Data Science Approach to Analysis of Lattes CV Data

Thiago Luís Viana de Santana
Applied Computing Graduate Program
INPE – Nat. Inst. for Space Research
thiago.santana@inpe.br

Rafael Santos
Applied Computing Graduate Program
INPE – Nat. Inst. for Space Research
rafael.santos@inpe.br

Abstract

The Lattes Platform is an online database of academical records. It is used by the research and educational community of Brazil (and some other countries), being of great value for identification of researchers and their relationships with other researchers and, for that, it can be considered a specialized kind of social network.

In spite of its usefulness, the main interface of access to its data does not allow any type of analysis, just basic reports. In this paper, we present a new tool and approach to analysis of groups of records from the Lattes Platform which is simpler and more flexible than other similar tools proposed in the past.

1 Introduction

Evaluation of researchers and students productivity, either individually or in groups, is an important task for universities, research centers, and funding agencies, and also for the students and researchers themselves: students may be interested in knowing more about the achievements, research areas and experiences of prospective advisors. Teachers and admission deans may also want to know about the academic history of candidates to a graduate program, for example. The usual method to evaluate academics and students is by analysis of the achievements and publications listed on his or hers curriculum vitae (CV).

The Brazilian National Council of Technological and Scientific Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*, CNPq) maintains an online system, the Lattes Platform[1] (named in honor of César Lattes, a Brazilian physicist) that provides an unified interface to a database that is used to collect, store and process information about academic achievements. Any researcher can create and maintain his or hers own Lattes CV, using the taxonomy, formats and fields defined by the platform. The unification of some fields and categories makes it easy to fill the forms that feed the database.

The Lattes Platform is also used by CNPq to generate reports about the current status of the academic production of the researchers and students, and to evaluate applications to several different types of grants. The data on the platform is also used by other government funding agencies and by the Ministry of Education, for evaluation of the production of professors and students in graduate programs.

The Lattes Platform public interface is a web-based system that allows the edition of the CVs by its owner and the search and retrieval of the CVs by anyone that knows either the researchers' names or IDs (a 16-digit unique identifier). In the end of 2016 there were more than 3.500.000 CVs stored in the database[2]. Of those, almost 1.500.000 were CVs of students.

Data in the Lattes Platform can also be used for other academic purposes: analysis of academic indicators' evolution (Perez-Cervantes et al., 2012), identification of communities based on similar interests or collaborations (Mena-Chalco et al., 2014; Araújo et al., 2014; Alves et al., 2011b), changes of research areas based on the publications records, etc. Analysis considering groups of researchers or students can be done in different scales, from the whole academic community to small groups, such as researchers in a group or students in a college. But although the Lattes CV data is considered public (being provided by the

[1] http://lattes.cnpq.br

[2] http://estatico.cnpq.br/painelLattes/

researchers and students themselves), retrieval of the data is limited: it is possible to download a full individual CV as a XML file with all the data entered on that CV, but it is not possible to retrieve subsets of the data for more than one CV at a time. This makes it hard to perform some specific types of analysis that requires the extraction of certain categories from several CVs at once.

In this paper we present our work on tools and techniques that allow the extraction and analysis of data from collections of Lattes CVs. One characteristic of these techniques is that it considers that extraction of data from a set of Lattes CVs as a part of a *data science* process (Schutt and O'Neil, 2013): raw data (the Lattes CV XMLs) is collected, processed and cleaned; allowing an analyst to use exploratory data analysis techniques and apply statistical or other models on it. The analyst has access to the data and the tools to process it in a simple but flexible environment, therefore he or she isn't limited to a packed set of tools. Results of the analysis are presented in charts, plots, reports and data products, derived from the set of Lattes CVs considered on the analysis, and which may be used as input to other analysis tasks.

