UNIVERSIDAD PRIVADA ANTENOR ORREGO

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

"DASHBOARD DE PROYECTOS DE INVESTIGACIÓN DE TESIS DE LA UNIVERSIDAD PRIVADA ANTENOR ORREGO BASADO EN INDICADORES EXTRAÍDOS A PARTIR DE PROCESAMIENTO DEL LENGUAJE NATURAL EN EL PERIODO 2013-2019"

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE INGENIERO DE COMPUTACIÓN Y SISTEMAS

LÍNEA DE INVESTIGACIÓN:
SISTEMAS INTELIGENTES

AUTOR(ES):

BR. Alcalde Escobedo Luis Alberto

BR. Aguilar Flores Juan Carlos

ASESOR: ING. Heber Gerson Abanto Cabrera

TRUJILLO - PERÚ

2020

FECHA DE SUSTENTACIÓN: 25/06/2020

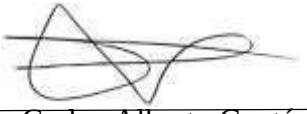
ACREDITACIONES

TÍTULO:

"DASHBOARD DE PROYECTOS DE INVESTIGACIÓN DE TESIS DE LA UNIVERSIDAD PRIVADA ANTENOR ORREGO BASADO EN INDICADORES EXTRAÍDOS A PARTIR DE PROCESAMIENTO DEL LENGUAJE NATURAL EN EL PERIODO 2013-2019"

AUTORES:

Br. Alcalde Escobedo, Luis Alberto
Br. Aguilar Flores, Juan Carlos



Ing. Carlos Alberto Gaytán Toledo
PRESIDENTE
N° CIP 84519

Ing. Agustín Eduardo Ullon Ramírez
SECRETARIO
N° CIP 137602

Ing. José Antonio Calderón Sedano
VOCAL
N° CIP 139198

Ing. Abanto Cabrera Heber Gerson
ASESOR
N° CIP 106421

PRESENTACIÓN

Señores Miembros del Jurado:

Dando cumplimiento y conforme a las normas establecidas en el Reglamento de Grados y Títulos y Reglamento de la Facultad de Ingeniería de la Universidad Privada Antenor Orrego, para obtener el título profesional de Ingeniero de Computación y Sistemas, se pone a vuestra consideración el informe del Trabajo de Investigación Titulado **"DASHBOARD DE PROYECTOS DE INVESTIGACIÓN DE TESIS DE LA UNIVERSIDAD PRIVADA ANTENOR ORREGO BASADO EN INDICADORES EXTRAÍDOS A PARTIR DE PROCESAMIENTO DEL LENGUAJE NATURAL EN EL PERIODO 2013-2019"** con la convicción de alcanzar una justa evaluación y dictamen.

Trujillo, 09 de Marzo de 2020

Alcalde Escobedo, Luis Alberto
Aguilar Flores, Juan Carlos

DEDICATORIA

Agradezco en primer lugar a Dios, mis Padres y mis amigos por su apoyo, consejos, amor y por darme la mano cada vez que los necesite en cada momento; especialmente por apoyarme a lograr esta meta trazada.

Especialmente también agradecer a mis abuelitos Sabina, Ulises, Chavelo que siempre me cuidan desde arriba iluminando mi camino en cada momento y también a mi abuelita Tomasita que siempre me impulsa a seguir adelante en cada paso que doy.

Este logro tan especial va dedicado a mis familiares, amigos del trabajo y compañeros de ruta y en especial a mi gran amigo Miguel Ángel en el Cielo que siempre me brindaron un apoyo en cada momento para llegar a ser un gran profesional.

LUIS ALBERTO

Esta tesis se lo dedico a Dios por ayudarmeen este largo camino tan importante de mi formación profesional. A mi madre Marcia, por su amor, trabajo y sacrificio en todosestos años. A mi abuelita Violeta por sus consejos, y palabras de aliento. A mi abuelito Gilberto que siempre me apoyo hasta sus últimos días y me dio las fuerzas para no rendirme y desde el cielo me sigue protegiendo y cuidando. A todas las personas que me acompañaron en esta etapa, gracias a ustedes he logrado llegar hasta aquí y convertirme en un Gran Profesional.

JUAN CARLOS

AGRADECIMIENTO

Agradecemos a nuestros padres por ser nuestro principal apoyo, gracias a ellos por confiar y creer en nosotros cada día, nos sentimos muy agradecidos con ustedes, a nuestro Asesor Heber Gerson Abanto Cabrera que nos ayudó y guio a realizar esta tesis y a nuestro gran amigo Oscar. No ha sido sencillo el camino, hubo momentos en que creíamos que no lo lograríamos, pero gracias al apoyo de nuestros padres, profesores y amigos pudimos lograr nuestro objetivo.

Realizar esta tesis no fue fácil, pero lo que puedo asegurar es que cada día nos divertimos, investigando, desarrollando y riendo por momentos, ha sido un gran trabajo el que hemos podido realizar con el apoyo de todos.

Todos estos años, son únicos e inolvidables, cada error cometido nos ayudó en nuestro aprendizaje y nos impulsó a seguir mejorando, y cada éxito nos dio la satisfacción de que nuestro esfuerzo valió la pena y poder decir "LO LOGRAMOS".

RESUMEN

En la presente tesis se desarrolló un Aplicativo Web para medir mediante indicadores las Tesis de un repositorio de la Universidad Privada Antenor Orrego. Esto a través de un Dashboard que permitirá al responsable del proceso de investigación desarrollo e innovación de la Escuela Profesional Ingeniería de Computación y Sistemas el acceso y manejo de las Tesis producidas y el análisis de esa producción.

Este proyecto se enfocó en la indagación adaptada a los Sistemas Inteligentes de cómo desarrollar una solución especifica que nos permita clasificar y visualizar toda la información de las Tesis por medio de bases de conocimientos, para su elaboración y procedimiento del Proyecto. Se recopilo 84 tesis del repositorio UPAO para tener una idea clara y precisa de como poder clasificar o buscar.

De esta forma se pudo trabajar la elaboración de un Aplicativo Web con la herramienta Shinny App de RStudio y por otro lado con el diseño e implementación del aplicativo web usando una metodología AGIL en la plataforma RStudio.

Posteriormente del desarrollo del aplicativo web se fijó un entorno de prueba para clasificar y medir la capacidad de las Tesis, seguidamente se utilizó el método de web scraping con el objetivo de obtener todas las Tesis mostradas del repositorio con la finalidad de conocer los resultados a través de la propuesta empleada.

Luego de obtener la los datos de las tesis pasamos a sacar los indicadores que necesitamos como: cuántas tesis pueden ver, cuantas líneas de investigación se encuentran, nombres de los jurados, con que continuidad los docentes participan como jurados o asesores.

Finalmente desarrollamos el dashboard el cual tendrá todos los indicadores obtenidos al realizar la minería de texto y además se expone la conclusiones y recomendaciones de la investigación

Palabras Claves: Aplicativo Web, Dashboard, Shinny App, RStudio, Clasificar, Sistemas Inteligentes, Medir, Web Scraping.

ABSTRACT

In this thesis a Web Application was selected to observe the classification and indicators of the Theses of a repository of the Antenor Orrego Private University. This is through a Board that is responsible for the development and innovation research process of the Professional School of Computer and Systems Engineering access and management of the theses produced and the analysis of that production.

This project focused on the investigation adapted to the Intelligent Systems of how to develop a specific solution that allows us to classify and visualize all the information of the Thesis through knowledge bases, for its elaboration and procedure of the Project. 84 theses from the UPAO repository were compiled which should have a clear and precise idea of how to classify or search.

In this way, it was possible to work on the development of a Web Application with the Shinny App tool from RStudio and on the other hand with the design and implementation of the web application using an AGIL methodology on the RStudio platform.

After the development of the web application, a test environment was established to classify and measure the capacity of the Thesis, then the web scraping method was used in order to obtain all the theses shown from the repository in order to know the results through of the proposal used.

After obtaining the data of the thesis we go on to get the indicators that we need such as: how many theses can see, how many lines of research are found, names of the jurors, with which continuity the teachers participate as jurors or advisors.

Finally, we develop the dashboard which will have all the indicators obtained when performing text mining and also exposes the conclusions and recommendations of the research.

Keywords: Web Application, Dashboard, Shinny App, RStudio, Sort, Intelligent Systems, Measure, Web Scraping.

ÍNDICE

ACREDITACIONES II
PRESENTACIÓN III
DEDICATORIA V
AGRADECIMIENTO VI
RESUMEN VII
ABSTRACT VIII
1. INTRODUCCIÓN 2
1.1. Realidad Problemática 2
1.2. Delimitación del Problema 3
1.3. Formulación del Problema 3
1.4. Formulación de la Hipótesis 3
1.5. Objetivos de Estudio 4
1.5.1. Objetivo General 4
1.5.2. Objetivo Especifico 4
1.6. Justificación del Estudio 5
1.7. Limitaciones del Estudio 5
2. MARCO TEORICO 5
2.1. Antecedentes 5
2.2. Marco Conceptual 12
2.2.1. Minería de Texto 12
2.2.2. Importancia de Minería de Texto 16
2.2.3. Modelo de Minería de Texto 16
2.2.4. Clasificación de Minería de Texto 17
2.3. Definiciones de Términos 20
2.3.1. Lenguaje R 20
2.3.2. Web Scraping 22
2.3.3. Dashboard: 22
2.3.4. Machine Learning 23
2.3.5. Text2vec 24
2.3.6. Tf-idf 25
2.3.7. Redes Neuronales Recurrentes 26
2.3.8. LSTM: 26
2.3.9. FASTTEXT 27

2.3.10. EXPRESIONES REGULARES REGEX (RegEx) 28
2.3.11. WORD EMBEDDING 29
2.4. Metodología Ágil 29
2.4.1. Objetivos de la Metodología Ágil 30
2.4.2. Características de Metodología Ágil 31
2.4.3. Ventajas de Metodología Ágil 31
2.4.4. Metodología Ágil más usadas 32
2.4.5. Proceso Ágil 33
2.4.6. Scrum 33
2.4.6.1. Roles que establece Scrum 34
2.4.6.1.1. El Project Owner o Dueño del Producto 34
2.4.6.1.2. El Scrum Master 34
2.4.6.1.3. El Equipo de Desarrollo o Scrum Team 34
2.4.6.2. Los pilares que provee Scrum permiten que el desarrollo del proyecto sea eficiente 35
2.4.6.3. Eventos de Scrum 35
2.4.6.3.1. Sprint 35
2.4.6.3.2. Planification de Sprint (Sprint Planning) 36
2.4.6.3.3. Objetivo del Sprint (Sprint Goal) 36
2.4.6.3.4. Scrum Diario (Daily Scrum) 36
2.4.6.3.5. Revision de Sprint (Sprint Review) 37
2.4.6.3.6. Retrospectiva de Sprint (Sprint Retrospective) 37
3. MATERIAL Y MÉTODO 38
3.1. Material 38
3.1.1. Población 38
3.1.2. Muestra 38
3.2. Método 38
3.2.1. Tipo de Investigación 38
3.2.2. Diseño de la Investigación 38
3.2.3. Variables de Estudio y Operacionalización 39
3.2.3.1. Variables 39
3.2.3.2. Operacionalización 39
3.2.4. Instrumentos de Recolección de Datos 40
3.2.5. Procedimientos y Análisis de Datos 40

4. RESULTADOS 41
4.1. ¿Por qué hacer un Dashboard con Metodología Ágil Scrum? 41
4.1.1. Artefactos de Scrum 41
4.1.1.1. Lista de Producto o Product Backlog 41
4.1.1.2. Sprint Backlog 42
4.1.1.3. Incremento 43
4.2. Desarrollo Del Proyecto 43
4.2.1. Etapas del desarrollo del Dashboard 43
4.2.1.1. Aplicación de la técnica Web Scraping 44
4.2.1.2. Diseño de Programa para el Dashboard 46
4.2.1.2.1. Extracción De Información De Los Documentos Pdf 47
4.2.1.2.2. Preprocesamiento De Datos 49
4.2.1.2.3. Análisis De La Extracción De Datos 51
4.2.1.3. Desarrollo e Implementación de gráficos para el Dashboard 54
4.2.1.4. Desarrollo del Aplicativo Web en Shiny Apps 56
5. DISCUSIÓN DE RESULTADOS 57
5.1. HIPOTESIS PLANTEADA 57
5.1.1. Muestra Aplicar 57
5.1.2. Validación de la Solución 57
5.1.3. Variable Independiente: Dashboard basado en Minería de Texto. 58
5.1.4. Variable Dependiente: Medición de los proyectos de investigación de tesis de la Universidad Privada Antenor Orrego 59
6. CONCLUSIONES 60
7. RECOMENDACIONES 61
8. REFERENCIAS BIBLIOGRAFICAS 62
9. ANEXOS 66

Índice de Tablas

Tabla 1. Operacionalización de Variables Independiente 39
Tabla 2. Operacionalización de Variables Dependiente 40
Tabla 3. Instrumentos de Recolección de Datos 40
Tabla 4. Formato de Valoración 58
Tabla 5. Valoración de Facilidad de Uso 58
Tabla 6. Valoración de la Satisfacción del Dashboard 59

Índice de Figuras

Figura 1. Modelo de Minería de Texto 3
Figura 2. Gestión de Riesgo 13
Figura 3. Servicio de Atención al Cliente en la Minería de Texto 14
Figura 4. Detección de Fraude en la Minería de Texto 14
Figura 5. Inteligencia de Negocio en la Minería de Texto 15
Figura 6. Análisis de Redes Sociales para la Minería de Texto 15
Figura 7. Modelo de cómo funciona la Minería de Texto 16
Figura 8. Modelo de Transformación de Minería de Texto **18**
Figura 9. Modelo de Extracción de Características de Minería de Texto 18
Figura 10. Lenguaje R 21
Figura 11. Web Scraping 22
Figura 12. Dashboard 23
Figura 13. Machine Learning 24
Figura 14. Text2vec 25
Figura 15. Tf-idf 25
Figura 16. Redes Neuronales Recurrentes 26
Figura 17. LSTM 27
Figura 18. FASTTEXT 27
Figura 19. Expresiones Regulares Regex 28
Figura 20. Word Embedding 29
Figura 21. Plan de la Metodología Ágil con sus principio 30
Figura 22. Metodología Ágil adaptada al Plan de Trabajo. 31
Figura 23. Metodología Ágil más usadas 32
Figura 24. Metodología Scrum en Proyecto de Desarrollo 33
Figura 25. Investigación postprueba 38
Figura 26. Lista de Productos 42
Figura 27. Tablero Físico De Scrum 43
Figura 28. Repositorio de la Escuela Ingeniería de Computación y Sistema de la Universidad Privada Antenor Orrego 44
Figura 29. Web Scraping 45
Figura 30. Proceso del Scraping 45
Figura 31. Código para la descarga de los PDF 46
Figura 32. Proceso de extracción de los documentos PDF 46
Figura 33. Código para la limpieza de los datos extraídos 47
Figura 34. Proceso para ver la línea de investigación 47
Figura 35. Proceso de búsqueda de Presidente, Vocal, Secretario 48
Figura 36. Arreglo de nombres y limpieza todos los datos. 48
Figura 37. Proceso de limpieza de datos de los jurados 49
Figura 38. Proceso de ejecución donde comenzaremos a ordenar los datos de asesor, autores, presidente, vocal, secretario 49
Figura 39. Código incluir a dos docentes 50

Figura 40. Incluir docente sin grados y títulos 50
Figura 41. Limpieza y creación de variables df_limpio 51
Figura 42. Datos limpios y eliminamos columnas que no interesan 52
Figura 43. Agrupamos los datos que requerimos 52
Figura 44. Tesis por Línea 53
Figura 45. Ordenamiento por año y comienzo de gráficos 53
Figura 46. Proceso de realización y ejecución del Dashboard 54
Figura 47. Datos ordenados con gráficos correspondientes 54
Figura 48. Gráficos de todas las tesis por año y fecha 55
Figura 49. Etapas de gráficos de cada panel por año, tesis y finalización 55
Figura 50. Ingreso del Shinyapps y visión general 56
Figura 51. Visión del Dashboard en la web 56
Figura 52.Dashboard con los datos extraídos de las tesis 57
Figura 53. Resultado de Facilidad de Uso 59
Figura 54. Resultado de la Satisfacción del Dashboard 60

1. INTRODUCCIÓN

1.1. Realidad Problemática

Para la realización de este Proyecto de Investigación se consideró lo dicho por el autor en el siguiente párrafo:

"El texto no tiene estructura, o, mejor dicho, su estructura es implícita es demasiado complicada y rica como para ser tratada computacionalmente sin un procesamiento previo. Esta aparente ausencia de estructura es el mayor problema de la Minería de Textos: la necesidad de preparar los documentos, pasarlos a una Forma Intermedia para que se les pueda aplicar algún algoritmo automático es una de las principales diferencias de la Minería de Textos con las Minería de Datos."(Torre, 2017)

Hoy en la actualidad manejamos información de formatos de textos no estructurados o semiestructurados, tales como los correos electrónicos, fuentes informativas de noticias, revistas, páginas web, etc. Las Compañías tienen una gran cantidad de Información que los perjudica a la hora de preguntar de como coleccionar, investigar y utilizar toda esta información.

El Texto Analítico, es el procedimiento del cual va inspeccionar grandes cantidades de requerimientos escritos, lo que genera una actual o moderna información y convirtiendo el texto no estructurado en datos estructurados para el posterior uso de sus análisis subsiguientes. Es decir, el proceso de originar información valiosa de texto, tiene como finalidad absoluta cambiar el texto en datos para un análisis respectivo, mediante la aplicación del proceso del lenguaje natural (PNL) y métodos analíticos. Además de extraer información necesaria del texto para que puedan ser relaciones con categorías adecuadas. (Vásquez & Mg. Hugo Vega Huerta, 2019)

Hoy en día la Escuela Profesional de Ingeniería de Computación y Sistemas no tiene una aplicación que ayude al responsable del proceso de

investigación, desarrollo e innovación en comprender como se maneja los proyectos de tesis.

Por consiguiente, se propone desarrollar a través de la Minería de Texto un Dashboard que contenga todas las tesis y mediante indicadores saber cuántas tesis pueden ver, cuantas líneas de investigación se encuentran, nombres de los jurados, con que continuidad los docentes participan como jurados o asesores. De esa manera se quiere satisfacer las necesidades de los docentes de la Universidad, con una gran calidad de fácil uso.

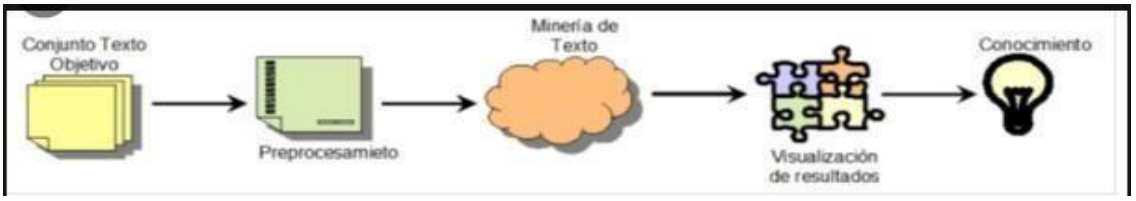


Figura 1. Modelo de Minería de Texto (Unknown, 2017)

1.2. Delimitación del Problema

- Delimitación Espacial: El proyecto se desarrollará e implantará en la ciudad de Trujillo, específicamente en la Universidad Privada Antenor Orrego
- Delimitación Temporal: La duración del proyecto es de 4 a 5 meses.
- Delimitación del Universo: Repositorio de Proyectos de Tesis de la Universidad Privada Antenor Orrego

1.3. Formulación del Problema

¿Como podremos medir mediante indicadores los Proyecto Investigación de Tesis para la Universidad Privada Antenor Orrego?

1.4. Formulación de la Hipótesis

Es posible que la implementación de un Dashboard basada en minería de texto ayudara a medir los proyectos de tesis en la universidad privada Antenor Orrego.

1.5. Objetivos de Estudio

1.5.1. Objetivo General

Desarrollar un Dashboard de Proyectos de Investigación de Tesis de la Universidad Privada Antenor Orrego que permita medir indicadores basados en Procesamiento de Lenguaje Natural.

1.5.2. Objetivo Especifico

- Utilizar la técnica de Web Scraping que nos permita recolectar todas las Tesis de Investigación.
- Diseñar un programa para el Dashboard.
- Desarrollar e implementar gráficos que analicen a partir de los indicadores las Tesis para el estudio por parte de la Universidad.
- Desarrollar un Aplicativo Web en el Framework Shinny App de RStudio, con el fin de que el usuario pueda interactuar y visualizar de manera profunda los datos o los indicadores de los Proyectos de Investigación de Tesis que fueron obtenidos a través de Minería de Texto. Esto bajo la forma de trabajo de Metodología Ágil para control de iteraciones.

1.6. Justificación del Estudio

La importancia de este proyecto se enfoca en obtener información con extensas proporciones de texto, donde el reporte es no estructurado, para realizar un análisis a los textos para finalmente observar los resultados obtenidos.

A través de la minería de texto el proyecto tiene como finalidad desarrollar en una plataforma de ShinyApp la cual nos ayudara a medir mediante indicadores, para ver la cantidad de proyectos de Investigación de Tesis que existen en el repositorio de la Universidad Privada Antenor Orrego, usando el lenguaje de programación R, que permite realizar el procesamiento de los textos y poder visualizarlos.

Desde el punto de vista de la investigación:

Este aplicativo web tiene la necesidad de medir los Proyectos de Investigación debido a una cierta cantidad de Proyectos.

Desde el punto de vista de legado:

El desarrollo de este aplicativo web incentiva a demostrar que se puede crecer a un nivel alto de tecnología estudiantil midiendo cada Proyecto de Investigación de Tesis utilizando inteligencia artificial.

1.7. Limitaciones del Estudio

El alcance del proyecto está delimitado a extraer los datos que nos proporcionan el repositorio de los Proyectos de Investigación de Tesis de la Universidad Privada Antenor Orrego, con la plataforma Shinny App de RStudio, con la implementación de un Dashboard para medir y reportar los datos obtenidos de los Proyecto de Tesis, posteriormente se desarrollara y se implementara en el tiempo de 5 meses.

2. MARCO TEORICO

2.1. Antecedentes

2.1.1. Antecedentes Locales

Título: Minería de Textos para construir un Modelo de Minería de Datos para Recomendación de Ofertas Laborales

Tipo de Ítem: Tesis

Autor: Honorio Apaza Alanoca

Año: 2019

Conclusión: Los sistemas de recomendación necesitan un valor numérico sobre los gustos y disgustos de los usuarios para hacer listas de recomendación basada en colaboración o similitud de contenido. Sin embargo, en la presente investigación la información a considerar son los términos o palabras relevantes que describen al candidato, experiencias y habilidades. Esta investigación se centra en el proceso de minería de textos, la cual consiste en el proceso de estructuración del modelo de datos partir de textos que describen a candidatos y ofertas laborales, el cual se considera como datos de entrada para los algoritmos de sistema de recomendación. Por tratarse de información textual y en lenguaje humano, se aplican las teorías de técnicas de minería de textos y procesamiento lenguaje natural para el desarrollo del presente trabajo, comenzando desde la recuperación de información hasta la estructuración de minería de datos a partir de datos de currículums vitae y ofertas laborales de sitios web. Como técnica relevante para el presente estudio se destaca la técnica frecuencia de término – frecuencia inversa de documentos -tf-idf, el cual permite a identificar términos relevantes de currículums vitae con relación a las ofertas laborales de sitios web. La misma que ayuda a determinar la relevancia de los currículums vitae con respecto a las ofertas laborales con un valor numérico a través del promedio de valores tf-idf. De tal manera que el valor ponderado pueda ser usado como valor rating de relevancia de currículums vitae para recomendación de ofertas laborales.

Título: Aplicación de Tecnologías Semánticas y Minería de Textos para la Comparación de Sílabos entre Distintas Universidades

Tipo de Ítem: Tesis

Autor: Abril, Esteban Sebastián Espinoza; Uguña, Noemi Elizabeth Sari

Año: 2018

Conclusión: Este trabajo se propone la utilización de tecnologías de la Web Semántica para la creación de una ontología de sílabos, la ontología será poblada con los datos que residen en los sistemas de información de la Universidad de Cuenca y se guardará en un repositorio. Posteriormente, sobre este repositorio se aplicará técnicas de minería de textos con el fin de identificar similitudes en el contenido académico de diferentes sílabos, de manera que se automatice el proceso de comparación de sílabos parala movilidad estudiantil y minimizar los errores en el proceso.

Para la evaluación experimental, se construyó un corpus especialmente diseñado para la tarea de la identificación de plagio parafraseado, y se implementó y evaluó un amplio conjunto de métodos de referencia provenientes de la detección de plagio y paráfrasis con el fin de permitir una comparación fiable con el estado del arte.

2.1.2. Antecedentes Internacionales

Título: Aplicación de Tecnologías Semánticas y Minería de Textos para la Comparación de Sílabos entre Distintas Universidades

Tipo de Ítem: Tesis

Autor: Abril, Esteban Sebastián Espinoza; Uguña, Noemi Elizabeth Sari

Año: 2018

Conclusión: Este trabajo se propone la utilización de tecnologías de la Web Semántica para la creación de una ontología de sílabos, la ontología será poblada con los datos que residen en los sistemas de información de la Universidad de Cuenca y se guardará en un repositorio. Posteriormente, sobre este repositorio se aplicará técnicas de minería de textos con el fin de identificar similitudes en el contenido académico de diferentes sílabos, de manera que se automatice el proceso de comparación de sílabos para la movilidad estudiantil y minimizar los errores en el proceso.

Para la evaluación experimental, se construyó un corpus especialmente diseñado para la tarea de la identificación de plagio parafraseado, y se implementó y evaluó un amplio conjunto de métodos de referencia

provenientes de la detección de plagio y paráfrasis con el fin de permitir una comparación fiable con el estado del arte.

Título: Un modelo de clasificación eficiente para no estructurados Documento de texto

Tipo de Ítem: Paper

Autor: Mowafy , Un rezk y El-Bakry

Año:2018

Conclusión: En este trabajo nos refleja la manera de que clasificar documentos se ha vuelto muy importante en el estudio de la investigación debido a la suma de documentos de textos no estructurados aceptables en forma digital, ya que se acepta una de la forma más principal utilizada para la Organización de datos digitales por medio de documentos en categoría predefinidas en función de su contenido, la forma de clasificar documentos en un etapa que consta en un conjunto de fases, las cuales se llevan mediante técnicas; no obstante, que para seleccionar una técnicaadecuada se debe utilizar la fase afectada para el rendimiento de la clasificación del texto, es decir, el objetivo principal de este trabajo es mostrar un Modelo de Clasificación de texto que puede soportar tanto la generalidad y eficacia. Donde nos dice que la generalidad es la secuencia lógica del proceso de clasificación de los documentos del texto estructurado. Los resultados obtenidos por más de 20 NEWGROUPS conjunto de datos que validad utilizando medidas estadísticas de precisión, recordar; han obtenido la forma de dar un modelo propuesto para un mejorrendimiento. En conclusión en este trabajo presentamos la clasificación de un texto mediante corriente de las fases a través de un clasificador automático de documento de texto y la relación entre ellos, en consistencia los Naives Bayes sirven para dar un enfoque clasificatorio a los documentos de texto, lo cual eligiendo técnicas en el modelo observamos que puede dar un mejor resultado mediante un proceso de clasificación , donde se verifica que existe una compatibilidad entre las técnicas seleccionadas de distintas fases.

Título: Uso de técnicas de minería de texto para identificar tendencias de investigación: un estudio de caso de investigación de diseño

Tipo de Ítem: Articulo de Revista

Autor: Binling Nie and Shouqian Sun

Año:2017

Conclusión: Uso de técnicas de minería de texto para identificar tendencias de investigación: un estudio de caso de investigación de diseño El objetivo de investigación de este artículo es identificar las principales ramas académicas y detectar tendencias de investigación en investigación de diseño utilizando técnicas de minería de textos. En este documento, se procesa la información sobre literatura científica en investigación de diseño. Una combinación de agrupamiento y análisis bibliométrico llevó a conformar cuatro ramas académicas y resumir cada rama académica. Luego, se exploran las tendencias de investigación y la evolución de cada rama académica. Realizamos un enfoque de minería de texto bidimensional, que incluye análisis bibliométrico y de red, para detectar tendencias de las principales ramas académicas.

las técnicas de minería de texto utilizadas en este estudio podrían ayudar a los investigadores a comprender el conocimiento de cierto campo oculto en una gran cantidad de literatura científica. El método de agrupamiento proporciona una descripción general de la arquitectura de un determinado campo con más detalle. Además, el análisis de redes sociales explora aún más los temas centrales para ayudar a los investigadores a comprender mejor la ganancia de desarrollo de un determinado campo.

