Damegender: Building an universal dataset about names, frequencies and gender

David Arroyo Menéndez

June 9, 2021

Presentation

- Thesis Student: David Arroyo Menéndez
- Title: Building an universal dataset about names, gender and frequencies
- Thesis Supervisor: Jesús González Barahona
- Objectives:

To facilitate to reproducible science in this domain. We love peer review.
To be free to measure gender gap
We want to help to the semantic web effort.
We want to help to choose names in diverse contexts: LGTB people, different nations and cultures, . . .

Steps in the development

1. To develop the packaging and Python tests with the basic features
2. To develop statistical tools and bash tests
3. To add datasets more than 20 countries and an international dataset unifiying csv files to name, gender and frequency

Machine Learning to guess nicknames and new names

Making experiments with English and Spanish we obtained results higher than 70% reaching results similar to commercial results with very long datasets.

With more languages and nations we could to improving the guessing.

Adapting names to non binary people

These dataset are allowing to choose names for babies or people changing the name by LGTB reasons. So, we can use more neutral names understanding number of people using a name with equilibrium.

Measuring gender gap

We are using these datasets to measure gender gap in:

- Webometrics (spanish academic people in 2020): 35% females
- GNU (maintainers people in 2020): 6.6% females
- Linux (maintainers people in 2020): 7.4% females
- Debian (developers in 2020): 7.9% females

The percentage of unknows are 5-12%.

Semantic Web

Sites such as Wikipedia or about scientific people, social networks, could improve the HTML markup with microformats such as

```
<div class="h-card">
  <span class="p-name">Emma Goldman</span>
  <span class="p-gender p-gender-female p-gender-female-es
  p-gender-female-us p-gender-female-inter">
      Female
  </span>
  <span class="p-street-address">123 Main St</span>
  <span class="p-locality">Some Town</span>
  <span class="p-region">CA</span>
  <span class="p-postal-code">90210</span>
  <a class="u-url" rel="me"
     href="https://twitter.com/emmagoldman">@emmagoldman</a>
</div>
```

Activities

We have presented this working in progress in:

Scientific events:

- Madrilenian Software Research
- Group Retreat 2019 Workshop
- SATToSE 2020: Seminar Series on Advanced Techniques & Tools for Software Evolution
- UC3M 2021

Event to master students:

- Periodismo de Datos (Medialab Prado)

Industrial events:

- Python Barcelona
- Open South Code
- esLibre

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