This paper is divided as follows: Section 2 presents related work, mainly on other tools used to extract information from Lattes CVs. Section 3 sets requirements for a Lattes CV exploration tool and proposes a data science based approach to build this tool. In section 4, the Lattes CV exploration tool is used and a few data Exploratory Data Analysis visualization examples are exhibited. Finally, section 5 enumerates problems outside the Exploratory Data Analysis framework that can be solved by extending the LattesLab tool.

2 Related Work

Access to subsets of data on Lattes CVs is desirable for several different types of analysis, and often these analyses must be done considering collections and not individual CVs. Considering this need, several different tools were created in the past to process and analyze data from the Lattes Platform.

One of the first tools that allowed the extraction of information from a collection of Lattes CVs is **scriptLattes** (Mena-Chalco et al., 2009). This tool allowed the extraction of data from the Lattes CVs of groups of researchers, creating reports, maps, graphs and other information from the collections of CVs. The tool could be deployed as a local application on computers running Linux, and its authors made the tool open so other groups could use it. Other research groups used scriptLattes as basis to create different types of analyses (Alves et al., 2016; Perez-Cervantes et al., 2012).

Initially scriptLattes downloaded the data from CNPq's servers as HTML files, but with later adoption of a CAPTCHA ("*Completely Automated Public Turing test to tell Computers and Humans Apart*") access control system, automatic download was made very difficult, so it was modified to use a local set of files that must be downloaded in advance.

scriptLattes is probably the most referenced tool in the bibliography we surveyed. It is quite complete, but the data is parsed from the HTML version of the Lattes CVs, which has changed in the past and may change in the future. Reports and graphics are also preprogrammed, so extensions and different layouts must be programmed separately.

LattesExtractor[3] is a tool developed by CNPq that allows the download of several Lattes XML CVs in batches. Although it seems to solve part of the problem at hand, namely, how to obtain subsets of the data, it is not as open or as flexible: only registered organizations can retrieve data with this tool, and organizations can only access data that is related to the organization itself. For example, an university may be able to download all the XML files with the CVs of its staff, teachers and students, but will not be able to download CVs of collaborators that don't currently work or study at the university. Similarly, when a student graduates and leaves the university, its CV will no longer be available after graduation (since he or she is not part of the university). This tool also does not perform any kind of analysis, providing only the XML files.

SUCUPIRA (Alves et al., 2011b) was developed as a tool that allowed both the semi-automatic extraction of the XML files from the Lattes Platform and the creation of reports and graphics that could answer questions about collaboration between researchers, their geographical location, their scientific production and its evolution, etc. SUCUPIRA is a web-based application that uses a list of names of researchers or students, managed by the system's user, to download the

[3] http://lattesextrator.cnpq.br/lattesextrator/

Lattes CVs (as HTML files), parse those and create reports based on the data extracted from the CVs on that list.

It seems that the development of that tool was discontinued, but changes on the Lattes Platform may have made it unusable: the structure of the HTML files changed, therefore parsers that could parse a version of the HTML generated by the platform had their usefulness restricted due to the new layout. Additionally, SUCUPIRA was written when access to the Lattes CVs was unhindered by CAPTCHAS – for some time the platform used simple CAPTCHAS that could be solved with tools such as Tesseract (Kay, 2007), but recently the CAPTCHAs were made more difficult to solve automatically.

SUCUPIRA used another tool developed by the same researchers: **LattesMiner** (Alves et al., 2011a). LattesMiner is a Domain-Specific Language (DSL) implemented as a set of Java classes that allows the manipulation of data on a set of Lattes CVs, defined by a programmer, with modules for data discovery (association of names and IDs), data extraction (parsing of the HTML files corresponding to the CVs with regular expressions), storage of data in a local database, visualization and analysis tools. As with SUCUPIRA, LattesMiner's development has stopped, since changes on the Lattes platform rendered some aspects of the tool unusable.

