Implementación de un Modelo Computacional de recomendación, utilizando Técnicas de Inteligencia Artificial para generar estrategias de investigación dentro de Grupos de Investigación

Alejandro Esteban Rendón Diosa, Deiner David Martínez Gutiérrez

Universidad
Tecnológica
de Pereira

Pereira, Risaralda

Implementación de un Modelo Computacional de recomendación, utilizando Técnicas de Inteligencia Artificial para generar estrategias de investigación dentro de Grupos de Investigación

Alejandro Esteban Rendón Diosa, Deiner David Martínez Gutiérrez

Maestría en Ingeniería de Sistemas y Computación
de la Universidad Tecnológica de
Pereira

Presentado por
Alejandro Esteban Rendón Diosa, Deiner David Martínez
Gutiérrez
en Pereira, Risaralda

Colombia. Noviembre, 2020

Director: Ramiro Andrés Barrios Valencia
Fecha de la Sustentación: Noviembre. 2020

Índice general

Índice de figuras

Índice de cuadros

Código Fuente y Listados

Capítulo 1

Introducción

Con el rápido crecimiento del hombre en la búsqueda de nuevo conocimiento, se ha presentado un incremento de publicaciones de artículos científicos durante los últimos años, por ejemplo.

"El incremento de la cantidad de artículos entre 1934 y 2009 tuvo una tasa de variación de 1,23 veces, lo que corresponde a una tasa media de crecimiento del 1 % anual. En el número de autores, la tasa de variación fue de 6,75 veces, con una tasa media de crecimiento del 2,7 % anual. La tasa de variación de la cantidad de autores por artículo mostró un incremento de 2,48 veces, lo que equivale a una tasa media de crecimiento del 1,6 % anual, mientras que el número promedio de mujeres por artículo mostró una tasa de crecimiento entre 1958 y 2009 de 18 veces, lo que corresponde a una tasa media de crecimiento anual del 5,8 %. [8]

Esto desemboca en una evolución del ser humano por participar en la construcción de la sociedad más autónoma y consistente en cuanto a bases estructurales de conocimiento se refiere; y si a eso se suma el crecimiento de estos datos en memorias de almacenamiento, podemos ver una paralelismo de crecimiento en el incremento del volumen mundial de datos [9].

Este crecimiento ha generado que se empleen técnicas de minería de datos para el descubrimiento de nuevo conocimiento, realizando búsquedas y descubrimientos de patrones que puedan predecir un comportamiento de los mismos. A pesar de que la minería de datos abarca un amplio rango de aplicaciones, muchas de estas técnicas también son utilizadas en el Aprendizaje Automático, los cuales tratan de extraer información o conocimiento de un conjunto de datos de ejemplo y a su vez generalizando estas similitudes con los otros ejemplos similares.

Capítulo 2

Problema de Investigación

La investigación es considerada uno de los motores de la sociedad ya que aporta avances en diferentes áreas y propone una eventual solución a la problemática planteada. Esta disciplina se materializa en diferentes campos y de los cuales devienen los denominados grupos de investigación.

Colombia no es ajena al incremento de estos procesos investigativos, basta con analizar las cifras proporcionadas por COLCIENCIAS, entidad que regula el funcionamiento de este tipo de agrupaciones, y por consiguiente quien establece, controla y guía los parámetros de investigación en el país, que en informe presentado para el año 2019 [10], se encontraron 5772 grupos de investigación reconocidos y el número sigue en ascenso hasta la fecha.

La presente tesis tendrá como referencia de estudio el grupo de investigación BIOTECNOLOGÍA - PRODUCTOS NATURALES de la Universidad Tecnológica de Pereira (que a la fecha presenta un total de 114 grupos investigativos avalados por COLCIENCIAS para el año 2020 [11]).

Contexto: El grupo de investigación BIOTECNOLOGÍA - PRODUCTOS NATURALES es uno de los más activos de la Escuela de Química de la Universidad Tecnológica de Pereira, lo cual significa que tienen una gran cantidad de producción al año. El director del grupo (Oscar Marino Mosquera) busca diferentes metodologías para establecer una ruta de acción referente a las nuevas actividades de investigación que se llevarán a cabo.

Después de un par de reuniones que se realizaron, se encontró que la herramienta más útil que tenia era una hoja de cálculo en Excel y con esto trataba de analizar los diferentes temas y la cantidad de productos que tenia relacionados. Aunque la herramienta utilizada funciona para algunos casos, se considera que hace falta herramientas especializadas en el tema, que sirvan como un apoyo a los grupos de investigación y que faciliten su trabajo.

Partiendo de la problemática presentada por estos grupos y la carencia de herramientas tecnológicas que faciliten trazar rutas de acción en los procesos investigativos que lleven

a la generación de nuevos conocimientos, es viable entonces proponer una herramienta tecnológica que permita agilizar a los investigadores el análisis documental de sus productos académicos que servirá de premisa para comprobar una hipótesis planteada.

Así las cosas, es pertinente formular un sistema de recomendación compuesto de dos fases de análisis (Primera, a nivel semántico y la segunda, a nivel de clasificación utilizando técnicas de Inteligencia Artificial) basándose en las diferentes producciones textuales recopilados por el grupo a lo largo de su carrera. Esto les proporcionará un apoyo a la hora de decidir cuáles pueden ser las mejores alternativas que ayuden a plantear estrategias de investigación.

Capítulo 3

Justificación

En los procesos investigativos se genera una gran cantidad de archivos que contienen información valiosa generada a través de revisión bibliográfica, análisis de datos, trabajo de campo, entre otros que permita al grupo investigador generar nuevo conocimiento, plantear nuevas tesis, defender o discutir una situación concreta.

Dicho proceso genera un gran esfuerzo físico y mental que devengan además la inversión de tiempo que en muchas ocasiones genera retrasos en la culminación de dichas investigaciones.

Uno de los puntos importantes para la clasificación, es saber cuáles son las características o atributos adecuados para poder distinguir los enfoques que tiene cada grupo de investigación con base a sus artículos publicados. El caso de estudio consiste en aplicar técnicas estadísticas y de inteligencia artificial sobre un conjunto de documentos digitales y/o producción textual de un grupo de investigación de la Escuela de Química de la UTP (ya que actualmente Febrero 2020 no cuentan con una herramienta que haga este tipo de análisis) que les permita visualizar características comunes entre los mismos, esto con el fin de que puedan solventar este tipo de situaciones que se traduce en mayor productividad y celeridad en este tipo de procesos que contribuyen de manera directa en la creación de un plan estratégico en la producción de nuevos artículos y productos.

Capítulo 4

Objetivos

4.1. Objetivo general

Formular un modelo computacional para recomendación de estrategias de investigación, aplicando técnicas de Inteligencia Artificial.

4.2. Objetivos específicos

- Obtener un conjunto de datos.
- Determinar técnicas de Inteligencia Artificial apropiadas para el conjunto de datos seleccionado.
- Implementar las anteriores técnicas.
- Evaluar los resultados obtenidos por el modelo formulado.

Capítulo 5

Marco Referencial

5.1. Marco Conceptual

A continuación se exponen algunos conceptos claves (teóricos y técnicos) para la comprensión del desarrollo del proyecto.

5.1.1. Modelo Computacional

Según el contexto del proyecto, se puede definir *Modelo Computacional* a un conjunto de pasos secuenciales que procesan y transforman cierto tipo de entradas. Dichos pasos pueden utilizar algoritmos o técnicas de diferentes tipos para completar un requerimiento.

5.1.2. Artículo Científico

Un artículo científico es un informe escrito que describe los resultados originales de una investigación ya realizada.

La característica principal de un artículo de investigación es que siempre debe producir avances en el conocimiento, por lo que resulta obvio que sólo puede cumplir su cometido cuando ha sido publicado y puesto a disposición de la comunidad científica para que pueda ser leído, entendido e incorporado por sus pares [12].

Estructura

Una estructura base para una publicación está compuesta por [13]:

- Título: factor determinante para que alguien se acerque al trabajo.
- Autor/autores: persona o personas en quienes recae la responsabilidad intelectual y el mérito.

- Resumen (Abstract): determina que los usuarios decidan si vale la pena leer lo que se ha hecho.
- Introducción: explica cual es el problema, el propósito de la investigación y su justificación.
- Materiales y métodos: señala la forma como se estudió el problema.
- Resultados: fruto de la investigación.
- Discusión: crítica o análisis de los resultados.
- Reconocimientos.
- Referencias.

5.1.3. Documentos Digitales

Un *Archivo Digital* es una representación virtual que puede contener diferentes tipos de información [14]. En el caso especifico en el que se hace referencia a *Documento Digital*, la representación contiene texto e imágenes que pueden representar documentos o archivos físicos.

5.1.4. El Procesamiento del Lenguaje Natural o NLP

El lenguaje es una herramienta para la transmisión de información, los seres humanos somos seres sociales por naturaleza y el lenguaje nos sirve para comunicarnos en cualquiera de sus formas (oral o escrito). Asimismo, el lenguaje humano es algo que está en constante cambio y evolución; y que puede llegar a ser muy ambiguo y variable. Por este motivo, entender y producir el lenguaje por medio de una computadora es un problema complejo de resolver. Ésta área de investigación, es el campo de estudio de lo que en inteligencia artificial se conoce como *Procesamiento del Lenguaje Natural* o *NLP* por sus siglas en inglés.[15]

Esta disciplina que se encuentra en la intersección de varias ciencias, tales como las Ciencias de la Computación, la Inteligencia Artificial y Psicología Cognitiva. Su idea central es la de darle a las máquinas la capacidad de leer y comprender los idiomas que hablamos los humanos. La investigación del Procesamiento del Lenguaje Natural tiene como objetivo responder a la pregunta de cómo las personas son capaces de comprender el significado de una oración oral/escrita y cómo las personas entienden lo que sucedió, cuándo y dónde sucedió; y las diferencias entre una suposición, una creencia o un hecho.

En general, en *Procesamiento del Lenguaje Natural* se utilizan seis niveles de comprensión con el objetivo de descubrir el significado del discurso. Estos niveles son:

- **Nivel fonético:** Aquí se presta atención a la fonética, la forma en que las palabras son pronunciadas. Este nivel es importante cuando procesamos la palabra hablada, no así cuando trabajamos con texto escrito.
- **Nivel morfológico:** Aquí nos interesa realizar un análisis morfológico del discurso; estudiar la estructura de las palabras para delimitarlas y clasificarlas.
- **Nivel sintáctico:** Aquí se realiza un análisis de sintaxis, el cual incluye la acción de dividir una oración en cada uno de sus componentes.
- **Nivel semántico:** Este nivel es un complemento del anterior, en el análisis semántico se busca entender el significado de la oración. Las palabras pueden tener múltiples significados, la idea es identificar el significado apropiado por medio del contexto de la oración.
- **Nivel discursivo:** El nivel discursivo examina el significado de la oración en relación a otra oración en el texto o párrafo del mismo documento.
- **Nivel pragmático:** Este nivel se ocupa del análisis de oraciones y cómo se usan en diferentes situaciones. Además, también cómo su significado cambia dependiendo de la situación.

5.1.5. Teoría de Grafos

Un grafo es un conjunto, no vacío, de objetos llamados vértices (o nodos) y una selección de pares de vértices, llamados aristas (edges en inglés) que pueden ser orientados o no.

Típicamente, un grafo se representa mediante una serie de puntos (los vértices) conectados por líneas (las aristas) y sirve para representar datos que tienen atributos comunes y permiten establecer una relación entre si.

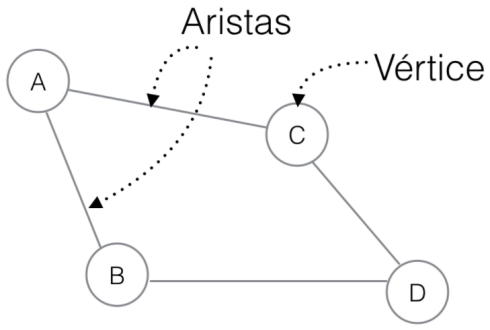

Figura 5.1: Modelo de Grafo no dirigido

5.1.6. Token

Palabra clave o con relevancia dependiendo del contexto.

5.1.7. Stop Word

Palabra común o repetitiva en un texto. Aunque aparece innumerables veces, no brinda valor o significado. *e.g.* Conectores, pronombres, etc.

5.1.8. Lematización

Inglés: Lemmatization. Transformación morfológica que cambia una palabra tal como aparece en el texto en ejecución en la forma básica o de diccionario de la palabra, que se conoce como lema, al eliminar la terminación de inflexión de la palabra [16][17]. En pocas palabras, consiste en reducir las palabras a su raíz, de modo que palabras derivadas de una misma se consideran como igual sin importar su tiempo verbal o si corresponden al singular o plural [18].

```
beautiful: beauty
corpora: corpus
```

Listing 5.1: Ejemplo de Lematización

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

Figura 5.2: Explicación sobre Lematización

5.1.9. Estemizado

Inglés: Stemming. Proceso de eliminar los últimos caracteres de una palabra determinada, para obtener una forma más corta, incluso si esa forma no tiene ningún significado. Reduce las palabras con la misma raíz a una forma común [16].

```
beautiful: beauti
```

```
corpora: corpora
```

Listing 5.2: Ejemplo de Estemizado

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Figura 5.3: Explicación sobre Estemizado

5.1.10. Agrupamiento de Datos o Clustering

La agrupación de datos es el proceso de identificar agrupaciones naturales o agrupaciones dentro de datos multidimensionales basados en alguna medida de similitud [19]

Clúster

Tomado del idioma inglés, *Clúster* hace referencia a un grupo con características comúnes.

5.1.11. Algoritmo Node2Vec

Por definición, *Node2Vec* es un marco algorítmico para el aprendizaje continuo de características representados por nodos en redes. En *Node2vec*, se aprende a interpretar los nodos a un espacio de bajas dimensiones o características que maximizan la probabilidad de preservar la vecindad de nodos [20].

5.1.12. Algoritmo K-means

K-means es un método de agrupamiento de observaciones dentro de un número específico de grupos separados. La k se refiere al número de clusters especificados. Existen varias medidas de distancia para determinar qué observación se va a agregar a qué grupo. El algoritmo tiene como objetivo minimizar la medida entre el centroide del grupo y la observación dada al agregar iterativamente una observación a cualquier grupo y terminar cuando se alcanza la medida de distancia más baja [21].

Elbow Mehod

Implica ejecutar el algoritmo varias veces en un bucle, con un número creciente de opciones de clúster y luego trazar una puntuación de agrupamiento en función del número de clústers [22].

Capítulo 6

Diseño Metodológico

La metodología propuesta para el desarrollo de la investigación se divide en tres aspectos fundamentales: *Delimitación del Modelo*, *Desarrollo*, *Análisis y Resultados*.

1. La etapa de *Delimitación del Modelo* consiste en describir la estrategia para el desarrollo e implementación, así como la explicación de un grupo de componentes técnicos a tener en cuenta para la comprensión de dicho planteamiento.

2. A partir de las bases de conocimiento propuestas, se muestra la implementación de algunos *Algoritmos y Técnicas*.

3. Luego de tener una implementación óptima que represente la solución planteada, se procede a realizar un *Análisis* para poder comparar los métodos a través de métricas y diferentes criterios de decisión. Posteriormente, se procede a sustentar los *Resultados* obtenidos del paso anterior.

Capítulo 7

Delimitación del Modelo

7.1. Plan

El sistema utilizará el enfoque basado en el *Procesamiento de Lenguaje Natural* para la extracción de información. Comparará la información extraída de cada documento y la relacionará a través de una red creada utilizando características similares encontradas. Este tipo de enfoque no solo reduce significativamente el esfuerzo humano para la clasificación de documentos, también proporciona un mecanismo para recopilar información y obtener radiografía general de las producciones del grupo de investigación.

7.2. Estrategia

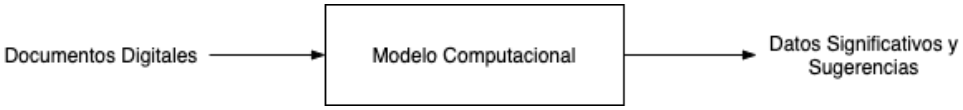

Figura 7.1: Propuesta general de Implementación

La manera en la que se decide avanzar en la sustentación de la propuesta planteada (Figura 7.1), se resumen en:

- Definición de tipos de datos de entrada requeridos.
- Selección y descripción de algoritmos/técnicas para tratamiento y manipulación de Documentos Digitales.
- Selección y descripción de algoritmos/técnicas de Inteligencia Artificial apropiadas para interpretación, clasificación y agrupamiento de datos.
- Implementar cada técnica sobre los conjuntos de datos elegidos.

- Establecer una métrica o criterio de aceptación para concluir sobre los resultados obtenidos.

7.3. Propuesta

Hasta esta sección, se tiene claridad sobre las entradas y las salidas que harán posible la validación de los objetivos que se proponen Figura 7.1.

A grandes rasgos, el *Modelo Computacional* es el encargado de recibir y procesar las entradas para arrojar unas salidas con información útil que sirva como soporte a los planteamientos propuestos, tomar decisiones y formular conclusiones.

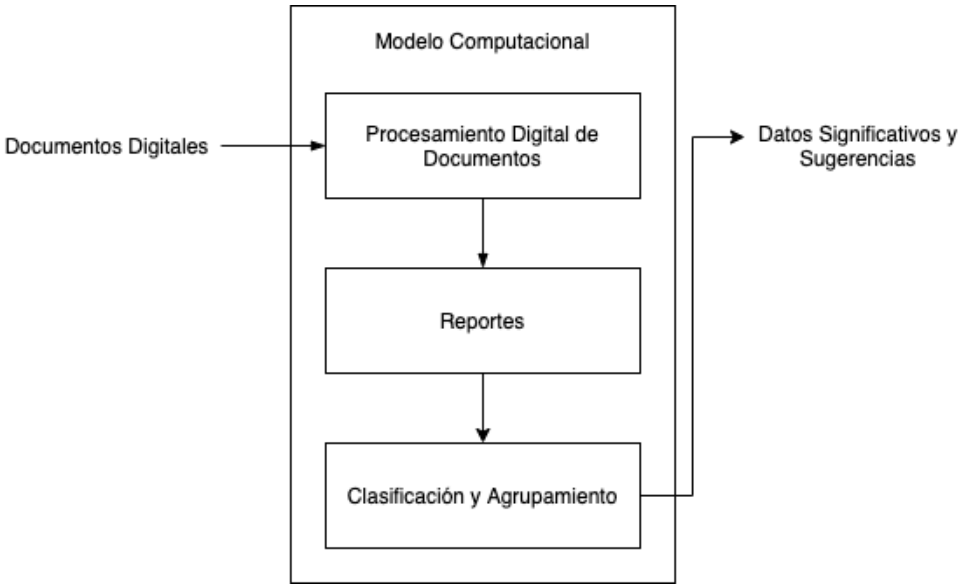

Figura 7.2: Propuesta de Implementación

De acuerdo a la idea anterior, se propone una serie de pasos para obtener el resultado esperado:

- **Procesamiento Digital de Documentos:** Extracción y manipulación de información que se encuentra contenida en las entradas (Específicamente, *Documentos Digitales* Subsección 5.1.3). Se trata cada documento individualmente y se adapta a una estructura que brinde fácil acceso y manipulación a los datos.
- **Reportes:** Luego de tener la información de cada entrada por individual, se trata de agrupar en conjuntos de datos que se puedan representar en alguna estructura que describa alguna relación.

- **Clasificación y Agrupamiento:** En este punto se busca encontrar factores comunes que digan como se relacionan los datos (si existe alguna relación). Este paso puede dar información de grupos y del conjunto de datos completos.

En las siguientes secciones se desarrollarán los puntos propuestos en Sección 7.2, con el fin de elegir las herramientas adecuadas y delimitar el alcance del desarrollo

7.4. Definición de Tipos de Datos de Entrada Requeridos

7.4.1. Contexto

En la actualidad existen múltiples maneras de almacenar información dentro de ordenadores. La información de un ordenador está almacenada en lo que se llaman archivos.

Dentro de los tipos de archivos de datos se pueden crear grupos, especialmente por la temática o clase de información que guarden.

- De texto: txt, doc, docx, etc.
- De imagen: jpg, gif, bmp, png, etc.
- De lectura: pdf, epub, azw, ibook, etc.

El tipo de archivos que pueden ser compatibles para almacenar información sobre Artículos(Papers) y que se selecciona para la implementación, es el tipo PDF.

PDF *(siglas en inglés de Portable Document Format - formato de documento portátil)* es un formato de almacenamiento para documentos digitales independiente de plataformas de software o hardware. Este formato es de tipo compuesto (imagen vectorial, mapa de bits y texto).

7.4.2. Estructura de los datos

Se requiere que los datos tengan ciertos valores que brinden información característica y única sobre cada publicación que se analiza. Según la estructura base que se define en Sección 5.1.2, se sugiere establecer los siguientes parámetros como requeridos:

- *Encabezado:* Título y Resumen (Abstract)
- *Cuerpo:* Introducción, Materiales y métodos
- *Desenlace:* Resultados y Discusión

7.4.3. Obtención de los datos

Datos Genéricos

Se denomina *Datos Genéricos* al conjunto de datos que cumplen con los requerimientos de Subsección 7.4.2 pero que no están relacionados necesariamente al caso de estudio o a las pruebas que se quieren sustentar.

El conjunto de *Datos Genéricos* seleccionados proviene de ARXIV y tiene las siguientes características:

- 31.000+ artículos/papers
- Temas relacionados con Aprendizaje de Máquina (Machine Learning), Ciencias de la Computación, Inteligencia Artificial (AI), Visión por Computador (Computer Vision), entre otros.
- Publicados entre los años 1992 a 2018.
- Idioma: Inglés

Datos Reales

El concepto de *Datos Reales* hace referencia a datos que provienen de una situación no simulada (*e.g.* un caso de estudio), los cuales brindan información característica de cierto problema a abordar.

El conjunto de *Datos Reales* seleccionados proviene del **Grupo de Investigación - Química UTP** y tiene las siguientes características:

- 56 artículos/papers
- Temas relacionados con Química, Biología y Matemáticas
- Idioma: 35 Español y 21 Inglés

Cada uno de los datos elegidos, contienen el formato y la información esperada:

```
author:[{name: ''}, {name: ''}, ...]
day: ''
id: ''
link: [
    {rel: '', href: '...', type: 'text/html'},
    {rel: '', href: '...', type: 'application/pdf'}
]
month: ''
```

```
summary: ''
tag: [ ... ]
title: ''
year: ''
```

Listing 7.1: Estructura del conjunto de datos

La información base se obtiene a partir de archivos digitales tipo PDF (Como se explica en Subsección 7.4.1).

7.5. Tratamiento y Manipulación de Documentos Digitales

La preparación de la información es un proceso clave. La preparación consiste en una serie de pasos que permiten obtener la datos útiles y desechar aquello que no es relevante para el resto de procesos.

Por ejemplo, se consideran valores relevantes a frases (exactamente las palabras que las componen) evitando signos de puntuación y conectores.

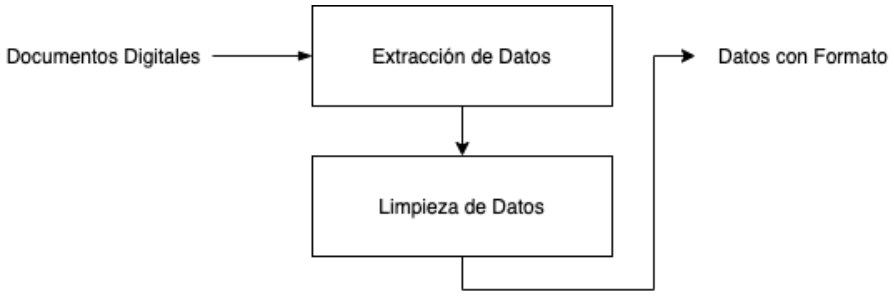

Figura 7.3: Flujo ideal de Procesamiento de Documentos Digitales

7.5.1. Extracción de Datos

Consiste en separar los tipos de datos de interés. En este caso se trabajara solo con caracteres alfanuméricos que representen texto. También se podrían incluir otros tipos de datos como imágenes pero no son del interés en este proyecto.

7.5.2. Limpieza de Datos

Aunque en el paso Subsección 8.1.1 se obtiene un tipo de datos con el cuál se podría trabajar, se debe aplicar una serie de técnicas para separar la información verdaderamente

útil y descriptiva del conjunto de datos:

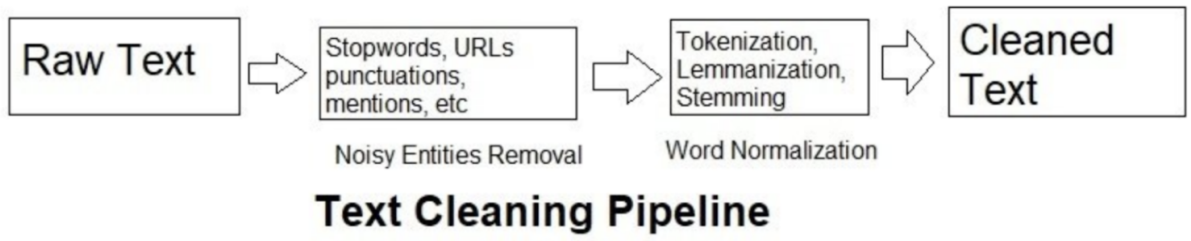

Figura 7.4: Flujo ideal propuesto para Limpieza de los Datos [1]

Formato del Texto

También conocido como *Eliminación de Entidades "Ruidosas"*

- Eliminación de hipervínculos
- Extracción de *tokens* (Subsección 5.1.6)
- Eliminación de signos de puntuación
- Eliminación de *stop words* (Subsección 5.1.7)

Normalización de Palabras

- Lematización (Subsección 5.1.8)
- Estemizado (Subsección 5.1.9)

7.6. Clasificación y Agrupamiento de Datos

Luego de extraer toda la información que se considera como valiosa de las entradas, se busca alguna forma de relacionar las diferentes entradas y tratar de entender las características que tienen en común.

Siguiendo la premisa anterior, se propone:

- Crear una red de conexiones apoyada en *Teoría de Grafos* (Subsección 5.1.5) para relacionar la información
- Transformar la red de datos de tal manera que cumpla con los requisitos:
  - Conserve las relaciones más fuertes y elimine aquellas poco significativas (evitando relaciones entre Documentos que no tienen características comunes).

    - Tenga una forma entendible y fácil de interpretar

- Agrupar los datos si se encuentran similitudes.

7.7. Criterio de Aceptación

Los siguientes elementos se colocan en consideración para poder concluir si el modelo propuesto soluciona el problema planteado (o da alternativas de solución):

1. Red conexiones construida con las entradas
2. Número de grupos construidos
3. Tamaño de los grupos
4. Proporción entre total de entradas y grupos
5. Características comunes de los grupos

Basados en los resultados que se obtengan en la observación de los puntos anteriores, se puede evaluar si el modelo es adecuado o no.

