepts emerge, but clusters 3 (sequence, nucleic, acid), 5 (receptor, antigen, cell) and 7 (vaccine, immunogenic, polypeptide) deserve special mention due to their specific nature and contribution. added value to identifying trends.

Identificador del clúster	Término 1	Término 2	Término 3
1	computing	group	data
2	composition	method	sample
3	sequence	nucleic	acid
4	vaccine	influenza	virus
5	receptor	antigen	cell
6	formula	salt	compound
7	vaccine	immunogenic	polypeptide
8	gripping	step	container
9	prevention	upper	infection
10	domain	binding	ligand
11	inhibitor	treatment	disease
12	treatment	formulation	combination
13	vaccine	composition	pharmaceutical
14	particle	vector	protein
15	specifically	bind	antibody

Figura 1. Clústeres obtenidos al aplicar Patent Radar a la búsqueda de tendencias en I< lucha contra el COVID-19

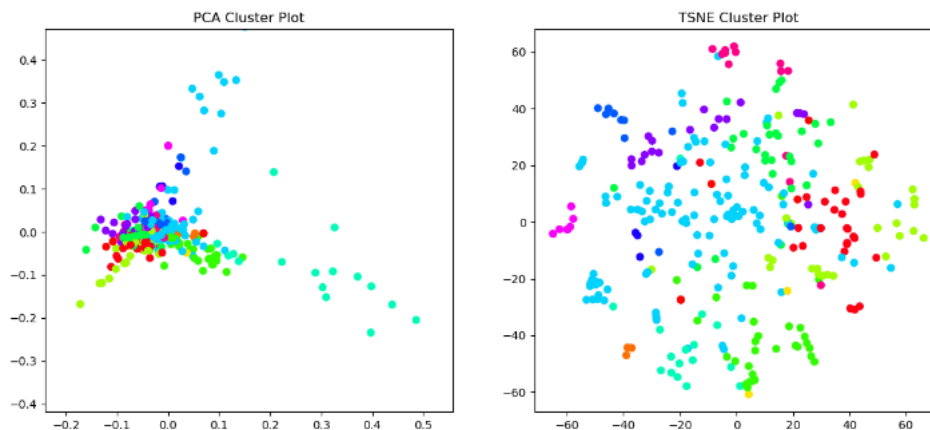


Figura 2. Clústeres obtenidos al aplicar Patent Radar a la búsqueda de tendencias en I< lucha contra el COVID-19

## Conclusions

In this report, a new way of approaching Technological Surveillance has been presented by analyzing trends in patents based on the analysis of all the fields of these. The method of controlling the research lines of the competition was formalized based on K-Means as a machine learning algorithm for word processing. This allowed to develop a framework for processing Abstracts and



Patent Titles, which leads to a new paradigm in the domain of competitive information when it comes to directing the innovation of a company or government. A series of algorithms were developed and an efficient architecture in the treatment of these was created by creating Patent Radar. Here are the most relevant conclusions:

1. Patents must become a central element of innovation
2. Patent Radar is useful in the fight against COVID-19
3. Patent Radar provides added value as an integrated innovation management platform
4. K-Means as NLP algorithm for topic processing is suitable for patent treatment
5. NLTK and Gensim as parallel Abstracts preprocessing algorithms

## **References**

- [1] Oldham, Paul. *WIPO*, "WIPO Manual on Open Source Patent Analytics" May 2020.  
Available at: <https://wipo-analytics.github.io/introduction.html>

# VIGILANCIA TECNOLÓGICA POR BIG DATA DE PATENTES Y ESPIONAJE INDUSTRIAL

**AUTOR: García López, Juan Carlos**

DIRECTOR: Ph. D. Antonio García de Garmendia

## RESUMEN DEL PROYECTO:

Actualmente vivimos y nos movemos en entornos VUCA<sup>1</sup> altamente cambiantes y que requieren de una constante adaptación a los mismos. Por consiguiente, la innovación es ya un valor de obligado cumplimiento en las organizaciones. Para innovar, a las empresas no les es suficiente su propia experiencia, ya que son conscientes de que el pilar central y anterior a la innovación es la amplia y exhaustiva información tecnológica. A partir de ella, se trata de edificar explorando, analizando, identificando y mezclando, ya que la generación de ideas no es sino una combinación de algunas previas. Ello sin duda facilitará la toma de decisiones estratégicas, operativas y tácticas en empresas, negocios y gobiernos, lo que hace que, a la información tecnológica, sea un pilar esencial para la innovación.

Las fuentes de información que habitualmente utilizan empresas, organizaciones y gobiernos son numerosas: contactos directos con *stakeholders*, científicos, ingenieros, empresarios, profesionales en un sector técnico específico, la literatura profesional y científica, participación en eventos, congresos, tesis académica y patentes. No obstante, muy frecuentemente no son estimadas estas últimas y sus tendencias, como fuentes de información a prospectar. No son por el momento, instrumentos comunes en todos los ámbitos innovadores de investigación. Quizás por desconocimiento, ya que, sin duda, pueden ser estímulo de nuevas ideas o pueden dar respuesta a diferentes problemas técnicos planteados. Se trata de una información de valor. Una de las más

---

<sup>1</sup> VUCA (Volatility, Uncertainty, Complexity y Ambiguity) hace referencia a entornos volátiles, inciertos, complejos y ambiguos

completas, accesibles, manejables, prácticas y actualizadas, disponibles para los diferentes usuarios.

Entre las ventajas de la utilización de la documentación de patentes y sus tendencias para la toma de decisiones de diferente nivel, se encuentra el tratarse de una información actualizada de reciente publicación. Al no ser documentos editados con fines comerciales, el mayor porcentaje de información no se ha publicado con anterioridad, por lo que poder disponer de la misma en tiempo real, es de valor para la estrategia de la gestión y los negocios. Otra de las ventajas que ofrece, consiste en que la documentación está disponible en su gran mayoría en internet, por lo que puede ser consultada de forma accesible cuando sea necesario.

La información sobre tendencias es una información de valor por sus posibles usos para la toma de decisiones en el terreno del marketing, análisis de riesgos, o la planificación estratégica en el ámbito de la investigación y desarrollo. A través de ella, podemos acceder a conocimiento y actualización sobre usos de tecnología, empresas y organismos que patentan, sectores emergentes, sectores en decadencia etc. Todo ello, es sin duda un elemento facilitador para la industria y las administraciones en la toma de decisiones estratégicas, tácticas y operativas. También lo es para la planificación de actividades, así como para el seguimiento de evolución de determinados sectores, y no sólo para el seguimiento, sino para la previsión y la planificación del proceso de desarrollo tecnológico, así como para la realización de investigaciones económicas, científicas y tecnológicas con diferentes fines.

Entre otros usos de la información de patentes que afectan a la toma de decisiones por se encuentran: fines legales, fines tecnológicos que incluyen la eficiencia en la resolución de problemas por experiencias previas reflejadas, ayuda a los planes de I+D y desarrollo tecnológico, impulso de políticas de innovación, control de líneas de investigación en la competencia y justificación ante tomadores de decisiones de determinadas inversiones financieras.

El análisis de información de las tendencias de patentes puede considerarse por consiguiente un flujo de conocimiento, una herramienta eficaz, que sin duda puede influir en el aprovechamiento económico de la ciencia e investigación para el desarrollo y potenciación de los negocios y organismos públicos.

El conocimiento de tendencias forma parte sin duda, del manejo de las herramientas de Inteligencia Competitiva que ayuda a la puesta en marcha de procesos innovadores y mejoras económicas y sociales. En las áreas de la dirección, son aspectos clave la búsqueda y localización de información relevante para definir la estrategia, planificar y gestionar empresas y negocios.

Podemos incluir su análisis como parte de la Inteligencia Competitiva de las diferentes entidades, al enfocarse la misma en el proceso de obtención, análisis, interpretación y difusión de la información de valor estratégico sobre industria, competidores, sectores innovadores etc. Cada organización deberá identificar aquella información relevante y clave para ella, obtenida del análisis de tendencias que tendrá prioridad diferente según cada caso. Entre algunos ejemplos figuran el identificar tecnología emergente que pueda suponer amenaza u oportunidad para poder reaccionar a tiempo si fuera conveniente, conocer qué hace la competencia y disponer de información sobre un mercado potencial.

Es indudable, por tanto, que el desarrollar un modelo de tratamiento de esta información y disponer de esta información relevante forma parte del amplio entramado de la Inteligencia competitiva en el mundo actual de negocios al dar respuesta a cuestiones como: ¿Qué empresas están patentando?, ¿En qué sectores?, ¿Qué tecnología está emergiendo?, ¿En qué países? Además, permitirá identificar posibles estrategias de los competidores detectando oportunidades, facilitar el prever futuras necesidades del mercado y clientes, así como conocer las tendencias de determinados sectores y estar actualizados en las innovaciones tecnológicas que pueden provocar un cambio en el mercado, en clientes y en proveedores.

## **Objetivos**

El objetivo principal de este trabajo es el plantear una metodología, asociada a las técnicas actuales de extracción, transformación y carga de datos mediante Machine Learning, con el fin de analizar tendencias en solicitudes de patentes y disponer de una analítica avanzada que nos ayude en el conocimiento para la toma de decisiones sobre Vigilancia Competitiva. Dicho objetivo puede desglosarse en profundidad en los siguientes:

1. Revisar el Estado del Arte en materia de analítica de datos de propiedad industrial

2. Dar una solución a medida, escalable y con un coste contenido a las principales inquietudes por parte de empresas y gobiernos en materia de Vigilancia Tecnológica
3. Optimizar del uso de los datos de acceso público de la EPO
4. Aplicar de manera práctica procesos de limpieza y análisis de datos

## **Metodología**

La metodología aplicada en este trabajo se basa en siete pasos diferenciados:

1. Plan de Trabajo: En este punto se fijaron las diferentes etapas del trabajo, con sus objetivos (entregables) claros para cada etapa.
2. Investigación sobre el “Estado del Arte” en materia de Analítica de Patentes desarrollando los procesos actuales relacionados con la patentabilidad y solicitudes de patentes. Se incluyeron textos, artículos, libros, páginas web y fuentes de datos cumpliendo con los siguientes criterios: Pertenencia, Actualidad e Internacionalidad.
3. Enfoque de la problemática a estudiar en el Trabajo: Marco teórico presente y casos y ejemplos explicativos de la situación actual.
4. Elección del algoritmo: para predecir y describir datos de salida, utilizando Inteligencia Artificial o más concretamente Machine Learning. Fue necesario formarse en técnicas de ML, asistiendo presencialmente y “on-line” a diferentes cursos y seminarios y en particular los impartidos por la Universidad de Standford. Se priorizó la intención de “clusterizar”, en la medida de lo posible, la información de patentes y determinar tendencias en función de esta “clusterización”. Así pues, se tuvo que trabajar en el ámbito de los algoritmos de tipo “Unsupervised Learning” en sus diferentes variantes, cada una de ellas con una aplicabilidad concreta.
5. Obtención de datos: se llevó a cabo un análisis de las principales fuentes de datos de patentes disponibles en la actualidad y se hizo una comparación exhaustiva de ellas para definir de cuales obtener la información que nutriría más adelante al algoritmo.

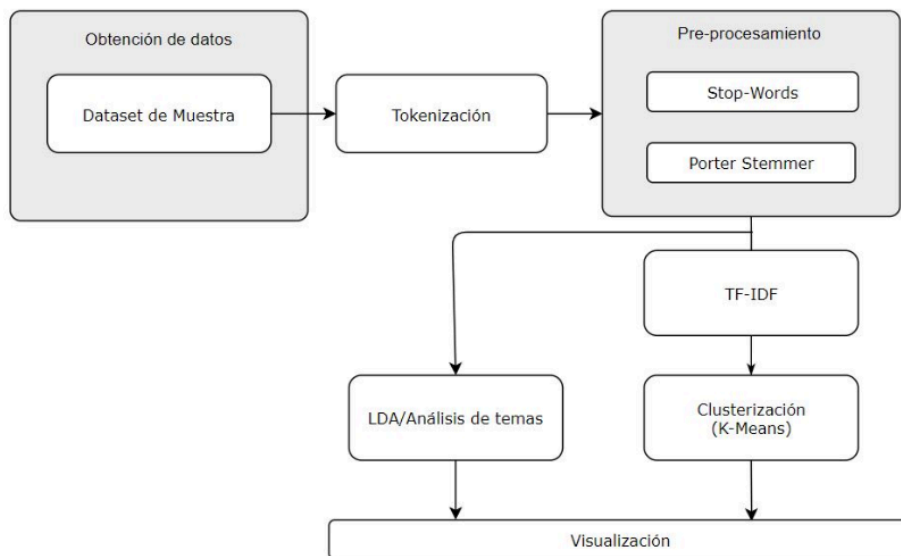


Figura 1. Diagrama de flujo de los datos de tipo string a través de Patent Radar

## Resultados

Se generaron clústeres de tres términos con el único objetivo de aportar valor añadido y no generar clústeres inútiles para la investigación, evitando por tanto que la clusterización se limitase a términos genéricos (i.e. invention, cure...). En la Figura 1 se muestran, los 15 clústeres trinomiales generados por el algoritmo K-Means. De ellos, como es evidente, surgen conceptos genéricos, pero merecen mención especial los clústeres 3 (sequence, nucleic, acid), 5 (receptor, antigen, cell) y 7 (vaccine, immunogenic, polypeptide) por su carácter específico y su aporte de valor añadido a la identificación de tendencias.

Identificador del clúster	Término 1	Término 2	Término 3
1	computing	group	data
2	composition	method	sample
3	sequence	nucleic	acid
4	vaccine	influenza	virus
5	receptor	antigen	cell
6	formula	salt	compound
7	vaccine	immunogenic	polypeptide
8	gripping	step	container
9	prevention	upper	infection
10	domain	binding	ligand
11	inhibitor	treatment	disease
12	treatment	formulation	combination
13	vaccine	composition	pharmaceutical
14	particle	vector	protein
15	specifically	bind	antibody

Figura 1. Clústeres obtenidos al aplicar Patent Radar a la búsqueda de tendencias en la lucha contra el COVID-19

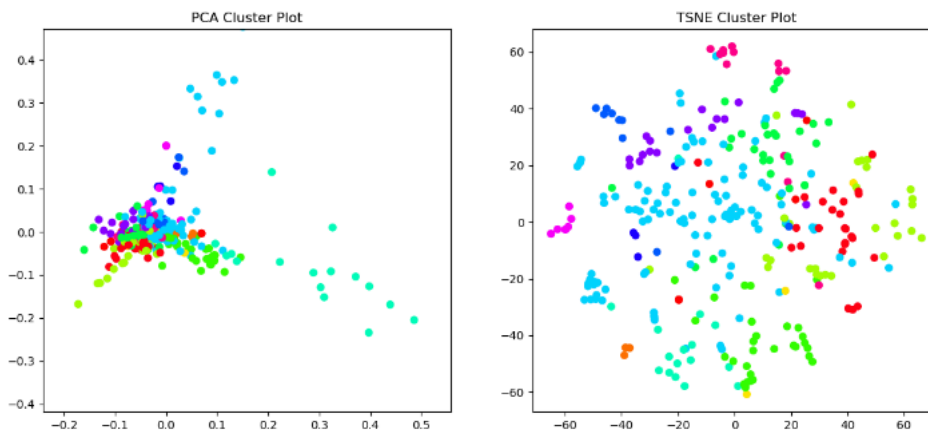


Figura 2. Clústeres obtenidos al aplicar Patent Radar a la búsqueda de tendencias en la lucha contra el COVID-19

## Conclusiones

En esta memoria, se ha presentado una nueva forma de enfocar la Vigilancia Tecnológica mediante análisis de tendencias en patentes basándonos en el análisis de todos los campos de estas. Se formalizó el método de control de líneas de investigación de la competencia basándonos en K-Means como algoritmo de aprendizaje automático para procesamiento de texto. Esto permitió desarrollar un marco de procesamiento de Abstracts y Títulos de patentes, que conduce a un nuevo paradigma en el dominio de información competitiva a la hora de dirigir la innovación de una empresa o gobierno. Se

desarrollaron una serie de algoritmos y se construyó una arquitectura eficiente en el tratamiento de estos creando Patent Radar. A continuación las conclusiones más relevantes:

1. Las patentes deben convertirse en elemento central director de la innovación
2. Patent Radar es útil en la lucha contra el COVID-19
3. Patent Radar aporta valor añadido como plataforma integrada de gestión de la innovación
4. K-Means como algoritmo NLP para procesamiento de temas resulta adecuado para el tratamiento de patentes
5. NLTK y Gensim como algoritmos paralelos de preprocesameinto de Abstracts

## **Referencias**

[1] Oldham, Paul. *WIPO*, “WIPO Manual on Open Source Patent Analytics” Mayo 2020.

Disponible en: <https://wipo-analytics.github.io/introduction.html>



# Vigilancia Tecnológica por Big Data de Patentes y Espionaje Industrial

Universidad Pontificia de Comillas - ICAI



Juan Carlos García López

Junio 2020

# Abstract

**Español** Los usuarios de patentes y documentos de solicitud, como gobiernos, inventores y corporaciones dedicadas a diferentes sectores industriales, se esfuerzan por remar a la cabeza en materia de innovación. Esta memoria analiza el potencial de los datos contenidos en solicitudes de patentes para identificar tendencias de investigación. Se presenta un modelo basado en la extracción de datos de patentes y generación de un algoritmo no supervisado de identificación de tendencias para la representación de temas relevantes relativos a la generación de nuevas invenciones. El modelo fue probado con datos extraídos de la Oficina Europea de Patentes (EPO). Los experimentos muestran que el método proporciona una visión general de las direcciones de las tendencias y una perspectiva detallada de las tendencias actuales.

**English** Users of patents and application documents, such as governments, inventors and corporations dedicated to different industrial sectors, strive to paddle out on the innovation front. This report analyzes the potential of the data contained in patent applications to identify research trends. A model based on the extraction of patent data and generation of an unsupervised trend identification algorithm for the representation of relevant issues related to the generation of new inventions is presented. The model was tested with data extracted from the European Patent Office (EPO). Experiments show that the method provides an overview of trend directions and a detailed perspective on current trends.

# Índice general

<b>1. Introducción</b>	<b>7</b>
1.1. Motivación . . . . .	7
1.2. Objetivos . . . . .	10
1.3. Metodología . . . . .	13
<b>2. Planteamiento del Problema</b>	<b>15</b>
2.1. Introducción a la propiedad industrial . . . . .	15
2.1.1. Patentabilidad . . . . .	19
2.1.2. Proceso de solicitud de una patente . . . . .	20
2.2. Patentometría y Analítica de Patentes . . . . .	23
2.3. Vigilancia Tecnológica . . . . .	24
2.3.1. Bases de Datos . . . . .	25
2.3.2. La patente como fuente de información para empresas y go- biernos . . . . .	27
<b>3. Estado del Arte</b>	<b>31</b>
3.1. Gestión de datos . . . . .	32
3.2. Literatura relacionada . . . . .	34
3.3. Valoración crítica y limitaciones . . . . .	38
<b>4. Descripción del modelo Patent Radar</b>	<b>41</b>
4.1. Introducción al caso de estudio . . . . .	41
4.2. Objetivos del modelo Patent Radar . . . . .	43
4.3. Arquitectura del modelo Patent Radar y Flujo de Datos . . . . .	43

4.4. Fuentes, Tipos y Herramientas de Gestión de Datos . . . . .	45
4.4.1. Fuentes de datos . . . . .	46
4.4.2. Tipos de dato . . . . .	47
4.4.3. Herramientas de gestión de datos . . . . .	52
4.5. Algoritmos . . . . .	54
4.5.1. Minería de datos . . . . .	54
4.5.2. Pre-procesamiento de datos . . . . .	55
4.5.3. Procesamiento de texto libre. NLP. . . . .	60
4.5.4. Clusterización K-Means . . . . .	65
4.6. Visualización y control de resultados . . . . .	69
<b>5. Resultados Empíricos</b>	<b>75</b>
5.1. Obtención de dataset de análisis . . . . .	76
5.2. Análisis de Resultados . . . . .	77
5.3. Limitaciones observadas . . . . .	82
5.4. Influencia de los resultados en decisiones corporativas . . . . .	83
5.5. Influencia de los resultados en decisiones gubernamentales . . . . .	86
<b>6. Conclusiones y Líneas Futuras de Investigación</b>	<b>89</b>
6.1. Conclusiones . . . . .	90
6.2. Trabajo Futuro . . . . .	92
<b>A. Objetivos de Desarrollo Sostenible de Naciones Unidas (ODS)</b>	<b>99</b>
<b>B. Código Relevante</b>	<b>102</b>
B.1. Obtención de dataset de muestra . . . . .	102
B.2. Algoritmo K-Means para procesamiento de Abstracts . . . . .	102
<b>C. Características de sesgo TF-IDF</b>	<b>109</b>
C.1. Gestión de dataset de muestra . . . . .	109
<b>D. Manifiesto sobre la lucha contra el COVID-19 (EPO)</b>	<b>114</b>
<b>E. Características principales de Power BI</b>	<b>117</b>

# Índice de figuras

2.1. Evolución regional de solicitudes de registro [WIPO Database, 2019]	19
2.2. Proceso de solicitud [Fuente: A. Alonso, 2017]	21
3.1. Proceso de Moehrle para tratamiento de datos [Elaboración Propia, 2020]	33
4.1. Matriz de actuación/observación. [Elaboración Propia, 2020]	42
4.2. Arquitectura del Modelo. [Elaboración Propia, 2020]	44
4.3. Diagrama de flujo de datos de tipo string. [Elaboración Propia, 2020]	46
4.4. Ejemplo de patente registrada por Dña. Margarita Salas. [WIPO, 2017]	50
4.5. Comparación de herramientas de visualización de datos. Gartner Institute, 2017]	53
4.6. Diagrama de flujo de obtención de dataset completo (con abstract). [Elaboración propia, 2020]	55
4.7. Proceso de truncado en el campo "Inventor". [Elaboración propia, 2020]	56
4.8. Top 3 librerías Python para limpieza de datos [Elaboración propia, 2020]	57
4.9. Tokenización de texto Stanford University, 2016]	58
4.10. Comparación de diferentes técnicas de tokenización Mehmet Kayaalp, 2015]	59
4.11. Diagrama simplificado: Flujo de texto libre. [Elaboración propia, 2020]	60
4.12. Reducción de número de palabras por patente. [Elaboración propia, 2020]	60

4.13. Función de especificidad de términos IDF genérica. [Open Commons, 2020] . . . . .	64
4.14. Estudio de frecuencias absolutas de términos simples. [Elaboración propia, 2020] . . . . .	65
4.15. Estudio de frecuencias absolutas de términos binomiales. [Elaboración propia, 2020] . . . . .	66
4.16. Estudio de frecuencias absolutas de términos trinomiales. [Elaboración propia, 2020] . . . . .	66
4.17. Definición gráfica de espacio vectorial de términos y abstracts. [Elaboración propia, 2020] . . . . .	68
4.18. Clusters para IPCs tipo A [Elaboración Propia, 2020] . . . . .	68
4.19. Vista A de Patent Radar (Panel Principal). [Elaboración propia, 2020]	71
4.20. Vista B de Patent Radar (Control Pre-procesamiento). [Elaboración propia, 2020] . . . . .	72
4.21. Vista c de Patent Radar (Rendimiento K-Means). [Elaboración propia, 2020] . . . . .	73
5.1. Evolución temporal de patentes con términos relativos al COVID-19 [Elaboración propia, 2020] . . . . .	77
5.2. Top 10 organsimos solicitantes relativos al COVID-19 [Elaboración propia, 2020] . . . . .	78
5.3. Top 10 inventores relativos al COVID-19 [Elaboración propia, 2020] .	79
5.4. Distribución global de solicitudes [Elaboración propia, 2020] . . . . .	79
5.5. Nube de términos simples ponderados por TF[Elaboración propia, 2020]	80
5.6. NLTK reducción de términos [Elaboración propia, 2020] . . . . .	80
5.7. Análisis PCA y TSNE [Elaboración propia, 2020] . . . . .	82
5.8. Análisis SSE [Elaboración propia, 2020] . . . . .	83
5.9. Top 10 solicitantes del sector farmacéutico 2019 [European Pharmaceutical Review, 2020] . . . . .	84
5.10. Presupuesto destinado a I+D en España (% PIB) [TradingEconomics, 2020] . . . . .	87

# Capítulo 1

## Introducción

En el Capítulo 1, se hará un breve repaso a modo introductorio sobre los factores que motivaron la redacción de esta memoria, los objetivos que se establecieron al inicio del proyecto, y la metodología seguida para alcanzarlos.

### 1.1. Motivación

Desde finales del siglo XVIII existe una clara correlación entre la capacidad de innovar en países y empresas y su desarrollo económico y social. No son los países con mayores recursos los que ocupan los primeros puestos en las listas de países con mayor capacidad para crecer, sino aquellos que aplican mejor la tecnología, desarrollan nuevos productos y servicios, innovan en el modo y manera de implantar nuevos procesos o comercializan y distribuyen de manera más eficiente.

Si atendemos a la definición de Innovación, consensuada por la OCDE, como “Todos los pasos científicos, comerciales, técnicos y financieros necesarios para el desarrollo e introducción en el mercado con éxito de nuevos o mejorados productos, el uso comercial de nuevos o mejorados procesos y equipos, o la introducción de una nueva aproximación a un servicio social.” vemos que se hace incidencia en los conceptos de novedad y/o mejora.

Para que exista un producto o servicio, o en definitiva una idea relacionada con

una novedad o mejora, es necesario que el esfuerzo asociado a esta consecución esté incentivado. Si se innova para conseguir una posición competitiva mejor, mediante una inversión en capital, tiempo o recursos y con la consiguiente asunción de riesgo, parece una obviedad que el retorno de esa inversión esté de alguna manera “protegido” por unas reglas justas y que den garantías jurídicas aquéllos que aportan más y penalicen a los que tratan de aprovecharse del esfuerzo ajeno. Aquí surge el concepto de patente como un título que reconoce el derecho exclusivo de explotación de una determinada invención. Se trata de un derecho de propiedad industrial.

No es baladí el debate sobre la relación, inversa según algunos autores, entre patentes e innovación, pero no es objeto de este estudio el análisis jurídico de cuándo el derecho de propiedad intelectual favorece o coarta la innovación, así que supondremos en adelante que las patentes son un incentivo a la innovación, así como una herramienta que fuerza al inventor y organismo solicitante a desvelar su invención.

Tampoco es objetivo del presente trabajo sobrepasar el ámbito académico y entrar en disquisiciones de índole política entre países, disquisiciones que van más allá del derecho y están más enfocadas a diferencias consustanciales entre diferentes regímenes políticos -USA Vs. China, por ejemplo-. Así pues, nos centraremos en la defensa de la propiedad intelectual como justa compensación al esfuerzo individual o colectivo, en un marco legal justo, estable y conocido por todas las partes. En este punto parece una obviedad el destacar la concepción común, y comprensible, de la utilización del espionaje industrial como mecanismo de aprovechamiento ilícito del trabajo de otros. Sin embargo, atendiendo puramente a su definición práctica, el Espionaje Industrial no es malo per se. A veces, esta línea tan clara de separación no es tan limpia. En la cultura asiática existe una tradición ancestral de absorción de la información externa, que no debe ser considerada como negativa. Esa captación de información les ha permitido en muchos casos mejorar productos de sus competidores y no es del todo cierto que siempre ha utilizado el espionaje industrial como palanca de su desarrollo. En algunos casos, es justo reconocerlo, simplemente lo han hecho mejor. El “benchmarking” es un concepto asiático en origen y nadie



lo considera un proceso ilícito. En occidente, sin embargo, no tenemos esa obsesión por fotografiarlo todo, aprender de todo, escuchar sin interrumpir y eso nos lleva a perder competitividad. Aprender de nuestros competidores es sano, copiar para acortar esfuerzo, no lo es, y ahí es donde esa línea imaginaria juega su papel más relevante. Tres países asiáticos lideran los los índices de innovación en su continente: China, Japón y Corea del Sur. En Europa lideran Suiza, Suecia, Países Bajos, Reino Unido y Finlandia. Estados Unidos el tercer país en cuanto a su capacidad de innovación. Si nos atenemos a la clasificación de solicitudes de patentes, China ocupa el primer lugar mundial, con U.S. en segundo lugar, Japón en el tercero, seguidos de Alemania, Corea, Francia, UK, Suiza y Suecia.

Como vemos existe una cierta correlación, aunque no necesariamente lineal, entre innovación y patentes. Lo más interesante es analizar la segmentación por áreas y tipos de compañías. China y Corea tienen un fuerte sustrato en el sector de las telecomunicaciones y así Huawei – empresa china de telecomunicaciones – es líder en el mundo en la presentación de solicitudes de patentes – o Samsung – empresa coreana de telefonía móvil ocupa el tercer lugar. En Japón existe un componente más centrado en el sector del automóvil, con Mitsubishi en el segundo lugar. En Suiza hay que destacar la relevancia del sector farmacéutico, con Novartis a la cabeza.

