Damegender Manual: Counting Males and Females in Internet Communities

for version 0.3.1, 4 Dec 2020

David Arroyo Menéndez (davidam@gnu.org)
Prologue by Lucía Santamaría (lucia.santamaria@ymail.com)

This manual is for Damegender (version 0.3.1, 4 Dec 2020).

The sources will be find in https://github.com/davidam/damegender/tree/master/manual

Table of Contents

1 Prologue 1

2 Introduction 2

3 Installation 3

4 Commands 4

5 Statistics 9

5.1 Measuring success and error 9
5.2 Principal Component Analysis (PCA) 13
5.2.1 Counting features in names 13
5.2.2 Choosing components 14

6 Use Cases 17

6.1 Introduction 17
6.2 Counting males and females in Debian 17
6.3 Counting males and females in Linux Kernel 19
6.4 Counting males and females in Forbes 20
6.5 Deciding for males and females in images 22
6.6 Webscraping and Damegender (counting scholars) 22
6.7 Counting males and females in a git repository 24
6.8 Counting males and females in Maps 25
6.9 Gender gap in science 26

7 Secondary Sources about the Gender Gap ... 27

7.1 Gender Inequality in the World 27
7.2 Gender Inequality in STEM 30
7.3 Gender Inequality in Free Software 30

8 Theorical Frameworks 34

8.1 Philosophies about software, market, freedom and gender 34
8.2 Multiculturalism, Interculturalism 37
8.3 Feminism, Ecofeminism and Intersectionality 39
8.4 Gender 40

9 Conclusions 41

Further reading 42

Appendix A License **45**

Index **46**

1 Prologue

The algorithmic assignment of a particular gender to a person based on their name has become a task of interest for sociological and gender studies as an useful research tool in multiple areas. Examples include the investigation of gender dynamics and gender biases in social media, scholarly output, and scientific and technical contributions, among others. Large-scale approximations based on data and Machine Learning as well as the rising availability of assorted corpora of labelled data names make it feasible to train and evaluate classification algorithms in order to produce models that can predict gender from names very accurately. Damegender is one such contribution to this space, where other commercial and non-commercial actors are also present.

Yet one fundamental caveat worth mentioning when approaching the gender prediction task pertains an ethical aspect: personal names are assigned to individuals at birth as part of a schema based on a binary, immutable, and physiologically determined definition of gender. Any approach that automates gender assignment denies the view that one's gender identity is profoundly subjective and cannot and should not be reduced to a binary label. It is virtually impossible to accurately assign a gender to every individual without misgendering a certain number of them; it is also non-inclusive and fundamentally unethical.

If gender studies based on individuals' names are to be carried out for the sake of answering a particular sociological question or hypothesis, inclusive and fair methods ought to be design to aid in this process. Self-identification is one simple avenue that is obviously not always practicable and is difficul to scale. Thus, at the very least, one should bear in mind that the question of ethical treatment of gender needs to be stressed every time that automatic assignments are made. Additionally, only results and conclusions based on aggregated data should be pursued, and never must an identifiable individual be assigned a gender automatically and quoted as such on a publication.

These caveats notwithstanding, the tool proposed in this manual, Damegender, offers a valuable contribution for powering gender analyses on assorted collections of names. It's compilation of examples and tests will be highly useful for the interested researcher that needs to apply automatic gender detection to their own data. We shall hope that those studies will eventually lead to a deeper understanding of existing gender dynamics in our society and to the development of fair and sustainable proposals to close the gender gap.

December 2020, Lucía Santamaría

2 Introduction

Damegender is a gender detection tool from the name coded by David Arroyo MEnéndez (DAME). See "*Damegender: Writing and Comparing Gender Detection Tools*", [Further reading], page 42.

The gender detection tools from the names are being used usually with commercial APIs. But many countries has been doing efforts in the last years for contribute names and a number of people using each name with Open Data Licenses. So, this software is collecting this effort in an industrial way and giving new original solutions to classify gender from the name (we are using Machine Learning algorithms for predict names that is not appearing in our database).

Damegender is giving measures to compare in any moment our solution with the commercial APIs. So, the user can understand when it's useful to invest money or not depending of the dataset. Damegender allows to the users download a big number of names from a csv file.

This software is written oriented to tests. So you can check the right behaivour of the software with python tests for the classes and methods and with shell tests for the python commands.

Damegender is using Perceval for count males and females in a lot of Internet Communities (wikis, mailing lists, software repositories, bug tracking systems, ...). "*Perceval: software project data at your will*", [Further reading], page 42, This manual show source for count males and females in different situations (Ex: `count-debian-gender.py`).

This software is taking into account the power to predict nations and ethnicity from the surnames (Ex: `surname.py`, `surnameincountries.py` and `ethnicity.py`).

This book starts explaining the installation. Later, it explains how to use Damegender from the commands. After, it explains Damegender from the commands. Next, it explains the statistical maths concepts to use Damegender with good results. The use cases showed allows to imagine a lot of applications about Damegender. We are giving some data sources about gender gap. And finally, the theorical frameworks are being showed to put data into discourses. We are added conclusions as summary and further reading for extend the interest.

3 Installation

Possible Debian/Ubuntu dependencies:

```
$ sudo apt-get install python3-nose-exclude python3-dev dict
dict-freedict-eng-spa dict-freedict-spa-eng dictd
```

Now, to install damegender with python package:

```
$ python3 -m venv /tmp/d
$ cd /tmp/d
$ source bin/activate
$ pip install --upgrade pip
$ pip3 install damegender
$ cd lib/python3.5/site-packages/damegender
$ python3 main.py David
```

To install apis extra dependencies:

```
$ pip3 install damegender[apis]
```

To install mailing lists and repositories extra dependencies:

```
$ pip3 install damegender[mails_and_repositories]
```

To install all possible dependencies

```
$ pip3 install damegender[all]
```

Currently you can need an api key from:

- https://store.genderize.io/documentation
- https://gender-api.com
- https://www.nameapi.org/
- https://v2.namsor.com/NamSorAPIv2/sign-in.html

To configure your api key you can execute:

```
$ python3 apikeyadd.py
```

4 Commands

You must start to check tests to understand that all is ok:

```
$ cd src/damegender
$ ./testsbycommands.sh         # It must run for you
$ ./testsbycommandsextralocal.sh  # You will need all dependencies
                                  # with: $ pip3 install damegender[all]
$ ./testsbycommandsextranet.sh    # You will need api keys
```

You can continue checking python tests:

Execute all tests:

```
$ nosetests3 tests
```

Execute one file:

```
$ nosetests3 tests/test_basics.py
```

Execute one test:

```
$ nosetests3 tests/test_basics.py:TestBasics.test_indexing
```

If you are in a fresh installation, perhaps you want regenerate by your own risk some files downloaded to understand how it has been generated:

```
$ python3 postinstall.py
```

You can find an big list of commands to execute this shell scripts. Now a detailed execution of some selected examples:

The first command to learn is `main.py`. You can play now with this command:

```
# Detect gender from a name (INE is the dataset used by default)
$ python3 main.py David
David gender is male
363559  males for David from INE.es
0 females for David from INE.es

# Detect gender from a name only using machine learning
$ python3 main.py Agua --ml=nltk
Agua gender is female
0 males for Agua from INE.es
0 females for Agua from INE.es

# Detect gender from a name (all census and machine learning)
$ python3 main.py David --verbose
365196 males for David from INE.es
0 females for David from INE.es
1193 males for David from Uruguay census
5 females for David from Uruguay census
26645 males for David from United Kingdom census
0 females for David from United Kingdom census
3552580 males for David from United States of America census
12826 females for David from United States of America census
David gender predicted with nltk is male
```

```
David gender predicted with sgd is male
David gender predicted with svc is male
David gender predicted with gaussianNB is male
David gender predicted with multinomialNB is male
David gender predicted with bernoulliNB is male
David gender predicted with forest is male
David gender predicted with tree is male
David gender predicted with mlp is male
```

The first Free Software for gender detection tool was created in C language program and you can look for a python version with the name genderguesser. Some people was working in a Free dataset called name_dict.txt with 48500 names. I want to give thanks to this effort with `nameincountries.py` due to the good work organizing many names in different countries.

```
$ python3 nameincountries.py David
grep -i " David " files/names/nam_dict.txt > files/grep.tmp
males: ['Albania', 'Armenia', 'Austria', 'Azerbaijan', 'Belgium',
'Bosnia and Herzegovina', 'Czech Republic', 'Denmark', 'East Frisia',
'France', 'Georgia', 'Germany', 'Great Britain', 'Iceland', 'Ireland',
'Israel', 'Italy', 'Kazakhstan/Uzbekistan', 'Luxembourg', 'Malta',
'Norway', 'Portugal', 'Romania', 'Slovenia', 'Spain', 'Sweden',
'Swiss', 'The Netherlands', 'USA', 'Ukraine']
females: []
both: []
```

To count gender from a git repository:

```
$ python3 git2gender.py
https://github.com/chaoss/grimoirelab-perceval.git
--directory="/tmp/clonedir"
The number of males sending commits is 15
The number of females sending commits is 7
```

You can see a verbose output using the spanish dataset (–language=es) for males and females with:

```
$ python3 git2gender.py https://git.drupalcode.org/project/orgmode.git
--directory=/tmp/orgmode --show=all --verbose --language=es
You are not using ml the process is not very slow, but perhaps
you are not finding good results
The number of males sending commits is 2
The list of males sending commits is:
['David Arroyo Menendez', 'David Arroyo']
David Arroyo Menéndez <davidam@es.gnu.org> (67 commits)
David Arroyo Menendez <davidam9@riseup.net> (49 commits)
David Arroyo Menéndez <davidam@gmail.com> (20 commits)
David Arroyo Menendez <david.arroyo@edoctores.com> (10 commits)
David Arroyo Menendez <davidam@es.gnu.org> (14 commits)
David Arroyo7 <davidam@es.gnu.org> (13 commits)
David Arroyo7 <davidam@gnu.org> (10 commits)
```

```
The number of females sending commits is 1
The list of females sending commits is:
['Miriam]
Miriam <miriam@xxxxxx.es> (23 commits)
The number of people with unknown gender sending commits is 0
The list of people with unknown gender sending commits is:
[]
```

To count gender from a mailing list:

```
$ cd files/mbox
$ wget -c
http://mail-archives.apache.org/mod_mbox/httpd-announce/201706.mbox
$ cd ../..
$ python3 mail2gender.py
http://mail-archives.apache.org/mod_mbox/httpd-announce/
You are not using ml the process is not very slow, but perhaps you are
not finding good results

The number of males sending mails is 24
The number of females sending mails is 2
The number of people with unknown gender sending mails is 5
```

