Revista Internacional de Aprendizaje
ISSN: 2575-5544 (versión impresa), ISSN: 2575-5560 (versión electrónica)
Volumen 9, Número 2, 2023
https://doi.org/10.18848/2575-5544/CGP/v09i02/75-95

Un *Dataset* Internacional Acerca de Nombres, Género y Frecuencias en Damegender. Contando Hombres y Mujeres en Debian GNU/Linux

David Arroyo Menéndez, URJC, España

Recibido: 29/07/2023; ***Aceptado:*** 25/08/2023; ***Publicado:*** 05/09/2023

Resumen: La igualdad de género es el quinto objetivo de desarrollo sostenible (ODS) para Naciones Unidas. Esta igualdad puede ser lograda midiendo, analizando datos y, creando buenas políticas con los resultados. Muchos estudios de género cuentan hombres y mujeres para explicar la posible desigualdad, por ejemplo, artículos de investigación, puestos de trabajo, calles, etc. El método tradicional de investigación es usar APIs comerciales con datos propietarios sin idea acerca de cómo los datos fueron recogidos. Los datos pueden también ser recogidos desde Wikipedia, estudios lingüísticos, sitios científicos, u oficinas estadísticas. Este enfoque está basado en recoger Datasets Abiertos (Open Datasets) que incluyen nombre, género y frecuencia desde muchas instituciones estadísticas. Así, las tareas abordadas están basadas en unificar formatos, procesar datos y, crear pruebas para medir la precisión de los nuevos *datasets*. El *dataset* usado cubre más de 20 países en el mundo occidental trayendo miles de nombres con una precisión de acierto mayor del 90%. Esto permitirá medir brecha de género a estudiantes y académicos interesados en el fenómeno sin costes y de una manera reproducible y más personas estarán contribuyendo a eliminar la brecha de género. El Software Libre y los datos provistos por instituciones estadísticas hacen posible producir investigación reproducible por pares.

Palabras Clave: *Herramientas de Detección de Género desde el Nombre, Brecha de Género, Feminismo de Datos, Libertad de Datos.*

Introducción

En este artículo abordaremos la pregunta de investigación: ¿Cuál es ratio de hombres y mujeres en la distribución Debian GNU/Linux? Pero para ello, primero un poco de información de contexto del interés y el enfoque qué tiene este trabajo para la sociedad. Naciones Unidas tiene como objetivo reducir la brecha de género , para eso, el primer paso es poder medirla ("si no se puede medir, no se puede mejorar") para medir brecha de género resulta económico utilizar ordenadores, así recordando la literalidad de (Boehm 2002) "Software Engineering Economics is an invaluable guide to determining software costs, applying the fundamental concepts of microeconomics to software engineering".

Así, el software y los datos libres reducen más los costes. Muchas personas e instituciones utilizan Software Libre como LibreOffice o Ubuntu GNU/Linux para evitar costes de licencias en productos similares como Microsoft Windows o Microsoft Word. Esto creará una competición en el mercado con algún ganador, evitando pagos y generando beneficios desde una marca, tal y como ocurre en los navegadores con Firefox y/o Chrome.

A través del uso de los nombres personales, uno puede inferir el sexo o género de una persona en artículos académicos, libros, periódicos y muchas interacciones en Internet. Así, detectar el género desde los nombres puede ser un camino estratégico para medir la (brecha de género, también.

Muchos usuarios y usuarias están, hoy en día, usando APIs tal y como Genderize, GenderAPI, Namsor, NameApi, Wikipedia, u otras soluciones libres: NLTK (Loper y Bird, 2002), R Gender, Gender Detector and Gender Computer .

Las soluciones libres tradicionales tienen un bajo número de nombres debido al uso de un solo fichero que quizás solo pertenece a un país siendo software no mantenido a través del tiempo. Por otro lado, Wikipedia almacena pocos nombres por país.

Sin embargo, la brecha de género es un problema reconocido en Naciones Unidas y el mercado IT está liderando grandes desigualdades en economía y brecha de género. Este artículo presenta datos recogidos para asistir en la solución a un número de problemas (buscadores, infiriendo género en ficheros CSV, nombres en diferentes países, creación de nuevos datasets) encarados por la industria así como otros problemas que carecen de muchas soluciones industriales como contar hombres y mujeres en GitHub, listas de correo, etc.

Un estudio previo (Karimi at al. 2016) mostraba una comparativa de datasets como un modo de mejorarla precisión, comparando herramientas que usan diferentes datasets públicos (SSA , IPUMS , namdict , etc).

Las posibles aportaciones a este trabajo realizadas en este artículo serían aumentar el número de nombres y la atención a la diversidad en cuestiones como el género no binario o minorías culturales.

El dataset o quizás más bien DameGender (el software para crear el *dataset*) podría ser aplicado en artículos de las diversas disciplinas donde operan las herramientas de detección de género a partir del nombre (Sun et al. 2019): ingeniería de software (Vasilescu, Capiluppi y Serebrenik, 2012), lingüística (Hutson 2016) y (Al–Zumor, 2009), bibliometría (Holman, Stuart–Fox y Hauser, 2018), periodismo (Mislove et al. 2011), (Niemi 2017) y (Sainz de Baranda 2014), ... Con vocación de ser puntuados muchos de ellos como de ciencia reproducible (Peng 2011) si se tienen las precauciones correctas de la definición. La estructura de este artículo es la siguiente:

La sección Marco Teórico presenta el vocabulario y las filosofías aplicadas en la investigación principal, para dotar de compresión al contexto del estudio realizado. La sección Estado del Arte presenta la investigación principal midiendo brecha de género y herramientas de detección de género desde el nombre.

La denominada Diseño da el vocabulario y la filosofía acerca de cómo elegir fuentes y encarar las problemáticas de diversidad construyendo un *dataset*.

La sección Midiendo Brecha de Género explica una aplicación para este *dataset*: medir la brecha de género en GNU/Linux.

La sección Conclusiones es un resumen acerca de este enfoque y apunta a posibles trabajos futuros.

Las contribuciones de este artículo son:

1. Una solución integrada en los diferentes campos de aplicación relativos a inferir género desde el nombre.
2. Una colección de *datasets* abiertos obtenidos desde fuentes estadísticas estandarizadas en un formato único y una clasificación por idiomas.
3. Un nuevo estudio para contar hombres y mujeres en GNU/Linux.
4. Un enfoque basado en resultados reproducibles.