Es una incógnita el casi seguro vuelco de investigación en estas listas, como consecuencia de la crisis global del coronavirus. La WIPO<sup>1</sup> es la encargada de elaborar estas listas y la encargada de evaluar este impacto.

Por último y no menos relevante, cabe reseñar que el registro de patentes no es sólo un mecanismo más o menos sofisticado de protección al agente que patenta. Hoy en día, tal y como se apuntaba en el párrafo sobre la obsesión asiática en la captura de información la analítica de datos, tanto descriptiva, como predictiva o prescriptiva es la base para la toma correcta de decisiones. El dato se transforma

---

<sup>1</sup>World Intellectual Property Organization

en información y ésta en conocimiento. Para tener un correcto conocimiento de las tendencias en innovación es necesario tener datos que sean fuentes en las que poder confiar. Estos datos se obtienen de los registros de patentes. La generación de la información procedente de estos datos se hace a través de herramientas, algoritmos, etc. Estos metadatos permiten identificar patrones, tendencias y predicciones basadas en conocimiento sin sesgos.

Cada país o cada empresa harán un uso más o menos eficiente de este conocimiento, pero es innegable que la aplicación de este conocimiento a la toma de decisiones, va a ser y de hecho ya es, una palanca fundamental para la obtención de ventajas competitivas y por ende de crecimiento. Es por ello, que este trabajo pretende crear una plataforma capaz de exprimir al máximo la información pública disponible en la actualidad para facilitar dicha toma de decisiones.

## 1.2. Objetivos

Una vez determinada la relación causa-efecto entre innovación, generación de patentes y desarrollo económico, se nos plantean las siguientes incógnitas:

- ¿Toda acción de innovación y/o de generación de patentes es buena de por sí? o ¿se necesita una priorización de propuestas en función de criterios medibles?
- ¿Cómo se determina el proceso de selección de acciones innovadoras? y lo que es más importante, ¿cuáles se desechan?
- ¿Porqué unas empresas o países tienen éxito en sus políticas de innovación y derechos de propiedad intelectual y otros no?
- ¿Se puede predecir si una determinada patente o acción innovadora puede tener más éxito que otra?
- ¿Qué influencia tiene el ecosistema cercano en el desarrollo de una patente? ¿Es recomendable promover acciones de innovación en telecomunicaciones en

un entorno preparado para el desarrollo de la industria del automóvil o de la cosmética?

Éstas y otras cuestiones similares nos llevan a buscar herramientas, datos de base, metodologías, modelos, etc. que nos permitan una toma de decisiones tanto en la selección de patentes como en los mecanismos de protección de éstas, de manera que se pueda optimizar unos recursos limitados en la decisión de inversión.

Uno de los errores más comunes en el ámbito de la innovación es el destinar ingentes cantidades de dinero y recursos a proyectos sin ningún atractivo o viabilidad. Las razones son múltiples: decisiones políticas sin fundamento, subvenciones encaminadas a otros fines distintos de la propia innovación, falta de conocimiento de las dificultades del desarrollo del proyecto en cuestión, no contar con los recursos con las suficientes competencias, etc. etc. Lo peor de estos casos es que, en muchas ocasiones, no se miden ni las previsiones ni los resultados, con lo que el derroche de medios sin sentido es un lastre que paraliza en su desarrollo a empresas y a países enteros de manera recurrente.

¿Cuál es el KPI más importante, entonces, que nos dirige a una decisión adecuada, profesional y eficiente? La contestación es obvia: el retorno económico de la inversión en innovación. No nos olvidamos de que hay proyectos con retorno social, ético o climático, etc. Sin embargo, el propio retorno económico de cualquier proyecto de innovación a mayor o menor plazo ya genera de por sí un beneficio social. El camino contrario no es evidente y en cualquier caso muy difícil de medir.

Una vez que determinamos la rentabilidad de un proyecto de innovación o de una patente con su KPI clave, nos queda la parte más importante. Es ingente la cantidad de proyectos fracasados con un retorno aceptable por tener un sesgo exclusivamente financiero sin estar soportado por lo que podríamos llamar: “datos relevantes del mercado de patentes”.

Para disponer de un mecanismo de toma de decisión es necesario disponer de:

- Datos adecuados segmentados de patentes por países y por áreas innovación
- Objetivos de cada patente encuadrados según una división conceptual preestablecida
- Una metodología de análisis de los datos disponibles
- Un modelo/herramienta de evaluación

Así pues, el objetivo principal de este trabajo es el plantear una metodología, asociada a las técnicas actuales de extracción, transformación y carga de datos, con el fin de disponer de una analítica avanzada que nos ayude en el conocimiento para la toma de decisiones sobre patentes. Dicho objetivo puede desglosarse en profundidad en los siguientes :

1. Revisar el Estado del Arte en materia de analítica de datos de propiedad industrial
2. Dar una solución a medida, escalable y con un coste contenido a las principales inquietudes por parte de empresas y gobiernos en materia de Vigilancia Tecnológica
3. Optimizar del uso de los datos de acceso público de la EPO<sup>2</sup>
4. Aplicar de manera práctica procesos de limpieza y análisis de datos

En el mercado de patentes debería y de hecho debe haber un proceso similar de análisis del entorno, utilizando datos reales, soportados por organismos internacionales sin ánimo de lucro.

---

<sup>2</sup>Oficina Europea de Patentes

### 1.3. Metodología

La metodología aplicada en este trabajo se basa en siete pasos diferenciados:

1. Plan de Trabajo: En este punto se fijaron las diferentes etapas del trabajo, con sus objetivos (entregables) claros para cada etapa. Aunque el plan de trabajo no permaneció invariable a lo largo del trabajo, se registraron las causas de las desviaciones para que cada una de éstas sea objeto de análisis particularizado y sirva de aprendizaje para corrección en la siguiente etapa. El plan de trabajo debe tener dos atributos de especial relevancia: “factibilidad” y “coherencia”
2. Investigación sobre el “estado del arte” en materia de Analítica de Patentes, desarrollando los procesos actuales relacionados con la patentabilidad y solicitudes de patentes. Se incluyeron textos, artículos, libros, páginas web y fuentes de datos cumpliendo con los siguientes criterios: Pertenencia, Actualidad e Internacionalidad.
3. Enfoque de la problemática a estudiar en el Trabajo: Marco teórico presente y casos y ejemplos explicativos de la situación actual.
4. Elección del algoritmo: para predecir y describir datos de salida, utilizando Inteligencia Artificial o más concretamente Machine Learning. Fue necesario formarse en técnicas de ML, asistiendo presencialmente y “on-line” a diferentes cursos y seminarios y en particular los impartidos por la Universidad de Standford. Se priorizó la intención de “clusterizar”, en la medida de lo posible, la información de patentes y determinar tendencias en función de esta “clusterización”. Así pues, se tuvo que trabajar en el ámbito de los algoritmos de tipo “Unsupervised Learning” en sus diferentes variantes, cada una de ellas con una aplicabilidad concreta.
5. Obtención de datos: se llevó a cabo un análisis de las principales fuentes de datos de patentes disponibles en la actualidad y se hizo una comparación exhaustiva de ellas para definir de cuales obtener la información que nutriría más adelante al algoritmo.

6. Contraste de hipótesis: se comprobó la relación entre causa y efecto con ejemplos prácticos y, reales. Se definió un protocolo de contraste con las suficientes garantías de rigor y control, evitando en lo posible todo tipo de sesgos.
7. Elaboración de conclusiones: se hizo una valoración objetiva de los resultados empíricos, reconociendo tanto las fortalezas como las debilidades del modelo y definiendo una hoja de ruta para futuras líneas de investigación.

# Capítulo 2

## Planteamiento del Problema

El Capítulo 2, constituye un breve repaso contextual desde la creación de patentes como documento legal, hasta la utilidad de la información contenida en ellas. Se realiza un breve repaso sobre el proceso a seguir para proteger una invención, así como los organismos dedicados a tal fin con el objetivo de ilustrar al lector en los pasos previos a la determinación de la importancia de la Vigilancia Tecnológica. La Vigilancia Tecnológica juega un papel fundamental a la hora de identificar tendencias en situaciones competitivas, en las que se encuentran actualmente numerosas empresas y gobiernos.

Una correcta Vigilancia Tecnológica del entorno competitivo, resulta un factor determinante para el desempeño de industrias críticas en la sociedad. Esta vigilancia, constituye un procedimiento sistemático de detección y análisis, así como de búsqueda de información científico-tecnológica, que ejerce de ayuda en la toma de decisiones, anteponiendo los retos y desafíos que favorezcan la estrategia de negocios y de investigación, independientemente del sector empresarial.

### 2.1. Introducción a la propiedad industrial

Las patentes tienen una importante finalidad en el mundo industrial, pues son las que reconocen el derecho exclusivo de uso, protección y explotación al creador de una invención. Éstas, facilitan que se pueda avanzar en diversos sectores, como

el de la industria o la tecnología, porque obligan al inventor a desvelar el invento.

Las patentes contemporáneas, posteriores al s.XIX y herederas de la firma del Estatuto de Venecia de 1474, se definen como:

*Un conjunto de derechos exclusivos concedidos por un Estado al inventor de un nuevo producto o tecnología, susceptibles de ser explotados comercialmente por un período limitado de tiempo, a cambio de la divulgación de la invención.*<sup>1</sup>

Si bien la aprobación de una patente, supone un monopolio artificial, el fin último de una patente es el de obligar al inventor a compartir sus descubrimientos sin dejar de lado como horizonte de sucesos el avance de la sociedad y, en retorno, obtener derechos exclusivos de forma temporal.

El concepto de patente, sin embargo, puede estudiarse desde dos prismas diferentes:

- Como elemento de derecho de propiedad industrial
- Como publicación científica de divulgación

Analizar de forma estructural los documentos de patentes y los campos de datos contenidos en ellas es el pilar central del análisis que ocupa esta memoria. Sin embargo, no deberemos dejar de lado las características clave de las patentes como elemento de derecho a la propiedad industrial. Según la *World International Patent Office*, *Una patente es una concesión temporal de un derecho exclusivo a un titular para evitar que otros hagan, usen, ofrezcan para la venta o importen, una invención patentada sin su consentimiento, en un país donde la patente esté vigente*".<sup>2</sup> Los derechos de patente, por regla general, son derechos territoriales: solo son válidos en el

---

<sup>1</sup>[RAE20] 2020 Real Academia Española y Asociación de Academias de la Lengua Española. «patente». Diccionario de la lengua española (23.<sup>a</sup> edición). Madrid: Espasa

<sup>2</sup>[WIPO3] 2003, WIPO, Training Course on Practical Intellectual Property Issues in Business



territorio o ámbito natural del país u organismo donde se otorgan y, por lo general, se otorgan durante un período de vigencia de 20 años a partir de la presentación de los datos de una solicitud, aún siendo susceptibles de oposición o revocación si se considera oportuno.

Resulta especialmente relevante en este punto de la memoria, profundizar en el concepto de *Patente Europea*. En la UE, existen tres formas<sup>3</sup> de proteger formalmente las invenciones:

1. Patente Nacional: se presenta una solicitud de patente individualizada en cada país o estado en el que se desee obtener la protección
2. Patente Europea: se presenta una solicitud a la Oficina Europea de Patentes (EPO), que resultará de forma conjunta en el registro en cada uno de los estados que suscriben el *Convenio sobre la Concesión de Patentes Europeas*<sup>4</sup>.
3. Patente Internacional PCT: se presenta una solicitud única de registro de la patente, válida para todos los estados<sup>5</sup> que forman parte del Tratado Internacional de Cooperación en Materia de Patentes (PCT).

Todas las solicitudes de patentes, independientemente del canal mediante el cual se hayan solicitado, han de cumplir las mismas características<sup>6</sup> para ser elegibles: involucrar temas patentables, implicar un paso inventivo y ser susceptibles de aplicación industrial. Aún cuando el contenido de las solicitudes de patentes difiere tanto en temática como en el estilo de redacción del solicitante, es crucial que la solicitud de registro conste de un documento formal con campos predefinidos formando una estructura de datos que puedan analizarse mediante aprendizaje automático.

Para el análisis de patentes y su contenido necesitamos focalizarnos en estudiar las solicitudes de protección industrial como tipo de documento y comprender:

---

<sup>3</sup>[MIT12] 2012, Ministerio de Industria, Energía y Turismo, La Patente Europea

<sup>4</sup>43 estados en Junio de 2020

<sup>5</sup>153 estados en Junio de 2020

<sup>6</sup>[OEP20], 2020, Oficina Española de Patentes y Marcas

- La definición estructural de los documentos de solicitud y los campos que puedan aportar datos de valor añadido en ellas
- Las características de las diferentes bases de datos disponibles como fuente inequívoca de datos.
- Las aplicaciones prácticas que esos datos nos brindan para ayudar en la vigilancia tecnológica, identificación de tendencias y control competitivo

En el último lustro, se ha podido apreciar cómo se solicitaban más de 3,3 millones de patentes en todo el mundo, lo que supone un aumento<sup>7</sup> del 5,2% durante un lustro de forma ininterrumpida y consecutiva. Asia, se encuentra a la cabeza en cantidad de solicitudes con China como principal potencia e India como país registrando el mayor crecimiento mundial. La polarización de las solicitudes de patentes desde los continentes occidentales al continente oriental ha generado una desestabilización en los procesos de compañías industriales y gobiernos, que tradicionalmente se limitaban a vigilar el entorno competitivo cercano. En la actualidad, dicho entorno competitivo se ha convertido en un vasto campo de análisis en el que, quien no es capaz de adaptarse a la globalización y rapidez de cambio, se queda atrás. Es por ello, que en el año 2002<sup>8</sup>, se preveía un fuerte incremento en los campos de análisis de tendencias que, sin duda, se ha cumplido. Grandes instituciones como Stanford University o el MIT comenzarán refinando las técnicas tradicionales de patentometría para, a partir del *boom* de la analítica de datos, pasar a estudiar la aplicación de algoritmos<sup>9</sup> más complejos mediante aprendizaje automático.

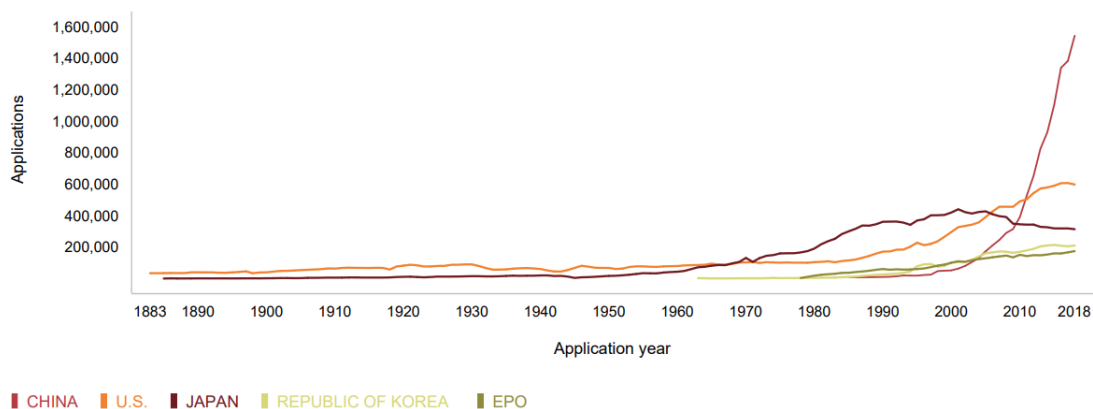
En paralelo, y por el peso histórico que conlleva, merece la pena pararse a reflexionar sobre los hechos que han acontecido en los meses previos a la elaboración de este informe. La irrupción de la pandemia asociada al COVID-19 prevé un cambio de paradigma en la innovación tecnológica del próximo siglo. Debido a la preocupación sobre factores sanitarios, es probable la migración de recursos por parte de

---

<sup>7</sup>[OMPI18] 2018, OMPI, Indicadores Mundiales de propiedad intelectual

<sup>8</sup>[BRE02] 2002, Breitzman, Anthony F.; Moguee, Mary Ellen. The many applications of patent analysis

<sup>9</sup>[TADU11] 2011, Taduri, S., Lau, G. T., Law, K. H., Yu, H. Developing an ontology for the US patent system.



**Figura 2.1:** Evolución regional de solicitudes de registro [WIPO Database, 2019]

Patentes	2017	2018	Crecimiento (%)	Distribución (%)
<b>Mundial</b>	<b>3.162.300</b>	<b>3.326.300</b>	<b>5.2</b>	<b>100</b>
China	1.381.594	1.542.002	11.6	46.4
U.S.	606.956	597.141	-1.6	18

**Cuadro 2.1:** Evolución del registro de patentes [Fuente: WIPO Database, 2019]

empresas de alta y baja capitalización hacia campos sanitarios o relacionados con la farmacología. La analítica de patentes puede convertirse en una gran aliada a la hora de dar visibilidad y agilizar la forma en la que se innova en estos campos. Además, la inevitable aversión al comercio chino, por ser el origen de la citada pandemia, puede dar lugar a nuevas reorientaciones en la fabricación y registro de propiedad intelectual hacia países más occidentales aunque con una como India, o al conjunto de países de Oriente Medio.

### 2.1.1. Patentabilidad

Como hemos comentado anteriormente, cualquier sector o campo de investigación es susceptible de innovación y patentabilidad. Existen, sin embargo, un conjunto de características necesarias y suficientes <sup>10</sup> necesarias al definir qué es patentable y qué no. A continuación, se resumen dichas características por A.Alonso<sup>11</sup>

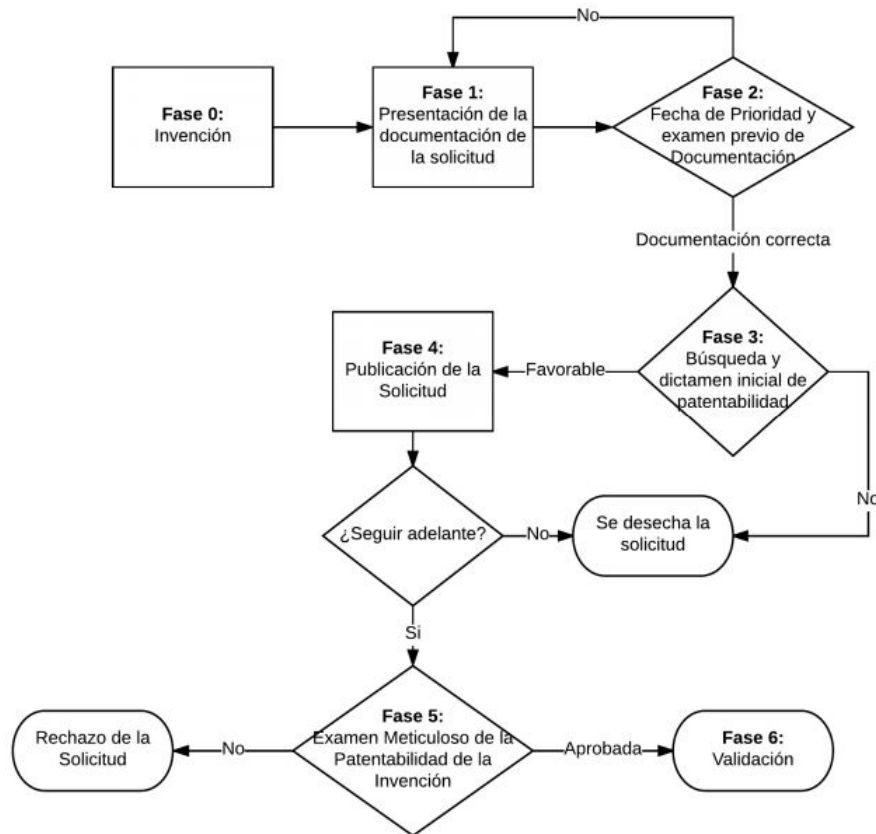
<sup>10</sup>Ley 10/2002, de 29 de Abril, BOE, Ministerio de Industria

<sup>11</sup>[ALON17] 2017, Sistema de Arrastre Innovador para deportes de deslizamiento, A. Alonso

1. "**Novedad:** Una invención es considerada nueva siempre y cuando no esté comprendida en el estado de la técnica. La LP especifica que el estado de la técnica está constituido por todo lo que antes de la fecha de presentación de la solicitud de patente se ha hecho accesible al público en España o en el extranjero por una descripción escrita u oral, por una utilización o por cualquier otro medio. De la misma forma se incluyen todos los contenidos de patentes solicitudes con fecha anterior. Para poder rellenar la solicitud es necesario que el solicitante se haya informado del previo estado del arte. Un meticuloso estudio por parte del solicitante puede evitar el desconcierto por un rechazo posterior por parte de la administración a la que se solicita la patente.
2. **Actividad Inventiva:** Se considera que una invención implica una actividad inventiva si aquélla no resulta del estado de la técnica de una manera evidente para un experto en la materia. En muchas ocasiones el paso inventivo se produce por una combinación de distintas características ya incluidas en el estado del arte, pero que indudablemente su combinación sin precedentes constituyen un producto totalmente nuevo. Es un resultado muy común cuando por eventos inventivos se descubren efectos técnicos inesperados o desconocidos hasta el momento.
3. **Aplicabilidad Industrial:** Según el artículo 9 de la LP se considera que una invención es susceptible de aplicación industrial cuando su objeto puede ser fabricado o utilizado en cualquier clase de industria. Esta restricción excluye todo intento de patentar invenciones que todavía no se hayan desarrollado con éxito, a pesar de que se lleve mucho tiempo investigando en ello."

### 2.1.2. Proceso de solicitud de una patente

Resulta de especial relevancia, hace un breve resumen del proceso que ha de llevar a cabo una persona interesada en la protección de un invento concreto. Cuando un solicitante se enfrenta a la petición formal de registro de una patente, ha de seguir una serie de pasos resumidos de en la Figura 2.2.



**Figura 2.2:** Proceso de solicitud [Fuente: A. Alonso, 2017]

La persona interesada y su apoyo legal si así lo requiere, deben presentar los documentos necesarios en el organismo correspondiente junto con la tasa definida para tal efecto. Se lleva a cabo, a continuación, una evaluación de formato para, en caso favorable, asignar al documento una fecha de prioridad, que cobra especial relevancia al evaluar la viabilidad de la solicitud: Todo contenido técnico publicado con anterioridad es considerado como parte del estado de la cuestión<sup>12</sup> y tiene la capacidad potencial de provocar que el tribunal de la EPO rechace la solicitud.

Más adelante, las persona interesada recibe un documento a modo resumen con la bibliografía relacionada a modo de orientación sobre el estado del arte para determinar el nivel de actividad inventiva de la propuesta de acuerdo a la solicitud.

<sup>12</sup>[ALON17] 2017, Sistema de Arrastre Innovador para deportes de deslizamiento, A. Alonso

Un año y medio después de prioridad existe obligación de publicación. En adelante, la solicitud estará reflejada en las cualquier base de datos de los organismos recogidos en el Apartado 2.2. Se tratará, por tanto, como Estado del Arte en cualquier registro/solicitud futura de solicitantes o inventores ajenos al interesado. Cuando la solicitud ha sido publicada, el interesado dispone de seis meses para decidir si quiere seguir adelante o no con la solicitud y el alcance geográfico de esta. Merece la pena reseñar, que la solicitud de protección aún no se ha aceptado: será, en cambio, la fecha legal de concesión de esta la que determinará la protección a efectos legales. Si el interesado ha elegido continuar el proceso, el paso posterior es solicitar una evaluación exhaustiva en la que la EPO decidirá si la solicitud cumple con las características necesarias para ser registrada mediante convenio CPC.

Actualmente existen tres organismos principales cuando un solicitante español desea obtener protección en una de sus invenciones. Por orden de granularidad, existen: la Oficina Española de Patentes y Marcas<sup>13</sup>, la Oficina Europea de Patentes<sup>14</sup> y la

---

<sup>13</sup>2020, Oficina Española de Patentes y Marcas. Definición oficial: La Oficina Española de Patentes y Marcas (OEPM) es un Organismo Autónomo del Ministerio de Industria, Energía y Turismo que impulsa y apoya el desarrollo tecnológico y económico otorgando protección jurídica a las distintas modalidades de propiedad industrial mediante la concesión de patentes y modelos de utilidad (invenciones); diseños industriales (creaciones de forma); marcas y nombres comerciales (signos distintivos) y títulos de protección de las topografías de productos semiconductores. Asimismo, difunde la información relativa a las diferentes formas de protección de la propiedad industrial. En el plano internacional, la OEPM es la encargada de representar a España en los distintos foros y organizaciones internacionales que se encargan de la propiedad industrial e intelectual. Los objetivos fundamentales de la OEPM son:

1. Proteger y fomentar la actividad de creación e innovación tecnológica en nuestro país, así como la identidad corporativa empresarial mediante la concesión de títulos de Propiedad Industrial.
2. Transmitir información que oriente la actividad investigadora a través del mantenimiento de fondos documentales y bases de datos que permiten un acceso rápido y sencillo al estado actual de la técnica mundial en cualquier sector.
3. Impulsar la circulación y el intercambio de bienes y servicios a través de la difusión de la información de los signos distintivos registrados.

<sup>14</sup>La European Patent Organization está compuesta por dos organismos: el Consejo de Administración y la EPO. El primer organismo se encarga de la supervisión de la Oficina de Patentes y su actividad operativa (EPOrg). La European Patent Office (EPO) es una institución internacional creada bajo amparo de la Convención Europea de Múnich en 1973. Como se comentó brevemente en el apartado 2.1, la EPO concede una única protección que no es sino un conjunto de protecciones en distintos países suscritos al convenio unificados en un solo proceso. Por tanto, la EPO no

Oficina Internacional de Patentes<sup>15</sup>

## 2.2. Patentometría y Analítica de Patentes

El afán de conocimiento o exploración. El análisis de los conocimientos contenidos en las patentes es el pilar principal de la patentometría.

Hace apenas 10 años, se necesitaban habilidades especiales, conocimiento y tenacidad para buscar en varias bases de datos internacionales y reunir la información de patentes necesaria para un propósito determinado. Eran necesarios procesos de corrección de ortografía, puntuación y transliteraciones para comenzar a dar sentido a lo que se estaba analizando. El lugar donde los competidores patentaban era una primera pista de dónde pretendían construir una presencia y potencialmente el alcance de una determinada tecnología. Con esa información, podía comenzar a formarse una estrategia y comenzar a desplegar recursos humanos y financieros.

Hoy en día, las citas, las normas de estilo y los códigos de clase se pueden usar de manera rutinaria para la identificación de tendencias, descargando gran parte del trabajo normativo que era inherente al trabajo de analista y que a menudo tomaba más tiempo que el análisis en sí. Con estos avances, las herramientas de búsqueda de patentes se convirtieron rápidamente en fuente de inteligencia crítica para el desarrollo de la innovación.

Tras la apertura de las bases de datos de los organismos comentados en el Apartado

---

tiene cobertura ni decide sobre procesos de infracción de la jurisdicción de la patente, quedando a competencia de los organismos nacionales. Las solicitudes de registro pueden estar cumplimentadas en cualquier inglés, alemán o francés (lenguas cooficiales de la UE) siempre que esta vaya adjunta a una traducción jurada.

<sup>15</sup>La WIPO<sup>16</sup> u Organización Mundial de la Propiedad Intelectual (OMPI) es el foro mundial para los servicios, las políticas, la información y la cooperación en materia de propiedad intelectual (PI). Son una agencia autofinanciada perteneciente al tratado de las Naciones Unidas, con 193 estados miembros. Pretenden liderar el desarrollo de un sistema internacional de propiedad intelectual equilibrado y eficaz que permita la innovación y la creatividad en beneficio de todos. Nuestro mandato, órganos rectores y procedimientos se establecen en el Convenio de la OMPI, que estableció la OMPI en 1967.

2.1.4 y desde que comenzó la aceleración en investigación de los campos relacionados con el análisis de datos masivos, o *Big Data* a partir de la década del 2010, el análisis tradicional de conteo y la clasificación estática de temática de patentes parece ir dejando lugar a una nueva corriente de análisis basados en Inteligencia Artificial<sup>17</sup>, Machine Learning y Deep Learning. El concepto clásico de patentometría se convierte ahora en la base para generar una corriente de análisis mucho más rica gracias al incremento en la capacidad de proceso de la máquinas actuales, bautizando por tanto un nuevo campo llamado *Analítica de Patentes*.

Es importante reseñar el número e impacto, cada vez mayor, de las herramientas *Open Source* que permiten al analista de datos gestionar grandes volúmenes de estos sin pasar por un proceso de pago. Estas permiten al usuario no profesional generar sistemas de análisis de información a un coste reducido o, incluso, gratuito. El número disponible de estas es, actualmente, bastante alto. Sin embargo, todas ellas cuentan con una curva de aprendizaje relativamente empinada, que hace que se necesite tiempo para manejarlas con soltura. Se necesitará, en la mayoría de los casos, conocimiento previo en materia de programación, lo que hace poco accesible al público general una metodología transparente de intercambio de información.

