Writing and Comparing Gender Detection Tools*

David Arroyo Menéndez[1][0000−0002−2986−5361] and Jesús González Barahona[2]

[1] Rey Juan Carlos University, Madrid, Spain
d.arrroyome@alumnos.urjc.es
http://www.davidam.com
[2] Rey Juan Carlos University, Madrid, Spain
jgb@gsyc.es
https://gsyc.urjc.es/jgb/

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, female and unknown 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.

Keywords: Gender gap · Gender detection tools · Software repositories.

1 Introduction

Santamaría and Mihaljević [6] compares and benchmarks five name-to-gender inference services by applying them to the classification of a test data set consisting of 7,076 manually labeled names. We are reproducing these experiments with our software using different apis and our own solution (DAMe Gender)

In this study we are using this dataset and performance metrics. We are giving a way to decide about male or female in undefined situations applying machine learning from some informative features. So, we are researching decisions about features and machine learning methods.

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.

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) [2]. Create free tools and improve the current state of art allows measure and later create policies with facts to fix the situation.

2 State of Art

We are reproducing Santamaría and Mihaljević [6] comparison bringing similar results. Generally, a good comercial solution is determined by a wide dataset. So,

* URJC - GSYC

the market is dominated by propietary solutions with money to invest in good datasets. Many names is determined by the geographic and cultural origin, it can be detected by surnames. Classify names and surnames in strings is detemined by good datasets, again.

3 Underlying Technologies

We have chosen Python free software tools with a good scientific impact. NLTK for Natural Language Processing [4]. Scikit for Machine Learning [5]. Numpy for Numerical Computation [7]. Matplotlib to visualize results [3]. And Perceval [1] to retrieve information in mailing lists and repositories.

4 Datasets and Census Open Data

Santamaría and Mihaljević [6] explains the different ways to create a dataset of 7000 labeled names. In summary, the names are retrieved from research articles and labeled by humans, checking in websites such as Wikipedia, or scientific web pages to decide if it's about a male, female or undefined.

Another approach is the census. The scientific value using Open Data is give a good explanation when we are asking about the gender from a name (number of males and females using a specific name in a country) versus a probability created by the way explained in Santamaría and Mihaljević [6] or similar.

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

We are using census Open Data from (Spain, USA and United Kingdom). That's a good point because the authors are acquainted with names in both languages.

A third approach is using a dataset from a popular free software solution. For instance, Natural Language Tool Kit (NLTK) is providing 8000 labeled english names. The classification is male or female. The problem again is about don't retrieve data with the social science quality of National Statistics Institutes.

We are using the census approach as base of truth because about of a name is male or female in a geographical area. Generally, a name has a strong weight to determine if it's a male or a female on this way. Although, if the census is not Open Data gender guesser dataset is a good dataset for international names.

5 Preliminary Results

5.1 Machine Learning

These results are experimental, we are improving the choosing of features and datasets. The datasets used in this experiment are INE.es and NLTK corpus names (this dataset is about english names). The features used are: first letter, last letter, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, vocals, consonants, first letter, first letter vocal, last letter vocal, last letter consonant, last letter a. We are improving the choosing of features with Principal Component Analysis. Take a look to the results:

ML Algorithm	**Accuracy**
Support Vector Machines	0.7049180327868853
Naive Bayes (nltk)	0.6677501413227812
Bernoulli Naive Bayes	0.5962408140192199
Gaussian Naive Bayes	0.5960994912379876
Multinomial Naive Bayes	0.5960994912379876
Stochastic Gradient Descendent	0.5873374788015828

Table 1. Machine learning algorithms accuracies

These results are demostrating that using Support Vector Machines english and spanish we are reaching results similar to another comercial solutions about gender detection tools. Our classifier is binary (only male and female)

We expect good results using gender guesser dataset and the Open Data census selected (Spain, USA and UK).

6 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.

7 References

References

1. Dueñas, S., Cosentino, V., Robles, G., Gonzalez-Barahona, J.M.: Perceval: Software project data at your will. In: Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings. pp. 1–4. ACM (2018)

2. Holman, L., Stuart-Fox, D., Hauser, C.E.: The gender gap in science: How long until women are equally represented? PLoS biology **16**(4), e2004956 (2018)
3. Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in science & engineering **9**(3), 90 (2007)
4. Loper, E., Bird, S.: Nltk: the natural language toolkit. arXiv preprint cs/0205028 (2002)
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
6. Santamaría, L., Mihaljević, H.: Comparison and benchmark of name-to-gender inference services. PeerJ Computer Science **4**, e156 (Jul 2018). https://doi.org/10.7717/peerj-cs.156, https://doi.org/10.7717/peerj-cs.156
7. Van Der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering **13**(2), 22 (2011)