Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Universidad del Bío-Bío
Facultad de Ingeniería
Depto. de Ingeniería Industrial

Profesor Guía:
Dr. Fredy Troncoso E.

Profesora Co-guía:
Dra. Macarena Valenzuela B.

“PREDICCION DE GENERO DE LOS AUTORES DE ARTICULOS CIENTIFICOS CON TECNICAS DE MINERIA DE DATOS PARA DETERMINAR LA PARTICIPACION EN LA CREACION DE CONOCIMIENTO CIENTIFICO POR GENERO”

“GENDER PREDICTION OF AUTHORS OF SCIENTIFIC ARTICLES USING DATA MINING TECHNIQUES TO DETERMINE THE PARTICIPATION IN THE CREATION OF SCIENTIFIC KNOWLEDGE BY GENDER”

Trabajo de Titulación presentado en conformidad a los requisitos
para obtener el título de Ingeniero Civil Industrial

Concepción, 20 de enero de 2023

Diego Yáñez Oyarce
Ingeniería Civil Industrial

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

DEDICATORIA

Este trabajo está dedicado a mis padres, a mi hermano y abuelos; los pilares fundamentales en mi vida, quienes me han acompañado incondicionalmente en todo momento y han hecho de mi la persona que soy el día de hoy. Cada paso que doy es gracias a ustedes, y sin su apoyo y cariño no todo sería de la misma manera. Los amo mucho.

A Ignacio, Matías y Raúl; con quien no solo conformamos un gran grupo de trabajo durante estos años de universidad, sino también una linda amistad que perdurará en el tiempo.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

AGRADECIMIENTOS

Agradezco al profesor Dr. Fredy Troncoso Escobar por su disposición, por acogerme como estudiante memorista e incorporarme como colaborador en el proyecto “Desarrollo de capacidades institucionales para la igualdad de género en I+D+i+e en la Universidad del Bío-Bío”, guiándome en todo momento en el camino para poder culminar mi proyecto de título.

Agradezco a mi profesora co-guía Dra. Macarena Valenzuela por la confianza al involucrarme y permitirme aportar en el proyecto. Por su dedicación y disposición a trabajar conmigo en cada una de las reuniones agendadas y su paciencia cuando se presentaban obstáculos en el camino.

Agradezco también a los profesores del Departamento de Ingeniería Industrial por la formación recibida en estos años de universidad. Al profesor Danilo Gómez, quien junto al profesor Fredy Troncoso han motivado el deseo de seguir el camino de la ciencia de los datos.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

RESUMEN

En este estudio, se documenta la creación de un algoritmo el cual es capaz de determinar la participación femenina y masculina en una base de datos extraída desde Web of Science para ayudar en la creación de herramientas que apoyen los estudios de cienciometría de las áreas de dirección de la universidad, algoritmo el cual fue programado en Rstudio.

Se contempla el uso de la metodología Knowledge Discovery in Databases (KDD) para realizar minería de datos con métodos estadísticos tradicionales (uso de un diccionario de nombres) y algoritmos de edición de cadenas de texto (distancia de Levenshtein) sobre el primer y segundo nombre de los autores.

La metodología se aplica sobre una base de datos de 12.000 artículos científicos filtrados por tema “género” en Sudamérica, identificando 50.300 autores. Al utilizar un diccionario de nombres considerado en otras investigaciones se dejan 7.975 autores sin clasificar (16%), sin embargo, incorporando un algoritmo de edición de cadenas de texto se puede disminuir ese número a 3.092 (6%), número el cual coincide con la cantidad de autores que no se les puede identificar su nombre, por lo que en realidad la metodología clasifica el 100% de los autores potenciales a ser clasificados con un porcentaje de acierto del 88,18%. El algoritmo entrega como resultado que la participación femenina corresponde a un 57% (26.825) y la participación masculina a un 43% (20.383) al omitir aquellos autores que no les identifica ni primer ni segundo nombre.

Palabras clave: metodología KDD, distancia de Levenshtein, creación de conocimiento científico por género, desigualdad de género.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

ABSTRACT

In this study, it is documented the creation of an algorithm of programming which is capable of determining the female and male participation in a database extracted from Web of Science to help in the creation of tools which would support cienciometry studies in areas directed by the university, algorithm wich was programmed in Rstudio. The use of the Knowledge Discovery in Databases (KDD) methodology is being considered to realize data mining with traditional statistical methods (the use of a name dictionary) and text distance edit (Levenshtein distance) over the authors first names. This methodology is applied on a database of 12.000 scientific articles filtered by the theme “gender” in South America, identifying 50.300 authors. By using a name dictionary found in others investigations 7.975 authors are left unclassified (16%), however, by incorporating text distance edit it is possible to diminish this number to 3.092 (6%), this number matches the amount of authors whose name can not be identified, so, this methodology classifies all of the potential authors to be classified with a percentage of success of 88,18%. The algorithm deliveries as a result that the female participation corresponds to a 57% (26.825) and the male participation to a 43% (20.383) by omitting those authors whose names are not identified.

Keywords: KDD methodology, Levenshtein distance, scientific knowledge creation by gender, gender inequality.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

TABLA DE CONTENIDO

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

INDICE DE FIGURAS

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

INDICE DE TABLAS

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPÍTULO 1: INTRODUCCIÓN

En el siguiente capitulo se expone la situación en la que actualmente se encuentra la Universidad del Bío-Bío: el origen del tema, contexto de la situación, problema y justificación y las preguntas, hipótesis de investigación, objetivos, ámbito del estudio y resultados esperados.

1.1 Origen del tema

El tema propuesto, se da gracias a la búsqueda de herramientas que puedan ayudar a realizar estudios cienciométricos incorporando el género de autores. Lo anteriormente expuesto, se da en el marco del proyecto “Desarrollo de capacidades institucionales para la igualdad de género en I+D+i+e en la Universidad del Bío-Bío”, en el cual participa como colaborador el profesor Fredy Troncoso Espinosa.

1.2 Contexto

La Vicerrectoría de Investigación y Postgrado de la Universidad del Bío-Bío (2021) señala que en el Primer Diagnóstico Institucional de Género (DIG) realizado durante el año 2020 por la Dirección General de Análisis Institucional (DGAI) y la Dirección General de Género (DIRGEGEN), se evidencian desigualdades muy grandes en un periodo de análisis de los últimos 10 años en la Universidad del Bío-Bío. Entre las principales brechas se pueden encontrar, por ejemplo, el conocimiento científico generado por mujeres versus el conocimiento científico generado por hombres y la composición del cuerpo académico de la universidad, donde el 69,32% corresponde a hombres y el 30,68% a mujeres.

Aunque la institución aún se encuentra débil en las dimensiones de perspectiva de género, durante el último trienio, la Universidad del Bío-Bío ha asumido responsablemente esta situación y ha estado avanzando en aspectos reglamentarios, en la difusión, formación y sensibilización de la equidad de géneros con la creación de la Dirección General de Género (DIRGEGEN) y de la Vicerrectoría de Investigación y Postgrado (VRIP). Estas dos entidades, asumen los procesos de generación y transmisión de conocimiento científico, creación artística, innovación y formación de capital humano avanzado, propiciando un accionar

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

transversal articulado con otras dependencias académicas-administrativas institucionales potenciando el desarrollo de actividades en I+D+i+e., contribuyendo a establecer las bases necesarias para la transversalización del enfoque de género, lo que se contempla en el Plan General de Desarrollo de la Universidad (2020-2029) en el cual se incorporó la perspectiva de género.

1.3 Justificación

En la última década, el movimiento feminista ha llevado a la agenda pública la temática de género en todas partes del mundo. Esta intervención, ha llevado a su auge los estudios de género.

El Estado, ha recogido las demandas ciudadanas y ha implementado una política pública que se enfoca en acortar las brechas de género en la participación de las mujeres en la ciencia. Es por ello por lo que abre un fondo para Innovación en Educación Superior en temas de Género llamado InES Género. Este proyecto se centra en abordar las principales dimensiones del Modelo de Madurez Huella de Género de Comunidad Mujer y el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación (2021), en el cual se señala que, la carrera científica de las mujeres se encuentra con obstáculos de diversa índole como por ejemplo: predominio masculino en la estructura del poder, poca valoración de la producción del conocimiento científico generado por mujeres o la permanencia de estereotipos de género arraigados en la comunidad científica.

Las universidades que participan para adjudicarse el fondo InES Género son 12, siendo la Universidad del Bío-Bío una de ellas. También se pueden encontrar instituciones interesadas asociadas al proyecto como: la Corporación Mujeres, el Servicio Nacional de la Mujer, el Instituto Nacional de Propiedad Industrial INAPI – Chile, la Corporación Chilena de la Madera (Ñuble y Bío-Bío), New Genesis, Denham Consulting, Asesorias e Inversiones SANTIBU, la Universitat Autónoma de Barcelona (España) y la Red de Ciencia, Tecnología y Género (México).

Dentro de la iniciativa con la cual participa la Universidad del Bío-Bío se estipulan tres objetivos, siendo la Dirección General de Análisis Institucional (DGAI) la

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

encargada de realizar estudios cienciométricos en I+D+i+e con enfoque de género y crear un modelo de indicadores de género para abordar uno de los objetivos.

La DGAI, al percatarse de la inexistencia de la variable que permitiese identificar la participación femenina y masculina en bases de datos extraídas desde bases de datos de revistas científicas como Scielo, Web of Science, Scopus, entre otras, se encuentra con la dificultad de no poder realizar la distinción de género por disciplina, tema, país, revistas más citadas, entre otros indicadores.

