DameGender

Towards an international and free dataset about name, gender and frequency

(David Arroyo Menéndez)
davidam@gmail.com

Equality of gender is the fifth objective of sustainable development for United Nations

This equality can be reached by measuring and analyzing data and making good politics with the results.

Many gender studies count males and females based on their names, for instance, research papers, job positions, streets, etc.

The traditional research method is to use commercial APIs with proprietary data without idea about how the data was collected.

genderapi, genderize, namsor, nameapi, ...

DameGender approach about Open Data is based collecting Open Datasets regarding name, gender and frequency from many statistical institutions.

Take a look about our mapamundi!

Take a look about our accuracies!

Dataset Test	Damegender Accuracy
Scientists Wikipedia	0.93
FIFA soccer	0.93
WTA tenis	0.91
National Leagues	0.91
Baby Names New York	0.98
Conseil Garonne	0.97
Paul Sebo	0.88
Lucia Santamaria & Helena Mihaljevic	0.88

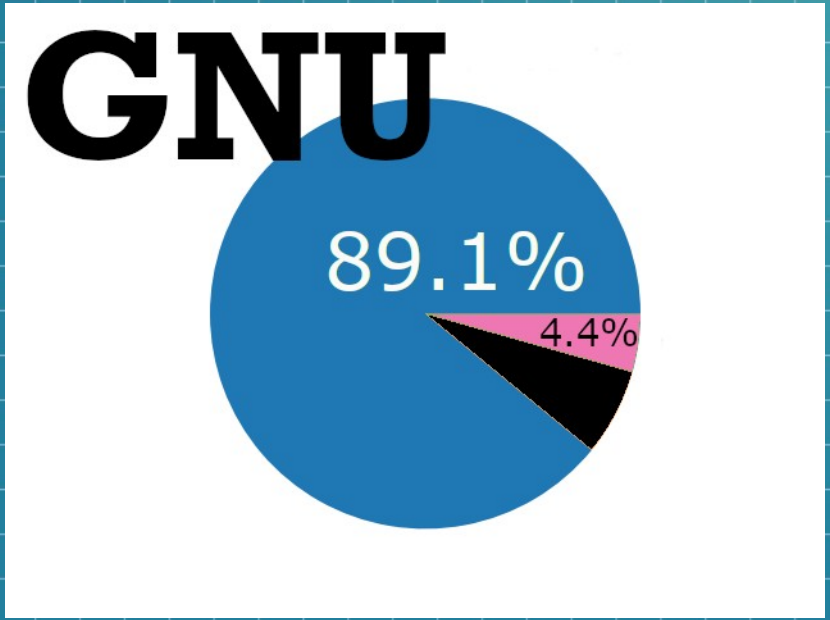
Comparing Damegender Free Dataset in number of names

Free Dataset	Number of males	Number of females
USA census (SSA)	91320	91320
Namdict (genderguesser)	48821	48821
NLTK	2943	5001
DameGender	257925	304553

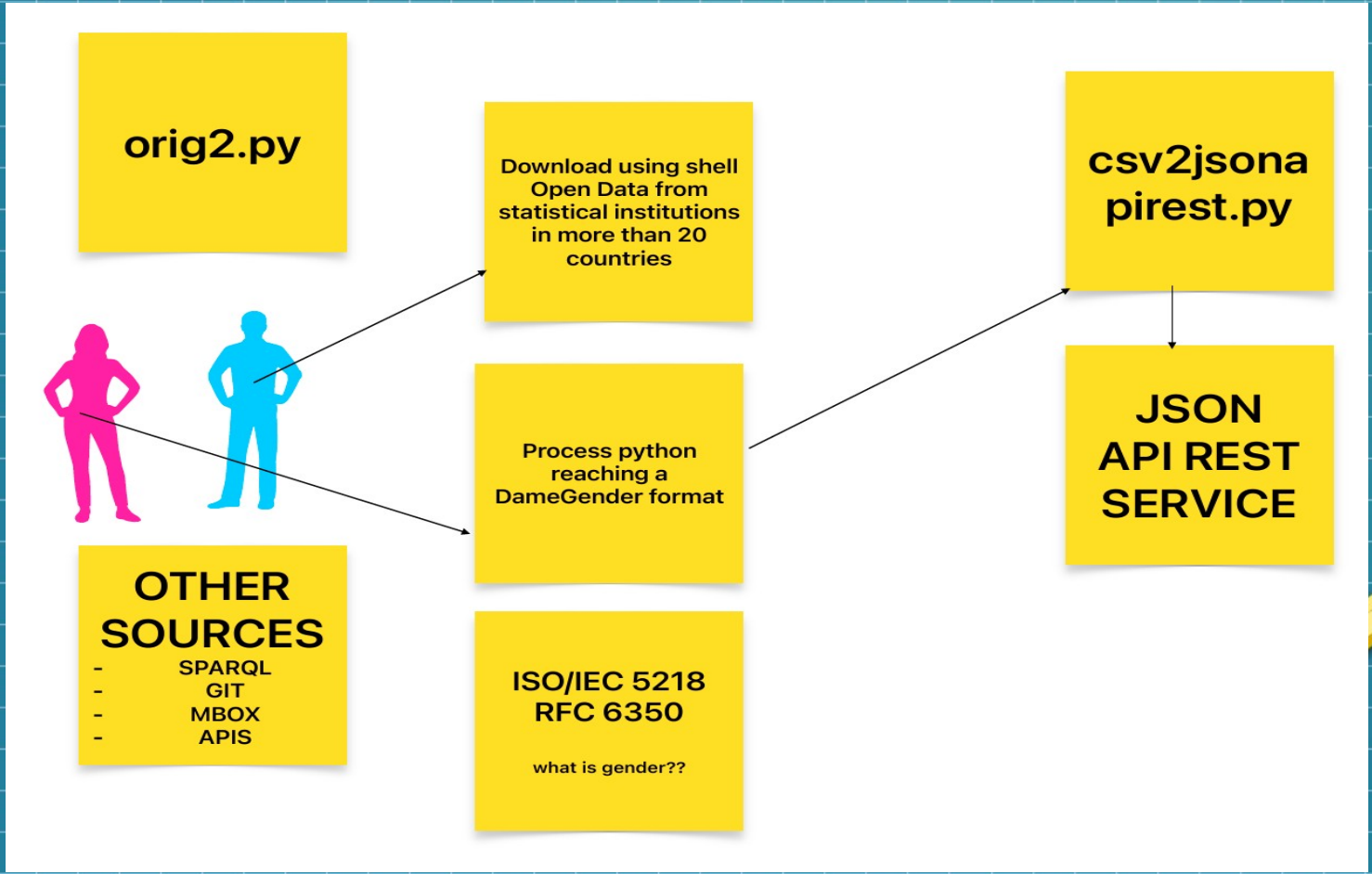
DameGender will allow to measure gender gap to students and academics interested on the phenomenon without costs and on a reproducible way and more people will be contributing to fix the gender gap.