Another tool that is concerned with processing data from the Lattes Platform is **XMLattes** (Fernandes et al., 2011), which converts the Lattes CVs in HTML to XML for further processing, and which was rendered unnecessary since the present version of the Lattes platform already exports the XML version of the curriculum vitae (although requiring CAPTCHAs for download of individual CVs).

As part of this research we reviewed several papers related to analysis of Lattes CVs data (Digiampietri et al., 2012; Mena-Chalco et al., 2014; Araújo et al., 2014; Perez-Cervantes et al., 2012). Most of these papers used a database with detailed information on academics, which was extracted from the Lattes Platform when it was possible to do so without the limitations imposed by the CAPTCHA currently in use. There were no references on whether that database was kept up to date.

3 A Data Science-based Approach to Analysis of Lattes CV Data

Up to now, a number of Lattes CV exploring tools have been listed. Some of them cannot be used as designed for the following reasons:

- The Lattes CV platform has updated its data publishing technology from HTML to XML. Therefore, all the tools that relied on that previous file distribution (often requiring complex procedures to parse HTML) no longer work properly. It seems a trivial technical issue, but the tools that extracted information from the Lattes CVs formatted as HTML documents had to deal with complex HTML structures and had to detect, from the HTML content itself, the categories of information being extracted (e.g. articles in journals, conferences, titles, authors, etc.) while the data represented as XML is properly formatted and tagged with this information. Therefore, even though the XML platform restricted the use of the HTML-based tools, it provided structural elements for development of new, more robust tools.

- Some of these tools were designed to automatically download a list of CVs from CNPq's site. This is not possible today due to the implementation of the CAPTCHA test, both to view and download the Lattes CVs.

- Some of the solutions are only available on specific platforms - such as Linux - and require a specific setup before use.

Considering the present status of the existing tools for Lattes CV analysis we consider that a new, functioning tool to extract, interpret, analyze and visualize the data in a simple but flexible way ought to comply with the following requirements:

1. Work with an offline set of Lattes CVs. Due to the current impossibility of automated batch download of Lattes CVs, the CVs must either be obtained manually or automatically through other authorized tools – such as LattesExtrator.

2. Be able to transparently transform the list of Lattes CVs' XML files into table-like data structures for further processing. To make this transformation, it is necessary to know

the structure of the Lattes CV XML, identify the parameters of interest to the researcher and migrate these parameters to the data structures. This transformation is made easier since CNPq publishes the XML Schema Dictionary (XSD) of the Lattes CV files.

3. Be agnostic with respect to the operating system used to run the tool. Each potential user has a limited number of resources and to require the user to learn a new programming language, install a new software or even a new whole operating system shall be avoided if the tool is to be widely used.

4. Require no specialized knowledge for operation. In the same way that requiring an specific environment limits the utilization of the tool, if specialized knowledge is not required to operate this tool, it will be easier to use, and for that, may be used by a bigger number of individuals. At the same time the tool must be extensible so more advanced users could do more with the tool.

5. The results produced by this tool should be reproducible by any user interested in analysis of a set of Lattes CVs. Reproducibility is ensured by the use of a common set of instructions that can be easily shared and build upon.

The first three requirements on that list are strictly technical, and must be met by a Lattes analysis tool to deal with the complexity of the Lattes CVs' data access and representation issues. More important are the fourth and fifth requirements, that ensure that such a tool can be extended for different ways to explore the data and that the results can be reproduced and shared.

By designing a tool that follows these requirements it is possible to apply a data science process to the problem of analysis of collections of Lattes CVs. The data science process (Schutt and O'Neil, 2013) is shown in Figure 1.