Muchos artículos relacionados muestran técnicas de aprendizaje automático aplicado, pero usando nuestro *dataset* libre mejoraría seguramente la precisión de los resultados contrastada con diversas fuentes abiertas como Wikipedia, Proyecto Gutenberg, Amazon, Forbes, instituciones deportivas, ...

También recogidas en DameGender y, mejorando el Estado del Arte en lo que respecta a *datasets* de libre copia en herramientas de detección de género a partir del nombre de manera significativa al cubrir más de 20 países (ver figura 1).

Figura 1: Países (En Azul Verdoso) Dónde sí hay Instituciones Estadísticas Proporcionando Datasets con Licencia de, al Menos, Libre Copia Acerca de Nombres, Género y Frecuencia a Damegender
Fuente: Elaboración Propia, 2023

Marco Teórico

Filosofías Acerca de Software, Mercado, Libertad y Género

En el contaje de hombres y mujeres en Internet, el suelo de este mundo es el software. Al igual que en las políticas públicas hay diferentes ideologías, al desarrollar software también las hay. Así en esta sección pretendemos dotar al lector de vocabulario y filosofía para reconocer las diferentes ideologías en la industria del software.

El software privativo, a veces traducido del inglés (*propietary software*) es todavía hoy, la idea más común de cómo encontramos el software distribuido en sistemas operativos, así Microsoft Windows, o MacOS se licencian con notas de copyright similares a:

Copyright (C) 2020 David Arroyo Menéndez

Author: David Arroyo Menéndez <davidam@gmail.com>
Maintainer: David Arroyo Menéndez <davidam@gmail.com>

 # All rights reserved

Esta idea está asociada a grandes compañías liderando mercado, pero cualquier persona puede usar esta filosofía.

La crítica al software privativo aparece con Richard Stallman acerca de la privacidad y la falta de libertad a las personas académicas, o hackers (personas quienes escriben y/o modifican software de terceras personas para adaptarlo a sus propias necesidades u objetivos globales)

Richard Stallman define la Libertad de Software con cuatro libertades:

- (0) Libertad para ejecutar el programa
- (1) Libertad para estudiar y cambiar el programa
- (2) Libertad para redistribuir copias exactas
- (3) Libertad para redistribuir las copias modificadas

Esta idea construye software como un bien social y queda motivado por valores éticos. La solución más recomendada por GNU es aplicar la licencia GPL a los programas software. A parte de incluir la licencia completa en el programa, seguramente cada fichero contenga una nota de copyright con un texto muy similar a:

;; This software is free software: you can redistribute it and/or modify.
;; it under the terms of the GNU General Public License as published by
;; the Free Software Foundation, either version 3 of the License, or
;; (at your option) any later version.

;; This software is distributed in the hope that it will be useful,
;; but WITHOUT ANY WARRANTY; without even the implied warranty of
;; MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
;; GNU General Public License for more details.

;; You should have received a copy of the GNU General Public License
;; along with GNU Emacs. If not, see <https://www.gnu.org/licenses/>.

El movimiento del Software Libre es anterior al movimiento *Open Source* que también cree en licencias libres como la GPL, pero dónde se piensa el desarrollo software como un campo de negocio y no de lucha de derechos civiles redefiniendo Software Libre por Código Abierto (en inglés *Open Source*) y creando una definición nueva que evita el lenguaje de valores y libertades de la otra definición.

1.Redistribución Libre

La licencia no impedirá que ninguna parte venda o regale el software como componente de una distribución de software agregada que contenga programas de varias fuentes diferentes. La licencia no exigirá regalías u otros derechos por dicha venta.

2. Código Fuente

El programa debe incluir código fuente y debe permitir la distribución en código fuente así como en forma compilada. Cuando algún tipo de producto no se distribuye con el código fuente, debe haber un medio bien publicitado para obtener el código fuente por no más que un costo de reproducción razonable, preferiblemente descargándolo a través de Internet sin cargo. El código fuente debe ser la forma preferida en la que un programador modificaría el programa. No se permite el código fuente deliberadamente ofuscado. No se permiten formas intermedias como la salida de un preprocesador o un traductor.

3. Trabajos Derivados

La licencia debe permitir modificaciones y trabajos derivados y debe permitir su distribución en los mismos términos que la licencia del software original.

4. Integridad del Código Fuente del Autor

La licencia puede restringir la distribución del código fuente en forma modificada sólo si permite la distribución de "archivos de parche" con el código fuente con el fin de modificar el programa en el momento de la compilación. La licencia debe permitir explícitamente la distribución de software creado a partir de código fuente modificado. La licencia puede requerir que los trabajos derivados lleven un nombre o número de versión diferente al del software original.

5. No Discriminación Contra Personas o Grupos

La licencia no debe discriminar a ninguna persona o grupo de Personas.

6. No Discriminación contra Campos de Actividad

La licencia no debe impedir que nadie haga uso del programa en un campo de actividad específico. Por ejemplo, no puede restringir el uso del programa en una empresa o para investigación genética.

7. Distribución de Licencia

Los derechos adjuntos al programa deben aplicarse a todos aquellos a quienes se redistribuya el programa sin necesidad de que esas partes ejecuten una licencia adicional.

8. La Licencia no debe ser Específica de un Producto

Los derechos adjuntos al programa no deben depender de que el programa forme parte de una distribución de software en particular. Si el programa se extrae de esa distribución y se utiliza o distribuye dentro de los términos de la licencia del programa, todas las partes a quienes se redistribuye el programa deben tener los mismos derechos que los que se otorgan junto con la distribución del software original.

9.La Licencia no debe Restringir otros Programas

La licencia no debe imponer restricciones a otro software que se distribuya junto con el software bajo licencia. Por ejemplo, la licencia no debe insistir en que todos los demás programas distribuidos en el mismo medio deben ser software de código abierto.

10. La licencia debe ser Tecnológicamente Neutral

Ninguna disposición de la licencia puede basarse en ninguna tecnología o estilo de interfaz individual.

En el número seis, se encuentra un punto de conflicto con las teorías feministas dónde la discriminación positiva es obviamente una buena idea para lograr la igualdad de género. Y, aunque, la Definición del Software Libre es acerca de valores y libertades en un camino práctico tiene el mismo problema debido a resultados similares en relación al contaje de personas que desarrollan software que son hombres o que son mujeres (ver figura 4).