Título: El uso de Minería de Texto para clasificar documentos de investigación

Tipo de Ítem: Paper

Autor: Radka Nacheva, Latinka Todoranova

Año: 2017

Conclusión: Este presente trabajo se encuentran en muchos sitios web, lo cual presenta un problema al momento al momento de ordenar, clasificar los documentos por característica o por campo científico; Con la manera

de hacer posible automatizando el proceso de organizar y clasificar con texto sin formato con la manera de hacerlo más fácil el proceso de ayudar y automatizar los artículos científicos enfocado en tecnología de lenguaje natural donde se aplicamos la Maquina de Vector de Soporte y los Naive Bayes. El tema propone en clasificar la literatura científica con su contenido, en conclusión el enfoque a utilizar a través del estudio que tiene el Minería de Texto para clasificar documentos de Investigación , nos basamos en proceso de vectores y Naives Bayes notación de apoyo, donde Los algoritmos escogidos se dan con reportes necesarios y confiables en el momento de clasificar datos de textos no estructurados, con la finalidad de mostrar y aprobar los que se ha presentado en resultado anteriores, donde el software a utilizar es RapidMiner ; de esta manera los formatos de artículos se clasifican en dos etapas, donde se asigna en categorías(Social , Tecnología, etc.) . Los resultados enseñan métodos propuestos para que pueda ser seleccionado con éxito para la clasificación de literatura científicas en categoría, de acuerdo a los investigadores eso nos permitirá una gran investigación científica que aumentar la efectividad de los trabajos en su identificación y buenos resultados con problemas de investigación.

Título: Minería De Texto Como Una Herramienta Para La Búsqueda De Artículos Científicos Para La Investigación

Tipo de Ítem: Paper

Autor: A.Arias, Y. Mattos, J. Heredia & D. Heredia

Año:2016

Conclusión: Esté presente trabajo comprende el estudio de análisis de información a través de la minería de texto (text mining) también conocida como minería de datos de texto, la cual pueden ser utilizadas para la extracción de información de manera eficaz y precisa desde las diferentes bases de datos cuyos textos no se encuentran estructurados o semiestructurados. Esta investigación se realizó mediante la recopilación de datos desde diferentes bases de datos con muy buenas referencias, la cual da como resultados factibles y de mucha confianza en la redacción del

contenido, además se contó el apoyo de personas idóneas en el tema de minería de texto.

Título: Enfoque de Minería de Texto para clasificar técnicos Documentos de investigación utilizando Naïve Bayes
Tipo de Ítem: Paper
Autor: Mahesh Kini, Saroja Devi, Prashant G Desai, Niranjan Chiplunkar
Año:2015
Conclusión: En este trabajo se habla que hay un gran volumen de documentos de texto que ha enfocado a los descubridores a enfocarse en la mejoría de la información clara y precisa que se encuentran en los grandes recursos, donde en gran parte la Minería de Texto es importante para la distribución de documentos de texto o determinación. En el proceso de recopilación de texto, metodología fuerte y eficaz de algoritmos como Naives Bayes es eficiente que cumplen con el rol de clasificar los documentos de texto para hacer una gran tarea firme y sólida, sin embargo al momento de clasificar los documentos de texto a través de clasificadores bayesianos que son algoritmos que llegan a que tengan un mejor éxito para el aprendizaje automático, nos relata que el documento especifica la implementación de Naive Bayesiano(NB) donde nos una prioridad de clasificar automáticamente a los documentos limitados a los documentos tecnológicos en función a su contenido de texto y a su análisis de resultado. En conclusión, nos dice que es un trabajo arduo con el procesamiento de la información, lo cual no es tan fácil adaptarnos. Los clasificadores de documentos son spam de correo electrónicos. Además, donde puede ser realmente sencillo para los humanos, pero difícil para una máquina. Donde el enfoque Bayesiano ha aumentado en el bienestar de características de un documento de investigación para su clasificación. En este proceso los motores de búsqueda son aprovechados por los sitios web para la construcción de un directorio de sitio web automatizado en el tipo de organización, donde el mismo algoritmo es utilizado para mejorar las paginas en categorías más precisas.

2.2. Marco Conceptual

2.2.1. Minería de Texto

La minería de texto o text mining es una forma de analizar información de datos no estructurados que son casi el 80% de datos del mundo; por ejemplo, revistas, páginas web, libros. estos datos. Este proceso puede realizarse al identificar patrones significativos, como el uso de palabras frecuentes, estructuras semánticas, etc.

En los últimos años el crecimiento de los datos está multiplicándose exponencialmente a medida que llegan información de varias fuentes. el resultado de este incremento de información propone nuevos desafíos para las empresas y organizaciones almacenar, procesar y analizar grandes cantidades de datos textuales, es aquí donde entra la minería de texto. Hay muchas aplicaciones para la minería de texto, extraer y analizar textos ayuda a las organizaciones a encontrar información útil en correos electrónicos, documentos corporativos, encuestas, publicaciones en redes sociales. esto reduce el tiempo que las empresas dedican a la hora de leer documentos y revisar información en redes sociales.

Las aplicaciones de minería de texto poden encontrarlo desde su uso en las academias hasta en las atenciones médicas. estos son algunos:

- **Gestión De Riesgo**: La principal causa de fracaso en las empresas es el poco análisis de riesgo adecuado o insuficiente. la adquisición de un software de gestión de riesgo con tecnología de

minería de datos, ayuda a las empresas a mantenerse actualizadas en tendencia actuales del mercado empresarial e incrementar las capacidades para moderar los peligros potenciales. de esta manera, como las tecnologías de minería de texto pueden recopilar información de varias fuentes de datos textuales y establecer vínculos entre los conocimientos extraídos, las empresas acceden a información precisa en el momento correcto, mejorando así todo el proceso de gestión de riesgo.

Figura 2. Gestión de Riesgo (GERENS, 2017)

- **Servicio De Atención Al Cliente**: Los métodos de minería texto, especialmente PNL, está ganando mayor relevancia en el campo de atención al cliente. las compañías invierten en software de análisis de texto para poder aumentar la experiencia general de los clientes a la hora de ingresar datos textuales de varias fuentes como encuestas, comentarios y llamadas de los clientes, etc. el análisis de texto tiene como principal función acortar el tiempo de respuesta de las compañías y apoyar a los clientes en sus quejas de manera eficiente y rápida.

Figura 3. Servicio de Atención al Cliente en la Minería de Texto (La atención al cliente del futuro:, 2017)

- **Detección De Fraude**: El análisis de texto apoyado por tecnologías de minería de texto brinda una gran oportunidad a los dominios que recopilan datos en forma de texto. las organizaciones de seguros y financieras combinan los resultados del análisis de texto con datos estructurados destacados, para atender reclamos rápidamente y también encontrar y evitar fraudes.

Figura 4. Detección de Fraude en la Minería de Texto (MeaningCloud, 2014)

- **Inteligencia De Negocios**: Las empresas y organizaciones están aprovechando la minería de texto como parte de su inteligencia empresarial. las técnicas de minería de texto además de proporcionar información de sus clientes, también ayuda a analizar las fortalezas y debilidades de sus rivales, esto ofrece una ventaja competitiva en el mercado.

Figura 5. Inteligencia de Negocio en la Minería de Texto
(MARTOS, 2018)

- **Análisis De Redes Sociales**: Hay bastante software de minería de texto para redes sociales, que ayudan a rastrear e interpretar los textos a partir de blog, noticias, etc. Estos softwares pueden realizar el análisis de gran cantidad de publicaciones, seguidores de marcas y me gusta en las redes sociales, esto permite a las organizaciones comprender mejor las reacciones de las personas (Rai, 2019)

Figura 6. Análisis de Redes Sociales para la Minería de Texto (IntelDig, 2019)

2.2.2. Importancia de Minería de Texto

- La Minería de Texto ayuda en la toma de decisiones mejor e inteligente.
- Ayuda a resolver problemas de conocimiento en diferentes áreas de negocio.
- Se puede visualizar los datos obtenidos a través de gráficos, cuadros. etc.
- La Minería de Texto es utilizada por organizaciones grandes y pequeñas que están basadas en conocimiento.
- Brinda un mejor resultado preciso y confiable que otras herramientas de software. (Botta-Ferret & Cabrera-Gato, 2017)

2.2.3. Modelo de Minería de Texto

La Minería de Texto muestra los siguientes elementos englobados:

Mencionamos como funcionaria La Minería de Texto donde decimos que se dividen en 5 pasos:

Figura 7. Modelo de cómo funciona la Minería de Texto (COMMUNICATIONS, 2018)

1. *Recopilación*: En este proceso hay diferentes recursos como sitio web, correos electrónicos, etc. Donde es motorizado y dirigido por alguien encargada de ejecutar el proceso

2. *Preprocesamiento:* El reconocimiento de argumento y el arrancamiento de una singularidad representativa

3. *Precisión de textos*: La exclusión de información insignificante o no preferida, tales como la publicidad de páginas

4. *Tokenización*: Se forma una cadena de frases, sin identificación de palabras o argumentos, donde rompe el texto en sociedades importantes tales como (frases, palabras, textos, etc.)

5. *Procedencia de características* (también llamada selección de atributos): es el proceso de caracterización. Es el proceso final de determinación y selección

(Jiakang, Christian O' Reilly, Gareth Owen, & Oudenhoven, 2018)

2.2.4. Clasificación de Minería de Texto

Es el proceso en el cual se asignan etiquetas o categorías a los textos. Debido a esto es fácil etiquetar una gran cantidad de documentos y obtener resultados en un corto tiempo, sin hacerlo manualmente. Esto se usa en diferentes áreas.

a) Sistemas basados en reglas

Este sistema es basado en reglas lingüísticas. Por reglas, se refiere a asociaciones que crean los humanos entre patrones lingüísticos y etiquetas. Una vez confirmados los algoritmos, los patrones lingüísticos y etiquetas son asignados. Este sistema es fácil de entender, porque las personas los desarrollan y mejoran, pero son

difíciles de escalar, ya que cuando se agrega nuevas reglas, se tiene que probar para que no afecte las demás predicciones.

b) Sistemas basados en aprendizaje automático

Estos sistemas tienen que aprender de datos anteriores, estos datos de entrenamiento deben ser constantes y representativas, para asíse haga predicciones precisas. Las maquinas deben poder entender estos datos y para eso se usa vectores, que son colecciones de números con datos codificados.

Los datos se transforman en vectores junto con las etiquetas, a esto se inserta el algoritmo de aprendizaje automático y se crea el modelo de clasificación.

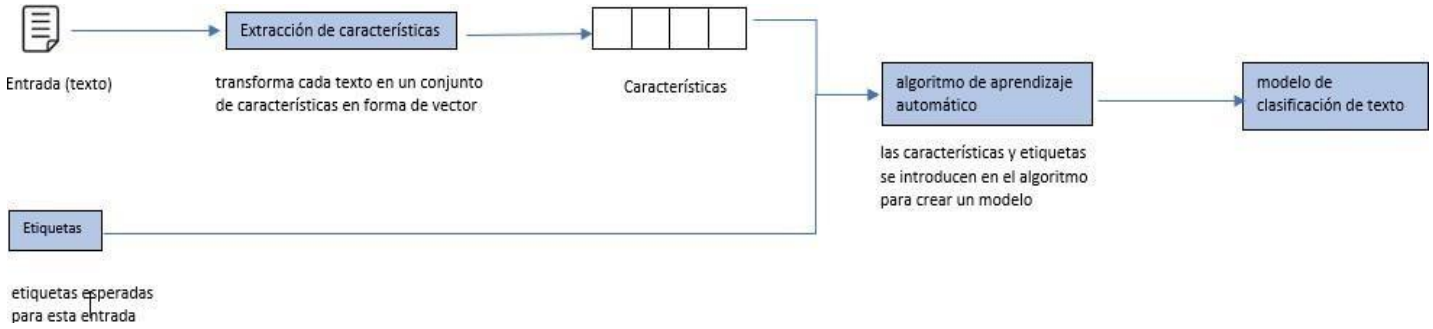


Figura 8. Modelo de Transformación de Minería de Texto(Learn, 2019)

Después, el modelo entrante hace la extracción de las características relevantes del nuevo texto invisible y hacer predicciones sobre información invisible.

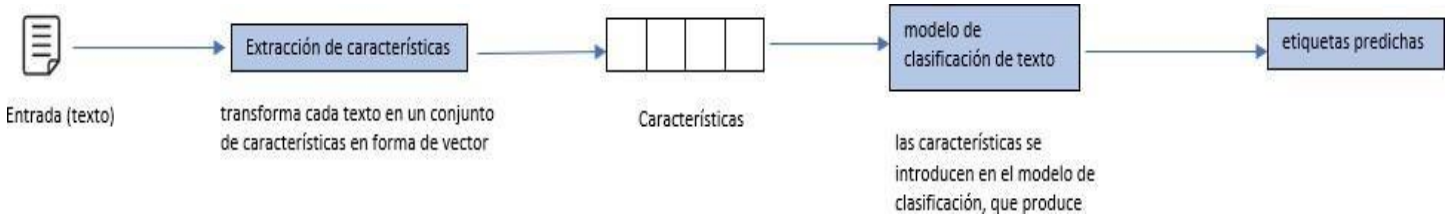


Figura 9. Modelo de Extracción de Características de Minería de Texto (Learn, 2019)

c) Algoritmos basados en aprendizaje automático

- Naive Bayes (NB): Utilizan el teorema de Bayes y la teoría probabilística para predecir la etiqueta de un texto. Los vectores descifran la información basados en la probabilidad de palabras de un texto que corresponden a las etiquetas de un modelo. Es modelo es aplicado cuando no hay muchos datos.

- Máquinas de Vectores de Soporte (SVM): Este modelo clasifica los vectores en dos grupos diferentes. Uno con casi todos los vectores que son de una etiqueta determinada y otro con los vectores que no pertenecen a esa etiqueta. Con este algoritmo se obtiene mejor resultado que con los algoritmos Naive Bayes, pero se necesita más codificación para realizar este modelo.

- Aprendizaje Profundo: son parecidos al cerebro humano. Se usan millones de ejemplos de entrenamiento, lo cual produce representaciones muy detalladas y sistemas extremadamente precisos. (Learn, 2019)

d) Sistemas híbridos

Son modelos que combinan los sistemas basados en reglas y los sistemas de aprendizaje automático, para aumentar la precisión de los resultados.

e) Métricas y Evaluación

El rendimiento de un clasificador de texto se mide en diferentes parámetros:

- ✓ Exactitud: Muestra el número de predicciones acertadas dividido entre el número total de predicciones. Pero a veces la precisión sola no es la mejor, porque existen varios ejemplos para una categoría que, para otra, esto se convierte en una paradoja de precisión, para este casoes

más útil usar otras métricas como precisión y recuperación.

- ✓ Precisión: Realiza la evaluación de un numero de predicciones correcta realizada por el clasificador entre el número total de predicciones para una etiqueta dada (incluida las predicciones correctas o incorrectas).
- ✓ Recuperación: Da como resultado el número total de textos que se predijeron de forma correcta sobre el número total que debería categorizar una etiqueta determinada. Esta métrica es útil cuando se necesita enrutar tickets de soporte a los equipos correctos.
- ✓ Puntación F1: Esta métrica combina la precisión y recuperación, para saber si el clasificador está funcionando bien. (Learn, 2019)

Esta en una de las tareas fundamentales de procesamiento de lenguaje natural. Las empresas están utilizando la clasificación de texto para mejorar la toma de decisiones y automatizar sus procesos (Learn, 2019)

> La minería de textos es una tecnología de recuperación y organización de la información que, aunque todavía es emergente y necesita ser mejor desarrollada, nos sirve para obtener un tipo de información muy útil en cualquier tipo de organización pública o privada.
>
> Económicamente es una técnica que puede utilizarse para ahorrar dinero y abrir oportunidades de negocio a las empresas.

2.3. Definiciones de Términos

2.3.1. Lenguaje R:

R es un lenguaje y entorno para computación estadística y gráficos. Es un proyecto GNU que es similar al lenguaje S. R proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no

lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento, ...) y gráficas, y es altamente extensible.

(¿Qué es la R?, 2016)

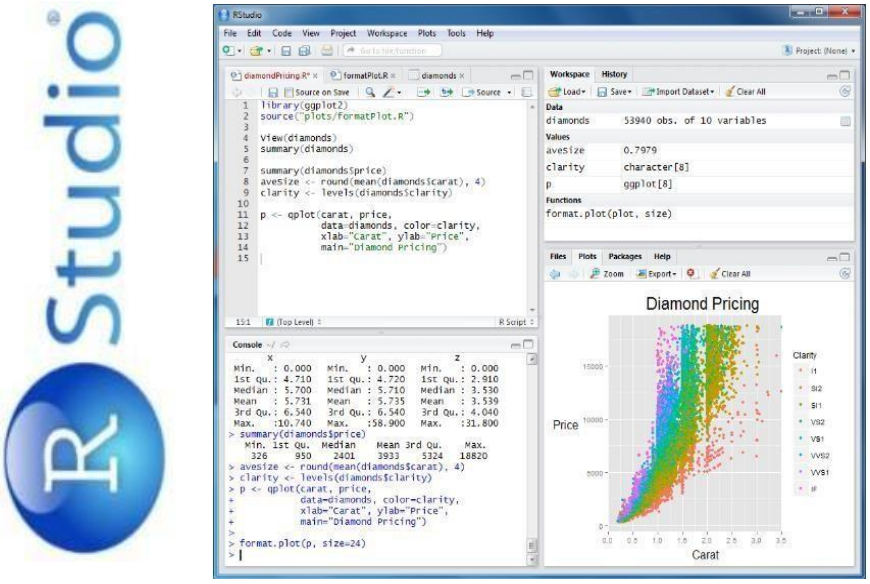


Figura 10. Lenguaje R (von, 2016)

2.3.2. Web Scraping:

El Web Scraping consiste en navegar automáticamente una web y extraer de ella información. Esto puede ser muy útil para muchísimas cosas y beneficioso para casi cualquier negocio. A día de hoy, no creo que exista una sola empresa de éxito que no lo haga o que no quiera hacerlo. De hecho, la empresa reina del scrapeo es Google, que para que su buscador funcione así de bien tiene que estar constantemente scrapeando la red entera. (Lafuente, 2019)

Figura 11. Web Scraping (Jarrell, 2018)

2.3.3. Dashboard:

Un Dashboard es una representación gráfica de las principales métricas o KPIs que intervienen en la consecución de los objetivos de una estrategia de Inbound Marketing.

Esta herramienta nos permite visualizar el problema y favorecer la toma de decisiones orientada a mejorar los posibles errores que podamos estar cometiendo. El fin último es transformar los datos en

información útil para orientar nuestra estrategia hacia la consecución de los objetivos planteados. (¿Qué es un dashboard?, s.f.)

Figura 12. Dashboard (SysAid, 2018)

2.3.4. Machine Learning:

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este

contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana. (Gonzalez, 2014)

Figura 13. Machine Learning (EUROFORUM, 2017)

2.3.5. Text2vec:

Text2vec Es Un Paquete R Que Proporciona Un Marco Eficiente Con Una API Concisa Para El Análisis De Texto Y El Procesamiento Del Lenguaje Natural (PNL).

Objetivos que pretendíamos alcanzar como resultado del desarrollo de text2vec:

- **Conciso**: exponga la menor cantidad de funciones posible
- **Consistente**: exponga interfaces unificadas, no es necesario explorar una nueva interfaz para cada tarea
- **Flexible**: permite resolver fácilmente tareas complejas
- **Rápido**: maximice la eficiencia por hilo único, escale de forma transparente a varios hilos en máquinas multinúcleo

- **Memoria eficiente**: use secuencias e iteradores, no mantenga los datos en RAM si es posible (Selivanov, 2016)

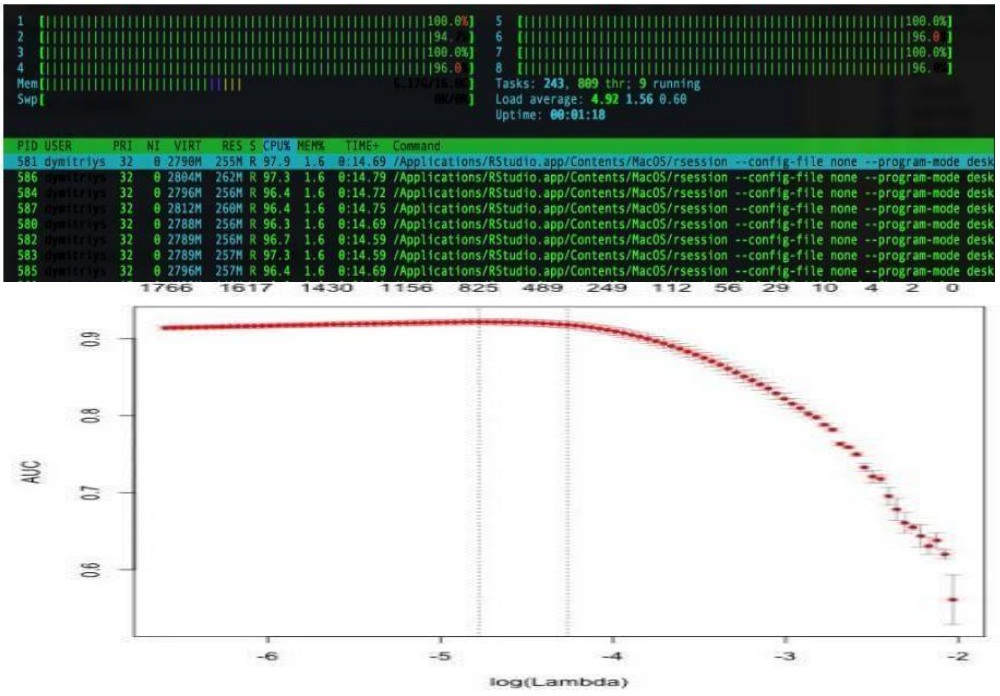


Figura 14. Text2vec (Selivanov, text2vec, 2016)

2.3.6. Tf-idf:

Tf-idf (del inglés Term frequency – Inverse document frequency), frecuencia de término – frecuencia inversa de documento (o sea, la frecuencia de ocurrencia del término en la colección de documentos), es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto. (Yatsko, 2019)

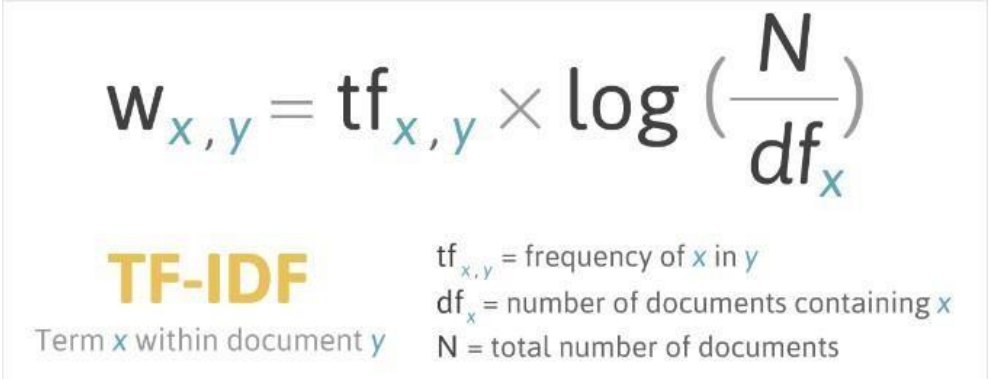


Figura 15. Tf-idf (Opengate, 2016)

2.3.7. Redes Neuronales Recurrentes:

Una red neuronal recurrente (RNN) es un tipo de red neuronal artificial que se usa comúnmente en el reconocimiento de voz y el procesamiento del lenguaje natural (PNL). Los RNN están diseñados para reconocer las características secuenciales de los datos y usar patrones para predecir el próximo escenario probable.

Los RNN se utilizan en el aprendizaje profundo y en el desarrollo de modelos que simulan la actividad de las neuronas en el cerebro humano. (Rouse, Redes Neuronales Recurrentes, s.f.)

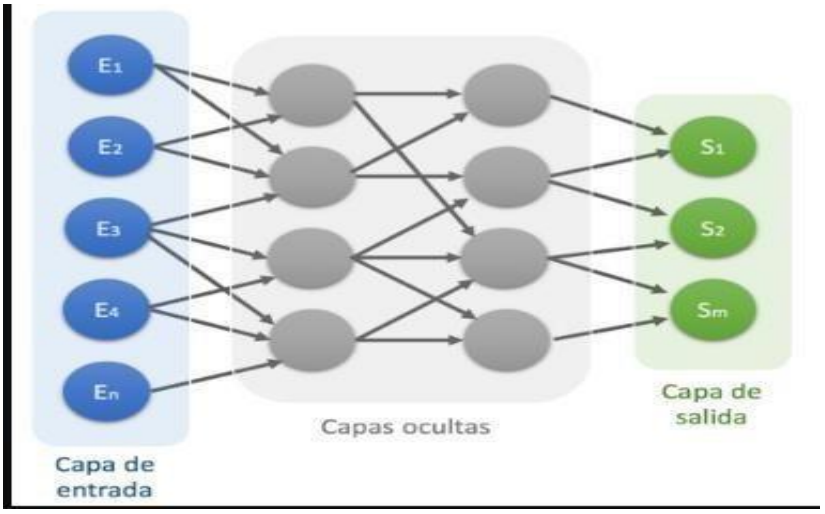


Figura 16. Redes Neuronales Recurrentes (Calvo, 2017)

2.3.8. LSTM:

Las LSTM son un tipo especial de redes recurrentes. Lacaracterística principal de las redes recurrentes es que la información puede persistir introduciendo bucles en el diagrama de la red, por lo que, básicamente, pueden «recordar» estados previos y utilizar esta información para decidir cuál será el siguiente. Esta característica las hace muy adecuadas para manejar series cronológicas. (GARZÓN, 2018)

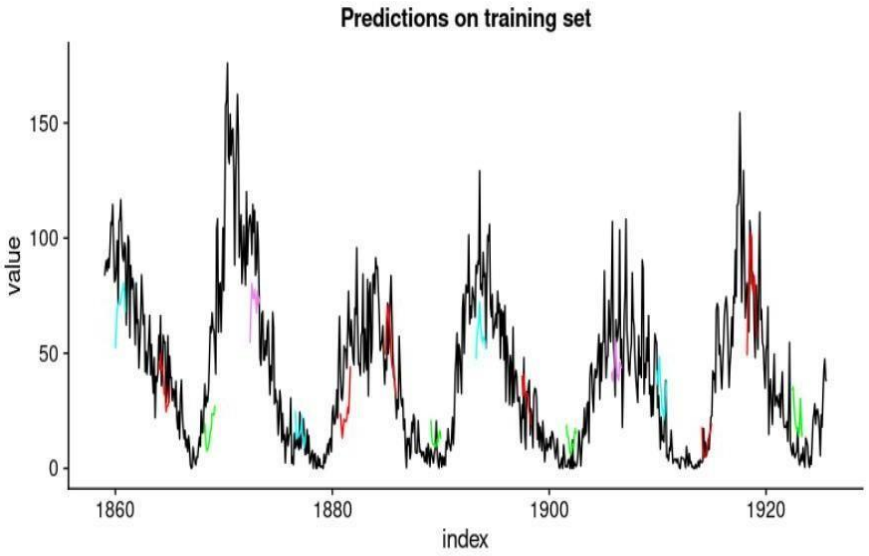


Figura 17. LSTM (Falbel, 2019)

2.3.9. FASTTEXT:

Es una biblioteca para el aprendizaje de incrustaciones de palabras y clasificación de texto creada por el laboratorio de investigación de AI (FAIR) de Facebook . El modelo permite crear un algoritmo de aprendizaje supervisado o de aprendizaje no supervisado para obtener representaciones vectoriales de palabras. Facebook pone a disposición modelos preentrenados para 294 idiomas, FastText utiliza una red neuronal para incrustar palabras. (Facebook, 2018)

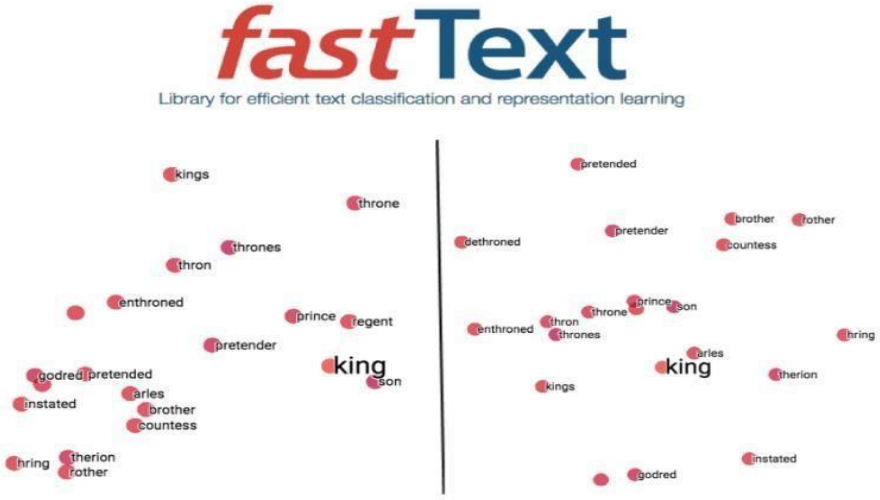


Figura 18. FASTTEXT (Fasttext, 2018)

2.3.10. EXPRESIONES REGULARES REGEX (RegEx):

RegEx es una herramienta en línea para aprender, construir y probar expresiones regulares (RegEx / RegExp).

