

Text Mining over Curriculum Vitae of Peruvian Professionals using Official Scientific Site DINA

Josimar Edinson Chire Saire

Institute of Mathematics and Computer Science (ICMC)
University of São Paulo (USP)
 São Carlos, SP, Brazil
 jecs89@usp.br

Honorio Apaza Alanoca

National University of Moquegua, Ilo
National University of Moquegua,
 Ilo, Moquegua, Peru
 honorio.apz@gmail.com

Abstract—During the last decade, Peruvian government started to invest and promote Science and Technology through Concytec(National Council of Science and Technology). Many programs are oriented to support research projects, expenses for paper presentation, organization of conferences/ events and more. Concytec created a National Directory of Researchers(DINA) where professionals can create and add curriculum vitae, Concytec can provide official title of Researcher following some criterion for the evaluation. The actual paper aims to conduct an exploratory analysis over the curriculum vitae of Peruvian Professionals using Data Mining Approach to understand Peruvian context.

Index Terms—Text Mining, Data Science, Peru, South America, Natural Language Processing, Curriculum Vitae, Research

I. INTRODUCTION

In the last decade, information and communication technology has innovated considerably, in the field of administration and selection of human resources in companies it has also evolved, Job applicants send their Curriculum Vitae (CV) through the Web or send them directly to a company. As an application, text mining is feasible for commercial use for the creation of knowledge, from the point of view of transformation of unstructured data [1], which assisted with a rational criterion of the human being. The e-procurement area is facing a growing number of these documents that are in different formats and contain a large amount of information [2]. The process of transforming unstructured candidate data into knowledge graphs has become a major challenge in machine learning [3]. A common problem in the academic field is to select professionals with a good research and development R&D profile.

Most of the time, identifying potential job candidates is an expensive and time-consuming task for human resources divisions [4]. In the academic field there is a need to select the best candidate, this problem is very common in Peruvian universities. In the research [5] an exploratory study of the information from Lattes, a specialized social network of researchers from Brazil, was carried out and presents a new approach to the analysis of regulated groups. The visualization components allow geographic exploration of collections, interpretation of the evolution of the topic [6], At present it is a very important component to visualize, analyze and

interpret the behavior of data that can express the inclination of research lines, languages, professions, among others. Recently, Artificial Intelligence has been successfully exploited and tools based on Data Mining, Multi-Agent Systems and Knowledge Representation Approaches (Ontologies) [7].

In context Peruvian in 2018, the investment reached 160 million soles and the amounts have advanced year after year [8]. The figure responds to the investment of Concytec and the National Fund for Scientific, Technological and Technological Innovation Development.

In Peru, an investigation of recommendation of resumes based on relevance of terms was carried out, natural language processing techniques and text mining were applied [9]. However, there is no study carried out exclusively in the academic field, while there is information about the researchers registered in the National Directory of Researchers (DINA). Therefore, the present investigation focuses on making an exploratory investigation and analysis of the data contained in curriculum vitae of the National Director of Researchers. In addition, it seeks that the results obtained can help to have a clearer picture of who evaluated the research and technological development in Peru.

II. PROPOSAL

The present paper is an exploratory study of Peruvian Professionals using available Curriculum Vitae, following the next steps:

- Selecting the scope
- Find the relevant terms to search
- Preprocessing
- Visualization

A. Selecting the scope

At the beginning, the motivation was related to study and understand Peruvian context about Research topic. Then, considering the existence of website gathering professional information in many countries, i.e. cv lattes(Brazil). The source of data is official website of Concytec.

B. Find the relevant terms to search

Considering the structure of website for each professional, html structure is analyzed to extract relevant data for the posterior phases. This step can use *div_class* or *html_tag*.

C. Preprocessing Data

The collected data has text format, then to make it readable for next step, the next steps are considered:

- Convert text to lowercase
- Remove non alphanumeric characters
- Remove stopwords
- Remove custom stopwords, i.e. experiencia(experience), inicio(start)

D. Visualization

The study is focusing on the analysis of Academic Information, Professional Experience, Scientific Publications and Languages. Then, many filtering/selections steps related are performed, i.e. bachiller(bachelor), maestria(masters) and so on. Besides, cloud of words are presenting to visualize the frequency of terms.