The requirements listed above are directly related to the Data Science Process: the first requirement makes reference to the collection of data, the second requirement is related to the processing and cleaning of data as well as storage of the "clean" data, so that it can be easily and transparently accessed. Even though the fourth and fifth requirements are not directly associated to

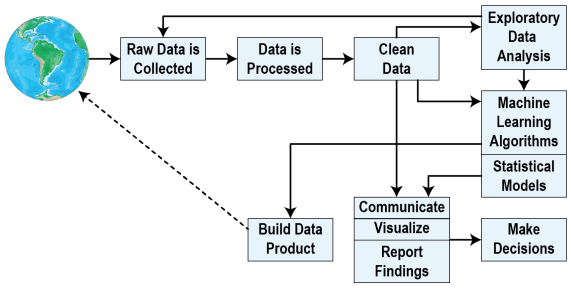

Figure 1: Data Science Process (Schutt and O'Neil, 2013).

the steps shown in the Data Science process (Figure 1), they serve as a guideline to ensure that the results attained by the tool are easy to acquire – and customize – and also reproducible.

The fourth requirement listed on this section is directly related to the concept of Exploratory Data Analysis (EDA), which intent is to allow the researcher to discover patterns on data by using visualization tools and statistics to understand "what is going on with this data".

Analysis of Lattes CVs can be done in different ways, using different metrics and algorithms, but if we consider that most of the analyses will be done considering thematic groups of researchers (e.g. researchers in a specific area of knowledge, or professors and students of a specific department) it becomes clear that methods and techniques applied to a particular analysis can be used into different contexts, depending on the group. A tool for analysis of Lattes CVs collections must make easy the reproduction of its results – reproducible research is also a concept closely linked to Data Science.

According to (Peng, 2011), there is a spectrum of reproducibility of research, that goes from a non-reproducible result to a fully reproducible one (Figure 2).

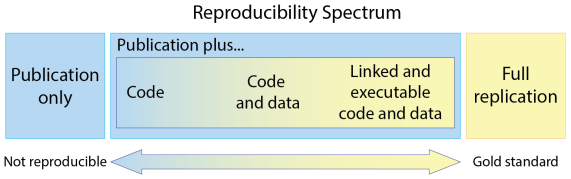

Figure 2: Reproducibility spectrum of a scientific publication (Peng, 2011).

Considering Figure 2, it is highly desirable that a tool performing Lattes CV Analysis allows the

full reproduction of the data analysis experiments with the corresponding code being publishable and applicable to different datasets of the same nature.

3.1 LattesLab

In order to address the requirements listed on the previous section, we propose a software stack solution, based on concepts and principles of Data Science, to tackle the generic problem of analyzing Lattes CVs. Our tool, named **LattesLab**, is based on the following components:

- A **library** that is able to scan a collection of Lattes CVs (stored as local files) and create a set of data frames (table-like structures) from that collection.
- A **deployment mechanism** for that library that allows its use, with minimal software installation requirements.
- A **set of live documents** that shows how to perform basic statistical analysis, visualizations and reports.

LattesLab is being developed in the Python language's (VanRossum and Drake, 2010), widely used platform – taking into account its large developer and user community. It is also one of the most used programming languages to solve Data Science problems due do the large amount of free and open analysis and visualization libraries.

Other languages were considered for implementation – a previous version was developed in Java, but we found out that the distribution of the library to use in other derived projects was too complex. R (Ihaka and Gentleman, 1996) was also considered: even though the R language is strongly supported by the community and is heavily used by the scientific community, Python was chosen for the readability of its code (over R code, at least) and its gradual learning curve.

To use LattesLab, it is necessary to have all the Lattes CV files of interest stored in a folder and pass the folder name as a variable to the LattesLab main library. Then, the tool reads the CV files as downloaded from the Lattes Platform, parsing the XMLs and storing the data available on data frames.

To extract data from the Lattes CV, XPath (part of XLST, a language for transforming XML documents) was used. One of the major advantages of the Lattes CV is its XML structure (which allows the extraction of semantic information from it) and the fact that the way its structure is posed is available in a public XML Schema Definition. By accessing the XML file through the XPath language one can extract the desired information and use it accordingly. In the presented case, the extracted information is used to generate a data frame which will the be used to perform a few basic analysis.