- ✓ Admite JavaScript y PHP / PCRE RegEx.
- ✓ Los resultados se actualizan en tiempo real a medida que escribe.
- ✓ Pase el cursor sobre una coincidencia o expresión para obtener detalles.
- ✓ Guarda y comparte expresiones con otros.
- ✓ Use Herramientas para explorar sus resultados.
- ✓ Referencia completa de RegEx con ayuda y ejemplos.
- ✓ Deshacer y rehacer con ctrl-Z / Y en editores.
- ✓ Busque y califique los patrones de la comunidad

(gSkinner, s.f.)

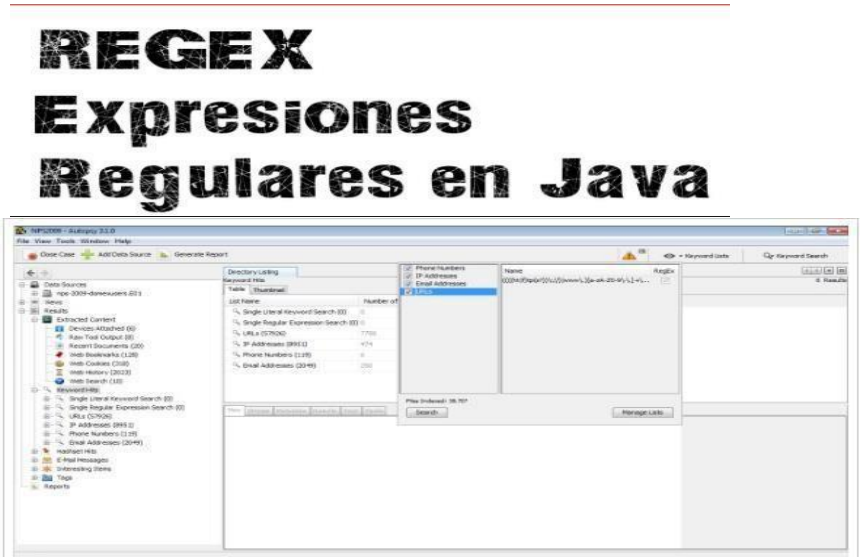


Figura 19. Expresiones Regulares Regex (ReYDeS, 2014)

2.3.11. WORD EMBEDDING:

Las incrustaciones de palabras son básicamente una forma de representación de palabras que une la comprensión humana del lenguaje con la de una máquina. Las incrustaciones de palabras son representaciones distribuidas de texto en un espacio n-dimensional. Estos son esenciales para resolver la mayoría de los problemas de PNL.

La adaptación del dominio es una técnica que permite que los modelos de Aprendizaje automático y Aprendizaje de transferencia mapeen conjuntos de datos de nicho que están escritos en el mismo idioma pero que aún son lingüísticamente diferentes. (Gupta, 2019)

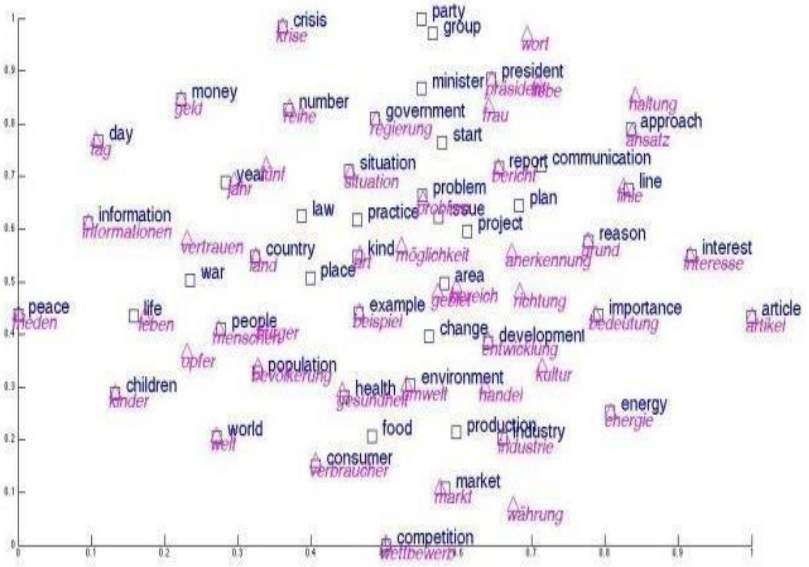


Figura 20. Word Embedding (Collis, 2017)

2.4. Metodología Ágil

Si hablamos de la Metodología Ágil decimos que es una de las Tecnologías de la Información que se una para explicar un método diferente de Gestión de Proyectos.

La Metodología Ágil es un procedimiento que acepta al equipo a entregar respuestas apresuradas a las formas que obtienen los proyectos, permite obtener oportunidades de calificar los Proyectos mediante su ciclo de desarrollo, esto es conocido cuando se evalúa un proyecto regular como Sprint o Iteraciones.

La Metodología Ágil es un procedimiento de empoderamiento que apoya a empresas a diseñar y elegir el producto adecuado. Es un proceso de gestión muy satisfactorio para las grandes compañías de software porque concede analizar y modernizar su producto mediante el desarrollo del mismo, esto permite que empresas tengan la capacidad de elaborar un producto provechoso, de manera eficiente para la competitividad en el mercado. (Gonçalves, 2019)

Figura 21. Plan de la Metodología Ágil con sus principios (Rodriguez, 2015)

2.4.1. Objetivos de la Metodología Ágil

Gestionar proyectos a través de Metodología Agiles tiene como objetivo bridar garantías a las 4 peticiones de la Industria que ha generado como son: Valor, Reducir tiempo del desarrollo, agilidad, garantizar calidad y satisfacción del cliente. (Rodriguez, 2015)

Figura 22. Metodología Ágil adaptada al Plan de Trabajo. (Marín, 2018)

2.4.2. Características de Metodología Ágil

- Agradar al cliente mediante su entrega de producto rapido, funcional y continuo.
- Cambio en los requerimientos permitidos
- Equipos auto organizados
- Entregar en el menor tiempo posible
- El equipo comprometido entre desarrollo y cliente deben trabajar durante todo el proyecto.
- El dialogo dentro del equipo siempre es eficiente y efectivo.
- Sencillez
- Aumentar Productividad
  Divulgación y Traspasó del conocimiento (Rodriguez, 2015)

2.4.3. Ventajas de Metodología Ágil

- Respuestas precisas y ágil para los cambios de requisitos durante el proyecto
- Reduce Costos
- Mayor velocidad y eficacia
- Identifica errores en el transcurso que se esta haciendo pruebas a medida que se avanza
- El equipo de desarrollo conoce el estado del Proyecto

- Aumenta la calidad del Producto
  (Rodriguez, 2015)

2.4.4. Metodología Ágil más usadas

- Programación Extrema (XP): Es adaptada a llevar limites extremos algunos fundamentos y actividades en la forma activa tradicional al momento de Programar

- Scrum: Metodología Ágil más usada y eficiente a los proyectos, que cuenta con una organización sin fines de lucro que lo difunde.

- Kanban: Es un Sistema de Información que dirige de modo armónico la elaboración de los productos necesarios con la cantidad y el tiempo que necesitan los procesos en el interior de una fábrica.

- Open Up: Proceso Ágil y rápido, que es un enfoque didáctico dentro de un ciclo de vida estructurado que tiene un conjunto de prácticas que permite ayudar al equipo a ser más eficiente en el desarrollo del software. (Rodriguez, 2015)

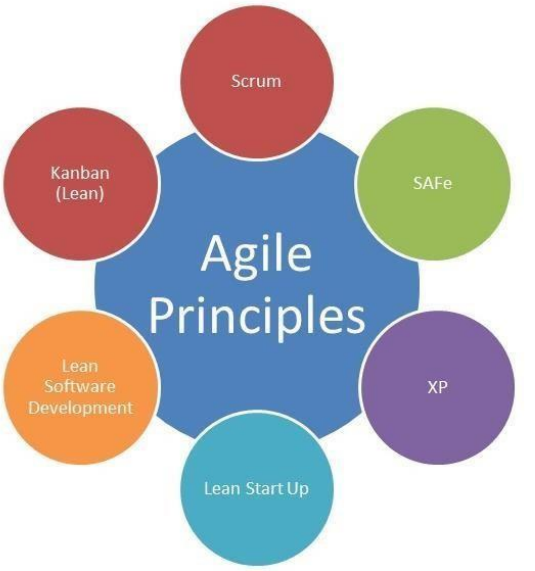


Figura 23. Metodología Ágil más usadas (Admin, 2017)

2.4.5. Proceso Ágil

El proceso ágil divide un proyecto de software en pequeñas partes, que son desarrolladas con incrementos o iteraciones, donde nos dice que reduce el tamaño del proyecto, elaborando muchos proyectos cortos esto distingue de los de más métodos agiles. En conclusión, es un programa de software que hace la entrega más rápida al cliente, donde pueden modificar sus cambios necesarios durante el proceso, lo que nos dice que el proceso iterativo se repite hasta cuando finalice el proyecto.

2.4.6. Scrum

Scrum es una Metodología Ágil para la gestión de proyectos de software, está compuesta por una serie de reglas, roles, artefactos y tiempos establecidos para el desarrollo de los proyectos. Estos pasos que se establecen deben respetarse y cumplirse.

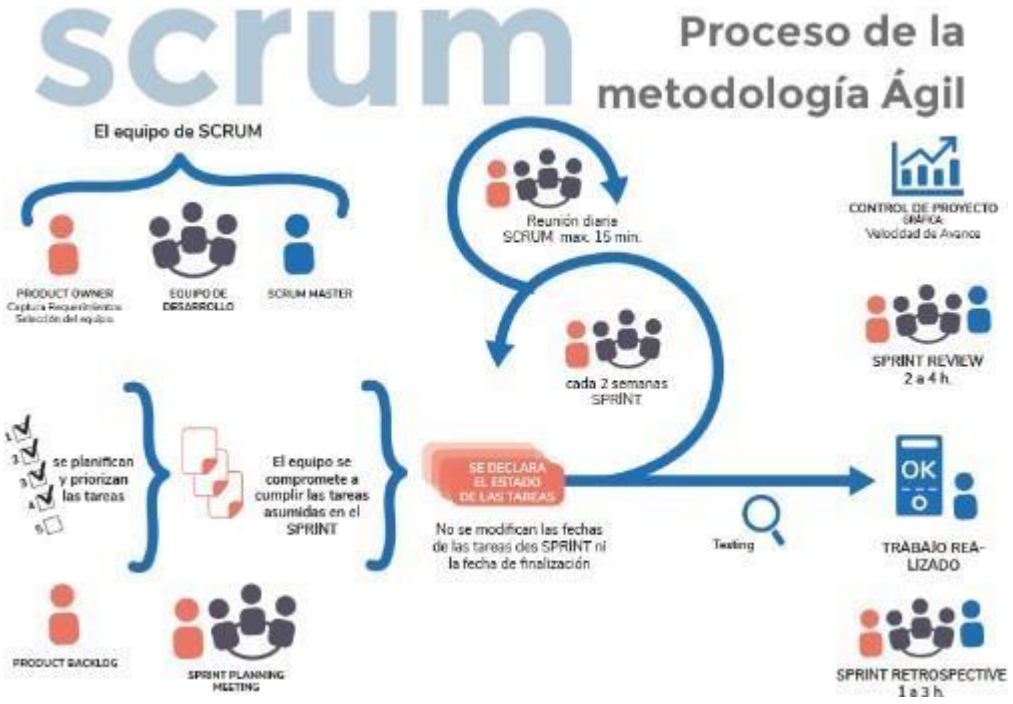


Figura 24. Metodología Scrum en Proyecto de Desarrollo (Garatu, s.f.)

2.4.6.1. Roles que establece Scrum

2.4.6.1.1. El Project Owner o Dueño del Producto

El Project Owner representa al cliente, usuarios del software y todas las partes interesadas en el producto. su función consiste en escuchar a los clientes y transmitir eso al equipo de desarrollo. Se encarga de que el proyecto se desarrolle acorde con la estrategia del negocio. es la persona que revisa el producto y puede ir ajustando las funcionalidades para obtener un mayor valor.

Tiene que tener una buena comunicación con las personas, conocimiento del negocio y poder analizar la relación costo/beneficio.

2.4.6.1.2. El Scrum Master

Es aquella persona capaz de trabajar con el equipo y ser parte de él, siendo capaz de ayudarlos en el proceso de desarrollo, pero sin estar al mando de este.

Ayuda al grupo de trabajo en los conflictos que se presenten en el desarrollo del proyecto, asegurando que el equipo aplique de forma correcta la Metodología Scrum, y motivando al equipo de desarrollo creando un buen clima laboral.

2.4.6.1.3. El Equipo de Desarrollo o Scrum Team

Son los encargados de crear el producto, conformado por diseñadores, arquitectos, programadores, etc.

Deben colaborar, con sus compañeros en algún problema que ellos necesiten y motivarlos para que puedan realizar un buen trabajo.

2.4.6.2. Los pilares que provee Scrum permiten que el desarrollo del proyecto sea eficiente:

- **Mejora la motivación e implicación del equipo**: Con el uso de la metodología ágil se permite a todo el equipo conocer el estado del proyecto en todo momento. Las ideas de todos los miembros se tienen en cuenta.

- **Mayor velocidad y eficacia:** Permiten hacer entregas del producto, por lo que se puede entregar una versión funcional del producto en breve intervalo de tiempo. Ayuda a mantener al proyecto por buen camino, dentro del alcance, respetando sus límites establecidos a través de la Ágil.

- **Mejora la calidad del producto:** Al trabajar con esta metodología hace que cada entrega del proyecto se detecten fallos o errores que tenga, y aplicar soluciones antes de su comercialización definitiva

2.4.6.3. Eventos de Scrum

2.4.6.3.1. Sprint

El sprint tiene una duración de un mes o menos tiempo. en este periodo de tiempo se crea un producto que se puede utilizar y desplegar. una vez terminado un sprint, se comienza inmediatamente con el siguiente sprint.

Los Sprints contienen la planificación del Sprint (Sprint Planning), los Scrum Diarios (Daily Scrum), el trabajo de desarrollo, la Revisión del Sprint(Sprint Review) y la Retrospectiva del Sprint(Sprint Retrospective).

En desarrollo de un sprint no se puede realizar cambios, los objetivos no disminuyen y el alcance se puede negociar entre el dueño de producto y el equipo de desarrollo.

2.4.6.3.2. Planification de Sprint (Sprint Planning)

Este evento se crea con todo el equipo de scrum completo, tiene una duración máxima de ocho horas, para un sprint de un mes. Sí es sprint es más corto, el tiempo también se reduce. el Scrum Master es el encargado de sprint se logre en el tiempo establecido y que los asistentes entiendan el objetivo del sprint.

Este evento responde a las siguientes preguntas:

- ¿Qué se puede entregar en el incremento resultante del Sprint que comienza?

- ¿Como se conseguirá hacer el trabajo necesario para entregar el incremento?

2.4.6.3.3. Objetivo del Sprint (Sprint Goal)

El objetivo del sprint, proporciona una guía al equipo de desarrollo, con esto se puede saber hacia dónde se debe llegar y que herramientas son necesarias para lograr esa meta. si el trabajo que se está realizando es diferente a lo esperado, el equipo de desarrollo colabora con el dueño del producto para negociar el alcance de la lista de pendientes del Sprint.

2.4.6.3.4. Scrum Diario (Daily Scrum)

- El Scrum Diario se realiza todos los días con un tiempo de 15 minutos, en este periodo de tiempo se realiza la revisión de los sprint que se han desarrollado y los que se van a desarrollar a continuación, también el equipo de desarrollo, planifica que es lo que se realizara en las siguientes 24 horas. este evento se realiza a la misma hora y en el mismo lugar todos los días.

- El Equipo de Desarrollo es el encargado de establecer la estructura de la reunión, evalúa como poder llegar al Objetivo del Sprint y crear el Incremento esperado hacia el final del Sprint.

- El Scrum Master orienta al Equipo de Desarrollo a mantener las reuniones en el plazo de 15 minutos, si otras personas están presentes en la reunión, el Scrum Master se asegura de que intervengan en la reunión.

- El Scrum Diario es una reunión beneficiosa para que el Equipo de Desarrollo se integre y se adapte.

2.4.6.3.5. Revision de Sprint (Sprint Review)

Al finalizar cada Sprint se lleva a cabo una revisión, para saber el Incremento. esta revisión se lleva a cabo con la colaboración de los interesados y el Equipo de Desarrollo. En esta revisión se determina las cosas que se pueden realizar para optimizar el valor. Es una reunión informal que tiene con objetivo la retroalimentación de información y la colaboración de las partes interesadas.

El tiempo de cada reunión, el máximo de cuatro horas para un Sprint de un mes. El Scrum Master tiene como propósito de que el evento se lleve a cabo y sé que los asistentes entiendan el propósito, también a que la reunión dese lleve a cabo el tiempo establecido.

2.4.6.3.6. Retrospectiva de Sprint (Sprint Retrospective)

Este evento es una ocasión para el Equipo de Desarrollo de revisar que mejoras se puede hacer durante el siguiente Sprint. Esta Retrospectiva de Sprint se realiza después de la Revisión de Sprint y antes de la siguiente Planificación del Sprint. Es una reunión de máximo tres horas para un Sprint de un mes. el Scrum Master se aseguró de que la reunión se lleve a cabo y que sepan cual es el propósito se este, también de que la reunión sea positiva y productiva. el Scrum Master participa en la reunión como un miembro más, porque en el recae la responsabilidad del proceso de Scrum.

Su propósito es:

- Inspeccionar como ha sido el último Sprint.
- Identificar y ordenar las cosas importantes que salieron bien y los cuales hay que mejorar.

- Crear un plan para implementar las mejorar para el desempeño del Equipo de Desarrollo en su trabajo.

(Schwaber & Sutherland, 2017)

3. MATERIAL Y MÉTODO

3.1. Material

3.1.1. Población

La investigación es realizada a los proyectos de investigación detesis en la Universidad Privada Antenor Orrego.

3.1.2. Muestra

El tipo de muestreo utilizado para la investigación es un muestreo probabilístico, debido a que utilizaremos 85 proyectos de investigación de tesis.

3.2. Método

3.2.1. Tipo de Investigación

Investigación Aplicada.

3.2.2. Diseño de la Investigación

Se aplico el diseño preexperimental postprueba, ya que se trabaja con un solo grupo antes y después

Figura 25. Investigación postprueba

Donde:

A: es la muestra control

A`: es la muestra resultante.

E: es el estímulo.

3.2.3. Variables de Estudio y Operacionalización

3.2.3.1. Variables

Variable Independiente: Dashboard basado en Minería de Texto

Variable Dependiente: Medición de los proyectos de investigación de tesis en la Universidad Privada Antenor Orrego.

3.2.3.2. Operacionalización

VARIABLE INDEPENDIENTE	DIMENSIÓN	INDICADORES	UNIDAD DE MEDIDA	INSTRUMENTO DE INVESTIGACIÓN
Dashboard basado en Minería de Texto	Dashboard para cuantificar los proyectos de investigación de tesis.	- Nivel de facilidad de uso. - Precisión de predicción de proyectos de investigación de tesis.	Numérico	Evaluación de resultados
		Predicción de uso	Numérico	

Tabla 1. Operacionalización de Variables Independiente

VARIABLE DEPENDIENTE	DIMENSIÓN	INDICADORES	UNIDAD DE MEDIDA	INSTRUMENTO DE INVESTIGACIÓN
Medición de los proyectos de investigación de tesis de la Universidad Privada Antenor Orrego.	Reporte de la medición de los proyectos de investigación de tesis.	- Grado de Satisfacción al usar el Dashboard. -Cantidad de proyectos de investigación de tesis que se puede precisar por el aplicativo web.	Número de elementos de precisión.	Evaluación de los términos encontrados son relevantes para el usuario

Tabla 2. Operacionalización de Variables Dependiente

3.2.4. Instrumentos de Recolección de Datos

TÉCNICAS	INSTRUMENTOS
Encuestas	Cuestionario

Tabla 3. Instrumentos de Recolección de Datos

3.2.5. Procedimientos y Análisis de Datos

Una vez recolectado toda la información proporcionada por los instrumentos de recolección y analizar estadísticamente los resultados obtenidos de la información.

Según la investigación realizada usaremos la prueba paramétrica t, para corroborar si la muestra evaluada antes y después difieren. Aplicaremos el método descriptivo, el cual se encarga de la descripción de datos y características de una población. (Anexo 12)

4. RESULTADOS

En este capítulo mostramos los resultados del desarrollo del proyecto de Tesis a través de un Dashboard para la medición de los Proyectos de Investigación de la Universidad Privada Antenor Orrego.

4.1. ¿Por qué hacer un Dashboard con Metodología Ágil Scrum?

El método Ágil es un proceso que permite al equipo dar respuestas rápidas e impredecibles a las valoraciones que reciben sobre su proyecto que se basa en el apoyo y la organización de equipos para mejor el rendimiento. Es un proceso de empoderamiento que ayuda a las empresas a diseñar y crear el producto idóneo. Este proyecto utiliza la Metodología Scrum debido a que proporciona a un fácil manejo del trabajo, ayuda al equipo a obtener mejores resultados. Es una Metodología diseñada para el desarrollo de software, la cual utilizan mucho las empresas de software porque les permite analizar y mejorar su producto durante el desarrollo del mismo, como se puede apreciar en los antecedentes de la investigación. Permite a todo el equipo conocer el estado del proyecto en todo momento, además permite controlar el proyecto de esta manera permite el desarrollo del software durante su ciclo de trabajo.

4.1.1. Artefactos de Scrum

4.1.1.1. Lista de Producto o Product Backlog

Es una lista ordenada de lo que se realizara en el producto y el responsable de esta lista es el Product Owner. Es una lista enumerada que contiene toda las características, funcionalidades, mejoras y requisitos. los elementos que conforman esta lista tienen la descripción, orden, estimación y valor.

La lista de producto cambia constantemente a

medida que el producto y su entorno evolucionan, por eso la lista es dinámica, para saber que producto necesita mejora y de esta forma sea competitivo, adecuado y útil.

Prioridad	Historia de Usuario	Valor	Esfuerzo estimado
1	como administrador del sistema necesito agregar productos al catálogo	10	
2	como usuario del sitio web quiero recorrer el catálogo de productos	10	
3	como cliente de la empresa quiero agregar productos a un pedido	10	
4	como cliente necesito enviar el pedido una vez que haya agregado todos los productos deseados	10	
5	como administrador de pedidos necesito ver la lista de pedidos efectuados por los clientes	7	
6	como administrador de pedidos necesito ver el detalle de los productos solicitados por el cliente en cada uno de los pedidos	10	
7	como administrador de pedidos necesito modificar el estado de cada pedido	3	
8	como cliente quiero poder visualizar el estado de mis pedidos	8	
9	como administrador del sistema necesito clasificar los productos por categorías	3	
10	como administrador del sistema necesito eliminar productos del catálogo	3	
11	como administrador del sistema necesito indicar cuáles productos y cuáles no, son visibles en el catálogo de productos que ven los usuarios	3	
12	como administrador del sistema necesito modificar productos del catálogo	3	
13	como cliente quiero poder modificar la cantidad de items de los productos de mi pedido	5	
14	como cliente de la empresa quiero poder cancelar mi pedido que aún no ha sido despachado	5	
15	como empleado de depósito necesito ver el detalle de los pedidos que aún no han sido enviados a empaque	1	
16	como empleado de depósito necesito indicar que un pedido ya ha sido enviado a empaque	5	
17	como cliente de la empresa necesito poder recuperar mi contraseña cuando la olvido	10	

Figura 26. Lista de Productos

4.1.1.2. Sprint Backlog

Es una lista más pequeña del Product Backlog, que se genera al comienzo de cada sprint y son las características que cada equipo de desarrollo tiene que realizar en un tiempo determinado no mayor a un día.

El Sprint Backlog se actualiza diariamente y muestra:

- Las tareas pendientes, en curso y terminadas.
- El nombre del miembro al que se le dio esa tarea.
- El esfuerzo pendiente de cada tarea sinconcluir.

Estas tareas se visualizan en un tablero, para poder controlar las tareas.

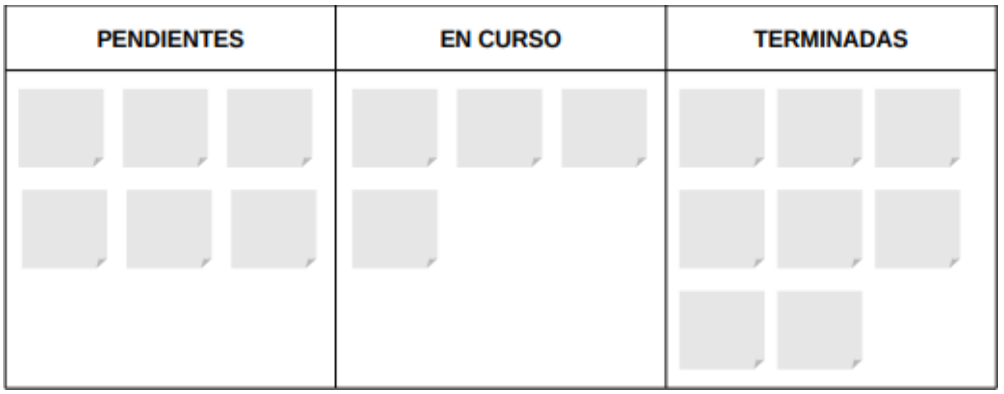


Figura 27. Tablero Físico De Scrum

4.1.1.3. Incremento

Cuando concluye un sprint, el equipo de desarrollo realizara una entrega de un incremento de funcionalidad para el sistema. este incremento debe estar terminado para poder ser implementado y utilizado en producción al 100%. (Bahit, 2015)

4.2. Desarrollo Del Proyecto

4.2.1. Etapas del desarrollo del Dashboard

Es el proceso en el cual nos mostrara paso a paso de la ejecución del Dashboard hasta la entrega final de su prototipado en general.

4.2.1.1. Aplicación de la técnica Web Scraping

En este paso utilizaremos un web scraping para extraer toda la información que nos proporciona el repositorio donde esta todos los proyectos de Tesis de la escuela profesional de Ingeniería de Computación y Sistemas de la Universidad Privada Antenor Orrego.

