Damegender: Hacia un Conjunto de Datos Internacional acerca de Nombres, Género y Frecuencia

David Arroyo Menéndez

October 25, 2021

Presentación

- Estudiante de Tesis: David Arroyo Menéndez
- Título de estas presentaciones: Hacia un Conjunto de Datos Internacional acerca de Nombres, Género y Frecuencia
- Director de Tesis: Jesús González Barahona

Motivación Ética

- La Igualdad de Género es el Quinto Objetivo en Naciones Unidas
- Solo cuando se mide un proceso, se puede mejorar
- Reduciendo los costes en el proceso de una manera de Software Libre, más personas académicas pueden contribuir a mejorarlo

Contribuciones al Estado del Arte:

- Una solución integrada dónde hacer experimentos en las diferentes aplicaciones relativa a inferir género desde el nombre
- Una colección de Conjuntos de Datos Abiertos utilizando fuentes de estadísticas oficiales y estandarizando en un único formato de género, nombre y frecuencia.
- Un nuevo estudio aplicando DameGender para contar hombres y mujeres en Linux.
- Un enfoque de Machine Learning clasificando género desde el nombre.
- Un enfoque basado en reproducibilidad.

Disciplinas Académicas que podrían estar beneficiándose en estos trabajos

- Ciencias Sociales: Fuentes Secundarias en Estudios de Género, e Indicadores acerca de Brecha de Género
- Ciencias de Computación: Ingeniería de Software, Procesamiento de Lenguaje Natural
- Lingüistica: Estudios acerca de nombres

Trabajos de Investigación usando estos Conjuntos de Datos Abiertos

- Brecha de Género en Conocimiento: Wikipedia, Twitter, Newspapers, Journal Papers, ...
- Contar hombres y mujeres en Ingeniería de Software: Git, StackOverflow, ...
- Lingüística: Estadísticas en cada lenguaje acerca de género en primeras y últimas letras, morfemas, fonemas, ...

Damegender: Un Kit de Herramientas midiendo Brecha de Género con un Enfoque en Reproducibilidad

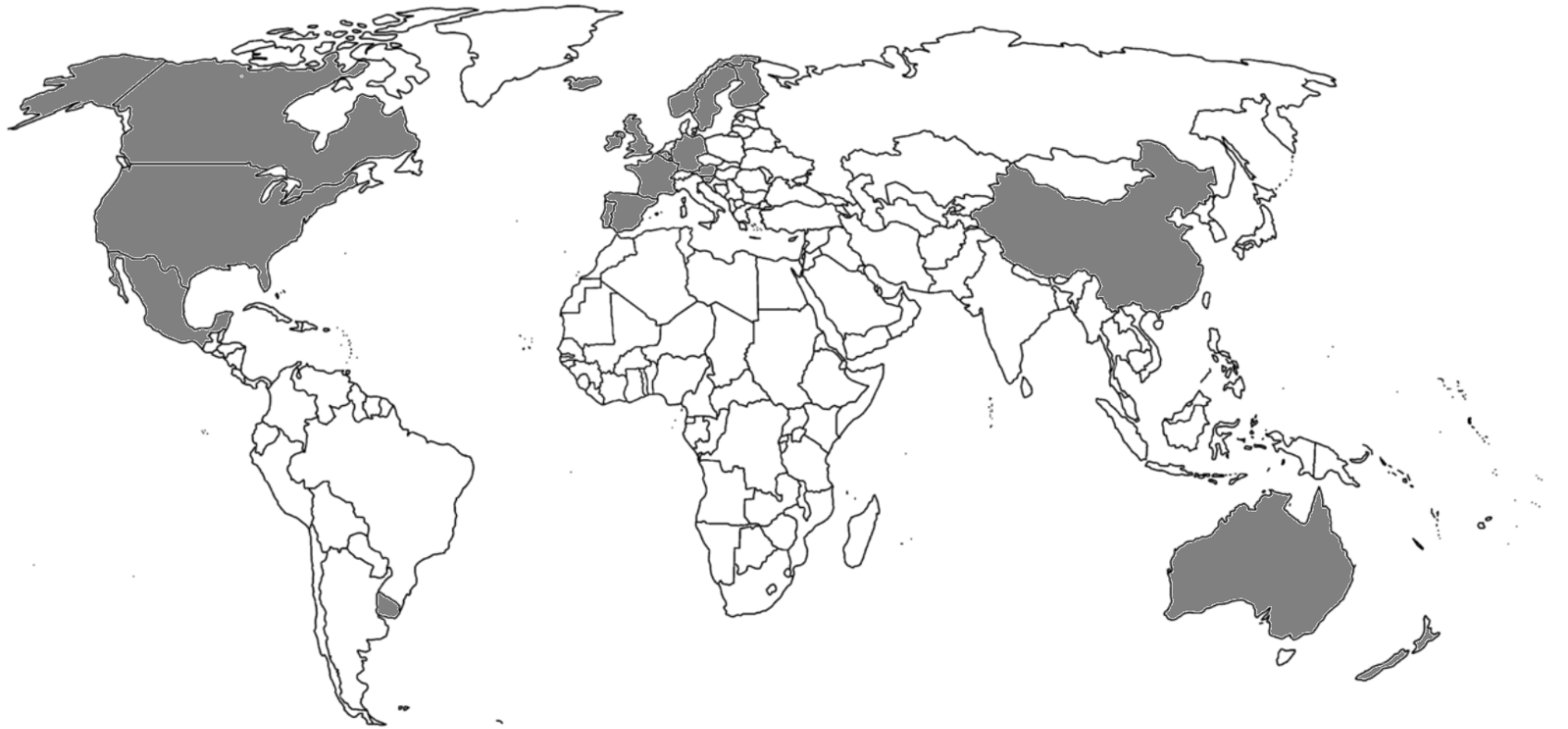
Estado Actual:

- Interfaz con APIs comerciales
- Automáticamente generar miles de conteos en ficheros CSV, Git, Listas de Correo, ...
- Permitir usar Aprendizaje Automático en apodos, diminutivos, nuevos nombres, ...
- Comparar usando herramientas estadísticas: propiedades en nombres, precisión, matrices de confusión, análisis de componentes, roc, ...

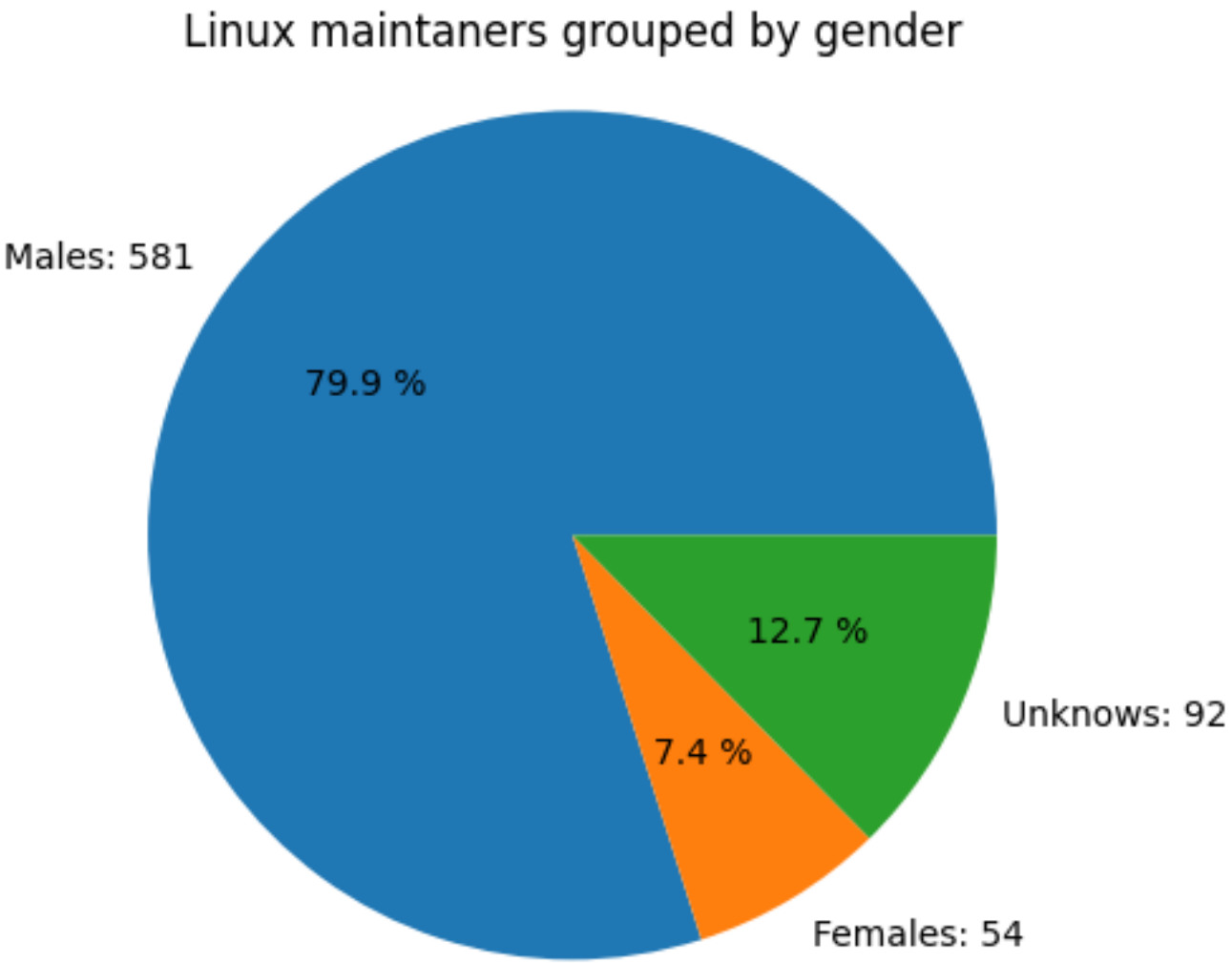
Damegender: Colección de Datos Abiertos

- Accuracy: 87.6%
- Precisión: 96.4%
- Número de nombres de mujer: 299870
- Número de nombres de hombre: 278981
- Datos Abiertos coleccionados desde instituciones estadísticas
- Más de 20 países

Damegender: Países de nuestra colección de datos



Aplicando Damegender para contar hombres y mujeres en Linux (I)



Aplicando Damegender para contar hombres y mujeres en Linux (II)

python3 csv2gender.py files/linux-maintainers.csv –first$_{\text{nameposition}}$=0
–title="Linux maintaners grouped by gender" –dataset="inter"
–outcsv="files/linux-maintainers.gender.csv"
–outimg="files/linux-maintainers.gender.png" –noshow –delete$_{\text{duplicated}}$

Actividades

Se ha presentado este trabajo en:

Eventos Científicos de Ingeniería de Software:

- Madrilenian Software Research
- Group Retreat 2019 Workshop
- SATToSE 2020: Seminar Series on Advanced Techniques & Tools for Software Evolution

Eventos dirigidos a la comunidad académica interdisciplinar:

- Periodismo de Datos (Medialab Prado)
- VI International Congress of Young Researchers with a Gender Perspective (UC3M 2021)
- I Congreso Internacional "Tecnologías I+D+i para la Igualdad: soluciones, perspectivas y retos" (UC3M 2021)
- Jornadas Online “Género y Ciencia de Datos en Deporte y Salud (UOC 2021)

Resultados

Software

Software Libre liberado con GPLv3 integrado en la industria

- git clone `https://github.com/davidam/damegender.git`
- pip3 install damegender

Publicaciones

- Damegender: Writing and Comparing Gender Detection Tools (CEUR)
- Damegender Manual: Counting Males and Females in Internet Communities

Licencia