Capítulo 8

Desarrollo

Resumen

Se planea construir un sistema que pueda servir como soporte de recomendación para la realización de investigaciones basadas en las temáticas manejadas por un Grupo de Investigación. Se compone de dos fases:

1. En la primera fase, se realiza un proceso extracción de información utilizando los artículos o publicaciones del grupo.

2. Una segunda fase, realiza un análisis de los datos obtenidos utilizando métodos computacionales y técnicas de Inteligencia Artificial. Esto con el fin de poder otorgarle al Grupo de Investigación un plan estratégico para trabajar en nuevas investigaciones y productos.

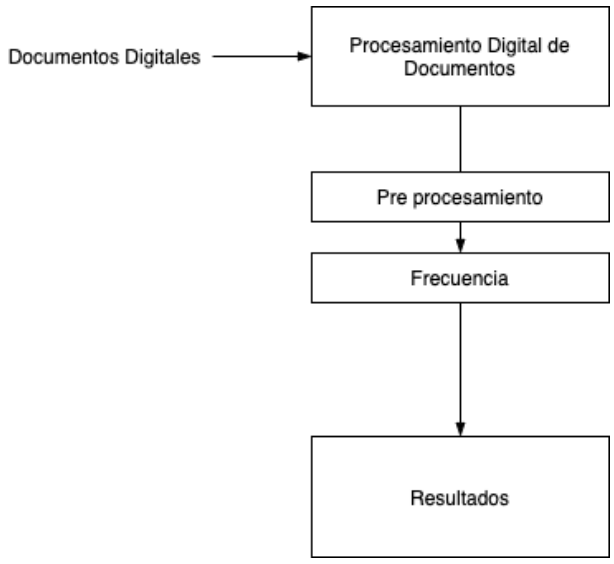

Figura 8.1: Primera parte del modelo propuesto

A partir del modelo propuesto en Sección 7.3, se procede al desarrollo de la primera parte (descrita en Figura 8.1).

8.1. Procesamiento digital de documentos

8.1.1. Obtención de Datos

Si la entrada al modelo cumple con los requerimientos planteados en Subsección 7.4.1, se procede a la extracción de los datos (cómo se plantea en Subsección 8.1.1). Para esto se utiliza alguna herramienta capaz de manipular archivos digitales (Apache Tika [23] o PyPDF [24] son ejemplos).

Como resultado, se esperaría obtener una cadena de caracteres con la información extraída. Algo parecido a lo que se conoce como un texto plano:

```
‘LONGSHORT–TERMMEMORYBASEDRECURRENTNEURALNETWORK
ARCHITECTURESFORLARGEVOCABULARYSPEECHRECOGNITION Has.imSak,
AndrewSenior,Franc.oiseBeaufays Google f hasim,andrewsenior,
fsb@google.com g ABSTRACT LongShort–TermMemory(LSTM)
isarecurrentneuralnetwork (RNN)architecturethat has
beendesignedtoaddressthevanish– ingandexploding gradient
problemsofconventionalRNNs.Unlike feedforwardneuralnetworks,
RNNshavecyclicconnectionsmak– ingthem powerful for
modelingsequences.Theyhavebeensuc–
cessfullyusedforsequencelabelinga’
```

8.1.2. Pre procesamiento

Como se sugiere en Subsección 7.5.2, este proceso se encarga de realizar una normalización de la información de cada archivo que se ha almacenado en el paso anterior. Es decir, se pasan todas las palabras a un formato común (en este caso específico a letras minúscula) de modo que se puedan evitar caracteres especiales y otras formas de escritura.

En una primera fase, luego de obtener el texto plano de cada documento en Subsección 8.1.1 (sin ningún tratamiento), se realiza una lectura del mismo para determinar palabras o partes de texto que no aportan al análisis. Entre estas se encuentran las “stopwords”Subsección 5.1.7.

Una vez efectuado el paso anterior, se procede con un procesamiento de Tokenización Subsección 5.1.6 (Las palabras son separadas entre sí, manteniendo solo una copia de

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Figura 8.2: Ejemplo de procesamiento de **Stop Words**

cada una existente en el cuerpo del texto), Lematización Subsección 5.1.8 y Estemizado Subsección 5.1.9.

Luego de efectuados estos pasos, se obtiene una lista de palabras para el siguiente paso que consiste en la obtención de frecuencia (número de apariciones de cada palabra en el texto)

Nota: Dentro de los procesos de Tokenización, Lematización y Estemizado se utilizan los algoritmos proporcionados por la librería nltk [25]

8.1.3. Frecuencia

Una vez obtenida la lista de palabras, se procede a realizar un conteo de cada palabra para obtener su frecuencia (número de repeticiones dentro del texto).

8.1.4. Resultados del Procesamiento

El resultado de esta sección es una estructura de datos que representa las palabras significativas y su respectiva frecuencia.

```
[
    'word_one': 90,
    'word_two': 8,
    ...
    'word_n': 100
]
```

Listing 8.1: Estructura de datos: Palabras significativas y Frecuencias

8.2. Reportes

Se denomina reporte al procesamiento de un grupo de documentos digitales usando los resultados obtenidos del Procesamiento Digital de Documentos

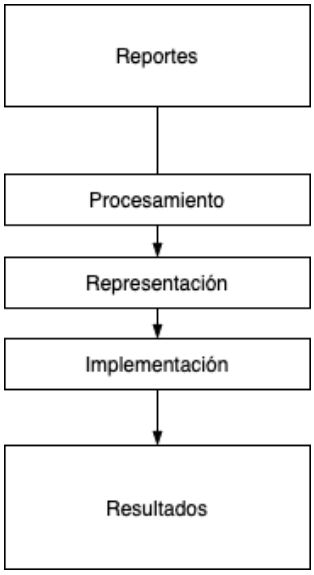

Figura 8.3: Segunda parte del modelo propuesto

8.2.1. Procesamiento

Es necesario crear una estructura de datos en la cual se agrupen los datos característicos obtenidos para cada uno de los documentos que conforman el reporte. Se selecciona una lista de objetos con llave-valor como ya que se adapta a las necesidades de manipulación de datos para los siguientes pasos.

```
[
    {
        'id': 0,
        'title': 'Paper One',
        'frequency': {
            'word_one': 50,
            'word_two': 5,
            ...
        }
    },
    {
        'id': 1,
```

```
        'title': 'Paper Two',
        'frequency': {
            'word_one': 5,
            'word_two': 25,
            ...
        }
    },
    ...
]
```

Listing 8.2: Estructura de datos: Representación de Reportes

- El identificador único de una publicación (id) es utilizado para representar que las publicaciones son singulares
- El título de la publicación (title) es utilizado en algunas representaciones para poder identificar fácilmente una publicación.
- *frequency* contiene el grupo de palabras representativas y el número de apariciones dentro del contenido del documento digital.

8.2.2. Representación

A partir de la estructura de datos generada en el Procesamiento de Reportes, es necesario plantear una representación que permita manipular y operar sobre la información obtenida de los documentos, con el fin de poder observar similaridades y posterior poder crear agrupaciones. Se encuentra en la Teoría de Grafos una alternativa viable.

Se elige un *Grafo* como medio para representar un reporte. Conceptualmente, un grafo es un conjunto de vértices(nodos) relacionados por enlaces aristas(arcos). Permiten representar relaciones binarias en un conjunto [26].

Partiendo de la definición anterior, se procede a identificar las partes/atributos con los cuales se construye el *Grafo*:

- **Nodos/Vértices:** Representan una Publicación/Documento.
- **Aristas/Arcos:** Conexiones entre las Publicaciones. Se busca que la relación entre las Publicaciones se de a través de elementos extraídos desde su contenido, por lo cual se utilizan las palabras características que se tengan en común.

En este punto, se tiene una representación de un Reporte usando un *Grafo*. Aún no es suficiente ya que existe una relación entre los nodos (Publicaciones) pero todas las

relaciones son iguales y no hay una variable que describa la relación.

Para solucionar el problema anterior, se aprovecha una característica de las aristas del *Grafo* conocida *Peso o Costo*, el cual sirve para determinar la fuerza entre la relación de dos nodos.

- **Pesos/Costos:** Como propuesta inicial, se establece una medida usando el número de palabras claves que se encuentran en ambos nodos (Representados como n en la gráfica).

Figura 8.4: Relación pesos y publicaciones

Es decir, se tiene el caso: 2 publicaciones relacionadas (*Publication 1* y *Publication 2*)

Publication 1	Publication 2
blockchain: 5	blockchain: 105
crypto: 100	crypto: 0
hash: 50	computer: 15
economy: 15	paper: 8
...	....

Cuadro 8.1: Contenido que relaciona dos Publicaciones

El peso para esta relación es $n = 2$.

8.2.3. Implementación Propuesta

Utilizando la librería networkx que permite hacer representaciones de redes [27] (entidades conectados), por lo cual se puede utilizar para la implementación de un *Grafo*.

La implementación crea un objeto del tipo nodo y se agrega una lista con los nodos y otra con las conexiones entre los nodos.

```
Node(s) = [node_1, node_2, node_3, ... , node_n]
Edge(s) = [(node_0, node_2, {'weight': 633}),
           (node_0, node_4, {'weight': 443}), ... ]
```

8.2.4. Resultados del Procesamiento

El resultado de esta sección es una estructura de datos que representa las Publicaciones que componen un Reporte y cómo se relacionan entre sí.

Figura 8.5: Relaciones existentes entre publicaciones

8.3. Clasificación y Agrupamiento

El siguiente paso en este proceso (Figura 8.6) tiene como objeto convertir la salida (Reporte representado en un Grafo) a una representación en la que se puedan agrupar los nodos con características similares, apoyándose en las diferentes características del Grafo.

El primer reto que se presenta, es crear una interpretación del *Grafo* ya que es una estructura considerada inicialmente como una representación no dimensional. El propósito de esta interpretación es facilitar la comprensión de la estructura y poder realizar diferentes operaciones en espacios conocidos.

Aunque existen variedad de algoritmos que podrían ayudar a interpretar una estructura como la requerida, el algoritmo seleccionado es Node2Vec. La decisión de elegir dicho algoritmo se toma por su característica de llevar el contenido de una estructura de tipo *Grafo* a un *Espacio Euclidiano*)

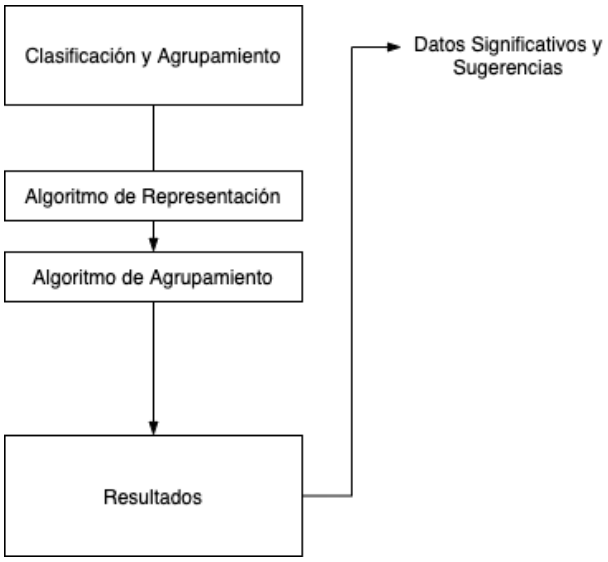

Figura 8.6: Tercera parte del modelo propuesto

8.3.1. Algoritmo (Node2Vec)

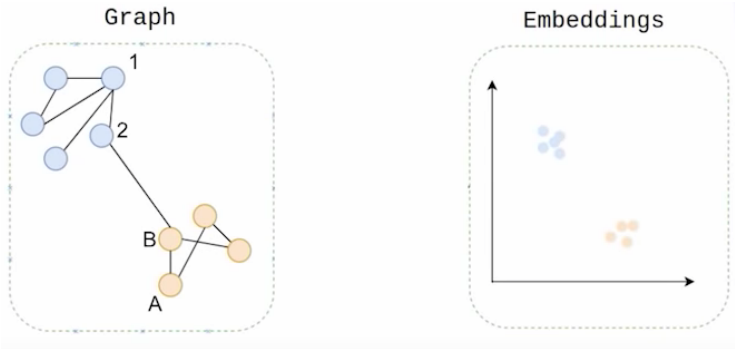

Figura 8.7: Agrupamiento de nodos

Una alternativa para implementar este tipo de algoritmo se encuentra en la librería node2vec. Además es compatible con la implementación del Grafo construido en Subsección 8.2.4

Cómo salida de esta operación se obtiene una lista con puntos (coordenadas cartesianas). Cada punto representa un nodo, por lo cual el número de puntos es equivalente al número de nodos (Publicaciones).

```
[(0, 1), (1, 2), (5, 2), ...]
```

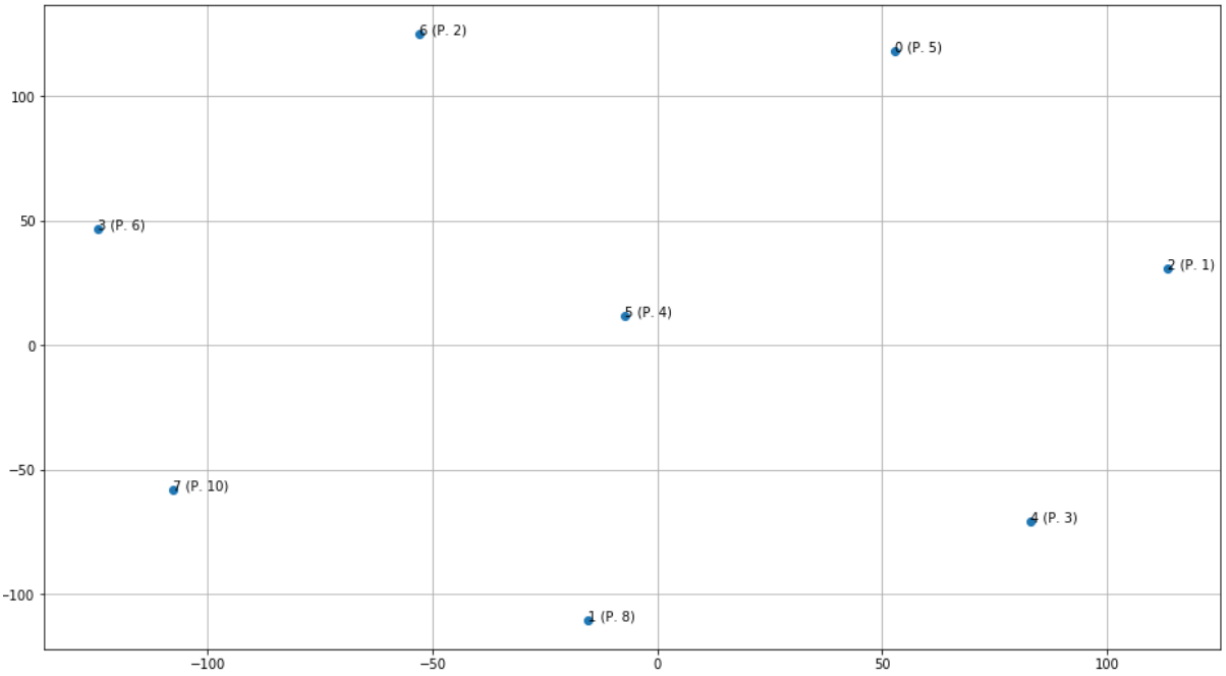

Figura 8.8: Representación del resultado de esta sección en un plano cartesiano de dos dimensiones: Cada punto representa una Publicación

Nota: Aunque la visualización resultante contiene un plano con coordenadas, los ejes no tendrán etiquetas ni significado ya que solo se utilizan para poder visualizar el conjunto de puntos que representan las publicaciones.

En este punto, ya se tiene una interpretación de las Publicaciones sobre un espacio donde se puede operar a través de operaciones matemáticas lo cual facilita el último paso del procesamiento (Figura 8.6): buscar una estrategia para agrupar las Publicaciones que posean características similares. Dicha estrategia es denominada como clustering o agrupamiento.

Existen varios métodos y algoritmos para este tipo de requerimiento. El elegido para el desarrollo de la sección es el denominado *K-means*.

8.3.2. Algoritmo (K-means)

La decisión más importante en este algoritmo es la de la elección del número de centroides. Existen diferentes técnicas de decisión y la más utilizada es *Elbow Method*.

La idea es comenzar con $K=2$ (centroides) y continuar incrementando el número en 1, calculando sus grupos y el costo que trae con el entrenamiento. Para algún valor para K, el costo cae dramáticamente, y después de eso alcanza una estabilidad a medida que van

aumentando los centroides. Este será el número de centroides adecuado [28].

Demostración - Elbow method

A continuación se visualiza los resultados de *Elbow Method* aplicado sobre un conjunto de 8 aplicaciones. Se tiene como premisa que mínimo se usa 1 y que máximo 8 centroides para agrupar dichas Publicaciones.

```
1 104555.3
2 64334.97010707855
3 37203.573181152344
4 24954.224899291992
5 16871.171875
6 11247.44287109375
7 5623.71875
8 0.0
```

Listing 8.3: Interpretación de resultados: *Número de Centroides* y *Valor del Costo de Entrenamiento*

El Costo de Entrenamiento hace referencia al cálculo de distancias entre las Publicaciones hasta el centroide correspondiente (utilizando el método de Mínimos Cuadrados [29]) Figura 8.12.

Una estrategia para definir el número correcto de centroides y que ayuda a tomar la decisión (Eliminando el concepto de método visual de decisión), es el análisis de las diferencias entre los costos del entrenamiento obtenidos con cada una de las variaciones del número de centroides. Para determinar que la medida se estabiliza, se compara la medida de diferencia *n* contra *n - 1*. Si el cambio entre ellos varía entre -10 % y 10 %, se toma como el punto de estabilización Figura 8.13.

```
Suggested number of clusters: 5
(5, 16871.171875)
```

8.3.3. Resultados del Procesamiento

La estructura de datos que arroja este algoritmo es una lista de listas. Cada una de las listas representa un cluster o grupo de Publicaciones y cada Publicación está representada

por un identificador único.

```
[[1 ,  4] ,  [3 ,  6] ,  [0 ,  2] ,  [5] ,  [7]]
```

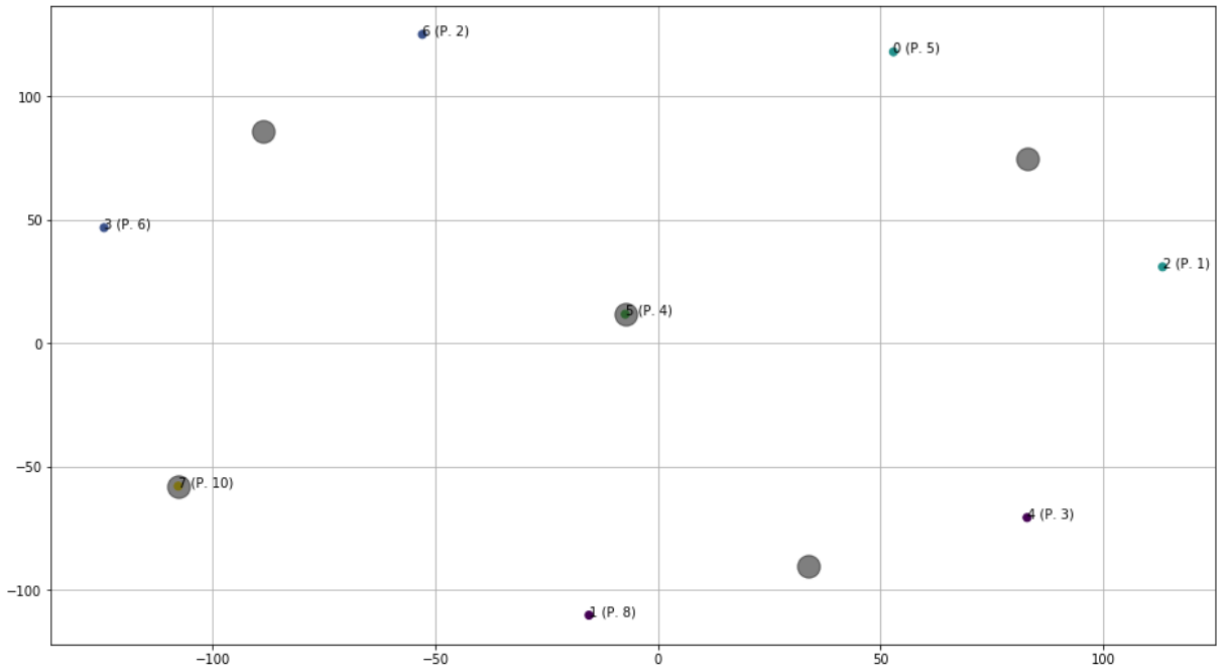

Figura 8.9: Clusters de información

8.3.4. Aplicación Web

Con el fin poder realizar un seguimiento y ver la interacción del proyecto, se creó una interfaz utilizando el framework django, el cual se compone de la diversas funcionalidades que se explicaran en el presente documento.

Más información en ANEXOS: Apéndice C

8.3.5. Resultados de la Sección

Para visualizar la salida del modelo (descrita en la última parte del modelo), se utiliza la herramienta Subsección 8.3.4 para crear una interfaz con la información relevante de todo el flujo.

De acuerdo a la interfaz (Figura 8.10), describimos los siguientes aspectos:

Figura 8.10: Interfaz - Salida del modelo

1. **Clusters/Grupos de Publicaciones:** Grupos de clasificación de las publicaciones en donde se agrupan por los criterios de los siguientes puntos (Puntos 3 y 4).

2. **Publicaciones:** Publicaciones que conforman el grupo. Se visualiza un identificador único asignado en las pruebas y el título.

3. **Palabras en común (Contenido):** Palabras que se encuentran en todas las Publicaciones y que después del procesamiento se consideran relevantes Figura 8.11.

   En el eje vertical se observa cada palabra/token contra la frecuencia o número de repeticiones(acumulada entre todas las Publicaciones del grupo) en el eje horizontal.

4. **Palabras clave en común (Tags):** En algunas publicaciones y documentos, los tags o palabras claves se encuentran en la sección de introducción y su función es describir con palabras claves el contenido.

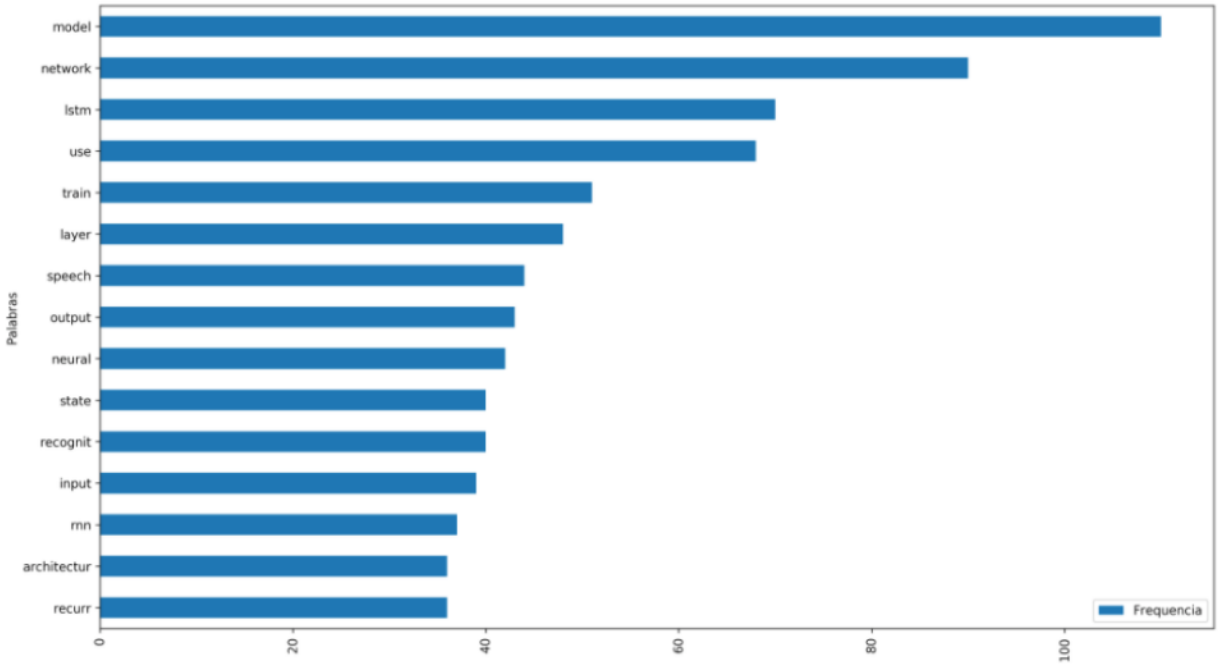

Figura 8.11: Palabras en común (Contenido)