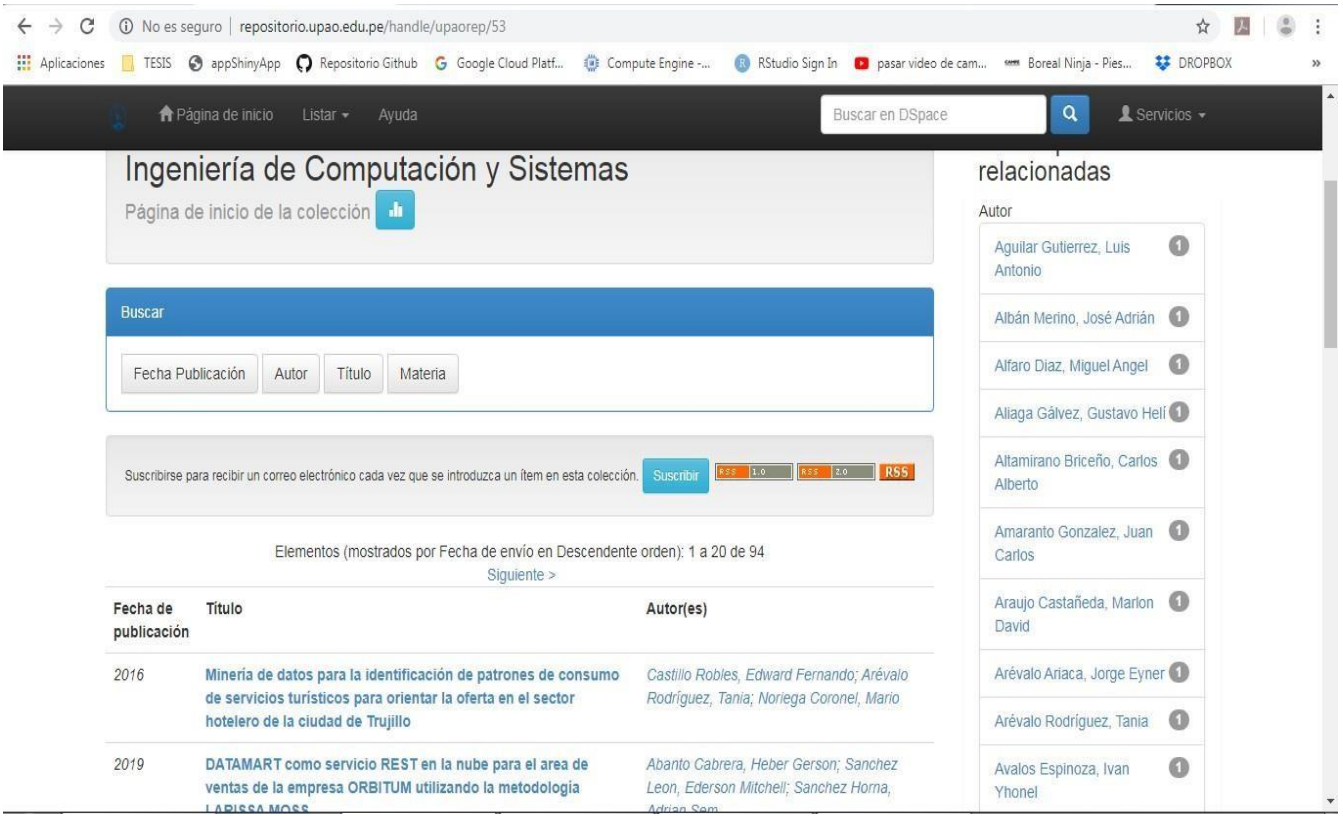


Figura 28. Repositorio de la Escuela Ingeniería de Computación y Sistema de la Universidad Privada Antenor Orrego

En este Scrapper realizamos la importacion de los siguientes paquetes Rvest, Tidyverse, Here. Para la obtener la informacion que necesitamos para realizar un correcto Dashboard.

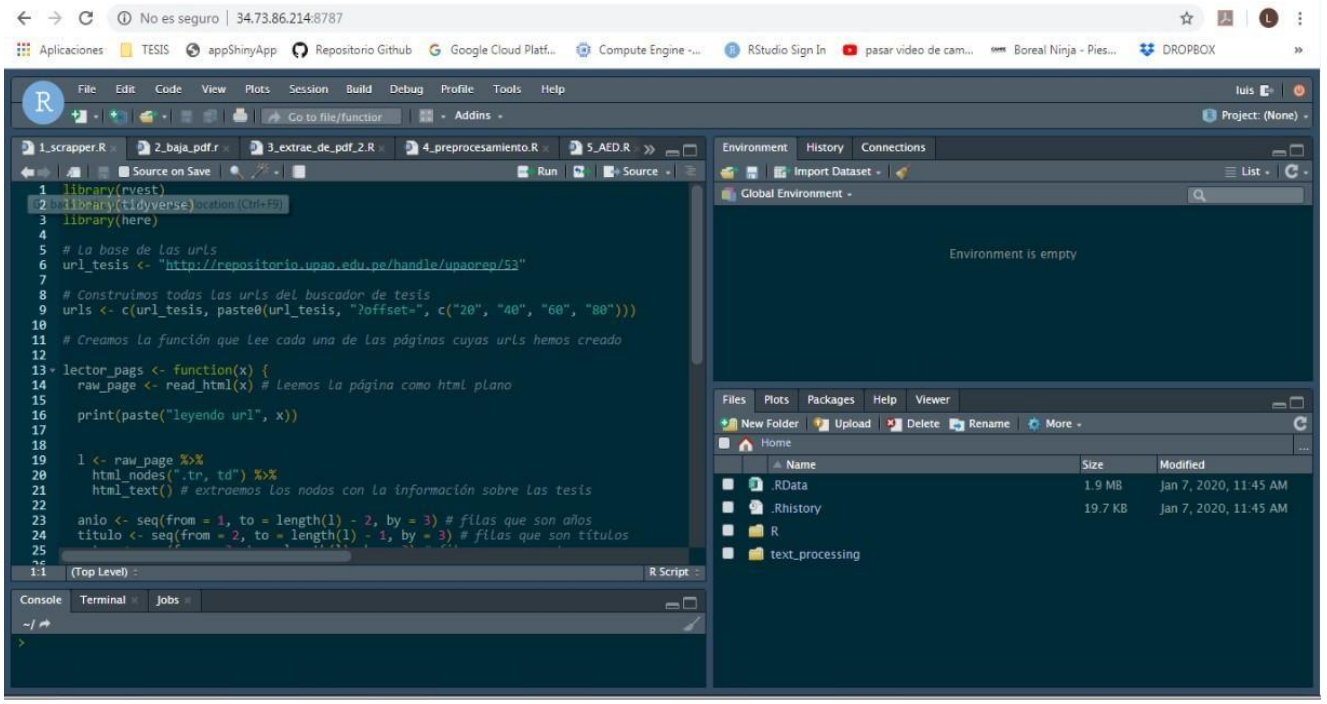


Figura 29. Web Scraping

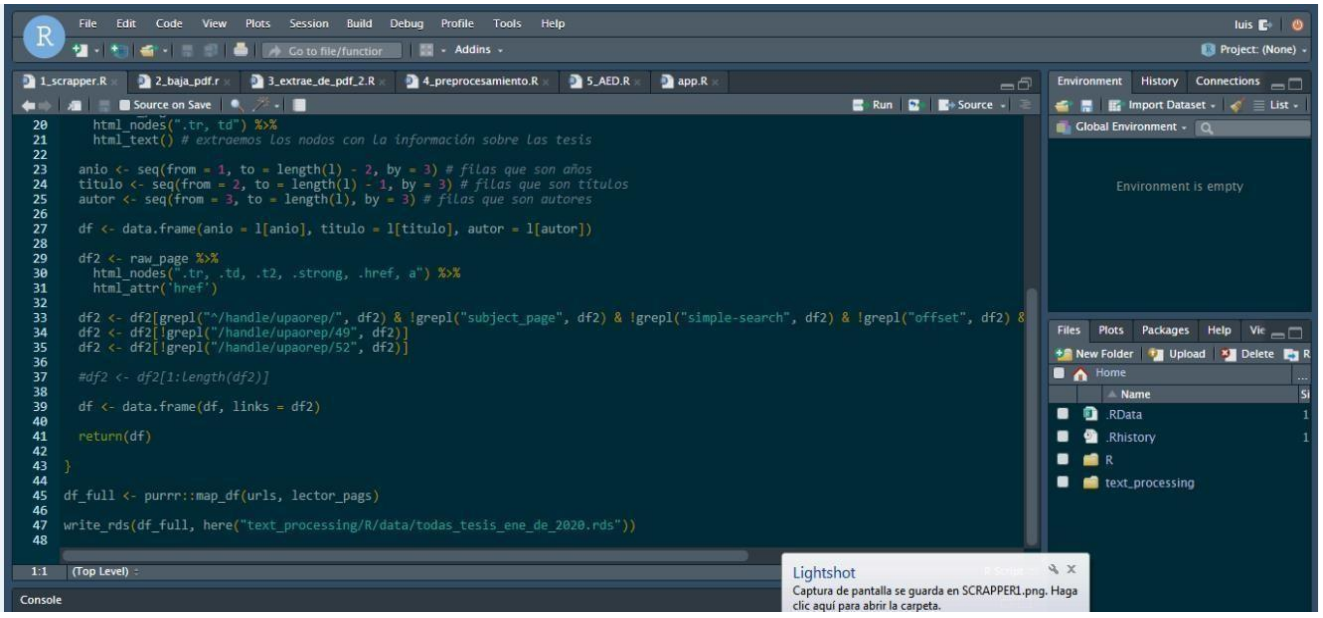


Figura 30. Proceso del Scraping

4.2.1.2. Diseño de Programa para el Dashboard

En este paso realizaremos la descarga de los documentos PDF, utilizando el documento guardado anteriormente por el web scraping, de la escuela de Ingeniería de Computación y Sistemas, el cual nos permitirá saber cuánto es la cantidad obtenida del repositorio.

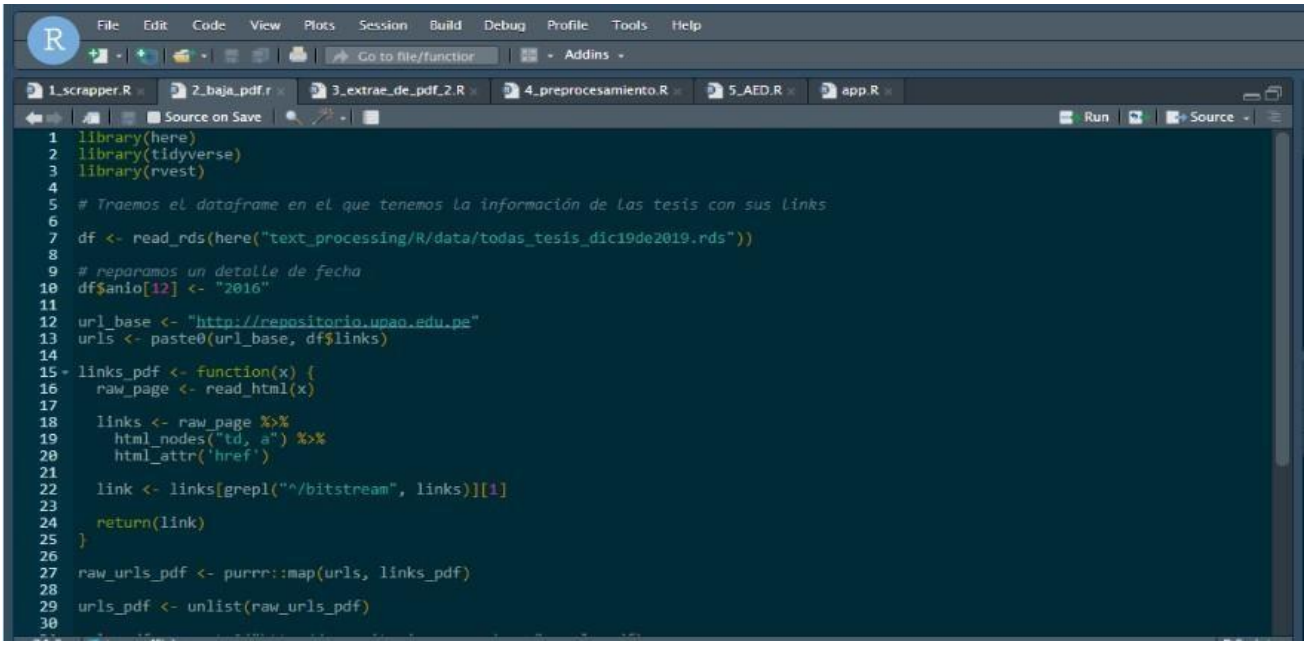


Figura 31. Código para la descarga de los PDF

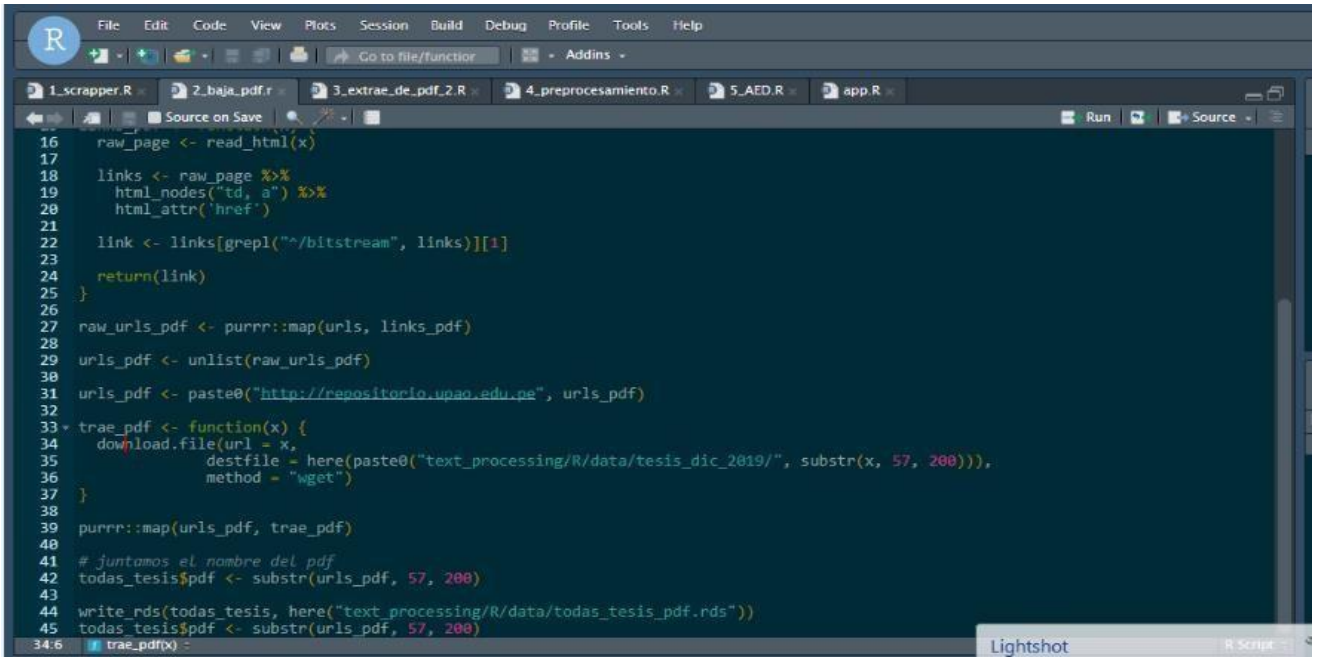
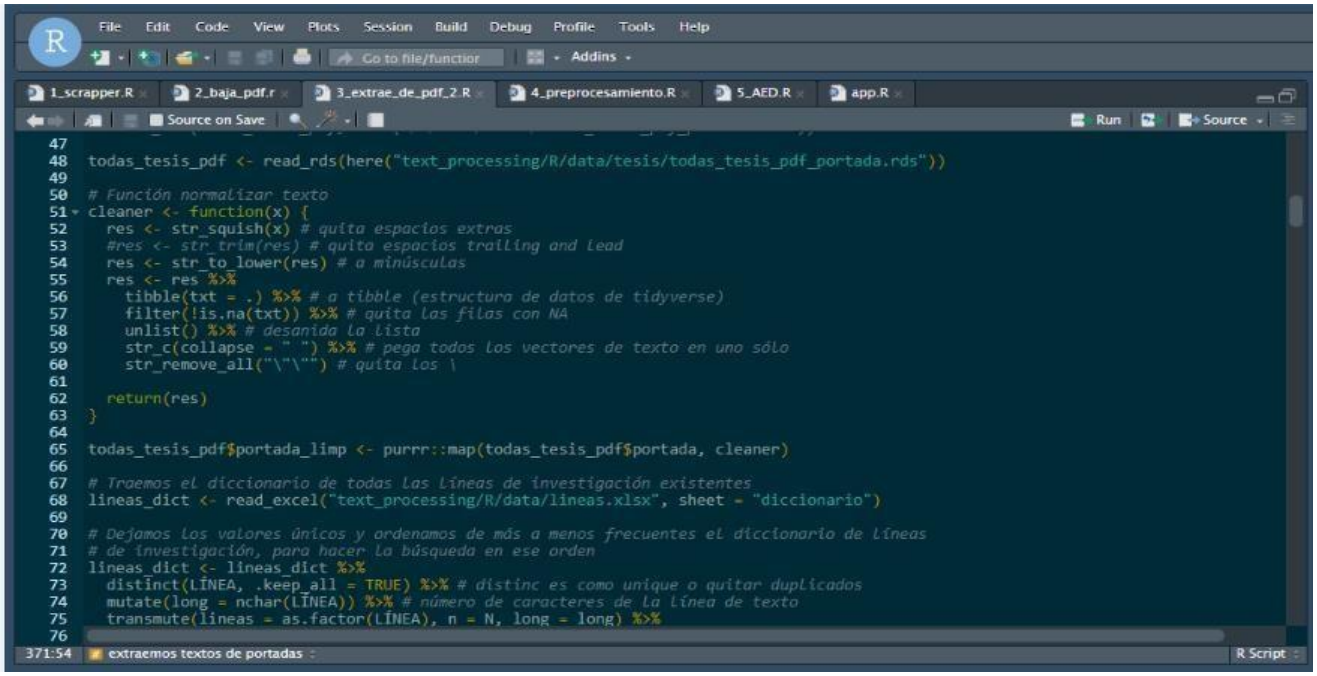


Figura 32. Proceso de extracción de los documentos PDF

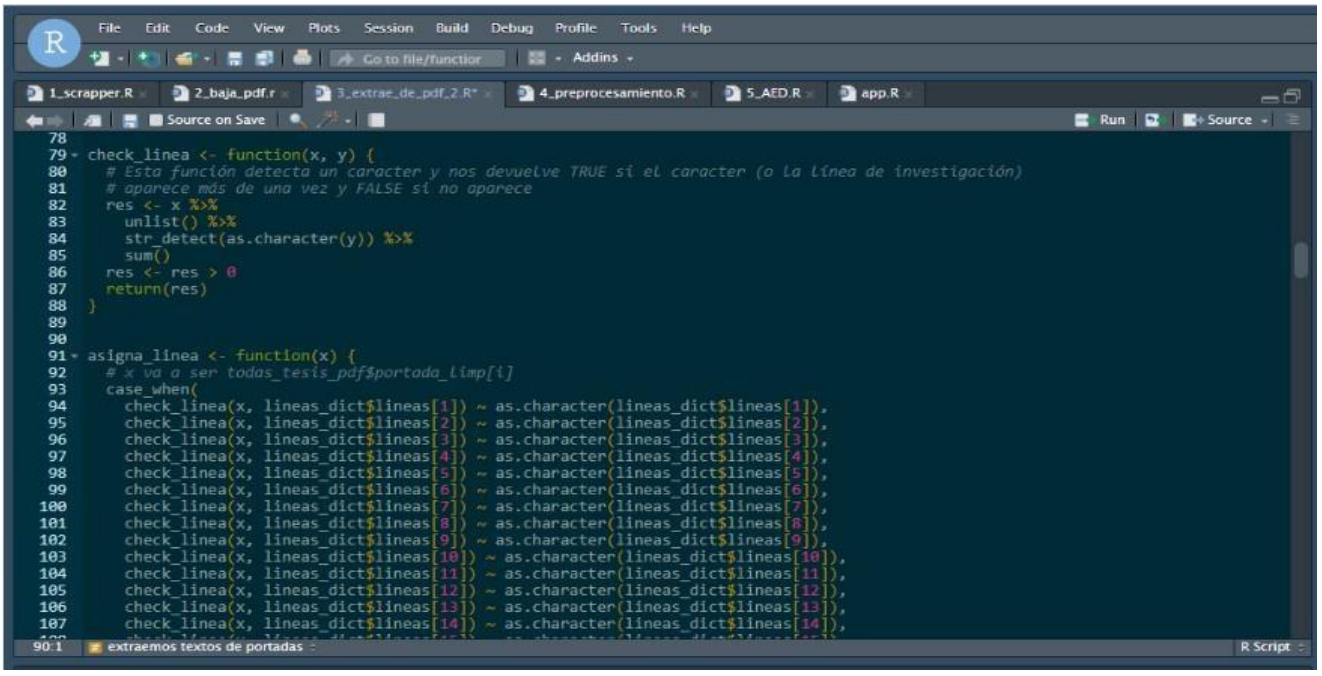
4.2.1.2.1. Extracción De Información De Los Documentos Pdf

Una vez obtenidos todos los PDF de la escuela de Ingeniería de Computación y Sistemas del repositorio, solo utilizaremos la información necesaria de dichos documentos.



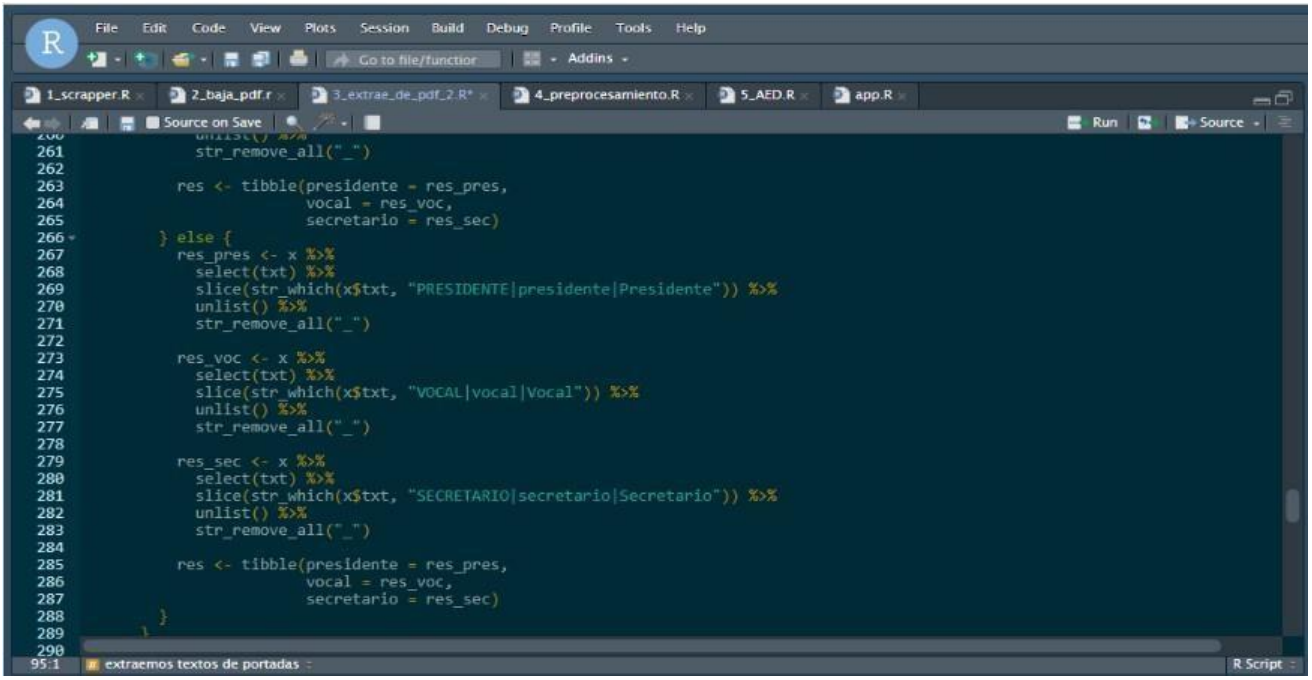
```
todas_tesis_pdf <- read_rds(here("text_processing/R/data/tesis/todas_tesis_pdf_portada.rds"))

# Función normalizar texto
cleaner <- function(x) {
  res <- str_squish(x) # quita espacios extras
  #res <- str_trim(res) # quita espacios trailing and lead
  res <- str_to_lower(res) # a minúsculas
  res <- res %>%
    tibble(txt = .) %>% # a tibble (estructura de datos de tidyverse)
    filter(!is.na(txt)) %>% # quita las filas con NA
    unlist() %>% # desanida la lista
    str_c(collapse = " ") %>% # pega todos los vectores de texto en uno sólo
    str_remove_all("\"\"") # quita los \

  return(res)
}

todas_tesis_pdf$portada_limp <- purrr::map(todas_tesis_pdf$portada, cleaner)

# Traemos el diccionario de todas las líneas de investigación existentes
lineas_dict <- read_excel("text_processing/R/data/lineas.xlsx", sheet = "diccionario")

# Dejamos los valores únicos y ordenamos de más a menos frecuentes el diccionario de líneas
# de investigación, para hacer la búsqueda en ese orden
lineas_dict <- lineas_dict %>%
  distinct(LÍNEA, .keep_all = TRUE) %>% # distinc es como unique o quitar duplicados
  mutate(long = nchar(LÍNEA)) %>% # número de caracteres de la línea de texto
  transmute(lineas = as.factor(LÍNEA), n = N, long = long) %>%
```