En este marco, es que esta tesis devela la necesidad de crear un algoritmo que permita identificar el género de los autores y autoras de acuerdo a su primer y segundo nombre. De esta forma, la creación de esta herramienta puede ayudar en los análisis para que la Dirección General de Análisis Institucional cree modelos autosustentables para el seguimiento y monitoreo de indicadores de género en la universidad con capacidad de ser replicados por otras instituciones.

La culminación exitosa de este estudio puede ayudar a generar conocimiento importante no solo para el proyecto mencionado, sino también para otras futuras investigaciones en género o incluso en otras áreas.

1.4 Pregunta de investigación

Lo expuesto anteriormente lleva a plantear la siguiente pregunta de investigación:

¿Es posible **determinar** la participación por género que hay en la creación de conocimiento científico de algún tema en específico?

1.5 Hipótesis de investigación

1. Aplicar técnicas de Minería de Datos, en específico, un diccionario de nombres y algoritmos de edición de distancias a bases de datos obtenidas desde Web Of Science, permite obtener una buena predicción del género de los autores que publican artículos científicos para conocer los desequilibrios por género en la creación de conocimiento.
2. Incorporar algoritmos de edición de cadenas de textos cómo la distancia de Levenshtein permite utilizar diccionarios de nombres menos extensos para la predicción de género.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

1.6 Objetivos

Los objetivos planteados para este proyecto son:

1.6.1 Objetivo general

Diseñar un algoritmo de programación que sea capaz de cuantificar la participación en la creación de conocimiento científico de algún tema en específico por género.

1.6.2 Objetivos específicos

1. Realizar una revisión bibliográfica referente a técnicas de minería de datos.
2. Extraer una base de datos de publicaciones de artículos científicos.
3. Plantear una metodología para predecir el género de los autores de artículos científicos.
4. Validar los resultados de la metodología.

1.7 Ámbito del estudio y fuente de los datos

El estudio abarca el área de la ciencia de los datos y considera la metodología Knowledge Discovery in Databases (KDD) para extraer conocimiento significativo de una base de datos estructurada obtenida a partir del motor de búsqueda de información Web of Science, por lo tanto, se contemplan fuentes secundarias de información para la obtención de los datos. Los artículos científicos que se analizan corresponden a aquellos filtrados por tema “género” y por región “Sudamérica”.

Se decide utilizar WoS ya que dentro de las variables que incorpora en su base de datos se puede encontrar el nombre completo del autor, variable que no se encuentra en bases de datos extraídas desde Scopus.

1.8 Resultados esperados

En lo que respecta a la investigación, se espera que aplicaciones del área de la ciencia de datos puedan aportar realizar estudios de cienciometría incorporando género a través del desarrollo de un algoritmo capaz de extraer información implícita de una base de datos obtenida a través de Web of Science y clasificar a los autores por género para que las áreas de dirección de la Universidad del Bío-Bío puedan analizar y potenciar sus estrategias y medidas destinadas a acortar las brechas y

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

avanzar hacia la igualdad de género, específicamente abordando la dimensión de las brechas, debilidades y desequilibrios en la investigación científica por género.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPÍTULO 2: REVISIÓN BIBLIOGRÁFICA

En esta sección se exponen los principales aspectos teóricos y estudios similares realizados para predecir el género de los autores de artículos científicos.

2.1 Estado del arte

La literatura sugiere dos formas de poder resolver el problema de la predicción de género: realizar minería de datos con métodos estadísticos tradicionales y realizar minería de datos con modelos de machine learning.

La mayor parte de los esfuerzos en la investigación acerca de género se habían centrado en el lenguaje hablado (Key 1972, Trudgill 1972, Labov 1990,Eckert 1997). Sin embargo, desde aproximadamente 1995 se ha visto una explosión en lo que es la investigación de categorización automatizada de textos (Sebastiani, 2001).

Según Koppel (2002), hasta ese momento, existían escasas evidencias en que la diferencia entre la escritura masculina y femenina era lo suficientemente pronunciada como para convertirse en un algoritmo de aprendizaje automático para categorizar el género de los autores. Koppel categoriza el género de los autores de acuerdo con la manera en que ellos escriben y las diferencias que pudiese haber entre el lenguaje empleado por hombres y el lenguaje empleado por mujeres en una base de datos de 566 documentos obtenidos del British National Corpus. Para ello se emplea una variante del algoritmo de Gradiente Exponencial de Kivinen & Warmuth (1997), el cual considera un modelo lineal de predicción que logra una precisión de 80%.

Argamon, Goulain, Horton, & Olsen (2009) tambien predicen el género del autor a partir de la forma en que escriben. Este enfoque, considera la implementación de algoritmos de aprendizaje automático que están estrechamente relacionados con métodos de análisis estadísticos, los cuales buscan explotar las diferencias entre la frecuencia de palabras utilizadas por hombres versus las que utilizan las mujeres. La ventaja de utilizar modelos de machine learning por sobre métodos estadísticos tradicionales radica en que se pueden crear modelos predictivos con desempeños comprobables, sin embargo, para entrenar y testear estos modelos se requiere tener

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

categorizada una base de datos con anterioridad, lo cual no es el caso de la base de datos que se puede obtener desde Web of Science.

La implementación del modelo contempla una aplicación de Super Vector Machine (SVM) sobre una base de datos de literatura francesa de 300 escrituras hechas por hombres y 300 escrituras hechas por mujeres. El clasificador logra una precisión de un 90% y de esta aplicación se puede resaltar, por ejemplo, que las mujeres utilizan de manera más frecuente pronombres y elementos de polaridad negativa, mientras que los hombres utilizan de manera más frecuente cuantificadores numéricos.

Bird, Klein & Loper (2009) crean una herramienta de procesamiento del lenguaje natural en donde incorporan un listado de 7.576 nombres distintos categorizados como “hombre”, “mujer” o “ambos” en caso de ser ambiguo. Son de los primeros autores en incorporar un diccionario de nombres para poder clasificar los nombres por género, sin embargo, no es muy práctico.

Figura 1:Diccionario de nombres introducido por Bird, Klein y Loper

```
genderdata::kantrowitz
 ## Source: local data frame [7,579 x 2]
 ##
 ##             name     gender
 ## 1          aamir       male
 ## 2          aaron       male
 ## 3          abbey     either
 ## 4          abbie     either
 ## 5          abbot       male
 ## 6         abbott       male
 ## 7           abby     either
 ## 8          abdel       male
 ## 9          abdul       male
 ## 10   abdulkarim       male
 ## ..           ...        ...
```

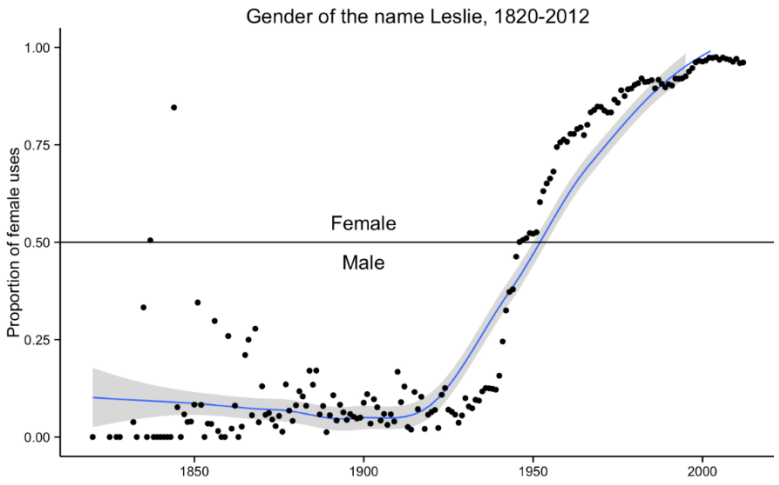
Fuente: (Bird, Klein, & Loper, 2009)

Mullens & Blevins (2015) se cuestionan de donde proviene el problema de la predicción de género a partir de los nombres. Señalan que, si hoy se tuviese que clasificar por género el nombre “Leslie”, casi con un 100% de seguridad todas las personas dirían que “Leslie” es un nombre de mujer, lo cual está en lo cierto. Sin embargo, si se hubiese tenido que clasificar el género de “Leslie” hace poco más de 70 años atrás, esto no sería de igual manera. Eso se debe a que el género

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

convencional del nombre Leslie cambió en el transcurso de los años. En 1900, alrededor del 92% de los bebés nacidos en los Estados Unidos que se llamaban Leslie se clasificaron como varones, mientras que en 2000 alrededor del 96% de los bebes llamados Leslie nacidos en ese año se clasificaron como mujeres. En la **Figura 2** se puede observar cómo ha ido variando la clasificación del género Leslie.

Figura 2: Cambio en el género del nombre Leslie



Fuente: (Mullen & Blevins, 2015).

Mullens & Blevins (2015) corrigen métodos existentes anteriormente como el de Bird et al. (2009). Señalan que los géneros de los nombres no son estáticos y que van cambiando con el tiempo y que, además, se debe utilizar una amplia base de datos. Es por lo anterior que, contemplan en su diccionario, 1.603.026 nombres del Social Security Administration de los Estados Unidos.

Wais (2016) documenta la creación de genderize.io, la cual es una de las API más prácticas desde el punto de vista del usuario, ya que se puede encontrar tanto en Rstudio como en Python. Esta API considera una versión de prueba y una versión de pago, lo cual dificulta su uso como tal. Wais (2016), utiliza una base de datos de 142.848 nombres distintos que recopiló, y su algoritmo incorpora la posibilidad de poder entregarle un texto e identificar cuantos nombres hay en ese texto, tal como se muestra en la **Tabla 1**

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 1: Ejemplo consulta al algoritmo de Kamil Wais

Texto	Resultado	
"Diego Antonio es un autor"	Diego	Masculino
	Antonio	Masculino
"Antonio Diego es otro autor"	Antonio	Masculino
	Diego	Masculino

Fuente: Elaboración propia haciendo uso de genderizeR de Kamil Wais.