## 2.3. Vigilancia Tecnológica

La Vigilancia Tecnológica es una ciencia que ayuda a enfocar y dirigir, desde un punto de vista estratégico, los procesos de I+D+i en cualquier sector industrial. Su pilar fundamental es la afirmación de que la tecnología es un elemento nuclear en los vaivenes competitivos de las corporaciones y por la innovación, en todos sus componentes, es una condición *Sine Qua Non* para el progreso.

---

<sup>17</sup>[ARIS18] 2018, Aristodemou, L., Tietze, F. The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data



### 2.3.1. Bases de Datos

Cuando centramos la Vigilancia Tecnológica en la analítica de solicitudes de propiedad intelectual, se trata generalmente con datos de siete tipos diferentes:

- Fechas (prioridad, fechas de solicitud y publicación)
- Números (número de prioridad, número de solicitud, número de publicación, miembros de la familia, citas)
- Nombres (solicitantes, también conocidos como inventores)
- Códigos de clasificación (por ejemplo, Clasificación internacional de patentes / Clasificación cooperativa de patentes)
- Campos de texto (Título, Resumen, Descripción)
- Imágenes (diagramas)
- Información adicional (estado legal, registro público, etc.)

Afortunadamente, estos datos vienen normalmente estructurados en bases de datos relacionales, que hacen que el trabajo de pre-procesamiento de los datos sea más sencillo. Sin embargo, el análisis de los datos de patentes no pretende, en su base, ser inferencial, sino describir patrones ocultos o tendencias en compañías y países.

El primer problema que se presenta al analizar solicitudes de patentes es la capacidad de acceder a los datos contenidos en estas en cantidades adecuadas para optimizar el proceso de análisis. Por lo general, este proceso implica gestionar millones de solicitudes. Los avances del último lustro, han acercado, cada vez más, contenidos en la solicitudes de propiedad intelectual al público haciendo disponible la descarga de aproximadamente 10,000 registros<sup>18</sup> de forma simultánea. Sin embargo, el acceso a grandes cantidades de títulos o abstracts sigue siendo limitado. Los organismos dedicados a la gestión de solicitudes, como la EPO, han jugado un papel nuclear a

---

<sup>18</sup>[WIPO16] 2016, Manual on Open Source Patent Analytics

la hora de hacer que los datos en masa sean accesibles por público general, y más específicamente, para los profesionales dedicados al *Patent Analytics*. Sin embargo, el proceso ha mejorado en cuanto al acceso a los datos, pero no del todo en los volúmenes en los que a los analistas de patentes les gustaría.

Desde hace aproximadamente diez años, se viene observando un cambio de paradigma en el uso de herramientas de investigación. Parece que los esfuerzos de la comunidad informática, han potenciado el uso de *software Open Source* y el fomento del acceso sin barreras a documentos científicos. La finalidad última de los organismos dedicados a la gestión de solicitudes es que el máximo volumen de información de estas esté disponible de un modo más abierto y, de esta forma, para una masa crítica mayor. Dichas organizaciones (WIPO, EPO...) han respondido a esto mediante la creación de bases de datos de acceso público<sup>19</sup>, como la base *Esp@cenet* de la Oficina Europea de Patentes. WIPO Patentscope, por su parte, permite el acceso a mas de medio centenar de millones de patentes suscritas al CPC. Existen otras herramientas menores procedentes de gigantes tecnológicos como Google o The Lens, aunque con menor repercusión. La mayoría de estas herramientas no requieren conocimientos de programación para la descarga de datos. Sin embargo, la EPO brinda acceso gratuito<sup>20</sup> a filas no pre-procesadas para interesados en gestores de aplicaciones (API) y análisis datos XML/JSON en bruto. Existe, por tanto, un ecosistema variado de fuentes y proveedores de información sobre patentes.

En la fecha de redacción de esta memoria, existen siete bases de datos principales disponibles para extraer datos relativos a patentes. Si bien este análisis sólo se centrará en datos que provengan de Esp@cenet (Oficina Europea de Patentes), en la Tabla 3.2 se puede apreciar en una síntesis de ellas y sus principales características. Los archivos de patentes extraídos de estas, por lo general, se gestionan en formato

---

<sup>19</sup>Resulta interesante la explicación de [PO20] 2020, I. Kitsara P. Oldham. The WIPO Manual on Open Source Patent Analytics. <https://wipo-analytics.github.io/>

<sup>20</sup>Para más información, es recomendable visitar la sección de datos en bruto de la EPO. <https://www.epo.org/searching-for-patents/data/bulk-data-sets.html>

XML.

Es importante comprender el proceso de análisis a seguir una vez se han obte-

BB.DD.	Cobertura geográfica	Actualización	Cobertura temporal
BBDD Local	Region	Variable	Variable
USPTO	U.S.	Semanal	Desde 1970
JAPIO	Japón	Variable	Desde 1976
PATSTAT	UE	Variable	Variable
Latindex	Variable	Semanal	Desde 1972
Esp@cenet	UE	Continúa	Desde 1800 (con limitaciones)

**Cuadro 2.2:** Bases de datos disponibles [Fuente: Elaboración Propia]

nido los datos de una fuente fiable. Trippe<sup>21</sup> presenta una guía, que identifica y explica una gran cantidad de conceptos sobre análisis de patentes y la metodología sobre cómo ejecutar los diferentes tipos de análisis sobre datos de estas. Con los recientes avances en inteligencia artificial, ha habido un nivel creciente de actividad en torno a las diferentes metodologías involucradas que podrían aplicarse a los datos de propiedad intelectual.

### 2.3.2. La patente como fuente de información para empresas y gobiernos

En un mundo conectado, donde el desarrollo tecnológico depende cada vez más de la colaboración de diferentes socios, la utilización efectiva de los datos de patentes tiene un potencial significativo. Si se aplican las técnicas correctas, los datos de patentes, en particular, pueden usarse para la toma regular de decisiones a nivel estratégico en todo tipo de organizaciones, ya sean pequeñas o grandes, pertenecientes al ámbito privado o público.

Los datos de patentes se han considerado durante mucho tiempo el repositorio más grande del mundo de información tecnológica <sup>22</sup>. En las últimas décadas, sin em-

---

<sup>21</sup>[TRIP15] 2015, WIPO Guide to Using Patent Information

<sup>22</sup>[OMP26] 2016 OMPI

bargo, se han vuelto cada vez más accesibles para un público no especializado. Si bien la calidad de los datos de patentes ha aumentado enormemente en los últimos 20 años, y se están desarrollando continuamente herramientas de software cada vez mejores para analizar los datos, aún hoy en día aún queda un potencial significativo para exprimir todo su potencial <sup>23</sup>.

Los cambios en materia de política de patentes<sup>24</sup> en los países de la OCDE durante las últimas dos décadas, han fomentado el uso y la aplicación de patentes con el objetivo de alentar las inversiones en innovación y mejorar la difusión del conocimiento. Los gobiernos actuales intentan, por tanto, pronosticar las principales áreas de investigación más atractivas susceptibles de ser sujetas a financiación. Del mismo modo, los investigadores pertenecientes a grandes corporaciones intentan mapear el conocimiento e identificar posibles lagunas relevantes para el avance de la ciencia.

La extracción de información relevante de las solicitudes de patentes permite el análisis de las principales áreas de investigación y el mapeo de los temas de interés actuales. Los desarrollos tecnológicos que facilitan el análisis de patentes pueden proporcionar a los tomadores de decisiones una visión general de alto nivel de la dirección que toman los nuevos inventos. Además, esta actividad puede ayudar a los investigadores a identificar esfuerzos de investigación que sean relevantes y fundamentados. Un mapa de conocimiento de las tendencias de patentes para vigilancia tecnológica puede permitir que un investigador identifique la necesidad de una investigación específica en un campo considerado como atractivo. Además, las instituciones de investigación y gubernamentales que proporcionan fondos pueden planificar previamente con un horizonte más largo y desviar los fondos de investigación a los campos necesarios para el desarrollo que demanda la sociedad. Los mapas de conocimiento de patentes pueden ayudar, a su vez, en la clasificación de las direcciones de investigación en el pasado y en el intento de predecir futuras direcciones en las

---

<sup>23</sup>[TIE15] 2015, Tietze, The future of patent analytics

<sup>24</sup>[OCDE04] 2004 Patents and Innovation: Trends and Policy Changes

que orientar la investigación. A continuación se enumeran los principales retos<sup>25</sup> en los que la vigilancia tecnológica puede ayudar a gobiernos en la toma de decisiones:

- **Mercados basados en innovaciones tecnológicas:** Las patentes juegan un papel fundamental en el desarrollo de transacciones financieras relacionadas con tecnología. Los gobiernos deben demostrar su conocimiento del funcionamiento de los mercados relativos a este campo y el efecto de dichos mercados en el desempeño económico estatal para apoyar su desarrollo en las direcciones más beneficiosas para la sociedad.
- **Organizaciones públicas de investigación:** Fomentar el desarrollo de la innovación por parte de las organizaciones públicas de investigación ha llevado a una mayor comercialización de invenciones derivadas de la investigación financiada con fondos públicos, lo que genera mayores beneficios para la sociedad aunque, a su vez, puede haber dificultado el acceso de los investigadores a ciertos tipos de tecnologías privadas. Los gobiernos deben garantizar el acceso a las invenciones básicas, por ejemplo, mediante la monitorización de las solicitudes de patentes.
- **Biotecnología:** Parece especialmente relevante reseñar que, tras la irrupción de la pandemia COVID-19, el aumento de la innovación, especialmente por parte de *Start-ups*, se beneficia enormemente de la posibilidad de obtener protección mediante patentes. En ciertos campos, como la investigación genética, hay casos en los que las patentes aún pueden impedir el acceso a la tecnología. Es necesario revisar la calidad (novedad) y la amplitud de las patentes en estas áreas. Los gobiernos pueden explorar formas de fomentar medios alternativos de difusión del conocimiento, como el dominio público, y mejorar la difusión de las invenciones patentadas, por ejemplo a través de la promoción de grupos de patentes.
- **Defensa:** La identificación de patentes en materia de defensa puede parecer

---

<sup>25</sup>[OCDE04] 2004 Patents and Innovation: Trends and Policy Changes

trivial para analistas cuyo marco de referencia se sitúa en la UE debido a la disminución de actividad bélica en el último siglo. Sin embargo, estados en plena situación de guerra o escalada de tensión, pueden ver la utilidad inmediata en la identificación de tendencias tanto en armamento como en elementos estructurales de protección.

- **Software:** El software y los servicios asociados a este son temas relativamente recientes en cuanto a patentes. La calidad y la amplitud de las patentes de software también deben ser monitorizadas, y las oficinas de patentes deben continuar sus esfuerzos para sistematizar su experiencia y base de conocimiento. También se debe evaluar el papel de las patentes en el mundo en expansión del software de código abierto.

Pese a su relevancia, parece no existir aún una corriente clara de investigación enfocada en la explotación de los datos contenidos en los solicitudes de protección de propiedad intelectual. Es por ello, que merece la pena una reflexión exhaustiva y una agrupación de la literatura actual en este campo. De esto nos ocuparemos en el capítulo siguiente.

# Capítulo 3

## Estado del Arte

En el Capítulo 3, se hará un repaso sobre la literatura disponible en la actualidad y la hoja de ruta tecnológica para explorar el futuro del análisis de patentes. Se identifican en este capítulo, once tecnologías de aprendizaje automático, que los expertos de la industria creen que es importante adoptar a un ritmo mayor para el dominio de análisis de patentes. Mientras que otras aplicaciones de la inteligencia artificial ya han adoptado tales tecnologías, el dominio de análisis de patentes parece estar poniéndose al día.

En las últimas dos décadas, han existido desarrollos sustanciales en el campo del análisis de patentes. Si bien los datos de patentes se han considerado durante mucho tiempo el mayor depósito de información tecnológica del mundo, con su digitalización desde el proyecto BACON<sup>1</sup> en 1984 y numerosas mejoras graduales y acumulativas de la calidad de los datos y las técnicas analíticas en las últimas décadas, los datos de patentes se han vuelto cada vez más accesibles y útiles para un público no especializado. Tras el boom de la inteligencia artificial (IA), aprendizaje automático (Machine Learning) y el aprendizaje profundo (Deep Learning), se han explotado varias de estas técnicas para analizar los datos de patentes locales e internacionales.

---

<sup>1</sup>[USCO84] 1984, Oversight of the Patent and Trademark Office

Los datos aportan valor añadido para permitir una economía competitiva basada en información, pilar de la revolucionaria Industria 4.0. El Big Data está cada vez más disponible en todos los campos relacionados con ingeniería y operaciones, y la Vigilancia Tecnológica no es una excepción. La mayor disponibilidad de estos presenta una oportunidad para una mejor toma de decisiones y desarrollo de estrategias, con el fin de introducir la próxima generación de tecnologías innovadoras y disruptivas.

### 3.1. Gestión de datos

Antes de comenzar a analizar la investigación disponible acerca del análisis de datos contenidos en solicitudes de patentes, merece la pena reflexionar sobre cómo se gestionan dichos datos y las líneas de investigación que llevaron al actual método de gestión de estos. La mayor parte de la literatura hace uso del proceso definido por Moehrle<sup>2</sup>, y consta de tres etapas principales: la etapa de preprocesamiento, la etapa de procesamiento y la etapa de análisis. Las tres fases se reflejan en la Figura 3.1

1. Preprocesamiento: los datos se recopilan y, después de la extracción de información, se limpian y se preparan para el tratamiento posterior
2. Procesamiento: el análisis de los datos extraídos en la etapa de preprocesamiento se lleva a cabo utilizando diferentes métodos para clasificar, agrupar e identificar ideas significativas de la información
3. Análisis: los resultados y la información de la etapa de procesamiento se visualizan y evalúan para respaldar la toma de decisiones estratégicas.

Raturi<sup>3</sup> es el primero en argumentar que la hoja de ruta planteada por Moehrle no es un simple diagrama de flujo para la analítica de patentes, sino que es un proceso que debe ser complementario a cualquier ciclo de innovación, y que el análisis de los datos de propiedad intelectual tiene muchas aplicaciones en diversos campos

---

<sup>2</sup>[MOER10] 2010. Patinformatics as a business process, M.G. Moehrle

<sup>3</sup>[RATU10] 2010, Patinformatics - an emerging scientific discipline, M.K. Raturi





**Figura 3.1:** Proceso de Moehrle para tratamiento de datos [Elaboración Propia,2020]

industriales.

Bonino<sup>4</sup> es el primero en hacer hincapié en los campos de fecha contenidos en las solicitudes de patente, así como en vincular el ciclo de vida de la patente con las fuentes de información. Argumenta, a su vez, que un proceso de análisis de patentes es un proceso impulsado por un propósito, que consiste en tareas de búsqueda (creación de un producto mínimo viable, posibles infracciones...), tareas de análisis (micro y macro evaluación del valor comercial, evaluación técnica y sugerencias de tecnología ) y tareas de monitorización y control posterior.

Fue también en 2013 cuando Baglieri y Cesaroni<sup>5</sup> sostuvieron que el análisis de patentes es la clave para la vigilancia tecnológica con el fin de apoyar la toma de decisiones. Utilizaron la investigación de Bonino<sup>6</sup> para relacionar las tres tareas de análisis de patentes: búsqueda de patentes, análisis de patentes y monitorización de patentes, con el valor de la información de estas en el proceso de generación de innovación libre (no restringida por licencias o cánones).

En los últimos años, sin embargo, el equipo de investigación más prolífico, a la cabeza en la dirección y coordinación de proyectos dedicados a la Vigilancia Tecnológica sin entrar en la generación de contenido es el liderado por L. Aristodemou <sup>7</sup>,

<sup>4</sup>[BONI13] 2013, Review of the state-of-the-art in patent information, D.Bonino

<sup>5</sup>[BACE13] 2013, Capturing the real value of patent analysis, D.Baglieri, F. Cesaroni

<sup>6</sup>[BONI13] 2013, Review of the state-of-the-art in patent information, D.Bonino

<sup>7</sup>[ARIS17] 2017, Aristodemou, Leonidas; TIETZE, Frank. Exploring the future of patent analytics. Cambridge, UK,

que ha tratado de resumir los desafíos que enfrentará la Vigilancia Tecnológica en el futuro que, aún dando un punto de vista general, dificulta la extracción de planes de acción específicos para empresas o gobiernos

## 3.2. Literatura relacionada

Actualmente, las tecnologías disponibles para el análisis de datos de patentes son numerosas dependiendo del tipo de información que se quiera obtener: regresión<sup>8</sup> (tanto lineal como no lineal) para inferencia de datos numéricos, regresión logística sigmoidea<sup>9</sup>) para clasificaciones de tipo binario, máquinas de soporte vectorial para clasificación múltiple<sup>10</sup>, o incluso redes neuronales<sup>11</sup> para análisis de textos y dibujos. Dado que la capacidad de procesamiento de los ordenadores domésticos y profesionales actuales permite gestionar grandes volúmenes de datos con relativa rapidez, es necesario identificar primero las líneas de investigación principales en el campo de la analítica de datos de patentes

Existen varios métodos analíticos en la literatura técnica del último lustro que han sido utilizados para tratar datos brutos de patentes y propiedad intelectual. Varios estudios se limitan a proporcionar una revisión exhaustiva de la literatura sobre las técnicas de análisis de patentes<sup>12</sup>, donde distinguen entre los enfoques de minería de texto y visualización y la aplicabilidad a datos estructurados y no estructurados. Todos ellos coinciden en la necesidad de desarrollar una tecnología capaz de gestionar los diversos campos que ofrecen las solicitudes de patente para encontrar patrones ocultos o tendencias que no sean apreciables a simple vista.

---

<sup>8</sup>[SJUN13] 2013, Emerging Technology Forecasting, S.Jun

<sup>9</sup>[SBAS10] Discovery of factor influencing patent value, S.Bas

<sup>10</sup>[JUN14] 2014 Jun, S., Park, S. S., Jang, D. S. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications

<sup>11</sup>[TRAP06] 2006, Development of a patent document classification platform using a back-propagation network, A.Trappey

<sup>12</sup>[ABB14] 2014, Abbas, A., Zhang, L., Khan, S. U. A literature review on the state-of-the-art in patent analysis. World Patent Information

Esta es la razón por la cual los analistas de datos, actualmente, se enfocan en la búsqueda de algoritmos no supervisados como los algoritmos de Clusterización. Es destacable entre estos, que la gran mayoría de artículos técnicos se centran en ofrecer soluciones basadas en Redes Neuronales Artificiales (RNA), o máquinas de soporte vectorial (SVM). Se pueden encontrar ejemplos de esto haciendo algunas investigaciones de los principales autores que han generado literatura recientemente: D. Malpure<sup>13</sup>, A. Supraja<sup>14</sup> y K.M. Zalanyi<sup>15</sup>.

De la revisión llevada a cabo en 2018 por L. Aristodemou y la Cátedra *IP analytics, artificial intelligence and machine learning for technology strategic decision making* de la University of Cambridge,<sup>16</sup>, surgen cuatro corrientes de investigación en las que se implementa el uso de métodos de aprendizaje automático en datos relacionados con la propiedad intelectual:

1. **Artículos referentes a la gestión del conocimiento:** se centran en la evaluación de patentes y la clasificación de la calidad de estas.
2. **Gestión de la tecnología:** incluye la patentabilidad tecnológica, la planificación de I + D dentro de las organizaciones, la inteligencia tecnológica, incluida la monitorización de los cambios tecnológicos, la identificación y el pronóstico de las tecnologías emergentes
3. **Valor económico de la propiedad intelectual:** trata, desde un punto de vista financiero, la importancia de la propiedad intelectual en organizaciones y su impacto en otras áreas como, por ejemplo, la ley

---

<sup>13</sup>[MALP17] 2017, Dinesh Malpure, Yogesh Botre, Darshan Bhansali, Rohan Bhagi, "Patent Trend Analysis and Future Prediction" 2017.

<sup>14</sup>[SUPR15] 2015, A. Supraja, S. Archana, S. Suvetha, "Patent Search and Trend Analysis", IEEE International Advance Computing Conference (IACC),

<sup>15</sup>[ZALA12] 2012, Zalányi, Kinga Makovi, Zoltán Somogyvári, Katherine "Prediction of Emerging Technologies Based on Analysis of the U.S. Patent Citation Network"

<sup>16</sup>[ARI18] 2018, Aristodemou, L., Tietze, F. The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data

4. **Categoría híbrida:** incluye la extracción de información y una gestión eficaz de la información, que se concentra en la extracción de características de patentes, como fórmulas y figuras químicas, o la clasificación efectiva de patentes en áreas tecnológicas

Los límites que propone L. Aristodemou entre estas categorías son permeables y pueden superponerse, ya que las categorías están relacionadas entre sí y un artículo podría pertenecer a más de una de ellas. Sin embargo, y dada la naturaleza de este proyecto, es el segundo campo el que nos ocupa a la hora de analizar el estado del arte.

Del mismo modo, otros autores<sup>17</sup> utilizan un enfoque holístico para analizar artículos, documentos y patentes sobre tecnologías en desarrollo, con el fin de identificar tendencias científicas y tecnológicas, pero que por su naturaleza *ad-hoc* pueden quedarse cortos en términos de escalabilidad.

Trabajos más recientes<sup>18</sup> discuten los beneficios y las limitaciones de los enfoques de aprendizaje automático en el análisis de patentes a nivel industrial para crear una visión general de las tendencias actuales dentro de la industria. De esta manera, pretenden identificar tendencias tecnológicas y pronosticar tendencias futuras. Este artículo resulta especialmente interesante para el caso que ocupa este proyecto, ya que, aunque centrado en la base de datos USPTO, puede ayudar a orientar una línea de trabajo a replicar en el caso de la EPO.

De los artículos centrados en la Gestión de la Tecnología, se desprende la importancia que la identificación de las tendencias tecnológicas tiene para los encargados de la toma de decisiones en la gestión del I+D+i. Existen trabajos académicos<sup>19</sup> a partir

---

<sup>17</sup>[JUN14] 2014 Jun, S., Park, S. S., Jang, D. S. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications*

<sup>18</sup>[SUO17] 2017, A. Suominen, H. Toivanen, M. Seppänen, Firms' knowledge profiles: mapping patent data with unsupervised learning

<sup>19</sup>[THO10] 2010, Thorleuchter, D., Van den Poel, D., Prinzie, A. A compared RD-based and patent-based cross impact analysis for identifying relationships between technologies

de 2010 que proponen una metodología para hacer que el impacto de tecnológico de una invención en los campos cercanos sea más transparente, basándose puramente en un análisis cuantitativo. Dicho método elabora un informe de situación tecnológica en el momento del análisis y las tendencias de I+D de los competidores de una organización, comparándolos con el impacto tecnológico y las tendencias de la propia I+D de las compañías interesadas.

Como se ha mencionado anteriormente, existen también aproximaciones a la identificación de tendencias utilizando redes neuronales<sup>20</sup>, donde se suele utilizar una red neuronal combinada con el algoritmo Girvan-Newman, para construir un mapa que visualice la evolución tecnológica dando información sobre el ciclo de vida de una industria concreta. El análisis de dicho ciclo de vida es importante para las estrategias de inversión relacionadas con la empresa, como la monitorización de tendencias tecnológicas. En esta misma línea, el grupo de trabajo de C.Lee<sup>21</sup> propone un análisis de ciclo de vida de tecnología que utiliza múltiples indicadores de la patente para examinar la progresión de una tecnología determinada. Los autores emplean un modelo oculto de Markov para estimar la probabilidad de que una tecnología se encuentre en una determinada etapa de su ciclo de vida e identificar patrones.

El auge de la Industria 4.0 parece haber creado también una tendencia al análisis profundo de los datos que se generan en maquinaria conectada. Gran ejemplo de ello es el artículo publicado por U.Govindarajan<sup>22</sup> proponiendo un enfoque de modelado de temas, basado en el algoritmo Latent Dirichlet Allocation (LDA) para construir una ontología de conceptos e identificar tendencias de desarrollo técnico y funcional para la Industria 4.0.

---

<sup>20</sup>[HYS17] 2017, H.-Y. Sung, H.-Y. Yeh, J.-K. Lin, S.-H. Chen, A visualization tool of patent topic evolution using a growing cell structure neural network

<sup>21</sup>[CLE16] 2016, C. Lee, J. Kim, O. Kwon, H.-G. Woo, Stochastic technology life cycle analysis using multiple patent indicators

<sup>22</sup>[GOV18] 2018, U. Govindarajan, A. Trappey, C. Trappey, Immersive technology for human-centric cyberphysical systems in complex manufacturing processes: a comprehensive overview of the global patent profile using collective intelligence

Todas estas corrientes, se reflejan en la Tabla 3.1 a modo de resumen:

Literatura técnica analizada		
Enfoque	Tecnología	Referencia bibliográfica
Regresión	Lineal Sigmoidea	[SJUN17] [SBAS10]
Clusterización	K-Means SVM	[SUO17] [JUN14]
Redes Neuronales	Girvan-Newman Backpropagation simple	[HYS17] [TRAP06]
Deep Learning	Aprendizaje Reforzado Redes Deep Belief	[TEN018] [JLEE17]
Árboles de Decisión	Regresión	[CHOI15]
Minería de Texto	NLP Ontology-based analysis	[QHAN17] [ZHAN16]
Reducción de dimensiones	LDA	[GOV18]

**Cuadro 3.1:** Literatura técnica analizada [Elaboración Propia, 2020]

### 3.3. Valoración crítica y limitaciones

Resulta evidente que existen una gran cantidad de técnicas disponibles y diferentes líneas de investigación a disposición del público interesado en la Vigilancia Tecnológica. Sin embargo, la literatura actual parece estar limitada por ciertos factores:

1. **Concentración de la investigación en Asia y Norteamérica:** el desequilibrio de investigación entre continentes, hace que la EPO se vea sujeta a un menor nivel de investigación, en favor de otras instituciones como WIPO o JPO.
2. **Deslocalización de las tecnologías:** el amplio número de ramificaciones en las técnicas de aprendizaje automático, hace que los investigadores tomen caminos variados a la hora de enfocar el problema de la Vigilancia Tecnológica.

Esto genera una atomización de las líneas de investigación que resulta en una clara falta de alineamiento entre la comunidad dedicada al análisis de datos contenidos en patentes.

3. **Falta de enfoque estratégico:** Si bien las organizaciones y gobiernos son cada vez más conscientes de la importancia de adoptar innovación basada en Inteligencia Artificial y de los beneficios positivos que esta ofrece, no logran abordarla desde una perspectiva estratégica. El resultado son, a menudo, proyectos dedicados de Inteligencia Artificial que no están planificados a nivel estratégico, no logran abordar objetivos comerciales y son inadecuados para las acciones generales de la compañía enfocadas al crecimiento y el desarrollo comercial.
4. **Falta de escalabilidad:** parece existir una clara focalización por parte de los autores en adoptar una tecnología concreta en vez de una herramienta que aúne las más útiles. Esto hace que las empresas y gobiernos no sepan apreciar el potencial de estas, ni aplicarlas a escala global o, por lo menos, en los territorios interesantes para dichas empresas y gobiernos.
5. **Falta de conexión con el mundo empresarial y gubernamental:** característico de las líneas de investigación en edad temprana, el sector académico para mostrar una fuerte desconexión con los sectores empresariales y gubernamentales, centrandose en exceso en la excelencia de la tecnología, en vez de en la mejora continua para aportar valor añadido a la sociedad.

Ningún autor ha dado una solución tangible, específica y fácilmente escalable a los problemas de vigilancia tecnológica que enfrenta el entorno competitivo de industrias de interés y las empresas dedicadas a esta actividad en la UE. Por lo tanto, se necesita un método integrado para resumir, establecer y resolver este nicho, satisfaciendo así a la demanda de rapidez en la toma de decisiones en los sectores industriales de la Unión Europea.

Este método es el que se propondrá en el capítulo siguiente y se desarrollará en

los sucesivos.



# Capítulo 4

## Descripción del modelo Patent Radar

### 4.1. Introducción al caso de estudio

El Capítulo 4 pretende proponer un sistema de Vigilancia Tecnológica mediante análisis y predicción de tendencias basado en aprendizaje automático que se pueda utilizar para proporcionar un análisis completo del estado del registro de patentes. El objetivo principal es confeccionar una herramienta útil, escalable y precisa para cualquier organización interesada en la Vigilancia Tecnológica. Además, Patent Radar pretende ayudar a cualquier usuario, ya tenga formación técnica o no, a comprender mejor el escenario de la propiedad intelectual y las solicitudes en curso, ya sea por su interés en presentar una patente o buscando una, así como ayudar a los usuarios más experimentados a obtener una revisión detallada y rápida de avances existentes en las áreas tecnológicas requeridas, obteniendo una idea de dónde y en qué área o dominio tecnológico se encuentra el mayor potencial en la futura presentación de patentes.

Los pasos clave en el desarrollo del modelo fueron:

1. **Identificación de ámbito de actuación:** se observó la necesidad creciente en gobiernos y organizaciones de contar con una herramienta capaz de identificar patrones/tendencias para Vigilancia Tecnológica con el fin de aumentar el nivel de control sobre el entorno competitivo y facilitar la toma de decisiones estratégicas.

