Damegender: Towards an International Dataset about Names, Gender and Frequency

David Arroyo Menéndez

October 22, 2021

Presentation (I)

- 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

- Secondary sources in gender studies
- To build indicators about gender gap
- Natural Language Processing
- Studies about Lingüistics

Research Examples about it

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

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

Current state:

- Interface with Commercial APIs
- Automatic counts: csv2gender, git2gender, mail2gender, ...
- Machine Learning for nicknames and diminutives
- Comparing tools with: infofeatures, roc, principal components analysis (pca), confusion matrices, accuracies, precision, ...

Damegender Dataset

- Accuracy: 0.876
- Females: 299870
- Males: 278981
- Names retrieved only from statistical institutions with free copy license

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