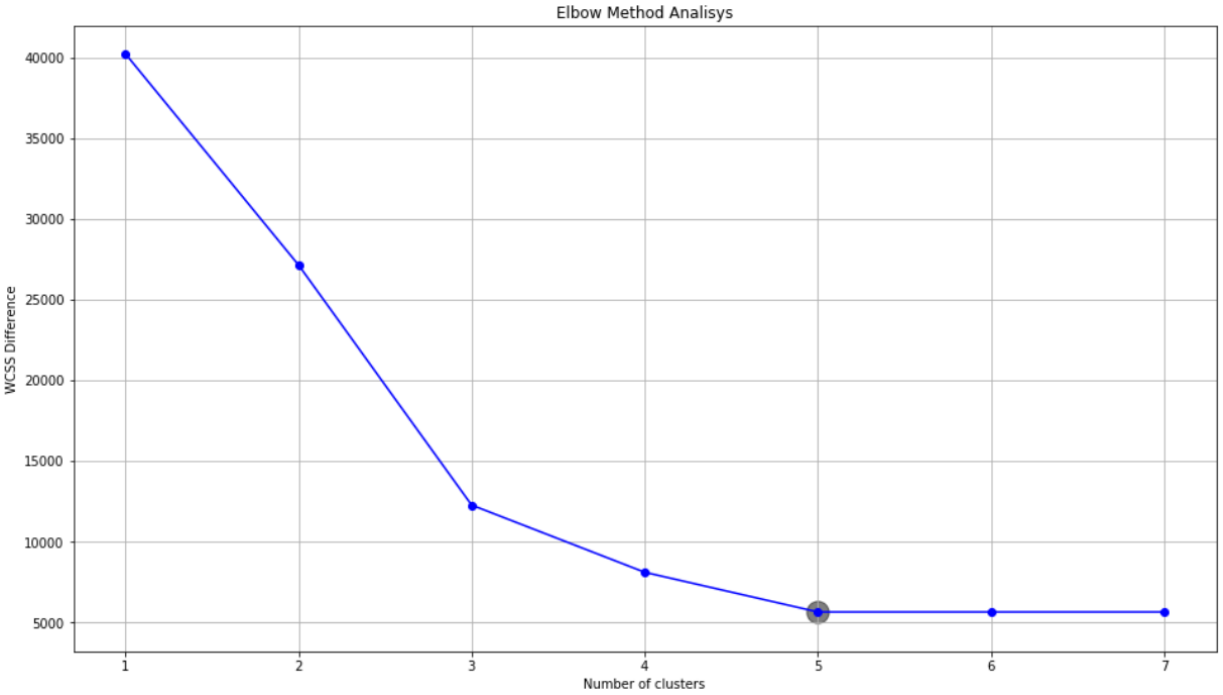

Figura 8.12: Costo de entrenamiento - Elbow method

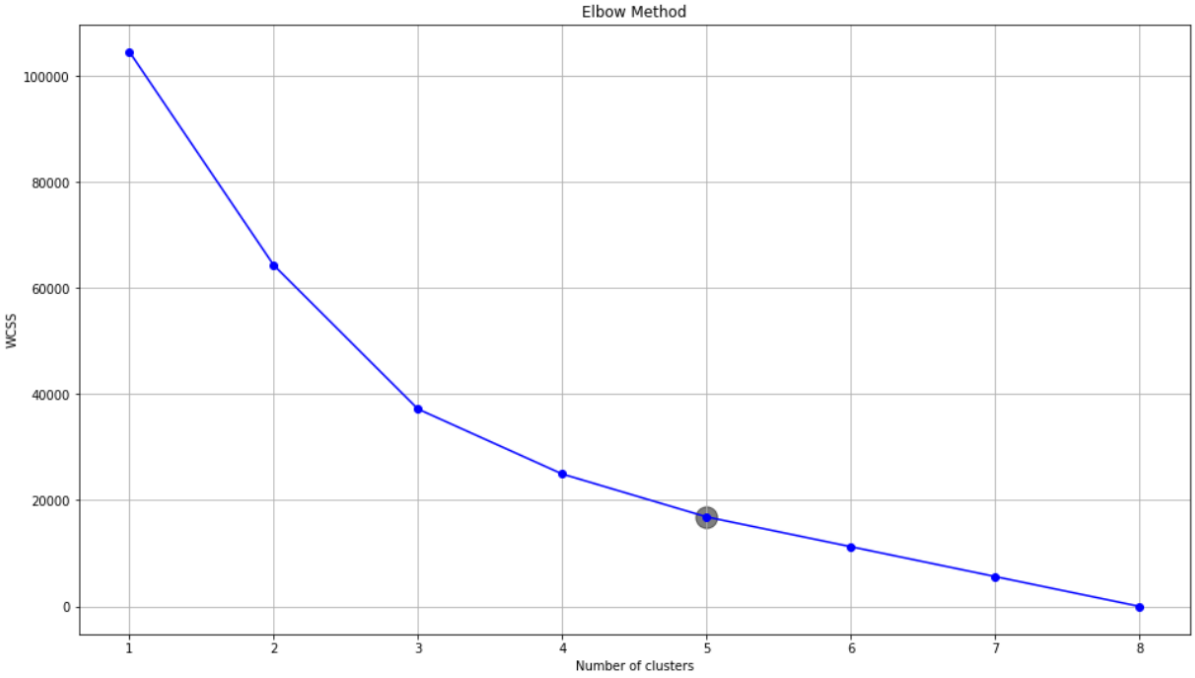

Figura 8.13: Punto de estabilización - Elbow method

Capítulo 9

Diseño y Ejecución de Pruebas

9.1. Plan

Diseñar y ejecutar una serie de pruebas sobre el modelo (Figura 7.2) desarrollado en Capítulo 8

9.2. Estrategia

Las condiciones propuestas para el desarrollo de las pruebas son:

- Segmentar las pruebas en dos partes:
  - **Etapa I: Procesamiento Digital de Documentos**
  - **Etapa II: Flujo de Trabajo Completo**
- Para cada una de las etapas propuestas, se evaluarán la mismas condiciones para ambos grupos de datos elegidos en Subsección 7.4.3

Nota: Se segmentan las pruebas (primera parte del modelo Sección 8.1 - Entradas y Procesamiento de Documentos) ya que existen variedad de casos posibles que se pueden presentar en esta etapa y se quieren exponer los más frecuentes.

9.3. Etapa I: Procesamiento Digital de Documentos

Pruebas relacionadas a Sección 8.1.

Las condiciones de prueba propuestas para esta etapa:

- Seleccionar 3 papers (Archivo tipo PDF)

- Aplicar el algoritmo propuesto sobre cada archivo
- Repetir el paso anterior por 3 iteraciones
- Comparar el resultado de las 3 iteraciones
- El resultado esperado es que la salida de todas las iteraciones sean iguales (Número de palabras y frecuencias)

9.3.1. Documentación

Dataset Grupo de Investigación - Química UTP

- **Caso 1: Production of scopolamine by normal root cultures of Brugmansia candida (Inglés)**
  Observaciones:
  - Salida: Tabla 9.5
  - Resultado: Esperado
  - Más información sobre el articulo: [3]
- **Caso 2: Cuantificacion de licorina en callos y raices cultivados in-vitro de crinum x powelli "album"(amaryllidaceae) por cromatografia liquida de alta eficiencia (hplc). (Español)**
  Observaciones:
  - Salida: Tabla 9.6
  - Resultado: No esperado
  - Más información sobre el articulo: [4]
- **Caso 3: Diospolysaponin A, a new polioxigenated spirostanol saponin from tubers of Dioscorea polygonoides. (Inglés)**
  Observaciones:
  - No se obtuvieron resultados de este procesamiento.
  - Resultado: No esperado
  - Más información sobre el artículo: [30]

Dataset ARXIV

- **Caso 4: Bitcoin (Inglés)**
  Observaciones:
  - Salida: Tabla 9.7

  - Resultado: Esperado
  - Más información sobre el articulo: [5]

- **Caso 5: A Clustering Method Based on K-Means Algorithm (Inglés)**
  Observaciones:

  - Salida: Tabla 9.8
  - Resultado: Esperado
  - Más información sobre el articulo: [6]

- **Caso 6: The COVID-19 epidemic (Inglés)**
  Observaciones:

  - Salida: Tabla 9.9
  - Resultado: Esperado
  - Más información sobre el artículo: [7]

Nota: Cada caso fue iterado 3 veces. Se publica únicamente 1 tabla ya que el resultado en todas las iteraciones fue idéntico.

9.4. Etapa II: Flujo de Trabajo Completo

Pruebas relacionadas a Sección 8.2 y Sección 8.3

Las condiciones de prueba propuestas para esta etapa:

- Seleccionar un grupo de papers (Archivo tipo PDF) para generar reportes
- No se documentarán las pruebas sobre la primera parte del modelo (Figura 8.1) ya que se hicieron en la anterior Sección 9.3 para algunos casos de documentos que pueden servir de entrada. Se asegura que para estos casos de prueba, la salida de esta sección fue exitosa y que es válida como entrada a la segunda parte del modelo (Figura 8.3)
- Aplicar el algoritmo (Sección 8.2) sobre cada grupo
- Repetir el paso anterior por 3 iteraciones
- Comparar el resultado de las 3 iteraciones
- Aplicar el algoritmo (Sección 8.3) sobre cada grupo
- Repetir el paso anterior por 3 iteraciones

- Comparar el resultado de las 3 iteraciones
- El resultado esperado es que la salida de todas las iteraciones sean iguales:
  - El grafo que representa un reporte (Sección 8.2)
  - El número de clusters o grupos de Publicaciones (Sección 8.3)

9.4.1. Documentación

Caso I: Dataset Grupo de Investigación - Química UTP

- **Entrada de 3 Publicaciones**
  Observaciones:
  1. Listado de Publicaciones: Tabla 9.10
  2. Reporte formado con las Publicaciones de entrada: Figura 9.1
  3. Clusters o grupos de Publicaciones formados: Iteración 1, Iteración 2 y Iteración 3
  4. Salida: Figura 9.10
  5. Resultados:
     - Reporte: Esperado
     - Clusters o grupos de Publicaciones: Aceptable
     - Salida del modelo: No esperado

- **Entrada de 4 Publicaciones**
  Observaciones:
  1. Listado de Publicaciones: Tabla 9.11
  2. Reporte formado con las Publicaciones de entrada: Figura 9.2
  3. Clusters o grupos de Publicaciones formados:
     - Iteración 1. Salida: Figura 9.14
     - Iteración 2. Salida: Figura 9.15
     - Iteración 3. Salida: Figura 9.16
  4. Resultados:
     - Reporte: Esperado
     - Clusters o grupos de Publicaciones: Esperado
     - Salida del modelo: Aceptable

- **Entrada de 6 Publicaciones**
  Observaciones:

1. Listado de Publicaciones: Tabla 9.12
2. No hay resultado del procesamiento para ninguna de las iteraciones.
3. Resultados:
   - Reporte: No esperado
   - Clusters o grupos de Publicaciones: No esperado
   - Salida del modelo: No esperado

Caso II: Dataset ARXIV

- **Entrada de 3 Publicaciones**
  Observaciones:

  1. Listado de Publicaciones: Tabla 9.13
  2. Reporte formado con las Publicaciones de entrada: Figura 9.4
  3. Clusters o grupos de Publicaciones formados: Iteración 1, Iteración 2 y Iteración 3.
     Salida: En este caso, todas las publicaciones quedan contenidas en el mismo grupo/cluster (Sin características comunes)
  4. Resultados:
     - Reporte: Esperado
     - Clusters o grupos de Publicaciones: Aceptable
     - Salida del modelo: No esperado

- **Entrada de 4 Publicaciones**
  Observaciones:

  1. Listado de Publicaciones: Tabla 9.14
  2. Reporte formado con las Publicaciones de entrada: Figura 9.5
  3. Clusters o grupos de Publicaciones formados:
     - Iteración 1. Salida: Figura 9.23
     - Iteración 2. Salida: Figura 9.24
     - Iteración 3. Salida: Figura 9.25
  4. Resultados:
     - Reporte: Esperado
     - Clusters o grupos de Publicaciones: Esperado
     - Salida del modelo: Aceptable

- **Entrada de 8 Publicaciones**
  Observaciones:

1. Listado de Publicaciones: Tabla 9.15
2. Reporte formado con las Publicaciones de entrada: Figura 9.6
   item Clusters o grupos de Publicaciones formados:
   - Iteración 1. Salida: Figura 9.29
   - Iteración 2. Salida: Figura 9.30
   - Iteración 3. Salida: Figura 9.31
3. Resultados:
   - Reporte: Esperado
   - Clusters o grupos de Publicaciones: Esperado
   - Salida del modelo: Aceptable

Nota:

- *Respecto a los Reportes:*
  - *Puntos o nodos: Representación de una Publicación y el label es un identificador asignado antes de aplicar el algoritmo*
  - *Número entero en conexiones: Cantidad de palabras en común entre Publicaciones. Se muestra dos veces al ser un grafo no dirigido (Igual cantidad de palabras al ser evaluado desde cualquier parte de la conexión)*
  - *Cada caso fue iterado 3 veces. Se publica únicamente 1 imagen ya que el resultado en todas las iteraciones fue idéntico.*
- *Respecto a los clusters o grupos de Publicaciones:*
  - *Puntos o nodos: Representación de una Publicación y el label es un identificador asignado antes de aplicar el algoritmo*
  - *Aunque la visualización resultante contiene un plano con coordenadas, los ejes no tendrán etiquetas ni significado ya que solo se utilizan para poder visualizar el conjunto de puntos que representan las publicaciones*

9.5. Resumen

De acuerdo a los criterios definidos en Sección 7.7, se calificaron las etapas de las pruebas y la salida del modelo usando la siguiente escala:

- Esperado: La salida es la esperada y contiene información correcta. Puede servir como entrada a un siguiente proceso (dependiendo del caso).
- Aceptable: La salida no es la esperada pero puede contener información correcta. el flujo de trabajo puede continuar.
- No Esperado: La salida no es la esperada o no hay salida. Bloquea el flujo de trabajo.

9.5.1. Etapa I: Procesamiento Digital de Documentos

Prueba	Procesamiento Digital de Documentos
Caso 1	Esperado
Caso 2	No Esperado
Caso 3	No Esperado

Cuadro 9.1: Resumen Etapa I: Datos reales

Prueba	Procesamiento Digital de Documentos
Caso 1	Esperado
Caso 2	Esperado
Caso 3	Esperado

Cuadro 9.2: Resumen Etapa I: Datos genéricos

9.5.2. Etapa II: Flujo de Trabajo Completo

	Modelo			
Prueba	**Procesamiento Digital de Documentos**	**Reportes**	**Clasificación y Agrupamiento**	**Salida**
Prueba 1: 3 Publicaciones	Esperado	Esperado	Aceptable	No Esperado
Prueba 2: 4 Publicaciones	Esperado	Esperado	Esperado	Aceptable
Prueba 3: 6 Publicaciones	Esperado	No Esperado	No Esperado	No Esperado

Cuadro 9.3: Resumen Etapa II: Datos reales

	Modelo			
Prueba	**Procesamiento Digital de Documentos**	**Reportes**	**Clasificación y Agrupamiento**	**Salida**
Prueba 1: 3 Publicaciones	Esperado	Esperado	Aceptable	No Esperado
Prueba 2: 4 Publicaciones	Esperado	Esperado	Esperado	Aceptable
Prueba 3: 8 Publicaciones	Esperado	Esperado	Esperado	Aceptable

Cuadro 9.4: Resumen Etapa II: Datos genéricos

Figura 9.1: Relación entre 3 Publicaciones - Prueba Sección 9.4.1

# Palabras Clave	Más representativas		Menos Representativas	
403	Palabra	Frecuencia	Palabra	Frecuencia
	root	28	zarat	1
	cultur	20	form	1
	scopolamin	18	format	1
	candida	15	fourth	1
	brugmasia	14	gallego	1

Cuadro 9.5: Etapa I: Datos reales. Salida Documento [3]

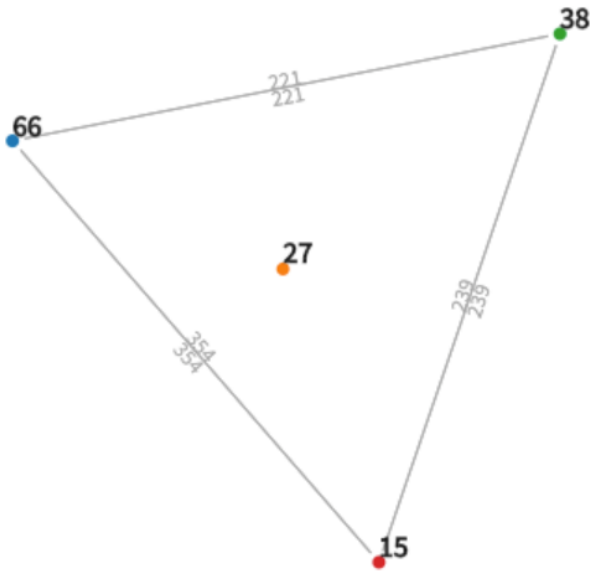

Figura 9.2: Relación entre 4 Publicaciones - Prueba Elemento 9.4.1

# Palabras Clave	Más representativas		Menos Representativas	
403	Palabra	Frecuencia	Palabra	Frecuencia
	de	252	optima	1
	la	108	hazekamp	1
	en	83	hasta	1
	lo	83	hairi	1
	se	69	haemantina	1

Cuadro 9.6: Etapa I: Datos reales. Salida Documento [4]

# Palabras Clave	Más representativas		Menos Representativas	
633	Palabra	Frecuencia	Palabra	Frecuencia
	transact	70	miss	1
	block	65	minut	1
	node	38	compet	1
	hash	33	minimum	1
	chain	27	abil	1

Cuadro 9.7: Etapa I: Datos genéricos. Salida Documento [5]

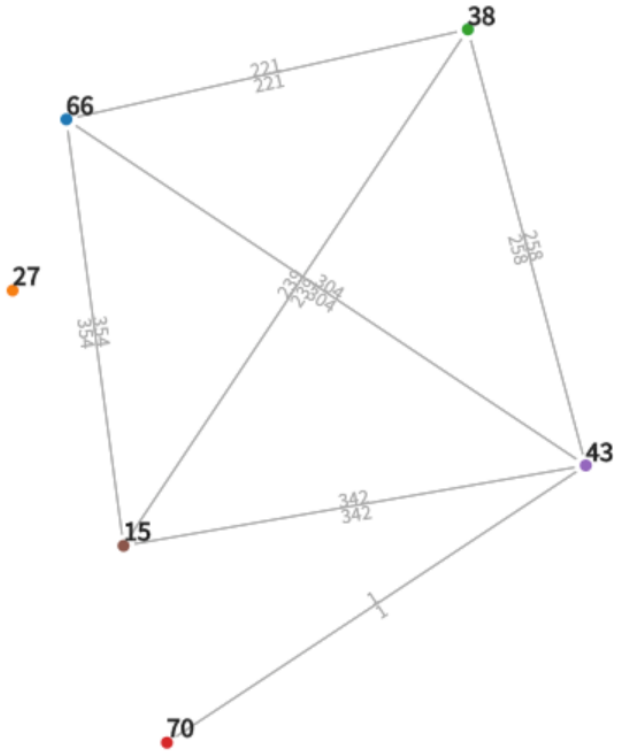

Figura 9.3: Relación entre 6 Publicaciones - Prueba Elemento 9.4.1

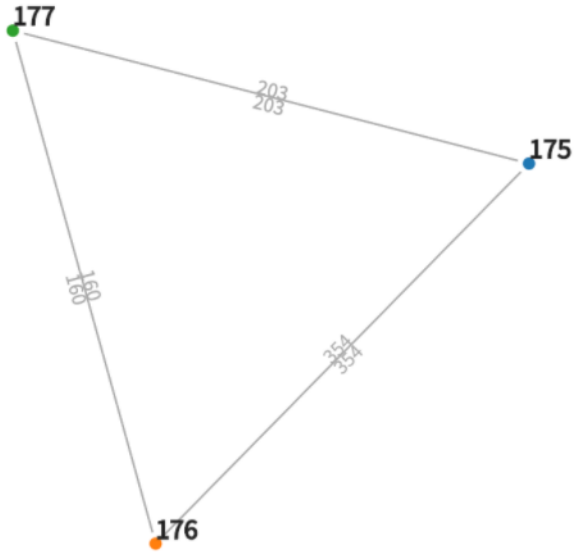

Figura 9.4: Relación entre 3 Publicaciones - Prueba Sección 9.4.1

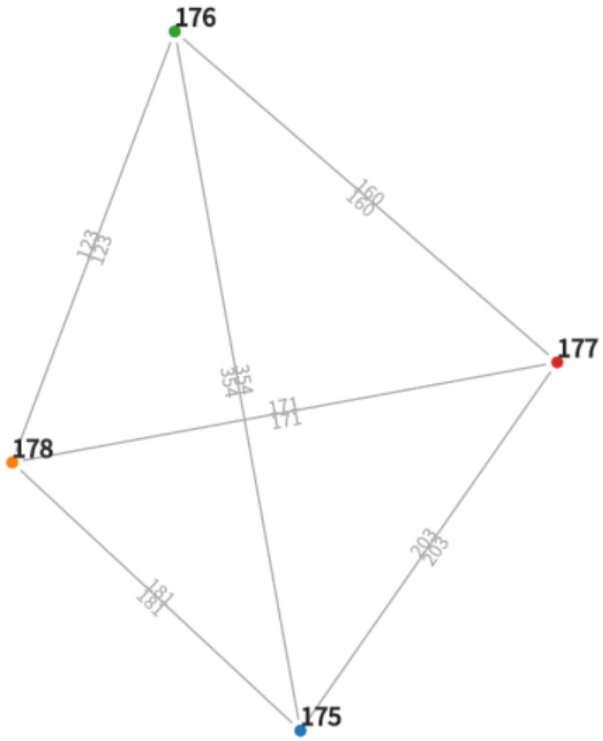

Figura 9.5: Relación entre 4 Publicaciones - Prueba Elemento 9.4.1

# Palabras Clave	Más representativas		Menos Representativas	
633	Palabra	Frecuencia	Palabra	Frecuencia
	cluster	52	duda	1
	algorithm	44	dube	1
	point	32	organ	1
	focal	28	doi	1
	j	21	kzzz	1

Cuadro 9.8: Etapa I: Datos genéricos. Salida Documento [6]

# Palabras Clave	Más representativas		Menos Representativas	
633	Palabra	Frecuencia	Palabra	Frecuencia
	infect	20	h	1
	coronaviru	13	grow	1
	health	10	ground	1
	china	9	griffin	1
	epidem	9	zhou	1

Cuadro 9.9: Etapa I: Datos genéricos. Salida Documento [7]

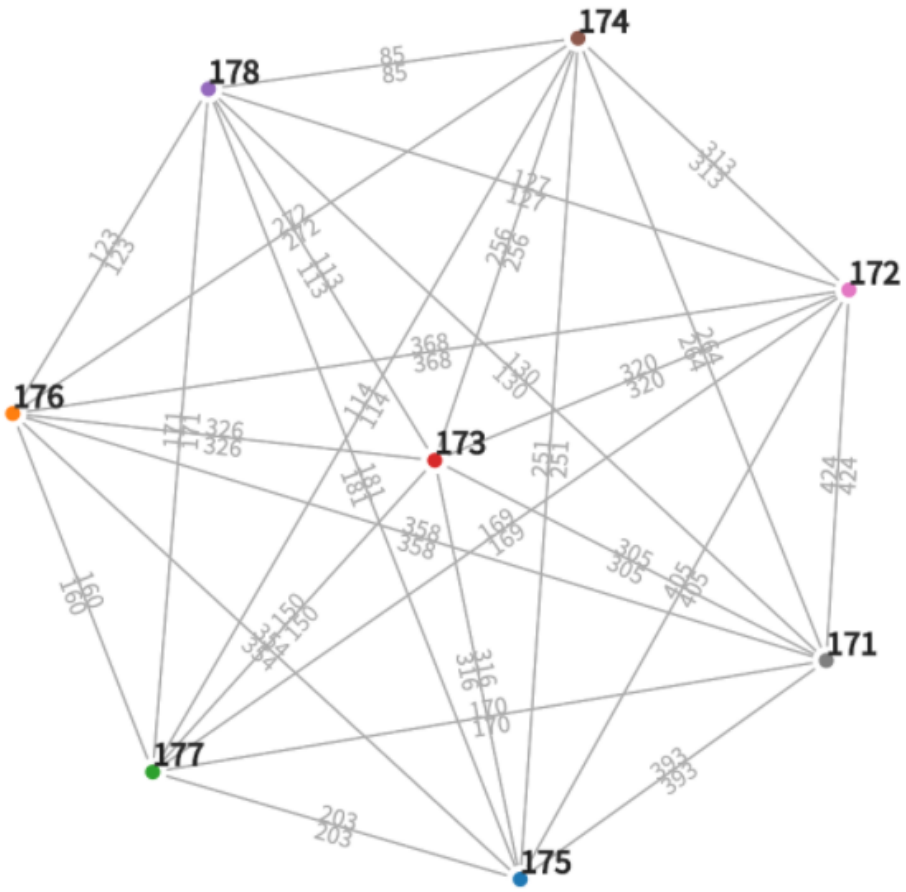

Figura 9.6: Relación entre 6 Publicaciones - Prueba Elemento 9.4.1

id	Título
66	Antibacterial and antifungal activities of crude plant extracts from Colombian biodiversity [31]
15	Antioxidant and antitopoisomerase activities in plant extracts of some Colombian flora from La Marcada Natural Regional Park [32]
27	Antimycotic activity of 20 plants from colombian flora

Cuadro 9.10: Etapa II: Datos reales. Lista de Publicaciones del Reporte Sección 9.4.1

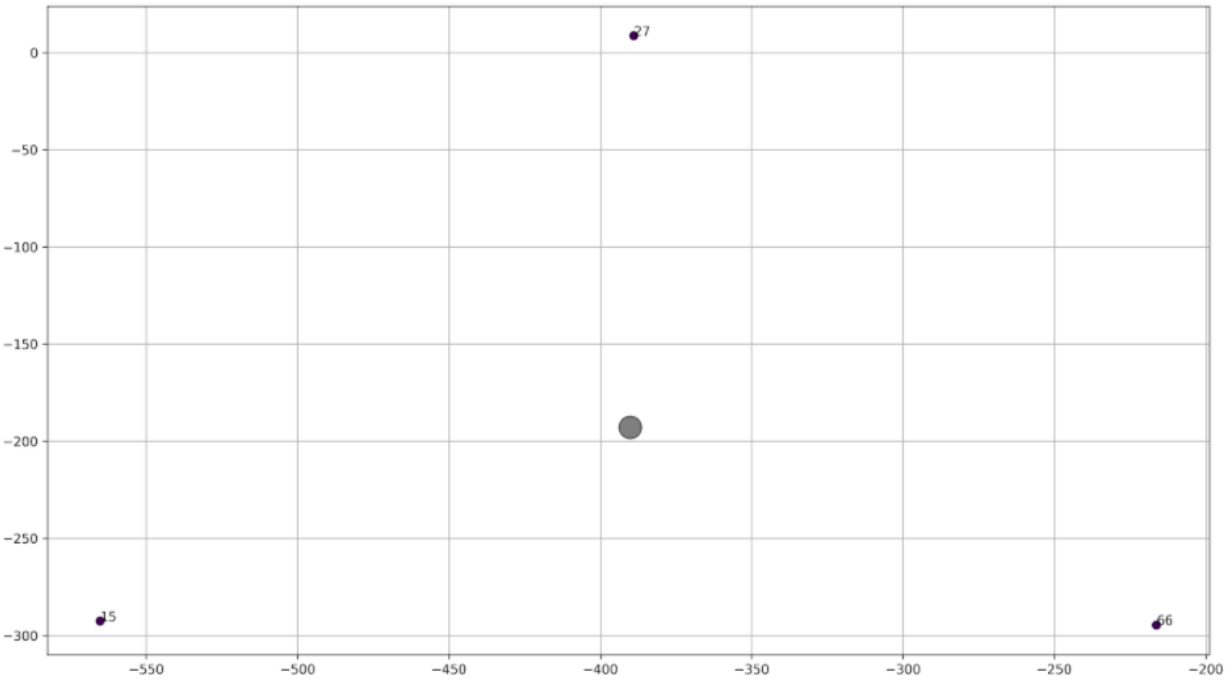

Figura 9.7: Clasificación con 3 publicaciones (Datos reales), iteración 1 - Prueba Sección 9.4.1

