Big Data for gender analysis

Author: Laura Guerra Rivas
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.[*]

Advisors: Sònia Estradé, Gemma Viscasillas

Abstract: The objective of this work is to use python-based big data tools to perform a gender study of doctoral theses carried out at the Faculty of Physics of the University of Barcelona. The data used for the study is a total of 561 theses that have been obtained from the Dipòsit Digital of the University of Barcelona through Web Scraping and have been analyzed using Python.

I. INTRODUCTION

A transversal point of view is necessary to understand what surrounds us. Even though higher education for women is currently normalized in Spain, it was not until 1910 that women were able to access university legally [1]. The legalization of higher education for women did not imply a paradigm shift, especially in technical and scientific education. A good example of this is that in Spain, there was no female PhD in physics until 1929 [2].

According to a study carried out by GENERA (Gender Equality Network in Physics in the European Research Area), the percentage of women full professors in the area of physics in universities in the USA and in Europe stands at 16% [3]. Furthermore, this percentage is closely related to the field of study. A study carried out by RFEF shows that in Spain, in the field of optics, women make up 40% of the total PDI, while in areas such as theoretical physics and electromagnetism they are only 15%. In addition, despite the fact that the percentage of PDI women in Spain is around 22%, this percentage decreases to 14% in full professor positions [4].

The objective of this work is to use Python to obtain and analyse data to carry out a gender study on the theses directed in the Faculty of Physics of the University of Barcelona, and the circumstances that accompany them in order to observe the evolution of the presence of the female gender in the faculty of physics of UB and thus be able to observe if there is segregation due to gender.

II. METHODOLOGY

In order to carry out this study, it has been necessary to obtain the data on the doctorates that have been completed. The theses that have been completed at this faculty are collected in different databases. Although these data are publicly accessible, none allows free downloading. That is why the data had to be obtained using the Web Scraping technique.

A. Web Scraping

Manual data collection can be a very slow technique if the number of data to be analysed is relatively large. In addition, it often involves human errors. That is why new techniques have been developed to obtain data in a much faster and more efficient way. In this study, the methodology used has been Web Scraping.

Web scraping consists of programming a bot with the aim of automating the collection of data from a web page. This data collection is done from the source code of the page through Parsing [5].

The source code of the page contains all visible information and is written in the HTML programming language. The different sections of the web are divided into HTML classes that serve to organise the text and, within these classes, the information is separated into HTML tags, which format the text and allows to find the information [6]. The Parser is a function that subdivides these classes and tags and allows a search for the required data through its class and tag.

B. Data collection

The data has been obtained from the Dipòsit Digital of the University of Barcelona [7]. This database organises doctorates by departments. The theses of each department are distributed in different URLs and, on each of these pages, you can preview 20 different theses. One advantage of this database is that each department is assigned a URL prefix, meaning that for viewing the entire list, it's only necessary to add a multiple of 20 to the URL assigned to the department, thus accessing each page.

In this case, the bot has been programmed using two Python libraries. The Request library allows establishing a connection between the web server and the computer from the URL. On the other hand, to obtain the data from the web server, Beautiful Soup has been used, which allows searching the web information through its Parser. The procedure followed consists of two parts.

First, using a loop, a connection is made with each of the pages that contains the list of theses and, by means of the Parsing function of BeautifulSoup, the hyperlink associated with each of the theses registered in the Astronomy and Meteorology, Electronics, Elec-

[*] Electronic address: lguerrri13@alumnes.ub.edu

tronic and Biomedical Engineering, Applied Physics, Applied Physics and Electronics, Applied Physics and Optics, Fundamental Physics, Quantum Physics and Astrophysics, Physics of Condensed Matter, Physics of the Earth and the Cosmos, Electronics Department and Faculty of Physics categories is obtained. As with the filter, all theses have the same URL except for a reference number. This number is the one that has been registered to be able to access the information of each thesis.

Once the URLs belonging to each thesis have been obtained, Request has been used to make a connection with each of the pages and, with the use of the parser, the author, director or directors, year of defense, department and, finally, the title of the thesis are obtained.

