Damegender: Writing and Comparing Gender Detection Tools

David Arroyo Menéndez

June 1, 2019

Abstract

Nowadays there are various APIs to detect gender from a name. In this paper, 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 a make a good tracing

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

Social Need (I)

The value to detect the gender in a name using machine learning is related with new names don't registered in census as male or female. On situations using nicknames, new names, diminutives, ... the humans knows the gender in an intuitive way. Lingüistic features and statistics about male or female, countries, ... in a name could be interesting to decide give a name to a baby.

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 Scientific Technologies

- Scikit
- NLTK
- Numpy
- Matplotlib
- Perceval

Datasets

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

- Poor explicative power

Census Open Data

```
$ 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

Give me informative features

```
$ python3 infofeatures.py
Females with last letter a: 0.4705246078961601
Males with last letter a: 0.04867256637168141 5
Females with last letter consonant: 0.2735841767750908
Males with last letter consonant: 0.6355328972681801
Females with last letter vocal: 0.726261299544 1552
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	0.753957037874505 4
Genderize	0.715375918598078
Support Vector Machines	0.704918032786885 3
Gender Guesser	0.690220463538722 5
NLTK Bayes	0.667750141322781 2
Gaussian Naive Bayes	0.596099491237987 6
Multinomial Naive Bayes	0.587620124364047 5
Stochastic Gradient Descendent	0.587337478801582 8
Bernoulli Naive Bayes	0.596240814019219 9

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-annour
$ cd ..
$ python3 mail2gender.py http://mail-archives.apache.org/mod_mb
The number of males sending mails is 6
The number of females sending mails is 0
```

Conclusions

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.