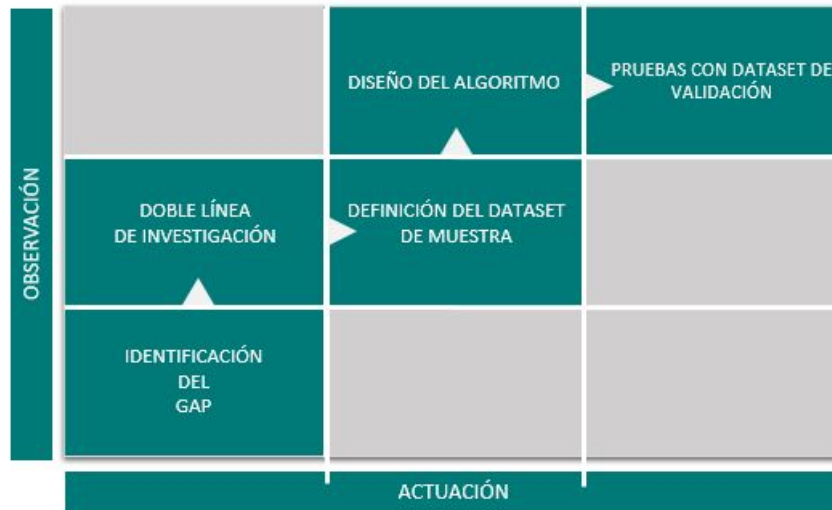


Figura 4.1: Matriz de actuación/observación. [Elaboración Propia, 2020]

2. **Doble Línea de Investigación. Patentes y Análisis de datos:** se analizó el estado del arte en el campo de la ciencia de datos y algoritmos más utilizados en la actualidad, así como un análisis detallado de los procesos que sigue la EPO, técnicas y problemas a los que se enfrenta actualmente.
3. **Definición del Dataset de Muestra:** Se estudió el conjunto de datos con el que trabaja actualmente la EPO, así como el estudio de si SQL sirve como tecnología de extracción de datos. Así mismo, se puso especial atención al proceso de limpieza de datos para, posteriormente, explotar de la forma más eficiente posibles los datos disponibles en las patentes (campos de fecha y texto plano)
4. **Diseño del algoritmo:** Tras establecer los requisitos de hardware y software, el resultado y las aplicaciones del proyecto, se realizó una estimación de tiempo. Además, se llevó a cabo la identificación, análisis y gestión de riesgos relacionados con el proyecto. Se diseñó un cronograma del proyecto que contiene tareas principales y un gráfico de línea de tiempo. La arquitectura de obtención, limpieza y explotación de datos, así como los diagramas de flujo, se confeccionaron en base a la solución del problema. Se utilizó Matlab / Octave como lenguaje para prototipado y pruebas para más adelante utilizar Python

en el diseño final del algoritmo

5. **Pruebas con el Dataset de Validación:** Se generó el análisis final de resultados con el dataset de validación (centrado en el COVID-19), así como las conclusiones del proyecto y la escalabilidad del código desarrollado.

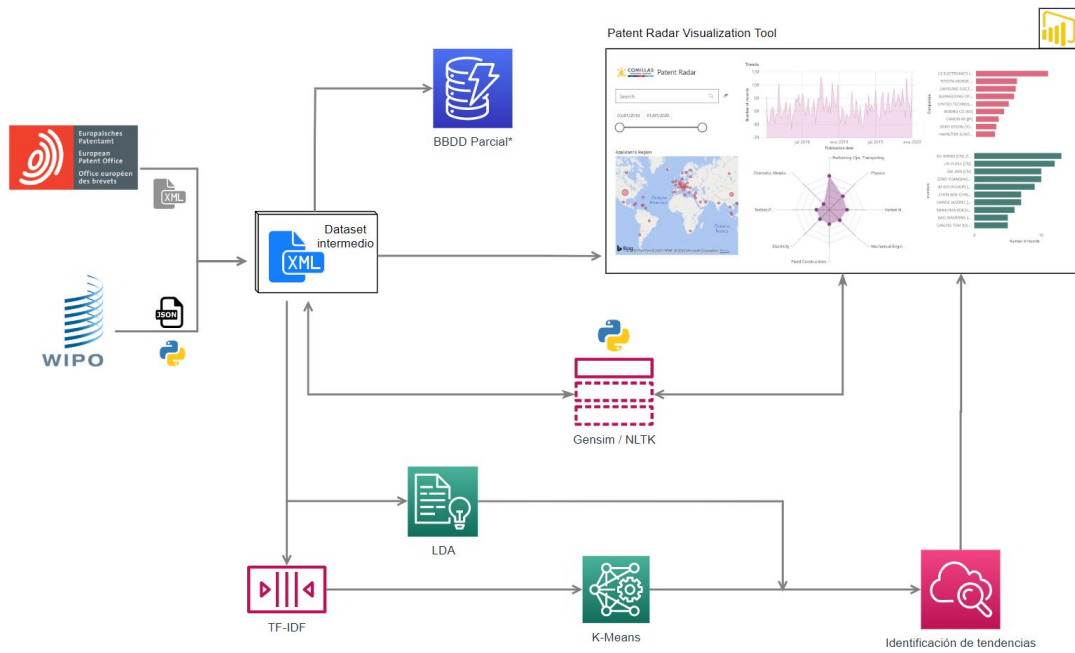
## 4.2. Objetivos del modelo Patent Radar

1. Dar una solución escalable a las principales inquietudes por parte de empresas y gobiernos en propiedad intelectual de la UE.
2. Unificar las corrientes de investigación en analítica de patentes más exitosas en una sola línea de trabajo
3. Optimizar y explotar el uso de los datos registrados en las patentes de la Oficina de Patentes Europea (EPO).
4. Definir un sistema de visualización para la identificación de tendencias explotable por cualquier organización interesada

## 4.3. Arquitectura del modelo Patent Radar y Flujo de Datos

Pare conveniente dar una imagen general de la arquitectura del modelo antes de profundizar en los detalles de cada nodo. Como puede observarse en la Figura 4.2, Patent Radar cuenta con varias etapas de procesamiento de datos. Dichas etapas no son, por definición, secuenciales sino que varias de ellas son cíclicas o redundantes. Dicha figura debe leerse, por tanto, como un sistema entrelazado y no un diagrama de flujo por pasos.

En ella, se aprecian los cuatro grandes grupos de actividades relacionadas con la gestión de los datos en las que se profundizará más adelante: la obtención (nodo compuesto por las bases de datos de EPO y WIPO), el pre-procesamiento (Gensi-



**Figura 4.2:** Arquitectura del Modelo. [Elaboración Propia, 2020]

m/NLTK), el procesamiento (LDA, y TF-IDF/K-Means) y el análisis o visualización (Power BI).

Tras comprender la arquitectura general sobre la que está construido Patent Radar, es importante ilustrar también al lector con un diagrama de flujo (Figura 4,3) sobre el proceso que siguen los datos provenientes del Abstract y el título de cada solicitud de patente, que son los campos en los que se basa Patent Radar para extraer información sobre las tendencias.

Pese a que se analizarán más en detalle cada uno de los pasos seguidos por los datos de tipo *string* en el modelo, es conveniente adoptar también una primera perspectiva general sobre el proceso que siguen desde que entran hasta que se reflejan en la etapa de visualización:

1. **Obtención de datos:** se creó un dataset de muestra basado en 5000 patentes de la EPO, con temática aleatoria y fecha de publicación entre 2018 y 2020.
2. **Tokenización de los campos de texto libre:** Los campos de texto se trajeron a lenguaje vectorial mediante tokenización, ocupando cada una de las

palabras un puesto en un vector de características.

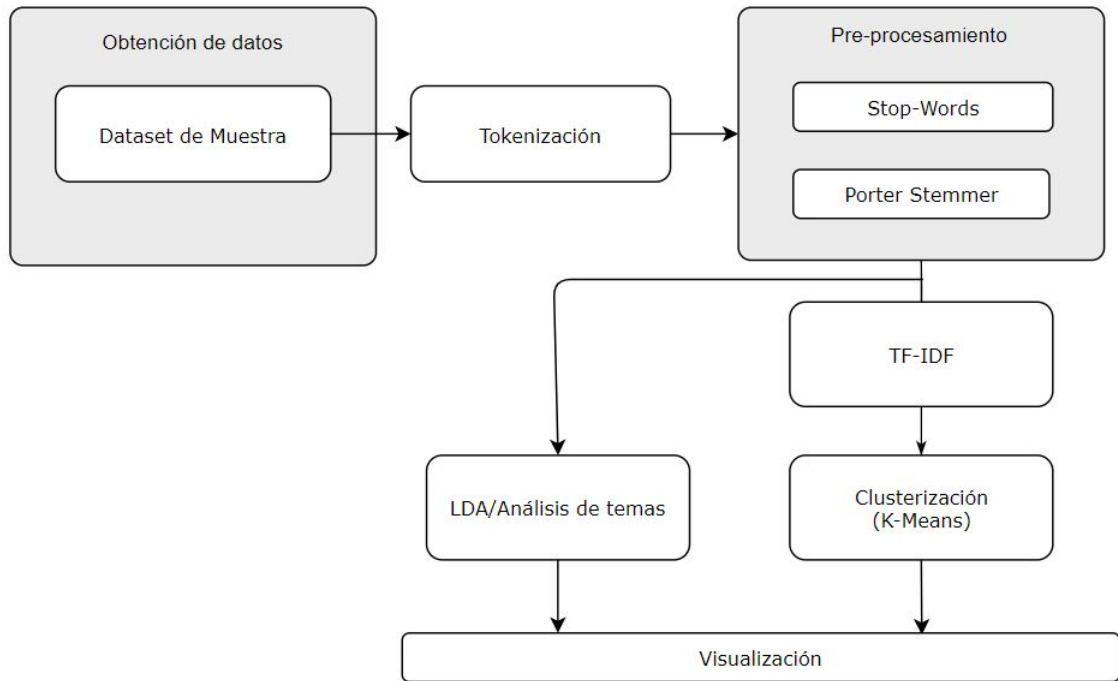
3. **Pre-pocesamiento de los campos de texto libre:** se utilizó Gensim para eliminar stop words y NLTK como algoritmo de Stemming<sup>1</sup>
4. **Análisis mediante Machine Learning:** con el fin de garantizar una utilidad doble: por un lado identificar tendencias en un dataset de patentes y por otro asignar una determinada patente a un campo concreto, se utilizaron dos algoritmos en paralelo: K-Means y Latent Dirichlet Allocation.
5. **Visualización de datos:** con el objetivo de no limitar el trabajo a la identificación de una tecnología aplicable a la detección de patentes, se utilizó Power BI como elemento central sobre el que construir la herramienta. En él podemos identificar tanto campos analizados con Machine Learning, como campos estructurados (Fechas, clasificación de la EPO, número de patentes...etc).

## 4.4. Fuentes, Tipos y Herramientas de Gestión de Datos

Las solicitudes de patentes, cuyo fin es el de proteger la propiedad intelectual, fuerzan a su vez al solicitante a desvelar su invento. En ese proceso, se solicitan y asignan una serie de campos de manera formal que ordenan y aseguran la correcta clasificación de las patentes. Son estos datos los que pueden y deben aprovecharse para analizar tendencias y patrones en sectores concretos.

---

<sup>1</sup>El Stemming es el proceso por el cual se analiza sólo la raíz de una palabra con el fin de evitar fragmentos que no aporten valor añadido



**Figura 4.3:** Diagrama de flujo de datos de tipo string. [Elaboración Propia, 2020]

#### 4.4.1. Fuentes de datos

La obtención de dichos datos se hizo mediante dos fuentes principales: La base de datos Esp@cenet de código abierto de la EPO<sup>2</sup> en combinación con la base de datos de código abierto de la WIPO<sup>3</sup>

1. Base de Datos Esp@cenet de la Oficina Europea de Patentes: si bien dicha institución ofrece registros de patentes de los últimos 25 años de forma abierta, su plataforma cuenta con una limitación de 500 registros máximos por descarga manual. Es por ello, que se realizaron 10 descargas de temas aleatorios para generar una base de muestra de 5.000 patentes sobre la que trabajar. Cabe destacar, que la institución también pone a disposición del usuario una API de pago para gestionar el conjunto completo de su base de datos. No obstante, esto es un modelo de pago y requiere una capacidad de proceso superior a la disponible para fines académicos.

<sup>2</sup>Oficina Europea de Patentes

<sup>3</sup>World Intellectual Property Organization

2. BB.DD. de la WIPO: una vez guardada la muestra de 5.000 patentes procedente de la Oficina Europea de Patentes, se procedió a utilizar la API de la WIPO para obtener los abstracts mediante .JSON. Mientras que la base de datos Esp@cenet sólo ofrece campos limitados en la descarga, WIPO permite la extracción de campos de texto de más de 50 caracteres.

#### 4.4.2. Tipos de dato

Una vez descritas las fuentes de obtención de los datos a procesar, es conveniente remarcar los tipos de dato que pueden extraerse de dichas bases. Al realizar el análisis de una solicitud de registro de patente, tratamos con varios modos de dato: Fechas (fechas de prioridad, fechas de solicitud y fechas de publicación); Números (identificador de familia, número de solicitud, número de publicación); Nombres (solicitantes, también conocidos como cesionarios o inventores); Códigos de clasificación (por ejemplo, Clasificación internacional de patentes / Clasificación cooperativa de patentes); Campos de texto libre (Título, Resumen, Descripción); Imágenes (diagramas) e Información adicional (estado legal, registro público)

En el análisis de solicitudes de patentes para Vigilancia Tecnológica, son especialmente relevantes la fecha de prioridad y la fecha de publicación. La fecha de prioridad, que es el eje central del modelo Patent Radar, tiene una doble importancia. Basándonos en el marco legal, establece el inicio de periodo de reclamo en caso de que el mismo invento sea presentado en diferentes estados adheridos al Convenio de París. En segundo lugar, y desde un punto de vista puramente basado en la identificación de tendencias, la fecha de prioridad es la mejor aproximación para identificar cuando se hizo una inversión en I+D y, por ende, la más relevante al adoptar un enfoque financiero<sup>4</sup>.

La fecha de publicación goza a su vez de especial relevancia ya que, como el *Pu-*

---

<sup>4</sup>[OCD19] 2019, Manual de Estadísticas de Patentes de la OCDE

*publication Number*, es, generalmente, la más utilizada al filtrar en bases de datos de patentes. Sin embargo, no utilizaremos dicha referencia por haber un retraso de dos a tres años desde la fecha de solicitud a la fecha de publicación. En analítica de propiedad industrial, esto quiere decir que los análisis centrados en la Fecha de Publicación muestran, generalmente, tendencias que tienen de dos a tres años de antigüedad<sup>5</sup>.

Otro de los pilares centrales del análisis es el IPC<sup>6</sup>, campo mediante el cual puede realizarse una primera aproximación estática de los temas más influyentes. Las oficinas de patentes de todo el mundo utilizan la Clasificación Internacional de Patentes (IPC) y hay diferentes códigos IPC para diferentes áreas técnicas. La Clasificación Cooperativa de Patentes (CPC) es una extensión del IPC y es administrada conjuntamente por la EPO y la Oficina de Patentes y Marcas de los Estados Unidos. CPC incluye una sección adicional Y relacionada con el etiquetado general de nuevos desarrollos tecnológicos, que también se subdivide en clases, subclases, grupos y sub-grupos. [Off20]Es importante hacer hincapié en el estatismo de dichas referencias, ya que este modelo pretende, precisamente, clusterizar nuevas tendencias o patrones que podrían no verse reflejados en el árbol estático de IPCs o CPCs.

Además de los campos relativos a fechas y el identificador IPC, existe un campo numérico asociado a la familia de patentes a la que se asocia dicha solicitud: "Family Number". Una familia de patentes es una colección de documentos entre los que se considera que cubren una sola invención. En las familias de patentes, el contenido técnico cubierto por las solicitudes se considera idéntico. Los miembros de una familia de patentes simple tendrán exactamente las mismas fechas de prioridad. Además, es importante reseñar el IPC como un indicador de áreas tecnológicas ex-

---

<sup>5</sup>2020, WIPO Manual on Open Source Patent Analytics

<sup>6</sup>La clasificación internacional de patentes, abreviada en inglés como IPC, fue establecida por el Acuerdo de Estrasburgo de 1971. Crea un sistema jerárquico de símbolos, independientes del idioma, para la clasificación de patentes y modelos de utilidad de acuerdo con los diferentes campos técnicos a los que pertenecen.



presas a través de códigos alfanuméricos (por ejemplo, C27B22 / 05 nos indica que la invención está relacionada con ADN/ARN biológico).

Respecto a los campos de tipo texto libre (en los que se centrará el análisis de aprendizaje automático referente a esta memoria) es importante destacar el título, el abstract, la información del solicitante y la información del inventor. Ellos son, sin duda, los campos que más información no visible a priori pueden proporcionarnos, ya que son generados en lenguaje natural y dependen del estilo e intención de solicitantes e inventores.

Se identificó el Abstract, o resumen, definido como un breve resumen de las ideas principales contenidas en la patente, como la fuente más potente de valor añadido por ser un gran término medio entre extensión y riqueza en contenido. El Abstract, sin embargo, no puede ser descargado automáticamente desde Esp@cenet, sino que tiene que ser obtenido de la BB.DD. de la WIPO, como veremos más adelante.

Tras una introducción a los campos absorbidos por el modelo, resulta adecuado resumir en el Cuadro 4.1 dichos elementos, para facilitar la comprensión.



Campo	Fuente	Descripción
No	EPO	Identificador orden en la descarga
Title	EPO	Título de la patente
Inventor (s)	EPO	Persona física
Applicant(s)	EPO	Persona jurídica
Publication No	EPO	País + Identificador + L + N
Earliest Priority	EPO	Fecha legal de actuación
IPC	EPO	Calsificación Internacional-
CPC	EPO/WIPO	IPC más granular. Internacional
Pub Date	EPO	Fecha de publicación
Earliest Priority	EPO	Fecha de prioridad
Family Number	EPO	Pertencia a familia
Abstract	WIPO	Resumen de la patente

**Cuadro 4.1:** Descripción de campos exportados en el análisis. [Elaboración Propia, 2020]

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau

(43) International Publication Date  
25 April 2019 (25.04.2019)

(10) International Publication Number  
**WO 2019/077119 A1**

---

(51) International Patent Classification:  
*C12Q 1/6844* (2018.01)    *C12Q 1/6862* (2018.01)

(21) International Application Number:  
PCT/EP2018/078737

(22) International Filing Date:  
19 October 2018 (19.10.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
P201731236    20 October 2017 (20.10.2017)    ES

(71) Applicants: **CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS** [ES/ES]; Serrano, 117, 28006 MADRID (ES). **INSTITUT PASTEUR** [FR/FR]; 25-28 Rue du Docteur Roux, PARIS CEDEX 15, 75724 PARIS (FR).

(72) Inventors: **SALAS FALGUERAS, Margarita**; Serrano, 117, 28006 Madrid (ES). **REDREJO RODRIGUEZ, Modesto**; Serrano, 117, 28006 Madrid (ES). **KRUPOVIC, Mart**; 25-28 Rue du Docteur Roux, PARIS CEDEX 15, 75724 Paris (FR). **FORTERRE, Patrick**; 25-28 Rue du Docteur Roux, PARIS CEDEX 15, 75724 Paris (FR).

(74) Agent: **PONS ARIÑO, Ángel**; Pon Patentes Y Marcas Internacional, S.L., Glorieta Rubén Dario, 4, 28010 Madrid (ES).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

---

(54) Title: PRIMER-INDEPENDENT DNA POLYMERASES AND THEIR USE FOR DNA SYNTHESIS

(57) Abstract: The present invention provides an isolated peptide of SEQ ID NO: 1 needed for primase active as well as new replicative DNA polymerase enzymes, preferably that of SEQ ID NO: 2, comprising said peptide. Thus, these DNA polymerases are endowed with priming activity and do not require externally provided primers for initiating and performing DNA amplification. These polymerases are able to carry out a faithful and processive de novo DNA synthesis of DNA templates in the absence of pre-synthesized primers. Therefore, these enzymes of the invention act both as primases and DNA polymerases. Furthermore, they show translesion synthesis capacity, so that they may be useful not only for whole-genome amplification but also for the amplification of damaged DNAs. The invention further refers to methods for amplifying templates DNAs using these enzymes.

Figura 4.4: Ejemplo de patente registrada por Dña. Margarita Salas. [WIPO, 2017

También se creyó conveniente adjuntar una solicitud de patente real en esta

memoria. Con ella, resulta más sencillo comprender los campos contenidos en el Cuadro 4.1. En la Figura 4.6, pueden observarse una solicitud registrada por Dña. Margarita Salas Falgueras, a quien este proyecto ha decidido rendir homenaje tras su reciente fallecimiento. Titulada "PRIMER-INDEPENDENT DNA POLYMERASES AND THEIR USE FOR DNA SYNTHESIS<sub>z</sub> registrada mediante PCT, la solicitud contiene como fecha de prioridad (primera presentación) 20.10.2017 (en formato DD.MM.AAAA), vinculada al identificador de prioridad P201731236. Además, puede apreciarse el número de solicitud de la EPO EP2018078737 y la fecha e identificador de publicación. El número de publicación( WO 2019077119 A1), es normalmente el más fácil de usar cuando se busca en una base de datos de patentes internacional y, en este caso, va asociado a una fecha de publicación 25.04.2019.

Otro tema relevante a tener en cuenta es que los campos Solicitante e Inventor (en este caso SALAS FALGUERAS, MARGARITA y CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS respectivamente) incluyen información de código de país (por ejemplo, ES) que utiliza códigos de país estándar de dos letras. Si bien esta información no siempre está disponible (especialmente para las presentaciones exclusivamente a nivel nacional), estos datos son muy útiles en el análisis de patentes para identificar colaboraciones entre inventores y solicitantes entre países y se explotan en el modelo Patent Radar para la confección de un mapa de calor en el que se profundizará más en detalle.

Se decidió también, adjuntar una extracción única en formato tabular procedente de Esp@cenet para facilitar la comprensión de cómo se generó el dataset de muestra (compuesto por más de 5000 patentes generando una base de datos relacional).

No	Title	Inventor(s)	Applicant(s)
1	Small Hydropower Generator	Kim Choong [KR]	500 VOLT INC

Earliest Priority	IPC	CPC	Pub Date	Earliest pri
10/10/2016	F03B17/06	F03B17/06 (EP,KR)	31/08/2017	31/08/2017

Publication No	Family No
KR101773513B1	59760928

**Cuadro 4.2:** Ejemplo de campos exportados a partir de la BB.DD de la EPO.  
[Elaboración Propia, 2020]

### 4.4.3. Herramientas de gestión de datos

Como se ha comentado anteriormente, se observa una tendencia creciente en el campo de la vigilancia competitiva por, no solo expertos en el tema, sino también nuevos interesados ocasionales (directivos, investigadores industriales, profesores o aficionados al análisis de datos) en el que cada uno necesita un conjunto diferente de funcionalidades y un nivel diferente de complejidad. La amplia oferta de herramientas para gestionar datos en pleno auge del *Big Data* parece abrumadora. Sin embargo, permite que los analistas puedan emplear diferentes técnicas y combinaciones de ellas para explotar el punto fuerte de cada una.

Para el análisis que ocupa esta memoria, se creyó conveniente analizar y sintetizar las herramientas revisadas y aprobadas por la WIPO<sup>7</sup> en la Tabla 3.3:

Herramienta	Propietario	Modelo	Ámbito
NLTK/Gensim	PSF	Libre	Limpieza
RStudio	RStudio	Libre	Minería
Tableau	Salesforce	De pago	Visualización
Power BI	Microsoft	Freemium <sup>8</sup>	Visualización
IBM Many Eyes	IBM	De pago	Visualización
Gephi	UTC	Libre	Redes
Python (lenguaje)	PSF	Libre	ML/IA

**Cuadro 4.3:** Herramientas aprobadas por la WIPO. [Elaboración Propia, 2020]

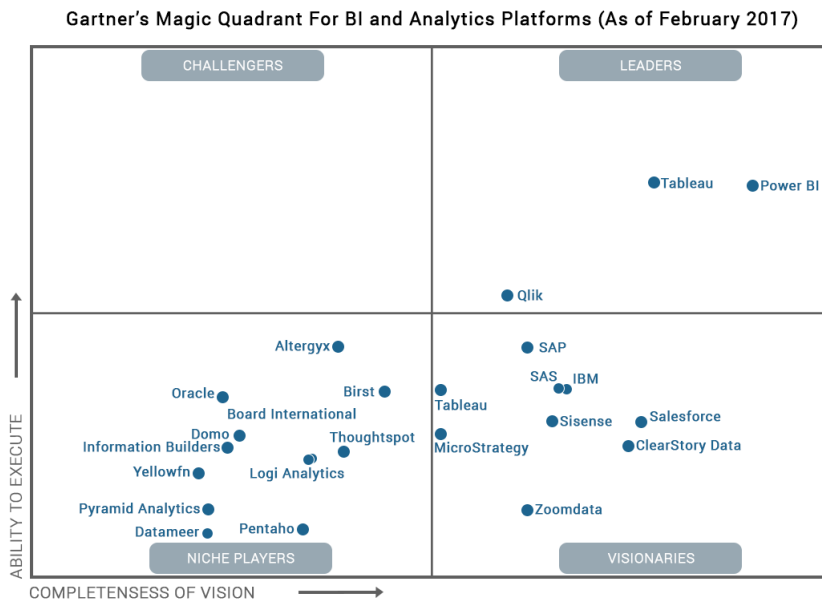
Merece una especial atención el último paso del modelo por ser en el sobre el que se construye la herramienta: La visualización de datos. Esta es una característica esencial del análisis moderno de patentes ya que el ser humano encuentra más sencillo absorber información visual que columnas y filas de números o ingentes cantidades de texto. Podemos ver y comprender más fácilmente las conexiones entre

<sup>7</sup>[WIPO16] 2016, Manual on Open Source Patent Analytics

múltiples piezas de información al inspeccionar las representaciones visuales de estas.

Al enfrentarse a un problema que requiere de una herramienta potente de visualización (capaz de gestionar grandes cantidades de datos) y observar la oferta de herramientas, uno puede verse abrumado por la cantidad, variabilidad y desequilibrio en precios. Por ello, una buena aproximación es utilizar fuentes fiables de estado de mercado como el Gartner Institute, conocido por realizar análisis exhaustivo de ciclo de vida de tendencias tecnológicas y comparaciones de herramientas en su aclamado *Magic Cuadrant*.

Parece evidente, tras analizar el cuadrante de Gartner (Ver 4.5), que una de las herramientas sobresale por encima de las demás tanto por rendimiento como por precio. Es por ello que para el análisis de este informe, se decidió utilizar Power BI como opción más razonable.



**Figura 4.5:** Comparación de herramientas de visualización de datos. Gartner Institute, 2017]

Tras un breve repaso a los campos y herramientas explotadas por el modelo, conviene resaltar que las técnicas de análisis descriptivo de datos en patentes solo son útiles para dar sentido a la información contenida en la base de datos de

partida. El verdadero valor añadido está en utilizar datos para encontrar patrones y correlaciones entre la información que se recopila y las tecnologías en auge. Una vez que hayamos encontrado esas tendencias, podemos usarlas para ayudarnos a mejorar nuestra comprensión de cómo funcionan las cosas en la actualidad, lo que lleva a solicitudes de patentes de propiedad intelectual más completas, precisas y mejor informadas. Es por ello, que se decidió adoptar un enfoque basado en aprendizaje automático para dicha identificación.

## 4.5. Algoritmos

En el apartado "Algoritmos" se pretende resumir al lector, de la forma más breve y clara posible, todos los procesos subyacentes con los que cuenta el modelo para la identificación de tendencias.

### 4.5.1. Minería de datos

Dependiendo de la arquitectura de hardware disponible, puede adoptarse un prisma u otro a la hora de decidir cómo obtener y guardar los datos. En el caso que aborda este proyecto, y por su naturaleza académica, se prefirió un nivel de complejidad no excesivamente elevado, en el que, como se ha comentado anteriormente, se descargó el dataset de muestra de forma semi-manual. Esp@cenet no cuenta con una API para desarrolladores, pero permitió la descarga manual de 500 patentes por extracción, confeccionando un total de 5000. El Abstract, sin embargo, no es un campo abierto a la descarga desde Esp@cenet de forma directa<sup>9</sup>, y por eso se decidió conectar el modelo, también, a la base de datos de la WIPO.