C. Data treatment

A Python program has been made to automatically classify the gender of the authors and thesis directors from a list of the Instituto Nacional de Estadística (INE) where all the names that have been registered a minimum of 20 times in Spain separated by gender are collected [8]. This program separates the first name of the authors and/or directors and checks if it is on the INE list. One of the problems associated with this methodology is that it is not capable of classifying all the names obtained during scraping. Those names that appear only associated with one gender are automatically assigned to its gender while both the names categorised as "undetermined" and those that were not on the INE list have been classified manually, searching for information on the doctor or director.

Another problem associated with this data is that the same name could appear written in different ways, so that the program interprets it as two different people. To avoid this error, these names have been manually corrected.

Once all the names have been categorised, they have been analysed. Most of the analysis has been carried out by counting the gender of the authors and/or directors based on the previous classification.

A visual representation of the data obtained has been made in the form of a tree using Pydot, which is a library that allows the creation of custom nodes and their connection. To do it, the nodes of all the people who appear in the study have been created. After this, a python list has been created with the directors and the authors of the theses they have directed. To finish, the nodes of the directors have been connected with the corresponding authors using an arrow that goes from the director to the authors.

All the code used in this project is in a GitHub repository [11].

III. RESULTS

The study has been performed with a total of 561 doctoral theses from the Faculty of Physics of the University of Barcelona between the years 1974 and 2022 and obtained from the Dipòsit Digital of the University of Barcelona [7].

A. Time evolution

FIG. 1 shows the evolution of the number of doctoral theses performed annually at the Faculty of Physics of the University of Barcelona.

From 1974 to 1985 there is no registered thesis performed by a woman. In addition, from 1985 to 2001, women only represent 24% of the theses carried out and, in absolute numbers, the average number of theses executed by women is 0.9 thesis/year, while in the case of men this figure is triplicated.

From 2002 to 2012, the percentage of PhDs awarded to women increased to 28%. This value, despite being much lower than 50%, coincides with the percentage of female physics graduates during those years [9]. This means that the deficit of PhD women is closely related to the low presence of women in the Physics degree.

From 2012 to the present, there has been an absolute growth in the number of theses presented compared to previous decades, with an average of 16.8PhD/year presented by women, but maintaining a female presence of 28%.

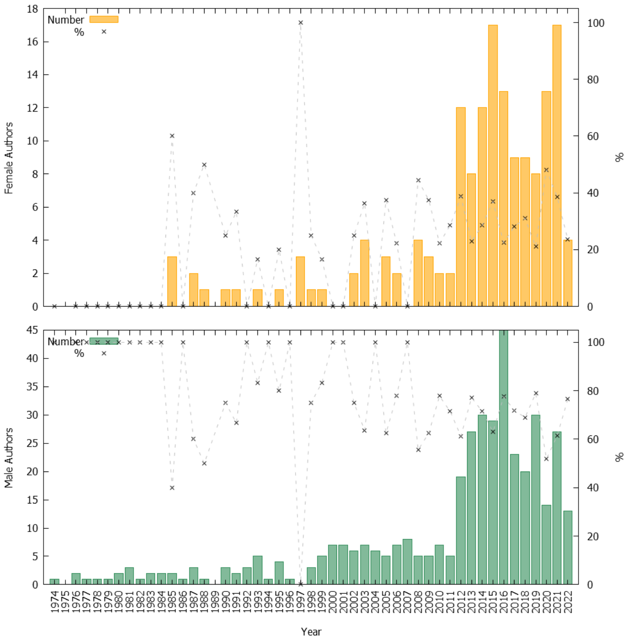

FIG. 1: Number and percentage of authors according to gender and year.

B. Data by department

If the gender of the PhD is studied according to the departments of the faculty, the results of Figure 2 are obtained. It should be noted that the names of the departments of the University have been changing in such a way that only those doctorates that are registered in the database with the current names have been represented.