You can execute a verbose output with:

```
$ python3 mail2gender.py
http://mail-archives.apache.org/mod_mbox/httpd-announce/
--verbose --show=all
You are not using ml the process is not very slow, but perhaps you are not
finding good results
The number of males sending mails is 24
The list of males sending mails is:
['Jim <jim@xxxxxx.es>', 'Jacob <jchampion@xxxxx.org>',
'DENNIS <balaranpillai@xxxxx.com>', '"Leonard (Jira)" <jira@xxxxx.org>',
'"Roy" <jira@xxxxx.org>', '"Roman (Jira)" <jira@xxxxx.org>',
'"Bertrand" <jira@xxxxx.org>', '"Mark (Jira)" <jira@xxxxx.org>',
'"Justin (Jira)" <jira@xxxxx.org>', '"Simon (Jira)" <jira@xxxxx.org>',
'"Chris (Jira)" <jira@xxxxx.org>', 'Jan <lahoda@xxxxx.com>',
'"Michael (Jira)" <jira@xxxxx.org>', '"Ralph (Jira)" <jira@xxxxx.org>',
'"Jens" <jensg@xxxxx.org>', 'Mark <markt@xxxxx.org>',
'"Ryan (Jira)" <jira@xxxxx.org>', 'Ismaël (Jira)  <jira@xxxxx.org>',
'"Shane (Jira)" <jira@xxxxx.org>', '"Kevin A. (Jira)" <jira@xxxxx.org>',
'"Gordon (Jira)" <jira@xxxxx.org>', 'Gary <garydgregory@xxxxx.com>',
'"Owen" <owen.omalley@xxxxx.com>', '"Sheng (Jira)" <jira@xxxxx.org>']
The number of females sending mails is 2
The list of females sending mails is:
['Riya <hellen.serviceweb@xxxxxx.com>',
'"Hannah (Jira)" <jira@xxxxx.org>']
The number of people with unknown gender sending mails is 5
The list of people with unknown gender sending mails is
```

```
['"SimpaticoTech" <web.info@xxxxxx.it>',
'Simpatico <web.info@xxxxxxxxx.it>',
'gmcdonald@xxxxx.org',
'Hen <bayard@xxxxx.org>',
'"Jean (Jira)" <jira@xxxxx.org>']
```

Perhaps you don't know a name, but you have obtained an free key for an api to retrieve it:

```
$ python3 api2gender.py Leticia --surname="Martin" --api=namsor
female
scale: 0.99
```

If you want to know the gender of a good number of names you can download results from an api and save in a file with `downloadjson.py`.

```
$ python3 downloadjson.py --csv=files/names/min.csv --api=genderize
$ cat files/names/genderizefiles_names_min.csv.json
```

Now we are going to learn some commands for measure the successful of our solution:

```
$ python3 accuracy.py --csv=files/names/min.csv
################### NLTK!!
Gender list: [1, 1, 1, 1, 2, 1, 0, 0]
Guess list:  [1, 1, 1, 1, 0, 1, 0, 0]
Dame Gender accuracy: 0.875
```

```
$ python3 confusion.py --csv="files/names/partial.csv" --api=nameapi
--jsondownloaded="files/names/nameapifiles_names_partial.csv.json"
A confusion matrix C is such that Ci,j is equal to the number of
observations known to be in group i but predicted to be in group j.
If the classifier is nice, the diagonal is high because there are true
positives Nameapi confusion matrix:

[[ 3, 0, 0]
 [ 0, 15, 1]]
```

```
$ python3 errors.py --csv="files/names/all.csv" --api="genderguesser"
Gender Guesser with files/names/all.csv has:
+ The error code: 0.22564457518601835
+ The error code without na: 0.026539047204698716
+ The na coded: 0.20453365634192766
+ The error gender bias: 0.0026103980857080703
```

You can generate a lot of logs about errors, accuracies and/or confusion:

```
$ ./logs-accuracies.sh
$ ./logs-confusion.sh
$ ./logs-errors.sh
```

Perhaps you are interested on reproduce experiments to determine features:

```
$ python3 infofeatures.py
# To determine number of components
$ python3 pca-components.py --csv="files/features_list.csv"
# To understand the weight between variables for a target
```

```
$ python3 pca-features.py
```

Now we can go to play with surnames:

```
$ python3 surname.py Gil --total=es
There are 140004 people using Gil in Spain

$ python3 surname.py Lenon --total=us
There are 837 people using Lenon in United States of America

$ python3 ethnicity.py Smith
In United States of America the percentages about the race
of Smith surname is:
White: 73.35
Black: 22.22
Hispanic: 1.56
Asian Pacific Indian American: 0.40
American Indian and Alaska Native: 0.85
Various races: 1.63
```

5 Statistics

In the last chapter we were learning to execute some commands such as `accuracy.py`, `confusion.py`, or `errors.py`, but perhaps you need to understand more theory about statistics to understand why this commands is being interesting for you.

5.1 Measuring success and error

To guess the sex, we have an true idea (example: female) and we obtain a result with a method (example: using an api, querying a dataset or with a machine learning model). The guessed result could be male, female or perhaps unknown. To remember some definitions about results about this matter:

True positive is to find a value guessed as true if the value in the data source is positive.

True negative is to find a value guessed as true if the the value in the data source is negative.

False positive is to find a value guessed as false if the the value in the data source is positive.

False negative is to find a value guessed as false if the the value in the data source is negative.

So, we can find a vocabulary for measure true, false, success and errors. We can make a summary in the gender name context about mathematical concepts:

Precision is about true positives divided by true positives plus false positives

```
(femalefemale + malemale ) /
(femalefemale + malemale + femalemale)
```

Recall is about true positives divided by true positives plus false negatives.

```
(femalefemale + malemale ) /
(femalefemale + malemale + malefemale +
+ femaleundefined + maleundefined)
```

Accuracy is about true positives divided by all.

```
(femalefemale + malemale ) /
(femalefemale + malemale + malefemale +
+ femalemale + femaleundefined + maleundefined)
```

The **F1 score** is the harmonic mean of precision and recall taking both metrics into account in the following equation:

```
2 * (
(precision * recall) /
(precision + recall))
```

In Damengender, we are using `accuracy.py` to apply these concepts. Take a look tothe practice:

```
$ python3 accuracy.py --api="damegender" --measure="f1score"
--csv="files/names/partialnoundefined.csv"
--jsondownloaded=files/names/partialnoundefined.csv.json
################## Damegender!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
```

```
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Damegender f1score: 0.9090909090909091

$ python3 accuracy.py --api="damegender" --measure="recall"
  --csv="files/names/partialnoundefined.csv"
  --jsondownloaded=files/names/partialnoundefined.csv.json
#################### Damegender!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Damegender recall: 1.0

$ python3 accuracy.py --api="damegender" --measure="accuracy"
  --csv="files/names/partialnoundefined.csv"
  --jsondownloaded=files/names/partialnoundefined.csv.json
#################### Damegender!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Damegender accuracy: 0.8571428571428571

$ python3 accuracy.py --api="genderguesser" --measure="accuracy"
  --csv="files/names/partialnoundefined.csv"
  --jsondownloaded=files/names/partialnoundefined.csv.json
#################### Genderguesser!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Genderguesser accuracy: 0.8571428571428571

$ python3 accuracy.py --api="genderguesser" --measure="precision"
  --csv="files/names/partialnoundefined.csv"
  --jsondownloaded=files/names/partialnoundefined.csv.json
#################### Genderguesser!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Genderguesser precision: 0.9090909090909091

$ python3 accuracy.py --api="genderguesser" --measure="recall"
  --csv="files/names/partialnoundefined.csv"
```

```
  --jsondownloaded=files/names/partialnoundefined.csv.json
################### Genderguesser!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Genderguesser recall: 1.0

$ python3 accuracy.py --api="genderguesser" --measure="f1score"
  --csv="files/names/partialnoundefined.csv"
  --jsondownloaded=files/names/partialnoundefined.csv.json
################### Genderguesser!!
Gender list: [1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Guess list:  [1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1,
              1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0]
Genderguesser f1score: 0.9090909090909091
```

Error coded is about the true is different than the guessed:

```
(femalemale + malefemale + maleundefined + femaleundefined) /
(malemale + femalemale + malefemale +
+ femalefemale + maleundefined + femaleundefined)
```

Error coded without na is about the true is different than the guessed, but without undefined results.

```
(maleundefined + femaleundefined) /
(malemale + femalemale + malefemale +
+ femalefemale + maleundefined + femaleundefined)
```

Error gender bias is to understand if the error is bigger guessing males than females or viceversa.

Weighted error is about the true is different than the guessed, but giving a weight to the guessed as undefined.

```
(femalemale + malefemale +
+ w * (maleundefined + femaleundefined)) /
(malemale + femalemale + malefemale + femalefemale +
+ w * (maleundefined + femaleundefined))
```

In Damegender, we have coded `errors.py` to implement several definitions in different apis.

The confusion matrix creates a matrix about the true and the guess. If you have this confusion matrix:

```
[[ 2, 0, 0]
 [ 0, 5, 0]]
```

It means, I have 2 females true and I've guessed 2 females and I've 5 males true and I've guessed 5 males. I don't have errors in my classifier.

```
[[ 2  1  0]
[ 2 14  0]]
```

It means, I have 2 females true and I've guessed 2 females and I've 14 males true and I've guessed 14 males. 1 female was considered male, 2 males was considered female.