Tanto el software libre como el privativo, estarían dentro de modelos capitalistas de producción de software, en cuanto que son acumulativos, generan desigualdades entre miembros de los proyectos. ¿Existen otros modelos teóricos o prácticos a la idea capitalista en el software? En sistemas operativos, no hay apenas software socialista o de Estado. En materia de buscadores el Estado Chino, desarrolló Baidu protegiendo este software con respecto su competidor Google de manera que se impedía utilizar Google a las personas en suelo chino, generando una controversia con respecto a la libertad de expresión, en cualquier caso sería un ejemplo de un software de interés general realizado por un Estado, a pesar de que más tarde se convierta en empresa o no. El sindicalismo revolucionario

desarrollaría software a través de la normalidad de la vida política de un sindicato, esto es, articulándolo desde las secretarías como una tarea técnica más para el disfrute de los miembros del sindicato y, no para la totalidad de las personas como normalmente se hace cuando se utiliza una licencia libre en un software, debido al carácter de clase obrera que tiene un sindicato en oposición a la clase burguesa.

Con esto se han planteado los modelos teóricos, con DameGender se reduce precios a lo que antes se hacía con software de pago, abaratando estudios relativos a un objetivo de Naciones Unidas, por lo que yo lo considero ético más allá del debate de la conveniencia del socialismo o el capitalismo.

Multiculturalismo, Interculturalismo

El *dataset* internacional de DameGender, se denomina *dataset* INTER, por razones de favorecer la interculturalidad ¿pero qué entendemos por interculturalidad? ¿por qué no decimos multiculturalidad?

Acercándonos a la literalidad de la Real Academia de la Lengua Española (RAE) por interculturalidad encontramos "Que concierne a la relación entre culturas" y, por multicultural "Caracterizado por la convivencia entre varias culturas". El contexto de detectar el género de nombres y apellidos, se refiere menos a convivencia y más entender que cada palabra de un nombre personal se relaciona con diferentes culturas, o al menos idiomas.

La idea de utilizar la palabra INTER en el sentido de interculturalidad, proviene de la educación (Aguado, Gil, Mata 2008) generando una guía para aplicar en las escuelas para aplicar esta perspectiva cómo modo de hacer una gestión enriquecedora de la diversidad para toda persona en el aula.

Género, Sexo

DameGender, utiliza la palabra género en su denominación, por una inercia social, debido a que todas las herramientas software tendían a utilizar esa palabra para referirse a la diferenciación de masculino, femenino, otros, … Pero reflexionemos brevemente sobre el significado y las implicaciones que tiene.

De nuevo, utilizamos la literalidad de la RAE, género tiene varias acepciones pero en nuestro contexto significa "Grupo al que pertenecen los seres humanos de cada sexo, entendido este desde un punto de vista sociocultural en lugar de exclusivamente biológico" y, por sexo, ocurre algo parecido "Condición orgánica, masculina o femenina, de los animales y las plantas.", "Órganos sexuales.".

En efecto, es la conveniencia de la conservadora industria del software la que nos hace pensar que género es mejor, aunque la palabra sexo es más precisa. Con DameGender, mi experiencia personal fue la siguiente: comencé utilizando la palabra sexo y sentía que las personas se enfadaban menos si utilizaba la palabra género. Esto es, razonable, debido a que

las acepciones de sexo, al referirse también al acto sexual y no solo a lo masculino y lo femenino, hacen que no metan al software en listados de pornografía, o spam. Son razones de posicionamiento en Internet la que llevan a las personas que desarrollan herramientas de detección de género desde el nombre, a utilizar género, en vez de sexo, esto es, se utiliza género como un eufemismo de sexo, para evitar posibles discriminaciones injustas. Así, en este artículo se puede considerar sinónimo género y sexo, aunque normalmente tienen significados distintos.

Un tema de candente actualidad en el feminismo de datos es la recogida del género no binario de las personas (D'ignazio y Klein 2020). Aparte del gran debate con respecto a la nomenclatura a utilizar para esta cuestión: género trans, transexual, no binario, LGTB, … en lo técnico los *datasets* Open Data no están dando a disposición del público estas opciones si es que las han recogido. No obstante, al analizar los porcentajes de personas masculinas y femeninas en cada nombre, sí es posible otorgar listados de nombres que no estén muy marcados a femenino o masculino, para las personas que por decisión de adoptar en sus vidas un género trans, puedan adoptar también nombres en rangos del 40% al 60% de masculinidad o feminidad, o similar.

Feminismo, Igualdad de Género, Ecofeminismo

Al llegar a la RAE encontramos que feminismo es el "Principio de igualdad de derechos de la mujer y el hombre", esto es, el objetivo 5 de Naciones Unidas. Así, igualdad de género es completamente sinónimo a feminismo, significa lo mismo. Quizás en el día de las conversaciones que tienen las personas al ver las noticias de la tele, se puede encontrar el feminismo más cercano al sujeto mujer que sufre injusticia y se la defiende, e igualdad de género más cercano a las macrocifras, el feminismo de datos (D'Ignazio Klein 2020) que nos dan la comprensión global de ese principio de igualdad. Para este artículo los consideraremos sinónimos.

Interseccionalidad

"La interseccionalidad se ha convertido en la expresión utilizada para designar la perspectiva teórica y metodológica que busca dar cuenta de la percepción cruzada o imbricada de las relaciones de poder" (Vigoya 2016). Este concepto nos lo vamos a encontrar muchas veces en el mundo del feminismo, debido a que se puede mejorar el feminismo con luchas antirracistas, de lucha de clases, ecologistas … Lo que viene a expresar es la interacción de esas luchas y la búsqueda de puntos de equilibrio para encontrar el bienestar de las personas.

Estado del Arte

Acerca de la Brecha de Género

Reducir la brecha de género se refiere a hacer actuar el principio de igualdad entre hombres y mujeres, primero midiendo con indicadores de desigualdad, después generando políticas de igualdad entre género y aplicándolas. Entregando, desde estadísticas oficiales de diferentes estados, datos acerca del género, la frecuencia y los nombres de personas, se facilita todo ese proceso de: medición – política – aplicación – revisión de medición.

Medir brecha de género requiere un conjunto de indicadores. Global Gender Gap Report (Chancel Piketty Saez y Zucman 2022) ha propuesto indicadores en economía, salud, educación y política. El sitio de Naciones Unidas Mujeres está mostrando indicadores para medir disparidades como leyes, educación, mortalidad maternal, participación política, pobreza, trabajo doméstico, paridad de género en el trabajo, acceso a la economía, acceso a estudios y/o trabajo, violencia contra las mujeres, justicia climática, acceso a justicia, salud, etc.