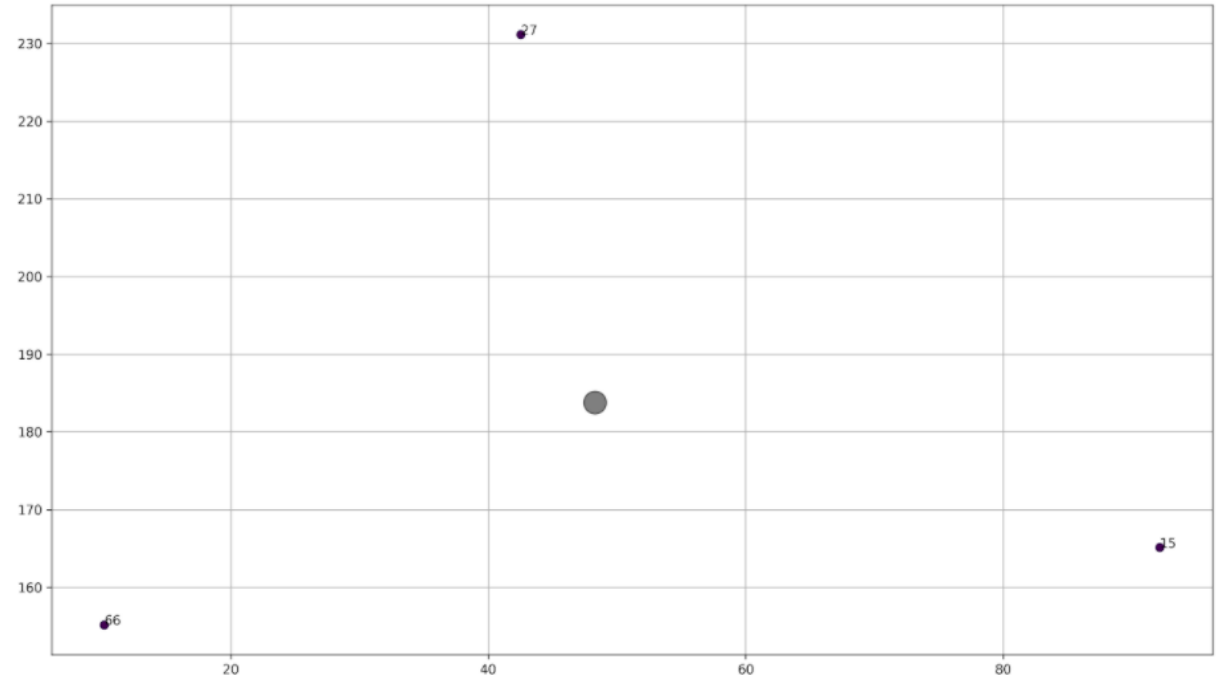

Figura 9.8: Clasificación con 3 publicaciones (Datos reales), iteración 2 - Prueba Sección 9.4.1

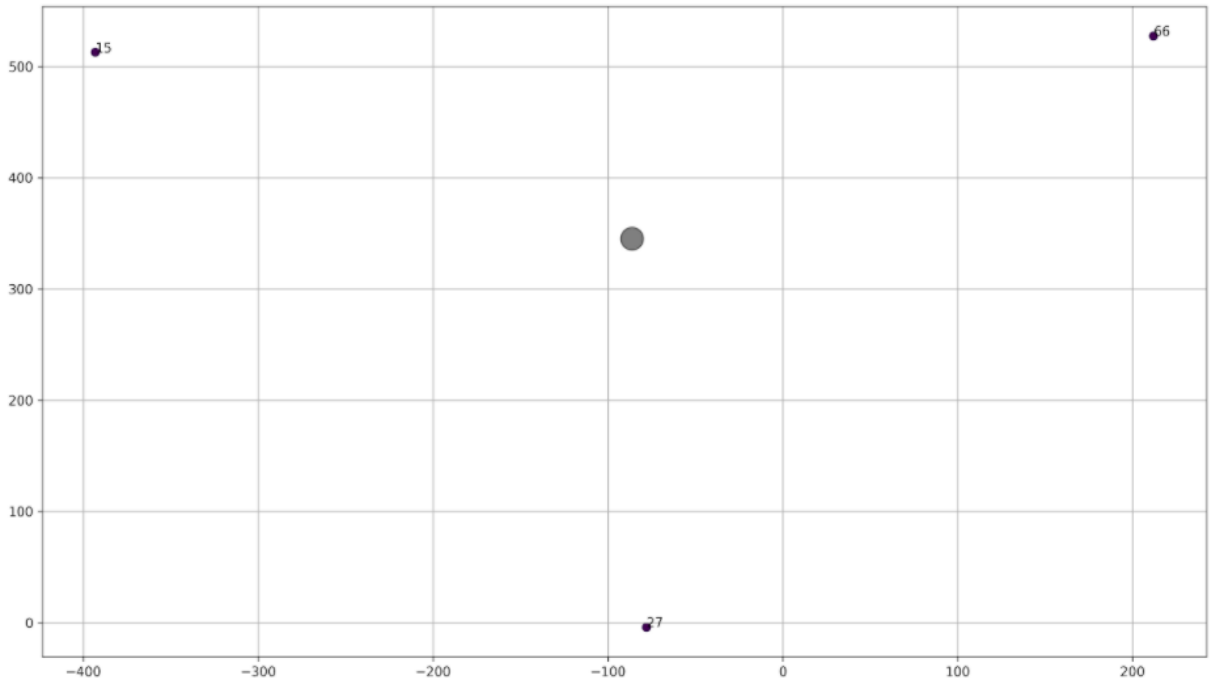

Figura 9.9: Clasificación con 3 publicaciones (Datos reales), iteración 3 - Prueba Sección 9.4.1

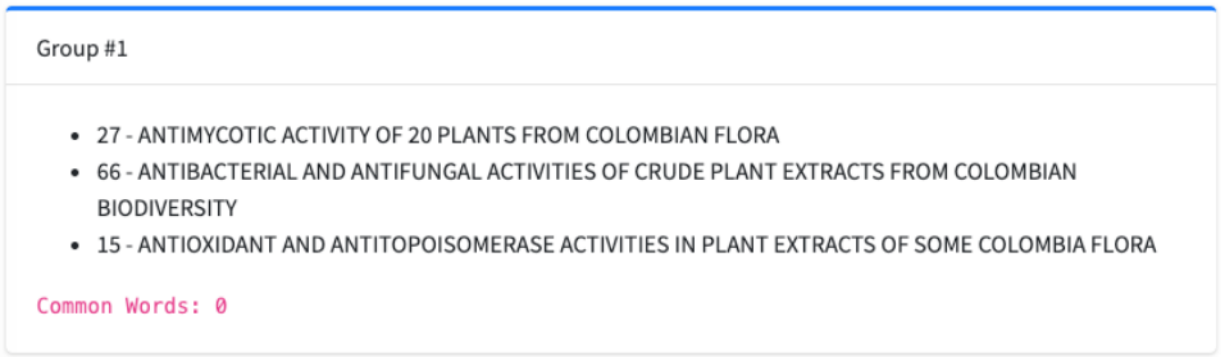

Figura 9.10: Clasificación con 3 publicaciones (Datos reales) - Prueba Sección 9.4.1

id	Título
66	Antibacterial and antifungal activities of crude plant extracts from Colombian biodiversity [31]
15	Antioxidant and antitopoisomerase activities in plant extracts of some Colombian flora from La Marcada Natural Regional Park [32]
27	Antimycotic activity of 20 plants from colombian flora
38	DNA interaction of plant extracts from Colombian flora [33]

Cuadro 9.11: Etapa II: Datos reales. Lista de Publicaciones del Reporte Elemento 9.4.1

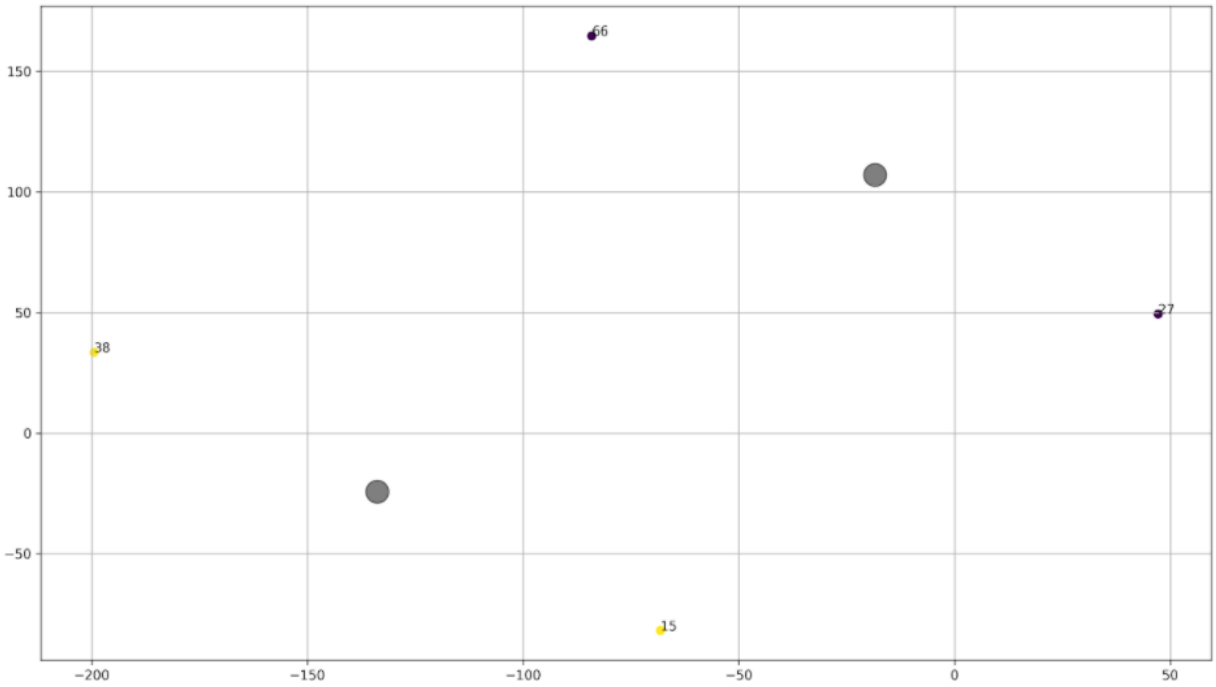

Figura 9.11: Clasificación con 4 publicaciones (Datos reales), iteración 1 - Prueba Elemento 9.4.1

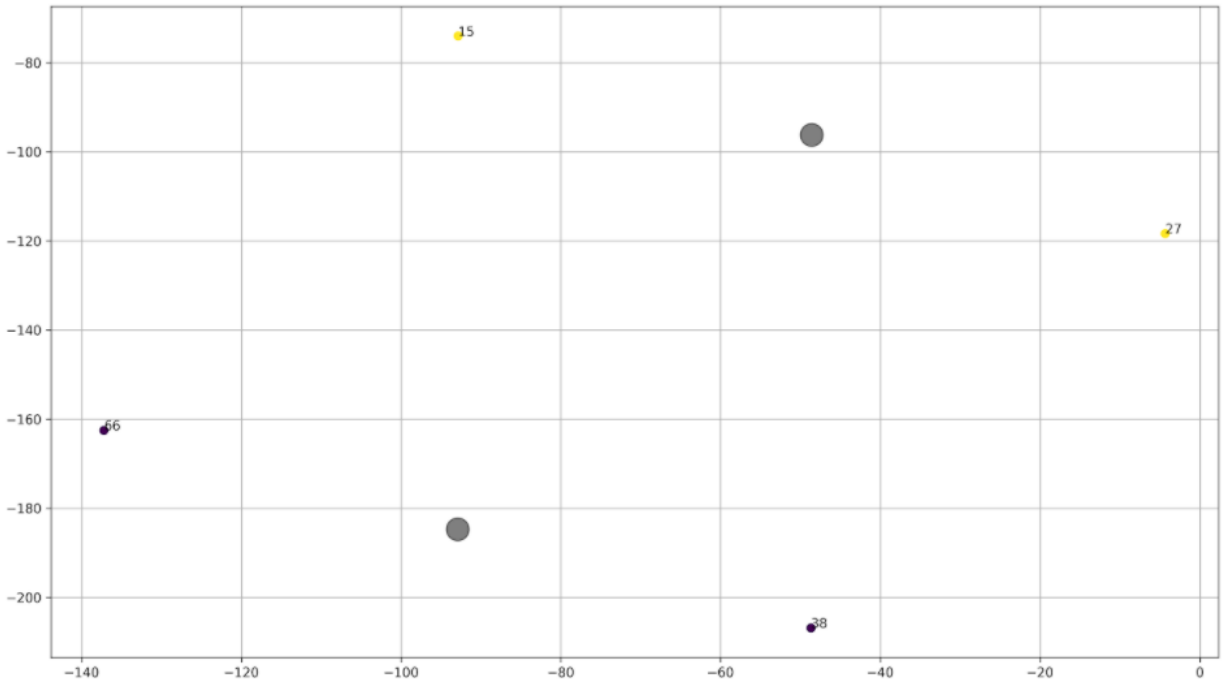

Figura 9.12: Clasificación con 4 publicaciones (Datos reales), iteración 2 - Prueba Elemento 9.4.1

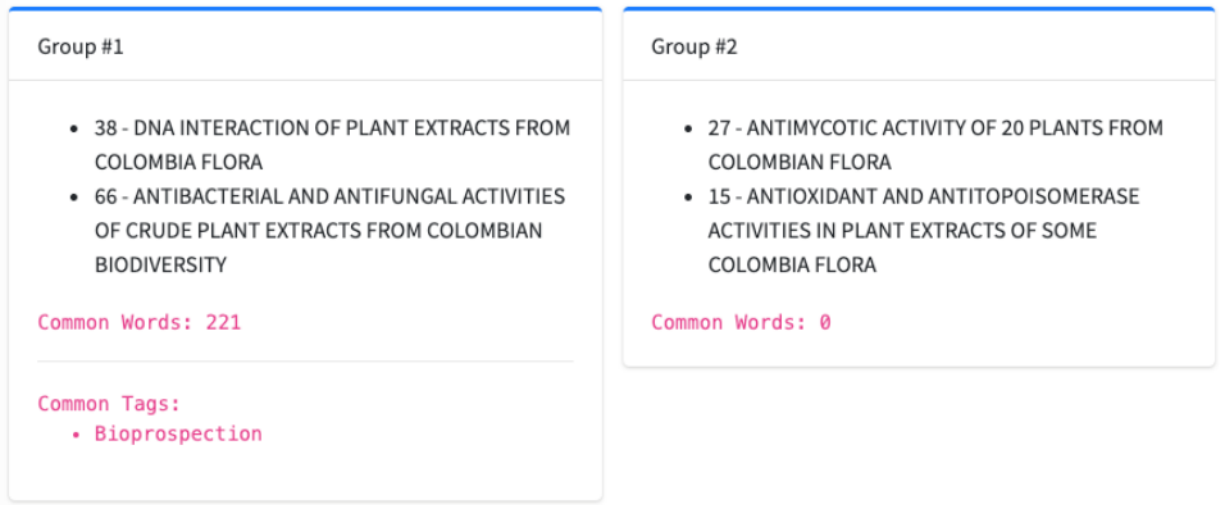

Figura 9.13: Clasificación con 4 publicaciones (Datos reales), iteración 3 - Prueba Elemento 9.4.1

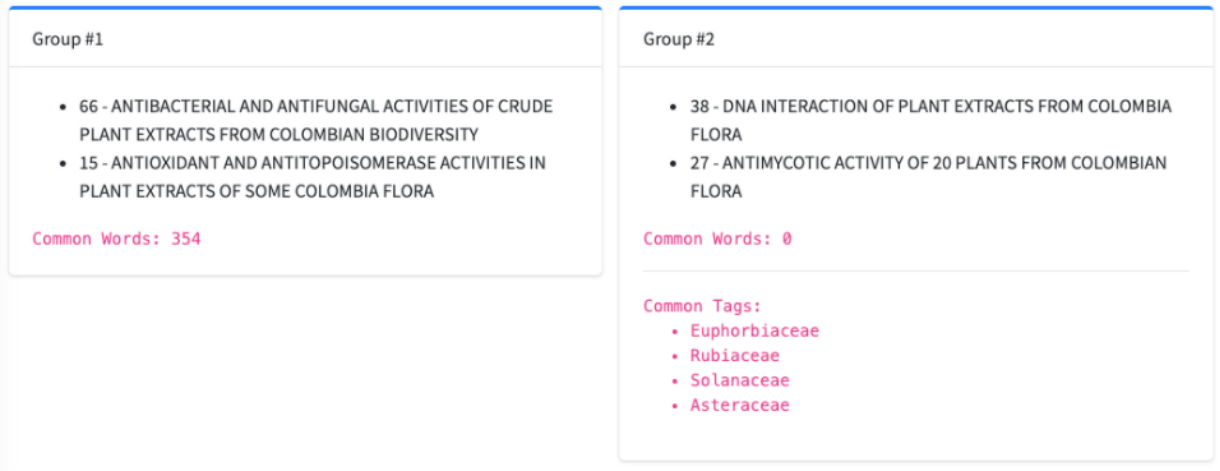

Figura 9.14: Clasificación con 4 publicaciones (Datos reales), Salida 1 - Prueba Elemento 9.4.1

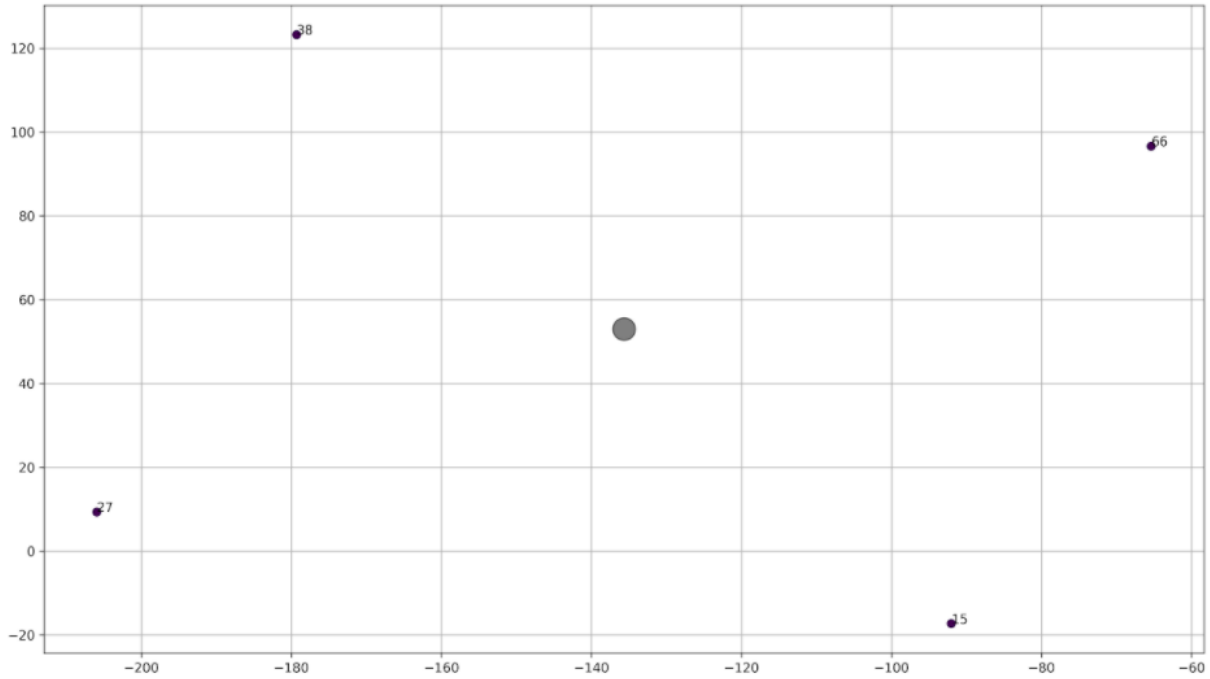

Figura 9.15: Clasificación con 4 publicaciones (Datos reales), Salida 2 - Prueba Elemento 9.4.1

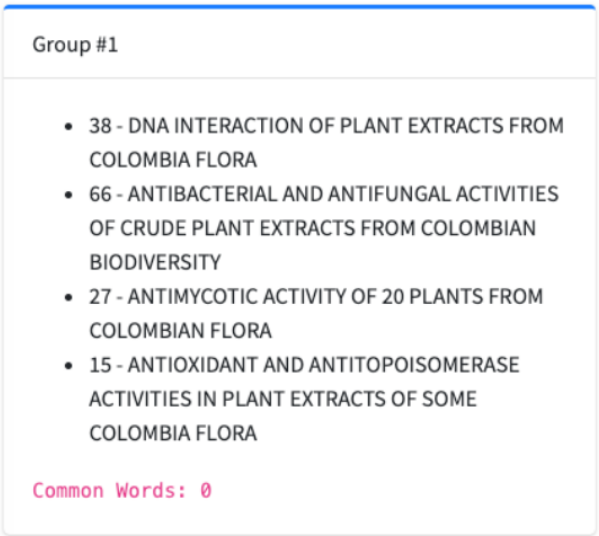

Figura 9.16: Clasificación con 4 publicaciones (Datos reales), Salida 3 - Prueba Elemento 9.4.1

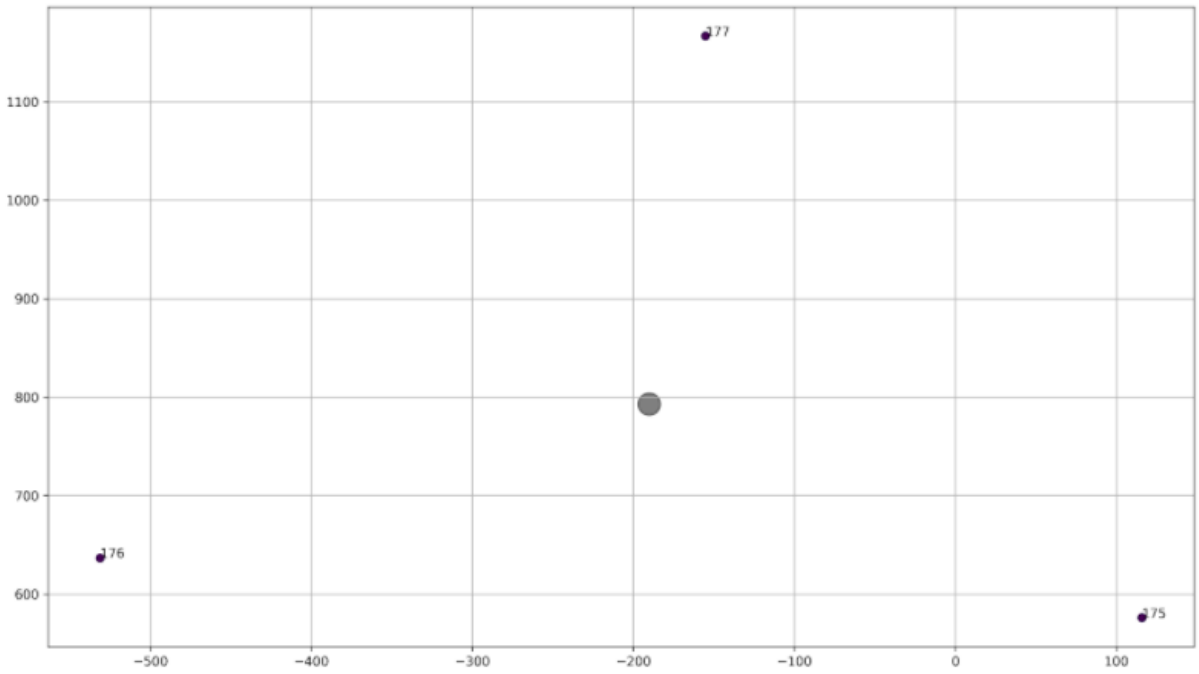

Figura 9.17: Clasificación con 3 publicaciones (Datos genéricos), iteración 1 - Prueba Elemento 9.4.1

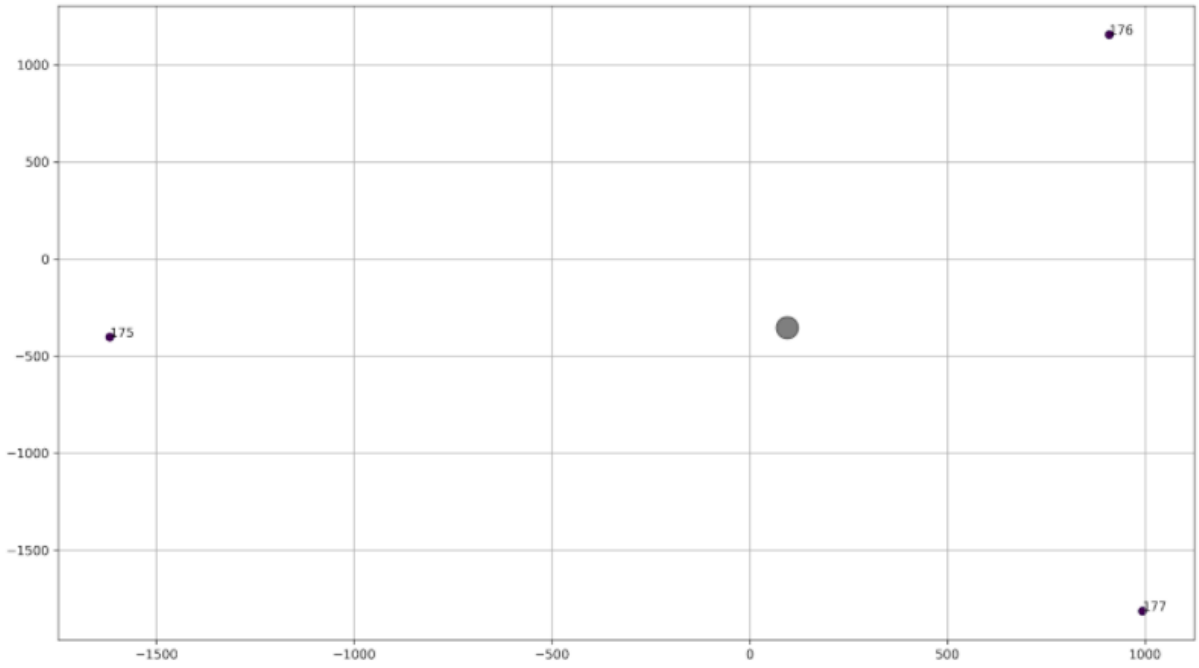

Figura 9.18: Clasificación con 3 publicaciones (Datos genéricos), iteración 2 - Prueba Elemento 9.4.1