In Damegender, we have coded `confusion.py` to implement this concept:

```
python3 confusion.py --csv=files/names/min.csv
--api=damegender --jsondownloaded=files/names/min.csv.json
A confusion matrix C is such that Ci,j is equal to the number of observations known to
If the classifier is nice, the diagonal is high because there are true positives
Damegender confusion matrix:

      M   F   U
M  [[ 5,  0,  0 ]
F   [ 0,  1,  0 ]]
```

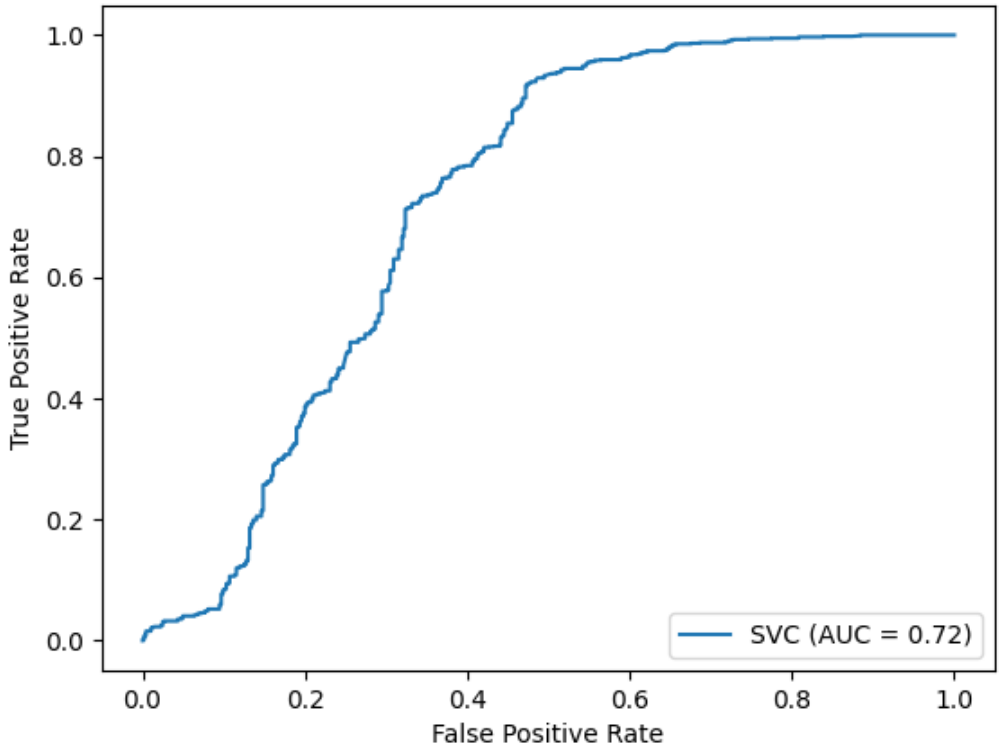
Remember that we can retrieve the json file from several apis having the names to be guessed in the csv file with `downloadjson.py`:

```
$ python3 downloadjson.py --csv="files/names/min.csv" --api="genderapi"
```

Similar to confusion is ROC (Receiver Operating Characteristic) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

In Damegender, you can use ROC relative to machine learning algorithms with the next command:

```
$ python3 roc.py svc
```



5.2 Principal Component Analysis (PCA)

5.2.1 Counting features in names

We have developed a script `infofeatures.py` with our datasets to visualize data about some features chosen by us.

```
$ python3 infofeatures.py ine
```

Take a look to the results with the different datasets:

Dataset	**Letter A**	**Last Letter A**	**Last Letter O**	**Last Letter Consonant**	**Last Letter Vocal**	**First Letter Consonant**	**First Letter Vocal**
Uruguay (females)	0.816	0.456	0.007	0.287	0.712	0.823	0.177
Uruguay (males)	0.643	0.249	0.062	0.766	0.234	0.771	0.228
Australia (females)	0.922	0.588	0.033	0.272	0.728	0.772	0.228

Australia (males)	0.818	0.03	0.269	0.57	0.43	0.763	0.237
Canada (females)	0.659	0.189	0.005	0.591	0.408	0.838	0.161
Canada (males)	0.752	0.22	0.025	0.54	0.456	0.818	0.181
Spain (females)	0.922	0.588	0.03	0.271	0.728	0.772	0.228
Spain (males)	0.818	0.03	0.268	0.569	0.43	0.763	0.236
United Kingdom (females)	0.825	0.374	0.013	0.322	0.674	0.765	0.235
United Kingdom (males)	0.716	0.036	0.039	0.78	0.218	0.799	0.2
USA (females)	0.816	0.456	0.007	0.287	0.712	0.823	0.177
USA (males)	0.643	0.02	0.061	0.765	0.234	0.84	0.159

The countries where the main language is spanish (Uruguay + Spain) and english (USA + United Kingdom + Australia) are having very similar variation with the features chosen between males and females with these datasets (remember is the datasets extracted from official statistics provided by the states). Canada, a country french centric has different rules with this features.

The letter a is varying 0.2 from males to females in (USA and Uruguay) and 0.1 from males to females (United Kingdom, Australia and Spain). The last letter a is varying 0.5 from males to females in (Australia, Spain) around 0.4 in (USA, United Kingdom) and 0.2 in Uruguay. The last letter o from females to males is varying 0.2 in (Spain, Australia) and is equal in (Uruguay, USA, United Kingdom). For the last letter consonant all countries is giving the result that is for males, with results from 0.2 to 0.5: Uruguay and USA (0.5), United Kingdom (0.4), Australia and Spain (0.3). So last letter vocal is reverse tha last letter consonant. First letter consonant or first letter vocal is a non significative feature due to so similar results in english and spanish.

Surely, the rules it's a coincidence but we think that is a coincidence between languages due to that there are a good number of names to think different.

5.2.2 Choosing components

After, to choose features for our machine learning task, we can understand if this features makes sense with Principal Component Analysis. We have written 2 scripts for this task `pca-components.py` and `pca-features.py`. With `pca-components.py` we are giving a csv (files/features_list.csv, files/features_list_no_cat.csv, ...) and the output is an image where

we can visualize a curve to determine when this curve stops the growth the number of components.

In the image, we can see that the curve stops the growth in the fourth component.

When you know the components you can execute `pca-features.py` so:

```
$ python3 pca-features.py  --categorical=both --components=4
The json file is created in files/pca.json
The html file is created in files/pca.html
```

first_letter	last_letter	last_letter_a	first_letter _vocal	last_letter _vocal	last_letter _consonant	target component
-0.2080025204	-0.3208958517	0.2352509625	0.2113242731	**0.6095269139**	**-0.6095269139**	-0.1035071139
-0.6037951881	**0.5174873789**	-0.4252467151	0.4278794455	0.0388287435	-0.0388287435	-0.0265942125
0.1049343046	0.1158117877	-0.2867605971	-0.3473950734	0.0901034539	-0.0901034539	-0.8697264971
0.2026467275	0.3142402839	**0.630802294**	**0.5325769702**	-0.1291229841	0.1291229841	-0.3811720011

To simplify and to learn, we can observe this analysis without letters. In this analysis, we can observe 4 components.

The first component is about if the last letter is vocal or consonant. If the last letter is vocal we can find a female and if the last letter is a consonant we can find a male.

The second component is about the first letter. The last letter is determining females and the first letter is determining males.

The third component is not giving relevant information.

The fourth component is giving the last_letter_a and the first_letter_vocal is for females.

6 Use Cases

6.1 Introduction

There are many research studies count males and females in specific communities such as Twitter, StackOverflow, ... A specific community has some clues to determine male or female, for example, in Twitter you can observe the photo, nickname, real name, ...

In this chapter we are going to apply the concepts to determine gender in real situations observing where the gender is provided.

6.2 Counting males and females in Debian

In the Debian community all members must have a gpg key to collaborate, so we can count males and females from the keyring. With gpg commands you can import a the debian keyring and dump the debian keyring in a csv file.

```
$ rsync -az --progress keyring.debian.org::keyrings/keyrings/ .
```

We have generated a script to count males and females:

```
~/git/damegender/src/damegender$ python3 count-debian-gender.py
Perhaps you need wait some minutes. You can take a tea or coffe now
debian males: 795
debian females: 24
```

In the dump of the debian keyring dataset we have divided name, surname and email in different fields. So, it's easy detect the name, although some names has several emails.

We have choosen the United States of America dataset and we are using the method name_frec to decide for male or female in the row. Take a look to the source:

```
import csv
import unicodedata
import unidecode
from pprint import pprint
import re
from app.dame_gender import Gender
from app.dame_utils import DameUtils

du = DameUtils()
g = Gender()

result=""
dm = []

with open('files/debian-maintainers-gpg-2020-04-01.csv') as csvfile:
    reader = csv.reader(csvfile, delimiter=',', quotechar='|')
    aux = ""
    cnt = 0
    for row in reader:
        cnt = cnt +1
```

```
        if (aux != row[0]):
            dm.append(row[0])
        aux = row[0]

print("Perhaps you need wait some minutes. You can take a tea or coffe now")

females = 0
males = 0
for rowdm in dm:
    if (int(g.name_frec(str(rowdm.upper()), 'us')['females'])
    > int(g.name_frec(str(rowdm.upper()), 'us')['males'] )):
        females = females + 1
    else:
        males = males + 1

print("debian males: %s" % males)
print("debian females: %s" % females)

csvfile.close()
```

The advantage using the method `name_frec` is about to understand how you are deciding male or female in the script counting males and females. In this script the decision is simple: a name is male if there are more males than females and female if there are more females than males.

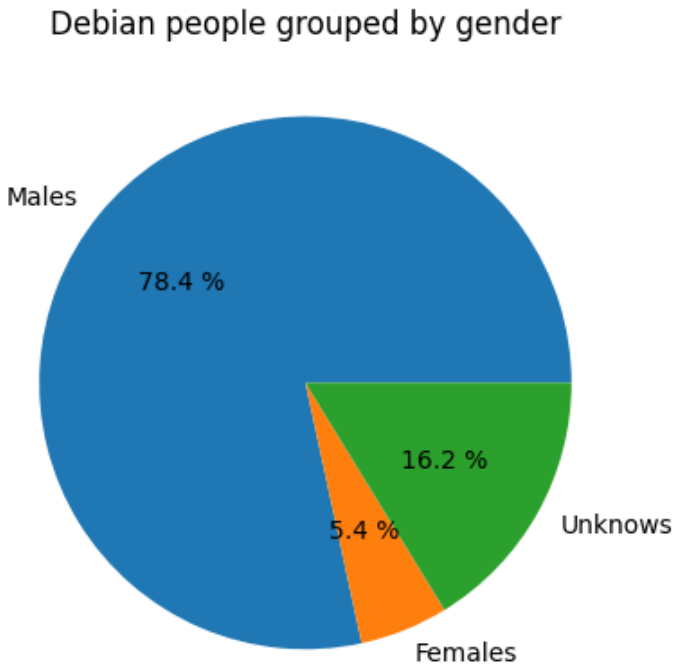
The United States of America dataset is a good choice for Free Software communities, due to that this communities is based on english as main language and United States of America is a leader country in software development. United States of America hosts people from different countries due to migrations towards good companies and universities.

In general, you can choose `csv2gender.py` to count males, females and unknows in a csv file. For example, doing this:

```
$ python3 csv2gender.py --first_name_position=0
files/debian-maintainers-gpg-2020-04-01.csv --verbose
```

But, a research must understand the source, too.

In the next diagram we can see (78.4% of males, 5.4% of females and 16.2% of unknowns).