El IPC sirve, por tanto, como llave primaria para acceder al campo más relevante de este proyecto. A continuación el código necesario en DAX (lenguaje utilizado por el

---

<sup>9</sup>El Abstract puede, sin embargo, descargarse de la EPO previo pago para acceso en PATSTAT o mediante API request

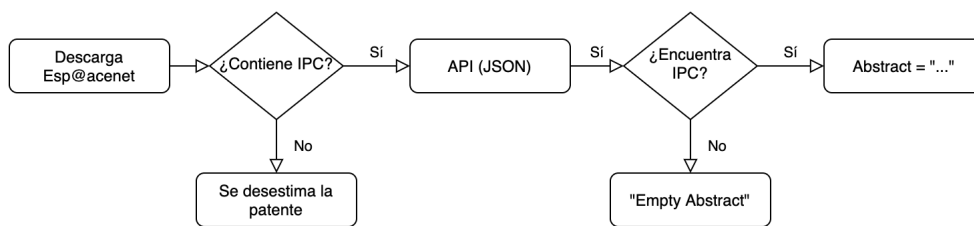
entorno Power BI) para llamar mediante JSON a la base de datos de la EPO para obtener los abstracts. Se incluye este fragmento de código por su especial relevancia:

```
(http_query_api as text) =>
  let
    Source = Json.Document(Web.Contents(https://data.epo.org/linked-data/data/publication/EP/1010425/A9.json)),
    result = Source[result],
    items = result[items],
    items1 = items{0},
    abstract = try items1[abstract] otherwise "Empty_Abstract"
  in abstract
```

La llamada a la base de datos de WIPO para obtener el abstract, es dependiente de campos ya exportados en la BBDD de EPO siguiendo la siguiente estructura:

$$https://data.epo.org/linked-data/data/publication/EP/1010425/A9.json \quad (4.1)$$

Como dicha URL es dependiente de campos contenidos en el registro de la patente, se debe generar código dinámico que busque en la URL asociada a cada patente.



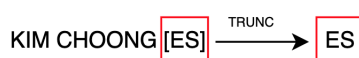
**Figura 4.6:** Diagrama de flujo de obtención de dataset completo (con abstract).  
[Elaboración propia, 2020]

### 4.5.2. Pre-procesamiento de datos

En cualquier análisis de datos, y una vez gestionado el proceso de minería u obtención de datos, la mayor parte del trabajo suele corresponder al periodo de limpieza previo a dicho análisis. En particular, los datos de diferentes bases de datos de

patentes generalmente implican diferentes desafíos a la hora de limpiar el dataset. La mayoría de estos retos consistieron en limpiar nombres de inventor y solicitante y limpiar los campos de texto libre para explotar con eficiencia los datos.

Los datos estructurados como el identificador de la patente, se truncaron para obtener el campo al que pertenecen según la EPO (extrayendo, por ejemplo, el país de origen del organismo solicitante con las 2 primeras letras de su identificador) para generar un mapa de calor.



**Figura 4.7:** Proceso de truncado en el campo "Inventor". [Elaboración propia, 2020]

Los datos de texto libre como el Título y el Abstract, sin embargo, contienen signos de puntuación, palabras mal escritas y abreviaturas, que deben eliminarse o corregirse. Las patentes analizadas, que como se explicó previamente constituyen un documento legal, contienen jerga, abreviaturas y lenguaje difícil de procesar por una máquina. Por lo tanto, el preprocesamiento<sup>10</sup>, era un paso necesario para el análisis de estas, y los procedimientos como la limpieza del texto, la expansión de abreviaturas y la eliminación de palabras se aplicaron para convertir los datos en una forma más efectiva y adecuada para el análisis.

A la hora de ejecutar un filtrado y limpieza de los datos en bruto mediante Python, nos encontramos con dos grandes y extendidas alternativas en la industria del tratamiento de datos:

1. GenSim<sup>11</sup>: principalmente usada para el modelado de temas y análisis de similitud en documentos

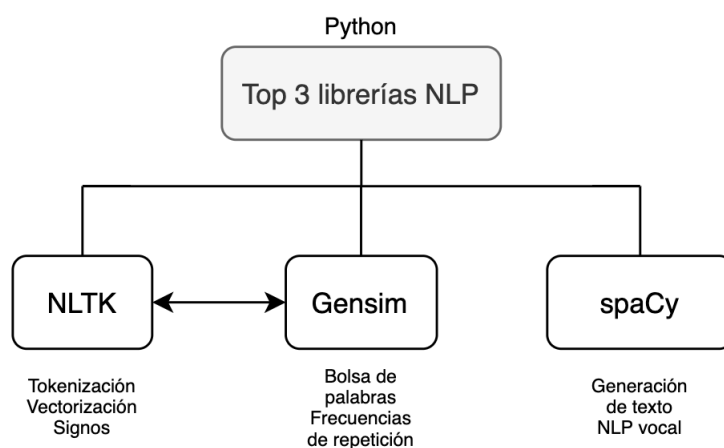
---

<sup>10</sup>Proceso que limpia y filtra el texto para cualquier análisis posterior

<sup>11</sup>Gensim, o *Generate Similar*, se desarrolló en 2008<sup>12</sup> como una colección de varios scripts de Python para personal académico. Su finalidad era ser el paso previo a generar una breve lista de los artículos más similares a un artículo determinado. Hoy en día, es ampliamente utilizada como librería NLP adicional a NLTK, para modelar temas y similitudes en documentos. Es una librería líder en el procesamiento de textos, y utiliza como modelos vectores o bolsas de palabras.<sup>13</sup> Este software depende de NumPy y Scipy, dos paquetes de Python para computación científica que deben estar instalados antes de utilizar gensim



2. NLTK: La librería de lenguaje natural NLTK o *Natural Language ToolKit*) es la biblioteca más popular<sup>14</sup> para el procesamiento del lenguaje natural (NLP) escrita en Python. NLTK destaca por su rápida curva de aprendizaje. Contiene bibliotecas de procesamiento de texto para tokenización, análisis y clasificación. NLTK se utiliza, por tanto, para tareas generales de NLP (tokenización y pre-procesamiento general del texto).



**Figura 4.8:** Top 3 librerías Python para limpieza de datos [Elaboración propia, 2020]

### Tokenización de los campos de texto libre

Una vez confeccionado el dataset de muestra con las 5000 patentes, fue necesario llevar a cabo un proceso de tokenización (Título y Abstract)<sup>15</sup> en los fragmentos de texto libre para vectorizar y ordenar partículas de texto sencillas de analizar. Un token es una instancia de una secuencia de caracteres en algún documento en particular que se agrupan como una unidad semántica útil para el procesamiento. El conjunto de tokens, permite vectorizar las unidades semánticas o palabras, para extraer conceptos. En la figura 4.10 se decidió ilustrar al lector con un ejemplo de tokenización. Si bien no todos los tokens aportan valor, este es el paso inicial para

<sup>14</sup>[SULI18] 2018, Susan Li, Topic Modelling in Python with NLTK and Gensim

<sup>15</sup>La tokenización es la tarea de dividir una cadena de texto en pedazos, llamados tokens, para al mismo tiempo desechar ciertos caracteres, como la puntuación y generar un vector que guarde

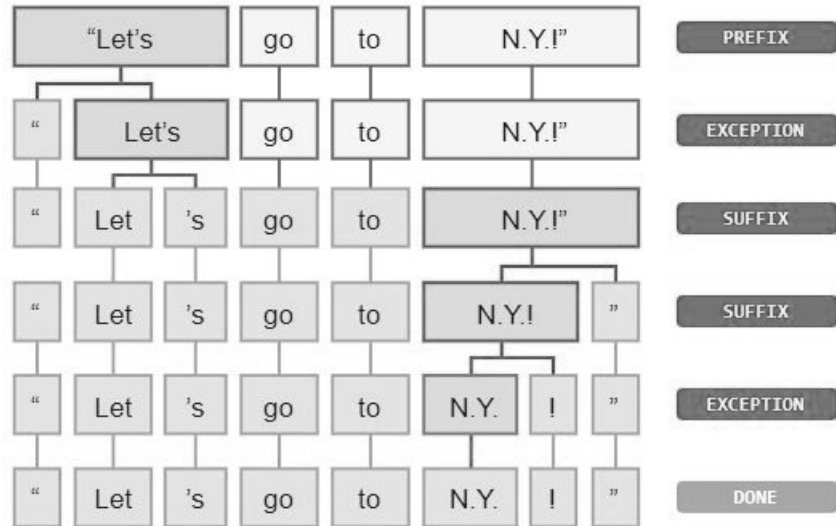


Figura 4.9: Tokenización de texto Stanford University, 2016]

sólo estudiar conceptos relevantes como se verá más adelante mediante "Bolsas de palabras".

Afortunadamente, y tras basarnos en la comparación de herramientas de tokenización llevada a cabo por Mehmet Kayaalp y reflejada en la Figura 4.8, se decidió utilizar NLTK como librería principal. Fue integrada en Patent Radar, por su capacidad para tokenizar grandes cantidades de texto con alta velocidad, estar programada en Python y ser de código abierto.

### Limpieza de datos

Una vez que se ha importado y vectorizado el conjunto de datos, el siguiente paso fue limpiar el texto. El texto de títulos y abstracts contenía números, palabras, caracteres especiales y espacios no deseados. En aras del proyecto que ocupa esta memoria, eliminamos todos los caracteres especiales, números y espacios no deseados de nuestros abstracts y títulos.

El proceso fue el siguiente:

---

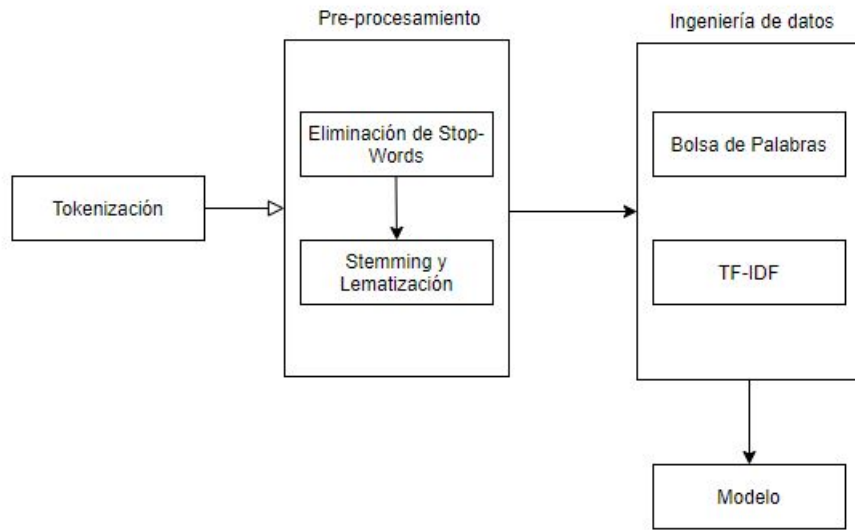
cada una de esas partículas

No	Name	Algorithm	Speed word/sec	Open Source	Language	Other
1	NLTK tokenizer	Simple splitter*	> 6000	Y	Python	
2	OpenNLP tokenizer	Maximum entropy	~ 400	Y	Java	Trainable
3	Mallet tokenizer	Simple splitter*	> 6000	Y	Java	
4	SPECIALIST NLP tokenizer	Simple splitter*	~ 600	Y	Java	
5	Gump tokenizer	Linguistic rules	> 6000	Y	Gump	
6	Dan Melamed's tokenizer	Linguistic rules	> 6000	Y	Perl	
7	Qtoken	Simple splitter*	~1500	N	Java	
8	UIUC word splitter	Linguistic rules	> 6000	Y	Perl	Sentence input
9	LT TTT tokenizer	Linguistic rules	~1000	Partial	Perl & others	Multiple input formats
10	MedPost tokenizer	Linguistic rules	~750	Y	Perl/C++	Trainable
11	Brill's POS tagger	Transformation-based error-driven learning	~300	Y	Java/C	Java wrapper
12	Stanford POS Tagger	Maximum entropy	~50	Y	Java	Trainable
13	MXPOST tagger	Maximum entropy	~200	N	Java	Trainable

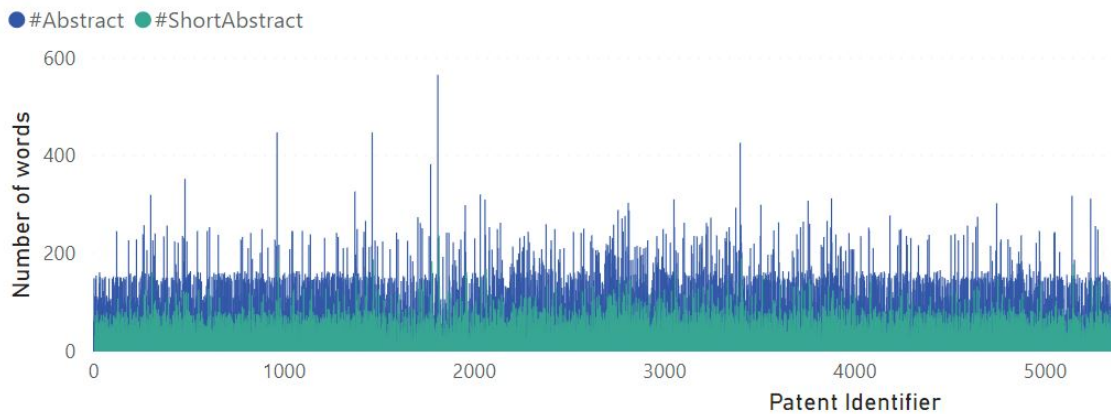
Figura 4.10: Comparación de diferentes técnicas de tokenización Mehmet Kayaalp, 2015]

1. Conversión a minúsculas: con el fin de evitar redundancia en la categorización de tendencias
2. Eliminación de signos de puntuación: por el escaso valor añadido que aportan al análisis de tendencias
3. Eliminación de símbolos incoherentes: comunes en solicitudes provenientes de países asiáticos o que, por cualquier razón, no hayan sido importadas correctamente..
4. Eliminación de palabras monocaracter: cuando eliminamos el signo de puntuación de *Engines.* y lo reemplazamos con un espacio, obtenemos *engine* y un solo carácter "s", que no tiene ningún significado. Para eliminar dichos caracteres individuales, usamos una expresión regular que sustituye todos los caracteres individuales que tienen espacios a cada lado, con un solo espacio.
5. Stop Words: proceso mediante el que eliminamos palabras que aportan escaso valor predefinidas en el diccionario de NLTK, como artículos o adjetivos. Este proceso es cíclico, y, de hecho, Patent Radar utiliza un diccionario de stop-words personalizado basado en conceptos repetitivos y de escasa utilidad.
6. Stemming mediante algoritmo Porter: o reducción de palabras a su raíz para

mejorar el recall<sup>16</sup>



**Figura 4.11:** Diagrama simplificado: Flujo de texto libre. [Elaboración propia, 2020]



**Figura 4.12:** Reducción de número de palabras por patente. [Elaboración propia, 2020]

### 4.5.3. Procesamiento de texto libre. NLP.

Tras el análisis preliminar, parece evidente que los campos que más información pueden proporcionar, son aquellos constituidos por texto libre (a elección del solicitante) y los diagramas. Dada la complejidad y el estado de maduración de las

<sup>16</sup>Aumentando el recall, conseguimos clusterizar más documentos en el mismo concepto

técnicas de aprendizaje automático, se creyó conveniente enfocar la línea de trabajo hacia los campos textuales.

### Bolsa de palabras

Las máquinas, a diferencia de los humanos, no pueden entender el texto sin formato. Las máquinas solo pueden gestionar números. Particularmente, las técnicas estadísticas como el aprendizaje automático solo pueden tratar con números. Por lo tanto, necesitamos convertir nuestro texto en valores numéricos.

Existen diferentes enfoques para convertir texto en la forma numérica correspondiente. El modelo de bolsa de palabras y el modelo de inclusión de palabras son dos de los enfoques más utilizados. En este artículo, utilizaremos el modelo de bolsa de palabras para convertir nuestro texto en lenguaje máquina.

#### Carbon Fan-Out Joint For Carbon Fiberoptic Cables

Carbon	Fan-out	Joint	For	Fiberoptic	Cables
2	1	1	1	1	1

**Cuadro 4.4:** Bolsa de Palabras. Ejemplo.

Cuando se convierten las palabras en números usando el enfoque de bolsa de palabras, las palabras únicas en todos los documentos se convierten en características (*features*). Todos los documentos pueden contener decenas de miles de palabras únicas. Pero las palabras que tienen una frecuencia muy baja de aparición no son inusualmente un buen parámetro para clasificar documentos. Por lo tanto, establecemos el parámetro *maxfeatures* en 1500, lo que significa que queremos usar 1500 palabras más comunes como características para entrenar a nuestro clasificador.

En el Cuadro 4.3 se ilustra con un ejemplo real de título de patente registrada en la EPO.

Tal y como se muestra en la matriz 4.3, y siendo  $m$  el número de patentes a analizar y  $n$  el número de términos tras el pre-procesado, puede reflejarse la bolsa de palabras como una matriz de frecuencias absolutas de dimensión  $n \times m$ .

$$\mathbf{BDP} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{bmatrix} \quad (4.2)$$

### TF-IDF

Con el fin de llevar a cabo un análisis más preciso del peso de los términos independientes contenidos en las patentes y el valor añadido que estos aportan, se decidió no sólo analizar las frecuencias absolutas de cada término, sino calcular el TF-IDF<sup>17</sup> utilizado a menudo como factor de ponderación en las búsquedas de identificación de información y minería de texto. El valor TF-IDF aumenta proporcionalmente con el número de veces que aparece una palabra en una patente y se compensa con el número de patentes en el corpus que contienen la palabra, lo que ayuda a ajustar el hecho de que algunas palabras aparezcan con una mayor frecuencia general.

- TF o Frecuencia de repetición de terminos** Supongamos que tenemos un conjunto de patentes de texto en español y deseamos clasificar qué patente es más relevante para la consulta con "the rotary valve". Una manera simple de comenzar es eliminar los documentos que no contienen las tres palabras "the", rotaryz "valve", pero esto todavía deja muchos documentos. Para distinguirlos aún más, podríamos contar el número de veces que cada término aparece en cada documento; La cantidad de veces que un término aparece en un documento se denomina frecuencia de término. Sin embargo, en el caso en que la longitud de los documentos varíe mucho, a menudo se hacen ajustes. La primera forma de ponderación de términos se debe a Hans Peter Luhn (1957), que puede resumirse en: *El peso de un término que aparece en un documento es simplemente proporcional a la frecuencia del término.*

<sup>17</sup>(Frecuencia de repetición de términos – Especificidad del término), es una metodología numérica que pretende ponderar las frecuencias de aparición intentando cuantificar la importancia de una palabra para un conjunto de strings o corpus

$$TF = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (4.3)$$

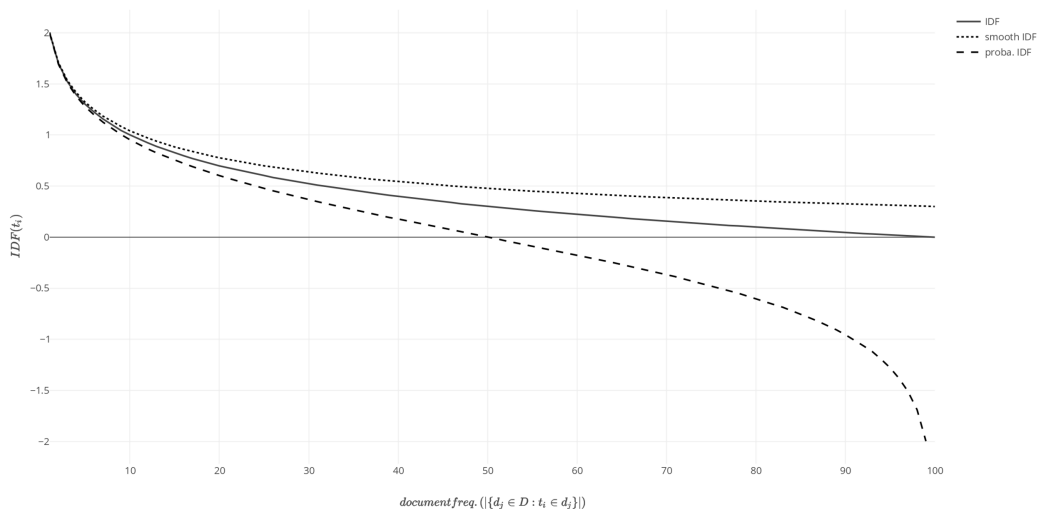
$$\mathbf{TF} = \begin{bmatrix} TF_{11} & TF_{12} & \cdots & TF_{1m} \\ TF_{21} & TF_{22} & \cdots & TF_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ TF_{n1} & TF_{n2} & \cdots & TF_{nm} \end{bmatrix} = \frac{1}{\sum_{t' \in d} f_{t',d}} * \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{bmatrix} \quad (4.4)$$

- IDF o Especificidad del término** Debido a que el término *the* es común, TF tenderá a enfatizar incorrectamente los documentos que usan la palabra *the* con mayor frecuencia, sin dar suficiente peso a los términos más significativos *rotary* y *valve*. El término "the" no es una buena palabra clave para identificar tendencias, a diferencia de las palabras menos comunes *rotary* y *valve*. Por lo tanto, se incorpora un factor de frecuencia inverso que disminuye el peso de los términos que ocurren con mucha frecuencia en el conjunto de patentes y aumenta el peso de los términos que ocurren raramente. La especificidad de un término puede cuantificarse como una función inversa del número de documentos en los que aparece, es decir: una medida de cuánta información proporciona la palabra, siendo N la suma de frecuencias absolutas de todas las patentes para un mismo término y n el número de patentes en las que aparece dicho término.

$$f^{-1} = \log\left(\frac{N}{n_t}\right) \quad (4.5)$$

$$\mathbf{IDF} = \begin{bmatrix} \log\left(\frac{N1}{n1_t}\right) \\ \log\left(\frac{N2}{n2_t}\right) \\ \vdots \\ \log\left(\frac{N3}{nn_t}\right) \end{bmatrix} \quad (4.6)$$

Con el fin de aclarar ambos términos, se cree conveniente ilustrar al lector de la memoria con un ejemplo de cálculo de frecuencias de términos y especificidad



**Figura 4.13:** Función de especificidad de términos IDF genérica. [Open Commons, 2020]

de estos tomando como dato de ejemplo el título de dos patentes reales:

Supongamos que tenemos tablas de recuento de términos de un corpus que consta de solo dos patentes. El cálculo de TF-IDF para el término *engine* se realiza de la siguiente manera:

En su forma de frecuencia sin procesar, TF es solo la frecuencia de *engine* para cada documento. En cada documento, la palabra *engine* aparece una vez; pero como el documento 2 tiene más palabras, su frecuencia relativa es menor. Un IDF, sin embargo es constante para cada término, y representa la proporción de patentes que incluyen la palabra *engine*. En este caso, tenemos un corpus de dos patentes y todos ellos incluyen la palabra *engine*. De esta forma, el término TF-IDF es nulo para la palabra *engine*, lo que implica que la palabra no es muy informativa dado que aparece en todos los documentos.

Al analizar abstracts de patentes, se descubrió que el análisis de términos simples conducía a la identificación de términos genéricos útiles para la clasificación. Sin embargo, se decidió seguir el mismo proceso de pre-procesamiento de textos con conceptos ligeramente más complejos mediante el estudio de duplas y trinomios de palabras. De esta forma, la identificación de tendencias



resultó aún más productiva, pasando de la identificación de frecuencias de términos con frecuencias altas como *engine* a términos mucho más ricos para el análisis y la clasificación de tendencias como *rotary engine valve*. Se reflejan las frecuencias absolutas en las Figuras 4.13 y 4.14.

$$\mathbf{TF-IDF} = \begin{bmatrix} TF_{11} & TF_{12} & \cdots & TF_{1m} \\ TF_{21} & TF_{22} & \cdots & TF_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ TF_{n1} & TF_{n2} & \cdots & TF_{nm} \end{bmatrix} * \begin{bmatrix} \log\left(\frac{N}{n_t}\right) \\ \log\left(\frac{N}{n_t}\right) \\ \vdots \\ \log\left(\frac{N}{n_t}\right) \end{bmatrix} \quad (4.7)$$

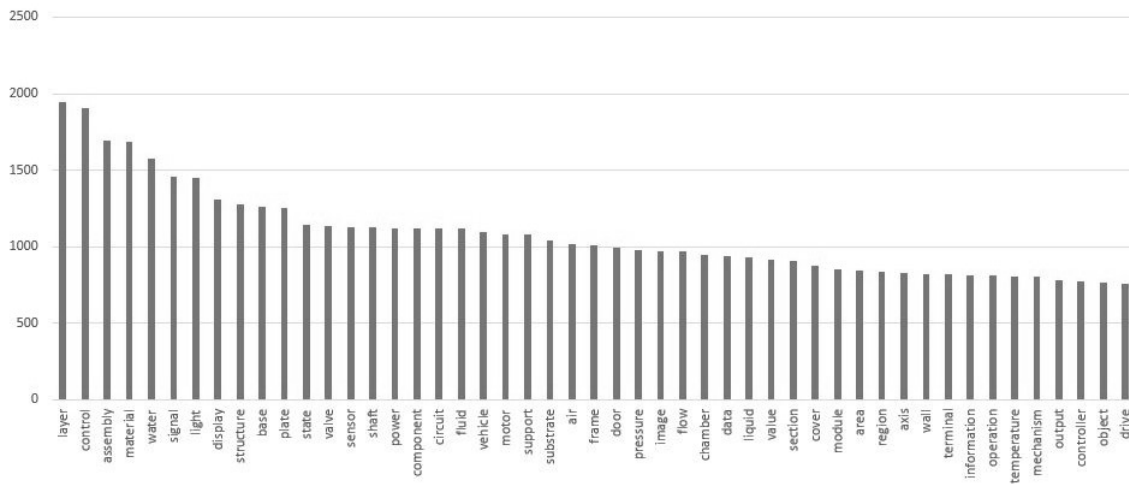


Figura 4.14: Estudio de frecuencias absolutas de términos simples. [Elaboración propia, 2020]

#### 4.5.4. Clusterización K-Means

Una vez generada la matriz TF-IDF, se utilizó K-means a modo de técnica exploratoria de análisis de datos contenidos en los campos de texto libre de las patentes por ser uno de los algoritmos no supervisados con más rendimiento en tareas de clusterización. Mediante K-Means pretendimos segregar un número N de patentes (N=5000) en un número K de clústers, haciendo una primera aproximación con una definición manual del número de clústers, para más adelante utilizar el método au-

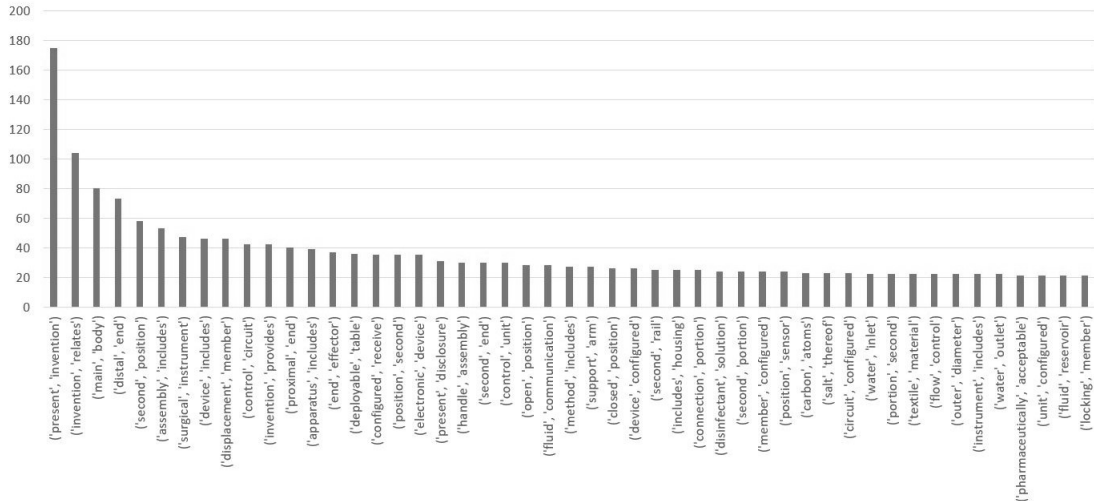


Figura 4.15: Estudio de frecuencias absolutas de términos binomiales. [Elaboración propia, 2020]

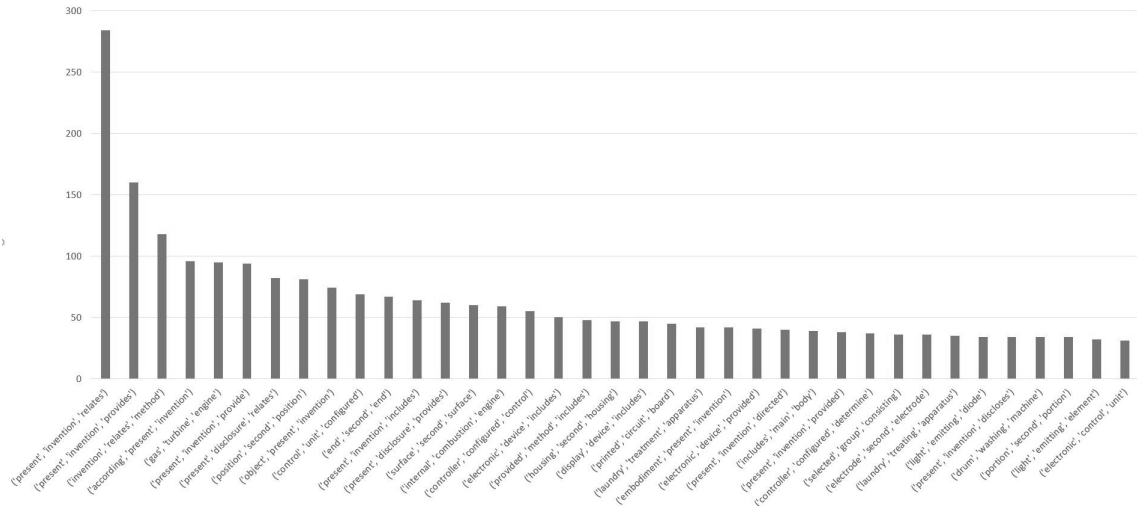


Figura 4.16: Estudio de frecuencias absolutas de términos trinomiales. [Elaboración propia, 2020]

tomatizado de detección del número óptimo mediante el *codo*<sup>18</sup> La K óptima para dividir nuestras N patentes en K clústers diferentes es aquella en las que las patentes del mismo clúster deben ser similares entre sí y diferentes de los otros clústers o clases usando algunas restricciones de similitud.