To obtain meaningful results from the analysis, it is desirable that the set of Lattes CVs share some characteristics. For example, to perform an analysis of the students and researchers of one institution, it is necessary to have the Lattes CVs of members of that institution stored and subject to analysis by LattesLab. Other sets of thematic collections could be researchers of all institutions sharing some CNPq classification (e.g. CNPq grantees), or researchers that stated that they work in a specific knowledge area.

The LattesLab main library is packed as a Python package, which simplifies its deployment. It can be downloaded and used in a standalone way, in this case the user must be able to at least install a Python IDE and then install and import the package.

Another deployment solution which is more straightforward is to deploy LattesLab as a Jupyter notebook (Kluyver et al., 2016). Jupyter notebooks are web applications that allow creation and sharing of documents that contain live code, equations, visualizations and explanatory text. The non-static parts of these documents are created by the execution of Python code.

Notebooks are an interesting solution not only to make access to the information easier, but also due to the fact that it runs in any operating system with a browser installed, and that it works not only with Python, but with R, Scala, Julia, and over 40 programming languages.

Jupyter notebooks allows not only the deployment of code, but, in the same document, formatted text (with different styles and allowing the use of hypertext, graphics, etc.) related to that code. Therefore, by using a Jupyter notebook, it is possible to run LattesLab code, to perform analyses and visualizations, to explain what was done and to give instructions to the users, so that he or she doesn't have to know the tool beforehand to use it and can change parameters or modify analysis to suit specific needs. Figure 3 shows a Jupyter

notebook running LattesLab.

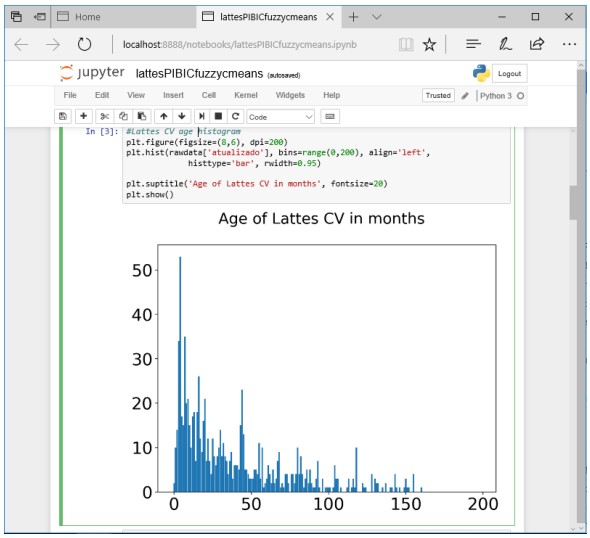

Figure 3: Example of a Jupyter Notebook window running the LattesLab tool.

When considering how to develop and deploy a tool for analysis of Lattes CV data we could opt for a standalone, GUI-based application. The reason to work on a tool with interactive lines of code is due to the flexibility that such a tool provides. Consider a simple analysis that requires the selection of a date range on a set of CVs: a GUI must provide a widget to allow the input of an initial and end year, which is quite simple to implement and use, and a programming approach would require one or two lines of code to implement the same functionality. But for more complex filters, e.g. to select a non-continuous range of dates, a GUI dialog would be more complex for the user (probably implemented as a list of checkboxes, one for each year) than one or two lines of code that filter a data set by a list of years.

A programming environment, while more complex, give more freedom to the user to implement filters, apply visual effects on graphics and use third-party tools, but the most important reason to avoid a GUI-based approach is the easiness of reproducibility: the chain of commands that provide an analysis from a dataset can be expressed in code, which can be documented and read by users, while GUI-based applications would require gestures (clicks, scrolls, inputs) that must be preserved somehow to allow reproduction of the analysis.