We can retrieve the names of unknows and to decide about the name in the sense of retrieve the gender from a commercial api (genderapi, genderize, namsor, nameapi, ...) or to classify the name as a software company or a bug. For now, the Open datasets contributed by states about names are very good but not all countries has the idea of contribute the names. Remember that you can download names from a csv file with `downloadjson.py`:

```
$ python3 downloadjson.py --api=genderize --csv=files/names/min.csv
```

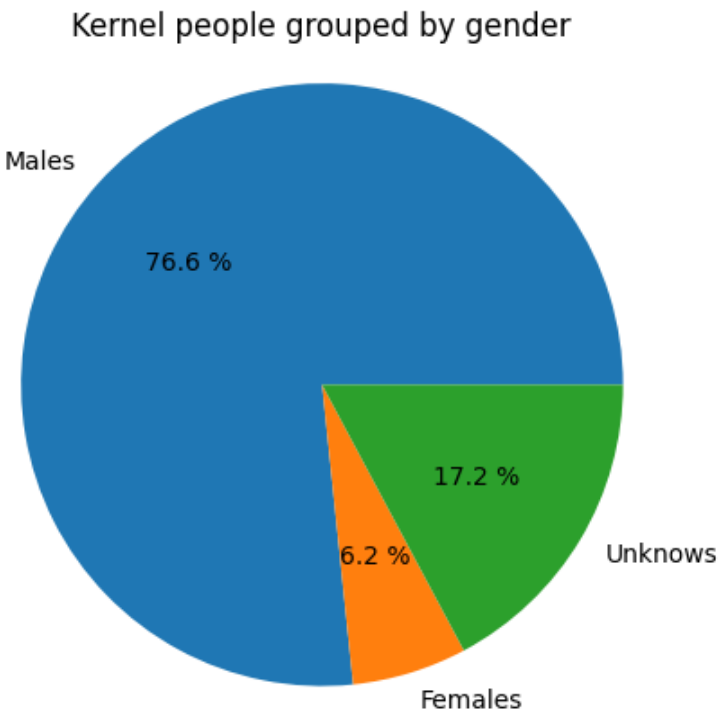
6.3 Counting males and females in Linux Kernel

When I'm writing this book the Linux Kernels maintainers appears in https://www.kernel.org/doc/html/latest/process/maintainers.html. Then I have downloaded the file and applied a single command:

```
cat maintainers.html | w3m -dump -T text/html
| grep "Mail:" > maintainers.txt
```

You can makes fixes to this command from GNU/Emacs or with shell scripting. Later, you can apply:

```
$ python3 count-kernel.py
```



6.4 Counting males and females in Forbes

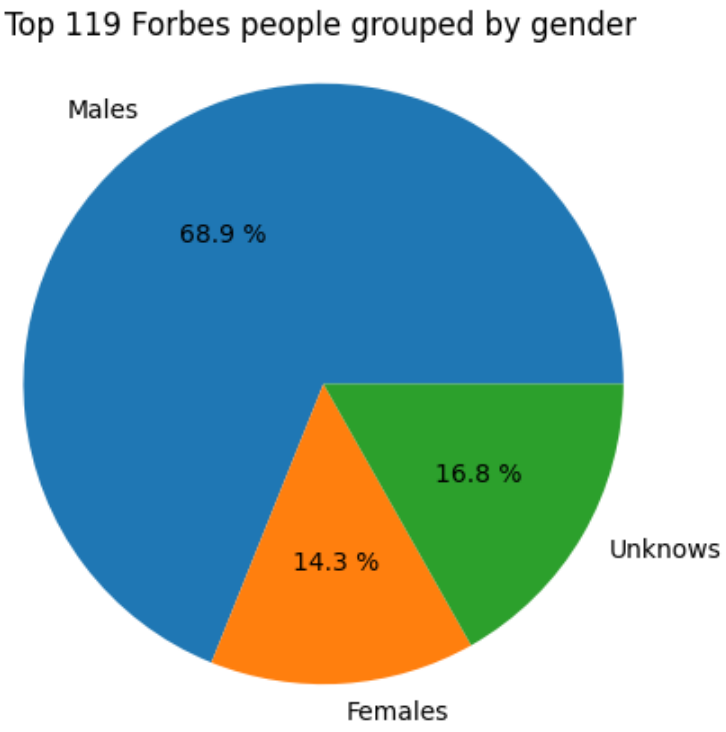
In the second example, we are using guess without machine learning instead of name_frec. If you are using guess you are trusting on damegender to take the decision, but perhaps you are not agree.

Please take a look about our guess method in the current state:

```
def guess(self, name, binary=False, *args, **kwargs):
    # guess list method
    dataset = kwargs.get('dataset', 'es')
    # guess method to check names dictionary
    guess = ''
    name = unidecode.unidecode(name).title()
    name.replace(name, "")
    dicc = self.name_frec(name, dataset)
    m = int(dicc['males'])
    f = int(dicc['females'])
    if ((m == 0) and (f == 0)):
        if binary:
            guess = 2
```

```
        else:
            guess = "unknown"
    elif (m > f):
        if binary:
            guess = 1
        else:
            guess = "male"
    elif (f > m):
        if binary:
            guess = 0
        else:
            guess = "female"
    else:
        if binary:
            guess = 2
        else:
            guess = "unknown"
    return guess
```

We are using the spanish dataset by default and the rest is the same idea that in the last script: more people using the name.



6.5 Deciding for males and females in images

There are many free software tools for decide gender in images files. We can use these tools to decide gender about images from Twitter, Github, ... We have selected the next tool:

```
$ git clone https://github.com/davidam/damefaces
$ cd damefaces/bin
$ python3 damefaces.py girl1.jpg
```

6.6 Webscraping and Damegender (counting scholars)

Sometimes, we can reach the database of names from a website, for example, we can retrieve a list of academics from Spain thanks to webometrics and the next script:

```
from lxml import html
import requests

print("Introduce an url from webometrics, for example,
       https://www.webometrics.info/en/GoogleScholar/Spain")

import argparse

parser = argparse.ArgumentParser()
parser.add_argument("url", help="display the gender")
args = parser.parse_args()

page = requests.get(args.url)
tree = html.fromstring(page.content)

academics = tree.xpath('//tr/td/a/strong/text()')

print('Academics: %s' % academics)
```

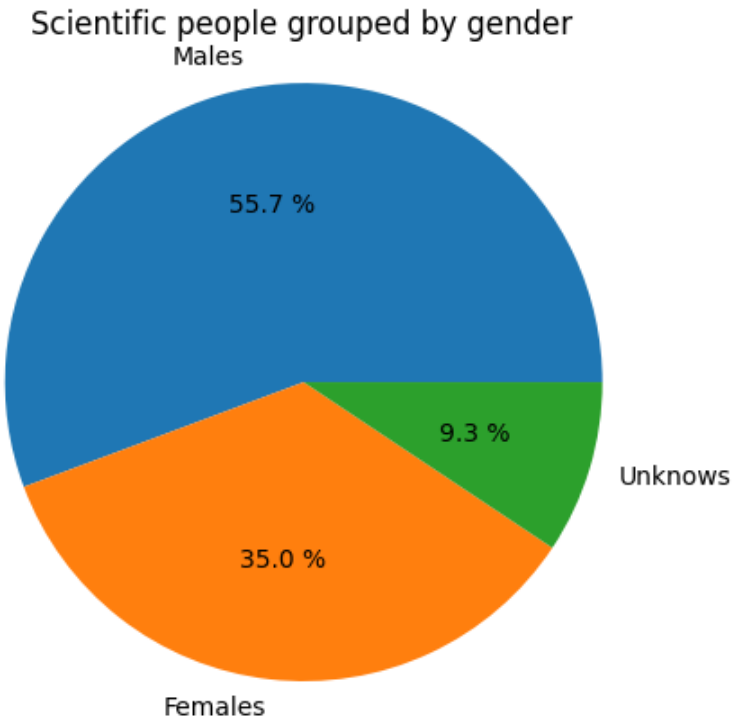
If you have retrieved the list of names in a file `files/scientifics.txt`, you could count males and females with the next script called `count-scientifics.py`:

```
import csv
import unicodedata
import unidecode
import re

from pprint import pprint
from app.dame_gender import Gender
from app.dame_utils import DameUtils
from ast import literal_eval
from app.dame_sexmachine import DameSexmachine

du = DameUtils()
g = Gender()
s = DameSexmachine()
```

```
with open('files/scientifics.txt') as f:
    mainlist = [list(literal_eval(line)) for line in f]

l = mainlist[0]

ll = []
for i in l:
    ll.append(i.split())


ten = ll[0:10]
hundred = ll[0:100]
thousand = ll[0:1000]

x = 0
y = 0
males = 0
females = 0
for j in hundred:
    if (len(j[0]) == 1):
        x = x + 1
    else:
        sex = g.guess(j[0], binary=False)
        y = y +1
        if (sex == "male"):
            males = males + 1
        elif (sex == "female"):
            females = females + 1

print("Number of scientifics with a single letter as first name: %s" % x)
print("Number of scientifics with the first name normal: %s" % y)
print("Number of females scientifics: %s" % females)
print("Number of males scientifics: %s" % males)

for j in thousand:
    if (len(j[0]) == 1):
        x = x + 1
    else:
        sex = g.guess(j[0], binary=False)
        y = y +1
        if (sex == "male"):
            males = males + 1
        else:
            females = females + 1

print("Number of females scientifics: %s" % females)
print("Number of males scientifics: %s" % males)
```

And the results are:

```
Number of females scientifics: 31425
Number of males scientifics: 47945
```



So, the percentage of academic people classified as females in Spain is bigger than Free Software people classified as females. Having a gender gap in both situations.