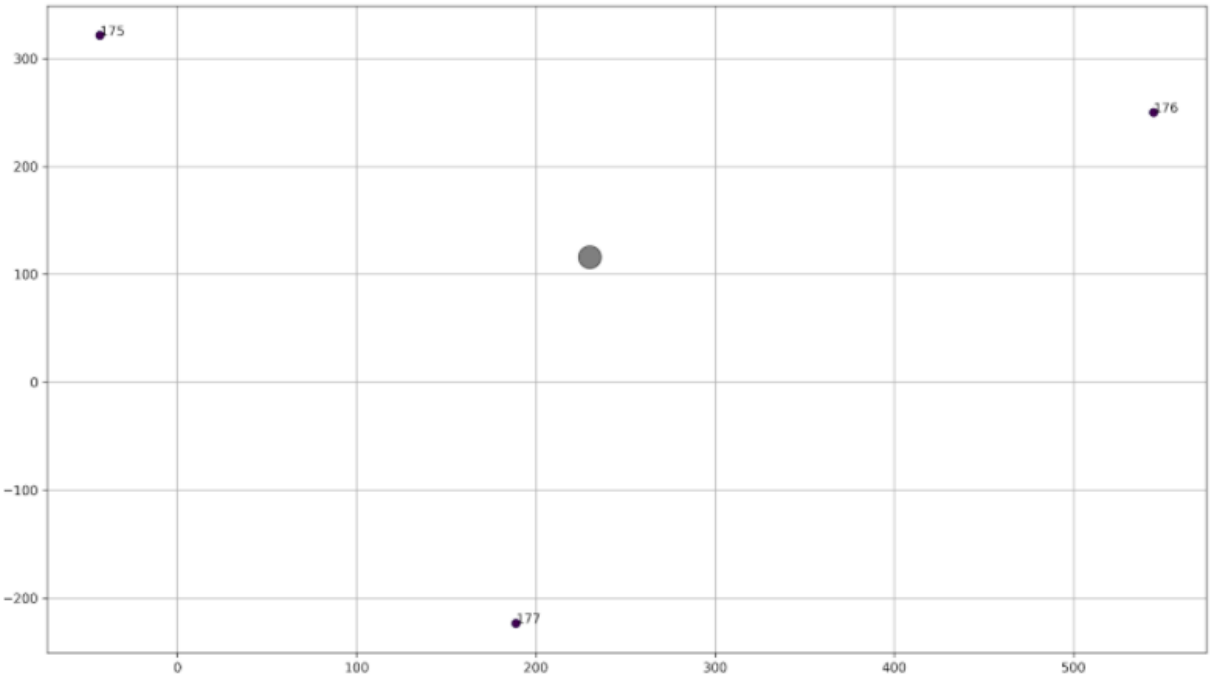

Figura 9.19: Clasificación con 3 publicaciones (Datos genéricos), iteración 3 - Prueba Elemento 9.4.1

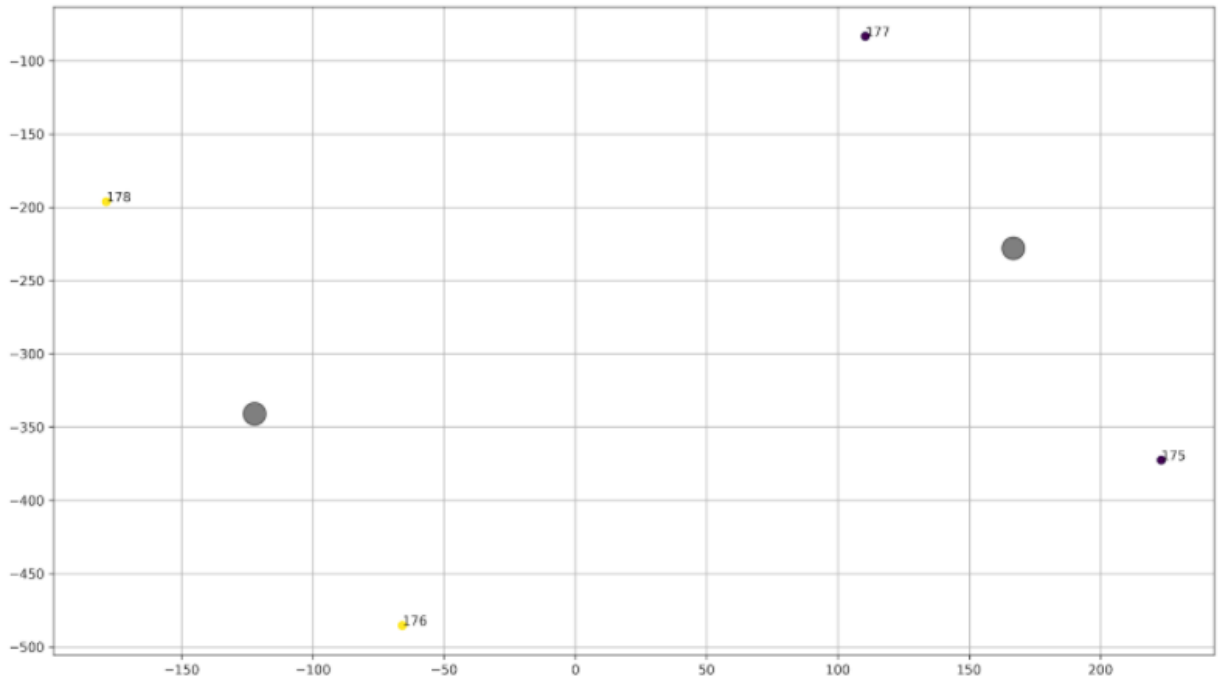

Figura 9.20: Clasificación con 4 publicaciones (Datos genéricos), iteración 1 - Prueba Elemento 9.4.1

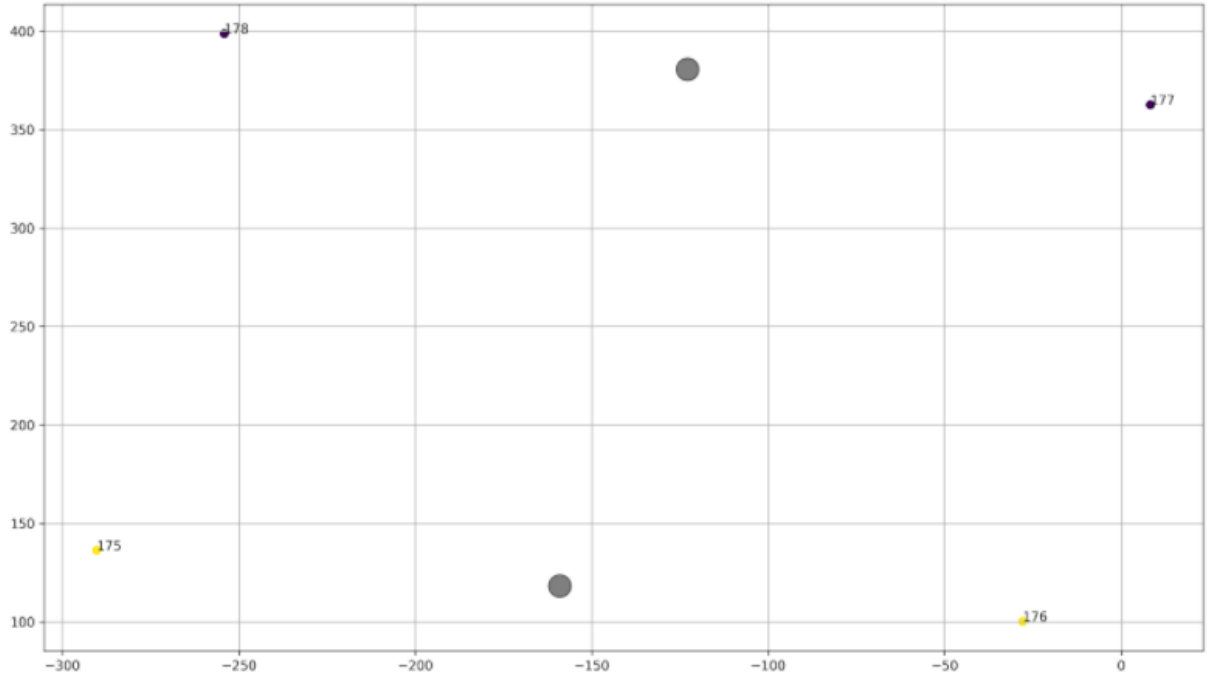

Figura 9.21: Clasificación con 4 publicaciones (Datos genéricos), iteración 2 - Prueba Elemento 9.4.1

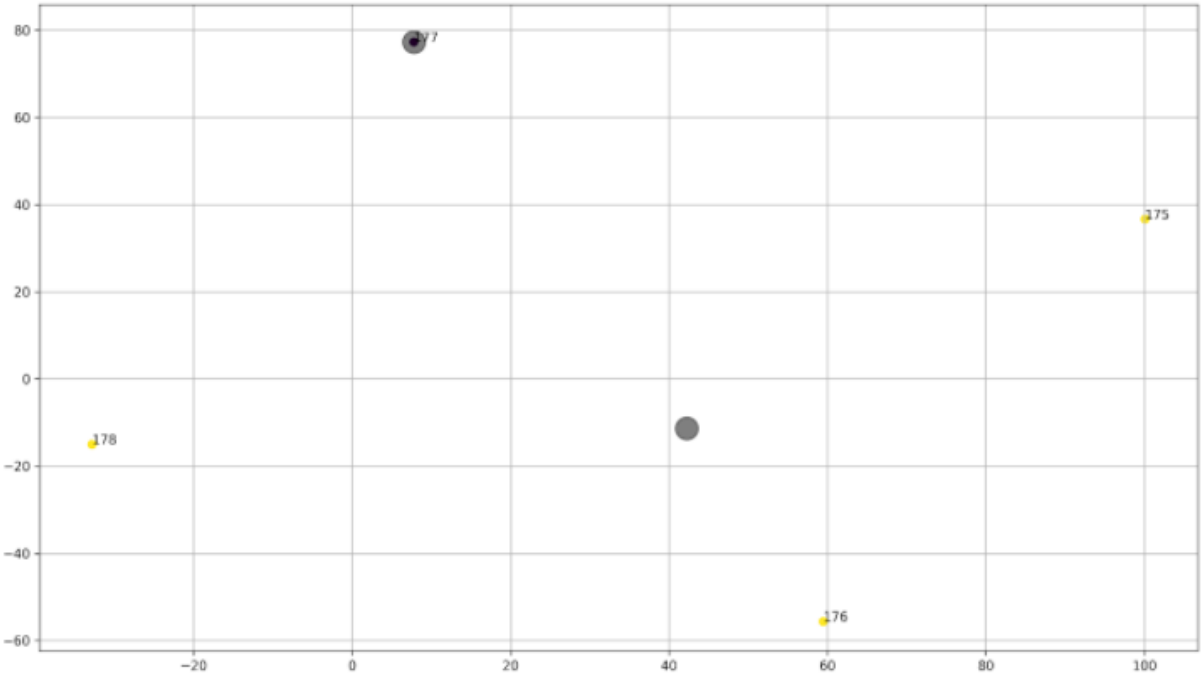

Figura 9.22: Clasificación con 4 publicaciones (Datos genéricos), iteración 3 - Prueba Elemento 9.4.1

Figura 9.23: Clasificación con 4 publicaciones (Datos genéricos), Salida 1 - Prueba Elemento 9.4.1

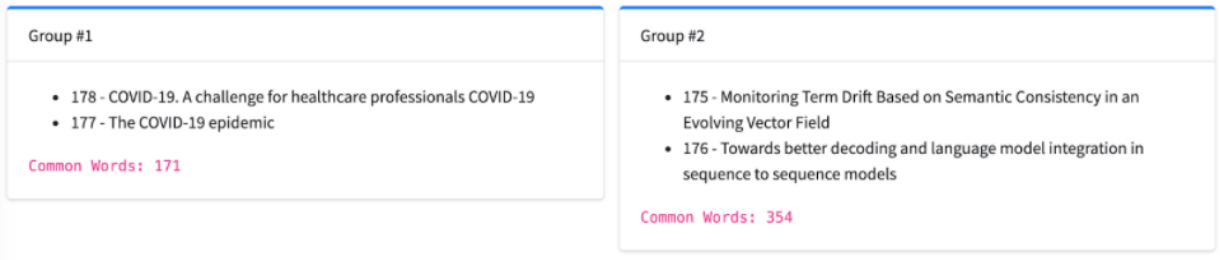

Figura 9.24: Clasificación con 4 publicaciones (Datos genéricos), Salida 2 - Prueba Elemento 9.4.1

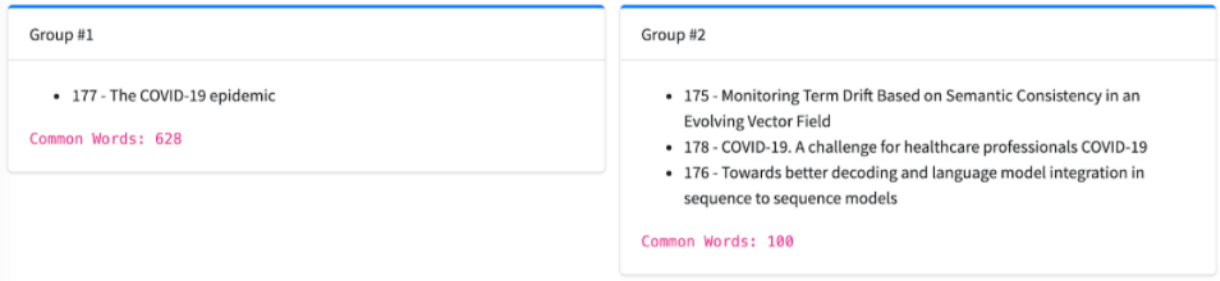

Figura 9.25: Clasificación con 4 publicaciones (Datos genéricos), Salida 3 - Prueba Elemento 9.4.1

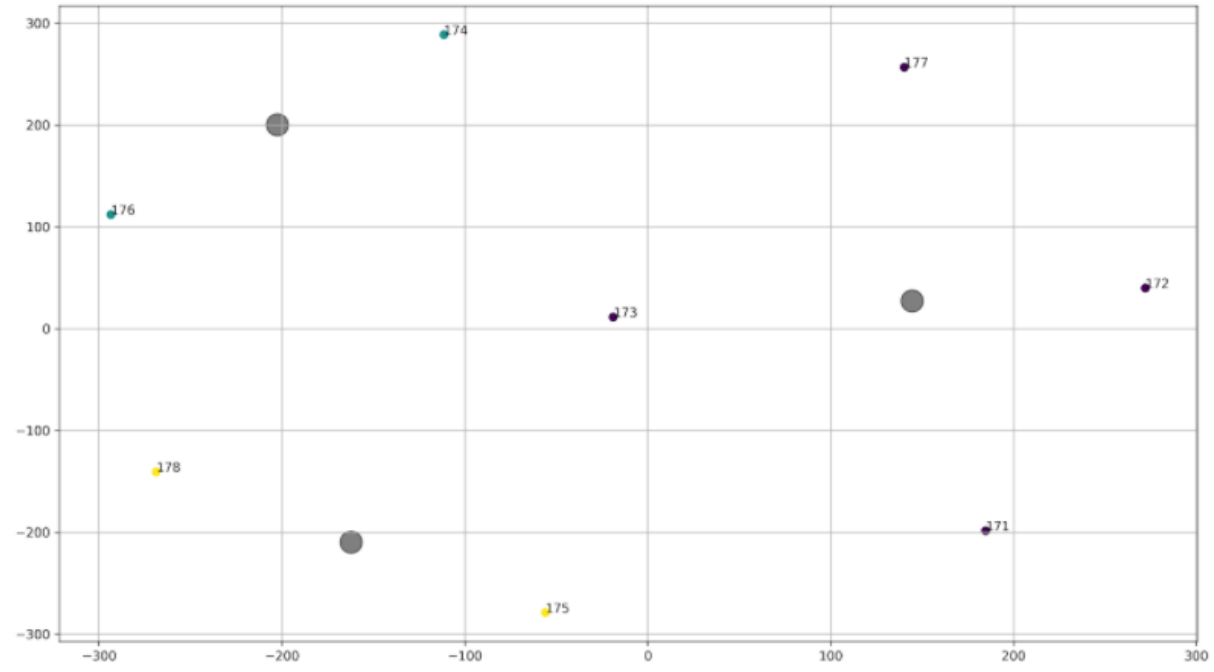

Figura 9.26: Clasificación con 8 publicaciones (Datos genéricos), iteración 1 - Prueba Elemento 9.4.1

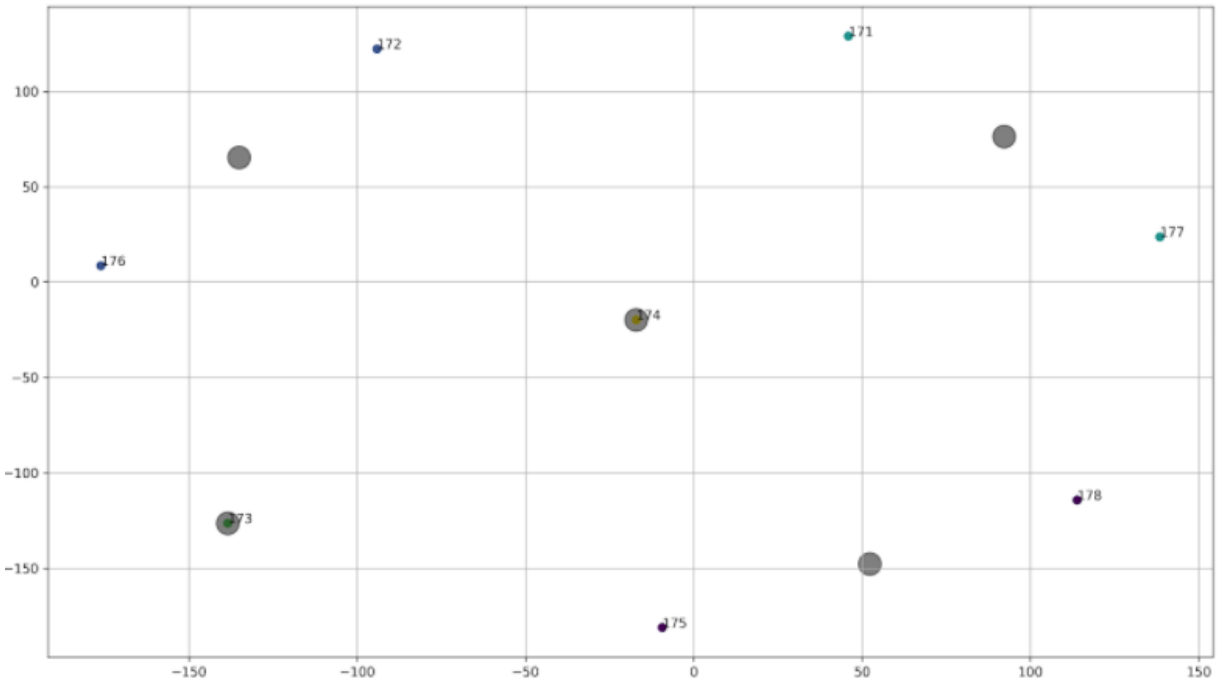

Figura 9.27: Clasificación con 8 publicaciones (Datos genéricos), iteración 2 - Prueba Elemento 9.4.1

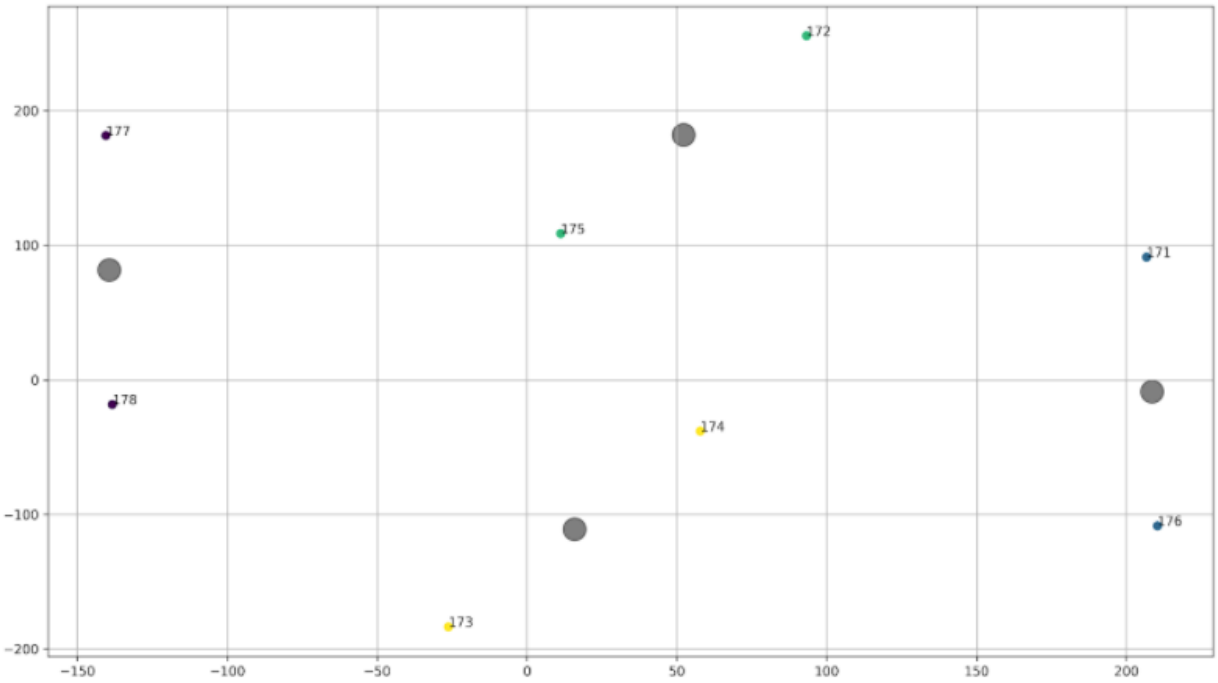

Figura 9.28: Clasificación con 8 publicaciones (Datos genéricos), iteración 3 - Prueba Elemento 9.4.1

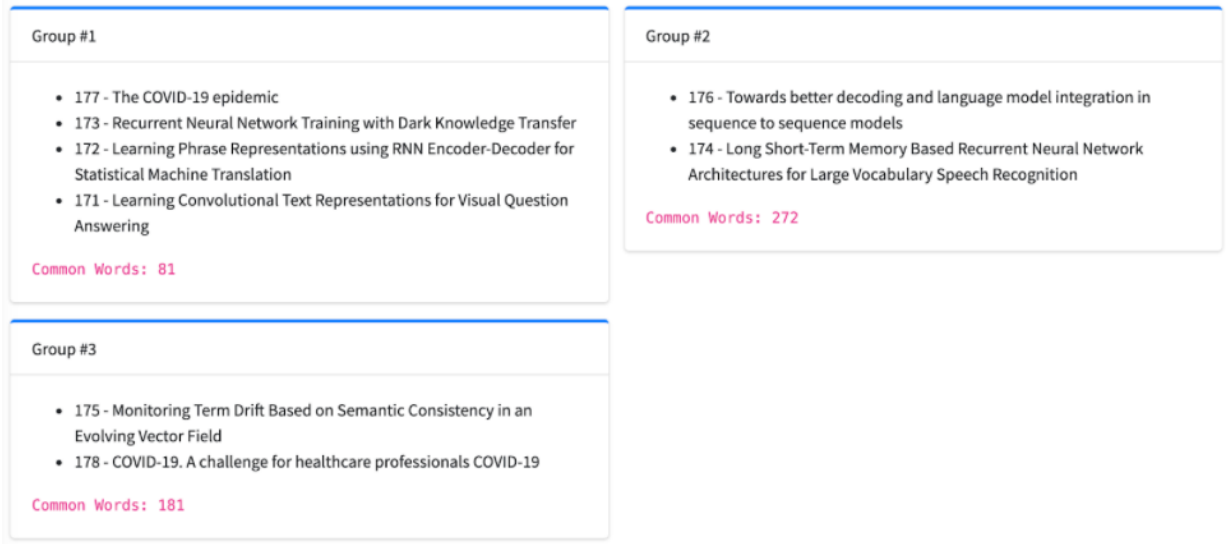

Figura 9.29: Clasificación con 8 publicaciones (Datos genéricos), Salida 1 - Prueba Elemento 9.4.1

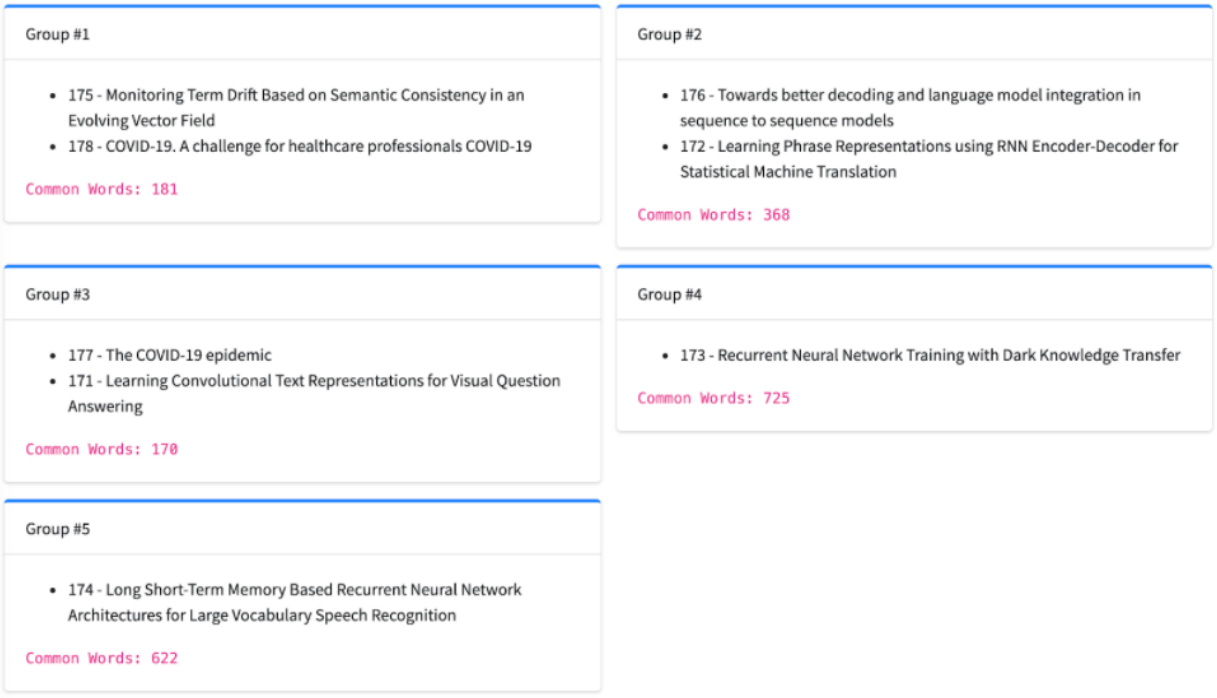

Figura 9.30: Clasificación con 8 publicaciones (Datos genéricos), Salida 2 - Prueba Elemento 9.4.1