Es posible crear decisiones de impacto en una cuestión a través de resultados de investigación que toman estos indicadores en consideración. Por ejemplo, (Miyake et al 2010) concluyeron que creando afirmaciones acerca de valores éticos la brecha de género se reducía en los colegios.

Relativo a medir brecha de género en investigación social, (Bimber 2000) presentaba dos factores que afectan a la brecha de género en Internet (acceso y uso), razones de socioeconomía y género en una encuesta que recoge datos a través de varios años…

Contando Hombres y Mujeres en Internet. ¿Por qué? ¿Dónde?

El software DameGender permite recuperar datos desde fuentes de datos diversas acerca de género y nombres: GitHub, Wikipedia, APIs, periódicos, páginas webs en general, listas de correo, ...

Por ejemplo, una persona de ciencias sociales estudiando brecha de género en periodismo puede contar hombres y mujeres en Twitter (Carrera Sainz–De–Baranda Herrero Limón 2012). Mientras tanto, algún artículo de ciencias de la computación puede estar mejorando la detección de género en imágenes de caras, en *nicknames*, o simplemente en nombres.

En un artículo se presentaron varias configuraciones de un clasificador independiente de lenguaje para predecir el género de usuarios Twitter (Burger et al. 2011) . El *dataset* usado para la construcción y evaluación de estos clasificadores se crea con usuarios Twitter que llegaban a tener blog en su perfil.

Este tipo de estudios permite analizar demografía de población Twitter (Mislove et al 2011), siendo del sexo una variable clave en cualquier estudio de demografía. En este caso,

el *dataset* escogido fue el del censo de Estados Unidos que funciona razonablemente bien para nombres occidentales.

En un estudio se analizaba la brecha de género en Wikipedia mostrando evidencia de pequeñas formas de desigualdad explicando cómo resolver estas evidencias (Wagner et al 2015). Para medir la desigualdad de género se han desarrollado las siguientes bias: visibilidad, léxico (ej.: palabras discriminatorias para mujeres), estructural y de cobertura… Un buen número de personas billonarias puntúan en Forbes como de Ciencias de la Computación y los repositorios públicos nos permiten entender tendencias de inclusión o no de mujeres, así como listados públicos de puestos de trabajo, directivos

Un acercamiento a entender estas cifras es utilizar la metodología de encuesta, así en 2014 se publicó un estudio de una encuesta de 2000 contribuidores en el que la participación femenina estaría entorno al 5% (Robles, Arjona, Serebrenik, Vasilescu y González Barahona 2014)

Más recientemente, se publicó un estudio longitudinal acerca de hombres y mujeres (Zacchiroli 2020) , que revelaba que las mujeres podrían ser el 8% en contribuciones de commits, esto es, código en repositorios públicos. Este estudio utilizó genderguesser para inferir género desde el nombre que a su vez usa namdict como dataset. Un grupo de investigadores exploraron StackOverflow un sitio muy popular de preguntas y respuestas concluyendo que los hombres representan la vasta mayoría de contribuciones (Vasilescu, Capiluppi y Serebrenik 2012)

Relativo a la brecha de género en ciencia, otro grupo de investigadores presentó un análisis bibliométrico confirmando que los desequilibrios de género persisten en investigación a través del mundo entero (Lariviere et al. 2013). Mientras que, otro grupo presentaron un código en R usando la API de genderize y proporcionando un buen enfoque acerca de cómo calcular brecha de género (Holman, Stuart–Fox y Hauser 2018)

Enfoques Automáticos para Inferir Género

Hay varios caminos para inferir género desde fuentes de Internet: texto escrito a mano, imágenes, expresiones en documentos y nombres de los autores. Así, se presentó un método de inferir género desde textos escritos a mano con un 67.5 % (Liwicki, Schlapbach y Bunke. 2011)

Con respecto a clasificación de género con imágenes, la aplicación suele venir desde varias fuentes, por ejemplo, un trabajo que detecta género a partir de una combinación de imágenes, edad e información cultural provista por primeros nombres con ejemplos no etiquetados y con una precisión del 60 % (Gallagher y Chen 2008)

También es posible inferir género a través de expresiones escritas en documentos, un estudio explica que las mujeres usan muchos más pronombres que los hombres, mientras que los hombres usan más modificadores de nombres (determinantes, o adjetivos) (Argamon et al 2003). Un grupo de investigadores presentaron un sistema de clasificación

de documentos por género con una precisión del 80 % aproximadamente (Koppel, Argamon y Shimoni 2002). Otros, mostraron una selección de características (features) y un modelo construido usando aprendizaje automático generando un acierto del 85.1 % para identificar género desde el texto (Cheng, Chandramouli y Subbalakshmi. 2011).

Infiriendo Género desde Nombres

Las herramientas usadas para inferir género desde un nombre están normalmente basadas en *datasets* que, como mínimo, incluyen género. Un estudio presentaba un método para inferir género desde los primeros nombres en Twitter (Liu, Ruths 2013). El *dataset* estaba codificado a mano por acuerdo entre tres trabajadores de Amazon con 50.000 usuarios Twitter seleccionados aleatoriamente con solo 12681 etiquetas de género. El objetivo de este estudio era determinar el valor incremental de usar el nombre de usuario (*username*) como una característica (*feature*) en la inferencia de género basado en tweets.

Otro estudio presentaba un trabajo de cómo inferir género en Twitter (Mueller y Stumme 2016). Ellos usaban el dataset namdict y el censo de los Estados Unidos como datasets. Para construir el modelo de aprendizaje automático, se escogieron características lingüísticas como "número de consonantes'', "número de vocales'', "número de sílabas'', ... El modelo de clasificación era creado usando SVM.

