Un dataset internacional acerca de nombres, género y frequencias en DameGender

David Arroyo Menéndez, Madrid, Spain
{davidamgmail}.com

DAMEGENDER

Abstract

La igualdad de género es el quinto objetivo de desarrollo sostenible (ODS) para Naciones Unidas[1].

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.

1 Introducción

Naciones Unidas tiene como objetivo reducir la brecha de género[2], para eso el primer paso es poder medir brecha de género ("if you cannot measure it, you cannot improve it" [Tho33]) para medir brecha de género resulta económico utilizar ordenadores ("Software Engineering Economics is an invaluable guide to determining software costs, applying the fundamental concepts of microeconomics to software engineering [B$^+$81]").

Además 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.

[1] https://www.un.org/sustainabledevelopment/gender-equality/

[2] https://www.un.org/sustainabledevelopment/gender-equality/

A través del uso de los nombres personales, uno puede inferir el sexo o género de una persona en artículos académicos, libres, periódicos y muchas interacciones en Internet. Así, detectar género desde los nombres puede ser un camino estratégico para medir 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[LB02], R Gender, Gender Detector and Gender Computer[3]).

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, inferiendo 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 [KWL+16] mostraba una comparativa de datasets como un modo de mejorarla precisión, comparando herramientas que usan diferentes datasets públicos (SSA [4], IPUMS [5], namdict [6], etc).

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 [SGT+19]: ingeniería de software [VCS12], lingüística [Hut16, AZ09], ciencias sociales [HSFH18, MLA+11, NP17, dBA14]), ... Con vocación de ser puntuados muchos de ellos como de ciencia reproducible[Pen11] si se tienen las precacuciones correctas de la definición.

La estructura de este artículo es la siguiente: La sección 2 presenta la investigación principal midiendo brecha de género y herramientas de detección de género desde el nombre. La sección **??** da el vocabulario y la filosofía acerca de cómo elegir fuentes y encarar las problemáticas de diversidad construyendo un dataset La sección **??** explica una aplicación para este dataset: medir brecha de género en GNU/Linux La sección **??** 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. 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 Open Data en herramientas de detección de género a partir del nombre de manera significativa al cubrir más de 20 países (figura 1)

2 Estado del Arte

2.1 Acerca de la Brecha de Género

Reducir la brecha de género se refiere a igualdad entre hombres y mujeres y políticas de no discriminación. El género se refiere al sexo de la persona determinado en su nacimiento, aunque éste pueda cambiarse en algún momento de su vida. Las discusiones acerca de las definiciones de género se refieren a estos problemas. Sin embargo, hay un consenso determinando de género, frecuencia y nombres con estadísticas oficiales entregadas por las instituciones de los estados.

Medir brecha de género requiere un conjunto de indicadores. Global Gender Gap Report [CPSZ22] han sido propuesto en economía, salud, educación y política. El sitio de Naciones Unidas[7] está mostrando indicadores para medir disparidades como leyes, educación, mortalidad maternal, participación política, pobreza, trabajo

[3] https://github.com/tue-mdse/genderComputer
[4] https://www.ssa.gov/oact/babynames/limits.html
[5] https://usa.ipums.org/usa-action/variables/NAMEFRST
[6] https://raw.githubusercontent.com/lead-ratings/gender-guesser/master/gender_guesser/data/nam_dict.txt
[7] https://www.unwomen.org/

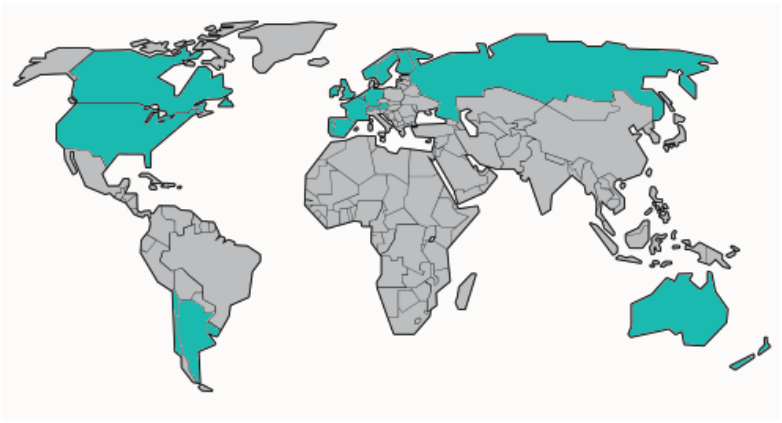

Figure 1: Green (Countries with Open Data provided by Statistical Institutions), White (Countries without Open Data or don't reproduced by the author). Thanks to wikimedia by the mapamundi base source `https://commons.wikimedia.org/wiki/File:Simple_world_map_edit.svg`

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 y otros autores [MKSF+10] 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 [Bim00] 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.

2.2 Contando hombres y mujeres en Internet. ¿Por qué? ¿Dónde?

El software DameGender focaliza en recuperar datos desde fuentes de datos diversas acerca de género y nombres. Lugares como GitHub, Wikipedia, APIs, periódicos, websites en general, listas de correo, ... Veamos cómo se estaban haciendo estos trabajos.

Por ejemplo, un científico social estudiando brecha de género

Acknowledgments

We would like to thank: the statistical institutions by release of the open datasets about names, gender and frequency. Luz Galvis for the software contributions, Daniel Izquierdo and Laura Arjona for starting this research field at URJC all those working with Jesús González Barahona and Gregorio Robles.

References

[AZ09] Abdul Wahed Qasem Ghaleb Al-Zumor. A socio-cultural and linguistic analysis of yemeni arabic personal names. *GEMA: Online Journal of Language Studies*, 9(2):15–27, 2009.

[B+81] Boehm Barry et al. Software engineering economics. *New York*, 197, 1981.

[Bim00] Bruce Bimber. Measuring the gender gap on the internet. *Social science quarterly*, pages 868–876, 2000.

[CPSZ22] Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World inequality report 2022*. Harvard University Press, 2022.

[dBA14] Clara Sainz de Baranda Andújar. 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):1225, 2014.

[HSFH18] Luke Holman, Devi Stuart-Fox, and Cindy E Hauser. The gender gap in science: How long until women are equally represented? *PLoS biology*, 16(4):e2004956, 2018.

[Hut16] ¿ Matthew Hutson. The gender of names. *Scientific American Mind*, 27(4):14–14, 2016.

[KWL+16] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54, 2016.

[LB02] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[MKSF+10] Akira Miyake, Lauren E Kost-Smith, Noah D Finkelstein, Steven J Pollock, Geoffrey L Cohen, and Tiffany A Ito. Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008):1234–1237, 2010.

[MLA+11] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[NP17] Mari K Niemi and Ville Pitkänen. Gendered use of experts in the media: Analysis of the gender gap in finnish news journalism. *Public Understanding of Science*, 26(3):355–368, 2017.

[Pen11] Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[SGT+19] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[Tho33] W Thompson. Electrical units of measurement, popular lectures, 1833.

[VCS12] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*, pages 332–338. IEEE, 2012.