<sup>18</sup>El comúnmente conocido como método del codo, o Elbow Method es una técnica heurística utilizada para determinar el número óptimo de clústers en un conjunto de datos. El método consiste en graficar la variación explicada como una función del número de grupos, y elegir el codo de la curva como el número de grupos para usar.

Desde un punto de vista técnico, el objetivo de K-means es disminuir la suma de la distancia cuadrada entre los puntos de datos (en este caso patentes y TF-IDF de términos) y sus respectivos centros de agrupación.<sup>19</sup> [ASu15]. A continuación se detallan los pasos seguidos por el modelo para la generación de clústers:

- Paso 1 (Input): Bolsa de palabras con puntuaciones TF-IDF generada en el paso de procesamiento de datos.
- Paso 2: Inicialización aleatoria del centroide del clúster.
- Paso 3 (Cálculo mediante similitud de coseno): El conjunto de patentes en el dataset se ve como un conjunto de vectores en un espacio vectorial.<sup>20</sup> Cada término tendrá su propio eje. Usando la fórmula dada a continuación, el modelo encuentra la similitud entre dos patentes.

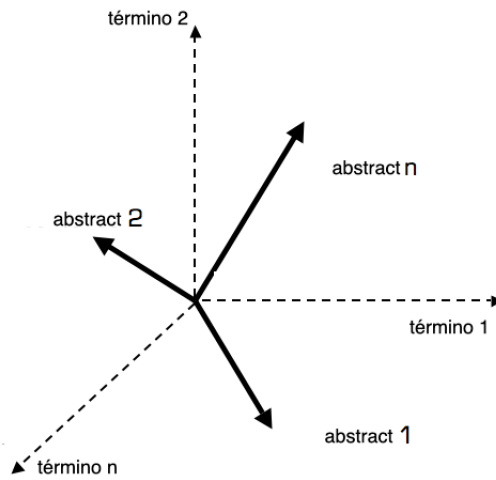
$$\cos(t, e) = \frac{te}{\|t\|\|e\|} = \frac{\sum_{i=1}^n t_i e_i}{\sqrt{\sum_{i=1}^n (t_i)^2} \sqrt{\sum_{i=1}^n (e_i)^2}} \quad (4.8)$$

Tras el proceso iterativo de asignación de clústeres, se procedió a la representación gráfica de los dichas agrupaciones. La Figura 4.15 muestra esto a modo de ejemplo para el conjunto de patentes pertenecientes al código A del IPC (Human Necessities) . Un gráfico usa PCA, que es mejor para capturar la estructura global de los datos. El otro usa TSNE, que es mejor para capturar las relaciones entre patentes con términos similares.

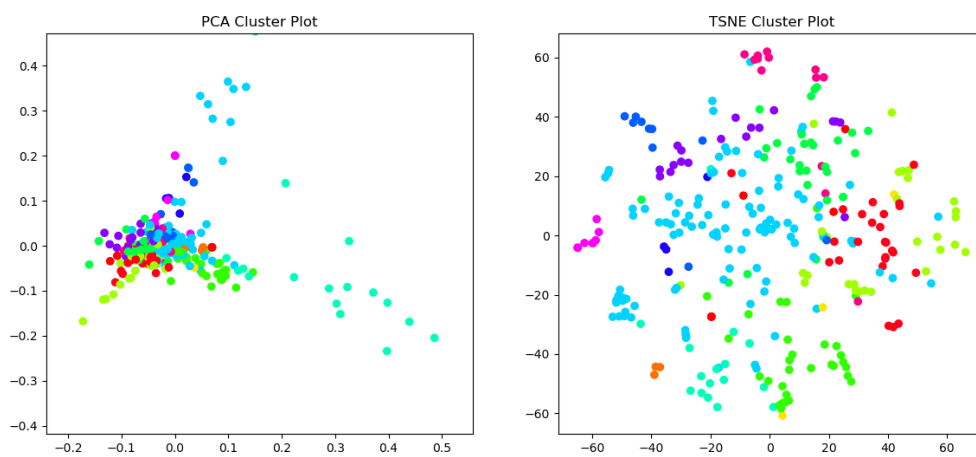
---

<sup>19</sup>La clusterización es la tarea de agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (llamado clúster) sean más similares entre sí que los de otros clústeres.

<sup>20</sup>La similitud de la del coseno es una métrica de similitud que pretende visualizar los términos como puntos en el espacio. Cuando dos términos son similares, tendrán calificaciones similares, por lo que estarán relativamente cerca en el espacio, al menos, estarán aproximadamente en la misma dirección desde el origen. El ángulo formado entre estas dos líneas será relativamente pequeño. Por el contrario, cuando los dos usuarios son diferentes, sus puntos serán distantes y probablemente en diferentes direcciones desde el origen, formando un gran ángulo. Este ángulo se puede usar como base para una métrica de similitud entre dos patentes. El coseno del ángulo conduce a un valor de similitud



**Figura 4.17:** Definición gráfica de espacio vectorial de términos y abstracts.  
[Elaboración propia, 2020]



**Figura 4.18:** Clusters para IPCs tipo A [Elaboración Propia, 2020]

Por último, se generó un paso automático para mostrar al usuario las palabras clave principales según su puntuación TF-IDF con el fin de detectar tendencias. Se hizo calculando un valor promedio en todas las dimensiones en Pandas, agrupadas por la etiqueta del clúster. Usando numpy, encontrar las palabras principales es simplemente ordenar los valores promedio para cada fila y tomar los de frecuencia superior superior. Como puede observarse en la Figura 4.15 a modo de ejemplo, Patent Radar muestra resultados tangibles y válidos. Los temas incluyen la explotación de niños, el fraude fiscal, los derechos civiles y los problemas ambientales se pueden inferir de las palabras clave principales. Sin embargo,

## 4.6. Visualización y control de resultados

Tras explicar los flujos de información que gestiona Patent Radar, resulta conveniente reflejar en la memoria, aunque perdiendo el dinamismo que caracteriza a un cuadro de mando de Power BI, la forma en la que los datos se muestran al usuario.

Se prestó una meticulosa atención a los formatos visuales escogidos para que la experiencia de usuario fuera la óptima y se empleó un amplio periodo de tiempo en el diseño de KPIs visuales. Se resumen todas las visuales o gráficos en el Cuadro 4.5.

Visual	Atributo
Buscador	PSF
Selector de Fecha	Fecha de prioridad
Mapa	Origen del aplicante
Contador	Identificador
Frecuencia por fecha	Fecha de prioridad
Radar de 8 puntos	IPC
Infografía de términos	Bolsa de Palabras
Frecuencia por aplicante	Aplicante
Frecuencia por inventor	Inventor
Lista resumen de títulos	Título
Gráfico de reducción de palabras por IPC	Short Abstract
Gráfico general de reducción de palabras	Short Abstract
Frecuencias absolutas de términos	Short Abstract
IPC Selector	Selector de campo IPC para K-Means
SSE by Cluster	Analizador de rendimiento
PCA Cluster Plot	Reductor de dimensiones
TSNE Cluster Plot	Clusteres

**Cuadro 4.5:** Resumen de visuales en Patent Radar. [Elaboración propia, 2020]

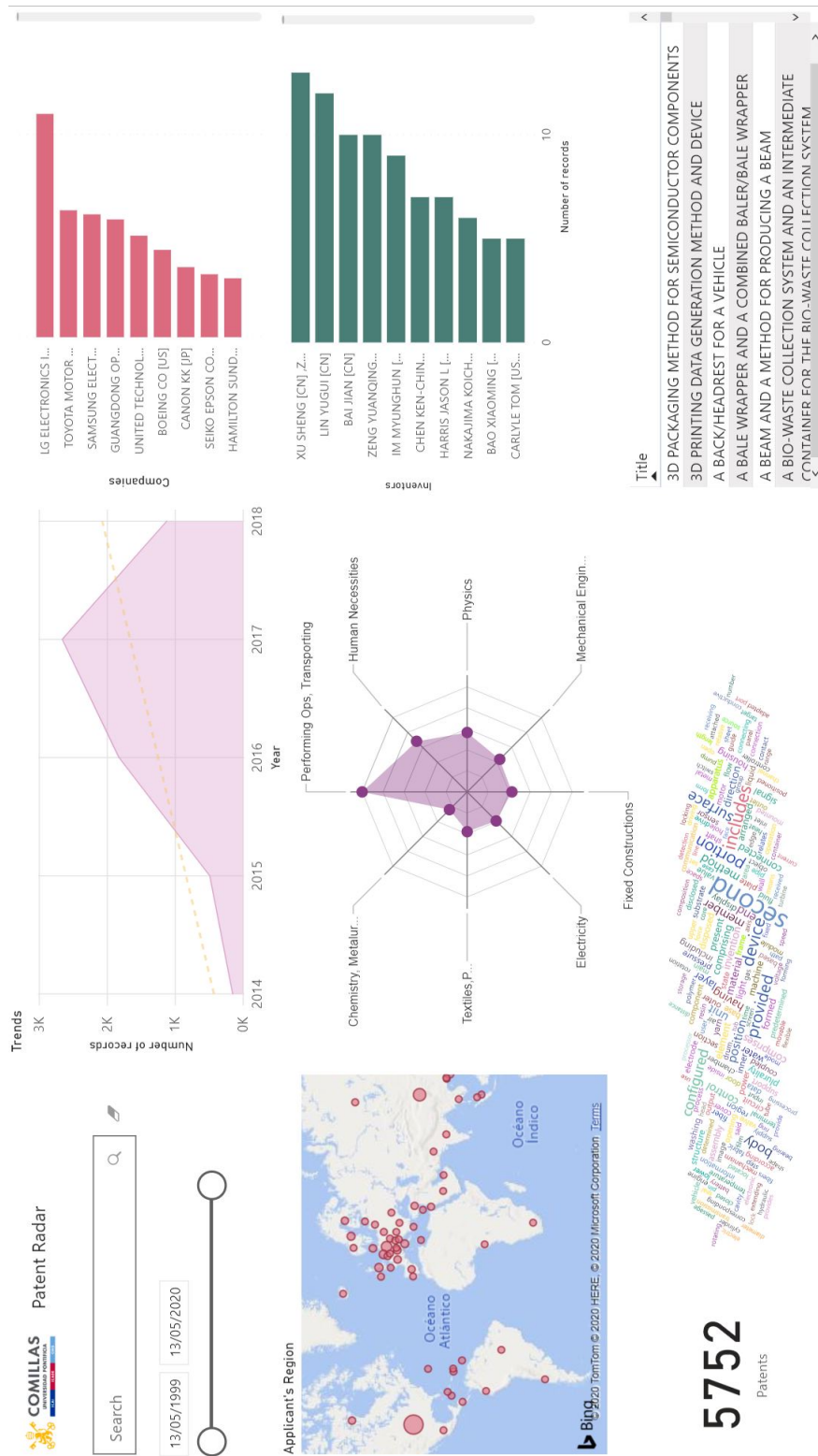


Figura 4.19: Vista A de Patent Radar (Panel Principal). [Elaboración propia, 2020]

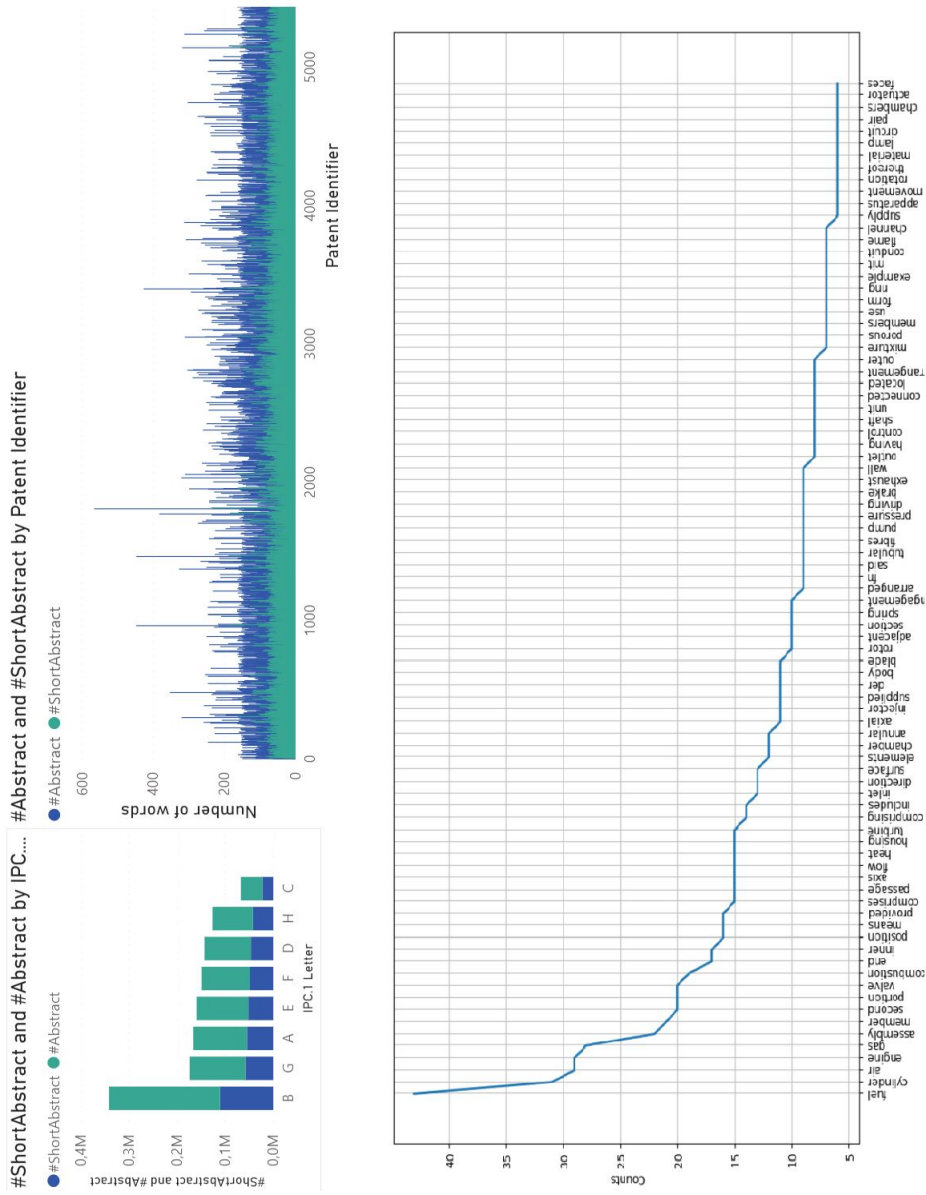


Figura 4.20: Vista B de Patent Radar (Control Pre-procesamiento). [Elaboración propia, 2020]



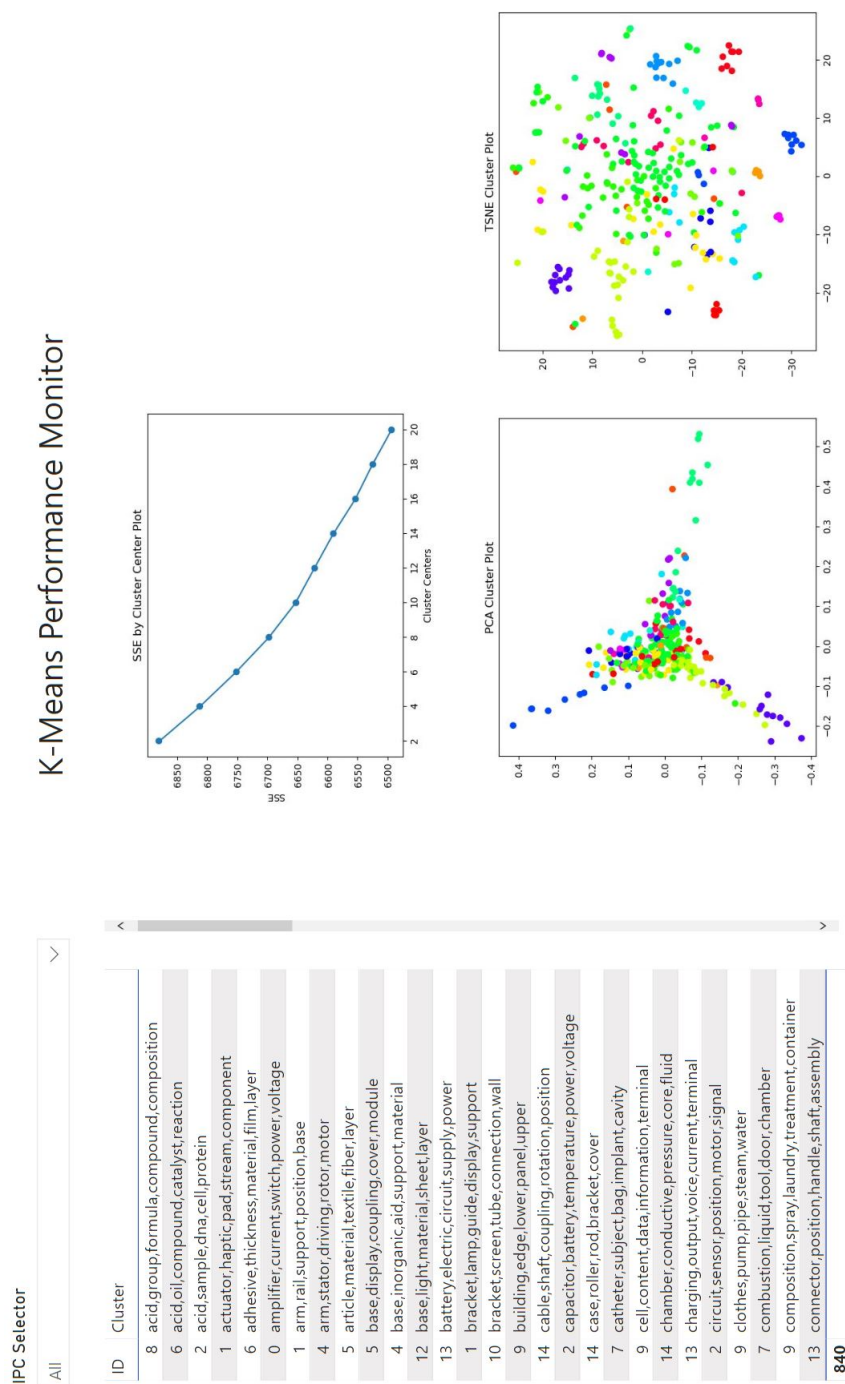


Figura 4.21: Vista c de Patent Radar (Rendimiento K-Means). [Elaboración propia, 2020]

Como puede apreciarse a lo largo del Capítulo 4, el presente Patent Radar constituye una herramienta integrada de gestión de la Vigilancia Tecnológica basada en Machine Learning y capaz de mostrar información de forma dinámica y con posibilidad de actualización en tiempo real. Dicha herramienta resulta de especial utilidad para empresas y gobiernos, en la que pueden apoyarse para identificar tendencias que pueden ser invisibles a priori.

Tras la definición de los flujos internos de trabajo que sigue la herramienta, se va a presentar un ejemplo de aplicación práctica con resultados empíricos, que ayudará al lector de esta memoria a comprender en profundidad su potencia, escalabilidad y alcance. Esto se llevará a cabo en el Capítulo 5.

# Capítulo 5

## Resultados Empíricos

El Capítulo 5 pretende servir, a modo de demostración práctica, como ilustración plausible de la potencia, escalabilidad y alcance de Patent Radar como sistema de Vigilancia Tecnológica para empresas y gobiernos. En él, se analizaron tendencias relacionadas con la pandemia asociada al COVID-19, por su relevancia social, su reciente irrupción y potencial ayuda al descubrimiento de un tratamiento o vacuna

La innovación es una de las mejores herramientas de la humanidad para combatir la amenaza del nuevo coronavirus SARS-CoV-2. Tecnologías que se han desarrollado en respuesta a pandemias anteriores, (e.g. La pandemia del VIH y la pandemia del virus influenza de 1968) pueden, gracias a la libertad a la hora de compartir información, salvar vidas hoy. Las grandes inversiones en I+D se dirigen ahora al desarrollo de nuevas respuestas tecnológicas al virus, y el salto en las innovaciones relacionadas con COVID-19 se refleja en todos los sectores imaginables, desde productos farmacéuticos hasta desinfectantes, pasando por elementos de protección o incluso mobiliario urbano

En su último manifiesto<sup>1</sup>, la EPO se comprometió a facilitar el flujo de información con el público, así como a compartir de forma rápida y veraz el conocimiento

---

<sup>1</sup>[EPOC20] 2020, Fighting Coronavirus, European Patent Office

de patentes más avanzadas sobre tecnologías que puedan ser útiles para combatir el virus o tratar la enfermedad. Para este propósito, los examinadores de patentes EPO trabajan ya, al momento de redacción de esta memoria, en mejorar el proceso de búsqueda de solicitudes de patentes para ayudar a los científicos a identificar los documentos más relevantes en estos campos técnicos.

## 5.1. Obtención de dataset de análisis

Para utilizar un dataset coherente con el objetivo de identificar tendencias en los últimos años en tratamiento de afecciones virales, se descargaron 803 patentes de Esp@cenet, cuyo desglose está recogido en el Cuadro 5.1. El dataset completo puede encontrarse en el Apéndice E.

Si bien es cierto que el COVID-19 es un tema candente en la actualidad, aún no pasado el suficiente tiempo como para poder aislar el análisis a las patentes solicitadas desde la irrupción de la pandemia. Es por ello, que el dataset contiene patentes solicitadas desde 2016 que, si bien diluyen la focalización en la pandemia de 2020, pueden incluso ayudar a identificar tendencias en la investigación contra diferentes virus de similar estructura en los últimos años. Merece la pena recordar también, que la mayor parte de patentes solicitadas, tienen como estado de solicitud China, imposibilitando aún el análisis de dichas solicitudes mediante Patent Radar hasta que un organismo oficial lleve a cabo una traducción legalmente vinculante de ellas.

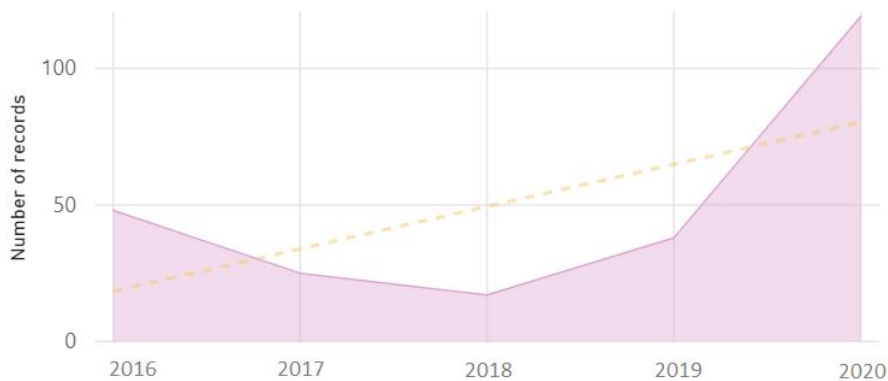
Concepto	Número de patentes
COVID-19	121
Coronaviridae	196
Sars-Cov-2	95
Coronavirus	196
Pandemic	196

**Cuadro 5.1:** Desglose del dataset de análisis. [Elaboración Propia, 2020]

## 5.2. Análisis de Resultados

Se centrará este apartado en comentar los diferentes *outputs* o elementos visuales que brinda Patent Radar con breves apartes para comentar procesos intermedios inherentes en la arquitectura del modelo.

Como puede apreciarse en la Figura 5.1, existe un fuerte repunte de las patentes solicitadas con los términos recogidos en el Cuadro 5.1 contenidos en cualquiera de sus campos de texto libre. Resulta lógico que dichos términos tengan una frecuencia superior de aparición en los últimos meses por la reorientación de los recursos económicos y gubernamentales al intento de frenado de la pandemia.

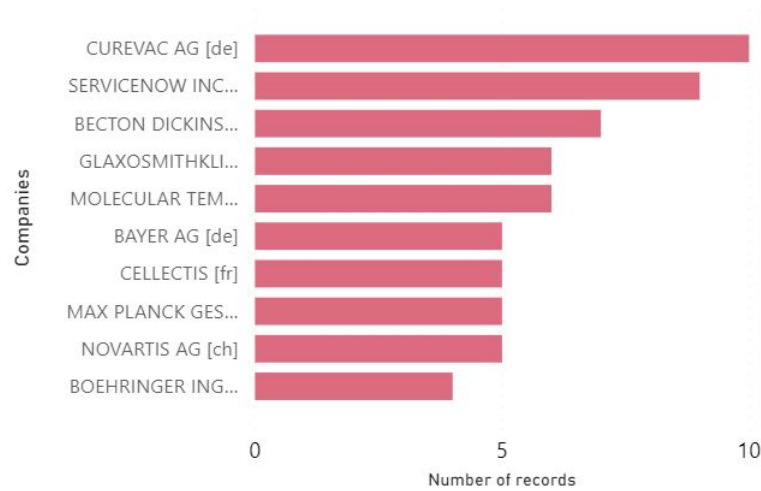


**Figura 5.1:** Evolución temporal de patentes con términos relativos al COVID-19  
[Elaboración propia, 2020]

Resulta especialmente relevante, por tanto, pararse a analizar qué organismos tanto públicos como privados, y con ellos las personas representantes de la solicitud, son los más prolíficos a la hora de pedir protección industrial. La Figura 5.2, extraída directamente de Patent Radar muestra, ordenados por volumen de solicitud, los organismos/compañías reincidentes en el dataset escogido. En él sobresalen Curevac<sup>2</sup>,

<sup>2</sup>CureVac es una compañía biofarmacéutica con sede en Tubinga, Alemania, que desarrolla vacunas basadas en ARN mensajero. El enfoque de la compañía es desarrollar vacunas para enfermedades infecciosas y medicamentos para tratar el cáncer y las enfermedades raras (Creative Commons - Wikipedia, 2020)

Servicenow INC<sup>3</sup>, y Becton Dickinson<sup>4</sup> como empresas con entre 5 y 10 solicitudes relativas al virus o a tratamientos potenciales a este.



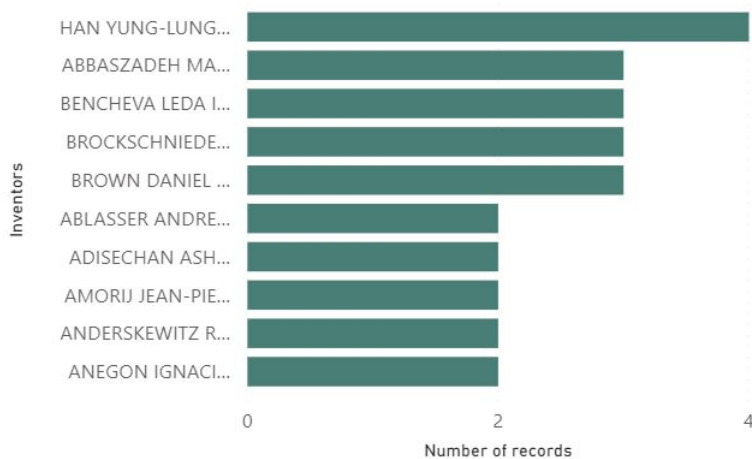
**Figura 5.2:** Top 10 organismos solicitantes relativos al COVID-19 [Elaboración propia, 2020]

Como se ha comentado anteriormente, sin embargo, no solo es importante el organismo tras la solicitud, sino la persona física representante de esta (o solicitante). La Figura 5.3 muestra, del mismo modo, y por orden de volumen solicitado, los solicitantes más relevantes en los campos relativos al virus. Destacan Han Yung-Lung (China), Abbaszadeh Masoud (Iran) y Bencheva Leda Ivanova (Italia). No es de extrañar que estos provengan de los países más azotados por la crisis sanitaria. Merece la pena mencionar en este punto, que si bien una empresa es capaz de solicitar más de diez patentes en un horizonte temporal relativamente corto, un valor alrededor de cuatro patentes para un único solicitante es extemadamente inusual por el volumen de trabajo que esto conlleva.