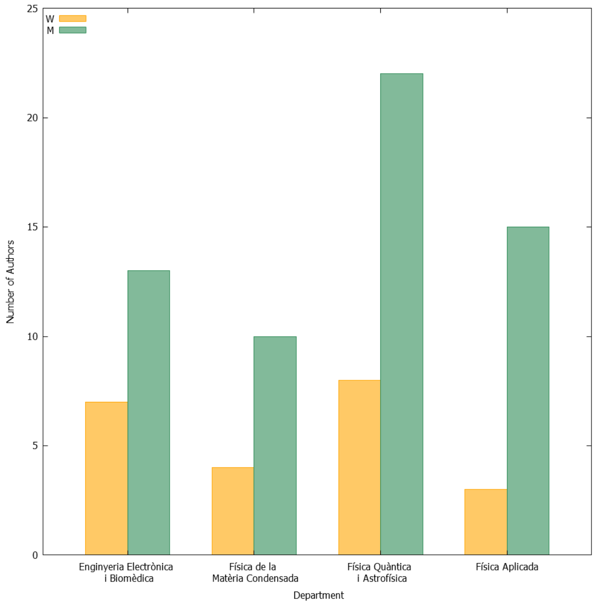

FIG. 2: Representation of the number of authors according to the department

Although in all the departments it is observed that the number of PhD of men is greater than that of women, the percentages vary in each department.

In three departments, the number of PhDs carried out by women exceeds 25% of the total number of theses presented. If these values are compared with the percentages of women and men that are part of the different departments [9, 10], it is observed that those departments with fewer women are those with more PhD awarded to women.

The Electronic and Biomedical Engineering department is the department with the fewest women (14%) and is the department with the highest percentage of women awarded PhDs (35%). In the case of Physics of Condensed Matter, the percentage of female staff (16%) is also lower than that of PhDs awarded to women (28%). In the department of Quantum Physics and Astrophysics, the percentage of PhD women (27%) is also higher than the percentage of women in the department. In contrast, the department of Applied Physics, which is the department with the highest female presence (23%), is the department with the lowest proportion of PhDs awarded to women, being only 17% of the total.

C. Doctoral Supervisors

For this study, in addition to making an analysis of the gender of the authors of the theses, a study has also been carried out on the supervisors of the theses to see if there are discrepancies and/or relationships between the genders beyond those found so far.

In FIG. 3 it is observed that the total percentage of theses supervised by women is much lower than that of men. Even so, this number agrees with the percentage of female supervisors and this indicates that the difference in theses supervised by one gender and another is only due to the fact that there are a greater number of male supervisors than female supervisors.

In relation to co-supervisions, it is observed that percentages varies with respect to those mentioned above. The co-directions with at least one woman amount to 20% and in the case of men this percentage is 80%.

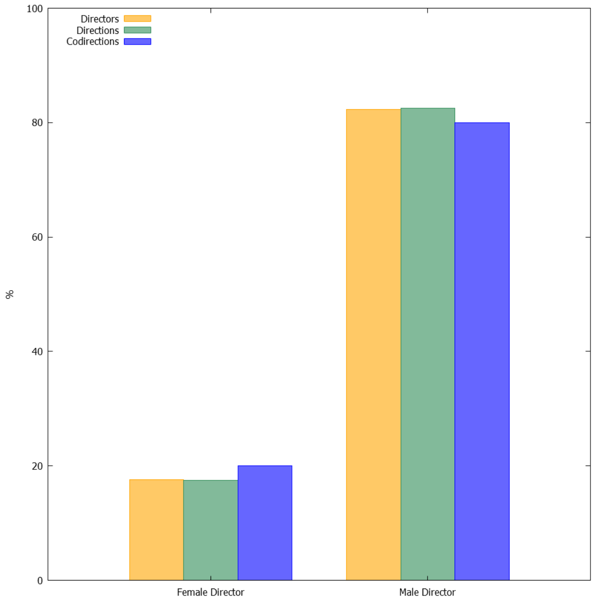

FIG. 3: Percentage of supervisors and supervisions and percentage of co-directions in relation to the total number of co-supervised directions.