4 Examples of Analysis Reports

For the following examples, a group of 876 Lattes CVs was downloaded from the Lattes Platform. These CVs belong to participants in the Scientific Initiation grant program at the Brazilian Institute of Space Research (INPE) – these grants are given to undergraduate students to participate in research and development at the institute. The LattesLab library scanned and parsed these CVs' files to create a data frame used in the analysis and examples in this section.

Lattes CVs are created and updated by the users, and some may stop updating theirs, specially if they are not involved in academic environments anymore. A basic but interesting question we could ask about our data is: how old are the CVs, i.e. what is the last time they were updated?

We used LattesLab to create a histogram of the age of the CVs (the "last updated" information is present on the CVs).

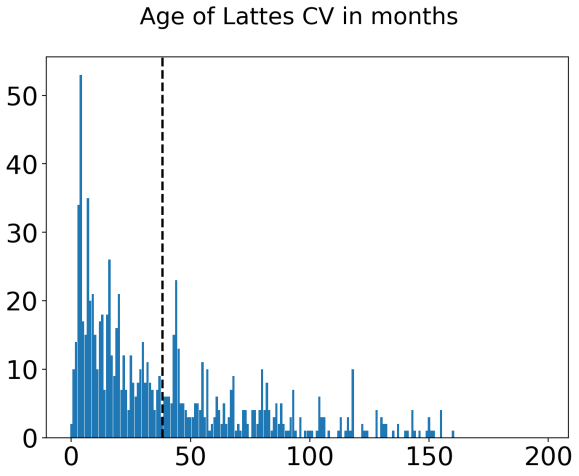

Figure 4: Age of Lattes CV in months.

In order to give an idea of the simplicity of using Jupyter notebooks and the LattesLab library, the plot in Figure 4 was created by three lines of code, once the data frame is created and loaded (to import data from the Lattes CVs and create the data frame, approximately two hundred lines of code were used, but these are not shown to users).

Scientific Initiation grants are given for a period of 12 months. It is possible for an undergraduate student to reapply for a grant, as long as he or she is enrolled in an undergraduate program. We knew that some of the students had held more than one grant, but wanted to get some statistics on it. A simple histogram was created, and it is shown

in Figure 5. Surprisingly, there were students that held grants for five and six years – that was unexpected since the average of the duration of undergraduate technical courses in Brazil is five years.

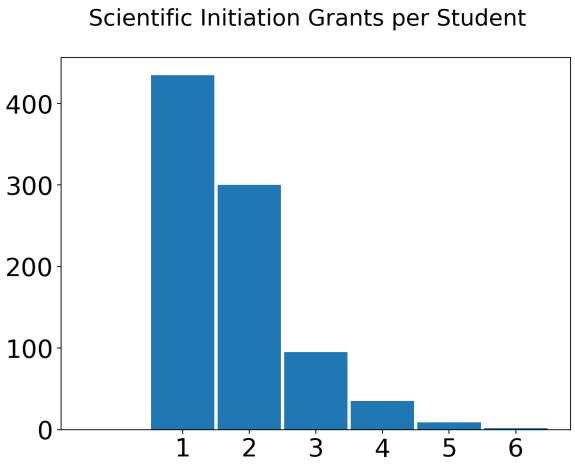

Figure 5: Number of Scientific Research Grants per Student.

The plot above was created with five lines of code in a Jupyter notebook.

With that thematic set of Lattes CVs data it is possible to analyze the academic achievements and get a glimpse on the careers of the students that held Scientific Initiation grants. How many of those decided on an academic career after their undergraduate studies? Figure 6 shows, of the Lattes CVs on our data set, how many individuals were part of different academic activities and achieved which academic degrees. An individual can be part of more than one graph bar, so individuals with Post-doctorate degrees are also found in the PhD degrees bar.