Kaushik, Gupta, Pratim Roy, & Prosad Dogra (2018) presentan un conjunto de experimentos sobre la predicción del género de los usuarios de Twitter basado en las caracteristicas de los twitters que publican (palabras y emojis que utilizan). Para ello contemplan modelos de machine learning y realizan estos estudios sobre seis idiomas diferentes. Los porcentajes de acierto son de alrededor del 65 por ciento.

Reddy, Vardhan, GopiChand, & Karunakar (2018) realizan un estudio para crear perfiles demográficos incorporando la edad, sexo, ubicación, idioma y formación académica de los autores. Los investigadores proponen caracteristicas como los caracteres que utilizan, palabras que utilizan, temas que abordan, sintactica que emplean o legibilidad de la escritura y es a través de ella que toman la decisión de clasificar a un autor como hombre o mujer.

2.2 Marco teórico

Al revisar lo expuesto en la **sección 2.1** se puede señalar que el utilizar un diccionario de nombres es la mejor opción para el problema planteado ya que la base de datos de Web of Science no se encuentra labelizada para diseñar algún modelo de machine learning. Sin embargo, dentro de los puntos débiles se puede encontrar que no se puede comprobar de manera automática el desempeño, se necesita un diccionario amplio y exacto de nombres y que se necesita utilizar un diccionario de nombres propios de la región de estudio.

Dentro de esta sección se exponen los principales aspectos teóricos de esta investigación.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

2.2.2 Web scraping

Según Zhao (2017), el web scraping es una técnica que recolecta datos desde la internet abierta y los guarda en un archivo de cualquier formato para su posterior análisis. Esta técnica es una de las más antiguas, pero se ha hecho cada vez más común debido a lo eficiente que es (Glez Peña, Lourenco, López Fernández, Reboiro Jato, & Fernández Riverola, 2014). La recolección de datos puede ser de manera manual o automática utilizando algún algoritmo de programación (comúnmente conocido como robots) creado por algún usuario (Zhao, 2017).

Delgado y Repiso (2013) señalan que existe un duopolio entre Scopus y Web Of Science en lo que es la cobertura de publicaciones científicas y que ambas opciones pueden cubrir gran parte de artículos creados en la región. Glez Peña et al. (2014), señala que las API son quienes abordan las principales tareas de recuperación de datos. Las API es una interfaz que reúne un conjunto de paquetes de software para que otro software pueda ejecutarlos e interactuar (Cámara de Comercio de Bogotá, 2019).

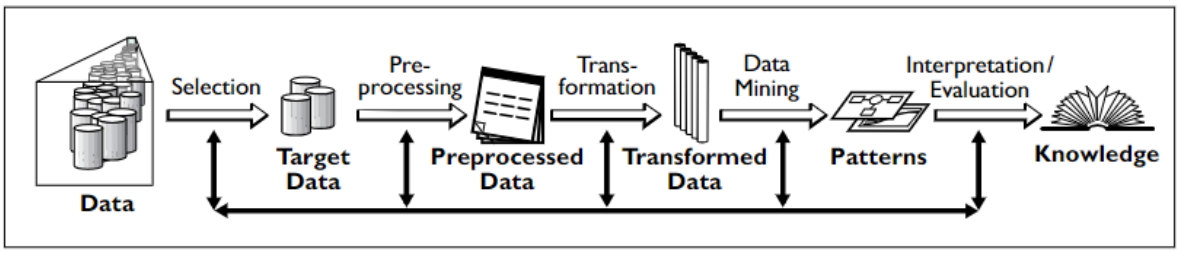
2.2.2 Knowledge Discovery in Databases

El Text Mining ha aumentado su auge en los últimos años debido a la gran cantidad de documentos disponibles en la web sobre los cuales existe una creciente necesidad de organizar y obtener el conocimiento contenido en textos (Guerrero, 2020).

Fayyad, Piatesky-Shapiro & Smyth (1996) señalan que la metodología “KDD” corresponde al proceso completo de descubrimiento de conocimiento en bases de datos. Dentro de esta metodología se encuentran las etapas de selección de atributos, preprocesamiento, transformación de atributos, minería de datos e interpretación de resultados.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Figura 3: Proceso KDD



Fuente: Fayyad et al. (1996)

Selección de variables

Timarán-Pereira, Hernández-Arteaga, Caicedo, Hidalgo-Troya, & Alvarado-Pérez, (2016) señalan que, una vez identificado el conocimiento relevante desde el punto de vista del usuario final, se selecciona solamente el conjunto de datos sobre el cual se realiza la extracción de conocimiento.

Preprocesamiento

En esta etapa se realiza la limpieza de las variables. Se remueven por ejemplo datos fuera de rango, se utilizan estrategias para tratar datos faltantes, datos duplicados, entre otras. La continua interacción del analista con el jefe o interesado en el proyecto es de suma importancia, ya que se utilizan tanto criterios estadísticos como personales (Timarán-Pereira et al.,2016).

Transformación de variables

En esta etapa se buscan características útiles en una variable para dar paso a otras (Timarán-Pereira et al., 2016).

Minería de datos

El objetivo de esta etapa es obtener el conocimiento de interés que se encuentra oculto tras los datos (Timarán-Pereira et al., 2016).

Validación e interpretación

Se debe comprobar la fiabilidad del modelo en un banco de ejemplos y luego de eso ser interpretado para establecer las principales conclusiones de los hallazgos (Beltrán-Martínez, 2001)

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

2.2.3 Distancia de Levenshtein

Corresponde al algoritmo de edición de textos más conocido (Viny Christiani, Rudy, & Dali S., 2018) e incluso grandes exponentes como Microsoft lo utiliza en sus correctores gramaticales (Mego Lizana & Cespedes Bravo, 2021).

La distancia de Levenshtein considera el costo de una determinada operación, como por ejemplo insertar, borrar o sustituir caracteres, por lo tanto, es adecuado para la tarea de encontrar similitudes entre dos cadenas de texto y corregir errores gramaticales (Viny Christiani et al, 2018). Corresponde al costo de operación necesario para transformar una palabra en otra (Damerau, 1964). En la **Ecuación 1** se puede visualizar la ecuación de la distancia de Levenshtein.

Ecuación 1: Distancia de Levenshtein

$$
\begin{aligned}
D(i,j) := \min[&D(i-1,j) + w_d, \\
&D(i,j-1) + w_i, \\
&D(i-1,j-1) + w_r] \\
D(i,0) &:= D(i-1,0) + w_d \\
D(0,j) &:= D(0,j-1) + w_i
\end{aligned}
\Bigg\} \forall i,j > 0
$$

$$D(0,0) := 0$$

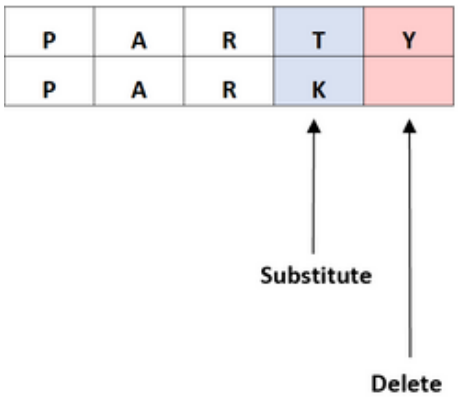
Fuente: (Schimke, Vielhauer, & Dittmann, 2004).

Schimke et al. (2004) explican el algoritmo de la siguiente forma. S_1 y S_2 corresponden a las dos cadenas de texto que se involucran en el algoritmo, donde i y j corresponden al tamaño de cada una de ellas y W_i, W_d y W_r, corresponden al costo de insertar, eliminar o reemplazar un carácter. El valor resultante D se vuelve más pequeño a medida que S_1 y S_2 se parecen más.

Para entender de mejor manera lo expuesto anteriormente se presenta un ejemplo en la **Figura 4.**

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Figura 4: Ejemplo distancia de Levenshtein



Fuente: (Statologos, s.f.)

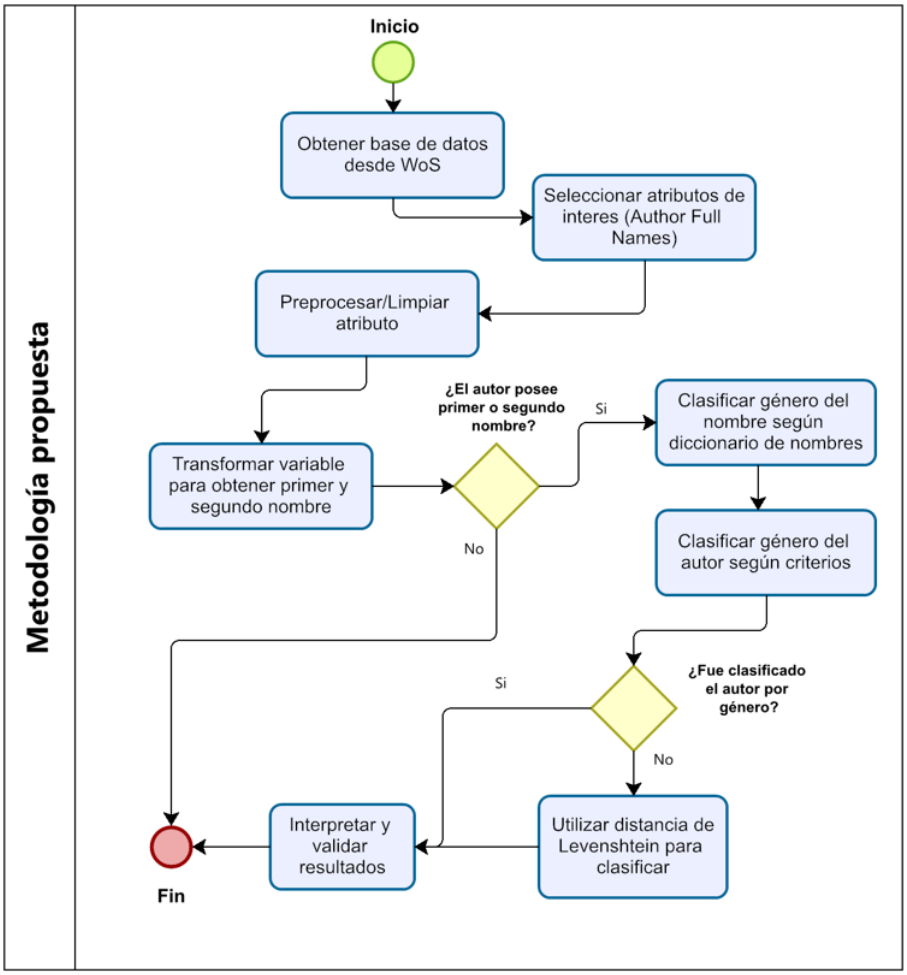
Como se puede apreciar, para llegar de la palabra “party” a “park”, es necesario sustituir la “T” por la “K” y eliminar la “Y”. En total, se ha realizado una sustitución, una eliminación y cero inserciones. Por lo tanto, al existir dos ediciones, la distancia de Levenshtein es igual a dos.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPÍTULO 3: DESARROLLO