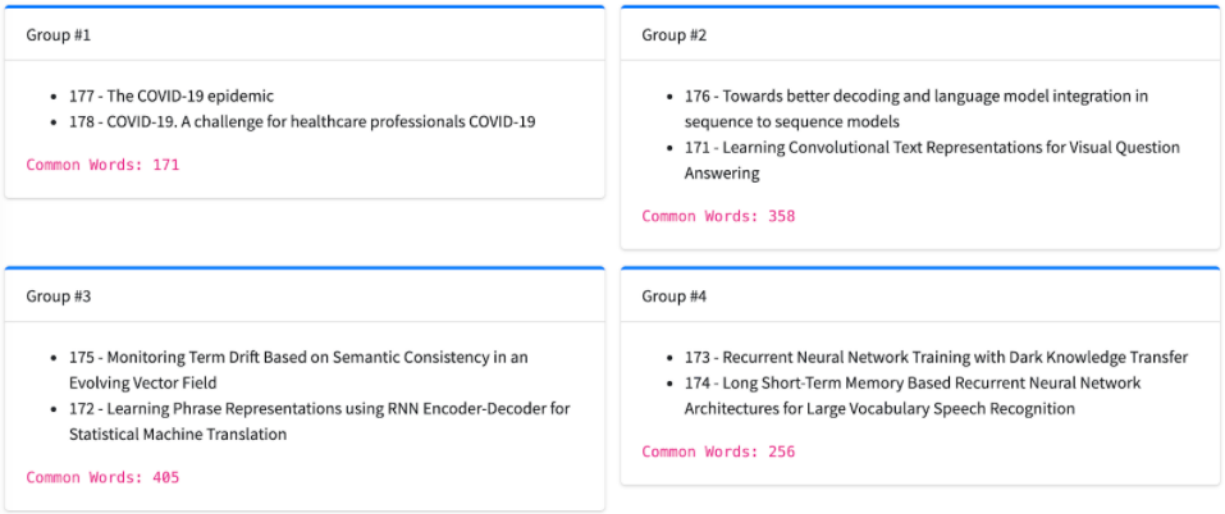

Figura 9.31: Clasificación con 8 publicaciones (Datos genéricos), Salida 3 - Prueba Elemento 9.4.1

id	Título
66	Antibacterial and antifungal activities of crude plant extracts from Colombian biodiversity [31]
15	Antioxidant and antitopoisomerase activities in plant extracts of some Colombian flora from La Marcada Natural Regional Park [32]
27	Antimycotic activity of 20 plants from colombian flora
38	DNA interaction of plant extracts from Colombian flora [33]
43	Biological activities of steroidal alkaloids isolated from solanum leucocarpum [34]
70	Bioprospección de hongos endofíticos asociados a especies del genero piper (piperaceae) del depto

Cuadro 9.12: Etapa II: Datos reales. Lista de Publicaciones del Reporte Elemento 9.4.1

id	Título
175	Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field [35]
176	Towards better decoding and language model integration in sequence to sequence models [36]
177	The COVID-19 epidemic [7]

Cuadro 9.13: Etapa II: Datos genéricos. Lista de Publicaciones del Reporte Sección 9.4.1

id	Título
175	Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field [35]
176	Towards better decoding and language model integration in sequence to sequence models [36]
177	The COVID-19 epidemic [7]
178	COVID-19. A challenge for healthcare professionals COVID-19 [37]

Cuadro 9.14: Etapa II: Datos genéricos. Lista de Publicaciones del Reporte Elemento 9.4.1

id	Título
175	Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field [35]
176	Towards better decoding and language model integration in sequence to sequence models [36]
177	The COVID-19 epidemic [7]
178	COVID-19. A challenge for healthcare professionals COVID-19 [37]
173	Recurrent Neural Network Training with Dark Knowledge Transfer [38]
171	Learning Convolutional Text Representations for Visual Question Answering [39]
172	Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [40]
174	Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition [18]

Cuadro 9.15: Etapa II: Datos genéricos. Lista de Publicaciones del Reporte Elemento 9.4.1

Capítulo 10

Análisis de Resultados

10.1. Etapa I: Procesamiento Digital de Documentos

- Algunas pruebas arrojaron resultados nulos ya que el contenido de los documentos eran fotos y el algoritmo aplicado solo utiliza metadata con caracteres digitales. Prueba referencia: Sección 9.3.1, entrada [30]
- Es de esperarse que al procesar un documento en idiomas como el español, no tenga un resultado significativo considerando que el corpus o diccionario de palabras utilizado en la implementación es del idioma inglés. Prueba referencia: Sección 9.3.1, entrada [4].
- Si no se encuentran palabras claves en un documento:
  - El diccionario de palabras utilizado en el documento no corresponde a un idioma relacionado
  - El contenido del documento no coincide con los parámetros establecidos para el procesamiento
  - Se puede considerar el archivo como vacío y no debería ser válido para continuar el procesamiento en las etapas siguientes para no afectar el resultado final
- El resultado es el esperado (Subsección 9.5.1): en cada uno de los casos, el número de palabras y su respectiva frecuencia son iguales después de cada iteración (3 iteraciones propuestas).

10.2. Etapa II: Flujo de Trabajo Completo

10.2.1. Reportes

- Si en el grafo final existe un nodo aislado (sin conexiones) y es una Publicación que no tiene palabras claves como resultado de la Etapa I del procesamiento, se

puede considerar el archivo como vacío y no debería ser válido para continuar el procesamiento en las etapas siguientes para no afectar el resultado final del análisis.

- El resultado es aceptable (Subsección 9.5.2): en cada uno de los casos, puede existir un *Grafo* que represente el reporte y contiene la misma estructura tras cada iteración (siempre que la etapa anterior arroje los resultados esperados).

10.2.2. Clasificación y Agrupamiento

- En la mayoría de los casos, las palabras comunes entre los grupos no son representativas de las Publicaciones.
- Si en el grafo final existe un nodo aislado (sin conexiones), no debería incluirse en grupos
- En el caso de que no existen palabras claves en común, las Publicaciones no deberían pertenecer al mismo grupo ya que no están relacionadas.
- El resultado no es el esperado (Subsección 9.5.2): en cada uno de los casos, el número de grupos/clusters y las publicaciones que lo conforman, varían tras cada iteración.

Capítulo 11

Conclusiones

- Las pruebas sobre la primera parte del modelo (Procesamiento Digital de Documentos) fueron exitosas, lo cual comprueba que el algoritmo de procesamiento de documentos funciona según lo planteado. Se debe considerar que su acierto depende de las librerías / métodos que se utilicen para la extracción de texto y del conjunto de palabras de un lenguaje (e.g. idioma Inglés) que se utilicen para identificar las palabras claves en el texto.
- Los resultados de las pruebas de la segunda parte del modelo (Reportes) fueron aceptables, sin embargo el modelo del reporte podría mejorar si se ajusta el algoritmo para procesar el contenido de los documentos ya que no se descartan todas palabras no claves y se toman en cuenta en la implementación (como conectores y sílabas sin significado).
- Los algoritmos utilizados en la tercera parte del modelo (Clasificación y Agrupamiento) son no deterministas, por esta razón, la distribución de las publicaciones y el número de grupos/clusters puede variar.
- Aunque analizar el contenido de los documentos brinda información importante para los algoritmos, es bueno considerar otros tipos de fuentes de datos del mismo documento que podrían ser valiosos al momento de hacer recomendaciones, *e.g. título, abstract, etc.*
- Las principales recomendaciones podrían ser basadas en Figura 8.10:
  - **Palabras clave en común (Tags):** Los tópicos que aparecen en esta sección al final del procesamiento, representan los temas comunes entre las Publicaciones y pueden significar que son un fuerte en el grupo de investigación. Se puede recomendar seguir trabajando en futuras Publicaciones.
  - **Palabras en común (Contenido):** Dependiendo de la frecuencia de las palabras
    - Alta frecuencia: Al igual que los tags, representan temas que se dominan y pueden significar que son un fuerte en el grupo de investigación. Se puede recomendar seguir trabajando en futuras Publicaciones.

    - Baja frecuencia: En su mayoría son palabras o temas no relacionados al dominio del grupo de investigación pero a su vez pueden mostrar algunos temas que no se han tratado y que pueden traer mejoras o descubrimientos. Su principal objetivo es mostrar posibilidades y opciones de diferentes temas que se pueden investigar.

- El conocimiento y la experiencia obtenida de este proyecto son base para futuras implementaciones que pueden hacer extensivas estas técnicas a otros problemas de clasificación e interpretación de textos, lo cual desembocó en el desarrollo de un prototipo web (Apéndice C) que integra todas las anotaciones poniéndolas a disposición de un usuario para que pueda efectuar consultas y visualizar los reportes resultantes de las mismas mediante una interfaz apropiada. El prototipo utiliza la herramientas web para el fácil manejo e implementación del mismo en cualquier servidor.

Capítulo 12

Trabajos Futuros

12.1. Procesamiento Digital de Documentos

- Implementar mediciones diferentes al número de palabras claves del texto. Sugerencias:
  - Palabras clave en los artículos
  - Extracción de palabras del abstract
  - Análisis de los títulos
- Obtener un Corpus o diccionario de palabras más robusto (que contenga palabras representativas y comunes al contexto de los textos que se van a probar).
- Habilitar la opción de manejar distintos idiomas en la evaluación de textos (De momento las pruebas se realizan utilizando un conjunto de palabras del idioma inglés)
- Implementación con diferentes librerías especializadas en el Procesamiento de lenguaje natural:
  - **Actual:** NLTK
  - **Recomendado:** Spacy
- Implementación con diferentes herramientas/librerías para la extracción de contenido (texto) en documentos digitales
  - **Actual** Apache Tika
  - **Opcional**: textract

12.2. Reportes

Respecto a la construcción del Reportes (Grafo de Reporte): Mejorar la medida de relación entre Publicaciones (Peso entre las conexiones de los nodos)

Problema

La medida actual de conexión entre Publicaciones (Peso) está dada por el número de palabras claves que tengan en común. Como se expresa en las conclusiones, el número de palabras claves es una medida poco significativa ya que puede contener palabras que no son representativas (conectores, sílabas, etc).

Propuesta de posibles soluciones

Usar una medida porcentual que represente la relación entre ambos nodos/Publicaciones, en donde:

- *100 %: Las publicaciones son equivalentes/Iguales*
- *0 %: Las publicaciones son diferentes - No están conectadas*

En los algoritmos usados en el desarrollo del proyecto, se recomienda usar valores pequeños para los pesos de los nodos en los grafos para facilitar la operación de transformación de grafo a puntos en dimensiones euclidianas. Tomando lo anterior en cuenta, se propone una medida entre intervalos, e.g. [0, 1]:

- *100 % - 1: Las publicaciones son equivalentes/Iguales*
- *0 % - 0: Las publicaciones son diferentes - No están conectadas*

Nota: la convención de la medida se puede adaptar para mejor entendimiento de la relación

12.2.1. Caso 1: Usando únicamente el número de palabras en común

$$publicacion\ a = 150\ palabras,\ publicacion\ b = 70\ palabras$$
$$total = 220\ palabras$$

Si las publicaciones tienen 30 palabras en común:

$$\frac{30\ palabras\ en\ comun * 100\,\%}{220\ palabras\ (total)} = \frac{3000}{220}\,\% = 13,63\,\% = 0,13$$

12.2.2. Caso 2: Usando el número de palabras en común y la frecuencia de repetición

Se considera una variante: la frecuencia de las palabras en común puede no ser la misma en las publicaciones.

$$publicacion\ a = 150\ palabras,\ publicacion\ b = 70\ palabras$$
$$total = 220\ palabras$$

Si las publicaciones tienen 30 palabras en común. Es probable que las palabras en común no tengan frecuencias similares:

$$palabra\ x\ en\ publicacion\ a = 5\ repeticiones,$$
$$palabra\ x\ en\ publicacion\ b = 30\ repeticiones$$

Se debe definir una medida de relación considerando este factor. Algunas ideas pueden ser:

Promedio de la cantidad de frecuencias

$$total\ frecuencia = frecuencia\ palabra\ x\ en\ publicacion\ a +$$
$$frecuencia\ palabra\ x\ en\ publicacion\ b = 35$$
$$medida\ de\ relacion\ palabra\ x = \frac{total\ frecuencia}{2} = 17,5$$

Medida porcentual de la cantidad de frecuencias

Lo primero, es considerar cuál es la equivalencia de una sola palabra. Si usamos el método propuesto en el Caso 1:

$$\frac{1\ palabras\ en\ comun * 100\,\%}{220\ palabras\ (total)} = \frac{100}{220\,\%} = 0,4545\,\% = 0,0045$$

Se sabe la equivalencia de una sola palabra, ahora se puede proponer un ajuste a esa medida basándose en la frecuencia de repeticiones de la palabra.

$$Si\ palabra\ x\ en\ publicacion\ a = 5\ repeticiones,$$
$$palabra\ x\ en\ publicacion\ b = 30\ repeticiones$$

- El caso ideal sería que ambas publicaciones tuvieran el mismo número de repeticiones para la palabra. Esto equivale a

$$100\,\% = 0,4545\,\% = 0,0045$$

- De lo contrario, si las medidas de frecuencia son diferentes

$$maximo\ repeticiones = max(palabra\ x\ en\ publicacion\ a = 5\ repeticiones,\ palabra\ x\ en\ publicacion\ b = 30\ repeticiones)$$

$$100\,\% = maximo\ repeticiones * 2$$

Para el caso de prueba:

$$maximo\ repeticiones = 30$$

$$\frac{(palabra\ x\ en\ publicacion\ a + palabra\ x\ en\ publicacion\ b) * 100\,\%}{maximo\ repeticiones * 2} = \frac{(5+30) * 100\,\%}{60} = \frac{3500}{60}\,\% = 58,34\,\%$$

Si consideramos que esta palabra es tan solo el 0,0045 de la medida de relación entre las publicaciones:

$$58,34\,\% * 0,045100\,\% = 0,02625\,\%$$

Entonces, este valor sería la representación de esta palabra dentro de la magnitud de peso o valor de la relación entre publicaciones.

Es importante decir que cualquiera que sea el cálculo definido, debe ser aplicado sobre cada una de las palabras comunes entre publicaciones.

12.3. Clasificación y Agrupamiento

Al usar algoritmos no determinísticos, es incierto predecir la clase de resultados que vamos a obtener o asegurar cuál va a ser la mejor alternativa a incluir en la implementación de nuestra solución.

En la solución propuesta, se utiliza un algoritmo no supervisado del tipo *K-means* o *K-medias* pero existen múltiples alternativas y sus funciones son las mismas, clasificar y agrupar. Algunos tipos de algoritmos propuestos son:

- Algoritmos de Agrupación por Jerarquías
- Algoritmos de Agrupación Espacial de Aplicaciones Basadas en la Densidad con Ruido (DBSCAN)
- Modelos de Mezcla Gaussiana (MMG)

Apéndice A

Guía de Desarrollo

A continuación se brinda más información acerca del desarrollo de los objetivos (Sección 4.2) propuestos en este trabajo:

- **Obtener un conjunto de datos.**
  - Planteamiento: Subsección 7.4.1
  - Ejecución: Subsección 8.1.1
- **Determinar técnicas de Inteligencia Artificial apropiadas para el conjunto de datos seleccionado.**
  - Planteamiento: Sección 7.5 y Sección 7.6
  - Ejecución: Capítulo 8
- **Implementar las anteriores técnicas:** Capítulo 8
- **Evaluar los resultados obtenidos por el modelo formulado.**
  - Criterio de aceptación: Sección 7.7
  - Resumen de pruebas: Sección 9.5
  - Análisis: Capítulo 10

Apéndice B

Modelo Propuesto

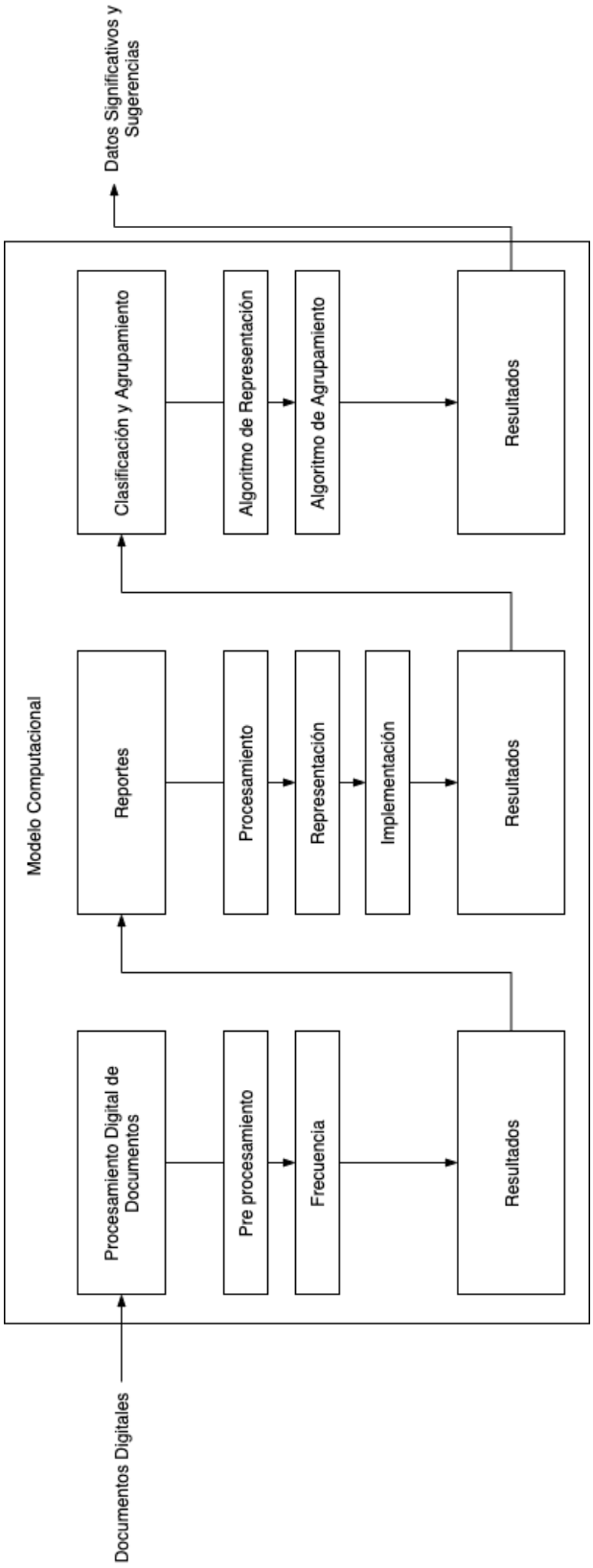

Figura B.1: Modelo Propuesto (Completo)

Apéndice C

Aplicación Web

C.1. Perspectiva del aplicativo

El sistema es un aplicativo para un grupo de investigación de la Escuela de Química de la UTP que busca gestionar toda la información de sus Papers, con el fin de proporcionarles una visión clara de sus puntos fuertes basándose en la información de los mismos.

El acceso a la información y funcionalidades se restringe de acuerdo a los perfiles que se pueden crear desde el administrador pero principalmente se destacan dos perfiles que harán uso de la plataforma, los cuales se enuncian en Tabla C.1.

C.2. Funcionalidades del proyecto

1. Interfaz de administración del proyecto
2. Inicio de sesión
3. Crear publicación
4. Crear reporte
5. Cierre de sesión
6. Detalle de publicación
7. Detalle de reporte
8. Eliminar reporte
9. Eliminar publicación
10. Buscar reporte
11. Buscar publicaciones

C.3. Restricciones

1. Los navegadores recomendados para utilizar el aplicativo son: Firefox 4.0 en adelante, Google Chrome 11.0 en adelante.

2. El motor de base de datos debe ser Relacional, para este caso se utilizó Postgres por el rendimiento que maneja para grandes cantidades de información, además de ser open source y gratuita.

3. Para la creación de un reporte es necesario seleccionar al menos dos publicaciones, ya que de esta manera se permite comparar el análisis de las dos publicaciones para poder luego sacar los cluster de información.

C.4. Requerimientos no Funcionales

C.4.1. Requisitos de Seguridad

- El sistema debe proporcionar seguridad a la información de cada publicación a través de la modificación de la política de contraseñas, donde se exija al Administrador cuando registre o cambie la contraseña de los usuarios sea de mínimo 8 caracteres, contenga mayúsculas y números.

- Política de seguridad de datos.

- El sistema debe alojarse en un proveedor de servicios de la nube que cuente con todas la garantías de seguridad para la administración de firewalls y backups de toda la información

C.4.2. Requisitos de Rendimiento

- La disponibilidad es de 7x24, puede garantizarse siempre y cuando el servidor esté en funcionamiento

- El aplicativo para poder ejecutar todas sus tareas asíncronas en el procesamiento de los documentos requiere mínimamente de la siguiente arquitectura

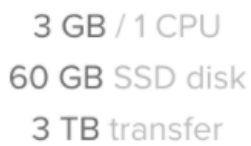

Figura C.1: Especificación de requisitos del servidor [2]

C.4.3. Restricciones de diseño

Se utilizará el Modelo de Diseño Vista Controlador, el cual es un patrón de arquitectura del software que separa los datos de la aplicación, interfaz de usuario y lógica de control en tres componentes distintos.

C.4.4. Atributos del sistema

El mecanismo de seguridad que restringe el acceso a los usuarios es por login y password. Puesto que el sistema es orientado a la web no dependerá de ningún hardware ni sistema operativo específico para su correcto funcionamiento, sólo será necesario el uso de un navegador web, lo que hará de este un sistema portable.

C.5. Diseño y arquitectura del aplicativo web

C.5.1. Diagramas de casos de uso

- Diagrama de caso de uso: Inicio de sesión (Figura C.10)
- Diagrama de caso de uso: Gestión de publicaciones (Figura C.11)
- Diagrama de caso de uso: Gestión de Reportes (Figura C.12)
- Diagrama de caso de uso: Buscar publicaciones (Figura C.13)
- Diagrama de caso de uso: Buscar reportes (Figura C.14)
- Diagrama de caso de uso: Administración de usuarios (Figura C.15)
- Diagrama de caso de uso: Cerrar sesión (Figura C.16)

C.5.2. Especificación de casos de uso

- Iniciar sesión (Tabla C.2)
- Gestión de publicaciones (Tabla C.3 y Tabla C.4)
- Gestión de reportes (Tabla C.5)
- Buscar publicaciones (Tabla C.6)
- Buscar reportes (Tabla C.7)
- Gestión de usuarios (Tabla C.8 y Tabla C.9)
- Cerrar sesión (Tabla C.10)

C.5.3. Diagrama relacional

Ver Figura C.17

Nota: el modelo entidad relación presentado anteriormente se encuentra en la tercera forma normal

C.5.4. Vistas

- Inicio de sesión: Figura C.2
- Dashboard: Figura C.3
- Agregar publicación: Figura C.4
- Lista de publicaciones: Figura C.5
- Publicación: Figura C.6
- Buscador de publicaciones: Figura C.7
- Reporte: Figura C.8
- Análisis de reporte: Figura C.9

C.6. Tecnologías

C.6.1. Python

Python es un lenguaje de programación poderoso y fácil de aprender. Cuenta con estructuras de datos eficientes y de alto nivel y un enfoque simple pero efectivo a la programación orientada a objetos. La elegante sintaxis de Python y su tipado dinámico, junto con su naturaleza interpretada, hacen de éste un lenguaje ideal para scripting y desarrollo rápido de aplicaciones en diversas áreas y sobre la mayoría de las plataformas.

El intérprete de Python y la extensa biblioteca estándar están a libre disposición en forma binaria y de código fuente para las principales plataformas desde el sitio web de Python [41], y puede distribuirse libremente. El mismo sitio contiene también distribuciones y enlaces de muchos módulos libres de Python de terceros, programas y herramientas, y documentación adicional.

El intérprete de Python puede extenderse fácilmente con nuevas funcionalidades y tipos de datos implementados en C o C++ (u otros lenguajes accesibles desde C). Python también puede usarse como un lenguaje de extensiones para aplicaciones personalizables.

La razón del uso de este lenguaje de programación para el desarrollo de la presente tesis se enfoca principalmente en las siguientes características:

- Amplia comunidad de desarrolladores de este lenguaje.
- Curva de aprendizaje
- Variedad de librerías enfocadas en el área de la Inteligencia artificial
- Fácil instalación de sus librerías.

C.6.2. Django

Django es un framework web de alto nivel que fomenta el desarrollo rápido y el diseño limpio y pragmático. Con Django se pueden obtener beneficios como:

Rapidez: Django nació en un ambiente periodístico, donde se subían noticias muy rápido, y como los desarrolladores no pudieron estar a ese ritmo decidieron crear algo que sí lo haga, y así fue como nace Django , es por eso que ha sido estructurado de tal manera que las aplicaciones web que se crearan sean muy rápidas.

DRY: "No te repitas", Django utiliza esta filosofía para no crear bloques de código iguales y fomentar la reutilización del mismo. Admin: Django es el único framework que "por defecto" viene con un sistema de administración activo, listo para ser utilizado sin ningún tipo de configuración.

ORM: Es una herramienta que te permite realizar consultas SQL a la Base de Datos, sin utilizar SQL.[42]

C.6.3. PostgreSQL

PostgreSQL es un potente sistema de base de datos relacional de objetos de código abierto que usa y amplía el lenguaje SQL combinado con muchas características que almacenan y escalan de manera segura las cargas de trabajo de datos más complicadas. Los orígenes de PostgreSQL se remontan a 1986 como parte del proyecto POSTGRES en la Universidad de California en Berkeley y tiene más de 30 años de desarrollo activo en la plataforma central.

PostgreSQL se ha ganado una sólida reputación por su arquitectura comprobada, confiabilidad, integridad de datos, conjunto de características robustas, extensibilidad y la dedicación de la comunidad de código abierto detrás del software para ofrecer soluciones innovadoras y de alto rendimiento. PostgreSQL se ejecuta en todos los principales sistemas operativos, cumple con ACID desde 2001 [43].

C.6.4. Celery

Es un sistema distribuido simple, flexible y confiable para procesar grandes cantidades de mensajes, mientras proporciona operaciones con las herramientas necesarias para mantener dicho sistema.

Es una cola de tareas centrada en el procesamiento en tiempo real, a la vez que admite la programación de tareas [44].

C.6.5. Docker

Docker es una herramienta diseñada para facilitar la creación, implementación y ejecución de aplicaciones mediante el uso de contenedores. Los contenedores permiten a un desarrollador empaquetar una aplicación con todas las partes que necesita, como bibliotecas y otras dependencias, y desplegarla como un paquete. Al hacerlo, gracias al contenedor, el desarrollador puede estar seguro de que la aplicación se ejecutará en cualquier otra máquina Linux, independientemente de cualquier configuración personalizada que pueda tener la máquina que podría diferir de la máquina utilizada para escribir y probar el código [45].

