Damegender: Writing and Comparing Gender Detection Tools

David Arroyo Menndez
GSYC
Universidad Rey Juan Carlos I
d.arroyome@alumnos.urjc.es

Jess Gonzlez Barahona
GSYC
Universidad Rey Juan Carlos I
jgb@gsyc.es

Abstract

The variable sex (male or female) is one of most used variables for any study in sociology, but this variable can be hidden in Internet communities. The gender detection from a name is an important problem in Natural Language Processing to decide if a string labeled as name is classified as male or female. An engineer will find useful make gender detection from a name retrieving information from social networks, mailing lists, instant messaging, software repositories, papers, etc. To achieve gender equality and empower all women and girls is a goal in sustanaible development in United Nations, so to measure the gender gap is a previous step to find solutions to reduce it.

Nowadays, there are several Application Programming Interfaces to guess gender from a name. This kind of software has the database based on propietary databases and the software is not free, so some scientific works are difficult to reproduce.

In this paper, we are envisioning how to solve these problems, offering a solution with a free license and open data names from official census useful to replace, use and/or compare these apis with very good results. This tool provides Machine Learning to predict strings, that's useful to guess diminutives or nicknames.

In: A. Editor, B. Coeditor (eds.): Proceedings of the XYZ Workshop, Location, Country, DD-MMM-YYYY, published at http://ceur-ws.org

1 Introduction

There are different ways to detect gender from a person name and perhaps a surname: census, wikipedia, self-references in trust websites, ... The most common way to detect gender from a name is the Application Programming Interfaces with a good popularity, for example, genderapi, namsor, genderize, ...

The problems addressed are:

- Evaluate quality/price with different commercial solutions.
- Think about solutions using free licenses.
- Treatment with names without census, for example, nicknames, diminutives, ...
- Massive gender detection from Internet, for example, mailing lists, software repositories, ...

In this paper, these problems are faced writing a Python solution for:

- To evaluate quality of different solutions applying metrics suggested by [**?**]
- To understand the current technology in detail, I have developed a tool guessing gender from a name giving support to Spanish and English from the open data census provides by the states.
- To fix the problem with nicknames and diminutives, we have developed a machine learning solution to strings not found in the census dataset.
- To do proof-of-concept tests applying Perceval to detect gender in mailing lists and software repositories.

2 First Level Heading

First level headings are all flush left, initial caps, bold and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

2.1 Second Level Heading

Second level headings must be flush left, initial caps, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

2.1.1 Third Level Heading

Third level headings must be flush left, initial caps and bold. One line space before the third level heading and 1/2 line space after the third level heading.

Fourth Level Heading

Fourth level headings must be flush left, initial caps and roman type. One line space before the fourth level heading and 1/2 line space after the fourth level heading.

2.2 Citations In Text

Citations within the text should indicate the author's last name and year[Knu73]. Reference style[Com79] should follow the style that you are used to using, as long as the citation style is consistent.

2.2.1 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page they appear on. Precede the footnote with a vertical rule of 2 inches (12 picas).

2.2.2 Figures

All artwork must be centered, neat, clean and legible. Do not use pencil or hand-drawn artwork. Figure number and caption always appear after the the figure. Place one line space before the figure, one line space before the figure caption and one line space after the figure caption. The figure caption is initial caps and each figure is numbered consecutively.

Make sure that the figure caption does not get separated from the figure. Leave extra white space at the bottom of the page to avoid splitting the figure and figure caption.

Figure 1 shows how to include a figure as encapsulated postscript. The source of the figure is in file `fig1.eps`.

Below is another figure using LaTeX commands.

[1]This is a sample footnote

Figure 1: Sample EPS figure

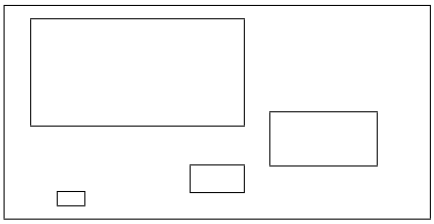

Figure 2: Sample Figure Caption

2.2.3 Tables

All tables must be centered, neat, clean and legible. Do not use pencil or hand-drawn tables. Table number and title always appear before the table.

One line space before the table title, one line space after the table title and one line space after the table. The table title must be initial caps and each table numbered consecutively.

Table 1: Sample Table

A	B	1
C	D	2
E	F	3

2.2.4 Handling References

Use a first level heading for the references. References follow the acknowledgements.

2.2.5 Acknowledgements

Use a third level heading for the acknowledgements. All acknowledgements go at the end of the paper.

References

D. Comer. The ubiquitous b-tree. *Computing Surveys*, 11(2):121–137, June 1979.

Santamara, Luca and Mihaljevi, Helena *Comparison and benchmark of name-to-gender inference services.* Addison-Wesley, 1973.