III. RESULTS

The next graphics are the result of the experiments on dataset. Subsection III-B is presenting Academic Information, subsection III-C presenting Professional Experience and Scientific Publications. Subsection III-D present information about languages, and inside of this part, III-D1 presents information about Peruvian Languages.

A. Description of dataset

- A collection of 25,000 registers were selected for this analysis.
- From this registers, only 14,504 has valid information.
- Fields: academic information(alphanumeric), professional experience(alphanumeric), scientific publications(alphanumeric), languages(alphanumeric)
- There are null values in some places, two options are possible: professional does not add any information or does not have any to add in one specific field.

B. Academical Information

Graphic 1 presents a cloud of words of the Academical Information, it is remarkable to see universidad nacional(national university), peru bachiller(peru bachelor). Then, first insight peruvian website has professionals who studied mainly in national universities. Peruvian national universities, are free of charge, this situation is different in other countries of South America, i.e. Chile.

An exploratory query to data is performed to know if there is more countries where professionals studied bachelor, master or phd degrees. Table I presents the distribution in South America, where people got his degree respectively. A previous affirmation related to Peruvian universities is confirmed, besides Brasil, Chile, Argentina are the top three in South American countries. At the other hand, considering the number of occurrences, it is possible people got his degree outside of South America. Checking, column 'Bachelor', 'Master', 'Phd', the next proportions are calculated: 54%, 77%, 28%. This last proportions indicates the potential of professional to work on Industry or Academy, most than 70% has a Master Degree and almost one third holds a Phd degree.

TABLE I
DISTRIBUTION OF COUNTRIES WHERE PROFESSIONAL GOT A DEGREE

Country	Ocurrences	Bachelor	Master	Phd
Peru	11795	7147	9940	3416
Argentina	101	48	89	49
Bolivia	12	2	11	3
Brasil	681	455	773	438
Chile	272	169	273	134
Colombia	58	17	59	22
Ecuador	29	8	28	10
Paraguay	7	5	5	2
Uruguay	4	0	4	3
Venezuela	23	12	23	15
Total	12982	7863	11205	4092

C. Professional Experience, Scientific Production

This subsection presents graphic Fig. 2 to present where people worked or works now. Left side, shows terms: 'universidad'(university), 'nacional'(national), 'actualidad'(actualidad), 'docente(docente)', 'instituto'(institute), those words indicates professional are working in Academic Centers(university, institute), working as teachers and most of them in national universities.

Right side of the graphic, present cloud of words about Scientific Production, terms are: 'journal', 'article', 'orcid', 'scopus', 'sevier'. These words are related to scientific publications(papers) in conference/journals with indexation Scopus, El Sevier. Besides, Orcid is a unique code to identify researcher and their contributions in personal website. Then, it is possible to add or import all relevant information to DINA website.

D. Languages

Considering results of subsection III-B, III-C, this subsection pretends to show how many languages or how is the level of languages of professional then this can be a way to export/show capacity through publications, talks, collaboration projects and more.

Most of the conferences/journals as English language as requirement, then it is necessary to write/read/speak it. Table III-D, can express the level of reading/speaking/writing using scale of basic/intermediate/advanced/superior advanced. Science, Technology are evolving everyday, therefore is necessary to read how this process happens, reading skill is vital 9,321 (64.26%). Write a publication/report or any scientific/industry document involves writing skill, 7,375 (50.84%), finally to communicate in conferences presentations speaking is necessary, 8270 (57.02%).

TABLE II
ENGLISH LEVEL

English Level	Reading	Speaking	Writing	Total_fil
Basic	2174	4120	3225	9519
Intermediate	4445	4194	4728	13367
Advanced	3722	2293	2664	8679
Superior Advanced	1154	888	878	2920
Total_col	11495	11495	11495	



TABLE VII
AYMARA LANGUAGE

Aymara Level	Reading	Speaking	Writing	Total_fil
Basic	31	23	40	94
Intermediate	29	26	30	85
Advanced	31	37	20	88
Superior Advanced	8	13	9	30
Total_col	99	99	99	

Phd. Most of them are working in National Universities and Institutes, then have the place to develop research groups, collaboration. Besides, has the access to Science through English, and Portuguese, French, Italian to open international projects, conferences and more. By other hand, Peruvian ancient languages requires preservation and can be the start of many studies, all this effort to integrate national community and foster studies from many fields, i.e. Social Sciences, Linguistics, Engineering and more.