Ideas Relacionadas

Un trabajo de interés presentaba un sistema de software libre para clasificar nombre y etnicidad usando aprendizaje automático utilizando una lista de nombres desde Wikipedia (Ambekar et al. 2009). Un trabajo más reciente, presentaba NamPrism dando ideas nuevas acercar de clasificar razas y, siendo aplicado a repositorios de software masivos (Nadri, Rodríguez Pérez y Nagappan 2021)

Con respecto a apodos o nicknames, se presentó un enfoque que usaba basado en patrones léxicos para extraer alias de un nombre dado, con un conjunto de nombres y sus alias como datos de entrenamiento. Los aliases candidatos entran en un ranking con varias puntuaciones. Support Vector Machine (SVM) fue usado para construir la función ranking (Bollegala y otros 2010)

Estándares Relacionados

ISO/IEC 5218 propone la siguiente norma acerca de la codificación de género: "0 como conocido'', "1 como masculino'', "2 como femenino'' y "9 como no aplicable''.

El RFC 6350 (vCard) tiene estas categorías "m como masculino'', "f como femenino'', "0 como otro'', "n como no aplicable'' y "u como indefinido (*undefined*)''. Basado en este estándar, con respecto a publicación web se puede usar el microformato h_card en el contexto de escribir interfaces de formularios web considerando lecturas de w3.

En DameGender se generó una codificación inicial de género: “0 como femenino”, “1 como masculino”, “2 como desconocido. Y en la actualidad se adapta a los otros dos estándares de esta sección

A Modo de Resumen

El primer nombre de un asunto es el factor clave usado para determinar género en el Estado del Arte de las herramientas de inferencia de género. Sin embargo, en muchos contextos hay más características: apellidos, texto, imágenes, *nicknames*, etc. El primer nombre puede ser útil para inferir otras cuestiones tal y como raza, etnicidad, o cultura, también, dónde es muy normal apoyarse más en apellidos.

El aprendizaje automático, dónde hay una selección de características previas a las que aplicar ese aprendizaje, está siendo usado en muchos trabajos, aunque hay una discusión acerca de cuál es el mejor enfoque.

Diseño

Verdad y Falsedad en Nombres, Género y Frecuencia

Usar nombre, género y frecuencia es una idea comúnmente admitida en vez de otros enfoques como nombre y género, o nombre y varios grados de masculinidad y feminidad. Esto es así, porque hay bastante industria e instituciones que utilizan esta manera de hacer las cosas. Así, las frecuencias nos permiten calcular porcentajes de masculinidad y, feminidad en cada nombre.

Con respecto a veracidad en las datos DameGender permite elegir a sus usuarios confiar en empresas o instituciones para inferir el género de cualquier nombre.

Y, se aportan *datasets* públicos construidos desde instituciones estadísticas que lo permiten. Este trabajo de observación de *datasets* de instituciones oficiales tiene el potencial de favorecer la participación ciudadana ante cambios desde los gobiernos…

Género, Lenguaje, Nación y Diversidad

Hay reglas y excepciones en diferentes lenguajes para predecir si un nombre es masculino o femenino. Por ejemplo, en español o inglés, hay más nombres que finalizan con “a'' clasificados como femeninos que, como masculinos. Sin embargo, Andrea es femenino en España y, masculino en Italia. Así, es útil comprender el lenguaje y cultura asociado con un nombre. El lenguaje nos acerca a naciones, pero con diferencias, por ejemplo, en España hay varios idiomas: vasco, catalán, castellano (español) y gallego. Español es la lengua oficial más utilizada en España, pero también en otros países como Argentina, Méjico, Ecuador, Bolivia, ...

En DameGender se han juntado los *datasets* oficiales (los entregados por instituciones estadísticas) para generar *datasets* de lenguajes. En lenguajes donde no había instituciones

estadísticas hemos valorado utilizar nombres descargados de Wikipedia descubriendo una baja cantidad de nombres una no muy buena calidad.

Obteniendo datos y datasets para DameGender

DameGender unificó los diferentes formatos descargados desde oficinas estadísticas oficiales para nombre, género y frecuencia en los siguientes países: Argentina, Austria, Australia, Bélgica, Canadá, Suiza, Alemania, Dinamarca, España, Finlandia, Francia, Gran Bretaña, Irlanda, Islandia, Noruega, Nueva Zelanda, Méjico, Portugal, Rusia, Eslovenia, Suecia, Estados Unidos de América y Uruguay.

Se han desarrollado diferentes scripts para la descarga de *datasets*:

- donwloadcsv.py: para la descarga de ficheros CSV desde las APIs
- donwloadjson.py: para la descarga de ficheros JSON desde las APIs
- get–wikidata–names.py: para la descarga de nombres desde Wikipedia
- get–wikidata–surnames.py: para la descarga de apellidos desde Wikipedia
- orig2.py: para la descarga de *datasets* desde instituciones estadísticas oficiales

En la figura 2, se puede el flujo de datos e interacciones:

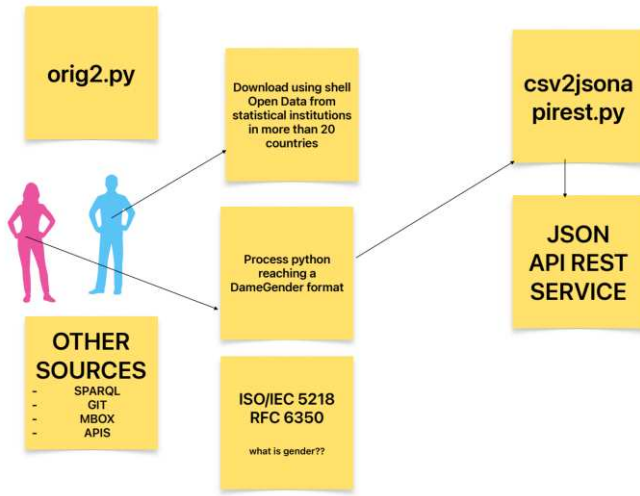

Figura 2: Flujo de Datos y las Herramientas DameGender para Recuperar y Procesar estos Datos
Fuente: Elaboración Propia, 2023

Una posible mejora a este diseño sería unificar todas estas funcionalidades en un solo comando, por ejemplo, orig2.py

Para generar las frecuencias de nombres ha sido aplicado el criterio de una persona usando un nombre es un voto que, normalmente puntúa en masculino o femenino, pero que podría puntuar en clasificaciones no binarias si las instituciones estadísticas las quieren utilizar. Así, cada país está entregando las frecuencias en nacimientos o totales construyendo un *dataset* internacional como una suma de frecuencias entregadas por países.

También se ha construido un *dataset* internacional de apellidos, aunque hay menos instituciones estadísticas entregando *datasets* de apellidos (España, Rusia, Estados Unidos de América y Argentina) en este momento. La fuente de datos fueron las instituciones estadísticas oficiales que daban esta información con licencias de libre copia. Aunque Wikipedia podría haber sido una fuente interesante para nombres o apellidos se comprobó que había pocos nombres o apellidos por nación y lenguaje.