La metodología que se propone para clasificar el género de los autores de una base de datos de artículos científicos se basa en la obtención de la base de datos y la aplicación de metodología KDD a la misma, la cual se puede dividir en las etapas de selección de variables, preprocesamiento o limpieza, transformación de variables, minería de datos y validación e interpretación de los resultados. En la **Figura 5** se puede visualizar gráficamente lo descrito anteriormente.

Figura 5: Diagrama de flujo de la metodología propuesta.



Fuente: Elaboración propia.

En esta sección, se exponen los mecanismos aplicados en el desarrollo del trabajo. Para no extender el tamaño de la investigación, no se adjuntarán líneas de código propias del lenguaje de programación utilizado.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.1 Obtención y descripción de la base de datos

Para la obtención de datos a partir de bases de datos que contienen artículos científicos, se selecciona como motor de búsqueda y extracción web la base de datos de *"web of science"*, la cual permite realizar una búsqueda por tema, región de publicación, lenguaje y varios filtros que puedan ser de interés. Para efectos de este estudio se filtran artículos científicos con el tema *"genero"* publicados en Sudamérica, obteniendo una base de datos en formato *".csv"* con un total de 12.000 registros y 71 variables.

Las variables o atributos que posee cada artículo científico se explican en la **Tabla 2**. Se omite la explicación de variables nulas.

Tabla 2: Descripción de la base de datos

N°	Variable	Descripción
1	**Publication Type**	Puede ser **J** si es un artículo de revista científica que aporta nuevos conocimientos o **S** si es una revisión sistemática de literatura ya existente (Jiménez Ávila, 2015).
2	**Authors**	Contiene el apellido de los autores separado por coma del primer y segundo nombre abreviados. Si hay más de un autor se separan por punto y coma (ej: Yáñez, DA; Oyarce, DA).
3	**Author Full Names**	Contiene el apellido de los autores separados por coma del primer y segundo nombre sin abreviar. Si hay más de un autor se separan por punto y coma (ej: Yáñez, Diego Antonio; Oyarce, Diego Antonio).
4	**Book Author Full Names**	Contiene el nombre completo de los autores de libros de revisiones sistemáticas.
5	**Group Authors**	Contiene el nombre del grupo el cual conforman los autores de artículos científicos en caso de que sean un grupo conformado de investigadores.
6	**Article Title**	Contiene el nombre del título del artículo científico.
7	**Source Title**	Contiene el nombre de la revista en cual fue publicado el artículo.
8	**Book Series Title**	Contiene el título de la serie a la cual conforma el libro de revisión sistemática.
9	**Book Series Subtitle**	Contiene el subtítulo de la serie a la cual conforma el libro de revisión sistemática.
10	**Language**	Contiene el lenguaje en el cual está escrito el artículo científico.
11	**Conference Type**	Contiene el nombre y número de edición del congreso en el cual fue expuesto el artículo científico (si es que aplicase).

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

12	**Conference Date**	Contiene la fecha del congreso en el cual fue expuesto el artículo científico (si es que aplicase).
13	**Conference Location**	Contiene el lugar en el cual fue desarrollado el congreso en el cual fue expuesto el artículo científico (si es que aplicase).
14	**Conference Sponsor**	Contiene el nombre de las instituciones que patrocina el congreso en el cual fue expuesto el artículo científico (si es que aplicase).
15	**Conference Host**	Contiene el nombre de las instituciones que organizan el congreso en el cual fue expuesto el artículo científico (si es que aplicase).
16	**Abstract**	Contiene el resumen la publicación.
17	**Researchers IDS**	Contiene el ID propio del autor del sistema de identificación ResearcherID. Permite asociar asociar producción científica a investigadores y cada base de datos de artículos científicos utiliza el suyo propio, Web Of Science utiliza este sistema (Borrego, 2013).
18	**ORCIDs**	Contiene un ID propio del autor que lo asocia a investigaciones. Es más general que el mencionado anteriormente ya que no depende de los sistemas de los cuales son propietarios bases de datos como Web Of Science (Denk, 2017).
19	**ISSN**	Es un número internacional normalizado propio de la publicación impresa el cual permite identificarla. No contiene ningún otro significado intrínseco (Centro Internacional de Registro de Publicaciones en Serie, 2017).
20	**eISSN**	Es un número internacional normalizado propio de la publicación digital el cual permite identificarla. No contiene ningún otro significado intrínseco (Centro Internacional de Registro de Publicaciones en Serie, 2017).
21	**ISBN**	Permite identificar el grupo, editor, titulo y digito de comprobación de un libro (Cruz Quintana, 2019)
22	**Publication Year**	Contiene el año de publicación del artículo científico.
23	**DOI**	Es el identificador más usado para identificar cualquier entidad digital. Conduce a la tipificación de un recurso incluso aunque su URL haya cambiado (Hospital a Domicilio , 2017).
24	**PubMed ID**	Contiene el ID propio de identificación del artículo en el portal PubMed (Trueba-Gómez & Estrada-Lorenzo, 2010).
25	**WOS ID**	Contiene el ID propio de identificación del artículo en el portal Web Of Science

Fuente: Elaboración propia.

3.2 Selección de atributos, preprocesamiento y transformación de variables

Se limpia la base de datos de variables que pudiesen no aportar información debido la presencia de una gran cantidad de datos nulos. Para ello se utiliza la función

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

"remove_empty" de Rstudio, la cual permite disminuir el espectro de 71 variables a 25, de las cuales se determinan aquellas que pudiesen contener información relevante para determinar el género de los autores. Dentro de las variables que pueden aportar información se pueden encontrar tres *"Researchers IDS", "ORCIDs"* y *"Author Full Names",* ya que estas tres contienen los nombres de los autores. En la **Tabla 3** se puede observar la disposición de los datos en las variables mencionadas de un registro de la base de datos. Se pueden encontrar cuatro autores, dos identificadores ResearcherID y dos ORCID, por lo tanto, las ID no sirven para poder determinar el género de una gran cantidad de autores ya que no todos los participantes poseen el identificador. La variable que se selecciona es *"Author Full Names".*

Tabla 3: Registro de la base de datos con tres potenciales variables para predecir género

ID	Author Full Names	Researchers IDS	ORCIDs
2	Salazar Torres, Virgilio Mariano; Goicolea, Isabel; Edin, Kerstin; Ohman, Ann	Mariano, Salazar/ABI-6855-2020	Mariano, Salazar/0000-0001-6935-9781; Goicolea, Isabel/0000-0002-8114-4705

Fuente: Elaboración propia.

Al estar trabajando con texto, se hace necesaria una estandarización de los datos y convertirlos a un formato que sea común para una manipulación más fácil en etapas posteriores (Doctor Bracho, 2018). Es por esto que se transforman los caracteres a minúscula, se remueven tildes, dígitos, dobles espacios y caracteres como, por ejemplo, guiones con las funciones *"tolower", "stri_trans_general", "str_replace_all"* y *"gsub".*

Se desanida cada autor presente en la publicación de manera que un registro de la base de datos sea un autor y no un registro sea un artículo científico. Para lo anterior, se utiliza como separador el punto y coma y es por esta razón que no se considera el punto y coma en la limpieza de caracteres anteriores. En la **Tabla 3** se puede observar la situación inicial y en la **Tabla 4** se puede observar la situación final de este proceso.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4: Situación final al desanidar los autores del conjunto de autores de un mismo artículo científico de la base de datos

ID Artículo Científico	**Nombre y Apellido**
2	Salazar Torres, Virgilio Mariano
2	Goicolea, Isabel
2	Edin, Kerstin
2	Ohman, Ann

Fuente: Elaboración Propia.

Al obtener cada autor separado del resto de autores de un mismo artículo científico, da paso a la transformación de variables para obtener variables latentes[1] y así clasificar el género a partir del nombre del autor. Para lo anterior, se utiliza como carácter separador la coma para desunir el nombre del apellido. Se utiliza el carácter espacio para separar el primer nombre del segundo nombre. En caso de que un autor no posea segundo nombre, esta variable se considera nula en el registro (tercer nombre o más se omiten). La situación que describe un ejemplo de la situación final del proceso descrito se puede visualizar en la **Tabla 5**.