Resulta también destacable el reparto de patentes solicitadas por país (como

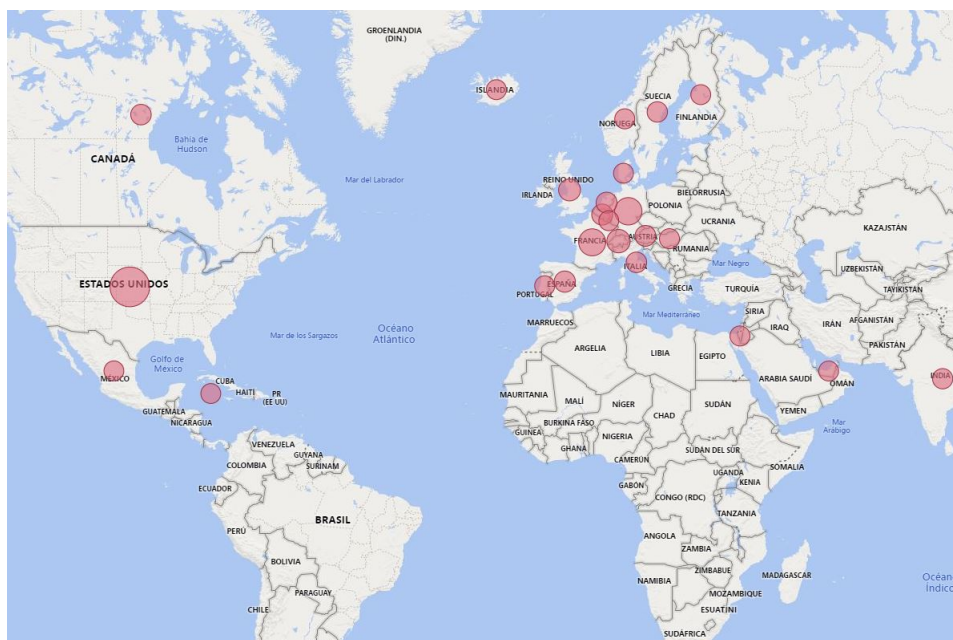
<sup>3</sup>ServiceNow es una compañía norteamericana de computación en nube con sede en Santa Clara, California. Su fundación se dio en 2004 por Fred Luddy (CreativeCommons - Wikipedia, 2020)

<sup>4</sup>Becton, Dickinson and Company, comúnmente conocida como BD, es una compañía estadounidense de tecnología médica con sede en Nueva Jersey que fabrica y vende dispositivos médicos, sistemas de instrumentos y reactivos. (CreativeCommons - Wikipedia, 2020)



**Figura 5.3:** Top 10 inventores relativos al COVID-19 [Elaboración propia, 2020]

se ha comentado anteriormente, las solicitudes Chinas han sido omitidas dada la imposibilidad de Patent Radar de tratar con caracteres no latinos). Destaca Estados Unidos a la cabeza en investigación, como era de esperar debido a la concentración de grandes corporaciones farmacéuticas. Merece la pena reflexionar también sobre la atomización de investigación en los países de la UE, en los que parece existir un interés por igual, en aportar conocimiento a la cura de virus de este tipo.



**Figura 5.4:** Distribución global de solicitudes [Elaboración propia, 2020]

Finalizado el análisis preliminar de datos, se profundizó en los términos más





valor añadido y no generar cústeres inútiles para la investigación, evitando por tanto que la clusterización se limitase a términos genéricos (i.e. invention, cure...). En el Cuadro 5.2 se muestran, por tanto, los 15 clústeres trinomiales generados por el algoritmo K-Means. De ellos, como es evidente, surgen conceptos genéricos, pero merecen mención especial los clústeres 3 (sequence, nucleic, acid), 5 (receptor, antigen, cell) y 7 (vaccine, immunogenic, polypeptide) por su carácter específico y su aporte de valor añadido a la identificación de tendencias.

Identificador del clúster	Término 1	Término 2	Término 3
1	computing	group	data
2	composition	method	sample
3	sequence	nucleic	acid
4	vaccine	influenza	virus
5	receptor	antigen	cell
6	formula	salt	compound
7	vaccine	immunogenic	polypeptide
8	gripping	step	container
9	prevention	upper	infection
10	domain	binding	ligand
11	inhibitor	treatment	disease
12	treatment	formulation	combination
13	vaccine	composition	pharmaceutical
14	particle	vector	protein
15	specifically	bind	antibody

**Cuadro 5.2:** Clústeres K-Means asociados al COVID-19. [Elaboración Propia, 2020]

Pese a que los clústeres parecen gestionar correctamente la información, el reducido número de patentes en la muestra hace que el número óptimo de clústeres sea complicado de encontrar mediante el método del codo. Esto queda reflejado en la Figura 5.6, en la que es difícil apreciar una estabilización de la curva en un punto concreto. Se eligieron 15 clústeres por resultar una medida cómoda y presentar una ligera estabilización en el gráfico de SSE<sup>5</sup>.

---

<sup>5</sup>Sum of Squared Errors, o Suma de Errores Cuadrados

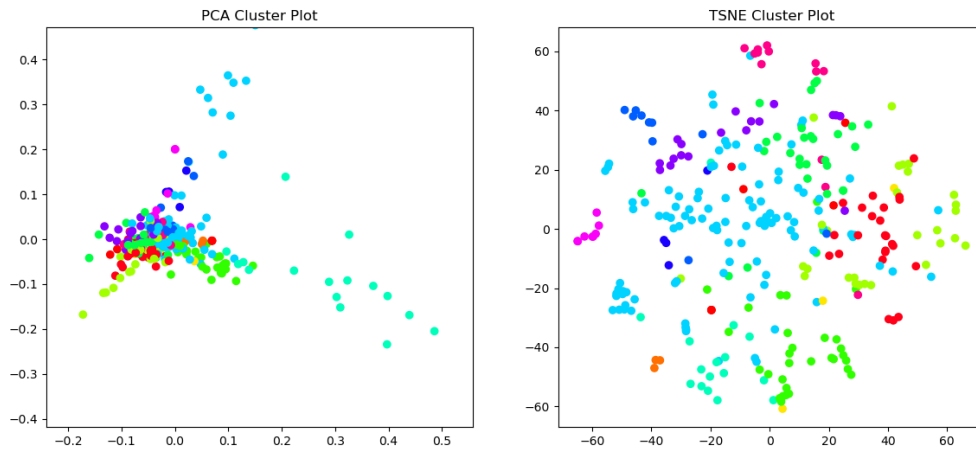


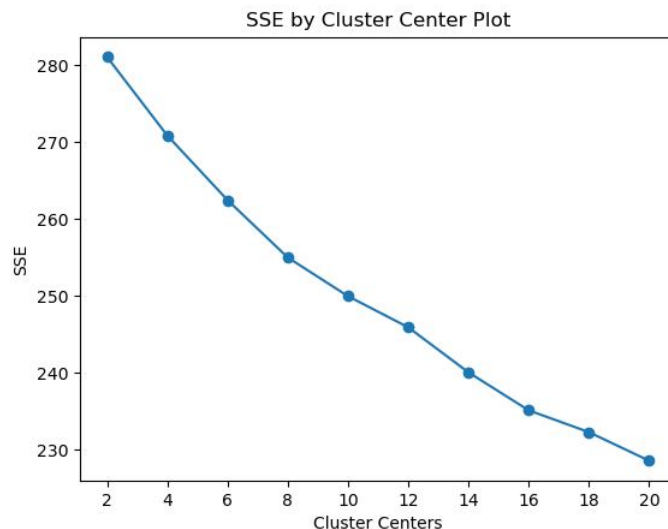
Figura 5.7: Análisis PCA y TSNE [Elaboración propia, 2020]

### 5.3. Limitaciones observadas

No debería finalizarse este análisis sin remarcar las limitaciones que se encontraron al aplicar Patent Radar a un caso práctico.

- **Necesidad de un dataset más grande para nutrir al algoritmo de clusterización:** Si bien los resultados obtenidos son concluyentes y útiles, mucho más lo serían si el modelo estuviese conectado a la base de datos total de la EPO. Recordemos que, en esta memoria, sólo analizamos un extracto reducido, haciendo que la potencia del modelo se vea ligeramente mermada.
- **Necesidad de un dataset más heterogéneo para alcanzar un número óptimo de clusters:** como puede observarse en la Figura 5.8, el algoritmo de clusterización no es capaz de calcular un número óptimo de clusters. Es por ello que se decidió adoptar un número cómodo: 15.
- **Necesidad de automatización de la conexión con el dataset de la EPO:** con el fin de disponer de todos los datos en tiempo real y con una masa crítica de patentes suficiente.
- **Generación dinámica de Stopwords:** ya que el diccionario de stopwords contenido en el modelo es estático y proviene del primer dataset de muestra.

En una versión enriquecida de Patent Radar, las stopwords han de calcularse de forma dinámica y relativa al dataset que se esté estudiando.



**Figura 5.8:** Análisis SSE [Elaboración propia, 2020]

## 5.4. Influencia de los resultados en decisiones corporativas

Partiendo de la base de que el análisis de patentes bien fundamentado es básico para la estrategia de innovación en cualquier empresa, se decidió centrar esta parte del capítulo en empresas farmacéuticas cuyos esfuerzos de investigación se dirigen hacia combatir el COVID-19. Las empresas del sector químico-farmacéutico que aparecen en la Figura 5.2 como Curevac, Glaxo Smith Kline o Bayer diseñan, de forma regular, tratamientos y productos médicos para hospitales y laboratorios, cuando estos necesitan validar procesos de producción de nuevos medicamentos. Se trata de un tema complejo que requiere alto nivel de especialización y precisión. De ahí, que sea clave utilizar el análisis de patentes como una de las estrategias de importancia para el diseño y comercialización de productos. En un momento tan convulso a nivel mundial provocado por la crisis, y en una carrera contra reloj por encontrar un medicamento capaz de frenar la evolución de la enfermedad, sino

su curación, las empresas del sector trabajan en equipos de trabajo continuos y multidisciplinares. El interés es doble: de una parte, poder contribuir a ayudar a la sociedad mundial y de otra obtener rentabilidad de su negocio.

Rank	Applicants	Applications
1	INSTITUT NATIONAL DE LA SANTE ET DE LA RECHERCHE MEDICALE (INSERM)	93
2	NOVARTIS	88
3	MERCK & CO	84
4	JOHNSON & JOHNSON	75
5	UNIVERSITY OF CALIFORNIA	64
6	BOEHRINGER INGELHEIM PHARMA GMBH & CO. KG	55
6	GLAXO SMITH KLINE	55
8	HOFFMANN-LA ROCHE	54
9	NESTLE	53
10	BAYER	46

**Figura 5.9:** Top 10 solicitantes del sector farmacéutico 2019 [European Pharmaceutical Review, 2020]

Merece la pena comentar cómo las empresas que tradicionalmente han mostrado más músculo al solicitar protección de la propiedad industrial, y que se recogen en la Figura 5.9 ordenadas por número de solicitudes en 2019, no son exactamente las mismas que las identificadas por Patent Radar, hecho que de forma inequívoca conduce a pensar que existen procesos internos en algunas de ellas que hacen su reorientación de la innovación algo lenta comparada con sus competidores.

Dichas empresas están comprobando, al momento de redactar este proyecto, cómo las ventas en Europa caen de manera notable debido entre otras razones al impacto del COVID-19 y, además, los pagos son efectuados con grandes retrasos por parte de sus clientes. Una de las estrategias sostenibles de las compañías puede y debe ser investigar en nuevos fármacos que puedan ampliar ventas a través de la captación, mediante una herramienta como Patent Radar, de nuevos clientes en el mercado global, donde algunas de ellas, ya obtenían porcentajes elevados de su facturación hasta principios de este año 2020 antes de la crisis. Se espera encontrar, en esa carrera contra reloj en la lucha contra el Coronavirus, un incremento en el número de directivos de empresas farmacéuticas que sean cada vez más conscientes de la necesidad de profesionalizar el análisis de información externa y más concretamente en el análisis de patentes.

Era común hasta ahora, encontrar en las grandes corporaciones farmacéuticas, proyectos que una vez iniciados, tienen que ser cancelados por una deficiente investigación previa de la actividad de la competencia. Estas cancelaciones suponen en ocasiones, desperdicios de recursos e incluso subvenciones con el consiguiente impacto negativo tanto en el aspecto financiero como en la imagen corporativa. Se necesita disponer de conocimientos sobre “Inteligencia y Vigilancia Tecnológica”. Algo no del todo sencillo debido a diferentes resistencias, al implantar cambios en las empresas en las nuevas formas de operar que requiere esta estrategia.

El conocimiento en profundidad del entorno del sector farmacéutico a nivel global, así como asignar un presupuesto para la implantación de un Sistema de Vigilancia Tecnológica, es clave para acometer y ejecutar nuevos proyectos de innovación e investigación. No se trata sólo de obtener los outputs del trabajo de los diferentes equipos focalizados en la innovación de una empresa, sino de dar visibilidad al resto de la compañía, de la necesidad de disponer de un sistema de vigilancia eficiente en aquellas organizaciones donde hasta ese momento, no se hubiera realizado de manera eficaz. Para ello, es necesaria una herramienta cuya experiencia de usuario, sea cuidada al máximo como Patent Radar. Cuando esto no sucede, es frecuente encontrar deficiencias como errores de diferentes índoles: duplicidades, re-trabajos, escasa comunicación o falta de acuerdos. Como puede observarse en la Figura 5.2, existe una alta atomización de las empresas que centran sus esfuerzos en la lucha contra el virus, lo que sugiere una falta de comunicación entre empresas competidoras. Si bien el sector farmacéutico es el sector competitivo por excelencia, las empresas

Es necesario, por tanto, disponer de herramientas adecuadas para tratar la robusta cantidad de datos que deben analizarse para el seguimiento eficiente de las nuevas patentes de los competidores de primer nivel. Patent Radar permite identificar los “gaps tecnológicos” en materia de innovación. Como consecuencia, esto puede ayudar a implantar una cultura de gestión basada en información objetiva de datos. Además, permite emitir alertas tempranas, al identificar en fases previas, las líneas

farmacéuticas hacia las que determinadas empresas dirigen su innovación respecto a tratamientos específicos para el tratamiento del COVID-19.

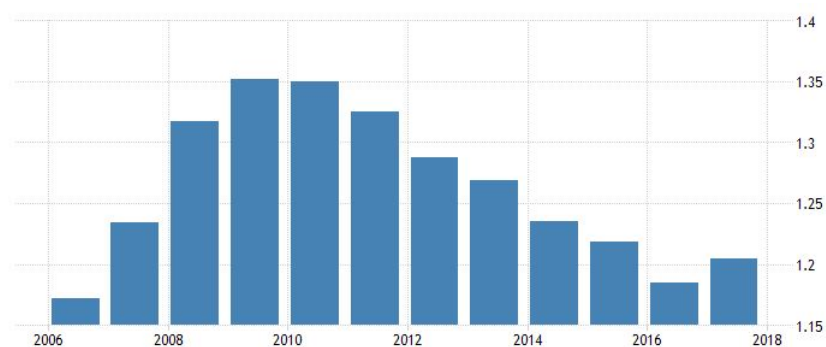
## 5.5. Influencia de los resultados en decisiones gubernamentales

Los gobiernos son los actores determinantes de la capacidad de innovación, aunque su papel y grado de participación en la innovación es discutible. La intervención gubernamental puede ser vital para apoyar el I+D y la innovación, ya que el mercado por sí solo no puede proporcionar incentivos adecuados para la producción de conocimiento. Sin embargo, los grados de intervención del gobierno varían en las diferentes economías y van desde la intervención directiva al asesorar activamente a la política industrial e invertir en áreas seleccionadas, hasta la intervención facilitadora creando un ambiente positivo y proporcionando bienes públicos para la industria y la sociedad.

Resulta curioso, cuanto menos, observar cómo en una sociedad altamente globalizada, seguimos recurriendo a los viejos estándares locales a la hora de evaluar con reticencia las líneas de investigación que siguen otros países y continentes. Patent Radar, en el pasado, hubiese arrojado datos significativos sobre la heterogeneidad de los campos de investigación en los distintos países dependiendo de sus circunstancias. Es común que países en guerra o cuyo presupuesto destinado a defensa muestren una clara predilección por conceptos relacionados con la actividad bélica, o que los países nórdicos registren una alta frecuencia de conceptos patentables relacionados con sus bajas temperaturas. Sin embargo, Patent Radar parece mostrar una cierta homogeneidad geográfica en el volumen de patentes registradas en la Figura 5.4. Parece haber sido el COVID-19 el elemento facilitador de unificación en líneas de investigación. Se espera, por tanto, una fuerte reorientación de los recursos económicos de todos los gobiernos en una misma dirección: encontrar una cura. Esto parece poder observarse ya en países del sur de Europa como España, y esta línea marcará

sin duda, tanto directa como indirectamente, los esfuerzos gubernamentales de los próximos años.

Se muestra en la Figura 5.10 una evolución temporal del presupuesto nacional español dedicado a I+D. Este presupuesto está reflejado en porcentaje sobre el PIB, eliminando de esta forma los efectos del crecimiento o decrecimiento económico. Es posible que la concienciación sobre la importancia de la innovación que ha generado el virus, ayudada de herramientas como Patent Radar, convierta este gráfico en creciente en los próximos años aunque, para ello, se necesite una gestión eficiente de los recursos económicos y una fuerte apuesta por el largo plazo.



**Figura 5.10:** Presupuesto destinado a I+D en España (% PIB) [TradingEconomics, 2020]

Parece evidente, que la redirección del PIB hacia el campo del I+D, es necesaria para que todos rememos en la misma dirección contra la pandemia. Para ello, sin embargo, los equipos gestores de gobierno necesitan más información sobre las formas en que inventores y solicitantes utilizan las patentes, por ejemplo, en lo que respecta a la implementación interna, los contratos de licencia y las estrategias comerciales con el fin de justificar dicha redirección. Es en este punto, donde deben conectarse los *outputs* de Patent Radar, con el saber hacer de los gobiernos. Gracias a este esfuerzo analítico, los políticos podrían fomentar el intercambio de experiencias entre países: existen diferencias significativas en los regímenes de patentes y muchos países han experimentado con diversos mecanismos, pero ha habido pocos intentos de sistematizar esta experiencia y difundir las mejores prácticas entre estados. Los análisis y políticas presentados en este informe también se aplican en cierta medida

a los países en desarrollo con una importante capacidad local de innovación. Estos países necesitan un sistema de patentes lo suficientemente fuerte como para atraer inversiones extranjeras directas, asegurar licencias internas y alentar la inversión local en investigación. Patent Radar es el elemento de unión necesario entre empresas cuyo núcleo de negocio es la innovación y los gobiernos, encargados de construir un sólido y sostenible tejido empresarial.

Como hemos podido apreciar en el Capítulo 5, Patent Radar es una plataforma que genera información de utilidad. Tanto la gestión de los datos generales, como la aplicación del algoritmo de Machine Learning en las cadenas de texto, pueden resultar cruciales para organismos interesados en un sector en concreto. Además, queda demostrada su utilidad en entornos de alta incertidumbre como el que vivimos al momento de redacción de esta memoria, en los que una gestión eficiente de los datos de propiedad intelectual puede agilizar la toma de decisiones y consecuentemente innovación hacia lo que podría ser una cura o tratamiento de una enfermedad a escala global. No obstante, este proyecto quedaría incompleto sin una recapitulación de conclusiones y una propuesta de hacia donde deberían dirigirse las futuras e inminentes líneas de investigación. Todo ello se aborda en el Capítulo 6.



## Capítulo 6

# Conclusiones y Líneas Futuras de Investigación

En este proyecto, se ha diseñado una herramienta innovadora a la hora de llevar a cabo diferentes niveles de Vigilancia Tecnológica: el modelo Patent Radar. Una nueva forma de enfocar la Vigilancia Tecnológica mediante análisis de tendencias en patentes basándonos en el análisis de todos los campos de estas. Se formalizó el método de control de líneas de investigación de la competencia basándonos en K-Means como algoritmo de aprendizaje automático para procesamiento de texto. Esto permitió desarrollar un marco de procesamiento de Abstracts y Títulos de patentes, que conduce a un nuevo paradigma en el dominio de información competitiva a la hora de dirigir la innovación de una empresa o gobierno. Se desarrollaron una serie de algoritmos y se construyó una arquitectura eficiente en el tratamiento de estos creando Patent Radar.

La tesis principal validada por esta disertación es que una comprensión más profunda de la naturaleza computacional de las patentes conduce a nuevos algoritmos prácticos que pueden ayudar. Dado que esta disertación tiene un componente teórico y práctico, se discuten a continuación nuestras principales conclusiones, el panorama general y las direcciones futuras para cada uno de los actores clave.

## 6.1. Conclusiones

El análisis de datos sobre patentes de empresas ha experimentado un continuo crecimiento durante varios años y está listo para ser utilizado por profesionales hasta ahora relativamente lejos de la propiedad intelectual, como aquellos dedicados al marketing, planificación estratégica e incluso análisis financiero.

Dentro de una empresa, si bien el uso de estudios de mercado y bases de datos financieras se ha vuelto tan común que ya no es realmente diferenciador, la integración de los datos de patentes disponibles antes de comenzar el ciclo de vida de los productos tecnológicos es ahora una forma de anticiparse a la competencia, al menos para las empresas activas en campos tecnológicos de alto valor añadido, así como para adaptar la respuesta táctica de uno antes de la fase de evolución de los mercados (especialmente en el caso de la aparición de tecnologías sustitutivas). El uso de datos de patentes, síntesis de indicadores y análisis del panorama de patentes mediante herramientas como Patent Radar al servicio de los tomadores de decisiones, probablemente crecerá en los próximos años.

Aunque la importancia de las patentes como elemento orientador de la innovación en entornos competitivos es bien conocida, hasta ahora no ha habido una aplicación profunda de análisis mediante métodos de aprendizaje automático. Esta disertación ha presentado uno de estos análisis en términos de procesamiento de señales arrojando las siguientes conclusiones:

- **Patent Radar** Se ha logrado una herramienta (Patent Radar) que aporta valor añadido a la hora de evaluar tendencias o patrones ocultos útiles para la Vigilancia Tecnológica. Se ha hecho el diseño lógico funcional, el desarrollo y se ha aplicado en un caso práctico con la pandemia asociada al COVID-19.
- **Las patentes como elemento central director de la innovación** Se confirma la teoría de que las solicitudes de patentes y las bases de datos actuales, permiten reorientar la forma en la que los organismos interesados en materia de innovación se gestionan. Con más de 150.000 patentes solicitadas en el mundo,

y con un crecimiento del 4% año tras año, parece que la actividad inventiva sigue una tendencia positiva. Esto, sin duda, experimentará un crecimiento aún más acusado tras la irrupción de la pandemia asociada al COVID-19 para lo que se necesitan tecnologías capaces de gestionar un volumen tan masivo de datos. Estas tecnologías permitirán una asignación eficiente de recursos por parte de empresas y gobiernos, para elevar la calidad de la innovación.

- **Utilidad de Patent Radar en la lucha contra el COVID-19** Ahora, más que nunca, cualquier desarrollo tecnológico en materia de vigilancia de la propiedad industrial, puede suponer días, semanas o incluso meses de ahorro en la búsqueda de una cura. Con unas escalofriantes cifras cercanas a las 600.000 muertes y más de dos millones de contagios totales, es el momento de que la tecnología y el Big Data entre, verdaderamente, al servicio de la humanidad. Patent Radar muestra esperanzadoras cifras de investigación y tendencias razonables en el sector farmacéutico como la focalización en ácidos nucleicos o polipéptidos inmunogénicos, que, si bien no son el objeto de estudio de esta memoria, están alineados con los últimos informes de las grandes corporaciones farmacéuticas.
- **Patent Radar como plataforma integrada de gestión de la innovación** El crecimiento continuo en volumen de solicitudes que experimentan las oficinas de patentes, combinado con la irrupción del virus, hacen que la dificultad de gestión y extracción de valor de tal volumen de datos aumente a la par. Si bien existen, por un lado, fuertes desarrollos en el campo del aprendizaje automático NLP y, por otro, herramientas de gestión de datos, es necesaria una conexión entre ambos mundos y los organismos de propiedad industrial. Patent Radar, o herramientas de índole similar, pueden ayudar a cualquier tipo de institución a adquirir un control superior sobre la adjudicación cuantitativa de sus recursos.
- **K-Means como algoritmo NLP para procesamiento de temas** Se valida la teoría de K-Means como algoritmo de clusterización semántica tras el correcto filtrado y limpieza de los datos de texto de los Abstracts. Si bien el

tamaño de la muestra es un elemento crucial a la hora de procesar los datos mediante K-Means, el algoritmo muestra clústeres razonables al tomar como *input* una muestra de 1000 patentes. Su rendimiento, por otra parte contrastado con procesamiento similares centrados en noticias o Tweets, es alto y obtiene resultados acordes con el alcance esperado en este proyecto.

- **NLTK y Gensim como algoritmos paralelos de preprocesamiento de Abstracts** Ambos algoritmos, cuando son combinados para preprocesar textos de menos de 150 caracteres, ofrecen un rendimiento óptimo y robusto. Tanto NLTK como algoritmo para tokenizar y eliminar Stopwords, como Gensim para crear una bolsa de términos y analizar frecuencias TF-IDF resultan extremadamente útiles en el procesamiento de patentes.

## 6.2. Trabajo Futuro

Aunque se considera que el objetivo principal de la tesis se ha logrado, hay muchas mejoras que podrían hacerse para lograr un mejor rendimiento. En las siguientes secciones presentamos, para cada uno de los aspectos de la investigación llevado a cabo en esta tesis, algunos de los temas abiertos que merecen más investigación.

Patent Radar, como toda herramienta *software*, viene determinada también por una serie de limitaciones asociadas al tiempo de desarrollo. Tanto la analítica de patentes, como Patent Radar, pueden potenciar su rendimiento basándose en las siguientes líneas de investigación.

- **Análisis de diagramas o croquis mediante Redes Neuronales Convolucionales** Con los recientes desarrollos en el campo del procesamiento de imágenes<sup>1</sup>, podría implementarse en Patent Radar una red neuronal convolucional en una unidad de procesamiento de gráficos. La red se entrenaría

---

<sup>1</sup>[ALLA17] 2017, Alla G. Kravets, Patents Images Retrieval and Convolutional Neural Network Training Dataset Quality Improvement

mediante Backpropagation, con núcleos convolucionales y matrices de sesgo modificadas en cada paso hacia atrás de acuerdo con un gradiente de una función de error, para clasificar croquis o planos contenidos en las solicitudes de patentes.

- **Escalabilidad de la arquitectura mediante hosting** Si bien el equipo de *hardware* no es el más adecuado para diseñar una herramienta a gran escala, uno de los siguientes pasos sería garantizar la robustez a la hora de gestionar los 100 millones de registros contenidos en Esp@cenet. Una posible solución, sería utilizar *hosting* de terceros para alojar un dataset intermedio de procesamiento. Cualquier modelo de base de datos actual (SQL Server, Postgres..) es válido para gestionar un volumen estimado de 2TB anuales de texto plano, aunque este paso debería ser diseñado por un arquitecto de sistemas informáticos.
- **Inclusión de información financiera** El uso de indicadores de patentes aún está poco desarrollado en el contexto del análisis financiero. El enfoque más publicitado es actualmente Ocean Tomo's, una compañía estadounidense que introdujo en 2017 su fondo Ocean Tomo 300 de las 300 compañías más innovadoras, utilizando sus métricas basadas en patentes. Ocean Tomo muestra en su sitio web que su índice de inversión supera año a año al SP 500. Además, es común toparse con cierto nivel de estanqueidad entre diferentes departamentos de empresas o unidades funcionales de gobiernos. La información analizada por Patent Radar es información puramente técnica cuya utilidad de explotación podría verse potenciada al complementarse con información financiera de dichas empresas o gobiernos. De esta forma, cuantificar la variación de PIB destinado a ciertas áreas de innovación, o modificar el presupuesto asignado a un proyecto podría ser el siguiente nivel de mejora de cualquier herramienta dedicada a la Vigilancia Tecnológica.
- **Conexión en paralelo de otras oficinas** Mientras que Esp@cenet, como base de datos de la EPO, cuenta con un número de patentes y una fiabilidad superior a cualquiera de las oficinas de otras regiones, la imposibilidad de conexión de Patent Radar a oficinas del continente asiático y americano limitan

su usabilidad. Especialmente en la actualidad, cuando parece que la globalización se acelera a pasos agigantados y Asia se convierte en locomotora, se hace necesaria una adaptación de la herramienta, con el consiguiente aumento de esfuerzo dedicado a la gestión masiva de datos, para poder capturar las tendencias de forma global.

- **Expansión a otros idiomas** En línea con la conexión de Patent Radar a oficinas de patentes de otras regiones, se hace necesario implementar un módulo de traducción de idiomas diferentes al inglés. Actualmente, gran parte de los esfuerzos dedicados a combatir el COVID-19 están concentrados en China, por lo que este módulo podría ser un gran paso en el valor añadido proporcionado por Patent Radar.