Una posible crítica al enfoque de nombre, género y frecuencia es el llamado problema Leslie (Blevins y Mullen 2015), un problema que se genera cuando en un mismo país hay un nombre que cambia de género a partir de una fecha. Muchas instituciones estadísticas no dan los totales, sino los nacimientos de cada año para tener en cuenta ese problema. El *dataset* construido está diseñado para la mayoría de los casos de uso y permite incluir *datasets* que no dan el desglose de frecuencias por año, pero se guardan los datos de quienes sí dan esa información.

Se ha medido la precisión del *dataset* internacional de DameGender con diferentes datasets de prueba obteniendo resultados similares y mejores (ver tabla 1). Estos datasets fueron proporcionados por científicos (Sebo, Santamaría y Mihaljevic), o proporcionados por ayuntamientos (Conseil Garonne), u hospitales (Baby names New York), o instituciones deportivas (FIFA, WTA) o Wikipedia.

Tabla 1: Nivel de Acierto del Dataset Internacional de Damegender con Respecto a otros Datasets con Licencias Libres

Datasets	*Nivel de acierto (accuracy)*
Wikipedia Scientists	*0.93*
FIFA soccer	*0.93*
WTA tenis	*0.91*
National League	*0.91*
Baby names New York	*0.98*
Conseil Garonne	*0.97*
Paul Sebo dataset	*0.88*
L. Santamaría & H. Mihaljevic dataset	*0.92*

Fuente: Portales Open Data y Fuentes Directas (Referencias desde Artículos, Sitios Web de Ayuntamientos, ...), 2023

Midiendo Brecha de Género

GNU/Linux como Caso de Uso

Con un *dataset* confiable de nombres, género y frecuencia es fácil medir la brecha de género. Estudiantes y académicos podrían medir la brecha de género a bajo coste y encontrar el quinto objetivo de desarrollo sostenible que es la eliminación de la brecha de género.

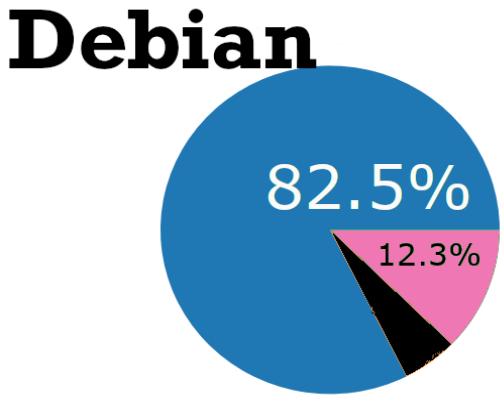

Figura 3: Debian Developers by Gender
Fuente: Elaboración Propia, 2023

Esta sección está dividida en contar hombres y mujeres usando Debian, GNU y Linux. Se han obtenido los ficheros csv usando diferentes métodos para determinar los nombres de personas en estas comunidades.

En la comunidad Debian todos los miembros deben colaborar con una clave gpg, así se pueden contar hombres y mujeres en el anillo de claves. El anillo de claves (gpg keyring) se importó con comandos gpg.

Tanto GNU como Linux, tienen *websites* colaborativos para estos proyectos. De este modo, se ha hecho *scraping* para descargar las personas implicadas y, que queden volcadas en un fichero csv. Después, con el comando de DameGender csvgender.py se infieren las personas que son hombres y las que son mujeres.

El *dataset* INTER fue creado sumando las frecuencias de los diferentes *datasets* de licencia libre otorgados por las instituciones estadísticas de diferentes naciones. El resultado es uno de los mejores *datasets* de nombres y frecuencias del mundo especialmente preciso en el mundo occidental que, es dónde hay más publicaciones académicas, software, ...

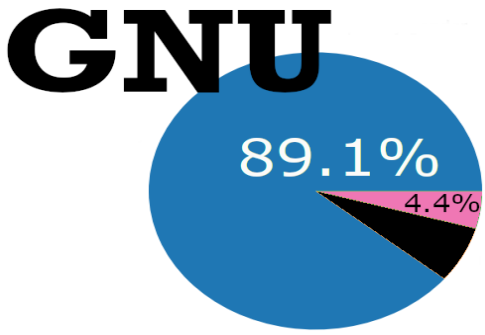

Figura 4: GNU Developers by Gender
Fuente: Elaboración Propia, 2023

En Linux (ver figura 5) podemos encontrar 1891 hombres desarrolladores (90,1 %), 78 mujeres desarrolladoras (3,7 %) y 130 desconocidos (6,2 %). El número de desconocidos es debido a diferentes razones, pero es común en Linux que, las personas hagan *commits* con la identidad de una compañía y no con una identidad personal.

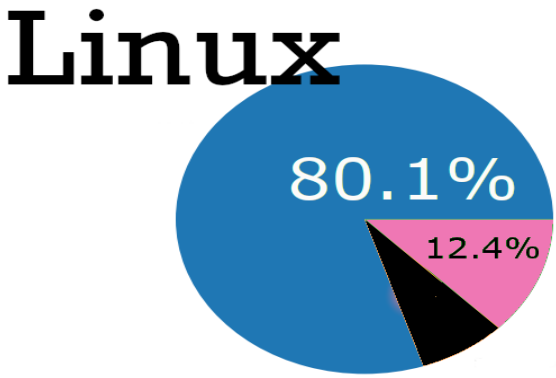

Figura 5: Linux *Developers by Gender*

Fuente: Elaboración Propia, 2023

En GNU (ver figura 4) se pueden encontrar 215 hombres desarrolladores (91,5 %), 12 mujeres desarrolladoras (5,1 %) y 8 identidades no clasificadas. Richard Stallman, el fundador de GNU y la FSF llegó a dimitir por su comportamiento con las mujeres durante unos meses.

Debian (ver figura 3) es una distribución de software libre, el proyecto que crea el CD/DVD y, el software listo para ser descargado desde Internet con las dependencias. Hay muchas distribuciones, tales como, Ubuntu o RedHat así no es representativo, pero es interesante comprender que los números son similares 863 hombres desarrolladores (85,5

%), 44 mujeres desarrolladoras (4,4 %) y 102 personas que desarrollan sin ser clasificadas (10,1 %).