Tabla 5: Obtención de las variables latentes "Primer Nombre" y "Segundo Nombre" a partir de "Author Full Names"

ID Artículo Científico	**Primer Nombre**	**Segundo Nombre**	**Apellido**
2	Virgilio	Mariano	Salazar Torres
2	Isabel		Goicolea
2	Kerstin		
2	Ann		Ohman

Fuente: Elaboración propia.

En caso de que la longitud del primer nombre o segundo nombre sea igual a uno significa que estos se encuentran abreviados con la inicial del nombre del autor y se elimina el dato de la variable correspondiente ya que no entrega información. Se eliminan 3.002 abreviaciones en primer nombre y 8.180 abreviaciones en segundo nombre. Este último paso, permite seguir con la siguiente etapa del proceso KDD.

[1] Variables Latentes: Variables no observadas en la base de datos que son inferidas a partir de otras.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.3 Minería de datos

La extracción del conocimiento de la base de datos se realiza a partir de los nombres de los autores, y, como sugiere la literatura, se determina la clasificación del género de un nombre según un diccionario de nombres ya clasificados anteriormente. Con los nombres ya clasificados por género, se clasifica el género de los autores considerando siete criterios conformados según las probabilidades que puedan tener tanto el primer como segundo nombre del autor. Sin embargo, no todos los nombres se pueden encontrar en el diccionario. Para corregir este problema, se recurre a la Distancia de Levenshtein para buscar similitudes de los nombres no clasificados con los pertenecientes al diccionario.

3.3.1 Obtención del diccionario de nombres

Para obtener un diccionario de nombres ya clasificados anteriormente se hace uso de listados de nombres entregados por el INE de España, Registro Civil de Chile, paquete “genero” y paquete “GenderizeR”.

Nombres más comunes en Chile y España

El Servicio de Registro Civil e Identificación de Chile (2015) y el Instituto Nacional de Estadísticas de España (2022) entregan, cada uno, una base de datos con los 100 nombres más frecuentes en los respectivos países. Estos listados son los primeros en incorporar al diccionario de nombres a elaborar y se considera una probabilidad de 100 por ciento para el género clasificado para un nombre en específico por el INE de España o el Registro Civil de Chile, es decir, si por ejemplo Diego se encuentra en los 100 nombres de hombre más frecuentes de Chile, se considerará que el nombre Diego tiene una probabilidad de 100 por ciento de ser masculino.

Paquete genero de Rstudio

Marin Diaz (2020) crea la función *“genero”* disponible en el paquete *“genero”* de Rstudio la cual, en su versión de prueba, contiene un diccionario de los 100 nombres más frecuentes en español y 100 nombres más frecuentes en portugués (la API de pago contiene un total de 9.000 nombres en español y 50.000 nombres en

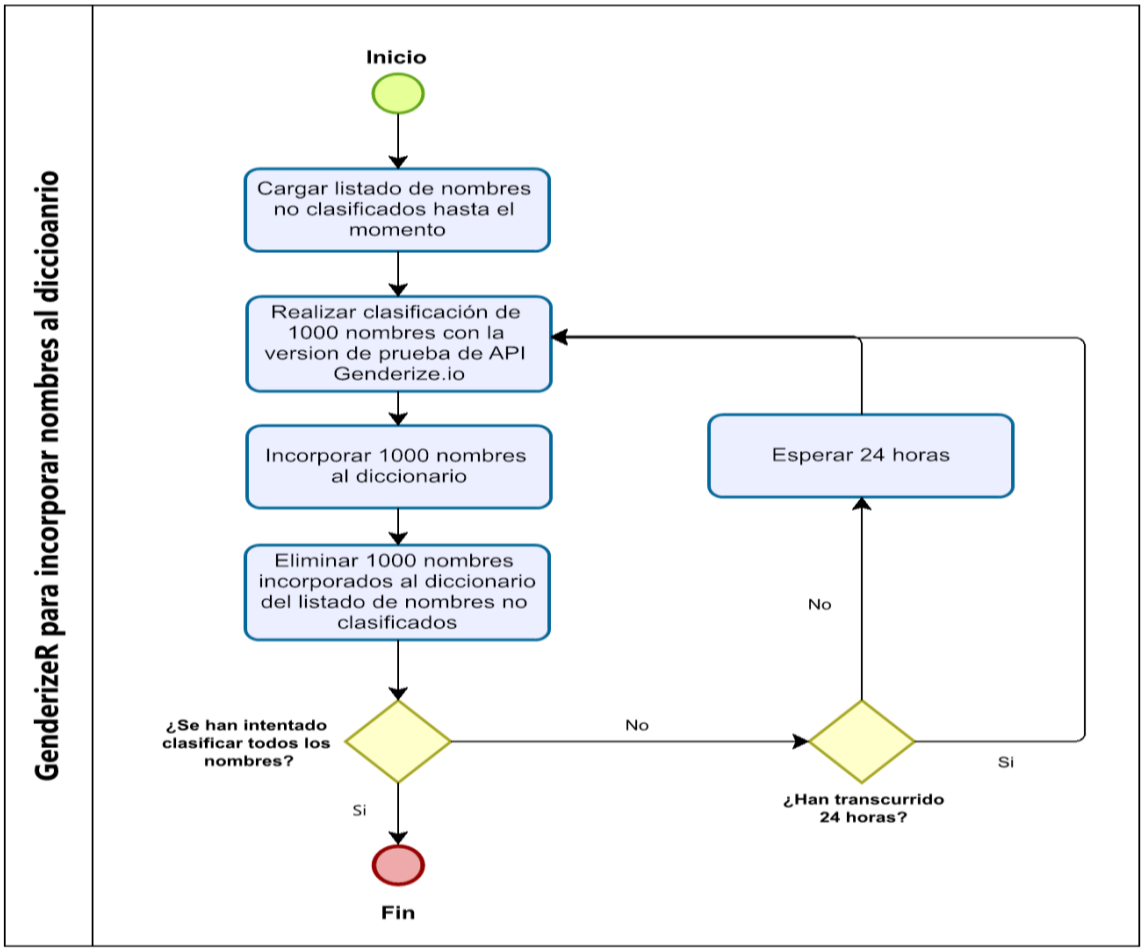
Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

portugués). Estos nombres también serán cargados al diccionario de nombres con una probabilidad de 100 por ciento para el género clasificado.

Paquete GenderizeR de Rstudio

Esta herramienta, corresponde a la API de Genderize.io la cual entrega la gran mayoría de la clasificación de nombres presentes en la base de datos extraída desde web of science. Sin embargo, esta API también es de pago, pero permite realizar 1.000 consultas cada 24 horas, por lo tanto, se elabora un segundo algoritmo que se ejecuta cada día. Lo anterior permite incorporar 1.000 nombres cada día al diccionario de nombres. La **Figura 6** detalla el algoritmo descrito anteriormente.

Figura 6: Diagrama de flujo del proceso de llenado de diccionario con el uso de API Genderize.io



Fuente: Ilustración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

En total, junto a los nombres presentes en la base de datos del INE de España, del Registro Civil de Chile, del paquete *"genero"* y del paquete *"genderizeR"* de la API de Genderize.io, se llega a un total de 9.966 nombres clasificados por género, entregando también en el diccionario la probabilidad de ser predichos de tal manera. En la **Tabla 6** se pueden visualizar 5 instancias pertenecientes al diccionario de nombres y cabe destacar que no todos los nombres presentes en la base de datos de web of science se encuentran en el diccionario.

Tabla 6: Diccionario de nombres elaborado (5 registros)

Nombre	**Género**	**Probabilidad**
Jose	Male	1
Marta	Female	1
Migueis	Male	0.67
Madia	Female	0.95
Damm	Male	0.88

Fuente: Elaboración propia.

3.3.2 Clasificación de los nombres

La función de Rstudio que ayuda a clasificar los nombres es *"agrep"*, y para hacerlo necesita como parámetros de entrada los datos de las variables *"Primer Nombre"* y *"Segundo Nombre"*. En la **Tabla 5** de la **sección 3.2** se puede apreciar una situación inicial. Se busca similitud entre los nombres y el diccionario. En la **Tabla 7** se puede encontrar, a modo de ejemplo, el resultado al encontrar la similitud entre los nombres en la base de datos y el diccionario en cinco instancias.

Tabla 7: Resultado al clasificar los nombres de cinco instancias según el diccionario de nombres

Primer Nombre	**Género Primer Nombre**	**Probabilidad Primer Nombre**	**Segundo Nombre**	**Género Segundo Nombre**	**Probabilidad Segundo Nombre**
Virgilio	Male	1	Mariano	Male	1
Marcelo	Male	1	Rosa	Female	1
Anazelia	NA	NA	NA	NA	NA
Milaynne	NA	NA	Christina	Female	1
Michele	Female	0.78	Alberto	Male	1

Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.3.3 Clasificación del género de los autores

Para clasificar el género, se conforman siete criterios para determinar finalmente el género del autor según el primer y segundo nombre de cada uno de ellos. En la **Tabla 8** se expone cada una de las reglas por orden en el que se clasifica.

Tabla 8: Criterios para predecir género de los autores.

N°	Condición	Género del autor
1	Probabilidad del género del primer nombre es igual a 100 por ciento.	Género del primer nombre.
2	Probabilidad del género del primer nombre es menor a 100 por ciento, pero la probabilidad del segundo es 100 por ciento.	Género del segundo nombre
3	Primer nombre no clasificado por el diccionario y probabilidad del género del segundo 100 por ciento.	Género del segundo nombre
4	Probabilidad del género del primer nombre mayor a 75 por ciento y segundo nombre no clasificado.	Género del primer nombre
5	Probabilidad del género del primer nombre mayor a 75 por ciento y probabilidad del segundo menor a 75%.	Género del primer nombre
6	Primer nombre no clasificado por el diccionario y probabilidad del género del segundo mayor a 75%	Género del segundo nombre
7	Probabilidad del género del primer nombre menor a 75 por ciento y probabilidad del segundo mayor a 75%.	Género del segundo nombre

Fuente: Elaboración propia.