Figura 33. Código para la limpieza de los datos extraídos



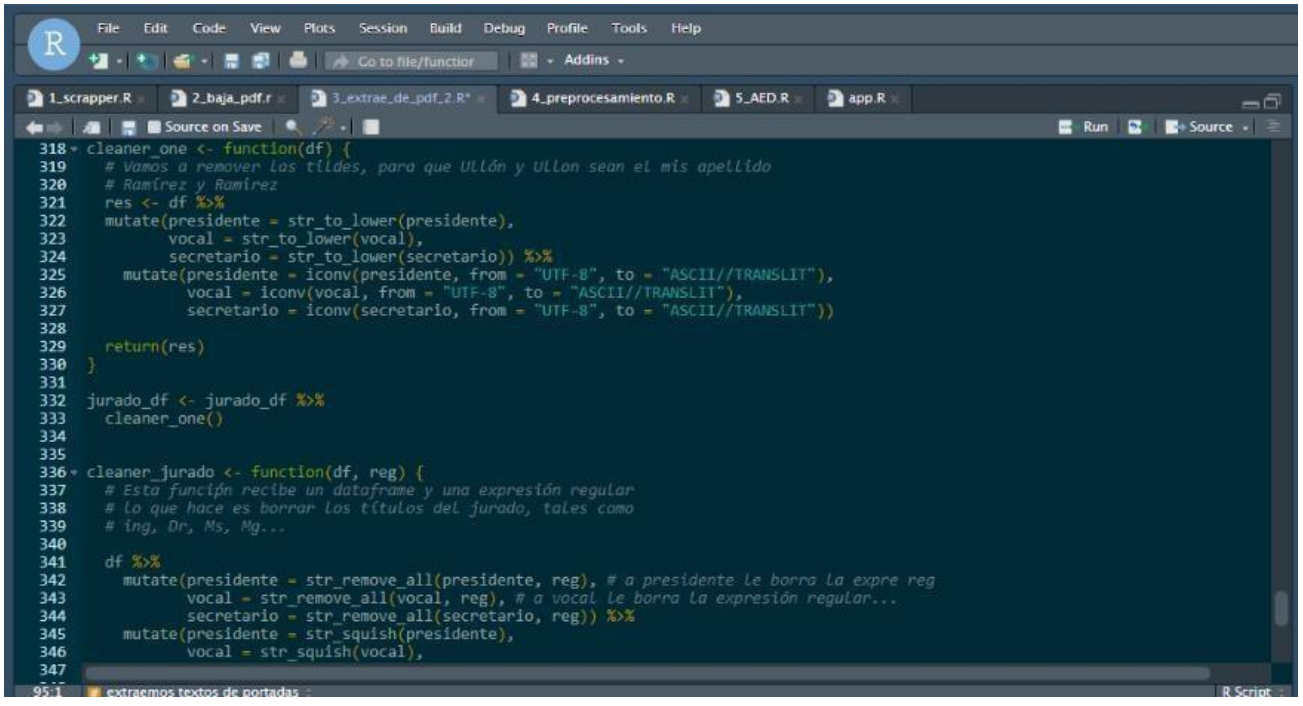
```
check_linea <- function(x, y) {
  # Esta función detecta un caracter y nos devuelve TRUE si el caracter (o la línea de investigación)
  # aparece más de una vez y FALSE si no aparece
  res <- x %>%
    unlist() %>%
    str_detect(as.character(y)) %>%
    sum()
  res <- res > 0
  return(res)
}


asigna_linea <- function(x) {
  # x va a ser todas_tesis_pdf$portada_limp[i]
  case_when(
    check_linea(x, lineas_dict$lineas[1]) ~ as.character(lineas_dict$lineas[1]),
    check_linea(x, lineas_dict$lineas[2]) ~ as.character(lineas_dict$lineas[2]),
    check_linea(x, lineas_dict$lineas[3]) ~ as.character(lineas_dict$lineas[3]),
    check_linea(x, lineas_dict$lineas[4]) ~ as.character(lineas_dict$lineas[4]),
    check_linea(x, lineas_dict$lineas[5]) ~ as.character(lineas_dict$lineas[5]),
    check_linea(x, lineas_dict$lineas[6]) ~ as.character(lineas_dict$lineas[6]),
    check_linea(x, lineas_dict$lineas[7]) ~ as.character(lineas_dict$lineas[7]),
    check_linea(x, lineas_dict$lineas[8]) ~ as.character(lineas_dict$lineas[8]),
    check_linea(x, lineas_dict$lineas[9]) ~ as.character(lineas_dict$lineas[9]),
    check_linea(x, lineas_dict$lineas[10]) ~ as.character(lineas_dict$lineas[10]),
    check_linea(x, lineas_dict$lineas[11]) ~ as.character(lineas_dict$lineas[11]),
    check_linea(x, lineas_dict$lineas[12]) ~ as.character(lineas_dict$lineas[12]),
    check_linea(x, lineas_dict$lineas[13]) ~ as.character(lineas_dict$lineas[13]),
    check_linea(x, lineas_dict$lineas[14]) ~ as.character(lineas_dict$lineas[14]),
```

Figura 34. Proceso para ver la línea de investigación



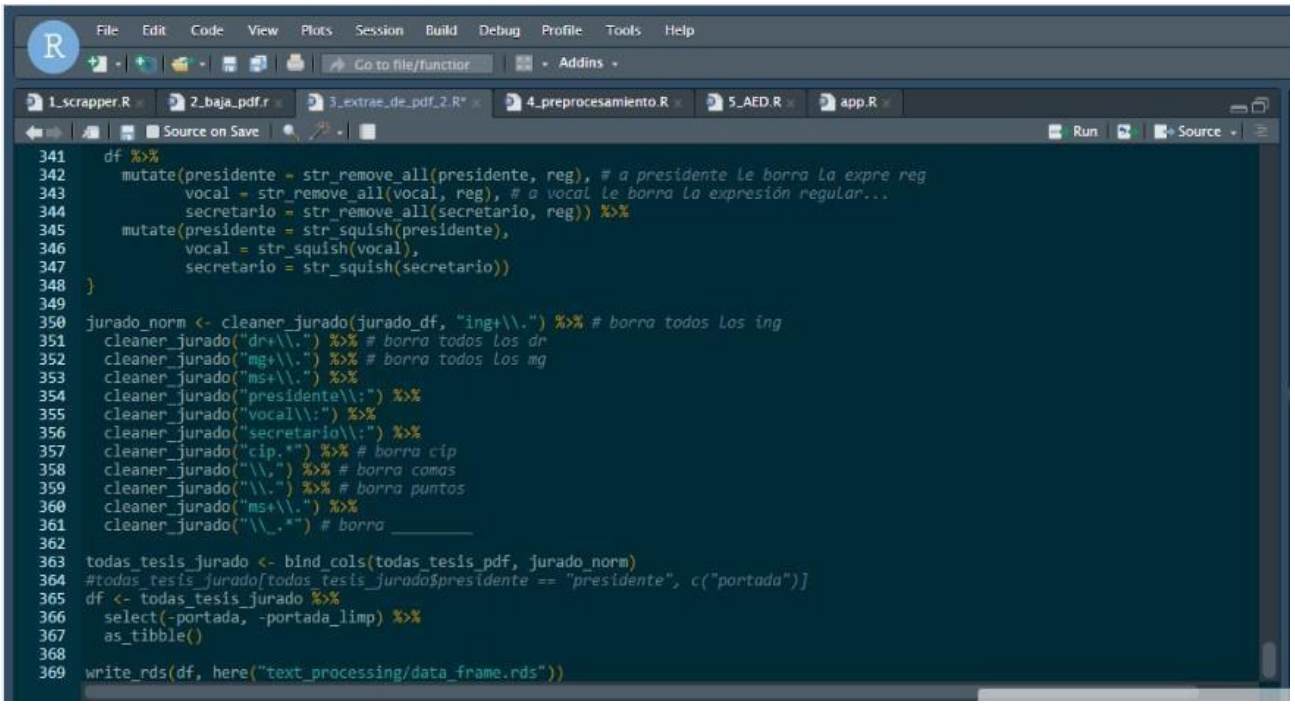
```
261            str_remove_all("_")
262
263          res <- tibble(presidente = res_pres,
264                        vocal = res_voc,
265                        secretario = res_sec)
266        } else {
267          res_pres <- x %>%
268            select(txt) %>%
269            slice(str_which(x$txt, "PRESIDENTE|presidente|Presidente")) %>%
270            unlist() %>%
271            str_remove_all("_")
272
273          res_voc <- x %>%
274            select(txt) %>%
275            slice(str_which(x$txt, "VOCAL|vocal|Vocal")) %>%
276            unlist() %>%
277            str_remove_all("_")
278
279          res_sec <- x %>%
280            select(txt) %>%
281            slice(str_which(x$txt, "SECRETARIO|secretario|Secretario")) %>%
282            unlist() %>%
283            str_remove_all("_")
284
285          res <- tibble(presidente = res_pres,
286                        vocal = res_voc,
287                        secretario = res_sec)
288        }
289      }
290
```

Figura 35. Proceso de búsqueda de Presidente, Vocal, Secretario



```
318 cleaner_one <- function(df) {
319   # Vamos a remover las tildes, para que Ullón y Ullon sean el mis apellido
320   # Ramírez y Ramirez
321   res <- df %>%
322   mutate(presidente = str_to_lower(presidente),
323          vocal = str_to_lower(vocal),
324          secretario = str_to_lower(secretario)) %>%
325     mutate(presidente = iconv(presidente, from = "UTF-8", to = "ASCII//TRANSLIT"),
326            vocal = iconv(vocal, from = "UTF-8", to = "ASCII//TRANSLIT"),
327            secretario = iconv(secretario, from = "UTF-8", to = "ASCII//TRANSLIT"))
328
329   return(res)
330 }
331
332 jurado_df <- jurado_df %>%
333   cleaner_one()
334
335
336 cleaner_jurado <- function(df, reg) {
337   # Esta funcipn recibe un dataframe y una expresión regular
338   # lo que hace es borrar los títulos del jurado, tales como
339   # ing, Dr, Ms, Mg...
340
341   df %>%
342     mutate(presidente = str_remove_all(presidente, reg), # a presidente le borra la expre reg
343            vocal = str_remove_all(vocal, reg), # a vocal le borra la expresión regular...
344            secretario = str_remove_all(secretario, reg)) %>%
345     mutate(presidente = str_squish(presidente),
346            vocal = str_squish(vocal),
347
```

Figura 36. Arreglo de nombres y limpieza todos los datos.

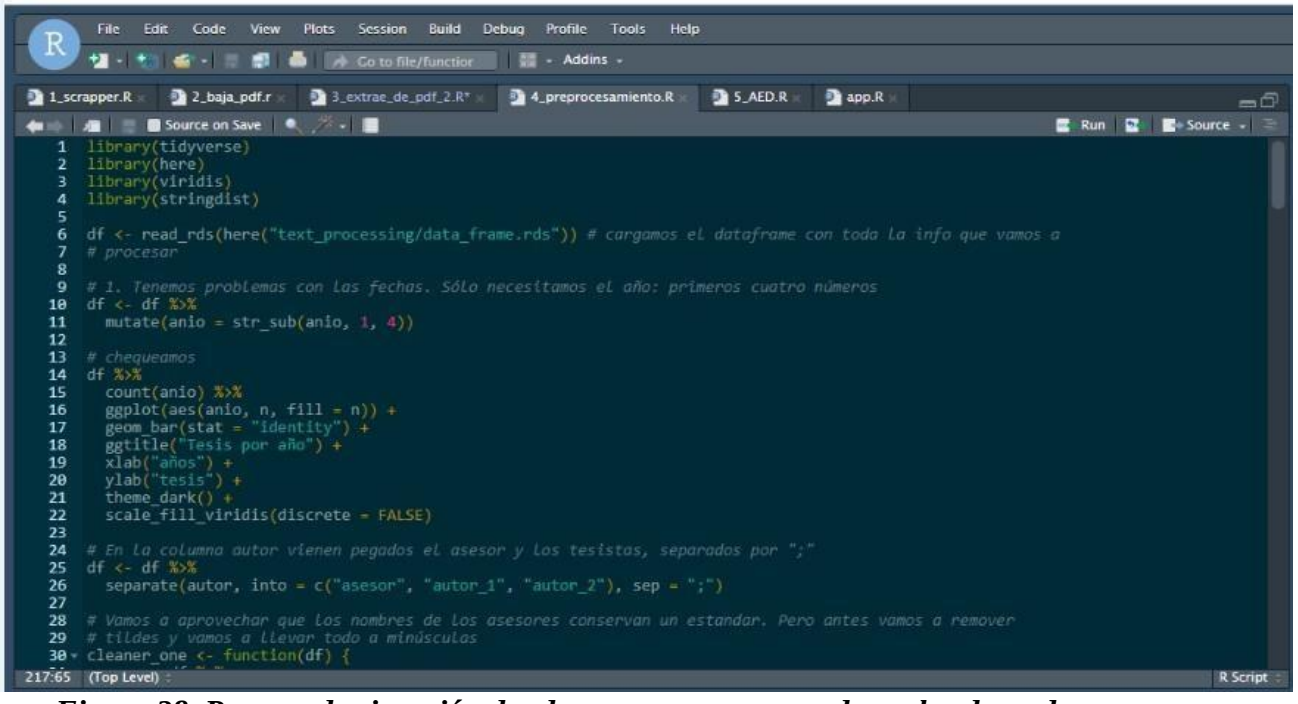


```
  df %>%
    mutate(presidente = str_remove_all(presidente, reg), # a presidente le borra la expre reg
           vocal = str_remove_all(vocal, reg), # a vocal le borra la expresión regular...
           secretario = str_remove_all(secretario, reg)) %>%
    mutate(presidente = str_squish(presidente),
           vocal = str_squish(vocal),
           secretario = str_squish(secretario))
}

jurado_norm <- cleaner_jurado(jurado_df, "ing+\\.") %>% # borra todos los ing
  cleaner_jurado("dr+\\.") %>% # borra todos los dr
  cleaner_jurado("mg+\\.") %>% # borra todos los mg
  cleaner_jurado("ms+\\.") %>%
  cleaner_jurado("presidente\\:") %>%
  cleaner_jurado("vocal\\:") %>%
  cleaner_jurado("secretario\\:") %>%
  cleaner_jurado("cip.*") %>% # borra cip
  cleaner_jurado("\\,") %>% # borra comas
  cleaner_jurado("\\.") %>% # borra puntos
  cleaner_jurado("ms+\\.") %>%
  cleaner_jurado("\\_.*") # borra _________

todas_tesis_jurado <- bind_cols(todas_tesis_pdf, jurado_norm)
#todas_tesis_jurado[todas_tesis_jurado$presidente == "presidente", c("portada")]
df <- todas_tesis_jurado %>%
  select(-portada, -portada_limp) %>%
  as_tibble()

write_rds(df, here("text_processing/data_frame.rds"))
```

Figura 37. Proceso de limpieza de datos de los jurados

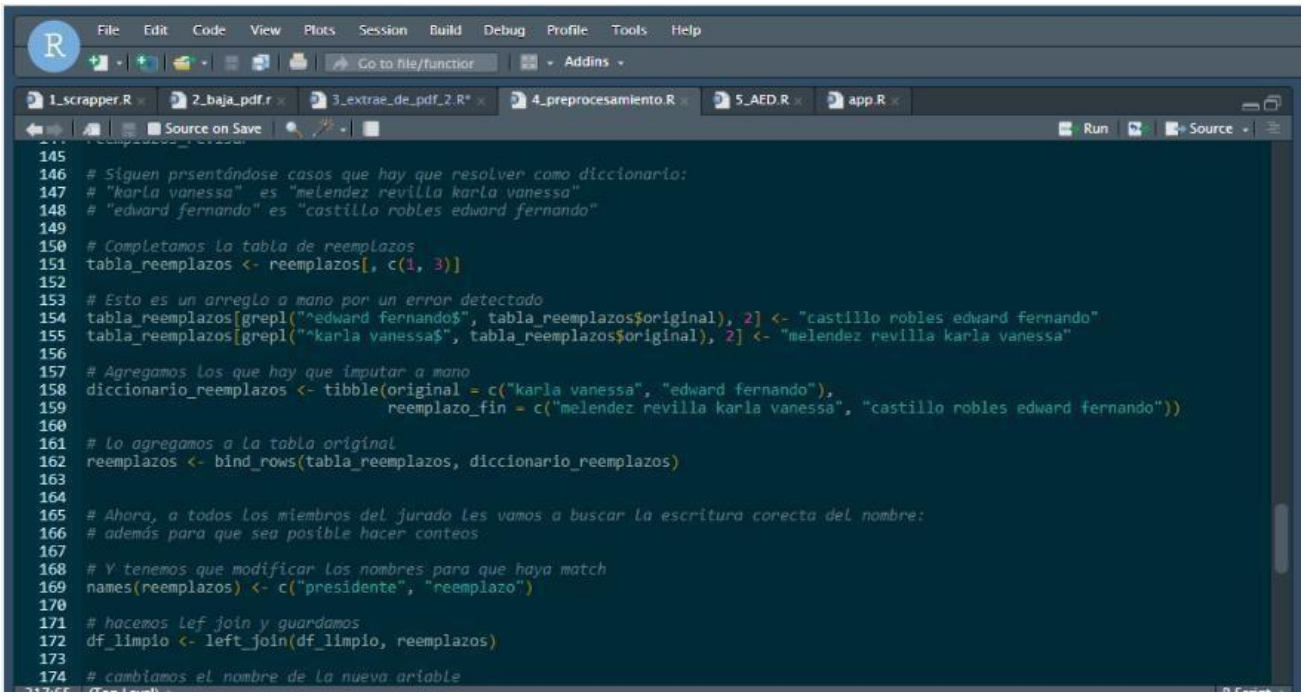
4.2.1.2.2. Preprocesamiento De Datos

Acá realizaremos la limpieza de los datos el cual nos va proporcionar los datos necesarios para pasar la siguiente fase.



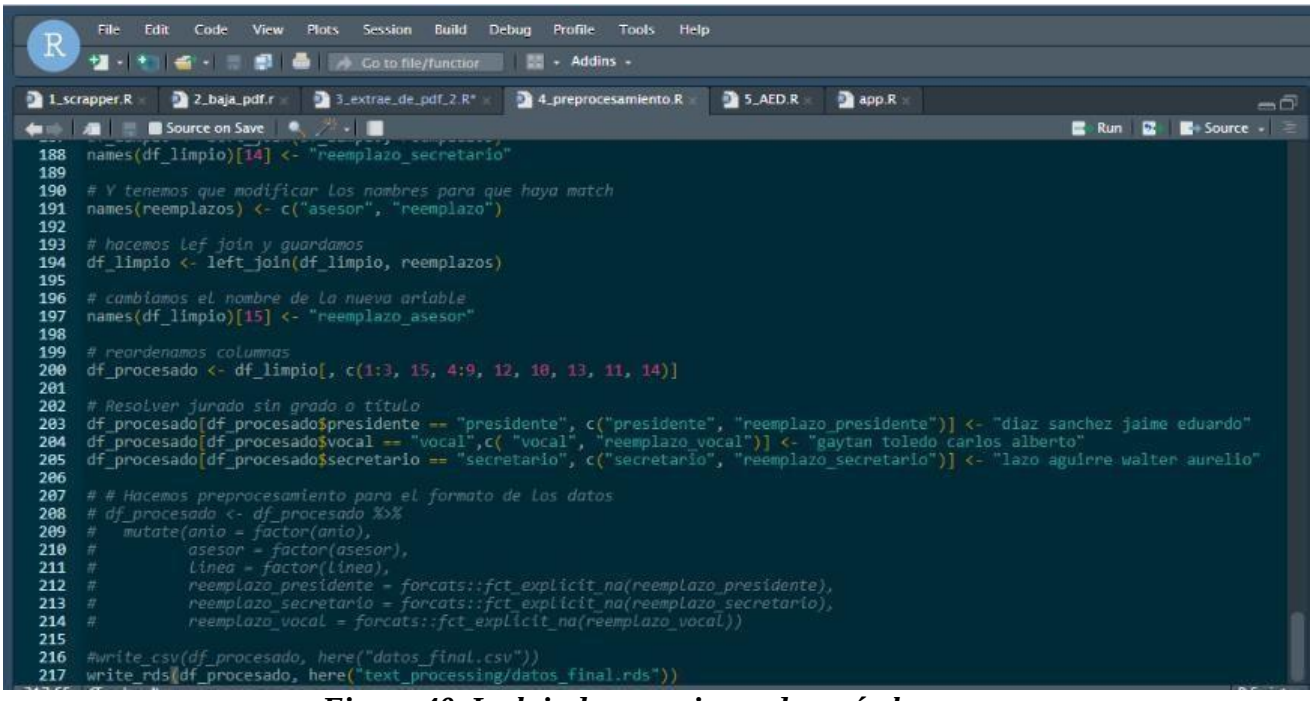
```
library(tidyverse)
library(here)
library(viridis)
library(stringdist)

df <- read_rds(here("text_processing/data_frame.rds")) # cargamos el dataframe con toda la info que vamos a
# procesar

# 1. Tenemos problemas con las fechas. Sólo necesitamos el año: primeros cuatro números
df <- df %>%
  mutate(anio = str_sub(anio, 1, 4))

# chequeamos
df %>%
  count(anio) %>%
  ggplot(aes(anio, n, fill = n)) +
  geom_bar(stat = "identity") +
  ggtitle("Tesis por año") +
  xlab("años") +
  ylab("tesis") +
  theme_dark() +
  scale_fill_viridis(discrete = FALSE)

# En la columna autor vienen pegados el asesor y los tesistas, separados por ";"
df <- df %>%
  separate(autor, into = c("asesor", "autor_1", "autor_2"), sep = ";")

# Vamos a aprovechar que los nombres de los asesores conservan un estandar. Pero antes vamos a remover
# tildes y vamos a llevar todo a minúsculas
cleaner_one <- function(df) {
```

Figura 38. Proceso de ejecución donde comenzaremos a ordenar los datos de asesor, autores, presidente, vocal, secretario



```
# Siguen prsentándose casos que hay que resolver como diccionario:
# "karla vanessa"  es "melendez revilla karla vanessa"
# "edward fernando" es "castillo robles edward fernando"

# Completamos la tabla de reemplazos
tabla_reemplazos <- reemplazos[, c(1, 3)]

# Esto es un arreglo a mano por un error detectado
tabla_reemplazos[grepl("^edward fernando$", tabla_reemplazos$original), 2] <- "castillo robles edward fernando"
tabla_reemplazos[grepl("^karla vanessa$", tabla_reemplazos$original), 2] <- "melendez revilla karla vanessa"

# Agregamos los que hay que imputar a mano
diccionario_reemplazos <- tibble(original = c("karla vanessa", "edward fernando"),
                                 reemplazo_fin = c("melendez revilla karla vanessa", "castillo robles edward fernando"))

# lo agregamos a la tabla original
reemplazos <- bind_rows(tabla_reemplazos, diccionario_reemplazos)


# Ahora, a todos los miembros del jurado les vamos a buscar la escritura correcta del nombre:
# además para que sea posible hacer conteos

# Y tenemos que modificar los nombres para que haya match
names(reemplazos) <- c("presidente", "reemplazo")

# hacemos lef join y guardamos
df_limpio <- left_join(df_limpio, reemplazos)

# cambiamos el nombre de la nueva ariable
```

Figura 39. Código incluir a dos docentes



```
names(df_limpio)[14] <- "reemplazo_secretario"

# Y tenemos que modificar los nombres para que haya match
names(reemplazos) <- c("asesor", "reemplazo")

# hacemos lef join y guardamos
df_limpio <- left_join(df_limpio, reemplazos)

# cambiamos el nombre de la nueva ariable
names(df_limpio)[15] <- "reemplazo_asesor"

# reordenamos columnas
df_procesado <- df_limpio[, c(1:3, 15, 4:9, 12, 10, 13, 11, 14)]

# Resolver jurado sin grado o título
df_procesado[df_procesado$presidente == "presidente", c("presidente", "reemplazo_presidente")] <- "diaz sanchez jaime eduardo"
df_procesado[df_procesado$vocal == "vocal",c( "vocal", "reemplazo_vocal")] <- "gaytan toledo carlos alberto"
df_procesado[df_procesado$secretario == "secretario", c("secretario", "reemplazo_secretario")] <- "lazo aguirre walter aurelio"

# # Hacemos preprocesamiento para el formato de los datos
# df_procesado <- df_procesado %>%
#   mutate(anio = factor(anio),
#          asesor = factor(asesor),
#          linea = factor(linea),
#          reemplazo_presidente = forcats::fct_explicit_na(reemplazo_presidente),
#          reemplazo_secretario = forcats::fct_explicit_na(reemplazo_secretario),
#          reemplazo_vocal = forcats::fct_explicit_na(reemplazo_vocal))

#write_csv(df_procesado, here("datos_final.csv"))
write_rds(df_procesado, here("text_processing/datos_final.rds"))
```

Figura 40. Incluir docente sin grados y títulos

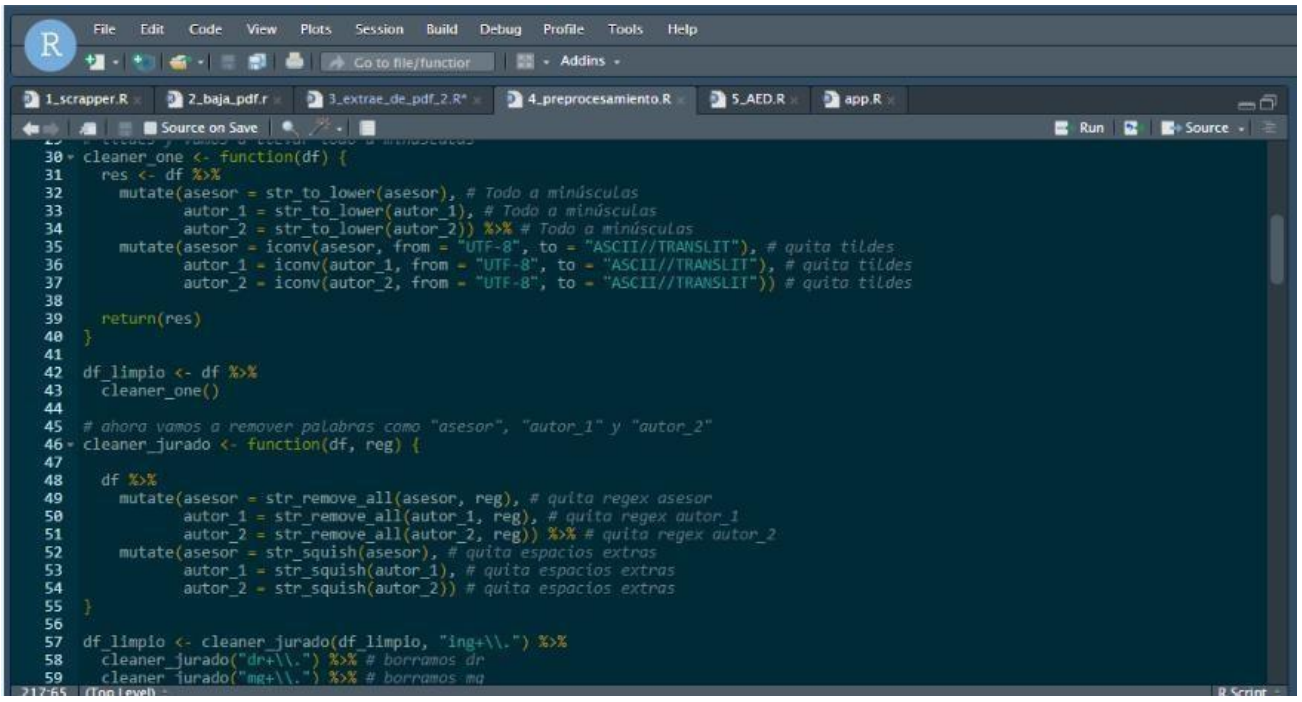
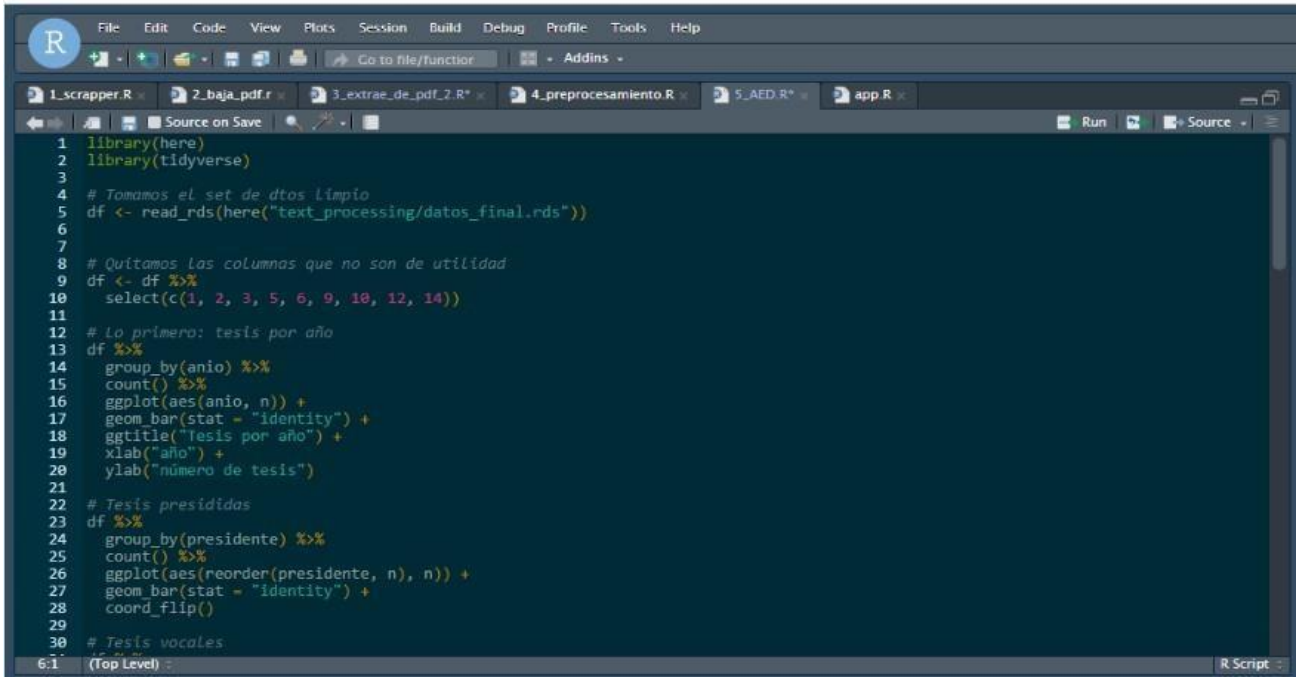


Figura 41. Limpieza y creación de variables df_limpio

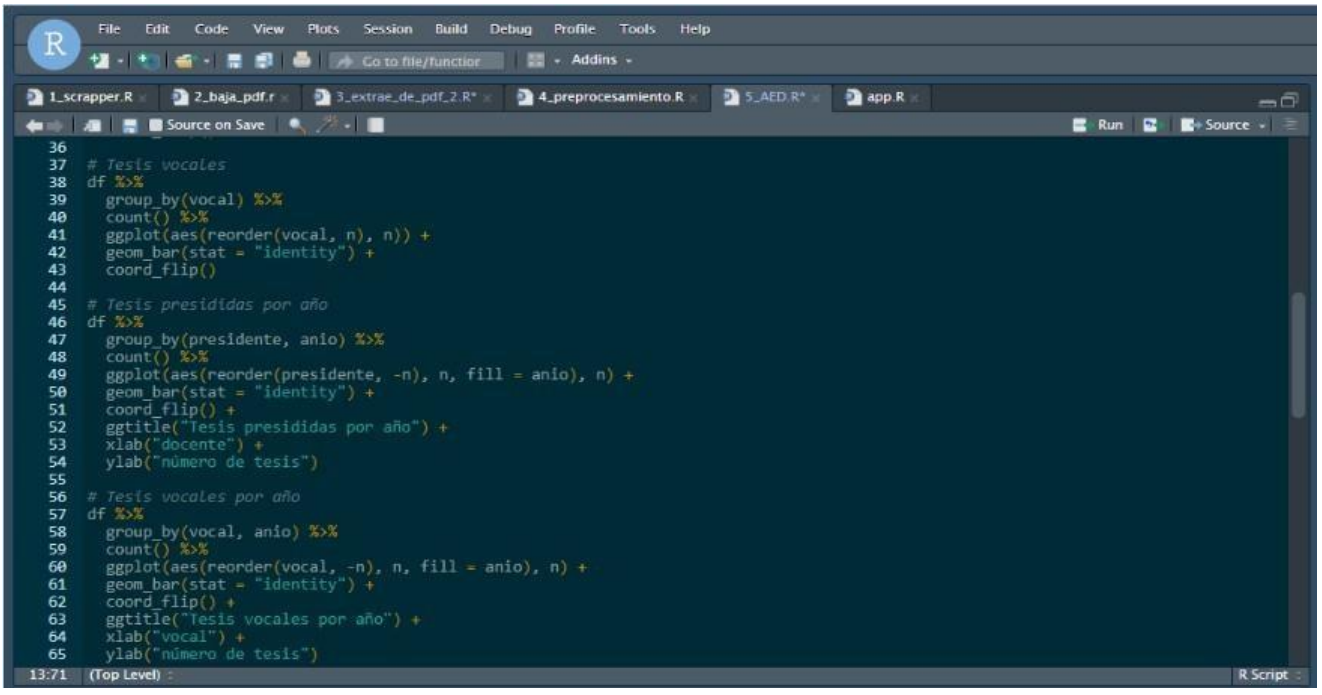
4.2.1.2.3. Análisis De La Extracción De Datos

Recopila la parte limpia del paso anterior y comienza a clasificar jurados, asesores, autores, etc. Este paso viene hacer preparación para lo que se va visualizar en el Dashboard.



