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

David Arroyo Menéndez

October 25, 2021

Presentation (I)

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

Ethical Motivation

- Gender Equality is the 5th Objective in United Nations.
- Only when you can measure a process, you can improve it.
- Reducing costs in the process in Free Software way, more academic people can measure it.

Contributions to the State of Art presented:

1. An integrated solution where make experiments in the different applications field relative to infering gender from the name.
1. A collection of Open Datasets retrieved from statistical sources and standarized in an unique format about gender, name and frequency
1. A new study applying DameGender to count males and females in GNU/Linux
1. A new Machine Learning approach classifyng gender from names
2. An approach based on reproducibility.

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

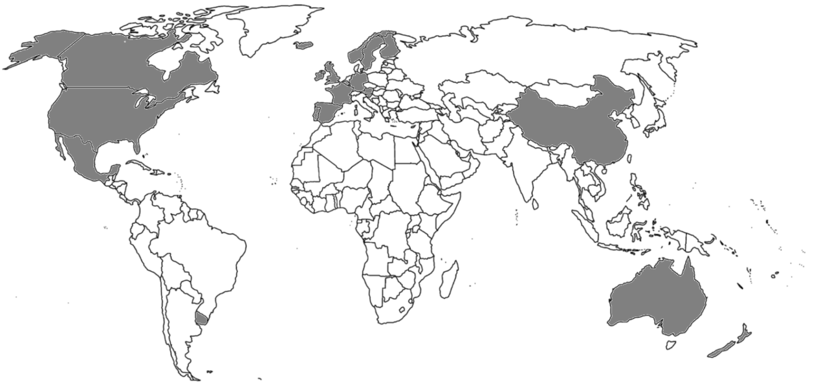
Current State:

- 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, . . .

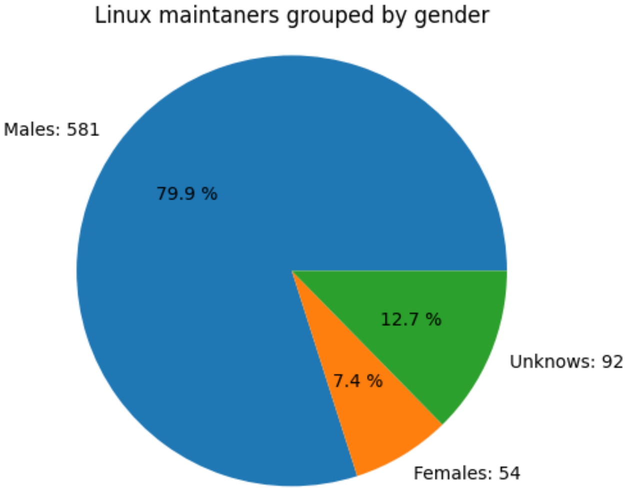
Damegender: Open Dataset

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

Damegender: Countries



Applying Damegender to count males and females in Linux (I)



Applying Damegender to count males and females in Linux (II)

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

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