Se decide no clasificar autores con nombres que posean una probabilidad menor a 75 por ciento, ya que viene siendo “la mitad de la mitad”. Lo anterior es por el motivo de cómo se conforman las probabilidades para cada nombre en el diccionario bajo concepto de probabilidad simple, en donde se dividen los casos posibles con los casos totales. En la **Tabla 9** se puede ver un ejemplo de lo expuesto anteriormente.

Tabla 9: Calculo de las probabilidades para género de nombres

Nombre:	**Personas llamadas Diego que son mujeres en una base de datos**	**Personas llamadas Diego que son hombres en una base de datos**
Diego	49	51
Resultado: El nombre Diego es clasificado como masculino con una probabilidad del 51 por ciento.		

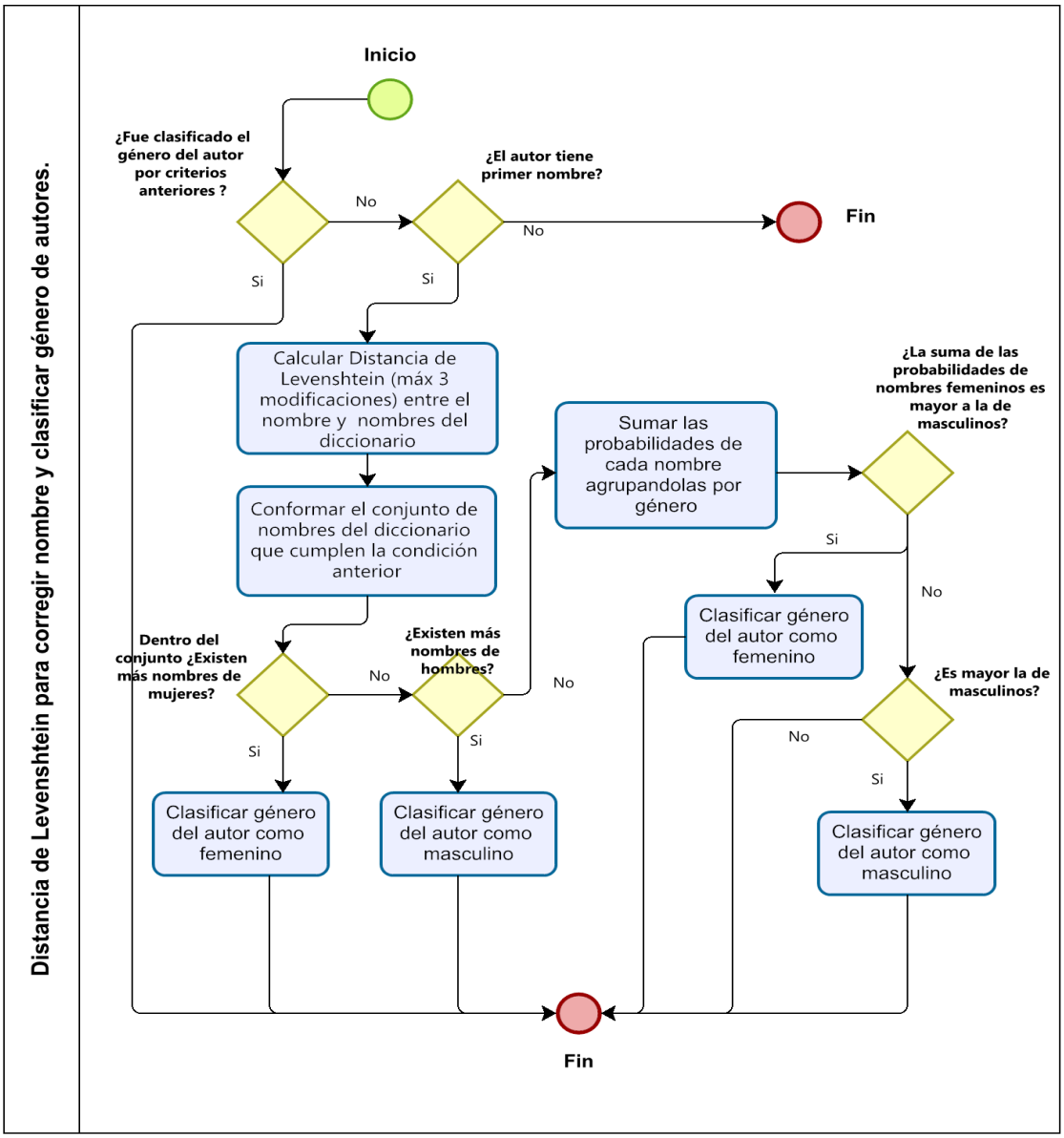
Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.3.4 Corrección del nombre para clasificar género del autor

Para clasificar el género de los autores que puedan tener un nombre no contenido en el diccionario o una baja probabilidad de clasificación, se utiliza la Distancia de Levenshtein. Esta herramienta permite encontrar nombres que puedan ser parecidos a otros, que se encuentren mal escritos por error de tipeo o que sean su símil en otro lenguaje o región. La **Figura 7** describe cómo funciona la edición de caracteres en el algoritmo.

Figura 7: Diagrama de flujo del proceso de clasificación de género por Distancia de Levenshtein



Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPÍTULO 4: RESULTADOS

4.1 Resultados de aplicación de la metodología

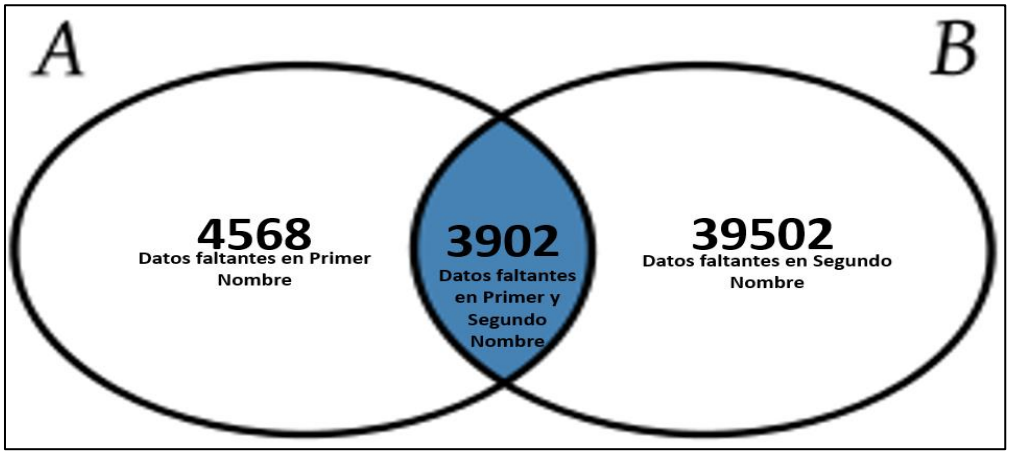
A través de la herramienta que proporciona el sitio web de Web of Science se obtiene de manera manual, una base de datos de 12.000 artículos científicos organizados de forma tabular en un archivo *".csv"* que pueden tener uno o más autores de diferentes partes del mundo. El haber filtrado por región *"Sudamérica",* no implica que no pudiese haber investigadores con distintas nacionalidades y diferentes lenguajes publicando en la región. Lo anterior puede resultar en desaciertos en la predicción realizada gracias a un diccionario de nombres que solo contemple nombres de Sudamérica. Zangwill (1908) introduce el concepto de *"Melting Pot"* y desde entonces se ha utilizado para referirse a Estados Unidos como un país con una gran mezcla cultural. Utilizar una robusta base de datos con nombres inscritos en los Estados Unidos que entrega Wais (2016) se considera la mejor opción para mitigar los efectos anteriores y tener una gran diversidad de culturas en el diccionario.

El preprocesamiento realizado en la **sección 3.2** permite identificar 50.300 autores en los 12.000 artículos científicos. El algoritmo es capaz de detectar y eliminar aquellos autores que han escrito de manera abreviada sus nombres, y se eliminan 3.002 abreviaciones en primer nombre sumando en total 4.568 con los datos faltantes en el primer nombre. Por otra parte, se encuentran 8.180 abreviaciones en segundo nombre, sumando 39.502 con los datos faltantes en segundo nombre.

Como se muestra en la **Figura 8**, resultan 3.092 autores que no poseen ni primer, ni segundo nombre. El género de estos autores no se puede clasificar bajo ningún criterio expuesto en la **sección 3.3.2.**

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Figura 8: Datos faltantes en Primer Nombre y Segundo Nombre después de limpieza



Fuente: Elaboración propia

Para efectos de orden, se traslada (cuando es nula) a la columna "*Primer nombre*", todos aquellos *"Segundo Nombre"* (que no son nulos).

En la **Tabla 10** se exponen los resultados al clasificar los géneros de los primeros y segundos nombres, en donde se puede encontrar una clasificación como nombre masculino, femenino o no clasificado en caso de que los nombres de los autores no se encuentren en el diccionario elaborado.