csv2gender: infer names to gender from a csv file

```
$ python3 csv2gender.py files/gnu-maintainers.csv
–first_name_position=0  --title="GNU"
--dataset="inter" --outcsv="files/gnu.gender.csv"
--outimg="files/debian.gender.png"
```



orig2.py and csv2jsonapirest.py



Many funny commands!

DameGender in a poster

Gender Detection Tool from the Name

David Arroyo Menéndez and Jesús González barahona

URJC University

Why?

- If you want to determine gender gap in Free Software projects or mailing lists
- If you don't know the gender about a name
- If you want research with statistics about why a name is related with males or females.
- If you want use the commercial apis in gender detection (genderize, genderapi, namsor, nameapi) from a command.

Reducing the gender gap. How? (I)

Create Free Software tools to determine gender in Internet can help to reduce the gender gap in the world

Basic usages

```
$python3 main.py David
$python3 main.py Mesa --ml=nltk
$python3 nameincountries.py David
$python3 surname.py Arroyo --total=es
$python3 surnameincountries.py Arroyo
$python3 api2gender.py Leticia --api=genderapi
```

Easy Installation and Testing

```
$ npm i damegender
$ pip3 install damegender[all]
$ python3 apikeyadd.py
$ cd src/damegender & nosetest3 test
$ ./testsbycommands.sh & ./testsbycommandsextralocal.sh
```

Statistics

```
$ python3 infofeatures.py
$ python3 confusion.py
$ python3 errors.py --csv="files/names/all.csv"
--api="genderguesser"
$ python3 pca-components.py --csv="files/features list.csv"
$ python3 pca-features.py
$ python3 top.py es
$ python3 accuracy.py --api="genderguesser" --measure="recall"
--csv=files/names/allnoundefined.csv
$ python3 csv2gender.py files/forbes2020.csv
--first name position=0
```

Open Datasets

- Official Statistics Offices
- Wikidata
- GenderGuesser

Reducing the gender gap. How? (II)

Measuring gender gap in software repositories and taking actions about it can help to reduce the gender gap in the world

{jgb}@gsyc.es https://github.com/davidam/damegender.git {davidam}@gnu.org

Machine Learning Approach (I)

First querying to the dataset
Later, infering the gender with Machine Learning
The features was calculated with Principal Component Analysis
You can decide to use several machine learning algorithms
The best result in the benchmarks was with SVC

```
% python3 main.py Mesala
Mesala gender predicted is female
0 males for Mesala from international statistics
0 females for Mesala from international statistics
% python3 main.py Ana --ml=svc
Ana's gender is female
probability: 0.999140241743792
Ana gender predicted is female
680 males for Ana from international statistics
790240 females for Ana from international statistics
```

Machine Learning Approach (II)

The current state of the ML models has been designed with French, Canada Spanish and USA

The next step is to include all main languages in the world

Enjoying the good work with the Open Dataset

Conclusions and Future Works

Open Datasets from statistical institutions is reaching accuracies bigger than 90%

These accuracies will be improved updating ML models

The discussions about gender concepts will be fixed applying standards

orig2.py will become the apt or npm of datasets about gender and names

Thanks!!