6.7 Counting males and females in a git repository

We can think a simple version of `git2gender.py`:

```
from app.dame_sexmachine import DameSexmachine
from app.dame_perceval import DamePerceval
from app.dame_utils import DameUtils
import sys
import argparse
parser = argparse.ArgumentParser()
parser.add_argument("url", help="Uniform Resource Link")
parser.add_argument('--directory')
parser.add_argument('--version', action='version', version='0.1')
args = parser.parse_args()
if (len(sys.argv) > 1):
```

```
ds = DameSexmachine()
du = DameUtils()
dp = DamePerceval()
l1 = dp.list_committers(args.url, args.directory)
l2 = du.delete_duplicated(l1)
l3 = du.clean_list(l2)

females = 0
males = 0
unknowns = 0
for g in l3:
    sm = ds.guess(g, binary=True)
    if (sm == 0):
        females = females + 1
    elif (sm == 1):
        males = males + 1
    else:
        unknowns = unknowns + 1

print("The number of males sending commits is %s" % males)
print("The number of females sending commits is %s" % females)
```

Try to execute this script:

```
$ python3 git2gender.py https://github.com/davidam/davidam.git
--directory="/tmp/clonedir"
The number of males sending commits is 3
The number of females sending commits is 0
```

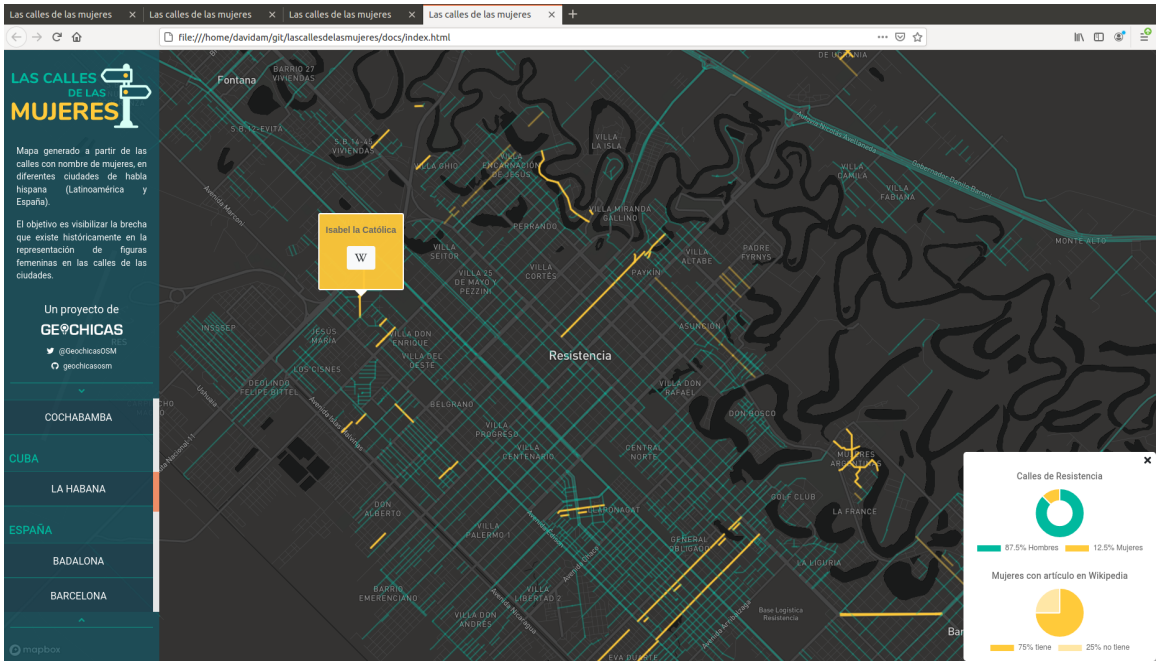
6.8 Counting males and females in Maps

Las calles de las mujeres is a project in NodeJS to display web streets with names about females using MapBox. You can download the project with:

```
$ git clone https://github.com/geochicasosm/lascallesdelasmujeres
```

It's licensed with a Creative Commons License (CC-BY-SA)

`https://creativecommons.org/licenses/by-sa/4.0/`.



It's giving statisticals about how many men and women has wikipedia article, too.

It has a strong community created by females (GeoChicas) `https://geochicas.github.io/` related with OpenStreetMap.

6.9 Gender gap in science

A good visual job can be found in `https://lukeholman.github.io/genderGap/`. This work is based on R retrieves the data from genderize and arxiv. The gap is especially large in authorship positions associated with seniority, and prestigious journals have fewer women authors. Additionally, they estimates that men are invited by journals to submit papers at approximately double the rate of women. Wealthy countries, notably Japan, Germany, and Switzerland, had fewer women authors than poorer ones. It concludes that the STEMM gender gap will not close without further reforms in education, mentoring, and academic publishing. There are a paper with a full explanation: "*The gender gap in science: How long until women are equally represented?*", [Further reading], page 42.

7 Secondary Sources about the Gender Gap

When a social researcher starts a new work, the first step is set an objective about the project with subobjectives, that's define the problem to solve with a methodology (quantitative, qualitative, or mixed). The second step is about sample decision who is the people, the population going to give us the data. The third step is about selection strategies about retrieve data, analysis and to show results. See "*Técnicas Cualitativas de Investigación Social*", [Further reading], page 42,

To read secondary sources about the objective and subobjectives of the project is to read previous works of another people (papers, books, data, news, ...) in science we refer to the state of the art or similar.

You can read secondary sources in different steps of a social research work, although is a task specially suggested in the first steps.

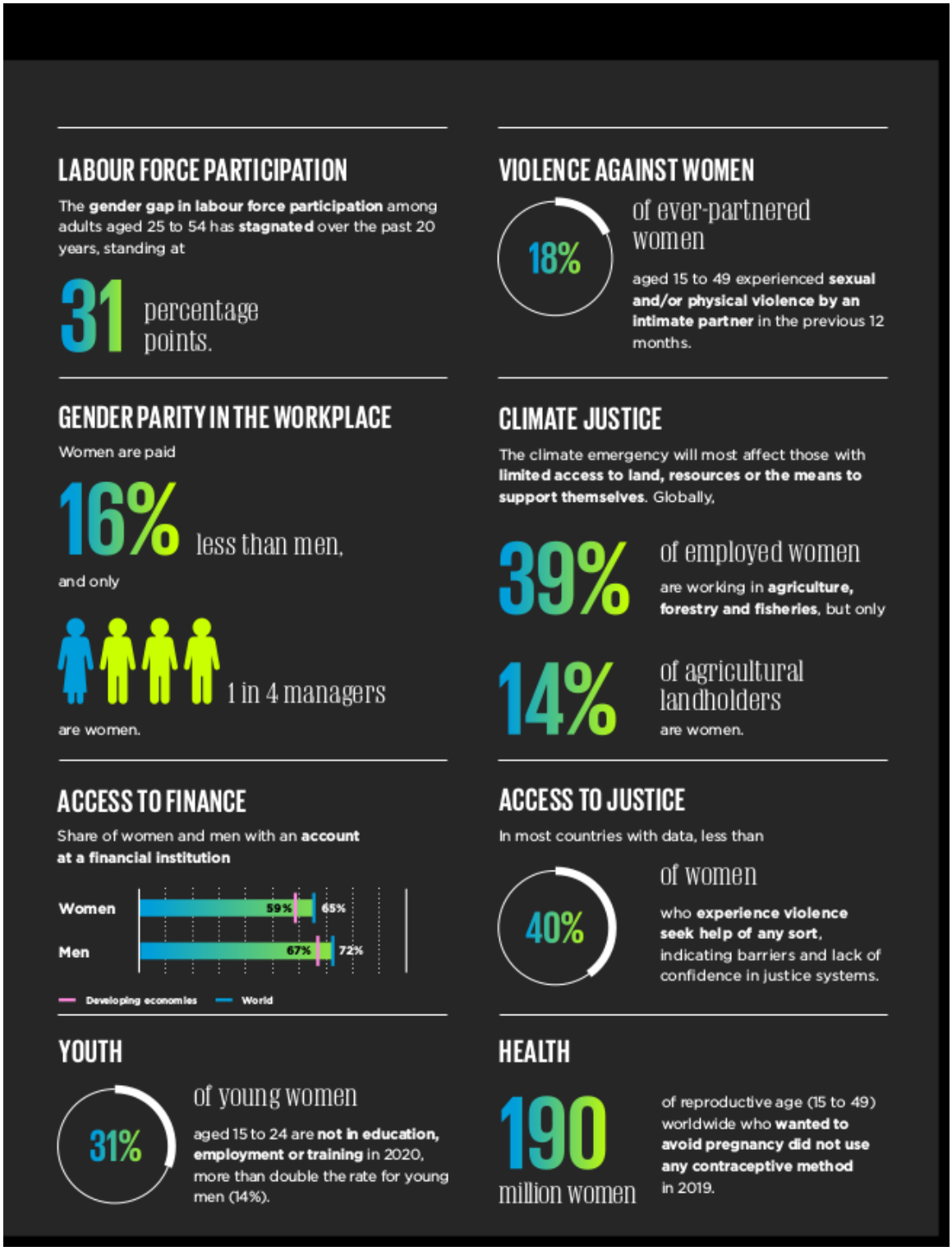
The objective of this chapter is to show some secondary sources about the gender gap in general, in STEM and in the Free Software. The idea is to reduce time to the people using Damegender understanding this kind of things. After of this, you can compare gender gap with the local problem: Kernel, Gnu/linux distribution, StackOverflow, Twitter, Forbes, Science, etc.

7.1 Gender Inequality in the World

Gender gap or gender inequality is the idea that men and women are not equal and that gender affects an individual's living experience. These differences arise from distinctions in biology, psychology, and cultural norms. Some of these types of distinctions are empirically grounded while others appear to be socially constructed. Studies show the different lived experience of genders across many domains including education, life expectancy, personality, interests, family life, careers, and political affiliations. Gender inequality is experienced differently across different cultures. (Source: wikipedia, 2020)

The women is underrepresent in the labour world (among adults aged from 25 to 54 has stagmated over the past 20 years, standing at 31 percentage points. The gender pay gap exists, too, so the women are paid 16% less than men. Share of women and men with an account at a financial institution is 65% of the total in women and 72% of the total in men. 31% of young women aged 15 to 24 are not in education, employment or training in 2020, more than double rate for young men (14%). Violence against women is 18% of ever-