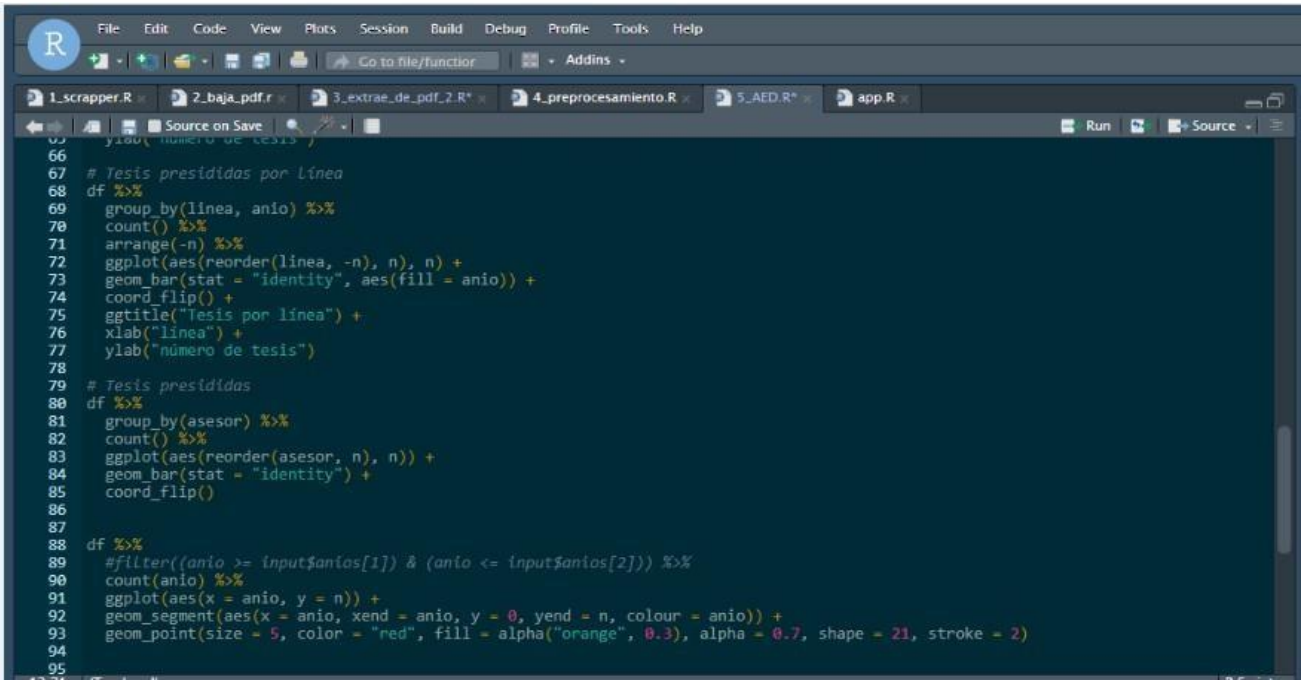
```
library(here)
library(tidyverse)

# Tomamos el set de dtos limpio
df <- read_rds(here("text_processing/datos_final.rds"))


# Quitamos las columnas que no son de utilidad
df <- df %>%
  select(c(1, 2, 3, 5, 6, 9, 10, 12, 14))

# Lo primero: tesis por año
df %>%
  group_by(anio) %>%
  count() %>%
  ggplot(aes(anio, n)) +
  geom_bar(stat = "identity") +
  ggtitle("Tesis por año") +
  xlab("año") +
  ylab("número de tesis")

# Tesis presididas
df %>%
  group_by(presidente) %>%
  count() %>%
  ggplot(aes(reorder(presidente, n), n)) +
  geom_bar(stat = "identity") +
  coord_flip()

# Tesis vocales
```

Figura 42. Datos limpios y eliminamos columnas que no interesan



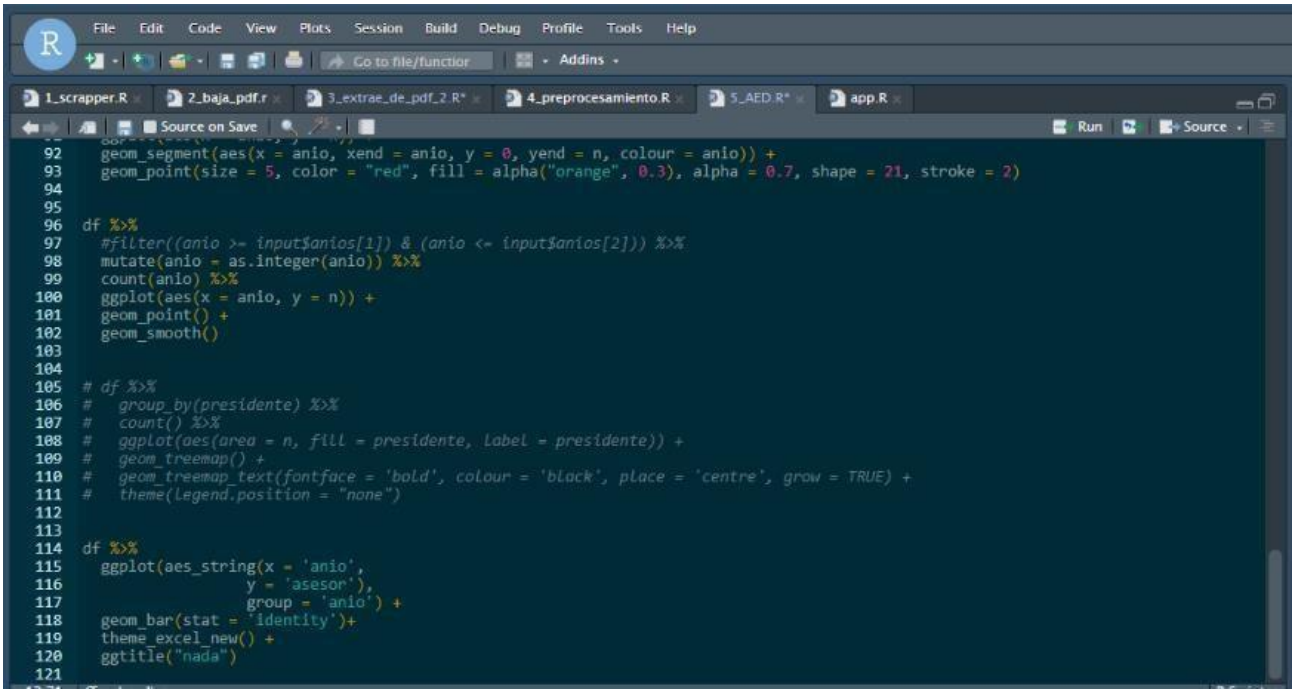
```
# Tesis vocales
df %>%
  group_by(vocal) %>%
  count() %>%
  ggplot(aes(reorder(vocal, n), n)) +
  geom_bar(stat = "identity") +
  coord_flip()

# Tesis presididas por año
df %>%
  group_by(presidente, anio) %>%
  count() %>%
  ggplot(aes(reorder(presidente, -n), n, fill = anio), n) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ggtitle("Tesis presididas por año") +
  xlab("docente") +
  ylab("número de tesis")

# Tesis vocales por año
df %>%
  group_by(vocal, anio) %>%
  count() %>%
  ggplot(aes(reorder(vocal, -n), n, fill = anio), n) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ggtitle("Tesis vocales por año") +
  xlab("vocal") +
  ylab("número de tesis")
```

Figura 43. Agrupamos los datos que requerimos



```
# Tesis presididas por línea
df %>%
  group_by(linea, anio) %>%
  count() %>%
  arrange(-n) %>%
  ggplot(aes(reorder(linea, -n), n), n) +
  geom_bar(stat = "identity", aes(fill = anio)) +
  coord_flip() +
  ggtitle("Tesis por línea") +
  xlab("línea") +
  ylab("número de tesis")

# Tesis presididas
df %>%
  group_by(asesor) %>%
  count() %>%
  ggplot(aes(reorder(asesor, n), n)) +
  geom_bar(stat = "identity") +
  coord_flip()


df %>%
  #filter((anio >= input$anios[1]) & (anio <= input$anios[2])) %>%
  count(anio) %>%
  ggplot(aes(x = anio, y = n)) +
  geom_segment(aes(x = anio, xend = anio, y = 0, yend = n, colour = anio)) +
  geom_point(size = 5, color = "red", fill = alpha("orange", 0.3), alpha = 0.7, shape = 21, stroke = 2)
```

Figura 44. Tesis por Línea



```
  geom_segment(aes(x = anio, xend = anio, y = 0, yend = n, colour = anio)) +
  geom_point(size = 5, color = "red", fill = alpha("orange", 0.3), alpha = 0.7, shape = 21, stroke = 2)


df %>%
  #filter((anio >= input$anios[1]) & (anio <= input$anios[2])) %>%
  mutate(anio = as.integer(anio)) %>%
  count(anio) %>%
  ggplot(aes(x = anio, y = n)) +
  geom_point() +
  geom_smooth()


# df %>%
#   group_by(presidente) %>%
#   count() %>%
#   ggplot(aes(area = n, fill = presidente, label = presidente)) +
#   geom_treemap() +
#   geom_treemap_text(fontface = 'bold', colour = 'black', place = 'centre', grow = TRUE) +
#   theme(legend.position = "none")


df %>%
  ggplot(aes_string(x = 'anio',
                    y = 'asesor'),
                    group = 'anio') +
  geom_bar(stat = 'identity')+
  theme_excel_new() +
  ggtitle("nada")
```

Figura 45. Ordenamiento por año y comienzo de gráficos

4.2.1.3. Desarrollo e Implementación de gráficos para el Dashboard

Es el diseño del Dashboard de todos los datos obtenidos desde Web Scraping hasta el Análisis de Extracción de Datos.

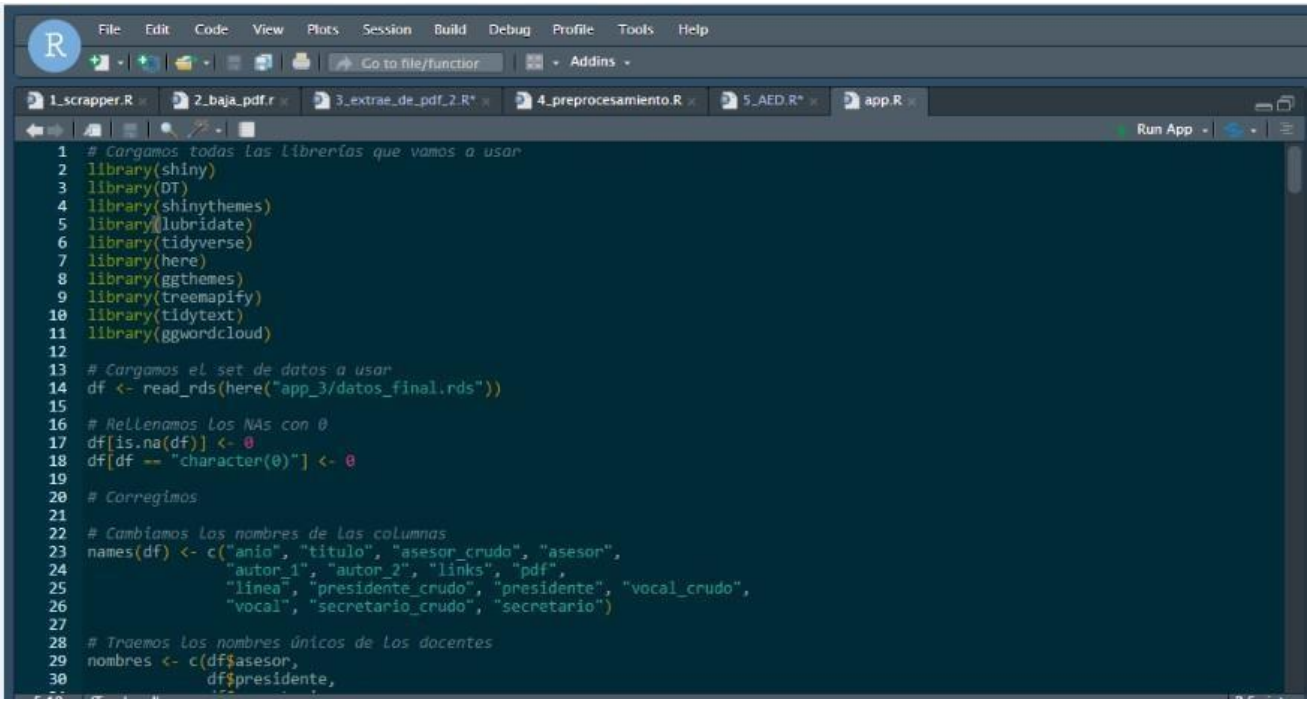


Figura 46. Proceso de realización y ejecución del Dashboard

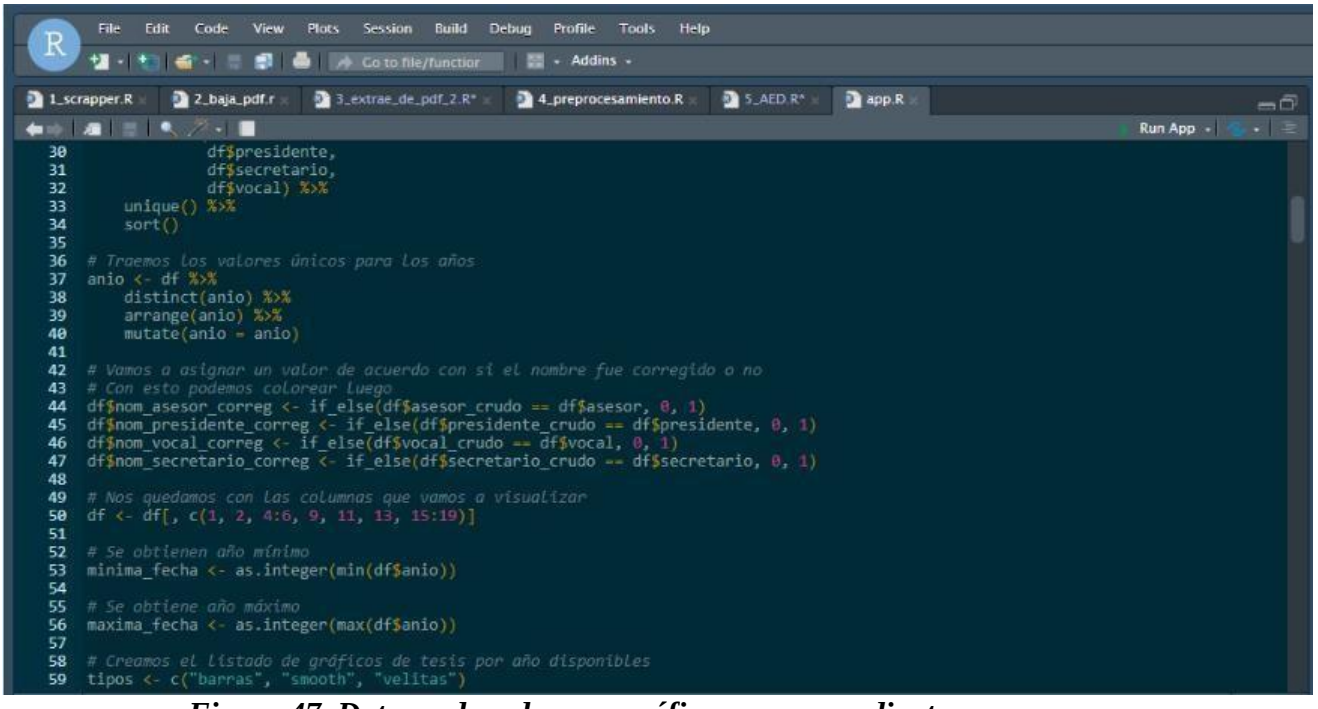
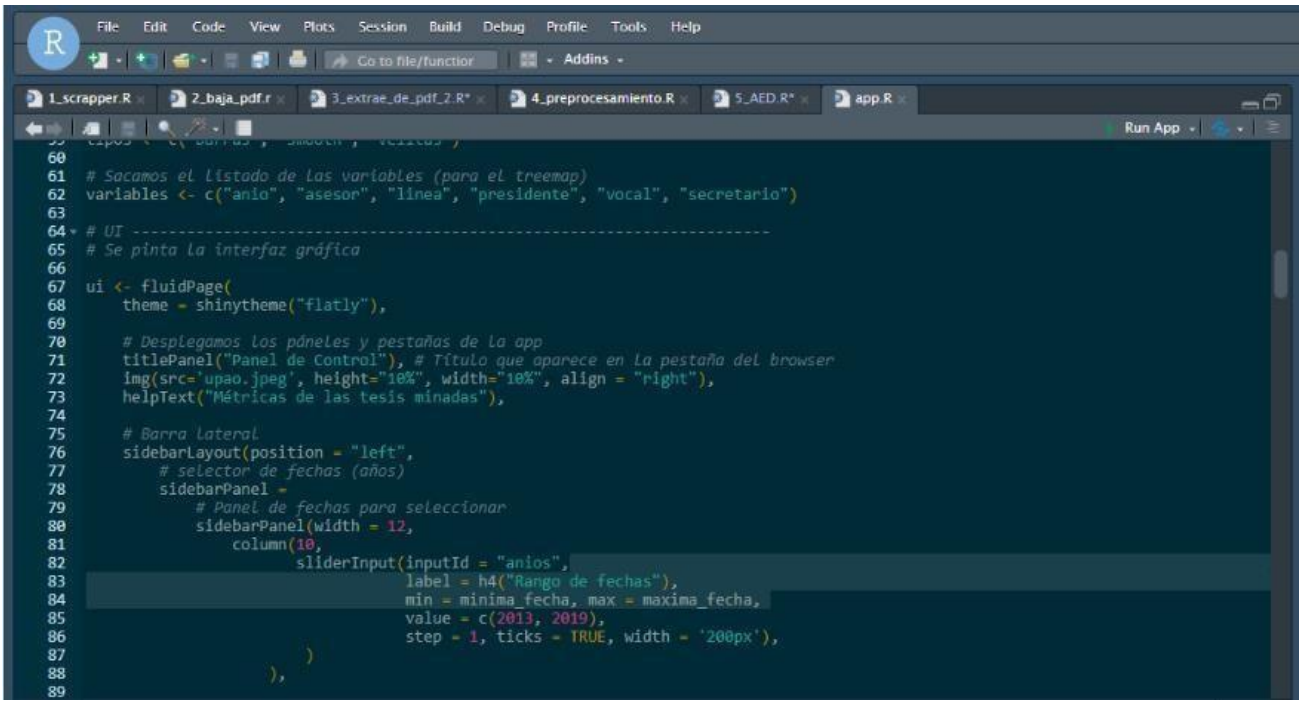
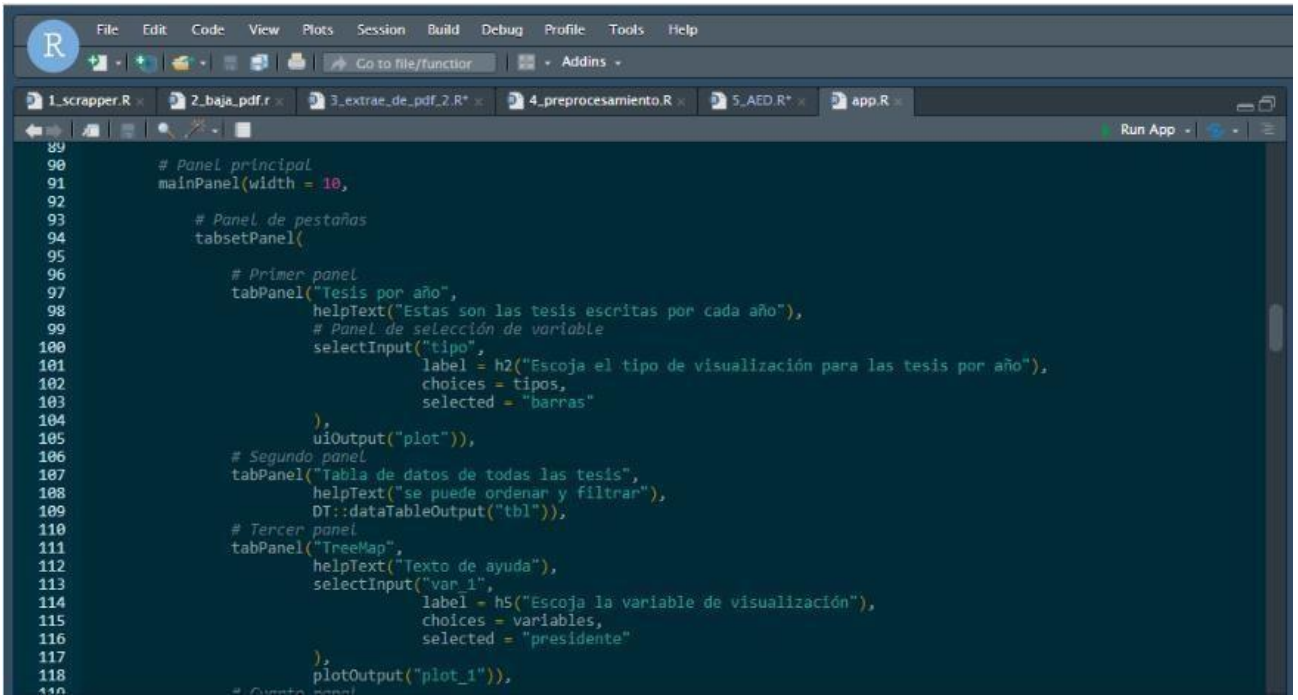


Figura 47. Datos ordenados con gráficos correspondientes



```
60
61  # Sacamos el listado de las variables (para el treemap)
62  variables <- c("anio", "asesor", "linea", "presidente", "vocal", "secretario")
63
64  # UI ------------------------------------------------------------------
65  # Se pinta la interfaz gráfica
66
67  ui <- fluidPage(
68      theme = shinytheme("flatly"),
69
70      # Desplegamos los páneles y pestañas de la app
71      titlePanel("Panel de Control"), # Título que aparece en la pestaña del browser
72      img(src='upao.jpeg', height="10%", width="10%", align = "right"),
73      helpText("Métricas de las tesis minadas"),
74
75      # Barra lateral
76      sidebarLayout(position = "left",
77          # selector de fechas (años)
78          sidebarPanel =
79              # Panel de fechas para seleccionar
80              sidebarPanel(width = 12,
81                  column(10,
82                         sliderInput(inputId = "anios",
83                                     label = h4("Rango de fechas"),
84                                     min = minima_fecha, max = maxima_fecha,
85                                     value = c(2013, 2019),
86                                     step = 1, ticks = TRUE, width = '200px'),
87                  )
88              ),
89
```

Figura 48. Gráficos de todas las tesis por año y fecha



```
89
90          # Panel principal
91          mainPanel(width = 10,
92
93              # Panel de pestañas
94              tabsetPanel(
95
96                  # Primer panel
97                  tabPanel("Tesis por año",
98                           helpText("Estas son las tesis escritas por cada año"),
99                           # Panel de selección de variable
100                          selectInput("tipo",
101                                      label = h2("Escoja el tipo de visualización para las tesis por año"),
102                                      choices = tipos,
103                                      selected = "barras"
104                          ),
105                          uiOutput("plot")),
106                 # Segundo panel
107                 tabPanel("Tabla de datos de todas las tesis",
108                          helpText("se puede ordenar y filtrar"),
109                          DT::dataTableOutput("tbl")),
110                 # Tercer panel
111                 tabPanel("TreeMap",
112                          helpText("Texto de ayuda"),
113                          selectInput("var_1",
114                                      label = h5("Escoja la variable de visualización"),
115                                      choices = variables,
116                                      selected = "presidente"
117                          ),
118                          plotOutput("plot_1")),
```

Figura 49. Etapas de gráficos de cada panel por año, tesis y finalización

4.2.1.4. Desarrollo del Aplicativo Web en Shiny Apps

En este paso final realizaremos la publicación del Dashboard.