Tras una revisión sobre las posibles futuras líneas que deben investigarse, merece la pena reseñar que la propiedad intelectual parece seguir siendo un campo complejo reservado a especialistas que tienen una doble formación técnica y jurídica. Sin embargo, la oferta de tecnológica continua, como la que ofrece este proyecto, sigue creciendo en los campos de licencias, transferencias de tecnología y evaluación de patentes, en línea con las necesidades del mercado.

Sólo los esfuerzos de los analistas de datos pueden aportar al sector de la propiedad industrial, a las grandes corporaciones y a los gobiernos una puerta para potenciar la utilidad de las solicitudes de patentes.

Los datos están disponibles para quien quiera explotarlos, sólo es necesario trabajar alineados para asegurar que estos son utilizados con seguridad, coherencia y destreza y permiten que el sector de la innovación coja la mayor tracción posible en beneficio de la sociedad.

# Bibliografía

- [RAE20] 2020, Real Academia Española y Asociación de Academias de la Lengua Española. *Diccionario de la lengua española 23a edición*, Madrid: Espasa
- [WIPO3] 2003, World Intellectual Property Organization, *Training Course on Practical Intellectual Property Issues in Business*
- [MIIT12] 2012, Ministerio de Industria, Energía y Turismo Gobierno de España, *La Patente Europea*.
- [OEP20] 2020, Oficina Española de Patentes y Marcas. *Patente Nacional, Modelo de Utilidad, Certificados Complementarios de Protección y Transferencias Contractuales y Licencias*
- [OMPI18] 2018, OMPI, *Indicadores Mundiales de propiedad intelectual*
- [BRE02] 2002, Breitzman, Anthony F.; Moguee, Mary Ellen, *The many applications of Patent Analysis*
- [TADU11] 2011, Taduri, S., Lau, G. T., Law, K. H., Yu, H., *Developing an ontology for the USpatent system*
- [ALON17] 2017, A. Alonso, *Sistema de Arrastre Innovador para deportes de deslizamiento*
- [ARIS18] 2018, Aristodemou L., *The state-of-the-art on Intellectual Property Analytics IPA: A literature review on artificial intelligence, machine*

- learning and deep learning methods for analysing intellectual property IP) data*
- [TRIP15] 2015, World Intellectual Property Organization, *WIPO Guide to Using Patent Information*
- [OMPI16] 2016, World Intellectual Property Organization, *Patents, Data and Future Perspectives in Patent Analytics*
- [TIE15] 2015, Tietze C, *The future of patent analytics*
- [OCDE04] 2004, OCDE, *Patents and Innovation: Trends and Policy Changes*
- [USCO84] 1984, United States Congress Senate, *Oversight of the Patent and Trademark Office*
- [MOER10] 2010, M.G. Moehrle, *Patinformatics as a business process*
- [RATU10] 2010, M.K. Ratur, *Patinformatics - an emerging scientific discipline*
- [BONI13] 2013, D.Bonino, *Review of the state-of-the-art in patent information*
- [BACE13] 2013, F. Cesaroni, *Capturing the real value of patent analysis*
- [ARIS17] 2017, Aristodemou, Leonidas; Tietze, Frank. *Exploring the future of patent analytics*
- [SJUN13] 2013, S.Jun, *Emerging Technology Forecasting*
- [SBAS10] 2010, S.Bas, *Discovery of factor influencing patent value*
- [JUN14] 2014 Jun, S., Park, S. S., Jang, D. S. *Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications*
- [TRAP06] 2006, A.Trappey, *Development of a patent document classification platform using a back-propagation network*
- [ABB14] 2014, Abbas, A., Zhang, L., Khan, S. U. *A literature review on the state-of-the-art in patent analysis.*



- [MALP17] 2017, Dinesh Malpure, Yogesh Botre, Darshan Bhansali, Rohan Bhagi, *PatentTrend Analysis and Future Prediction*
- [SUPR15] 2015, A.Supraja, S.Archana, S.Suvetha, “*Patent Search and Trend Analysis*”, *IEEE International Advance Computing Conference IACC*)
- [ZALA12] 2012, Zalányi,Kinga Makovi, Zoltán Somogyvári,Katherine ”*Prediction of EmergingTechnologies Based on Analysis of the U.S. Patent Citation Network*”
- [ARI18] 2018, Aristodemou, L., Tietze, F. *The state-of-the-art on Intellectual Property Analy-tics: A literature review on artificial intelligence, machine learning and deep learning methodsfor analysing intellectual property*
- [JUN14] 2014, Jun, S., Park, S. S., Jang, D. S. *Document clustering method using dimensionreduction and support vector clustering to overcome sparseness. Expert Systems with Applications*
- [SUO17] 2017, A. Suominen, H. Toivanen, M. Seppänen, *Firms’ knowledge profiles: mappingpatent data with unsupervised learning*
- [THO10] 2010, Thorleuchter, D., Van den Poel, D., Prinzie, A. *A compared RD-based andpatent-based cross impact analysis for identifying relationships between technologies*
- [HYS17] 2017, H.-Y. Sung, H.-Y. Yeh, J.-K. Lin, S.-H. Chen, *A visualization tool of patenttopic evolution using a growing cell structure neural network*
- [CLE16] 2016, C. Lee, J. Kim, O. Kwon, H.-G. Woo, *Stochastic technology life cycle analysisusing multiple patent indicators*
- [GOV18] 2018, U. Govindarajan, A. Trappey, C. Trappey, *Immersive technology for human-centric cyberphysical systems in complex manufactu-*

*ring processes: a comprehensive overview of the global patent profile using collective intelligence*

- [OCD19] 2019, OCDE, *Manual de Estadísticas de Patentes de la OCDE*
- [SULI18] 2018, Susan Li, *Topic Modelling in Python with NLTK and Gensim*
- [WIPO16] 2016, World International Patent Office, *Manual on Open Source Patent Analytics*
- [EPOC20] 2020, European Patent Office, *Fighting Coronavirus Manifesto*,
- [ALLA17] 2017, Alla G. Kravets, *Patents Images Retrieval and Convolutional Neural Network Training Dataset Quality Improvement*

## Apéndice A

# Objetivos de Desarrollo Sostenible de Naciones Unidas (ODS)

El objetivo 9 de desarrollo sostenible de la UN dice textualmente: *“Desarrollar infraestructuras resilientes, promover la industrialización inclusiva y sostenible, y fomentar la innovación”*

La industrialización inclusiva y sostenible, junto con la innovación e infraestructura, pueden poner en marcha diferentes fuerzas económicas que generen empleo e ingresos. Estos 3 ejes: industrialización, innovación e infraestructura juegan un papel clave en la introducción y promoción de nuevas tecnologías, facilitando el comercio internacional y utilización de uso eficiente de recursos. Sin embargo, al mundo le queda un largo camino para cumplir este objetivo. Los países menos desarrollados, en particular, necesitan acelerar el desarrollo de estos tres pilares de crecimiento.

Aunque parezca un contrasentido, la gestión en la innovación y en las patentes necesita precisamente de innovación. Para acelerar el desarrollo de los adelantos tecnológicos que surjan en cualquier parte del mundo y en cualquier contexto, se necesita avanzar más en dos materias de obligado cumplimiento por todas las empresas y por ende de los organismos oficiales de regulación. Éstas son: La gestión eficiente de los datos maestros: Master Data Management MDM y la Analítica Avanzada.

El presente TFM contribuye a este punto con la introducción y adopción de tecnologías de predictibilidad analítica, lo que acortará los tiempos de desarrollo de las patentes y de su implantación. Conociendo y trabajando sobre variables de corte – “features”- datos para modelización, KPIs, etc. se puede predecir y/o prescribir tanto el éxito de una patente, como su capacidad de retornar la inversión, su grado de aplicabilidad, etc. Todo ello gracias a la “clusterización” de los datos históricos –“training data”- gestionados a su vez por tecnologías de “Big Data”.

Con una información fiable y unos datos bien gestionados, se podrá saber a priori el grado de éxito que una patente en el sector logístico, por ejemplo, puede tener en un país en desarrollo que necesita de una determinada tecnología aplicable en su ecosistema de distribución de productos y servicio, lo que evita múltiples pruebas de ensayo y error.

El objetivo 17 de desarrollo sostenible de la Un dice textualmente: *“Fortalecer los medios de ejecución y reavivar la alianza mundial para el desarrollo sostenible”*

El desarrollo sostenible sólo puede ser viable con alianzas y cooperación. El desarrollo con éxito de la agenda para el desarrollo sostenible, requiere de alianzas “inclusivas” a todos los niveles: globales, nacionales, regionales y locales, construidas sobre principios y valores y sobre una visión y objetivos compartidos colocando a la gente y al planeta en el centro del ecosistema. Se necesita una cooperación internacional fuerte para asegurar que los países se recuperen de la pandemia de la COVID19. Es ciertamente curioso cómo se aplican se aplican técnicas de ML como “Content Based Recommendations” y filtros colaborativos para predecir los gustos de un consumidor sobre una película o producto determinado, pero no se conozcan herramientas de aprendizaje de máquina para asociar iniciativas de innovación o patentes con su aplicación en determinados países, grupos sociales u otras empresas que puedan aportar su valor en una alianza multilateral que acelere la aplicabilidad del producto o servicio en cuestión.

A través de algoritmos “K-Means” pertenecientes al segmento de algoritmos “no supervisados” de Machine Learning, cuyo propósito básico es distribuir en “clusters” homogéneos las iniciativas de innovación y patentes, se pretende paliar este contrasentido. Cada uno de estos “clusters” contiene miles de ejemplos que dependen a su vez de cientos de variables. Una vez distribuidos estos ejemplos, se determinan unos ejemplos “tipo” por cada “cluster” denominados “centroides”.

Si una empresa o país u otro tipo de organismo dispone de la información adecuada y de la herramienta presentada en este TFM puede plantearse alianzas con aquellas iniciativas que más se adapten a sus necesidades con el consiguiente ahorro en tiempo y recursos y desechar aquéllas que obtenga una baja puntuación en el resultado probabilístico del algoritmo.

Sin duda, esta herramienta contribuirá a la consolidación de alianzas, que a su vez aceleren todas las iniciativas de la agenda de sostenibilidad.

## Apéndice B

### Código Relevante

#### B.1. Obtención de dataset de muestra

```
let EPO_REQUEST = (http_query_api as text) =>
  let
    Source = Json.Document(Web.Contents(http_query_api)),
    result = Source[result],
    items = result[items],
    items1 = items{0},
    abstract = try items1[abstract] otherwise "Empty_Abstract"
  in
    abstract
  in
    EPO_REQUEST
```

#### B.2. Algoritmo K-Means para procesamiento de Abstracts

```
import pandas as pd
from pandas import ExcelWriter
```

```
from pandas import ExcelFile
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.util import ngrams
from nltk.stem import WordNetLemmatizer
from gensim.parsing.preprocessing import remove_stopwords
from gensim.parsing.preprocessing import STOPWORDS
import matplotlib.pyplot as plt
from sklearn.cluster import MiniBatchKMeans
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE

#Letter -> IPC Category, Key -> One, Bin, Tri Gram
def Corpora_Function(letter, key, df):
    mylist = ""
    Corpora = []
    for ind in df.index:
        #Checks that the letter corresponds to any of the IPC
        Categories
        if df['IPC'][ind] == letter:
            words = df['Short_Abstract'][ind]
            #Function Call
            Document = PreProcessingData(words, key)
            Corpora.append(Document)
        #Special Case for all categories
```

```
        if letter == "":
            words = df['Short_Abstract'][ind]
            #Function Call
            Document = PreProcessingData(words, key)
            Corpora.append(Document)
    return Corpora #3 distintos para una misma letra = 24; Vector
    de abstracts pre-procesados.

def PreProcessingData(document, key):
    lemmatizer = WordNetLemmatizer()
    document = document.split()
    if key == "Tri":
        #adding new words to stop words
        NewList='present,relates,present,provides,according,
        disclosure,position,second,end,provided,method,
        includes,main,body,discloses,following,having,
        relate, consisting, washing'
        NewList = NewList.split(',')
        my_stop_words = STOPWORDS.union(set(NewList)) #
        #Stop Words Function
        mylist = [word for word in document if not word in
        my_stop_words]
        #Lemmatize Process
        mylist = [lemmatizer.lemmatize(word) for word in
        mylist]
        #Creates 3-N Gram Tokens
        mylist = ngrams(mylist, 3)
    elif key == "Bin":
        #adding new words to stop words
        NewList='present,relates,present,surface,portion,
        comprises,layer,edge,configured,provides,according,
```



```

        disclosure,position,second,end,provided,method,
        includes,main,body,discloses,following,having,
        relate, consisting, washing'
NewList = NewList.split(',')
my_stop_words = STOPWORDS.union(set(NewList))
#Stop Words Function
mylist = [word for word in document if not word in
        my_stop_words]
#Lemmatize Process
mylist = [lemmatizer.lemmatize(word) for word in
        mylist]
#Creates 2-N Gram Tokens
mylist = ngrams(mylist, 2)
else:
    #adding new words to stop words
    NewList='state, position,machine,surface,member,
        plurality,user,apparatus,inner,present,direction,
        present,outer,invention,element,body,device,second,
        end,provided,configured,thereof,based,use,mounted,
        main,opening,arranged,having,comprising,comprises,
        coupled,disposed,connected,formed,disclosed,
        including,said,portion,method,includes,containing,
        obtained,relates,dislcosed,producing,cotaing,
        provide,provides,washing,drying,rotating,heating,
        according,housing,forming,locking,connecting,fixed,
        adapted,cooling,bearing,presents,determined,
        received,processing,corresponding,associated,
        generated,unit,input,receiving'
    NewList = NewList.split(',')
    my_stop_words = STOPWORDS.union(set(NewList))
    #Stop Words Function

```

```
        mylist = [word for word in document if not word in
                  my_stop_words]
        #Lemmatize Process
        mylist = [lemmatizer.lemmatize(word) for word in
                  mylist]
        mylist = (" ").join(mylist)
    return mylist

def find_optimal_clusters(data, max_k):
    iters = range(2, max_k+1, 2)

    sse = []
    for k in iters:
        sse.append(KMeans(n_clusters=k).fit(data).inertia_)
        print('Fit {} clusters'.format(k))

    f, ax = plt.subplots(1, 1)
    ax.plot(iters, sse, marker='o')
    ax.set_xlabel('Cluster Centers')
    ax.set_xticks(iters)
    ax.set_xticklabels(iters)
    ax.set_ylabel('SSE')
    ax.set_title('SSE by Cluster Center Plot')

def plot_tsne_pca(data, labels):
    max_label = max(labels)
    max_items = np.random.choice(range(data.shape[0]), size=100,
                                  replace=False)
```

```

pca = PCA(n_components=2).fit_transform(data[max_items,:].
    todense())
tsne = TSNE().fit_transform(PCA(n_components=50).
    fit_transform(data[max_items,:].todense()))

idx = np.random.choice(range(pca.shape[0]), size=60, replace=
    False)
label_subset = labels[max_items]
label_subset = [cm.hsv(i/max_label) for i in label_subset[idx
    ]]

f, ax = plt.subplots(1, 2, figsize=(14, 6))

ax[0].scatter(pca[idx, 0], pca[idx, 1], c=label_subset)
ax[0].set_title('PCA Cluster Plot')

ax[1].scatter(tsne[idx, 0], tsne[idx, 1], c=label_subset)
ax[1].set_title('TSNE Cluster Plot')

def get_top_keywords(data, clusters, labels, n_terms):
    df = pd.DataFrame(data.todense()).groupby(clusters).mean()

    for i,r in df.iterrows():
        print('\nCluster {}'.format(i))
        print(', '.join([labels[t] for t in np.argsort(r)[-n_terms:]])
            )

df = pd.read_excel('export_covid.xlsx')
corpus = Corpora_Function("", "One", df)

```

```
tfidf = TfidfVectorizer( #cogeme terminos con unos threshold de tfidf
    que hagan que sean utiles en ingles).
    min_df = 5,
    max_df = 0.95,
    max_features = 8000,
    stop_words = 'english'
)
tfidf.fit(corpus)
text = tfidf.transform(corpus)
print(text)

find_optimal_clusters(text, 20)

clusters = MiniBatchKMeans(n_clusters=15, init_size=1024, batch_size
    =2048, random_state=20).fit_predict(text)

plot_tsne_pca(text, clusters)

get_top_keywords(text, clusters, tfidf.get_feature_names(), 3)
plt.show()
```

# Apéndice C

## Características de sesgo TF-IDF

```
tfidf = TfidfVectorizer( .
    min_df = 5,
    max_df = 0.95,
    max_features = 8000,
    stop_words = 'english'
)
tfidf.fit(corpus)
text = tfidf.transform(corpus)
print(text)
```

### C.1. Gestión de dataset de muestra

```
let
    Source = Csv.Document(File.Contents("C:\Users\XXX\Desktop\TFM\
        DataTFM\
    DataSet.csv"),
    [Delimiter=";", Columns=11, Encoding=1252, QuoteStyle=QuoteStyle.
        Csv]),
    #"Promoted Headers" = Table.PromoteHeaders(Source, [
        PromoteAllScalars=
```

```

true]),
#"Duplicated Column" = Table.DuplicateColumn("#Promoted Headers",
"Publication number",
"Publication number - Copy"),
#"Split Column by Character Transition" = Table.SplitColumn
("#Duplicated Column", "Publication number - Copy",
Splitter.SplitTextByCharacterTransition({"0".."9"}, (c) =>
not List.Contains({"0".."9"}, c)),
{"Publication number - Copy.1", "Publication number - Copy.2"}),
#"Split Column by Character Transition1" = Table.SplitColumn("#
Split
Column by Character Transition",
"Publication number - Copy.1", Splitter.
SplitTextByCharacterTransition
((c) => not List.Contains({"0".."9"}
, c), {"0".."9"}), {"Publication number - Copy.1.1", "Publication
number
r - Copy.1.2"}),
#"Added Custom" = Table.AddColumn("#Split Column by Character
Transition1"
, "Custom", each "https://data.epo.org/linked-data/data/
publication/" &
["Publication number - Copy.1.1"]& "/" &
["Publication number - Copy.1.2"]& "/" &["Publication number -
Copy.2"]&
.json"),
#"Removed Columns" = Table.RemoveColumns("#Added Custom",{
Publication
number - Copy.1.1",
"Publication number - Copy.1.2", "Publication number - Copy.2"}),

```

```

#"Renamed Columns" = Table.RenameColumns("#Removed Columns",{{"
    Custom",
    "HTTP API Request"}}),
#"Invoked Custom Function" = Table.AddColumn("#Renamed Columns",
    "Abstract", each A9([HTTP API Request])),
#"Filtered Rows" = Table.SelectRows("#Invoked Custom Function",
    each true),

#"Replaced Value" = Table.ReplaceValue("#Filtered Rows", "#(cr)#(
    lf)", "",
    ,Replacer.ReplaceText,{"Inventors"}),
#"Replaced Value1" = Table.ReplaceValue("#Replaced Value", "#(cr)
    #(lf)", "",
    ,Replacer.ReplaceText,{"Applicants"}),
#"Replaced Value2" = Table.ReplaceValue("#Replaced Value1", "#(cr)
    #(lf)", "",
    ,Replacer.ReplaceText,{"IPC"}),
#"Replaced Value3" = Table.ReplaceValue("#Replaced Value2", "#(cr)
    #(lf)", "",
    ,Replacer.ReplaceText,{"CPC"}),
#"Replaced Value4" = Table.ReplaceValue("#Replaced Value3", "#(lf)
    ", " ",
    Replacer.ReplaceText,{"Abstract"}),
#"Added Custom1" = Table.AddColumn("#Replaced Value4", "#Abstract
    ", each
    List.Count(Text.Split([Abstract], " "))),
#"Run Python script" = Python.Execute("# 'dataset' holds the
    input data
    for this script#(lf)# 'dataset' holds the input data for this
    script#(lf)from nltk.tokenize import word_tokenize#(lf)from
    nltk.stem.porter import PorterStemmer#(lf)from nltk.stem

```

```

import
WordNetLemmatizer#(lf)from nltk.probability import FreqDist#(lf)
from gensim.parsing.preprocessing import remove_stopwords#(lf)
from gensim.parsing.preprocessing import STOPWORDS#(lf)
from gensim import corpora#(lf)import nltk#(lf)import string#(lf)
import re#(lf)import matplotlib as ml#(lf)import matplotlib.
    pyplot as plt#(lf)
#(lf)mylist = """"#(lf)table = str.maketrans(' ', ' ', string.
    punctuation)#(lf)
porter = PorterStemmer()#(lf)lemmatizer = WordNetLemmatizer()#(lf
    )
ArrayOfAbstracts = dataset.loc[:, "Abstract"].values#(lf
)ArrayOfAbstracts = [str(i) for i in ArrayOfAbstracts]#(lf)
    TiltArray = []#(lf)
for i in range(len(ArrayOfAbstracts)):#(lf) words =
    ArrayOfAbstracts[i] #(lf)
# convert to lower case#(lf) words = words.lower()#(lf) # remove
    punctuation#(lf) stripped = [w.translate(table) for w in words
    ]#(lf) stripped = ("").join(stripped)#(lf)
#keep alphabet characters#(lf)
stripped = re.sub('[^a-z ]', '', stripped)#(lf)
#remove single letters#(lf) stripped = re.sub('(:^| ) [a-z] (=? |$)
    ', ' ', stripped)#(lf)
#stop words#(lf) filtered_sentence = remove_stopwords(stripped)#(
    lf)
#Save to DataFrame#(lf) TiltArray.append(filtered_sentence)#(lf)
#stemming process #(lf) tokens = word_tokenize(filtered_sentence)
    #(lf)
stemmed = [lemmatizer.lemmatize(word) for word in tokens]#(lf)
stemmedtemp = (" ").join(stemmed)#(lf)#(lf)#Add New Abstracts
    to DataFrame#(lf)dataset["ShortAbstract"] = TiltArray#(lf)",

```



```

[dataset=#"Added Custom1"])),
dataset = #"Run Python script"{[Name="dataset"]}[Value],
#"Changed Type" = Table.TransformColumnTypes(dataset,{{"No",
    Int64.Type}, {"Title", type text}, {"Inventors", type text},
    {"Applicants", type text}, {"Publication number", type text},
    {"Earliest priority", type date}, {"IPC", type text}, {"CPC",
    type text}, {"Publication date", type date}, {"Earliest
    publication", type date}, {"Family number", Int64.Type}, {"
    HTTP API Request", type text}, {"Abstract", type text}, {"#
    Abstract", Int64.Type}, {"ShortAbstract", type text}}),
#"Added Custom4" = Table.AddColumn(#"Changed Type", "#
    ShortAbstract", each List.Count(Text.Split([ShortAbstract], "
    "))),
#"Split Column by Delimiter" = Table.SplitColumn(#"Added Custom4
    ", "IPC", Splitter.SplitTextByDelimiter(",", QuoteStyle.Csv),
    {"IPC.1", "IPC.2"}),
#"Changed Type1" = Table.TransformColumnTypes(#"Split Column by
    Delimiter",{{"IPC.1", type text}, {"IPC.2", type text}}),
#"Added Custom2" = Table.AddColumn(#"Changed Type1", "IPC.1
    Letter", each Text.Start([IPC.1],1)),
#"Added Custom3" = Table.AddColumn(#"Added Custom2", "IPC.2
    Letter", each Text.Start([IPC.2],1)),
#"Changed Type2" = Table.TransformColumnTypes(#"Added Custom3
    ",{{"#ShortAbstract", Int64.Type}})
in
#"Changed Type2"

```

## Apéndice D

# Manifiesto sobre la lucha contra el COVID-19 (EPO



30 April 2020

### **Responding to COVID-19**

#### **A joint message of support for inventors from the USPTO and EPO**

The coronavirus outbreak has changed our daily lives almost beyond recognition. The staff of the European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) stand in solidarity with everyone affected.

Beyond the instances of personal hardship, we are now seeing a significant negative impact on the global economy. Every day, we learn further of the economic effects in nations in various degrees of lockdown.

In these challenging times, it is our innovators, inventors and all those involved in pioneering research who will play a central role in the recovery of our economies and societies. In both the US and Europe, industries that make intensive use of intellectual property rights (IPR) generate approximately 40% of GDP and are directly and indirectly responsible for around 30% of jobs. They are the driving force behind exports, amounting to more than one trillion USD or EUR each year.

Among these IPR-intensive industries are the innovative companies, research centers and universities that are researching and developing tests, vaccines, and treatments that could put an end to the coronavirus pandemic. Their work is fundamental to us as people and as a society fighting to take care of its vulnerable members and those in need of medical care. Indeed, such innovation has long served as the driving engine of human development, and will continue to do so.

To support innovation during this crisis, the USPTO and the EPO stand shoulder-to-shoulder with the innovation community. Our Offices are now offering assistance through time extensions and fee deadlines, as well as flexibility on hearings, such as offering video conferences or postponements.<sup>1</sup> By doing so, we hope to continue

supporting inventors with high-quality intellectual property rights that help them attract investment and license technology, create jobs, and enter new markets with confidence and predictability.

Our Offices will spare no effort to give users the support they need. We will develop our capacity to respond to the difficult circumstances that applicants face. At a time when the dissemination of knowledge is crucial, we will persevere in developing the tools that can help scientists all over the world. We will continue to enrich our public patent databases, which are free to access and contain hundreds of millions of documents from all over the world. They present a wealth of technological knowledge that can help inventors and researchers everywhere build on previous developments, gain new insight and help identify potential suppliers, technology partners, and customers.

The USPTO and the EPO stand united in our effort to support the public in this crucial time, and we will build on our longstanding relationship to provide everyone in the IP community with the support they need.

Andrei Iancu  
USPTO Director

Antonio Campinos  
EPO President

## Apéndice E

# Características principales de Power BI

Microsoft Power BI es una colección de herramientas de explotación de datos, que cuenta con servicios de software, aplicaciones y conectores de datos.<sup>1</sup> Es una plataforma basada en la nube utilizada para consolidar datos de diversas fuentes en un solo conjunto de datos. Estos conjuntos de datos se utilizan para la visualización, evaluación y análisis de datos al hacer informes, paneles y aplicaciones compartibles. Microsoft ofrece tres tipos de plataformas Power BI: Power BI Desktop (una aplicación de escritorio), Power BI Service (SaaS, es decir, Software as a Service) y Power BI Mobile (para dispositivos iOS y Android).

Power BI se puede implementar tanto en equipos locales como en la nube. También puede importar datos de bases de datos locales / fuentes de datos, fuentes de datos basadas en la nube, SQL, archivos simples de Excel y otras fuentes híbridas. Por lo tanto, Power BI, un líder entre muchas otras herramientas de BI, demuestra ser una herramienta eficiente y fácil de usar para el análisis de datos. Además, permite a los usuarios consolidar datos de múltiples fuentes, crear paneles interactivos, evaluar datos, crear informes y compartirlos con otros usuarios en tiempo real.

---

<sup>1</sup>2020, Microsoft

A continuación se enumeran las características principales que hacen de dicha herramienta la mejor alternativa:

1. Amplio Rango de herramientas visuales: Power BI ofrece más de 100 formas de visualización o *Visuals*. Esto hace que el muestreo y explotación de datos sean sencillos al ojo humano, y se siga un proceso iterativo haciendo partícipes a los usuarios para ir mejorando las formas de representación.
2. DAX como lenguaje para formulas: Las funciones DAX son las expresiones de análisis de datos que se encuentran en Power BI. Estas funciones de análisis son códigos predefinidos para realizar tareas analíticas específicas en los datos.
3. Flexibilidad en la elección de fuentes de datos: Power BI puede gestionar de forma nativa 13 tipos de fuente de datos: Excel, Power Query, Power Query Dataflows, Servidores SQL, BBDD MySQL, Azure, Texto/CSV, BBDD Oracle, PDF, Access, XML, y JSON.
4. Potencia de filtrado en Datasets: El usuario Puede filtrar los conjuntos de datos y tener subconjuntos más pequeños que contengan solo los datos importantes y su relevancia contextual. Power BI proporciona a los usuarios una amplia gama de conectores de datos integrados, como Excel. Como se ha comentado anteriormente, los usuarios pueden conectarse fácilmente a dichas fuentes de datos y crear conjuntos de datos importando datos de una o más fuentes combinadas.
5. Escalabilidad del cuadro de mando: los cuadros de mando típicos en Power BI se componen de múltiples visualizaciones en forma de mosaico. Las páginas de dichos cuadros/informes se pueden compartir de forma automática y periódica o, incluso, si alguno de los valores gestionados por Power BI alcanza un valor frontera, generando un informe automático a modo de alarma.
6. Ayuda mediante preguntas y respuestas en lenguaje natural: El cuadro de preguntas y respuestas mediante lenguaje natural es una característica única

de Power BI. Con el cuadro de preguntas y respuestas, un usuario puede hacer preguntas en lenguaje natural para buscar datos e información disponibles en el sistema Power BI. Los motores cognitivos de Power BI buscan los datos o una parte del informe que se ha buscado y se lo devuelven al usuario de forma gráfica.