IV. CONCLUSIONS

The Text Mining tasks involves many steps, from capture, cleaning, processing to visualization. It is important to notice that real data is not clean, follows a specific format or even uses the same language. Filtering data and organizing it, can help understand how is Peruvian situation and how professionals are ready to grow up and evolve in Industry/Academy. Most of them are working in National Academic Institutions, then it must be a good place to learn. collaborate and promote Science. At the end, languages are a key to connect with other countries, continents and create, organize international projects where researchers can collaborate/support each other and build a community to promote Science, Technology and progress to their countries.

V. CONSIDERATIONS

This lines are suggestions to improve your work:

- First, if you want to research and answer a question, maybe to solve a problem, you need data to answer and support the analysis.
- If data is not available, check alternative sources, i.e. you want to analyze government policies but there is no data, maybe it is possible to get data from other similar countries with policy of open data and it is possible to do a extrapolation or you can find social data to analyze the impact over the population. This step is key to develop next steps.
- Real data is not clean, clear for exploration then you are going to invest time, reading, testing ideas, cleaning until this data is ready, besides you can lose data, i.e. 20% of the dataset. This evaluation must be think and considered for next steps.
- Finally, maybe there is no specific data about your question, then you need to create artificial variables to understand the data, this involves much creativity and imagination. Remember, it is possible to cross dataset to get more meaningful data.

VI. FUTURE WORK

The authors explore Peruvian situation considering only 25000 curriculum vitae, the next step is to replicate the analysis over the entire existent registers. And, create a tool to support and foster collaboration projects between Peruvians and foreigners to catapult Science in Latin America.

ACKNOWLEDGMENTS

This final sections is to thank Concytec for the role of promoting Science, Technology in Peru. A country supported by Science and Technology can be a sustainable nation in the short future and influence in South America region. Finally, the authors want to mention Research4Tech, an Artificial Intelligence community of Latin American Researchers for promoting Science and collaboration in Latin American countries, his roles as integrator between Professional, Researchers, Technology communities is key to develop the Latin American region as a strong body.

REFERENCES

- [1] J. Dörre, P. Gerstl, and R. Seiffert, "Text mining: Finding nuggets in mountains of textual data," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 398–401. [Online]. Available: <https://doi.org/10.1145/312129.312299>
- [2] S. Amdouni and W. Ben abdessaïem Karaa, "Web-based recruiting," in *ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2010*, 2010, pp. 1–7.
- [3] J. Pujara and S. Singh, "Mining knowledge graphs from text," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ser. WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 789–790. [Online]. Available: <https://doi.org/10.1145/3159652.3162011>
- [4] T. C. Sandanayake, G. A. I. Limesha, T. S. S. Madhumali, W. P. I. Mihirani, and M. S. A. Peiris, "Automated cv analyzing and ranking tool to select candidates for job positions," in *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, ser. ICIT 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 13–18. [Online]. Available: <https://doi.org/10.1145/3301551.3301579>
- [5] T. L. V. de Santana and R. Santos, "Data science approach to analysis of lattes cv data," in *SIMBig*, 2017.
- [6] P. Accuosto, F. Ronzano, D. Ferrés, and H. Saggion, "Multi-level mining and visualization of scientific text collections: Exploring a bi-lingual scientific repository," in *Proceedings of the 6th International Workshop on Mining Scientific Publications*, ser. WOSP 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 9–16. [Online]. Available: <https://doi.org/10.1145/3127526.3127529>
- [7] T. C. Sandanayake, G. A. I. Limesha, T. S. S. Madhumali, W. P. I. Mihirani, and M. S. A. Peiris, "Automated cv analyzing and ranking tool to select candidates for job positions," in *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, ser. ICIT 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 13–18. [Online]. Available: <https://doi.org/10.1145/3301551.3301579>
- [8] S. Pichihua, "Concytec apunta a mejorar inversión en ciencia y tecnología para el 2020," 2020.
- [9] H. A. Alanoca, A. A. R. de Celis Vidal, and J. E. C. Saire, "Curriculum vitae recommendation based on text mining," 2020.