On the other hand, the number of male doctors who have subsequently supervised a thesis is also much higher than that of women (FIG. 4). Even so, this fact could occur because the number of male doctors is much greater than that of female doctors. When calculating the percentage of women who have done a thesis at the faculty and subsequently directed one, it is obtained that this value only represents 8%. This same figure for men amounts to 15%.

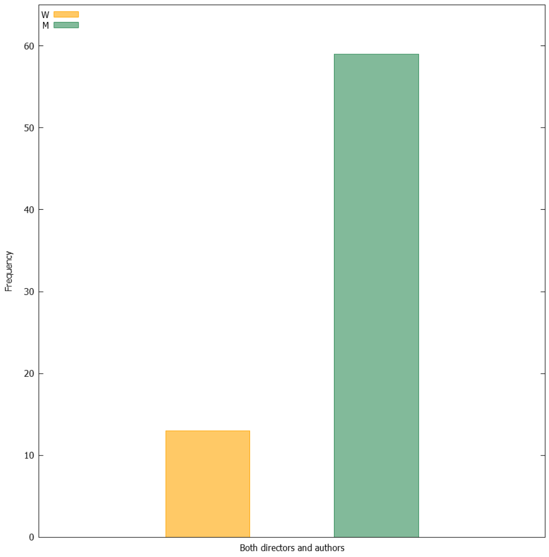

FIG. 4: Number of PhD women and men who are also directors.

It is observed that, in the data related to theses directed by a single person (FIG. 5), the percentage of female unique directors is 12% of the total, a value below the 17% of theses directed by women. On the other hand, as far as the gender of the author is concerned, the percentage of male supervisions with a male author is 87%, while in the case of female author this percentage is 78%.

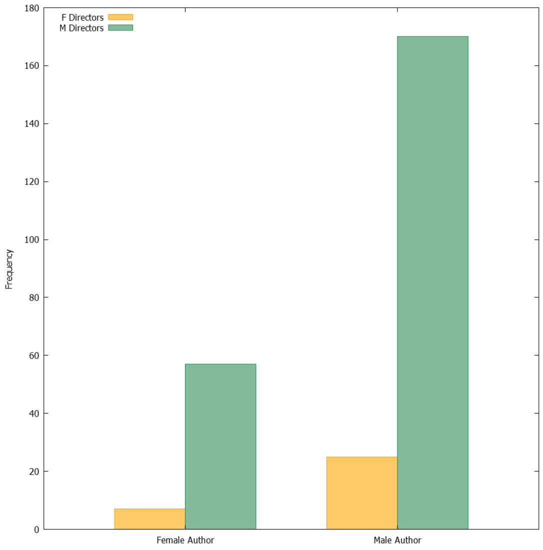

FIG. 5: Number of unique supervision of authors and authors according to the gender of the director.

If the number of co-supervisions of male and female authors is compared (FIG. 6), the proportion of theses co-supervised by women of the total theses supervised or co-directed by women represents 79% and in the case of men, this percentage is 68%.

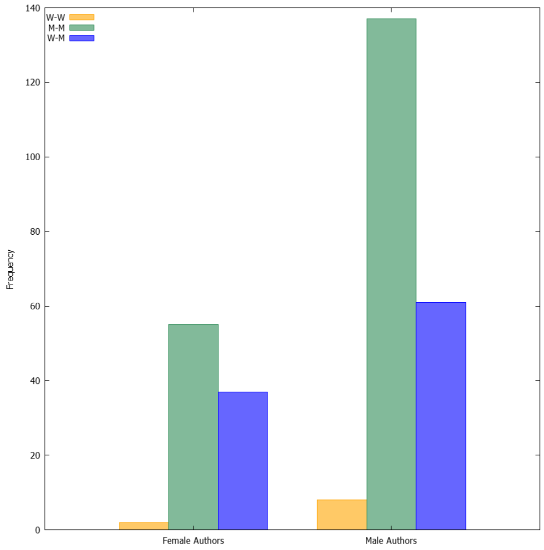

FIG. 6: Number of co-supervisions of authors according to the genders of the co-supervisors.