En 2009 se publicó un estudio que daba un 24% de trabajos en STEM a las mujeres norteamericanas (Beede et al, 2011). Así, resulta llamativo el bajo número de mujeres en GNU/Linux, debido a que había sido considerado como una alternativa de valores en el mundo del software. Una hipótesis para investigaciones posteriores podría ser que el mundo de sistemas operativos, intérpretes, … esté más masculinizado que otras áreas informáticas más cercanas a las ciencias sociales (paquetes estadísticos) o el periodismo (programación web).

Conclusiones

El Feminismo de Datos (D'ignazio y Klein, 2020) es un área de interés creciente. Se ha explicado la aplicación DameGender, las motivaciones (investigación reproducible, eliminar la brecha de género para resolver el quinto objetivo de Naciones Unidas, campos de aplicación, incluyendo la lingüística, ciencias sociales, ingeniería de software, procesamiento de lenguaje natural y periodismo).

Una aportación es construir un *dataset* internacional, universal de *datasets* libres de nombres, género y frecuencia para el correcto diseño, atendiendo a la diversidad (LGTB, minorías culturales, etc.) Además, este artículo ha explicado las tecnologías involucradas en la reducción de costes relacionado al estudio de la brecha de género (detección de género desde los nombres, APIs, web semántica, etc.). Así, quienes investigan usando DameGender encuentran las ventajas del Software Libre para la investigación. Esto es, que se puede usar, copiar, modificar y redistribuir el software, favoreciendo la confianza de la revisión por pares.

El estado actual de este trabajo es uno de los más grandes *datasets* científico de nombres personales con más de 20 países dando estadísticas oficiales y generando una excelente representación de nombres en el mundo occidental. También hemos incluido *datasets* de apellidos y nombres personales clasificados por un gran número de idiomas y, por supuesto, países. Habiendo generado un esfuerzo software notable en recuperación y mantenimiento de nombres para el *dataset*. Estos datos pueden proporcionar una visión muy precisa del mundo real para lograr bajos números de nombres desconocidos.

Ante la pregunta de investigación relativa al ratio de hombres y mujeres en Debian GNU/Linux gracias a una herramienta de detección de género a partir del nombre se puede deducir de manera precisa que el número de mujeres estaría oscilaría del 4 al 13 por ciento, con un número de hombres que iría del 80 al 90 por ciento y el resto podrían estar en una categoría de otras opciones o sexo desconocido. Resaltar que tras las mediciones son las afirmaciones en valores en las aulas y, las políticas generales de equidad de género las que pueden cambiar estas situaciones a favor de la igualdad de género.

Agradecimientos

Muy agradecido a: Clara Sainz de Baranda y Milagros Sainz Ibáñez por las oportunidades en la diseminación. A las instituciones estadísticas por la entrega de *datasets* con licencias de dominio público o similares. A Luz Galvis por su interés y buen hacer en el proyecto DameGender. A Daniel Izquierdo y Laura Arjona por iniciar este campo de investigación en la URJC (GSYC) y a todas las personas que lo han continuado como Gema Rodríguez Pérez, Jesús González Barahona y Gregorio Robles. A las mujeres de mi familia. A Regina Quero por la asistencia en la publicación.

REFERENCIAS

Abdul Wahed Qasem Ghaleb Al–Zumor. 2009. "A socio–cultural and linguistic analysis of Yemeni Arabic personal names". GEMA: Online Journal of Language Studies. 9(2): 15–27

Aguado, Teresa, Gil Jaurena, Inés, Mata, Patricia. 2008. "El Enfoque Intercultural en la Formación del Profesorado: Dilemas y Propuestas" Revista Complutense de Educación

Al–Zumor, Abdul Wahed Qasem Ghaleb. 2009. "A socio–cultural and linguistic analysis of Yemeni Arabic personal names" GEMA: Online Journal of Language Studies

Ambekar, Anurag, Ward, Charles, Mohammed, Jahangir, Male, Swapna y Steven Skiena. 2009. "Name–ethnicity classification from open sources". Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining.

Argamon, Shlomo, Koppel, Moshe, Fine, Jonathan y Anat Rachel Shimoni. 2003. "Gender, genre, and writing style in formal written texts. Text & Talk, 23(3):321–346.

Beede, David, Tiffany, Julian, Langdon, David, McKittrick, Khan, Beethika y Mark Doms. 2011. "Women in STEM: A Gender Gap to Innovation". Economics and Statistics Administration Issue Brief.

Benard Odoyo Okal. 2018. "A linguistic overview of the patronymic and gender names amongst the selected African Communities" American Journal of Linguistics.

Bimber, Bruce. 2000. "Measuring the gender gap on the Internet". Social science quarterly.

Blevins, Cameron y Lincoln Mullen. 2015. "Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction." DHQ: Digital Humanities Quarterly

Boehm, Barry. 2002. "Software engineering economics" Springer Berlin Heidelberg.

Bollegala, Danushka, Yutaka Matsuo y Mitsuru Ishizuka. 2010. "Automatic discovery of personal name aliases from the web". IEEE Transactions on Knowledge and Data Engineering

Buchbinder, Felipe y Walter Sande. 2015. "O papel do gênero no desenvolvimento da habilidade geral no uso de computadores: um estudo com alunos universitarios". Revista Internacional de Tecnología Ciencia y Sociedad, 4 (1), 69–83.

Burger, Henderson, Kim y Zarrella. 2011. “Discriminating gender on Twitter”, Association for Computational Linguistics.

Carrera Álvarez, Pilar, Sainz De Baranda Andújar, Clara, Herrero Curiel, Eva, y Nieves Limón Serrano. “Journalism and social media: How spanish journalists are using twitter”. Estudios sobre el mensaje periodístico, 18(1):31, 2012.

Chancel, Lucas, Thomas Piketty, Emmanuel Saez y Gabriel Zucman. 2022. “World inequality report 2022”. Harvard University Press.

Cheng, Na, Rajarathnam Chandramouli y KP Subbalakshmi. 2011. “Author gender identification from text”. Elsevier. 8(1): 78–88.

D'ignazio, Catherine y Lauren F Klein. 2020. “Data Feminism”. MIT Press.

Esparza Rodríguez, Saúl Alfonso, Apolinar Martínez–Arroyo, Jaime, García Tapia, Gabino y Marco Alberto Valenzo–Jiménez. 1996. “Análisis de densidad bibliométrica en la estructura de artículos científicos: caso de estudio”. Revista Internacional de Ciencias Sociales, 8 (1).