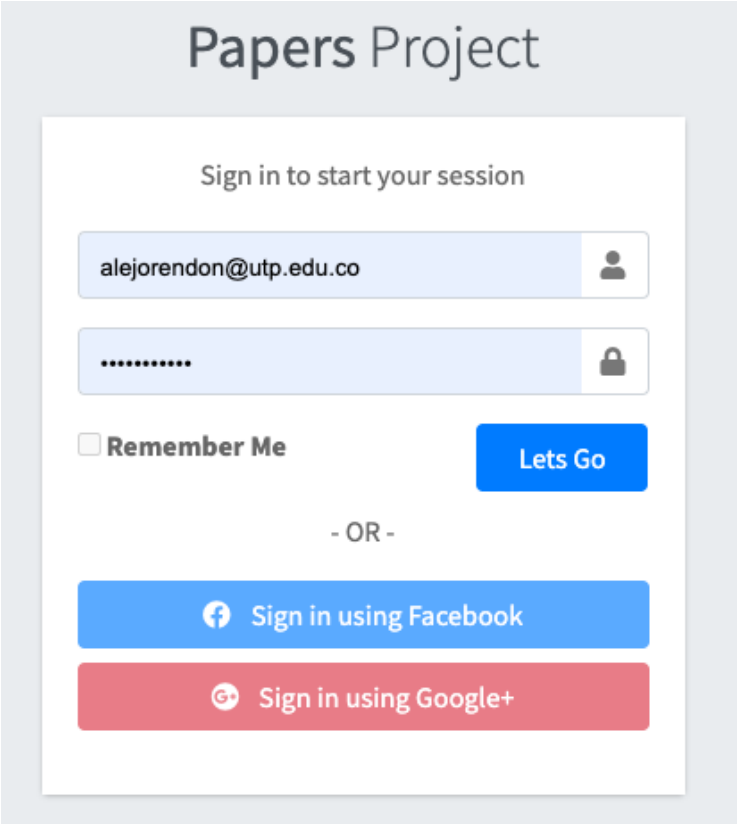

Figura C.2: Vista - Inicio de sesión

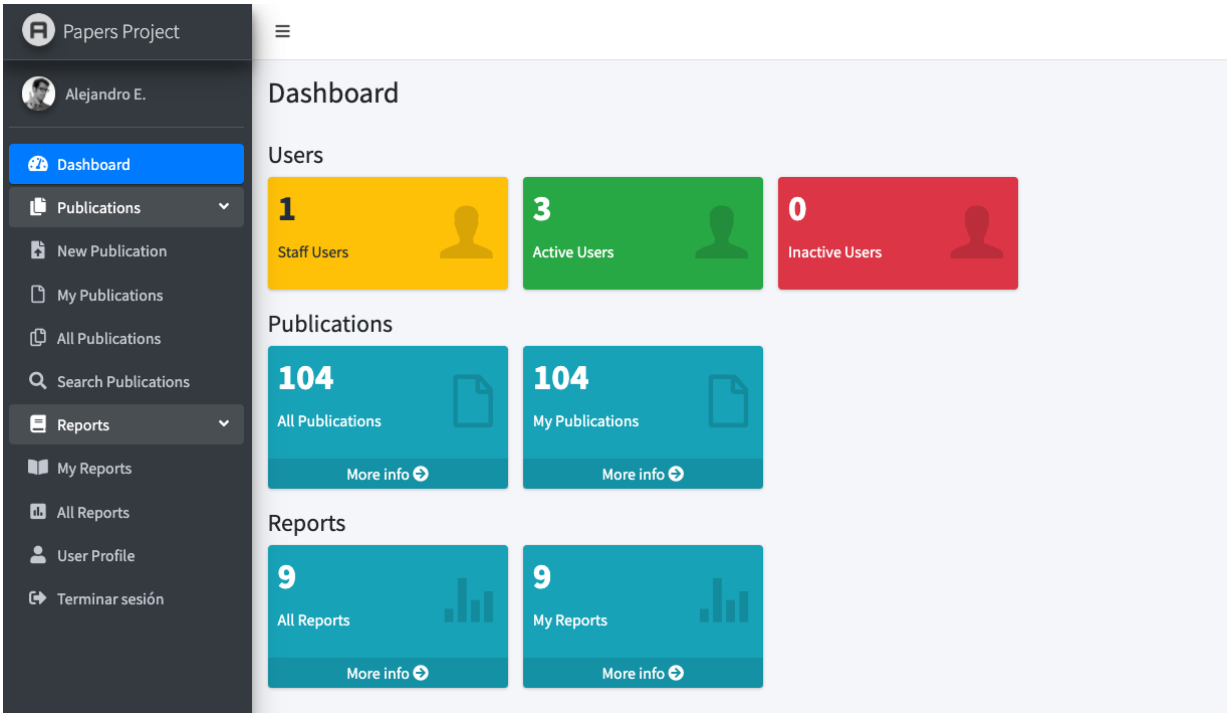

Figura C.3: Vista - Dashboard

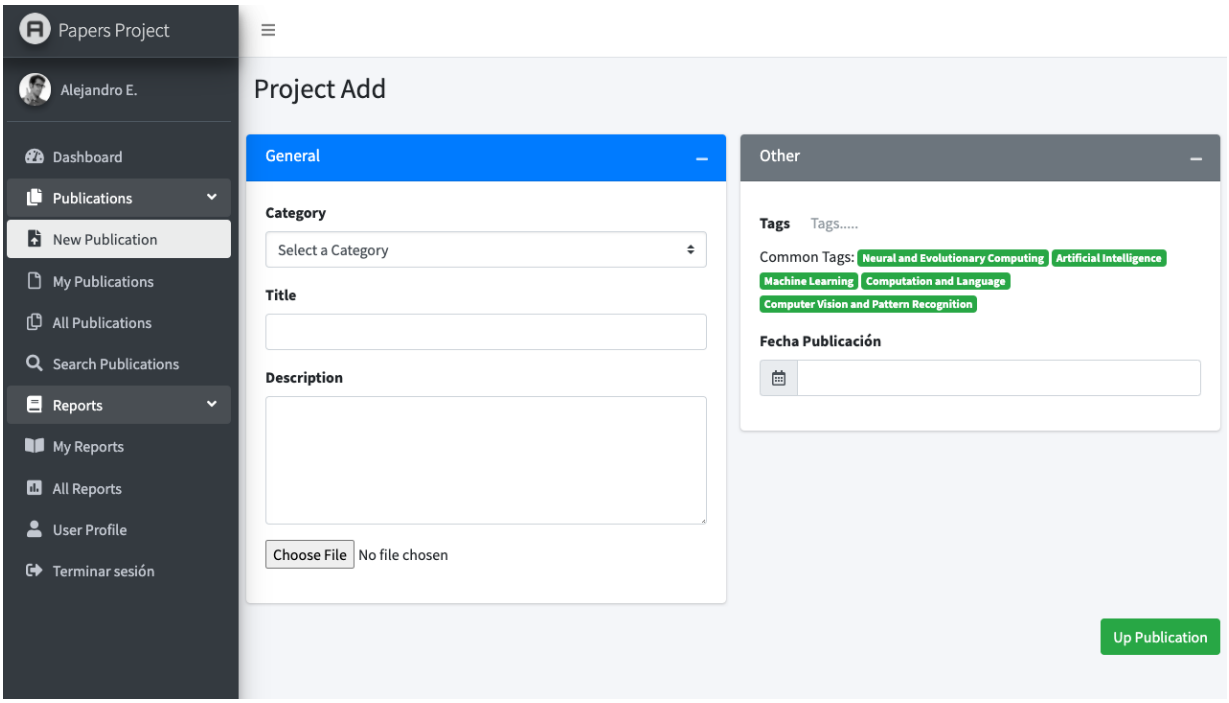

Figura C.4: Vista - Agregar Publicación

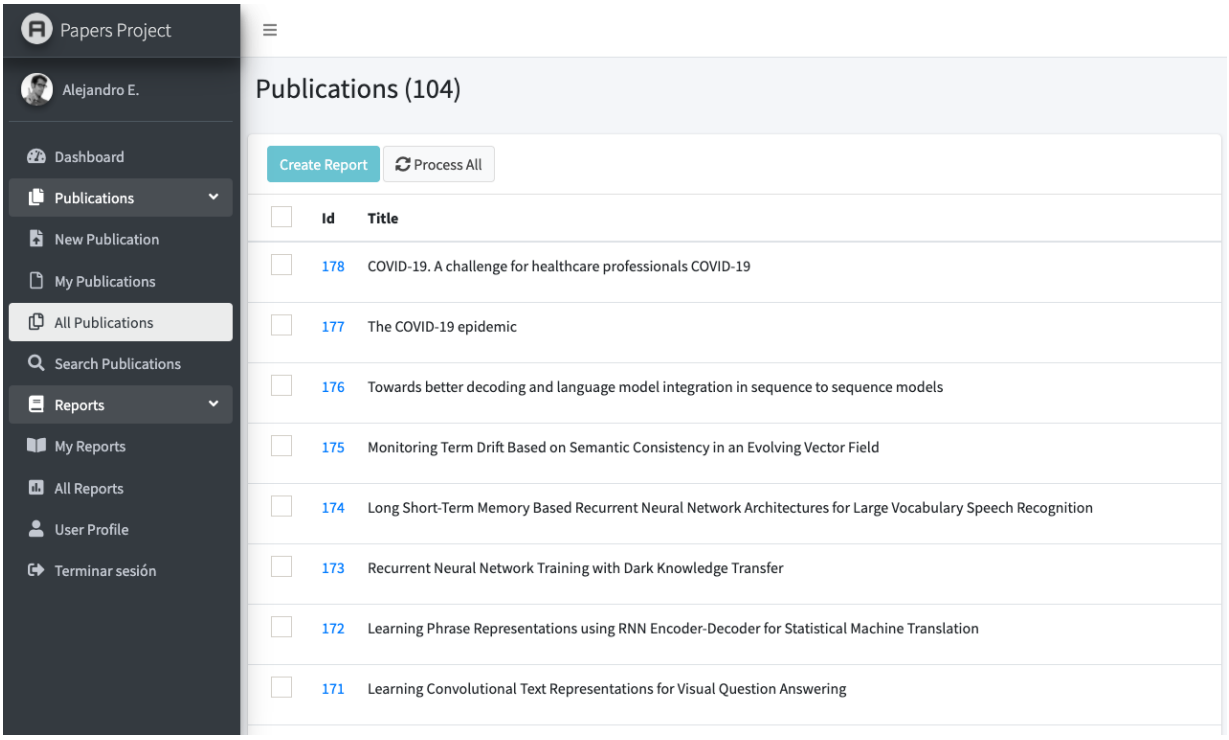

Figura C.5: Vista - Lista de Publicaciones

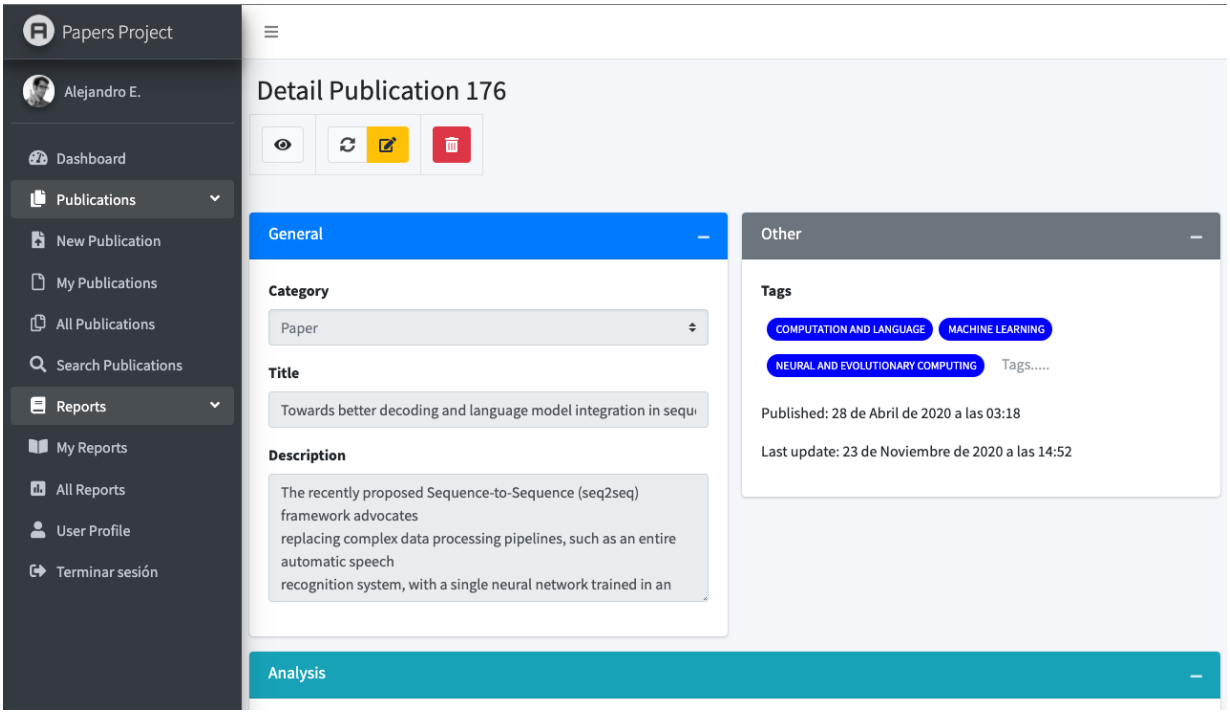

Figura C.6: Vista - Publicación

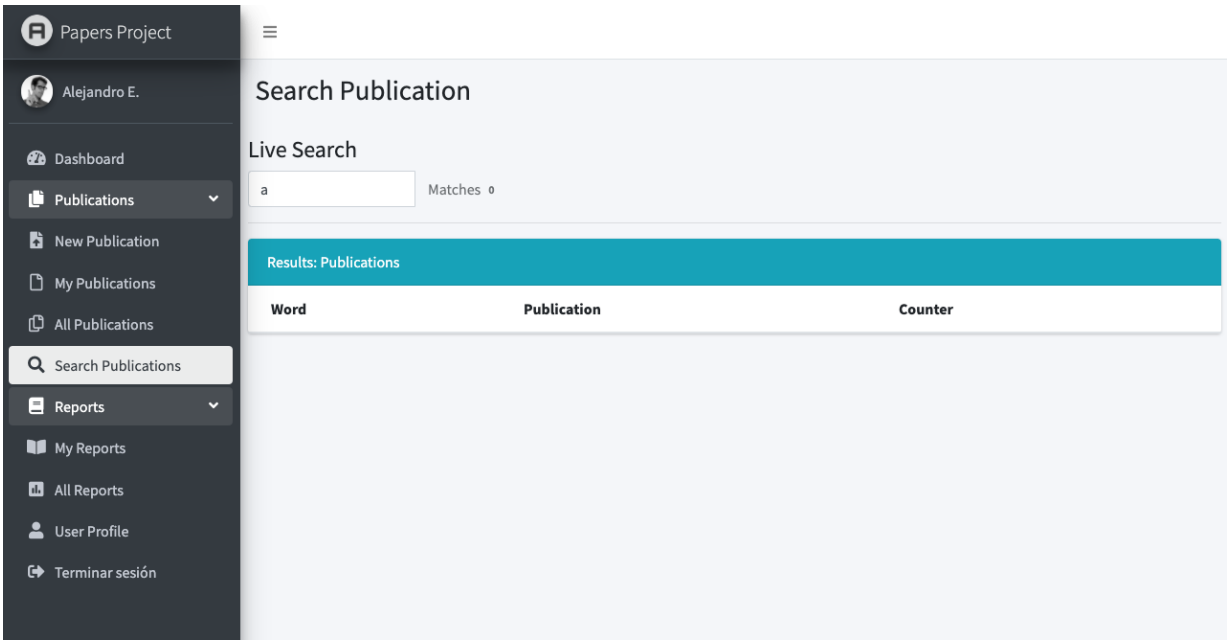

Figura C.7: Vista - Buscador de Publicaciones

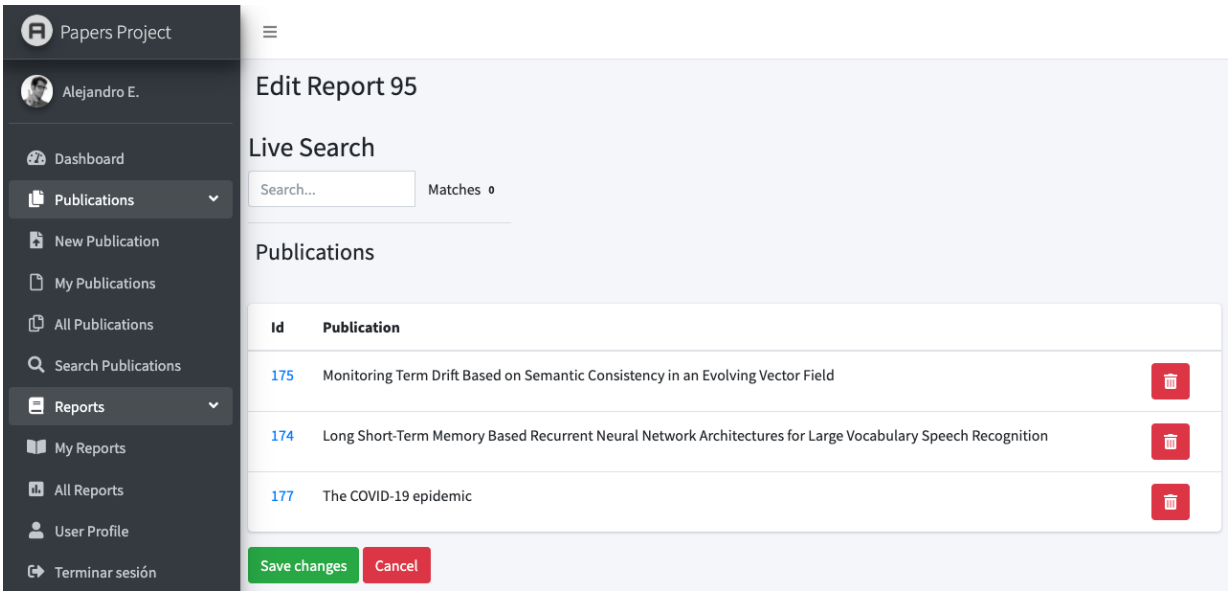

Figura C.8: Vista - Reporte

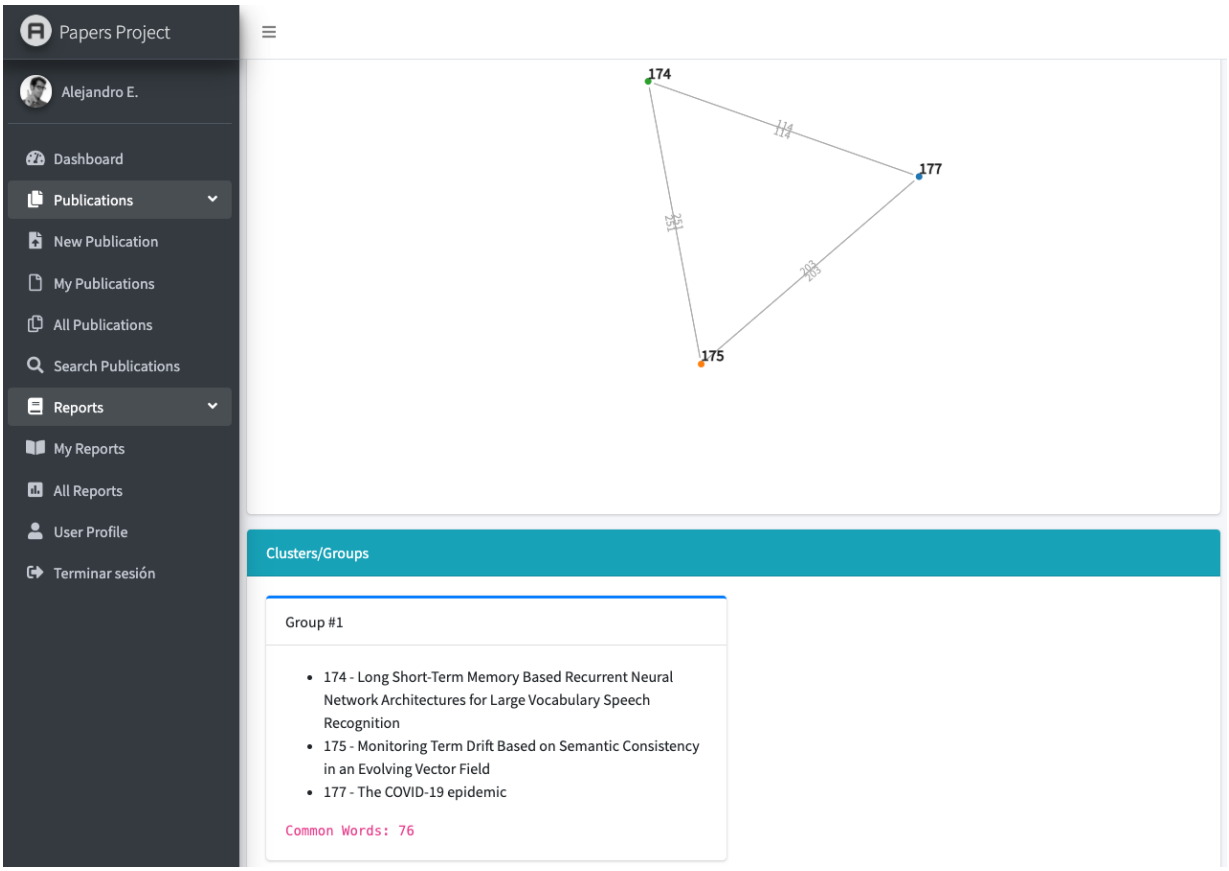

Figura C.9: Vista - Análisis de reporte

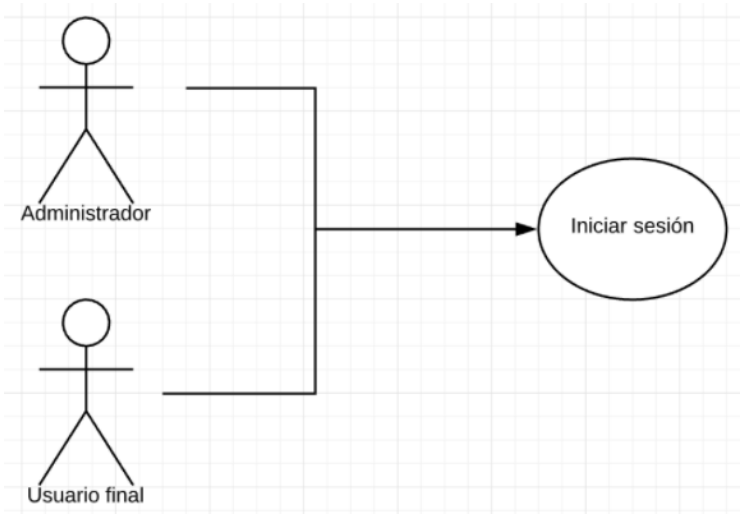

Figura C.10: Inicio de sesión

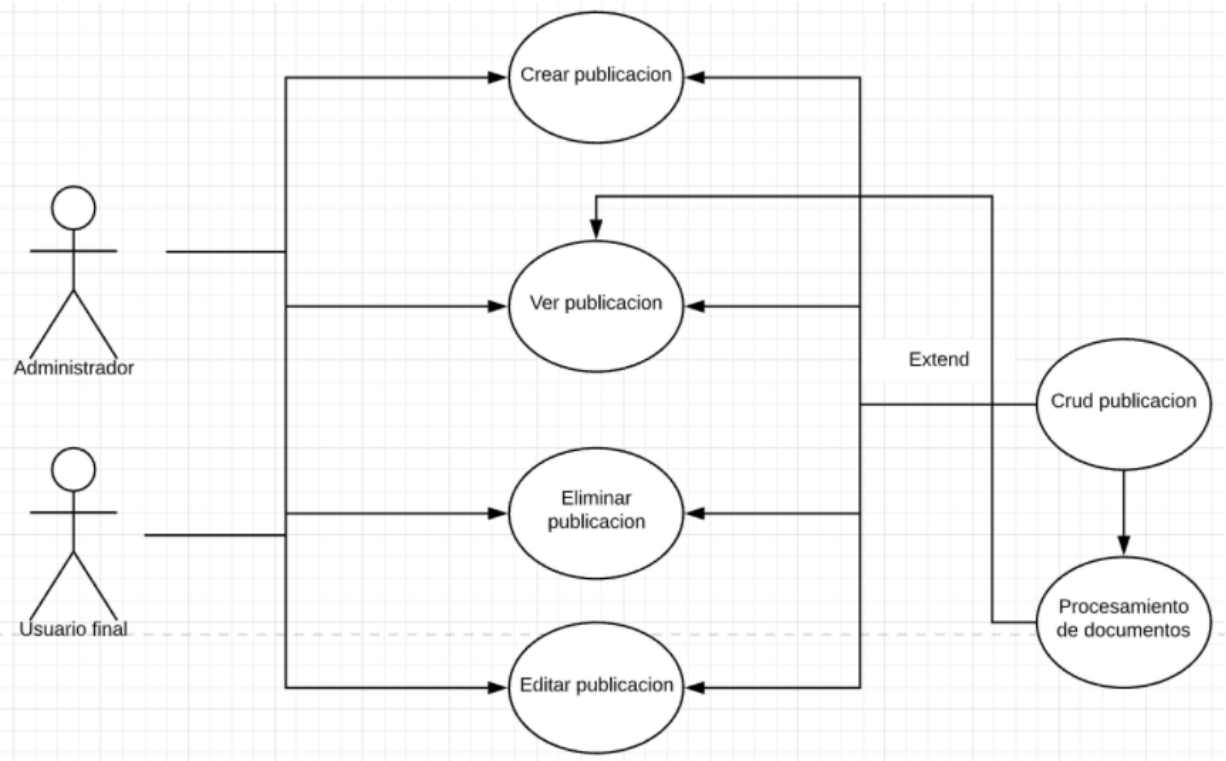

Figura C.11: Gestión de publicaciones

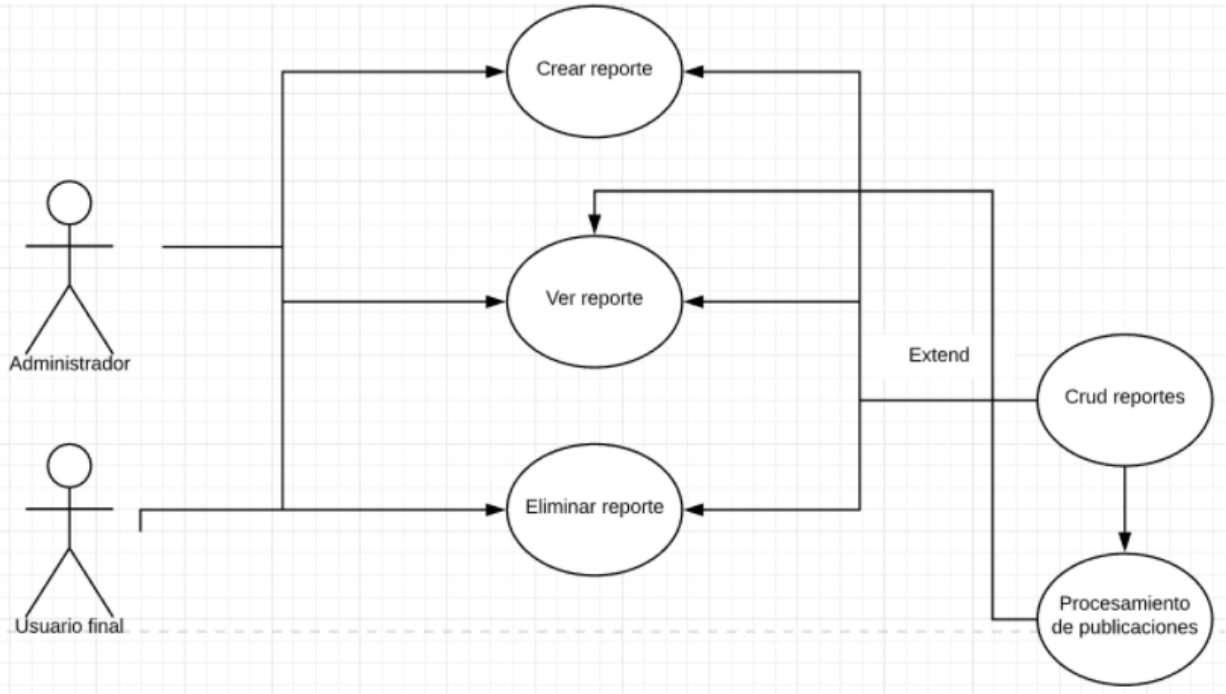

Figura C.12: Gestión de Reportes

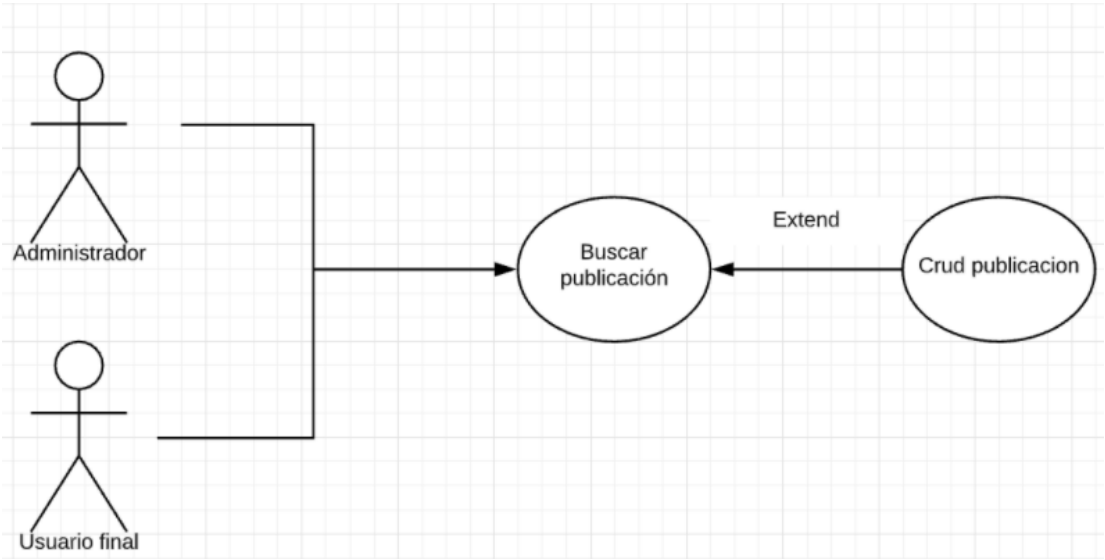

Figura C.13: Buscar publicaciones

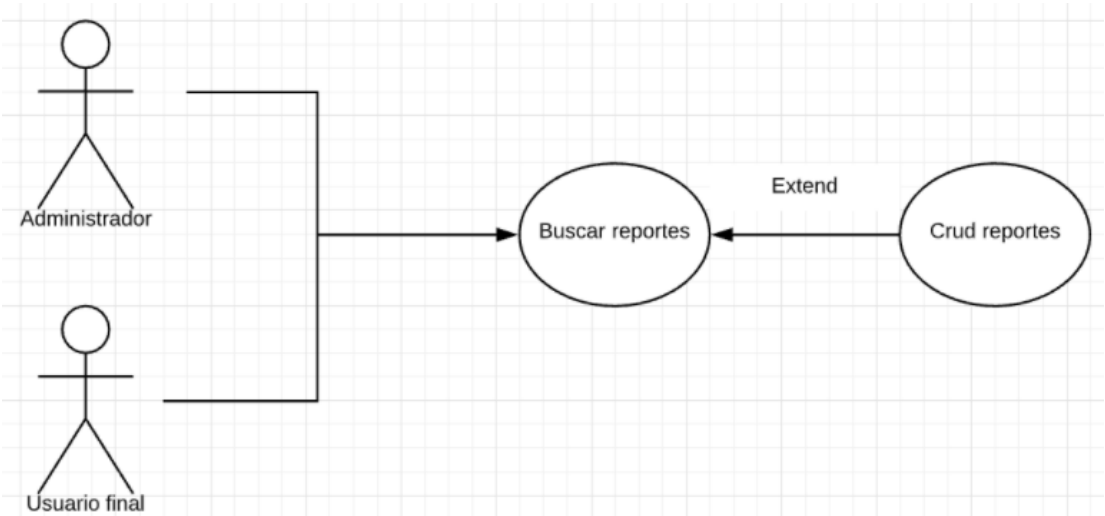

Figura C.14: Buscar reportes

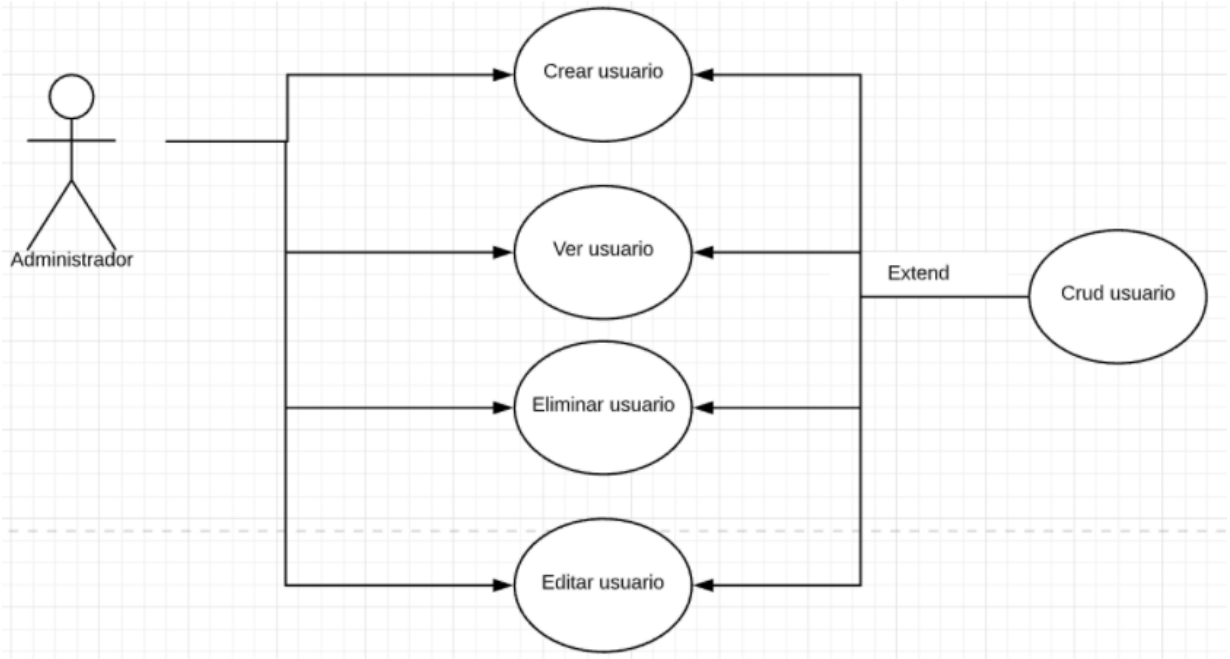

Figura C.15: Administración de usuarios

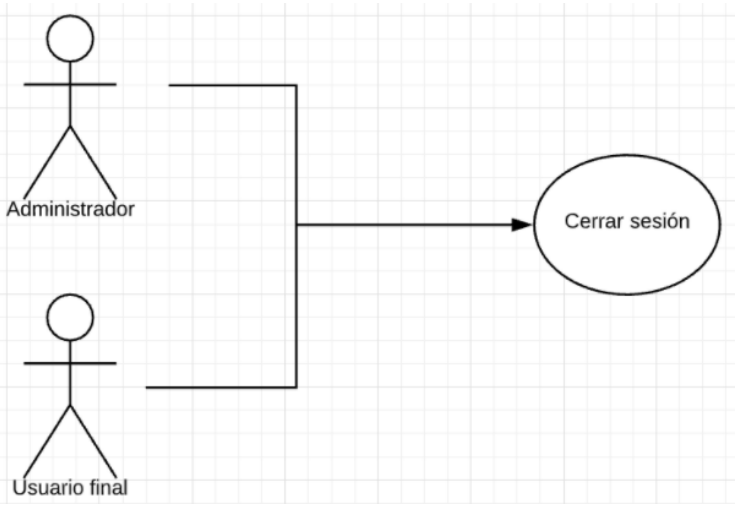

Figura C.16: Cerrar sesión

Figura C.17: Base de datos de archivos de procesamiento

Perfil	Lista de permisos
Root o Administrador	Administración de Usuarios Creación de publicación Asignación de usuarios a una publicación Cambio de contraseña de los usuarios Asignación de permisos a un publicación Eliminar Publicación Crear de reportes Eliminar Reportes Actualizar reportes Actualizar publicación Visualización de detalle de publicación Visualización de detalle de reporte
Usuario final	Creación de publicación Creación de reportes Eliminar Reportes Actualizar reportes Actualizar publicación Visualización de detalle de publicación Visualización de detalle de reporte

Cuadro C.1: Definición de Perfiles

Especificación de caso de uso			
Nombre	Iniciar sesión		
Actores	Administrador,Usuario final		
Tipo	Esencial		
Propósito	Proporcionar un mecanismo de autenticación por usuario		
Resumen	Los usuarios pueden acceder al aplicativo por medio de un usuario y/o contraseña		
Curso normal de eventos			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario ingresa a la vista inicial del proyecto ingresando la url del proyecto	2	El sistema carga el formulario para autenticación
3	Ingresa usuario y contraseña	4	Sistema valida datos de entrada: -Si son correctos crea variables de sesion para el usuario y automáticamente lo redirecciona al panel principal del proyecto -Si son incorrectos devuelve mensaje de error
Desviación del curso normal de los eventos			
5	Super usuario ingresa por la vista principal del proyecto	6	-Si el usuario es el administrador devuelve mensaje “Administradores utilizar la url de autenticación para administradores”

Cuadro C.2: Caso de uso: Iniciar sesión

Especificación de caso de uso			
Nombre	Gestión de publicaciones		
Actores	Administrador,Usuario final		
Tipo	Esencial		
Propósito	Proporcionar un mecanismo para el manejo de las publicaciones que se almacenarán en el proyecto		
Resumen	Los usuarios podrán administrar las publicaciones que existan en la plataforma		
Sección: crear publicaciones			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario ingresa a la vista para crear una nueva publicación	2	El sistema carga el formulario para cargar el archivo al sistema
3	El usuario completa los campos: -Categoría -Título -Descripción -Tags	4	Sistema valida datos de entrada: -Si son correctos crea la nueva publicación en el sistema, manda a encolar una nueva tarea de procesamiento del nuevo archivo a Celery y devuelve mensaje de éxito -Si son incorrectos devuelve mensaje de error
Sección: Editar publicaciones			
1	El usuario selecciona la publicación a editar	2	El sistema carga la información solicitada por el usuario, y le devuelve el formulario con los campos de la publicación para editar
3	El usuario edita cualquiera de los siguientes campos: -Categoría -Título -Descripción -Tags	4	Sistema valida datos de entrada: -Si son correctos actualiza la publicación en el sistema, manda a encolar una nueva tarea de procesamiento del archivo adjuntado a Celery y devuelve mensaje de éxito- Si son incorrectos devuelve mensaje de error

Cuadro C.3: Caso de uso: Gestión de Publicaciones (Parte 1)

Sección Eliminar publicación			
1	El usuario selecciona la publicación a eliminar	2	El sistema valida si la publicación a eliminar existe en la base de datos -Si existe, elimina el registro en la base de datos y devuelve mensaje de éxito -Si no existe devuelve mensaje de error
Sección Ver detalle de publicación			
1	El usuario selecciona la publicación a visualizar	2	El sistema valida si la publicación a visualizar existe en la base de datos -Si existe, carga toda la información en un template de html para ser visualizado por el usuario -Si no existe devuelve mensaje de error

Cuadro C.4: Caso de uso: Gestión de Publicaciones (Parte 2)

Especificación de caso de uso			
Nombre	Gestión de reportes		
Actores	Administrador,Usuario final		
Tipo	Esencial		
Propósito	Proporcionar un mecanismo para el manejo de los reportes que se almacenarán en el proyecto		
Resumen	Los usuarios podrán administrar los reportes que existan en la plataforma		
Sección: crear reporte			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario ingresa a la vista para crear una nuevo reporte	2	El sistema carga la lista de todas las diferentes publicaciones procesadas en el sistema
3	El usuario seleccion mínimo dos publicaciones para la creación de un nuevo reporte	4	Sistema valida datos de entrada: -Si son correctos crea un nuevo reporte en el sistema, manda a encolar una nueva tarea de procesamiento de las publicaciones seleccionadas a Celery y devuelve mensaje de éxito -Si son incorrectos devuelve mensaje de error
Sección: Ver reporte			
1	El usuario selecciona el reporte a visualizar	2	El sistema carga la información solicitada por el usuario, y le devuelve la información procesada por celery
Sección Eliminar reporte			
1	El usuario selecciona el reporte a eliminar	2	El sistema valida si el archivo a eliminar existe en la base de datos -Si existe, elimina el registro en la base de datos y devuelve mensaje de éxito -Si no existe devuelve mensaje de error

Cuadro C.5: Caso de uso: Gestión de Reportes

Especificación de caso de uso			
Nombre	Buscar publicaciones		
Actores	Administrador,Usuario final		
Tipo	Básico		
Propósito	Proporcionar un mecanismo para la búsqueda de las publicaciones que se manejan en el proyecto		
Resumen	Los usuarios podrán buscar las publicaciones que existan en la plataforma		
Sección: Buscar publicación			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario ingresa a la vista para buscar publicaciones	2	El sistema carga un formulario de búsqueda de las publicaciones
3	El usuario ingresa caracteres en el buscador	4	Sistema valida datos de entrada y busca si la información ingresada en el formulario existe dentro de la lista de publicaciones (se filtra por el título de las mismas), en caso de encontrar coincidencias devuelve la lista de publicaciones que concuerdan con el criterio de búsqueda

Cuadro C.6: Caso de uso: Buscar publicaciones

Especificación de caso de uso			
Nombre	Buscar reportes		
Actores	Administrador,Usuario final		
Tipo	Básico		
Propósito	Proporcionar un mecanismo para la búsqueda de reportes que se manejan en el proyecto		
Resumen	Los usuarios podrán buscar los reportes que existan en la plataforma		
Sección: Buscar reporte			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario ingresa a la vista para buscar reportes	2	El sistema carga un formulario de búsqueda de los reportes
3	El usuario ingresa caracteres en el buscador	4	Sistema valida datos de entrada y busca si la información ingresada en el formulario existe dentro de la lista de reportes (se filtra por los tag de las publicaciones que conforman el reporte y a nivel de frecuencia de las mismas), en caso de encontrar coincidencias devuelve la lista de publicaciones que concuerdan con el criterio de búsqueda

Cuadro C.7: Caso de uso: Buscar reportes

Especificación de caso de uso			
Nombre	Gestión de usuarios		
Actores	Administrador		
Tipo	Esencial		
Propósito	Proporcionar un mecanismo para el manejo de los usuarios que se harán uso de la plataforma		
Resumen	El administrador tendrá una interfaz para poder crear, visualizar eliminar y editar los usuarios que formarán parte de la plataforma.		
Sección: crear usuario			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El administrador ingresa a la vista para crear un nuevo usuario	2	El sistema carga el formulario para la creación de un nuevo usuario
3	El administrador completa los campos:-Usuario-Contraseña	4	Sistema valida datos de entrada: -Si son correctos crea el nuevo usuario en la plataforma- Si son incorrectos devuelve mensaje de error
Sección: Editar usuario			
1	El administrador selecciona el usuario a editar	2	El sistema carga la información solicitada por el usuario, y le devuelve el formulario con los campos del usuario para editar