Regarding the type of co-supervision, in the case of the co-supervisions made up entirely of women, there are no significant differences in relation to the author's gender, in the case of female authors this percentage represents 3.9% while in the case of men represents 4.3%. On the other hand, in the co-directions formed only by men there is a greater difference according to the gender of the author. In the case of male authors, the percentage represents 69.6% of co-supervised theses, while in the case of female authors, this percentage decreases to 60.8%.

In the following QR (FIG. 7), all the authors and directors analysed in this work are shown in tree form. Each node corresponds to a person and has been coloured according to their gender. The nodes are joined from the directors of the theses to the authors.

FIG. 7: QR to access to the tree graphic form that contains the data used in the study [12].

IV. CONCLUSION

Web Scraping has proved to be a very good technique for obtaining data. Despite this, it is important to know the limitations, not only associated with this technique, but also with the sources of data collection that make the subsequent analysis complicated due to errors in the data. Therefore, it has been possible to create a program capable of obtaining all the data, classifying most of the names by gender and allowing subsequent data analysis.

This analysis shows that, despite the general increase in doctorates in the Physics Faculty of the UB, the evolution of the presence of women does not seem to be comparable to that of men. On the other hand, it has been seen that this doctorate bias (28%) is the same as that observed in physics graduate students [9], which indicates that this imbalance may be due to a lower female presence in the degree.

On the other hand, it has been observed that the departments with the fewest PDI women are the ones with the highest number of PhDs awarded to women and that the proportion of male PhDs who have subsequently supervised a thesis is double that in the case of female PhDs. women.

One thing to note is that the difference in the volume of works supervised by men and women seems to come only from a difference in the number of supervisors of each gender. In relation to the gender of the authors and directors, no very significant correlations have been seen, but it could be noted that the percentage of men who co-supervise works are 11% lower than in the case of women and the co-supervisions entirely carried out by two men are a 9% more frequent if the author is also a man.

Finally, mention that the number of data analysed is not very large. To obtain a more rigorous study, would be necessary a larger sample than the 561 data used for this study.

Acknowledgments

I would like to thank my advisors Sonia and Gemma for their support and guidance on this project. Lastly, thanks to my family and friends for their daily support.

[1] Díaz, Nieto and E S Caceres Marcela. “La Mujer en la Universidad en 1910.” (2010).

[2] Valdés, Juan Núñez, and Carmen Carbonell Coronado. ”100 años de derechos: la primera mujer española doctora en Física.” Investigaciones multidisciplinares en género: II Congreso Universitario Nacional” Investigación y Género”: Sevilla, 17 y 18 de junio de 2010. (2010).

[3] Skibba, R. ”Women in physics. Nature Reviews Physics”, 1(5), 298-300. (2019).

[4] RSEF. ”Las fisicas en cifras: Universidad”. (2020).

[5] vanden Broucke S. and Baesens B., ”The Web Speaks HTTP”, Practical Web Scraping for Data Science, pp. 25-48. (2018).

[6] vanden Broucke S. and Baesens B., ”Introduction”, Practical Web Scraping for Data Science, pp. 3-23. (2018).

[7] Dipòsit Universitat de Barcelona. `http://diposit.ub.edu/dspace/handle/2445/34657`

[8] INE, ”Apellidos y nombres más frecuentes”. (2022). `https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990.`

[9] Cantero B., Grau Torre-Marín V.and Viscasillas Valls G., ”Desequilibrio estacionario. La perpetuación del sesgo de género en Física”. (2022).

[10] Universitat de Barcelona, ”Departaments”. `https://www.ub.edu/dyn/cms/continguts_ca/universitat/campus_fac_dep/departaments/cercador_departaments.html`

[11] Guerra Rivas, Laura, ”Python code used for this project”. `https://github.com/Laura-Guerra/TFG`

[12] Guerra Rivas, Laura, ”Tree Graph”. `https://drive.google.com/file/d/1yilB1MCkTzYOx-MQNuTvOnGBvUEEUuH5/view?usp=sharing`