Tabla 10: Resultados de la clasificación de los géneros de primeros nombres y segundos nombres

Primeros Nombres		**Segundos Nombres**	
Primeros nombres clasificados como femeninos por diccionario	22515	Segundos nombres clasificados como femeninos por diccionario	4251
Primeros nombres clasificados como masculinos por diccionario	19086	Segundos nombres clasificados como masculinos por diccionario	3455
Primeros nombres no existentes en el diccionario (no clasificados por diccionario)	4797	Segundos nombres no existentes en el diccionario (no clasificados por diccionario)	8906
Total	**46398**	**Total**	**16612**

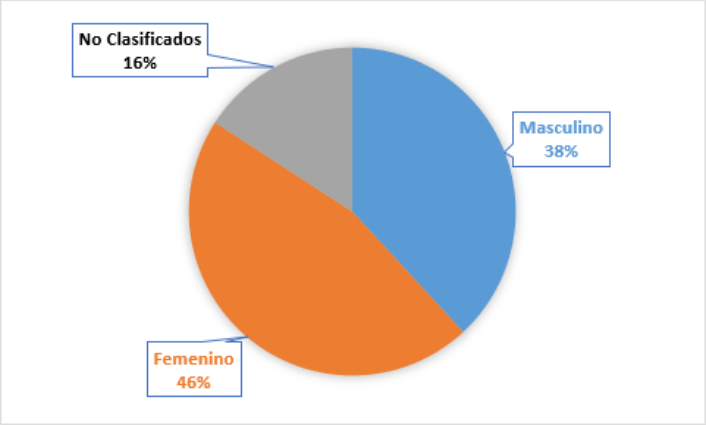
Fuente: Elaboración propia.

Al clasificar utilizando los criterios expuestos en la **sección 3.3.3** para conseguir finalmente la variable latente *"género del autor"*, se puede obtener que existe un total de 23.145 autores de género femenino, 19.180 de género masculino y 7.975

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

autores que no se han clasificado como ninguno de los dos. La proporción anterior se puede visualizar en la **Figura 9.**

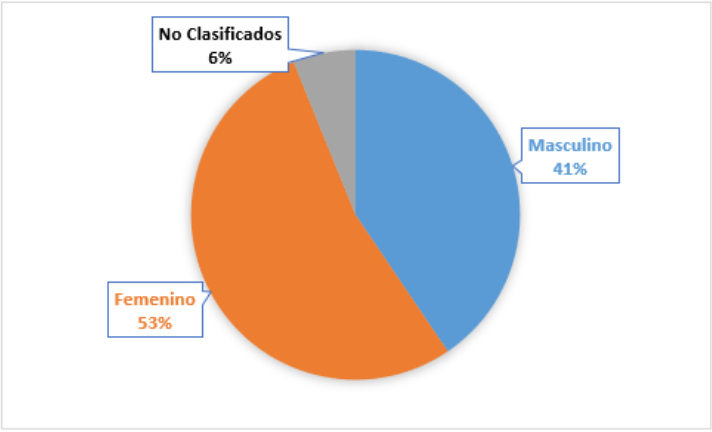
Figura 9: Resultado de la clasificación del género de los autores por criterios



Fuente: Elaboración propia.

Después de aplicar la distancia de Levenshtein en el algoritmo se obtiene un total de 26.825 autores que se han clasificado como femenino, 20.383 que se han clasificado como masculino y se logra reducir a 3.092 la cantidad de autores que no se han clasificado por género, lo que significa un 6 por ciento del total de autores en la base de datos y reducir en un 38,77 por ciento la cantidad de autores no clasificados por género antes de aplicar esta herramienta. Lo anterior se puede visualizar en la **Figura 10.**

Figura 10: Resultado de la clasificación del género de los autores por criterios y Distancia de Levenshtein



Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

4.2 Validación de resultados

Dado que no se sabe desde un comienzo la verdadera categorización del género real de los autores, la única manera de comprobar los resultados es a través de una búsqueda manual del artículo científico por internet y llegar a sitios web que puedan dar pistas del género del autor (cómo, por ejemplo, páginas de universidades, perfiles de LinkedIn, perfiles de investigador, etc.).

Para validar los resultados se ha extraído una muestra aleatoria de 110 autores de la base de datos clasificados por cada uno de los criterios. De la muestra aleatoria se han considerado 10 autores clasificados por género bajo el primer criterio, 10 clasificados por el segundo y tercero, 10 clasificados por el cuarto, 9 clasificados por el quinto, 10 clasificados por el sexto, 10 clasificados por el séptimo y 51 clasificados por la Distancia de Levenshtein. Al realizar la validación de manera manual, se obtiene un desempeño del 88,18 por ciento y si se analiza solo los 51 registros clasificados por la Distancia de Levenshtein se obtiene un porcentaje de acierto del 88,24 por ciento. El desempeño de la metodología se puede visualizar en la **Tabla 11.**

Tabla 11: Validación de resultados

Acierto	97
No Acierto	13
Desempeño	88,18%

Fuente: Elaboración propia.

4.3 Prueba eliminando nombres del diccionario

Se realiza un experimento en el algoritmo eliminando del diccionario de nombres todos los nombres de los autores utilizados para validar los resultados en la **sección 4.2** pero solo con la distancia de Levenshtein. El resultado al clasificar nuevamente a los 110 autores se muestra en la **Tabla 12.**

Tabla 12: Desempeño del algoritmo al eliminar todos los nombres del diccionario

Acierto	85
No Acierto	25
Desempeño	77,27%

Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPITULO 5: CONCLUSIONES GENERALES, DISCUSIÓN DE RESULTADOS Y RECOMENDACIONES

La contribución de este trabajo consiste en el diseño de una herramienta para la Dirección General de Género y la Vicerrectoría de Investigación y Postgrado de la Universidad del Bío-Bío que permite cuantificar la participación por género en la creación de conocimiento científico para que las áreas de dirección de la universidad puedan realizar sus estudios cienciométricos incorporando la variable género.

La literatura analizada propone el uso de dos técnicas para poder predecir el género de los autores. El uso de Machine Learning y trabajar sobre el contenido del informe (abstract o el contenido completo) para identificar patrones en la forma de escribir de los autores. También sugieren el uso de estadística tradicional para realizar análisis y minería de datos sobre los nombres de los autores. Se concluye que utilizar un diccionario de nombres y trabajar sobre la variable “Author Full Names” es la mejor opción para poder finalmente clasificar el género de los autores, ya que todas las personas en el mundo poseen un nombre que podría ser asociado a un género. Se da cumplimiento al objetivo específico número uno.

El motor de búsqueda de la meta base de datos de artículos científicos Web of Science resulta ser una buena herramienta para poder *“scrapear”* datos desde la web. Sin embargo, al utilizar la versión gratis que posee en su sitio web, no se automatiza la obtención de la base de datos. Para incorporar la API en Rstudio se requiere de una versión de pago que permita la conexión entre el entorno de programación y la base de datos de artículos científicos. Al evaluar la posibilidad de utilizar otros motores de búsqueda (como Scopus) para hacer más amplio el alcance del algoritmo, se obtiene que las bases de datos que entregan no poseen el nombre de los autores para ser rescatados de alguna variable. Lo expuesto anteriormente, da culminación al objetivo específico número dos.

La metodología expuesta en la **sección 3** permite realizar un análisis de los datos presentes en fuentes de información con la necesidad de ser organizados para luego obtener un conocimiento a partir de ellos.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Uno de los grandes aportes en la literatura es el de Wais (2016), sin embargo, esta es solo una herramienta que clasifica nombres, no realiza algo similar a lo expuesto en este trabajo y no serviría por sí solo para el problema de cuantificar la participación en la creación del conocimiento científico por género. Como se puede apreciar en la **Tabla 1**, la herramienta de Wais es capaz de identificar nombres y clasificarlos, sin embargo, el algoritmo no es capaz de identificar que en realidad Diego Antonio y Antonio Diego son dos personas en ese ejemplo, y se estarían reconociendo cuatro personas cuando en realidad son dos. A pesar de ello, utilizar el diccionario que emplea, es una buena alternativa para lograr un diccionario de nombres más extenso.

Con respecto a resultados, si se compara la **Tabla 10** con lo explicado para la **Figura 9** se puede apreciar que no existe una diferencia muy significativa con respecto a si se clasificara y sacaran conclusiones solamente considerando el género del primer nombre. El considerar un diccionario principalmente compuesto por nombres del diccionario de Wais, se puede utilizar para compararlo con la metodología expuesta en esta investigación. Al incorporar la distancia de Levenshtein para poder clasificar aquellos autores que poseen nombres no contenidos en el diccionario o nombres clasificados con baja probabilidad, se obtiene una gran mejora en los resultados.

En la **Figura 9** se puede apreciar que el diccionario de Wais deja muchos autores sin clasificar mientras que en la **Figura 10** se puede evidenciar que, si se incorpora un algoritmo de edición de caracteres como la distancia de Levenshtein, se puede recuperar un gran porcentaje de clasificaciones.

Tabla 13: No utilizar distancia de Levenshtein versus si utilizar

	Sin distancia de Levenshtein	**Con distancia de Levenshtein**
Autores no clasificados por género	7.975	3.092
Autores no clasificados por género (%)	16%	6%

Fuente: Elaboración propia.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

El algoritmo, al incorporar la edición de caracteres deja sin clasificar 3.092 autores, sin embargo, como se expone en la **Figura 8** de la **sección 3.3.4** estos 3.092 autores corresponden a aquellos que no poseen ni primer, ni segundo nombre en la base de datos que no se pueden categorizar bajo ningún criterio. Al coincidir el número de autores sin clasificar con el número de autores que no poseen primer nombre ni segundo nombre, se concluye que se recupera el 100% de los nombres que no han sido clasificado por los criterios. Además de eso, los clasifica con un gran porcentaje de acierto (88,24%). Lo anterior, lleva a la pregunta de qué ocurriría en una base de datos en donde se da la situación que ningún nombre coincide con los contemplados en el diccionario ya que evidentemente nombres existen millones en el mundo y no siempre estarán contenidos en el diccionario. Al clasificar los mismos casos utilizados para validar los resultados en la **sección 4.2** pero yendo al extremo eliminando todos los nombres del diccionario y utilizando solo la edición de caracteres se obtiene que el desempeño del algoritmo es de un 77,27%. Lo anterior permite concluir que incorporar algoritmos que busquen similitudes de cadenas de texto posibilita utilizar un diccionario de nombres menos extenso que los expuestos en la literatura, manteniendo un buen desempeño en la predicción.