Fraser, Peter Marshall, Matthews, Elaine, Osborne, Michael J, Byrne, Sean G, Catling, Richard WV, Balzat, J–S, Chiricat, E, Corsten, Thomas y Fabienne Marchand. “A lexicon of Greek Personal Names”. 1987. 5.

Gallagher, Andrew C y Tsuhan Chen. 2008. “Estimating age, gender, and identity using first name priors”. IEEE.

Gonzalez–Barahona, Jesus M, Robles, Gregorio, Andradas–Izquierdo, Roberto y Rishab Aiyer Ghosh. 2008. “Geographic origin of libre software developers”. Information Economics and Policy

Holman, Luke, Devi Stuart–Fox y Cindy E. Hauser 2018. “The gender gap in science: How long until women are equally represented?” PLoS biology 16(4).

Hutson, Mathew. 2016 “The Gender of Names” Scientific American Mind

Izquierdo, Daniel, Huesman, Nicole, Serebrenik, Alexander y Gregorio Robles. 2018. “Openstack gender diversity report”. IEEE Software.

John R Krueger. “Mongolian personal names”. 1962. 10(2): 81–86.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi y Markus Strohmaier. 2016. “Inferring Gender from Names on the Web: {A} Comparative Evaluation of Gender Detection Methods” ArXiv.

Kofi Agyekum . “The sociolinguistic of Akan personal names”. 2006. 15(2): 206–235.

Koppel, Moshe, Shlomo Argamon y Anat Rachel Shimoni. 2002. “Automatically categorizing written texts by author gender”. Literary and linguistic computing. 17(4): 401–412.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin y Cassidy R Sugimoto. 2013. “Bibliometrics: Global gender disparities in science”. Nature. 504(7479): 211–213.

Lariviére, Vincent, Ni, Chaoquin, Gingras, Yves, Cronin, Blaise y Cassidy R. Sugimoto. 2013. "Bibliometrics: Global gender disparities in science". Nature. 504 (7479), 211–213.

Lawson, Edwin D y Natan Nevo. 2005. "Russian Given Names: Their Pronunciation, Meaning, and Frequency". Names. 53(1–2): 49–77.

Liu, Wendy y Derek Ruths. 2013. "What's in a name? using first names as features for gender inference in twitter". AAAI Spring Symposium Series

Liwicki, Marcus, Andreas Schlapbach y Horst Bunke. 2011. "Automatic gender detection using on–line and off–line information". Springer. 14(1): 87–92.

Lockhart, Jeffrey W, King, Molly y Christin Munsch. 2023. "Name–based demographic inference and the unequal distribution of misrecognition". Nature. 1–12.

Loper, Edward y Steven Bird. "NLTK: the natural language toolkit" 2002.

Mislove, Alan, Sune Lehmann, Yong–Yeol Ahn, Jukka–Pekka Onnela y J Niels Rosenquist. 2011. "Understanding the demographics of twitter users". Fifth international AAAI conference on weblogs and social media

Miyake, Akira, Lauren E. Kost–Smith, Noah D. Finkelstein, Steven J Pollock, Geoffrey L Cohen y Tiffany A. Ito. 2010. "Reducing the gender achievement gap in college science: A classroom study of values affirmation". American Association for the Advancement of Science. 330(6008): 1234–1237.

Mueller, Juergen y Gerd Stumme. 2016. "Gender inference using statistical name characteristics in twitter" Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016.

Murray, Dakota, Siler, Kyle, Lariviére, Vincent, Chan, Wei Mun, Collings, Andrew M, Raymond, Jennifer y Cassidy R. Sugimoto. 2018. "Gender and international diversity improves equity in peer review" BioRxiv.

Nadri, Reza, Rodríguez Pérez, Gema y Nagappan, Meiyappan. 2021. "On the Relationship Between the Developer's Perceptible Race and Ethnicity and the Evaluation of Contributions in OSS". IEEE Transactions on Software Engineering

Niemi, Mari K y Ville Pitkanen. 2017. "Gendered use of experts in the media: Analysis of the gender gap in Finnish news journalism". Public Understanding of Science

Peng, Roger D. 2011. "Reproducible research in computational science" American Association for the Advancement of Science. 334(6060): 1226–1227.

Robles, Gregorio, Laura Arjona Reina, Alexander Serebrenik, Bogdan Vasilescu y Jesús M González Barahona. 2014. "FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing, and combining". ACM.

Sainz de Baranda Andújar, Clara. 2014. "El género de los protagonistas en la información deportiva (1979–2010): noticias y titulares/The gender of the main characters in sports reporting (1979–2010): News and headlines". Estudios sobre el mensaje periodístico 20(2)

Santamaria, Lucía, Helena Mihaljevic. 2018. "Comparison and benchmark of name–to–gender inference services". PeerJ Computer Science

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai–Wei Chang y William Yang Wang. 2019. "Mitigating gender bias in natural language processing: Literature review". ArXiv

Thompson, W 1833 "Electrical Units of Measurement, Popular Lectures" London: Macmillan

Van Rossum, Guido, y Fred L Drake. 2011. "The python language reference manual". Network Theory Ltd.

Vasilescu, Bogdan, Capiluppi, Andrea y Alexander Serebrenik. 2012. "Gender, representation and online participation: A quantitative study of stackoverflow" IEEE.

Vasilescu, Bogdan, Daryl Posnett, Basihakhi Ray, Mark GJ van der Brand, Alexander Serebrenik, Premhumar Devanbu y Vladimir Filkov. 2015. "Gender and tenure diversity in GitHub teams". Proceedings of the 33rd annual ACM conference on human factors in computing systems

Vigoya, Mara Viveros. 2016. "La interseccionalidad: una aproximación situada a la dominación". Debate feminista. Elsevier.

Wagner, Claudia, David Garcia, Mohsen Jadidi y Markus Strohmaier. 2015. "It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia".

Zacchiroli, Stefano. 2020. "Gender differences in public code contributions: a 50–year perspective". IEEE. 38(2):45–50.

SOBRE EL AUTOR

David Arroyo Menéndez: La investigación como estudiante de doctorado en la URJC, en el Departamento Grupo de Señales y Comunicaciones (GSYC), Universidad Rey Juan Carlos, Madrid, España; actualmente de manera independiente.

Email del autor: https://www.davidam.com