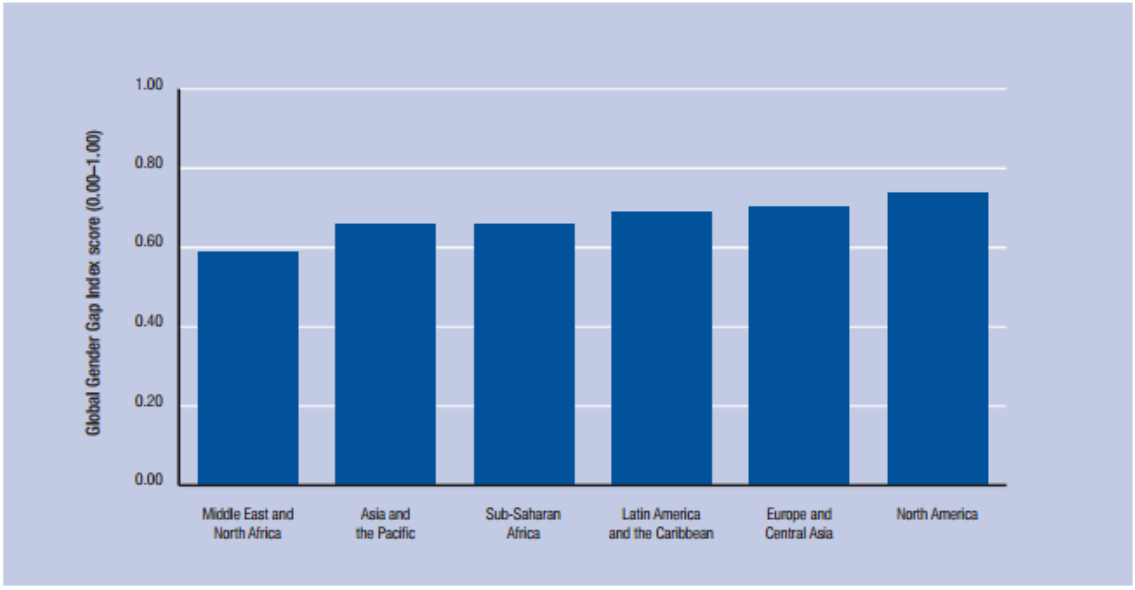
partnered women aged 15 to 49 experienced sexual and/or physical violence by an intimate partner in the previous 12 months.[1]



[1] https://www.unwomen.org

In the next image you can observe a graphic about gender gap in several continents (Source: Global Gender Gap Index 2012):

Figure 2: Regional performance on the Global Gender Gap Index 2012

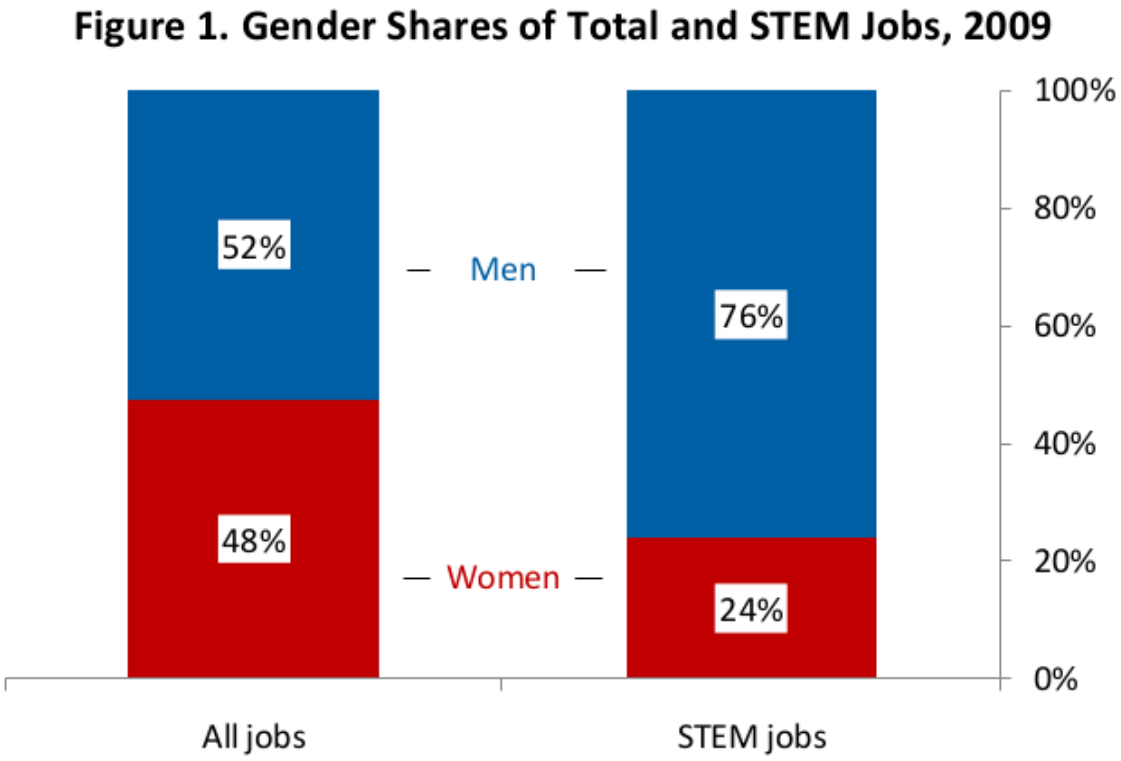


Source: Global Gender Gap Index 2012; details of regional classifications in Appendix B.
Scores are weighted by population; population data from the World Bank's *World Development Indicators (WDI)* online database 2011, accessed July 2012.

From the best score to the worst score. We can find: North America with the best score, Europe and Central Asia, Latin America and the Caribbean, Sub-Saharan Africa, Asia and the Pacific, and the worst score Middle East and South Africa.

7.2 Gender Inequality in STEM

In the graphic we can understand gender gap in stem in 2009 and compare with gender gap in the market:



So the gender gap in STEM is bigger than the labour market in general.

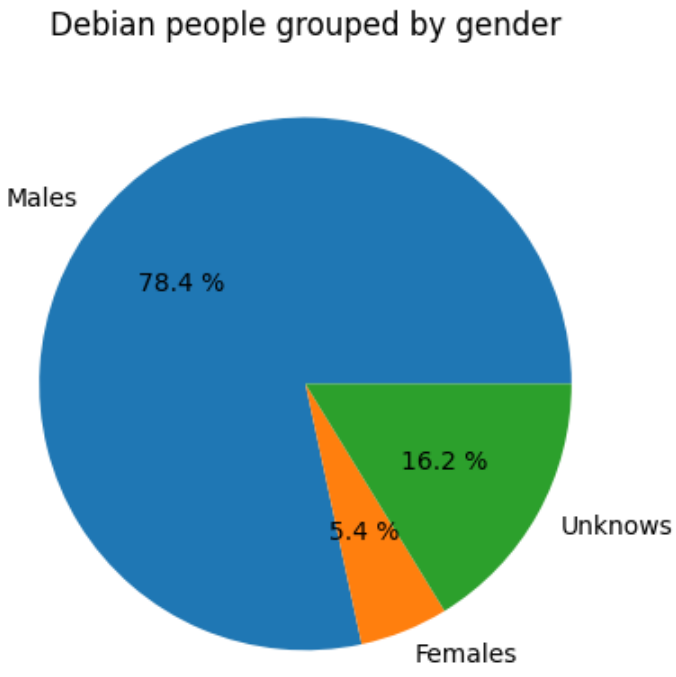
7.3 Gender Inequality in Free Software

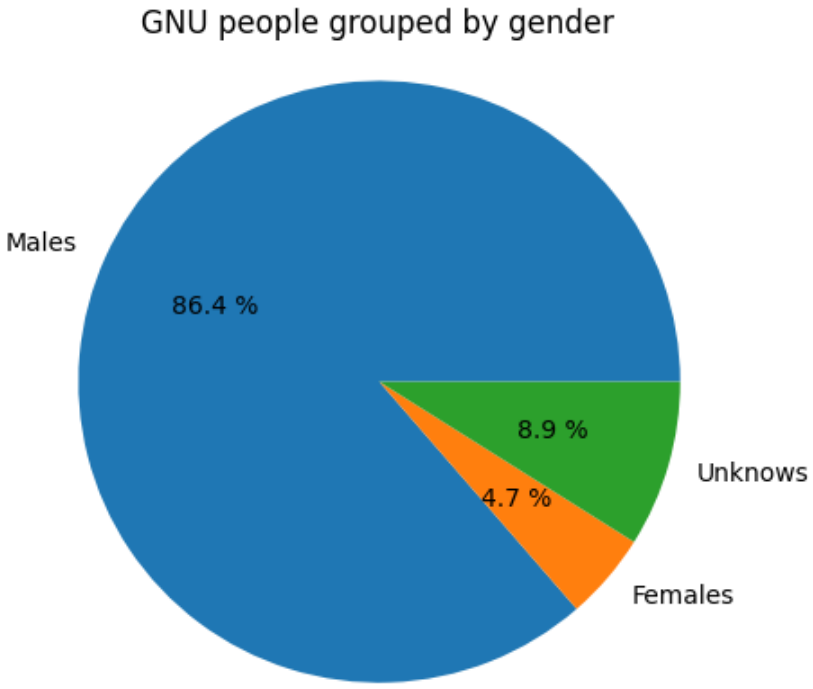
The gender gap in the Free Software world is so high we are presenting preliminar due to that we can reduce the percentage of unknowns to male or female. But we can observe that the gender gap is bigger in the Free Software world than in STEM.

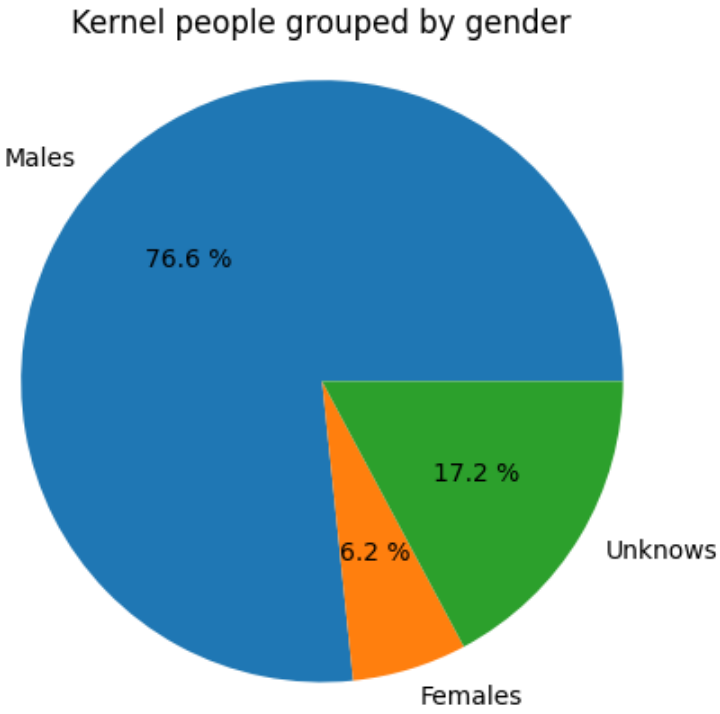
The context of the operatings systems is not feminist by different reasons. In the Free Software world there are many males developing the software. But in another Operating Systems there are another problems about the male domination, for example, Microsoft will create the most rich men in the world for many years[2]. Apple was classificated as the most valuable company in the world[3]. By intersectionality, to create companies more powerful

[2] https://www.forbes.com/profile/bill-gates/?sh=79472717689f
[3] https://forbes.es/listas/32245/apple-la-marca-mas-valiosa-del-2017/

than states is bad for the democracy and in a world where exists the gender gap the change to the gender equity is a pressure of values.