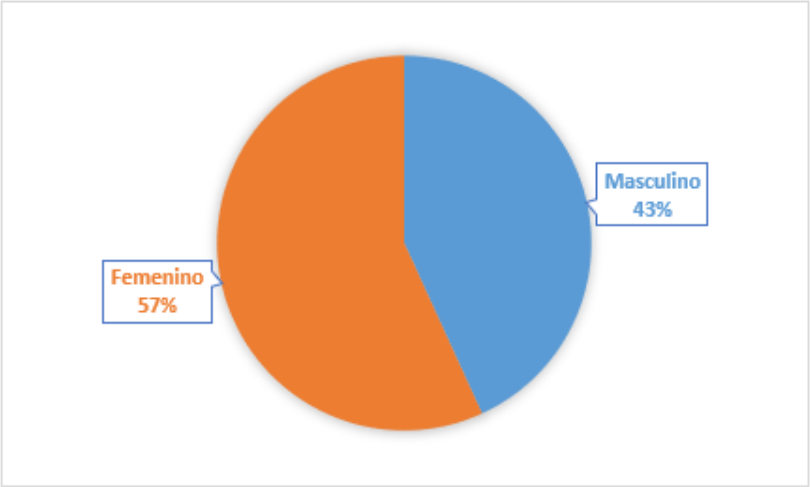
Todo lo expuesto anteriormente da cumplimiento a cada uno de los objetivos específicos planteados en la **sección 1.4**. y se da cumplimiento al objetivo general.

Aplicar la metodología a la base de datos permite concluir con un acierto del 88,18% que la participación en la creación de conocimiento científico por género al filtrar por tema “género” y región “Sudamérica” está dada por 57% participación femenina y 43% participación masculina (omitiendo autores no clasificados). Evidentemente la base de datos utilizada fue a modo de ejemplo y se puede utilizar cualquier base de datos proveniente desde Web of Science. A pesar de omitir de la estadística aquellos autores no clasificados, estos datos solo corresponden al 6% de los autores pertenecientes a la base de datos, por lo que, clasificar el otro 94% sigue siendo significativo para que las áreas de la Universidad del Bío-Bío puedan conocer la participación femenina y masculina y realizar los estudios de cienciometría para elaborar sus estrategias frente a la desigualdad de género en la creación de

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

conocimiento científico. La clasificación por género en la base de datos de WoS se puede visualizar en la **Figura 11.**

Figura 11: Participación femenina y masculina en artículos científicos relacionados a género en Sudamérica.



Fuente: Elaboración propia.

Se hace mención en que se podría mejorar la forma de obtener la base de datos y en incorporar más nombres al diccionario. La DIRGEGEN y la VRIP podrían evaluar e incorporar el uso de la API de pago para de esta forma, incluir esta herramienta en el código de programación y automatizar el proceso de la obtención de datos al hacerse dentro del entorno de programación. Por otro lado, se puede ir actualizando constantemente el diccionario de nombres para hacerlo más amplio y variado.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

CAPÍTULO 6: REFERENCIAS

Argamon, S., Goulain, J.-B., Horton, R., & Olsen, M. (2009). Vive la Différence! Text Mining Gender Difference in French Literature. *Digital Humanities Quarterly*, 1.

Beltrán-Martínez, B. (2001). *Minería de datos.* Puebla: Benemérita Universidad Autónoma de Puebla.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly.

Borrego, Á. (abril de 2013). *Sistemes d'identificació unívoca d'investigadors.* (U. d. Barcelona, Ed.) Barcelona, España: Edicions de la Universitat de Barcelona.

Cámara de Comercio de Bogotá. (2019). El mundo conectado por las API. Obtenido de http://hdl.handle.net/11520/22728

Centro Internacional de Registro de Publicaciones en Serie. (2017). *¿Qué es el número ISSN?* Obtenido de International Standard Serial Number International Centre: https://www.issn.org/es/comprender-el-issn/que-es-el-numero-issn/

Centro Internacional de Registro de Publicaciones en Serie. (24 de Mayo de 2017). *El e-ISSN se convierte en una mención obligatoria según la nueva lista de características de calidad editorial de Latindex*. Obtenido de International Standard Serial Number International Centre: https://www.issn.org/es/el-e-issn-se-convierte-en-una-mencion-obligatoria-segun-la-nueva-lista-de-caracteristicas-de-calidad-editorial-de-latindex/

Comunidad Mujer. (2021). *Huella de Género*. Obtenido de Comunidad Mujer: https://comunidadmujer.cl/huella-de-genero/

Cruz Quintana, F. (2019). El ISBN y su utilidad para la investigación bibliográfica. *Bibliographica*, 172-188.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 171-176.

Delgado, E., & Repiso, R. (2013). The Impact of Scientific Journals of Communication: Comparing Google Scholar Metrics, Web of Science and Scopus. *Scientific Journal of Media Education*, 45-52.

Denk, C. (2017). *ORCID, ResearcherID, Scopus Author ID - ¿Qué son y para qué sirven?* Sevilla: Editorial Universidad de Sevilla.

Doctor Bracho, E. A. (2018). *Técnicas Estadísticas en Mineria de Textos [Tesis para optar al grado de licenciatura en Matemáticas, Universidad de Sevilla].* Recuperado el 6 de enero de 2023, de https://idus.us.es/bitstream/handle/11441/77508/Doctor%20Bracho%20Elena%20TFG.pdf?sequence=1&isAllowed=y.

Eckert, P. (1997). Gender and sociolinguistic variation. *Linguistics*, 64-75.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the acm*, 50-62.

Glez Peña, D., Lourenco, A., López Fernández, H., Reboiro Jato, M., & Fernández Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics, Volume 15, Issue 5*, 788–797.

Guerrero, J. (2020). Aplicación de técnicas de minería de texto para el descubrimiento de relaciones conceptuales entre los trabajos de grado de la Universidad de Nariño. Buenos Aires, Argentina.

Hospital a Domicilio . (2017). La importancia y necesidad del Digital Object Identifier (DOI). *Hospital a Domicilio* , 185-187.

Instituto Nacional de Estasticas de España. (17 de Mayo de 2022). *Instituto Nacional de Estadisticas*. Obtenido de Apellidos y nombres más frecuentes. Resultados:

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990

Jiménez Ávila, J. M. (2015). Tipos de publicaciones científicas. *Medigraphic*, 58-67.

Kaushik, P., Gupta, A., Pratim Roy, P., & Prosad Dogra, D. (2018). EEG-Based Age and Gender Prediction Using Deep BLSTM-LSTM Network Model. *IEEE Sensors Journal*, 2634 - 2641.

Key, M. R. (1972). inguistic behavior of male and female. *De Gruyter*. doi:https://doi.org/10.1515/ling.1972.10.88.15

Kivinen, J., & Warmuth, M. (1997). Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 1-64.

Koppel, M. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 401-412.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*.

Marin Diaz, J. P. (9 de Septiembre de 2020). *Estimate Gender from Names in Spanish and Portuguese*. Obtenido de https://cran.rstudio.com/web/packages/genero/genero.pdf

Mego Lizana, J. A., & Cespedes Bravo, S. M. (2021). *Algoritmo para la corrección de textos en español basados en los algoritmos Metaphone y Distancia de Levenshtein [Tesis para optar al grado de Ingeniero de Sistemas].* Recuperado el 9 de enero del 2023 de https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/92285/Mego_LJA-Cespedes_BSM-SD.pdf?sequence=1&isAllowed=y.

Mullen, L., & Blevins, C. (2015). Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction. *Digital Humanities*, 3.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Reddy, T. R., Vardhan, B. V., GopiChand, M., & Karunakar, K. (2018). Gender Prediction in Author Profiling Using ReliefF Feature Selection Algorithm. *Intelligent Engineering Informatics*, pp 169–176.

Schimke, S., Vielhauer, C., & Dittmann, J. (2004). Using Adapted Levenshtein Distance for On-Line Signature Authentication. *17th International Conference on Pattern Recognition* (págs. 931-934). Cambridge: IEEE Computer Society.

Sebastiani, F. (2001). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*.

Servicio de Registro Civil e Identificación. (30 de octubre de 2015). *Datos.gob*. Obtenido de Servicio de Registro Civil e Identificación: https://datos.gob.cl/dataset/9438

Statologos. (s.f.). *Cómo calcular la distancia de Levenshtein en R (con ejemplos)*. Obtenido de Statologos: Obtenido el 10 de enero de https://sites.google.com/site/algoritmossimilaridaddistancia/distancia-de-levenshtein

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo, Z. S., Hidalgo-Troya, A., & Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, 63-86. Universidad Cooperativa de Colombia.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*.

Trueba-Gómez, R., & Estrada-Lorenzo, J.-M. (2010). La base de datos PubMed y la búsqueda de información científica. *Elsevier*, 49-63.

Vicerrectoría de Investigación y Postgrado. (2021). Concurso de desarrollo de capacidades institucionales para la igualdad de género en el ámbito de la I+D+i+e en instituciones de educación superior. Concepción: Universidad del Bío Bío.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Viny Christiani, M., Rudy, R., & Dali S., N. (2018). Fast and Accurate Spelling Correction Using Trie and Damerau-levenshtein Distance Bigram. *TELKOMNIKA, Vol.16, No.2*, 827-833.

Wais, K. (2016). Gender Prediction Methods Based on First Names with genderizeR. *The R Journal*, 17-37.

Zangwill, I. (1908). *The Melting Pot.* Nueva York: DigiCat.

Zhao, B. (2017). Web Scraping. *Encyclopedia of Big Data*, 1-3.