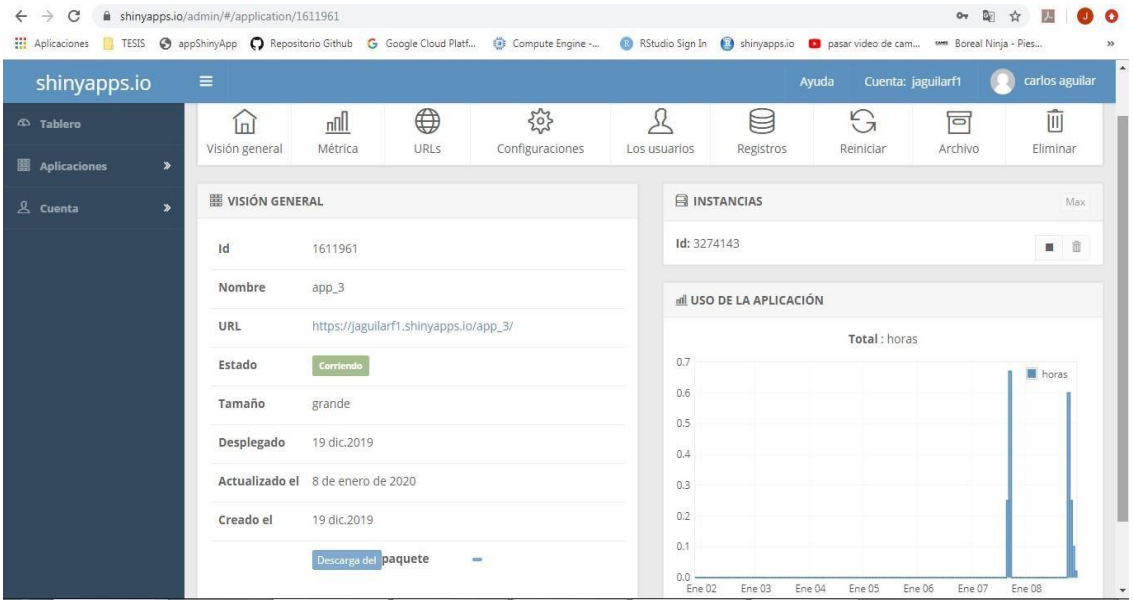


Figura 50. Ingreso del Shinyapps y visión general

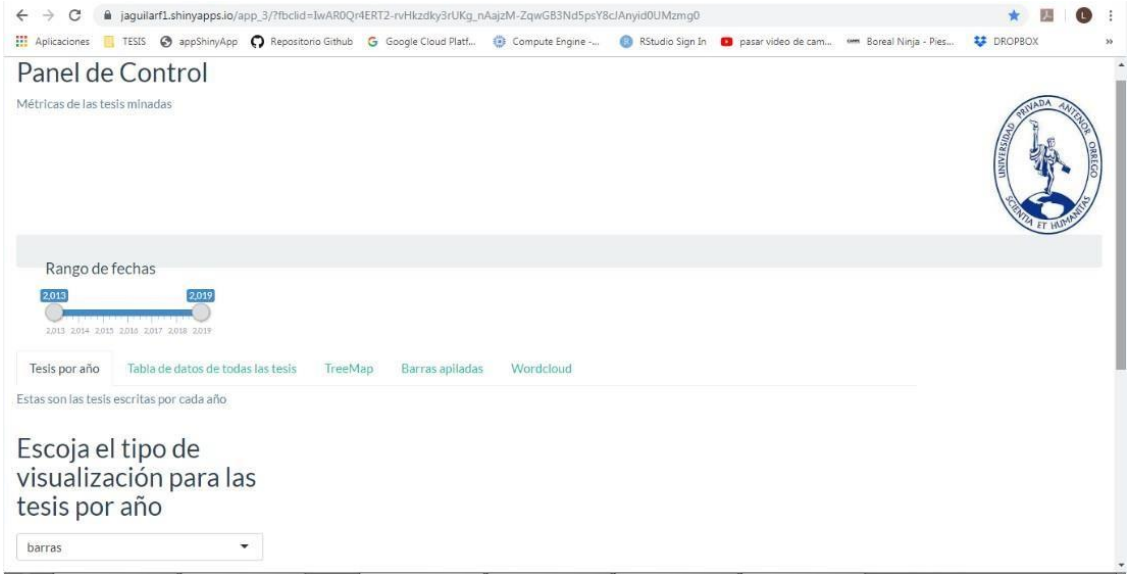


Figura 51. Visión del Dashboard en la web

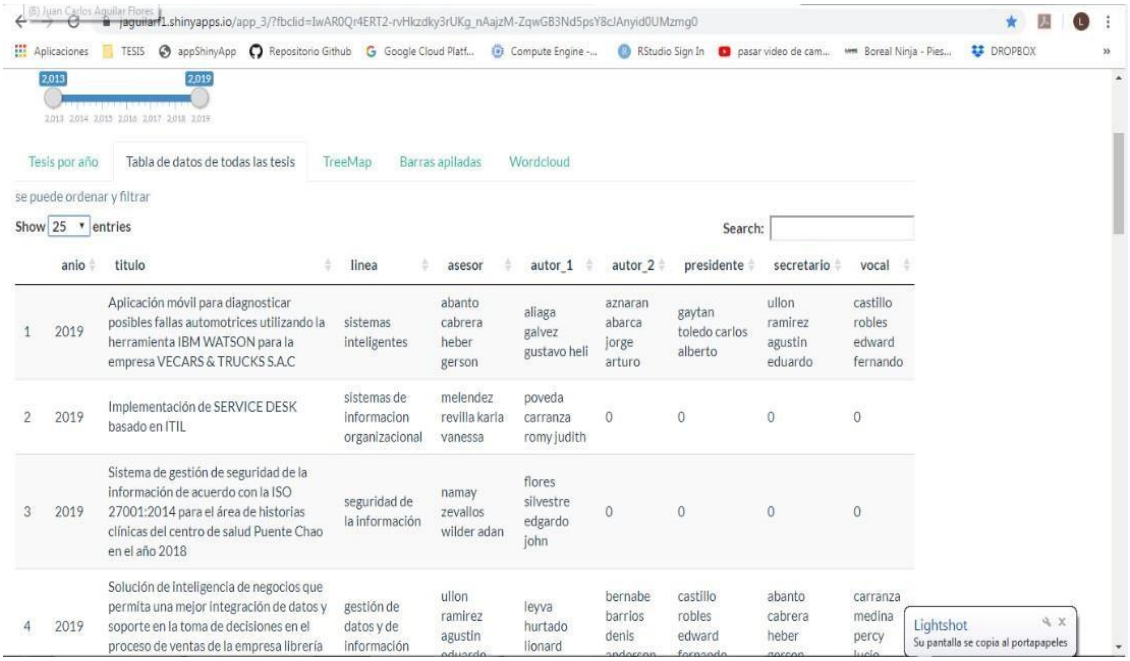


Figura 52.Dashboard con los datos extraídos de las tesis

5. DISCUSIÓN DE RESULTADOS

5.1. HIPOTESIS PLANTEADA

"Es posible que la implementación de un Dashboard basada en minería de texto ayudara a medir los proyectos de tesis en la universidad privada Antenor Orrego."

5.1.1. Muestra Aplicar

Se realizó una encuesta a 5 docentes de forma aleatoria dentro de la Universidad Privada Antenor Orrego ubicado en Trujillo. Para realizar una evaluación de que tan eficiente puede ser el Dashboard.

5.1.2. Validación de la Solución

Se hizo una encuesta para poder recolectar la información necesaria con una escala de 1 al 5, siendo 1 Muy Malo y 5 Muy bueno, de esa manera poder conocer si el Dashboard es útil para la Universidad.

VALORACION	ESCALA
Muy Bueno	5
Bueno	4
Regular	3
Malo	2
Muy Malo	1

Tabla 4. Formato de Valoración

5.1.3. Variable Independiente: Dashboard basado en Minería de Texto.

Para obtener los valores de calificación de los resultados se aplicó una cuesta sobre el Dashboard y se determinó por la opinión de los docentes para obtener la facilidad de uso obteniendo como resultado el valor 4 lo que se interpreta según nuestra tabla de valoración como BUENO. Esto quiere decir que nuestro Dashboard es fácil de usar.

Facilidad con que el Usuario Maneja el Dashboard					
Personas / Preguntas	DOCENTE 1	DOCENTE 2	DOCENTE 3	DOCENTE 4	DOCENTE 5
PREGUNTA 1	5	5	5	4	4
PREGUNTA 2	4	3	3	3	5
PREGUNTA 3	4	4	4	4	5
PREGUNTA 4	4	5	3	5	3
PREGUNTA 5	5	3	3	3	4
TOTAL	22	20	18	19	21
PROMEDIO	4.4	4	3.6	3.8	4.2
Facilidad de Uso	4				

Tabla 5. Valoración de Facilidad de Uso

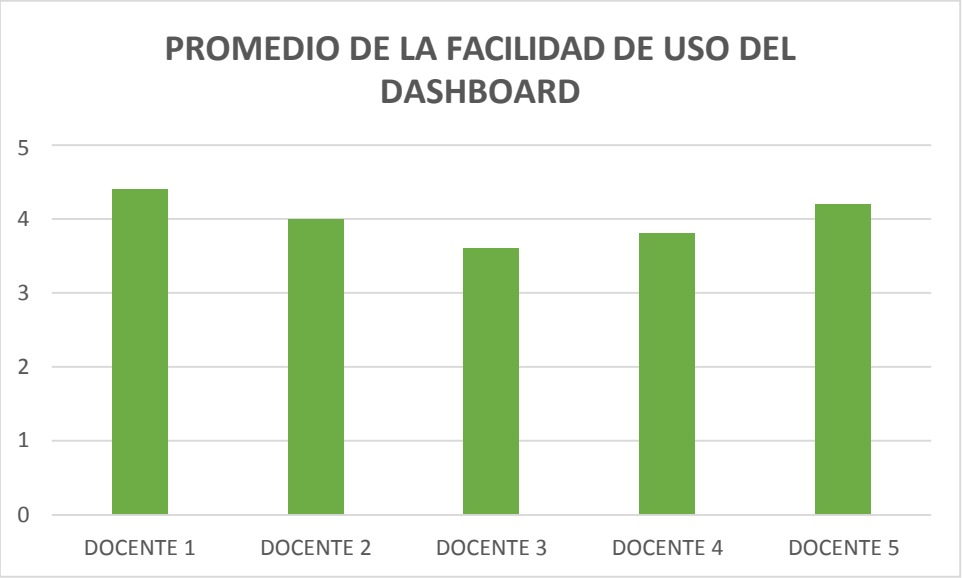


Figura 53. Resultado de Facilidad de Uso

5.1.4. Variable Dependiente: Medición de los proyectos de investigación de tesis de la Universidad Privada Antenor Orrego.

Para obtener los valores de calificación de los resultados se aplicó una encuesta sobre el Dashboard y se determinó por la opinión de los docentes para obtener el grado de satisfacción obteniendo como resultado el valor 4.04 lo que se interpreta según nuestra tabla de valoración como BUENO. Esto quiere decir que nuestro Dashboard tiene un grado de satisfacción aceptable.

Satisfacción del Usuario al Usar el Dashboard					
Preguntas \ personas	DOCENTE 1	DOCENTE 2	DOCENTE 3	DOCENTE 4	DOCENTE 5
PREGUNTA 1	4	4	4	4	3
PREGUNTA 2	5	3	5	4	4
PREGUNTA 3	3	5	5	5	4
PREGUNTA 4	3	3	3	4	5
PREGUNTA 5	4	5	4	4	4
TOTAL	19	20	21	21	20
PROMEDIO	3.8	4	4.2	4.2	4
Grado de Satisfacción	4.04				

Tabla 6. Valoración de la Satisfacción del Dashboard

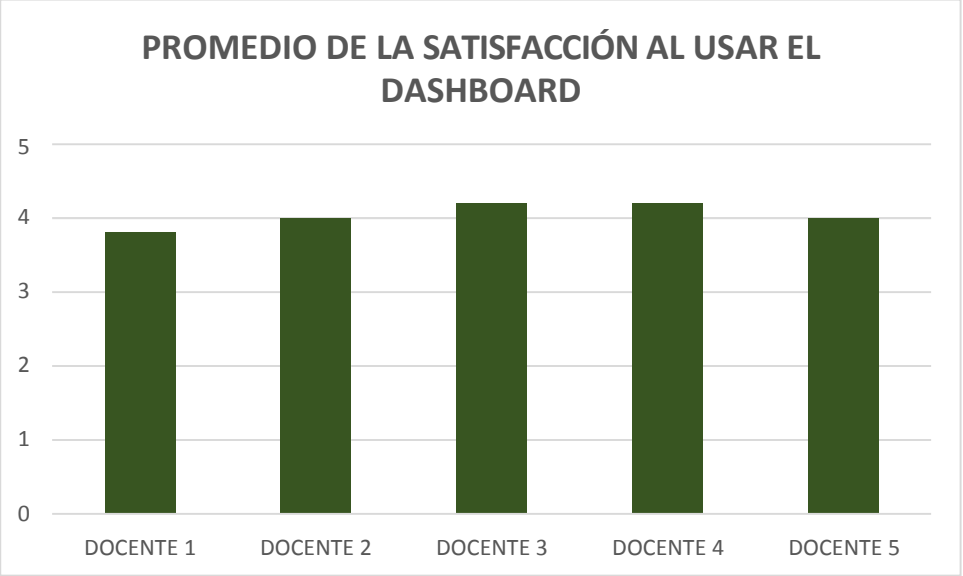


Figura 54. Resultado de la Satisfacción del Dashboard

6. CONCLUSIONES

- Se realizó la técnica de Web Scraping de los 84 Proyectos de Tesis del repositorio de la Escuela Profesional de Ingeniería de Computación y Sistemas, que nos permitió obtener datos como los años, profesores, líneas de investigación, a través; de los paquetes Rvest, Tidyverse, Here.
- Se empleó el lenguaje de programación llamado R y el entorno de trabajo RStudio, el cual nos permite descargar una gran variedad de paquetes que se utiliza para el desarrollo del Dashboard, y posteriormente utilizar el framework Shinny Apps para publicarlo en la web.
- Se realizó mediante los modelos Regex y Word Embbeding la extracción de datos para el desarrollo del Dashboard el cual contiene gráficos de barras, treemap, barras apiladas, tablas, para visualizar los indicadores las cuales son, cuántas tesis pueden ver, cuantas líneas de investigación se encuentran, nombres de los jurados, con que continuidad los docentes participan como jurados o asesores, utilizando el paquete ggplot2.
- Finalmente, para poder visualizar el Dashboard, se utilizó el framework Shinny Apps, que nos permite subir nuestro programa a la

web. Esto gracias a que se realizó el programa en la plataforma RStudio, el cual contiene los indicadores y los gráficos deseados, de esa manera los docentes pueden interactuar con dicho Dashboard ya que es de fácil uso.

7. RECOMENDACIONES

- Controlar que todos los Proyectos de Tesis tengan una estructura fija de esa manera realizar una Minería de Texto más precisa.

- Revisar los documentos de Proyectos de Tesis antes de ser subidos al repositorio de la Universidad, ya que existen algunos Proyectos de Tesis que no tienen todo su contenido y eso dificulta la Minería de Texto.

- Comprar RStudio con licencia, para de esa manera realizar una minería de texto más eficiente, ya que la versión gratuita tiene sus limitaciones.

- Desarrollar un Dashboard a futuro donde se pueda incluir los Proyectos de Tesis de las diferentes carreras que dicta la Universidad, para saber cómo están llevando sus proyectos las demás escuelas profesionales.

8. REFERENCIAS BIBLIOGRAFICAS

¿Qué es la R? (2016). Obtenido de ¿Qué es la R?: https://www.r-project.org/about.html

¿Qué es un dashboard? (s.f.). Obtenido de ¿Qué es un dashboard?: https://www.40defiebre.com/que-es/dashboard

Admin. (12 de Marzo de 2017). *Centrarse en el método ágil*. Obtenido de Centrarse en el método ágil: http://www.portage-emploi.com/2017/03/12/focus-sur-la-methode-agile/

Bahit, E. (2015). *Scrum y eXtreme Programming para Programadores* . Obtenido de Scrum y eXtreme Programming para Programadores : http://umh2818.edu.umh.es/wp-content/uploads/sites/884/2016/02/Scrum-y-eXtrem-Programming-para-programadores.pdf

Calvo, D. (12 de Julio de 2017). *Definición de red neuronal artificial*. Obtenido de Definición de red neuronal artificial: http://www.diegocalvo.es/definicion-de-red-neuronal/

Chang, J., Christian O' Reilly, N. P., Gareth Owen, K. H., & Oudenhoven, M. (21 de Febrero de 2018). *¿Qué es la minería de textos, cómo funciona y por qué es útil?* Obtenido de ¿Qué es la minería de textos, cómo funciona y por qué es útil?: http://openminted.eu/text-mining-101/

Collis, J. (19 de Abril de 2017). *Glosario de aprendizaje profundo: incrustación de palabras*. Obtenido de Glosario de aprendizaje profundo: incrustación de palabras: https://medium.com/deeper-learning/glossary-of-deep-learning-word-embedding-f90c3cec34ca

EUROFORUM. (24 de Abril de 2017). *Machine Learning, un paso más hacia la Inteligencia Artificial*. Obtenido de Machine Learning, un paso más hacia la Inteligencia Artificial: https://www.euroforum.es/blog/machine-learning-un-paso-mas-hacia-la-inteligencia-artificial/

Facebook, L. d. (19 de Diciembre de 2018). *FastText*. Obtenido de FastText: https://en.wikipedia.org/wiki/FastText

Falbel, D. (19 de Junio de 2019). *Ejemplo de regresión con redes Keras LSTM en R*. Obtenido de Ejemplo de regresión con redes Keras LSTM en R: https://www.datatechnotes.com/2019/01/regression-example-with-lstm-networks.html

Fasttext. (12 de Abril de 2018). *Redes neuronales: ¡es fácil! O clasifique el texto usando fasttext*. Obtenido de Redes neuronales: ¡es fácil! O clasifique el texto usando fasttext: https://gosha20777.github.io/tutorial/2018/04/12/fasttext-for-windows/

Garatu, G. (s.f.). *El corazón del Scrum en proyectos de desarrollo de IoT y SmartFactory*. Obtenido de El corazón del Scrum en proyectos de desarrollo de IoT y SmartFactory: https://development.grupogaratu.com/metodologia-scrum-desarrollo-software/

GARZÓN, J. I. (6 de Noviembre de 2018). *Cómo usar redes neuronales (LSTM) en la predicción de averías en las máquinas*. Obtenido de Cómo usar redes neuronales (LSTM) en la predicción de averías en las máquinas: https://blog.gft.com/es/2018/11/06/como-usar-redes-neuronales-lstm-en-la-prediccion-de-averias-en-las-maquinas/

GERENS. (2017). *Gestión de riesgos: ¿Qué es? ¿Por qué emplearla? ¿Cómo emplearla?* Obtenido de Gestión de riesgos: ¿Qué es? ¿Por qué emplearla? ¿Cómo emplearla?: https://gerens.pe/blog/gestion-riesgo-que-por-que-como/

Gonçalves, L. (25 de Enero de 2019). *QUÉ ES LA METODOLOGÍA ÁGIL*. Obtenido de QUÉ ES LA METODOLOGÍA ÁGIL: https://luis-goncalves.com/es/que-es-la-metodologia-agil/

Gonzalez, A. (1 de Julio de 2014). *¿Qué es Machine Learning?* Obtenido de ¿Qué es Machine Learning?: https://cleverdata.io/que-es-machine-learning-big-data/

gSkinner. (s.f.). *Regexr*. Obtenido de Regexr: https://regexr.com/

Gupta, S. (2 de Enero de 2019). *Incrustaciones de palabras en PNL y sus aplicaciones*. Obtenido de Incrustaciones de palabras en PNL y sus aplicaciones: https://hackernoon.com/word-embeddings-in-nlp-and-its-applications-fab15eaf7430

IntelDig. (4 de Setiembre de 2019). *Cómo el NLP ayuda a detectar sentimientos en Redes Sociales*. Obtenido de Cómo el NLP ayuda a detectar sentimientos en Redes Sociales: https://www.inteldig.com/2019/09/como-el-nlp-ayuda-a-detectar-sentimientos-en-redes-sociales/

Jarrell, E. (5 de Mayo de 2018). *Web Scraping 101 with Python & Beautiful Soup*. Obtenido de Web Scraping 101 with Python & Beautiful Soup: https://codeburst.io/web-scraping-101-with-python-beautiful-soup-bb617be1f486

La atención al cliente del futuro:. (Abril de 4 de 2017). Obtenido de La atención al cliente del futuro:: https://blog.prodware.es/atencion-cliente-futuro-menos-gastos-mejor-experiencia/

Lafuente, A. (21 de Enero de 2019). *Qué es el web scraping*. Obtenido de Qué es el web scraping: https://aukera.es/blog/web-scraping/

Learn, M. (s.f.). *Introducción a la minería de texto*. Obtenido de Introducción a la minería de texto: https://monkeylearn.com/text-mining/

Marín, C. (15 de Diciembre de 2018). *Adaptar nuestra metodología de trabajo a una metodología Ágil*. Obtenido de Adaptar nuestra metodología de trabajo a una metodología Ágil: https://justdigital.agency/metodologia-trabajo-agil/

MARTOS, J. (19 de Setiembre de 2018). *TOMA MEJORES DECISIONES CON LA MINERÍA DE TEXTOS*. Obtenido de TOMA MEJORES DECISIONES CON LA MINERÍA DE TEXTOS: https://www.techedgegroup.com/es/blog/toma-mejores-decisiones-con-mineria-textos

MeaningCloud. (2 de Octubre de 2014). *El papel de la Minería de Texto en el Sector de Seguros*. Obtenido de El papel de la Minería de Texto en el Sector de Seguros: https://www.meaningcloud.com/es/blog/mineria-de-texto-en-sector-de-seguros

Opengate. (14 de Abril de 2016). *TF-IDF en lenguaje R*. Obtenido de TF-IDF en lenguaje R: https://mropengate.blogspot.com/2016/04/tf-idf-in-r-language.html

Rai, A. (1 de 06 de 2019). *Qué es la minería de texto: técnicas y aplicaciones.* Obtenido de upGrad: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/#Applications_Of_Text_Mining

ReYDeS. (28 de Octubre de 2014). *Expresiones Regulares utilizando la Sintaxis Java Regex*. Obtenido de Expresiones Regulares utilizando la Sintaxis Java Regex: http://www.reydes.com/d/?q=Expresiones_Regulares_utilizando_la_Sintaxis_Java_Regex

Rodriguez, C. (20 de Mayo de 2015). *Metodologías Ágiles, objetivos, características, ventajas*. Obtenido de Metodologías Ágiles, objetivos, características, ventajas: https://comunidad.iebschool.com/metodologiasagiles/general/concepto-metodologias-agiles/

Rouse, M. (s.f.). *Redes Neuronales Recurrentes*. Obtenido de Redes Neuronales Recurrentes: https://searchenterpriseai.techtarget.com/definition/recurrent-neural-networks

Schwaber, K., & Sutherland, J. (Noviembre de 2017). *La Guía de Scrum*. Obtenido de La Guía de Scrum: https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-Spanish-SouthAmerican.pdf

Selivanov, D. (Febrero de 2016). *text2vec*. Obtenido de text2vec: http://text2vec.org/index.html

Selivanov, D. (2016). *Text2vec*. Obtenido de Text2vec: http://text2vec.org/

SysAid. (2018). *Manager Dashboard*. Obtenido de Manager Dashboard: https://www.sysaid.com/it-service-management-software/analytics/manager-dashboard

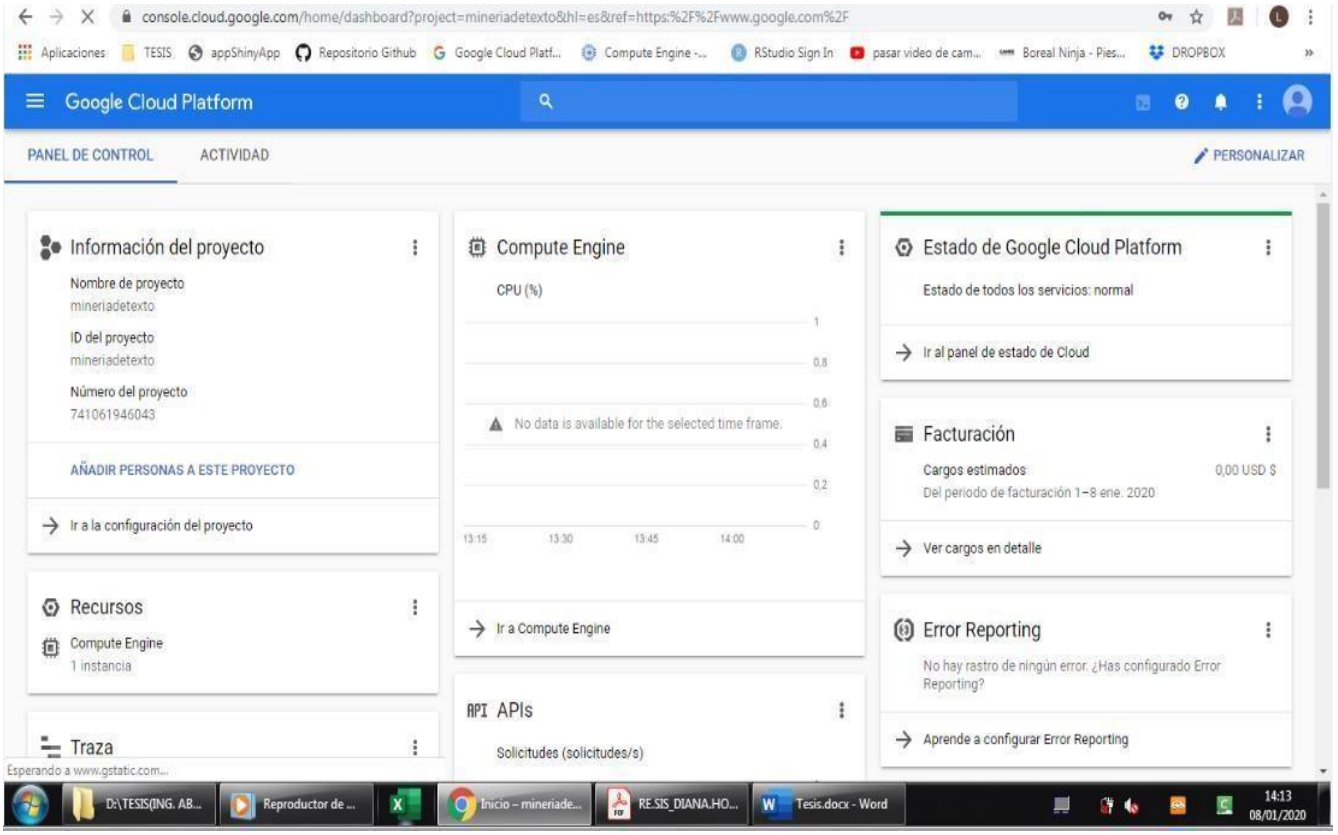
Tf-idf. (s.f.). Obtenido de Tf-idf: https://es.wikipedia.org/wiki/Tf-idf

Torre, M. d. (2017). *Nuevas Tecnicas de Minería de Textos*. Universidad de Granada. Obtenido de Nuevas Tecnicas de Minería de Textos: http://digibug.ugr.es/handle/10481/46975