GNU people grouped by gender
Males
86.4 %
8.9 %
Unknows
4.7 %
Females


Kernel people grouped by gender
Males
76.6 %
17.2 %
6.2 %
Unknows
Females

8 Theorical Frameworks

If you want do a social research for a quantitative study, such as, count males and females you can:

- To generate objectives about the research study
- To read and to understand previous works.
- To choose some theorical framework.
- Perhaps, with a qualitative study (interviews, focus groups, ...) you could understand better the problems, the specific vocabulary used by the people, the reasons about the decisions, ...
- To retrieve data with Damegender or make surveys (online, offline, ...)
- An analysis quantitative with maths, graphics and interpretations
- Conclusions

A theoretical framework consists of concepts and, together with their definitions and reference to relevant scholarly literature, existing theory that is used for your particular study. In the last chapters you have learnt to count males and females, but you need give meaning to the words that you are using about your gender study. That is the point in this chapter.

We present some theorical frameworks that you can use as example in your works:

- Philosophies about software market and freedoms
- Interculturalism and Multiculturalism
- Feminism, Ecofeminism and derivatives
- Gender terms and philosophies

8.1 Philosophies about software, market, freedom and gender

There are different philosophies developing software and we are counting males and females in Internet, so the floor is the software in this world. If we must analyze gender in a country the ideology is changing in the place where you are. In the software world is the same problem. So, we are giving the vocabulary and the philosophy for speak about software and ideologies.

The propietary software is the most common idea for the common people, operating systems such as Microsoft Windows or Mac OS. If you are using software with propietary licenses, the source files will be containing copyright notes such as:

```
# Copyright (C) 2020  David Arroyo Menéndez

# Author: David Arroyo Menéndez <davidam@gmail.com>
# Maintainer: David Arroyo Menéndez <davidam@gmail.com>

# All rights reserved
```

This idea is associated to big companies leading the market but any people can use this philosophy. The criticism appears with Richard Stallman about privacy and lack of freedom

to the academic people, or hackers (people who knows read and write software and they do it for his objectives or global objectives). I could to say the monopoly is too strong with this license and the current social inertia and now nobody can change the market, we need another licenses to preserve the free market with an ethical strategy for startups and students.

Richard Stallman defines the Free Software with four freedoms: (0) to run the program, (1) to study and change the program in source code form, (2) to redistribute exact copies, and (3) to distribute modified versions. See "*Free Software Free Society*", [Further reading], page 42,

This idea to build software as a social good and motivated by ethical values. The solution is to apply the GPL license and to request to GNU to include the software.

The copyright note in GNU would be similar to:

```
;; This software is free software: you can redistribute it and/or modify
;; it under the terms of the GNU General Public License as published by
;; the Free Software Foundation, either version 3 of the License, or
;; (at your option) any later version.

;; This software is distributed in the hope that it will be useful,
;; but WITHOUT ANY WARRANTY; without even the implied warranty of
;; MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
;; GNU General Public License for more details.

;; You should have received a copy of the GNU General Public License
;; along with GNU Emacs.  If not, see <https://www.gnu.org/licenses/>.
```

On opposition the Open Source movement believes in free licenses, but they thinks that the software is business and they want to develop Free Software by economy, so they prefer change the word Free Software by Open Source claiming their philosophy.

They redefines the Free Software Definition by the Open Source Definition[1].

```
1. Free Redistribution

The license shall not restrict any party from selling or giving away
the software as a component of an aggregate software distribution
containing programs from several different sources. The license shall
not require a royalty or other fee for such sale.

2. Source Code

The program must include source code, and must allow distribution in
source code as well as compiled form. Where some form of a product is
not distributed with source code, there must be a well-publicized
means of obtaining the source code for no more than a reasonable
reproduction cost, preferably downloading via the Internet without
charge. The source code must be the preferred form in which a
```

[1] https://opensource.org/osd

programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

3. Derived Works

The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

4. Integrity of The Author's Source Code

The license may restrict source-code from being distributed in modified form only if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.

5. No Discrimination Against Persons or Groups

The license must not discriminate against any person or group of persons.

6. No Discrimination Against Fields of Endeavor

The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

7. Distribution of License

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

8. License Must Not Be Specific to a Product

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

```
9. License Must Not Restrict Other Software

The license must not place restrictions on other software that is
distributed along with the licensed software. For example, the license
must not insist that all other programs distributed on the same medium
must be open-source software.

10. License Must Be Technology-Neutral

No provision of the license may be predicated on any individual
technology or style of interface.
```

In the point six, we find the conflict with the feminist theories due to the possitive discrimination is a good idea to reach gender equity.

The GNU philosophy has the same problem explained on a different way. Only the Free Software is a good idea, if the software is not Free Software, then it's Propietary Software (the bad idea to avoid).

In Damegender, we want to deliver Free Software by practical reasons released with GPLv3 trough pypi.org and github.com the very popular sites to distribute Free Software written in Python. But we understand that we can to make positive discrimination in the development in favor to the women as an experiment with this copyright note in the development branch:

```
# You can share, copy and modify this software if you are a woman or you
# are David Arroyo Menéndez and you include this note.
```

This book will be published with this license, too.

8.2 Multiculturalism, Interculturalism

The term multiculturalism has a range of meanings within the contexts of sociology, of political philosophy, and of colloquial use. In sociology and in everyday usage, it is a synonym for "ethnic pluralism", with the two terms often used interchangeably, for example, a cultural pluralism in which various ethnic groups collaborate and enter into a dialogue with one another without having to sacrifice their particular identities. It can describe a mixed ethnic community area where multiple cultural traditions exist (such as New York City or Trieste) or a single country within which they do (such as Switzerland, Belgium or Russia). Groups associated with an indigenous, aboriginal or autochthonous ethnic group and settler-descended ethnic groups are often the focus.

In reference to sociology, multiculturalism is the end-state of either a natural or artificial process (for example: legally-controlled immigration) and occurs on either a large national scale or on a smaller scale within a nation's communities. On a smaller scale this can occur artificially when a jurisdiction is established or expanded by amalgamating areas with two or more different cultures (e.g. French Canada and English Canada). On a large scale, it can occur as a result of either legal or illegal migration to and from different jurisdictions around the world (for example, Anglo-Saxon settlement of Britain by Angles, Saxons and Jutes in the 5th century or the colonization of the Americas by Europeans, Africans and Asians since the 16th century).

In reference to political science, multiculturalism can be defined as a state's capacity to effectively and efficiently deal with cultural plurality within its sovereign borders. Multiculturalism as a political philosophy involves ideologies and policies which vary widely. It has been described as a "salad bowl" and as a "cultural mosaic", in contrast to a "melting pot". (Source: wikipedia, 2020)

Interculturalism refers to support for cross-cultural dialogue and challenging self-segregation tendencies within cultures. Interculturalism involves moving beyond mere passive acceptance of a multicultural fact of multiple cultures effectively existing in a society and instead promotes dialogue and interaction between cultures.

Interculturalism has arisen in response to criticisms of existing policies of multiculturalism, such as criticisms that such policies had failed to create inclusion of different cultures within society, but instead have divided society by legitimizing segregated separate communities that have isolated themselves and accentuated their specificity. It is based on the recognition of both differences and similarities between cultures. It has addressed the risk of the creation of absolute relativism within postmodernity and in multiculturalism. (Source: wikipedia, 2020)

Aguado proposes these principles (See "*La Educación Intercultural: Concepto, Paradigmas, Realizaciones*", [Further reading], page 42.

1. Promote the respect by all cultures together and condemn the politics to change the culture of the people towards the culture dominant. (Borrelli y Essinger, 1989)

2. The intercultural education is relevant for any student, not only for the foreigners and minorities (Borrelli and Essinger, 1989)

3. The troubles created by the ethnic and cultural diversity of the society has many solutions, there not an only magic solution. The politics in education there are partials because we are in a global society (Galino, 1990).

4. It's based in the perception about to accept cultures in contact, it's near to the form of life of societies with a poor cultural context instead of societies with more rich, more structure and high social control.

5. We need develop a scheme of concepts with many cultures demonstrating in the education that the knowledge is the common property of all people (Walking, 1990).

So, interculturalism and multiculturalism are the same concept in many uses, both recognize the cultural diversity in the contexts where there are the diversity, but interculturalism is doing an emphasis in the enrichment of all cultures respecting the diversity.

Damegender understands has an international and intercultural perspective about guess the gender about the name in the sense that in many countries are existing many different cultures determining names, surnames with a gender. So, in Spain are living 4 so important cultures (no foreigners):

- Castillian (culture dominant)
- Catalan
- Basque
- Galician

These cultures has correlations with names and surnames.

8.3 Feminism, Ecofeminism and Intersectionality

Feminism is a range of social movements, political movements, and ideologies that aim to define and establish the political, economic, personal, and social equality of the sexes. Feminism incorporates the position that societies prioritize the male point of view, and that women are treated unjustly within those societies. Efforts to change that include fighting against gender stereotypes and establishing educational, professional, and interpersonal opportunities and outcomes for women that are equal to those for men. (Source: wikipedia, 2020).

Ecofeminism is a branch of feminism that sees environmentalism, and the relationship between women and the earth, as foundational to its analysis and practice. Ecofeminist thinkers draw on the concept of gender to analyse the relationships between humans and the natural world. The term was coined by the French writer Françoise d'Eaubonne in her book Le Féminisme ou la Mort (1974). Ecofeminist theory asserts a feminist perspective of Green politics that calls for an egalitarian, collaborative society in which there is no one dominant group. Today, there are several branches of ecofeminism, with varying approaches and analyses, including liberal ecofeminism, spiritual/cultural ecofeminism, and social/socialist ecofeminism (or materialist ecofeminism). Interpretations of ecofeminism and how it might be applied to social thought include ecofeminist art, social justice and political philosophy, religion, contemporary feminism, and poetry. (Source: wikipedia, 2020)

The goal 5 in United Nations in 2020 is "Achieve gender equality and empower all women and girls". (Source: United Nations website)

Damegender don't reduce the gender gap per se. It's a tool to measure gender gap in Internet. The data is the basis to do politics to reduce the gender gap. These data must be used in contexts helping to the women, such as, feminism asociations, political parties or trade unions with accomodation in favor to the gender equity, etc.

Damegender is giving more oportunities to the women to reduce the gender gap than another tools due to the license system.

