Damegender: A Toolkit for to Measure Gender Gap with an Approach on Reproducibility

David Arroyo Menéndez

October 25, 2021

- Thesis Student: David Arroyo Menéndez
- Title of this slides: Towards an International Dataset about Names, Gender and Frequency
- Thesis Director: Jesús González Barahona

Academic Disciplines that could be finding profit on it

- Social Sciences: Secondary Sources in Gender Studies, Indicators about Gender Gap
- Computer Science: Software Engineering, Natural Language Processing
- Lingüistics: Studies about names

Research Works using these Open Datasets

- Gender Gap in Knowledge: Wikipedia, Twitter, Newspapers, Journal Papers, . . .
- Gender Gap in Software Engineering: Git, StackOverflow, . . .
- Lingüistics: Statistics in each language about number of last letters, first letters, phonems, . . . in males females

Damegender: A Toolkit measuring Gender Gap with an Approach on Reproducibility

Current State:

- Interface with Commercial APIs
- Automatically generate thousands of counts in CSV, Git, Mailing Lists, . . .
- To allow Machine Learning in nicknames and diminutives.
- Comparing using statistical tools: features in names, accuracies, confusion matrices, roc, principal component analysis, . . .
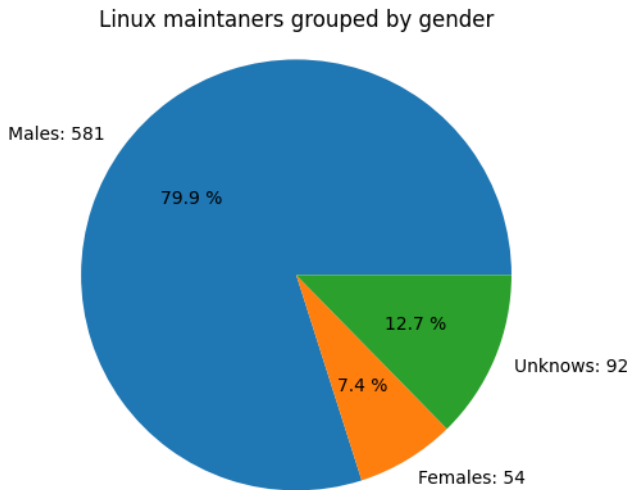
Damegender: Open Dataset

- Accuracy: 87.6%
- Number of females names: 299870
- Number of males names: 278981
- Open Data retrieved only from statistical institutions
- More than 20 countries

A new study applying Damegender to count males and females in GNU/Linux

Linux maintaners grouped by gender

Activities

We have presented this work in:

Scientific events on Software Engineering:

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

Event to master students and researchers in another disciplines:

- 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)

Results

Software

Free Software released with GPLv3 integrated in the industry

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

Publications

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