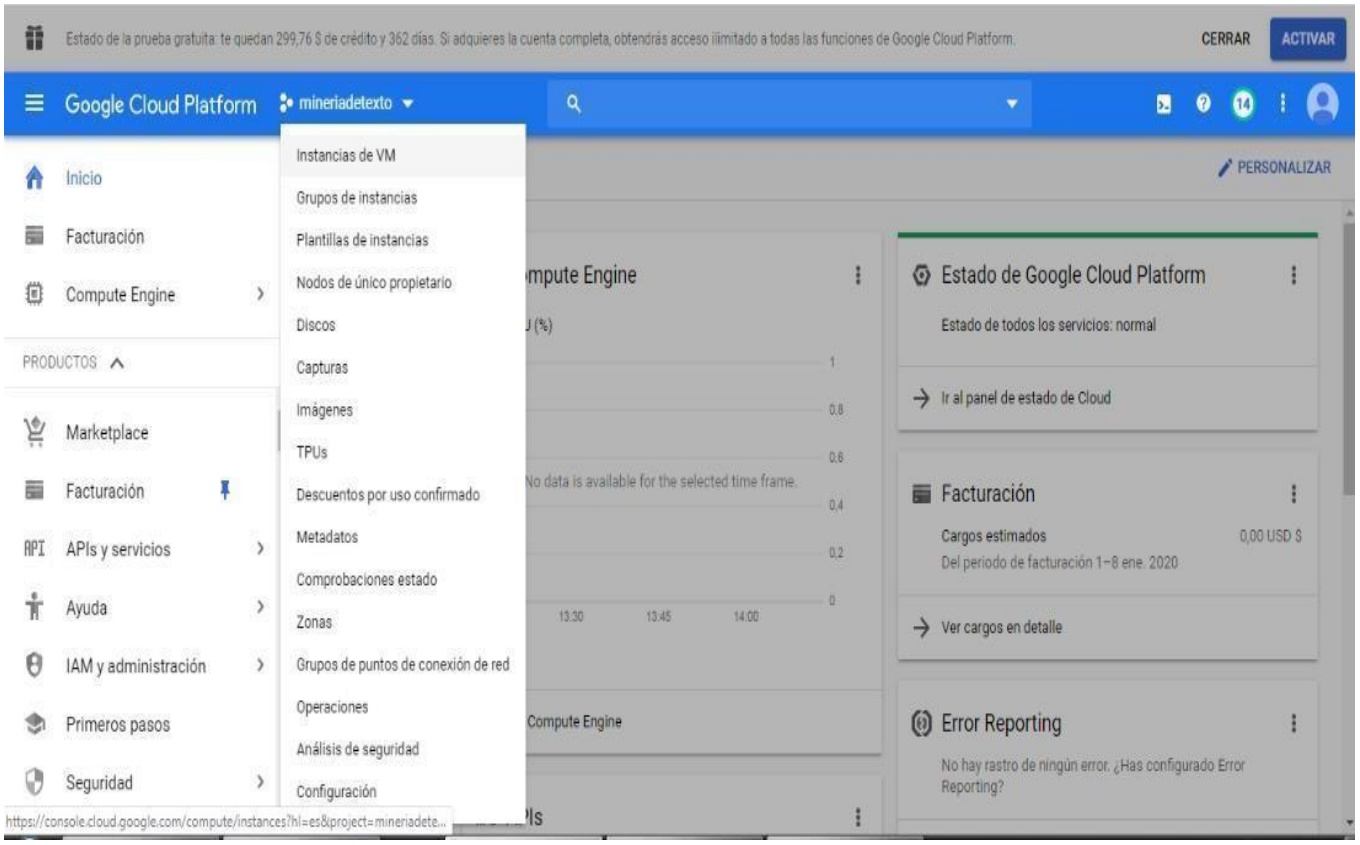
von, J. G. (23 de Setiembre de 2016). *RStudio*. Obtenido de RStudio: https://proyectosbeta.net/2016/09/curso-gratuito-sobre-introduccion-al-tratamiento-de-datos-con-r-y-rstudio/

9. ANEXOS

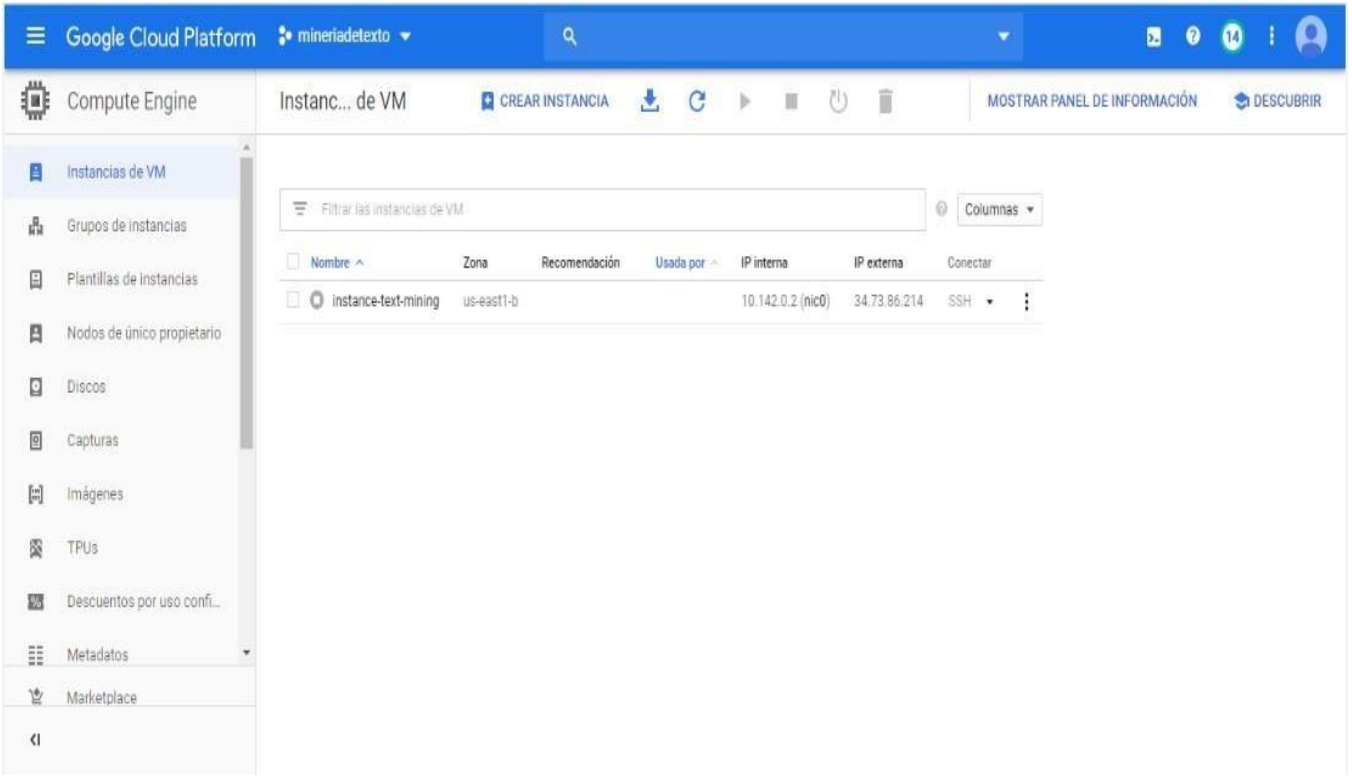
ANEXO 01: Plataforma de Google Cloud Platform



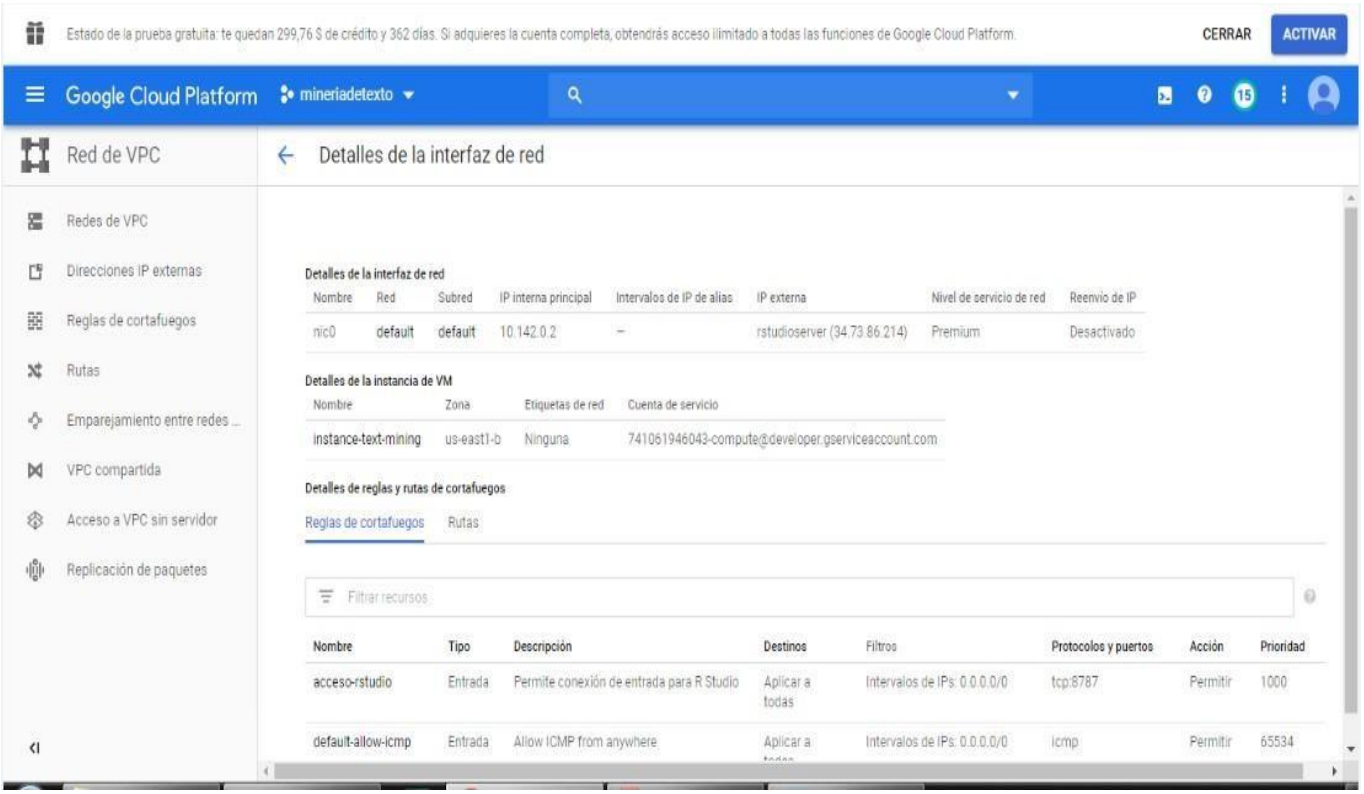
ANEXO 02: Detalle de Plataforma de Google Cloud



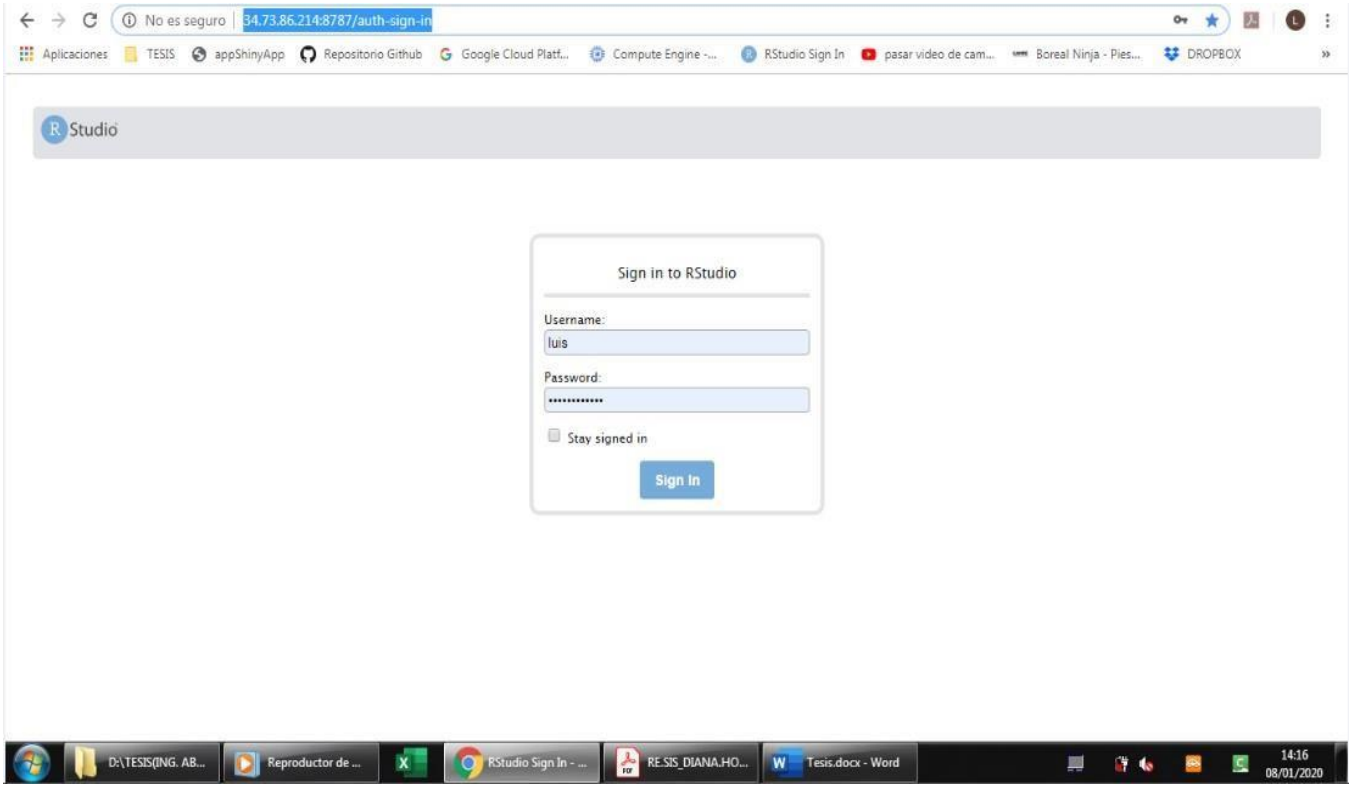
ANEXO 03: Creación de Máquina “Instance-text-mining”



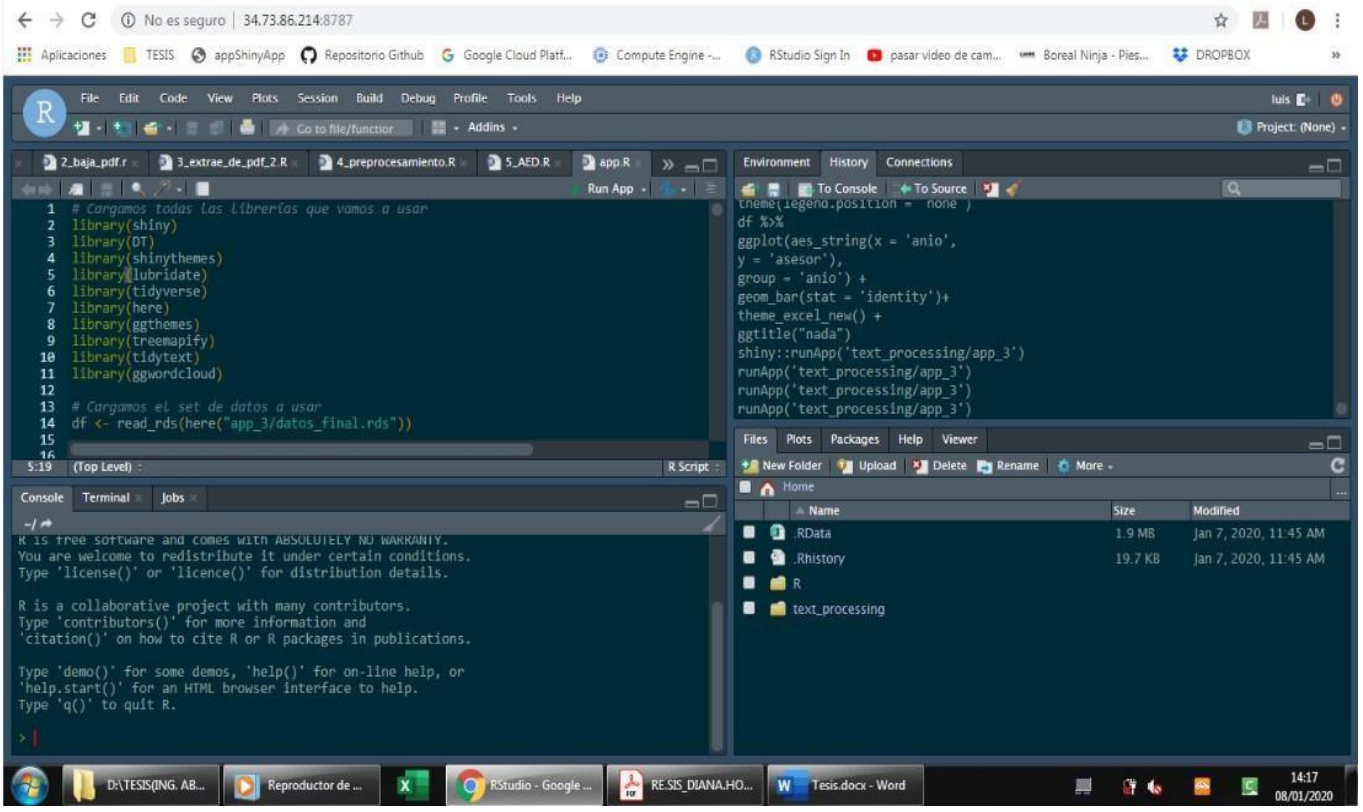
ANEXO 04: Creación de Máquina “Instance-text-mining”



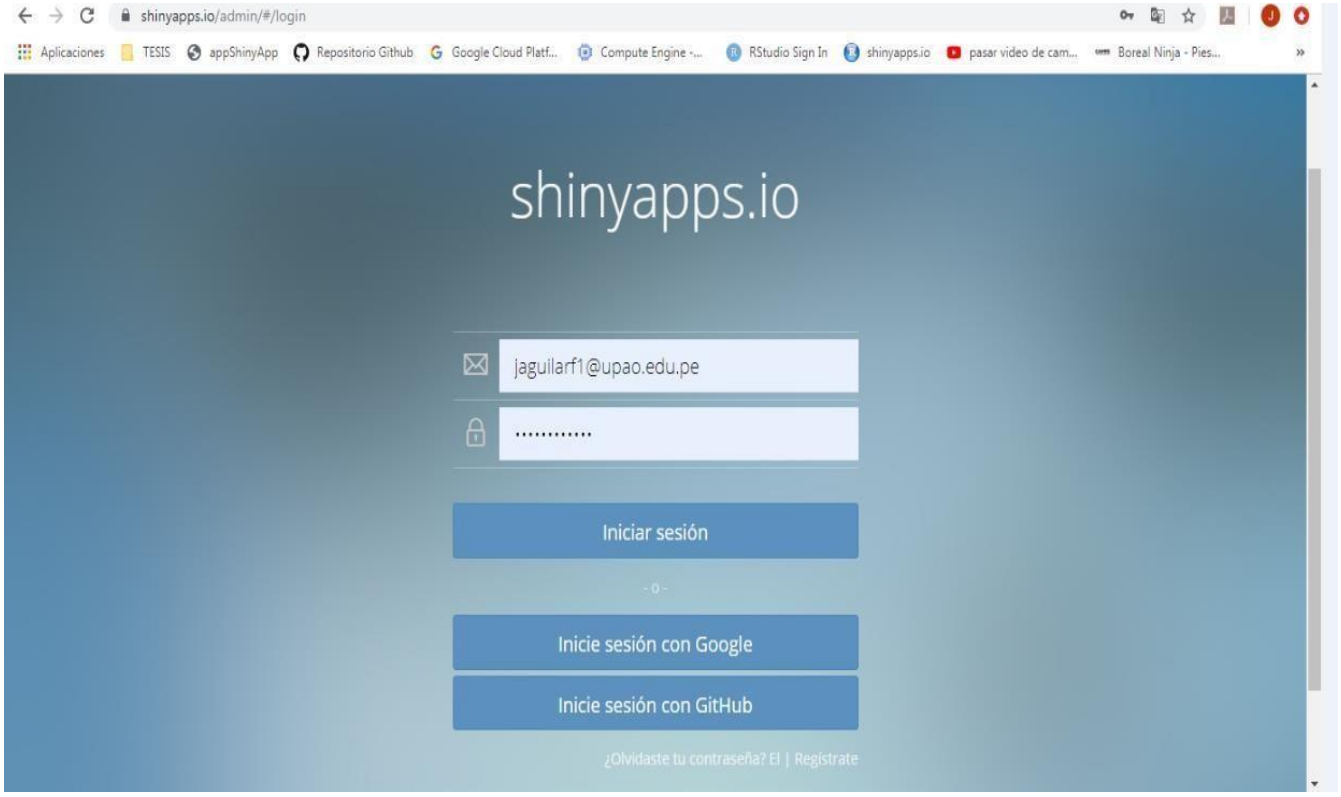
ANEXO 05: Registro de Entrada para el RStudio



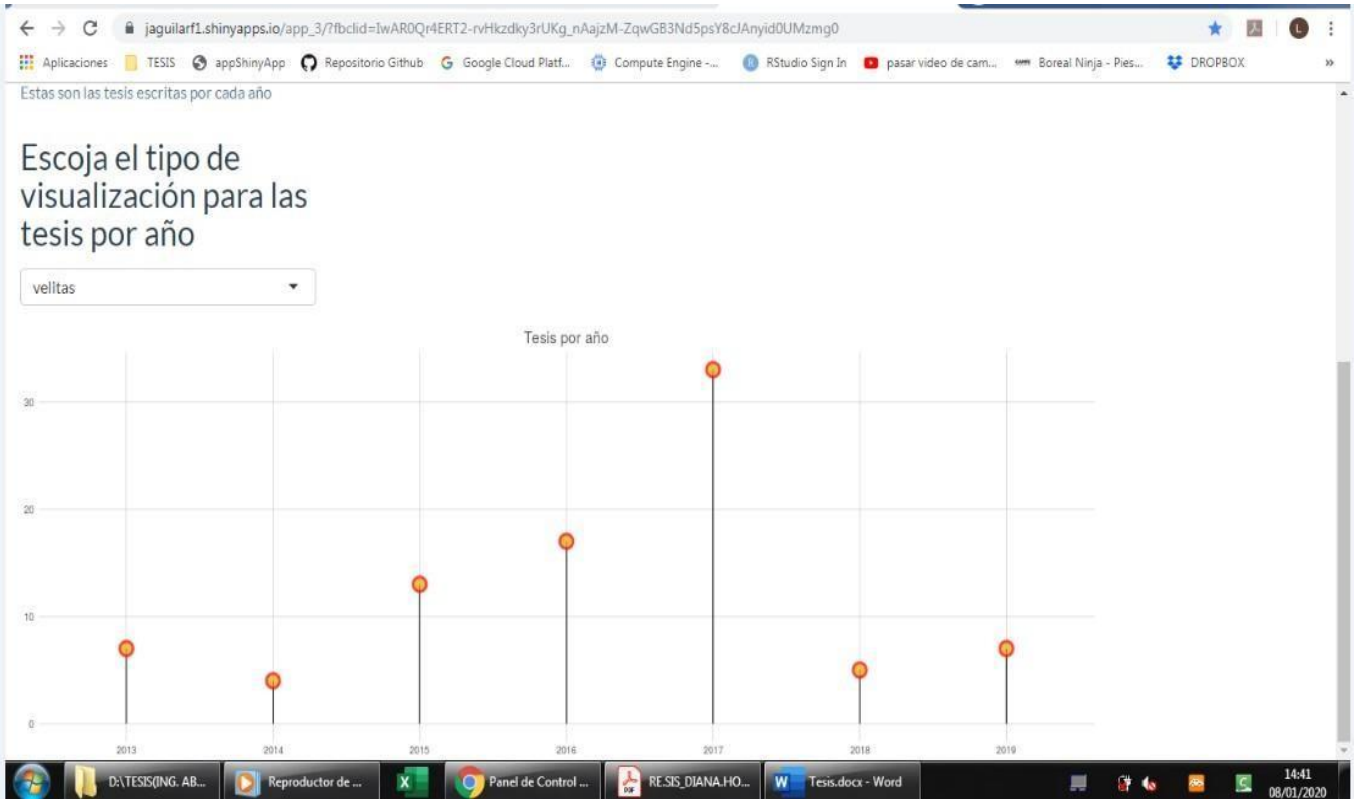
ANEXO 06: Inicio de la Plataforma en la Nube de RStudio



ANEXO 07: Creación de la Web en Shinyapps.io un framework de RStudio

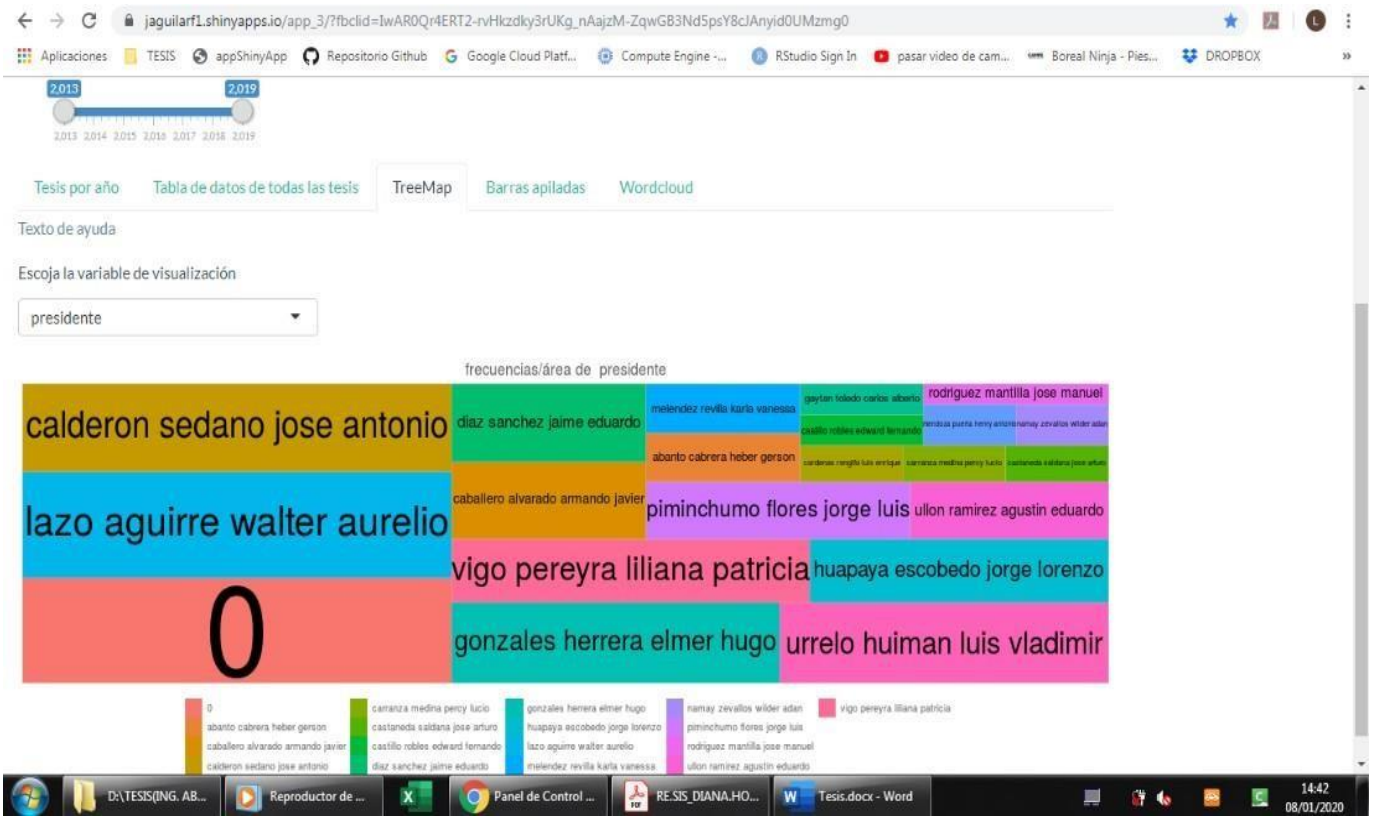


ANEXO 08: Gráfico en Shinyapps



ANEXO 09: Gráfico en Shinyapps

ANEXO 10: Gráfico en Shinyapps.



ANEXO 11: Gráfico en Shinyapps

ANEXO 12: Cuestionario sobre el Dashboard

CUESTIONARIO

El siguiente cuestionario trata sobre el funcionamiento del Dashboard. El cual tiene una escala del 1 al 5, siendo 1 muy malo y 5 muy bueno.

VALORACION	ESCALA
MUY BUENO	5
BUENO	4
REGULAR	3
MALO	2
MUY MALO	1

Facilidad con que se usa el Dashboard:

1. Para usted es fácil navegar dentro del Dashboard.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
2. La apariencia del Dashboard es agradable y fácil de entender.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
3. Cuando se solicita información al Dashboard, este despliega la información en el tiempo esperado.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
4. La búsqueda de información es sencilla
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
5. Como es la Interacción con el Dashboard.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo

Grado de Satisfacción del usuario

1. Califique la solución ofrecida
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
2. Que tan útil resulto la solución ofrecida.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
3. Califique el desempeño del Dashboard al momento de brindar la información ofrecida.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
4. La información mostrada le sirvió para conocer mejor el manejo de los proyectos de tesis en la universidad.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo
5. La información es confiable.
   ( ) Muy Bueno ( ) Bueno ( ) Regular ( ) Malo ( ) Muy Malo