Intersectionality is a theoretical framework for understanding how aspects of a person's social and political identities (e.g., gender, sex, race, class, sexuality, religion, disability, physical appearance, height, etc.) combine to create unique modes of discrimination and privilege. Intersectionality identifies advantages and disadvantages that are felt by people due to a combination of factors. For example, a black woman might face discrimination from a business that is not distinctly due to her race (because the business does not discriminate against black men) nor distinctly due to her gender (because the business does not discriminate against white women), but due to a unique combination of the two factors. (Source: wikipedia, 2020)

So, to understand discrimination, we must understand multiple factors. For example, the free software communities with the principles about no discrimination and share code seems a good place to advance in the rights of the women in the software world, but if the reality is a place dominated by men then we must look for another ideas. Intersectionality is about to find the best formula to advance in values understanding that the rights of the women depends of another values such as democracy, free speech, labor rights, ...

8.4 Gender

Gender is the range of characteristics pertaining to, and differentiating between, masculinity and femininity. Depending on the context, these characteristics may include biological sex, sex-based social structures (i.e., gender roles), or gender identity.

Most cultures use a gender binary, having two genders (boys/men and girls/women); those who exist outside these groups fall under the umbrella term non-binary or genderqueer. Some societies have specific genders besides "man" and "woman", such as the hijras of South Asia; these are often referred to as third genders (and fourth genders, etc.). (Source: wikipedia, 2020)

Transgender people have a gender identity or gender expression that differs from their sex assigned at birth. Some transgender people who desire medical assistance to transition from one sex to another identify as transsexual. Transgender, often shortened as trans, is also an umbrella term. In addition to including people whose gender identity is the opposite of their assigned sex (trans men and trans women), it may include people who are not exclusively masculine or feminine (people who are non-binary or genderqueer, including bigender, pangender, genderfluid, or agender). Other definitions of transgender also include people who belong to a third gender, or else conceptualize transgender people as a third gender. The term transgender may be defined very broadly to include cross-dressers. (Source: wikipedia, 2020)

In Damegender, we are applying binary ideology classifying people as male or female only due to that the free datasets provided by the states only supports this idea in the moment writing this book. But we respect the non binary philosophies due to this philosophies are describing a reality.

9 Conclusions

There are many options to count males and females in Internet, a good idea is to retrieve a dataset about males and females. Damegender is giving the most modern open datasets and it provides a good toolkit for many solutions:

- To count males and females in git repositories, mailing lists, csv files, ...
- To predict gender with machine learning if the name is not in the dataset
- To guess the country about the surname
- To understand how is used a name with different cultural regions
- To retrieve names from commercial apis
- To view relationships between names and races

These techniques can help to research or to visualize gender gap.

With this manual we have understood:

- How to use Damegender.
- How to compare different solutions with a scientific perspective, that is to manage mathematical vocabulary for to apply the concepts.
- How to apply this software to different use cases.
- How to find the main external resources about gender gap.
- To explain some philosophies for to give explanations to the data.

Further reading

La Máquina Reaccionaria by María Ávila (Published by Tirant Humanidades)

The objective of this book is to analyze two social strengths. First, the fight of the women changing her position in the society and the structure of the patriarchy. Second, the resistances to decrease and to stop this change in a clear, sistematic and aware form or in a silly and involuntary form.

Measuring the Gender Gap on the Internet by Bruce Bimber (Published by University of Texas Press)

This paper evaluates differences in men's and women's presence on the Internet, testing for the presence of gender-specific causes for different rates of Internet use. Methods. The paper presents new survey data collected by the author in 1996, 1998, and 1999 showing trends in Internet use, and presents regression models of Internet access and use. Results. Two statistically significant gender gaps exist on the Internet: in access and in use. The access gap is not the product of gender-specific factors, but is explained by socioeconomic and other differences between men and women. The use gap is the result of both socioeconomics and some combination of underlying gender-specific phenomena. Conclusions. Around one-half of the "digital divide" between men and women on the Internet is fundamentally gender related. Several possible causes may explain this phenomenon.

Women in STEM: A Gender Gap to Innovation by various authors (Published by Economics and Statistics Administration Issue Brief No. 04-11)

This executive summary very good referenced presents data about males and females in STEM. Very useful to compare data (for example, in this manual we have compared this data with males and females in the Free Software world).

Colapso. Capitalismo terminal. Transición Ecológica. Ecofascismo. by Carlos Taibo (Published by Catarata, ISBN 978-84-9097-203-8)

This book is a good explanation about politics chances related with ecology, climatic change, etc.

Free Software Free Society by Richard Stallman (Published by GNU Press, ISBN 1-882114-98-1)

Richard Stallman is the best philosopher about Free Software being the father of the most used free software licenses and the founder of GNU project being developed by him and many other people. This book explains all ethical ideas that is inspiring the free software community called GNU.

Comparison and benchmark of name-to-gender inference services by Lucía Santamaría and Helena Mihaljevic (Published by PeerJ Journal)

This paper has inspired the development of Damegender and many ideas about name-to-gender software is implemented in this book. Thanks a lot by this job.

The Effect of Gender in the Publication Patterns in Mathematics by Helena Mihaljevic, Lucía Santamaría, Marco Tullney (Published by Plos One Journal)

A very good scientific job about gender gap in science. This paper is suggested to people who has learnt to use Damegender and they want develop a good scientific job.

Damegender: Writing and Comparing Gender Detection Tools by David Arroyo Menéndez (Published by EasyChair 2020).

This paper presents the scientific perspective about Damegender.

Perceval: software project data at your will by Santiago Dueñas, Valerio Consentino, Gregorio Robles, Jesús M. González Barahona (Published by ICSE 2018).

This paper presents Perceval a software to retrieve information from different sources. Damegender is using perceval to retrieve data from git in the moment to write this book.

The gender gap in science: How long until women are equally represented? by Luke Holman, Devi Stuart-Fox, Cindy E. Hauser.

Using the PubMed and arXiv databases, the paper estimated the gender of 36 million authors from >100 countries publishing in >6000 journals, covering most STEMM disciplines over the last 15 years, and made a web app allowing easy access to the data (https://lukeholman.github.io/genderGap/).

La Educación Intercultural: Concepto, Paradigmas, Realizaciones by Teresa Aguado Odina

This paper is a very good reference to understand Interculturalism focused on education. The vocabulary explained in this paper can be so useful to apply in concepts related with feminism, or social sciences in general.

Guía INTER: una guía práctica para aplicar la educación intercultural en la escuela by Teresa Aguado Odina, Inés Gil Jaurena y otras personas. (Published by Universidad Nacional de Educación a Distancia UNED).

This document allows to apply the interculturalism concepts in the schools.

Machine Learning by Tom M. Mitchell (Published by Mc Graw Hill, ISBN: 0-07-042807-7)

Tom Mitchell is a father of the Machine Learning. This book explains in a simpley way the main concepts understanding the need about the machine learning. So useful to learn algorithms.

Periodismo y Social Media: como estan usando Twitter los periodistas españoles by Clara Sainz de Baranda (Published by Estudios sobre el Mensaje Periodistico, UCM)

Understanding the Demographics of Twitter Users by Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. published by Icwsm

This paper could have been written with Damegender. In this paper the author calculates gender of twitter users using first names and race/ethnicity using last names.

Discriminating Gender on Twitter by John D. Burger and John Henderson and George Kim and Guido Zarrella

This paper explains classify gender in Twitter obtaining very good results with tweet texts, screen name, description and full name. The best classifier was obtaining the 92% of accuracy.

Galaxia Internet by Manuel Castells (Published by Plaza & Janés, ISBN: 978-8401341571)

This book explains Internet with the point of view of a big sociologist. So it is regarded as a good introduction to Social informatics. That is the study of information and communication tools in cultural or institutional contexts.

Global Gender Gap Index 2012 (Published by World Economic Forum)

The Global Gender Gap Report was first published in 2006 by the World Economic Forum. The 2020 report (published in 2019) covers 153 countries. The Global Gender Gap Index is an index designed to measure gender equality.

Has Feminism Changed Science by Londa Schiebinger (Published by The University of Chicago Press Journals)

Feminism and Science by Evelyn Fox Keller & Helen E. Longino (Published by Oxford University)

Teoría Feminista by Ana de Miguel & Celia Amoros (Published by Minerva Ediciones)

Constructing Grounded Theory by Kathy Charmaz (Published by SAGE)

Grounded Theory is the most important theory to apply qualitative research. You can learn to classify and to count discourses (for example, feminist discourses) and to reach conclusions with this book.

Técnicas Cualitativas de Investigación Social by Miguel Valles (Published by Síntesis Sociología)

A good first book about qualitativism for sociology, politics or social work students, it's giving theorical contents with practical examples done in Spain. If you want learn to put Damegender data into discourses this book can learn you to do it.

Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares by INE.es (https://www.ine.es/prensa/tich_2020.pdf)

This document explains differences in gender using Internet and computers in Spain.

Appendix A License

You can share, copy and modify this manual if you are a woman or you are David Arroyo Menéndez and you include this note.

The sources will be find in https://github.com/davidam/damegender/tree/master/manual

Index

A

Accuracy 9
APIs 2

B

Bibliography 42

C

Choosing components 14
Commands 4
Commands about Statistics 4
Conclusions 41
Configuring Api Keys 3
Confusion matrix 9
Counting features in names 13
Counting males and females in a git repository 24
Counting males and females in Debian 17
Counting males and females in Forbes 20
Counting males and females in Linux Kernel ... 19
Counting males and females in Maps 25

D

Damegender License 45
Debian 17
Deciding for males and females in images 22

E

Ecofeminism 34, 39
Error coded 9
Error coded without na 9
Error gender bias 9
Executing tests 4

F

F1 score 9
False negative 9
False positive 9
Feminism 34, 39
Forbes 17
Free Software 34
Further reading 42

G

Gender 27, 34, 40
Gender Detection Tools from the Name 2
Gender Gap 27
Gender gap in science 26
Gender Inequality in Free Software 30
Gender Inequality in STEM 30
Gender Inequality in the World 27
Git 17

I

Installation 3
Interculturalism 34, 37
Intersectionality 39
Introduction 2

K

Kernel 17

L

Las Calles de la Mujeres 17

M

Maps 17
Measuring success and error 9
Multiculturalism 34, 37

O

Open Source 34

P

Perceval 4
Precision 9
Principal Component Analysis (PCA) 9, 13
Prologue 1
Python Virtual Environment 3

Q

Qualitative Research 34

R

Recall 9
Regenerating files in post installation 4
Richard Stallman 34
ROC 9

S

Secondary Sources 27

Social Research 27, 34

Software Philosophies 34

T

Theorical Frameworks 34
True negative 9
True positive 9

W

Webscraping 17
Webscraping and Damegender
(counting scholars) 22