3	El administrador edita cualquiera de los siguientes campos: -Usuario -Nombres -Correo -Permisos	4	Sistema valida datos de entrada: -Si son correctos actualiza la información del usuario en el sistema y devuelve mensaje de éxit -Si son incorrectos devuelve mensaje de error

Cuadro C.8: Caso de uso: Gestión de usuarios (Parte 1)

Sección Eliminar usuario			
1	El administrador selecciona el usuario a eliminar	2	El sistema valida si el usuario a eliminar existe en la base de datos -Si existe, elimina el registro en la base de datos y devuelve mensaje de éxito -Si no existe devuelve mensaje de error
Sección Ver detalle de usuario			
1	El administrador selecciona el usuario a visualizar	2	El sistema valida si el usuario a visualizar existe en la base de datos -Si existe, carga toda la información en un template de html para ser visualizado por el administrador -Si no existe devuelve mensaje de error

Cuadro C.9: Caso de uso: Gestión de usuarios (Parte 2)

Especificación de caso de uso			
Nombre	Cerrar sesión		
Actores	Administrador,Usuario final		
Tipo	Esencial		
Propósito	Proporcionar un mecanismo para salir de la plataforma		
Resumen	Los usuarios pueden salir de la plataforma cerrando sesión		
Curso normal de eventos			
Nº	Acciones de los actores	Nº	Respuesta del sistema
1	El usuario en el panel principal, selecciona la opción cerrar sesión	2	El sistema valida la sesión activa del usuario y cierra su sesión en el sistema

Cuadro C.10: Caso de uso: Cerrar sesión

Bibliografía

[1] "NLP Text Preprocessing and Cleaning Pipeline in Python," disponible en https://towardsdatascience.com/nlp-text-preprocessing-and-cleaning-pipeline-in-python-3bafaf54ac35.

[2] "Digital Ocean," disponible en https://cloud.digitalocean.com/.

[3] J. Niño, C. M. Gallego, Y. M. Correa, and O. M. Mosquera, "Production of scopolamine by normal root cultures of Brugmansia candida," *Plant Cell, Tissue and Organ Culture*, vol. 74, no. 3, pp. 289–291, 2003.

[4] J. Niño Osorio, Y. Correa Navarro, O. Mosquera Martínez, and L. Ramirez Q., "Cuantificacion de licorina en callos y raices cultivados in-vitro de crinum x powelli "album"(amaryllidaceae) por cromatografia liquida de alta eficiencia (hplc)." *Scientia et Technica*, vol. 3, no. 29, pp. 83–88, 2005.

[5] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash SyNakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. Consulted, 1–9." *Journal for General Philosophy of Science*, no. 1, pp. 1–9, 2008.

[6] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.phpro.2012.03.206

[7] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical Medicine and International Health*, vol. 25, no. 3, pp. 278–280, 2020.

[8] "La publicación científica y algunos fenómenos emergentes," disponible en http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-24362016000300251.

[9] "El Universo Digital se expande acelerado por el crecimiento de los datos," disponible en https://www.computerworld.es/tendencias/el-universo-digital-se-expande-acelerado-por-el-crecimiento-de-los-datos.

[10] "Grupos de Investigación en Colombia," disponible en https://minciencias.gov.co/la-ciencia-en-cifras/grupos.

[11] "Grupos de Investigación en UTP," disponible en https://www.utp.edu.co/vicerrectoria/investigaciones/investigaciones/grupos.html.

[12] P. Blanco Altozano, "EL ARTÍCULO CIENTÍFICO: PUNTUALIZACIONES ACERCA DE SU ESTRUCTURA Y REDACCIÓN," *Revista de Investigación en Educación*, vol. 7, no. 6, pp. 1–25, 2012.

[13] P. Calvo-Soto and V. M. Whizar-Lugo, "Como escribir un artículo científico," *Revista de Investigación en Educación*, no. 6, pp. 124–162, 2007.

[14] B. Rodriguez-Bravo, "El análisis documental de documentos digitales y/o multimedia," *Revista Códice*, no. 2, pp. 9–20, 2005.

[15] A. Cortez, H. Vega, and J. Pariona, "Procesamiento de lenguaje natural robusto," *Primer encuentro de grupos de investigación sobre Procesamiento del lenguaje*, vol. 2013, no. 3, p. 147, 2011.

[16] V. Balakrishnan and L.-Y. Ethel, "Stemming and lemmatization: A comparison of retrieval performances," *Lecture Notes on Software Engineering*, vol. 2, pp. 262–267, 01 2014.

[17] H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor, "BioLemmatizer: A lemmatization tool for morphological processing of biomedical text," *Journal of Biomedical Semantics*, vol. 3, no. 1, p. 3, 2012. [Online]. Available: http://www.jbiomedsem.com/content/3/1/3

[18] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," no. Cd, 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[19] M. G. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 583–605, 2007.

[20] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 855–864, 2016.

[21] L. Morissette and S. Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 1, pp. 15–24, 2013.

[22] M. A. Syakur, B. K. Khotimah, E. M. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, 2018.

[23] "Apache Tika," disponible en https://tika.apache.org/.

[24] "PyPDF2," disponible en https://github.com/mstamy2/PyPDF2.

[25] "Python NLTK," disponible en https://www.nltk.org/.

[26] A.-L. Barabasi and M. Pósfai, "Network science graph theory," 2016.

[27] D. Zhou, S. Niu, and S. Chen, "Efficient Graph Computation for Node2Vec," 2018. [Online]. Available: http://arxiv.org/abs/1805.00280

[28] A. Coates and A. Y. Ng, "Learning Feature Representations with K-Means," pp. 561–580, 2012.

[29] B. Jaeger, "The method of least squares," *Handbook of Research on Informatics in Healthcare and Biomedicine*, pp. 181–184, 2006.

[30] J. Niño Osorio, Y. Correa Navarro, O. Mosquera Martínez, and L. Ramirez Q., "Cuantificacion de licorina en callos y raices cultivados in-vitro de crinum x powelli .ªlbum"(amaryllidaceae) por cromatografia liquida de alta eficiencia (hplc)." *Scientia et Technica*, vol. 3, no. 29, pp. 83–88, 2005.

[31] J. Niño, O. M. Mosquera, and Y. M. Correa, "Antibacterial and antifungal activities of crude plant extracts from Colombian biodiversity," *Revista de Biologia Tropical*, vol. 60, no. 4, pp. 1535–1542, 2012.

[32] J. Niño, Y. M. Correa, G. D. Cardona, and O. M. Mosquera, "Antioxidant and antitopoisomerase activities in plant extracts of some Colombian flora from La Marcada Natural Regional Park," *Revista de Biologia Tropical*, vol. 59, no. 3, pp. 1089–1097, 2011.

[33] Y. M. Correa, J. Niño, and O. M. Mosquera, "DNA interaction of plant extracts from Colombian flora," *Pharmaceutical Biology*, vol. 45, no. 2, pp. 111–115, 2007.

[34] J. Niño, Y. M. Correa, and O. M. Mosquera, "Biological activities of steroidal alkaloids isolated from solanum leucocarpum," *Pharmaceutical Biology*, vol. 47, no. 3, pp. 255–259, 2009.

[35] P. Wittek, S. Darányi, E. Kontopoulos, T. Moysiadis, and I. Kompatsiaris, "Monitoring term drift based on semantic consistency in an evolving vector field," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2015-Septe, 2015.

[36] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, no. 5, pp. 523–527, 2017.

[37] H. Garcell, "Covid-19. a challenge for healthcare professionals covid-19." *Revista Habanera de Ciencias Médicas*, vol. 19, 04 2020.

[38] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 5900–5904, 2016.

[39] Z. Wang and S. Ji, "Learning convolutional text representations for visual question answering," in *SIAM International Conference on Data Mining, SDM 2018*, 2018, pp. 594–602.

[40] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[41] "Lenguaje de Programación Python," disponible en http://docs.python.org.ar/tutorial/pdfs/TutorialPython3.pdf.

[42] "Django Framework," disponible en https://devcode.la/blog/por-que-usar-django/.

[43] "PostgreSQL," disponible en https://www.postgresql.org/about/.

[44] "Celery," disponible en https://docs.celeryproject.org/en/stable/index.html.

[45] "What is Docker," disponible en https://opensource.com/resources/what-docker.

[46] A. Srivastava and M. Sahami, "Text mining: Classification, clustering, and applications," *Boca Raton*, 06 2009.

[47] E. L. S.Vidhya , D.Asir Antony Gnana Singh, "Feature Extraction for Document Classification," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 04, no. 06, pp. 50–56, 2016. [Online]. Available: www.ijirset.com

[48] A. De Sitter and W. Daelemans, "Information extraction via double classification," *Proceedings of International Workshop on Adaptive Text Extraction and Mining*, no. September 2003, pp. 66–73, 2000. [Online]. Available: http://www.aifb.uni-karlsruhe.de/WBS/pci/ontolearning.pdf

[49] A. R. Afonso and C. G. Duque, "Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Portuguese: Analysis and Comparison of Methods," *Journal of Information Systems and Technology Management*, vol. 11, no. 2, pp. 415–436, 2014.

[50] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv*, no. July, 2017.

[51] V. Renganathan, "Text mining in biomedical domain with emphasis on document clustering," *Healthcare Informatics Research*, vol. 23, no. 3, pp. 141–146, 2017.

[52] A. Téllez-Valero, M. Montes-y Gómez, and L. Villaseñor-Pineda, "A machine learning approach to information extraction," *Lecture Notes in Computer Science*, vol. 3406, no. May 2014, pp. 539–547, 2005.