

# Damegender: Writing and Comparing Gender Detection Tools

David Arroyo Menéndez

June 10, 2019

Nowadays there are various APIs to detect gender from a name. In these slides, we offer a tool to use and compare these apis and a method to classify male and female applying machine learning and using a free license. The gender detection from a name is useful to make gender studies from social networks, mailing lists, software repositories, articles, etc.

# Download source and article to make a good tracing

- `git clone https://github.com/davidam/damegender.git`

# Social Need (I)

Traditional approaches for inferring the gender given a name are based on the use of census data and specific APIs that tend to use also census data. I propose to use ML techniques to complement them, so that it can be applied to **nicknames, new names, diminutives**, etc. that usually do not appear in census data. The underlying assumption that lead me to this approach is that we humans tend to have a certain intuition, that commonly works, to infer the gender of a name even if it is the first time we see it

## Social Need (II)

In this moment there are a **gender gap** between males and females in computer science and science in general (STEMM: Science, Technology, Engineering, Mathematics and Medicine). Create **free tools** and improve the current state of art allows measure and later create policies with facts to fix the situation.

# Underlying Technologies

- Scikit
- NLTK
- Numpy
- Matplotlib
- Perceval

```
$ python3 api2gender.py Laura --surname="Cornejo" --api=namsor  
female  
scale: 1.0
```

- Poor explicative power
- Peer review only with black box tests
- You need pay if you want compute many data (more expensive)
- You can compute from local massive data (fast)

```
$ python3 main.py David --total="ine"  
David gender is male  
363559 males for David from INE.es  
0 females for David from INE.es
```

- Precise explanation from a Statistical Institute
- You have the possibility to make peer review in a specific geographical area



# Selecting components with PCA

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.

# Give me informative features

```
$ python3 infofeatures.py  
Females with last letter a: 0.4705246078961601  
Males with last letter a: 0.048672566371681415  
Females with last letter consonant: 0.2735841767750908  
Males with last letter consonant: 0.6355328972681801  
Females with last letter vocal: 0.7262612995441552  
Males with last letter vocal: 0.3640823393612928
```

- A female distinguish feature is the last letter a.
- A male distinguish feature is the last letter consonant.

## Some accuracies

Way to guess a string	Accuracy
Namsor	0.7539570378745054
Genderize	0.715375918598078
Support Vector Machines	0.7049180327868853
Gender Guesser	0.6902204635387225
NLTK Bayes	0.6677501413227812
Gaussian Naive Bayes	0.5960994912379876
Multinomial Naive Bayes	0.5876201243640475
Stochastic Gradient Descendent	0.5873374788015828
Bernoulli Naive Bayes	0.5962408140192199

With Machine Learning we can guess nicknames, new names, or diminutives

# Proof of Concept in Repositories

```
$ python3 git2gender.py https://github.com/chaoss/grimoirelab-p
The number of males sending commits is 15
The number of females sending commits is 7
```

# Proof of Concept in Mailing Lists

```
# Count gender from a mailing list
$ cd files/mbox
$ wget -c http://mail-archives.apache.org/mod_mbox/httpd-announ
$ cd ..
$ python3 mail2gender.py http://mail-archives.apache.org/mod_m
The number of males sending mails is 6
The number of females sending mails is 0
```

We have presented this working in progress in one scientific event:

- <http://gregoriorobles.github.io/MadSESE/201906.html>

We have presented this working in progress in different industrial events:  
Python Barcelona, Open South Code and Medialab Prado.

- <https://www.youtube.com/watch?v=dvN0lMgQ9Pc>
- <https://www.opensouthcode.org/conferences/opensouthcode2019/program/proposals/183>
- <https://www.medialab-prado.es/actividades/procesamiento-de-lenguaje-natural-con-python-nltk>

Our idea is present an article finished this academic year.

Damegender is a tool to research in gender gap. So, the future work is to understand the massive gender gap with an empirical approach. The public mailing list and software repositories is a big public data source in this sense.

The market of gender detection tools is dominated by companies based on **payment services through APIs**. This market could be changed thanks to **free software tools and open data** due to give more explicative results for the user. Although the **machine learning** techniques is not new in this field, it's **an incentive for researchers** in computer science create free software tools.

These advances in computer science could be giving support to study the gender gap in repositories and mailing lists.