Many different types of visualizations can be easily achieved using LattesLab: Figure 7 shows how many of the students in our thematic data set obtained his/hers Masters' degree per year. Of course the interpretation of the results of these graphics is heavily dependent on the environment where the data has been collected: it could be possible to infer from Figure 7 that the number of degrees awarded is increasing over the years, or that there are in general more degrees awarded in even years, but the data itself does not answer why this is happening. It must be pointed out that we're only showing some examples of analysis/visualizations as examples of Exploratory Data Analysis that can be achieved with LattesLab.

Data frames generated with the LattesLab library also contain statistics on publications as declared in the Lattes CVs. Publications counts are stored by type and year of publication. One indicator of interest would be the evolution of the number of papers of a given researcher, or the entire researcher group, over the years.

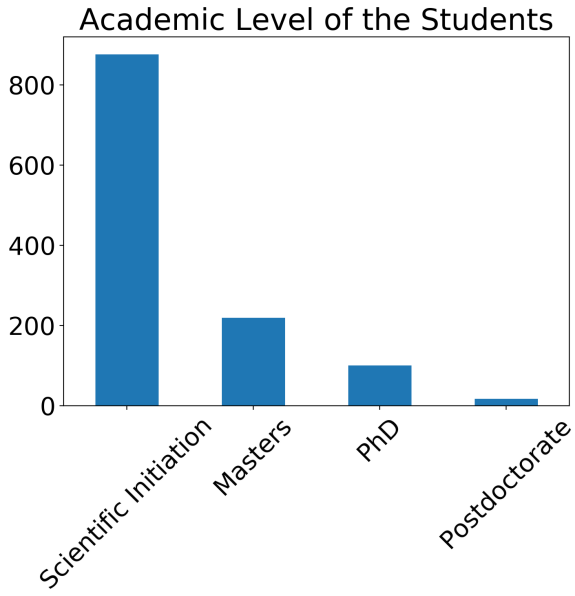

Figure 6: Academic Level of the Researchers.

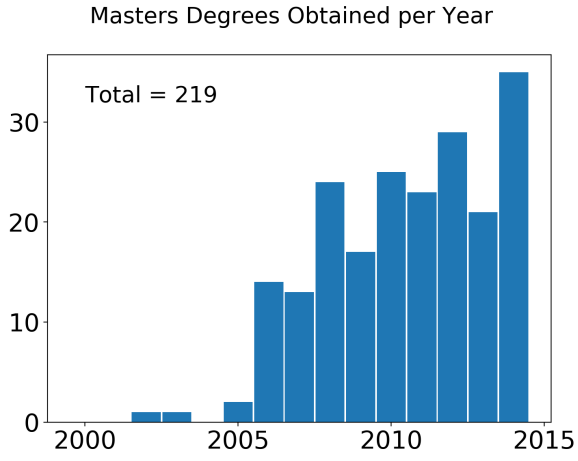

Figure 7: Master Degrees obtained per year.

We could consider that there are different profiles for Scientific Initiation grantees, depending on whether the grantee was able to publish his/her work some time after receiving the grant, and depending on the number of papers published per year.

In order to perform this type of analysis, a feature vector – that counted the number of papers published in conferences and journals each year after the student received his or hers first grant –

was used. This is one of the many possible ways to analyze individual and group publication indexes, and can be easily extracted from the data frame obtained from the LattesLab library.

We know that some Scientific Initiation grantees did not pursued further academic activities – it is to be expected that these students did not publish their results, or, if they did, did that one or two years after being awarded the grant. This is one profile we expect from our data and feature vector – are there others? What are the most interesting or unexpected profiles?

We used the Fuzzy C-Means clustering algorithm (Bezdek et al., 1984) to group the 876 profiles extracted from the grantees CVs into nine different groups (there are metrics that can be used to indicate the best number of groups for clustering a data set, but these metrics are sometimes conflicting and inconclusive (Morais et al., 2015)).

The centroids obtained from the profiles are shown in Figure 8. That figure shows some interesting patterns – one, already expected, shows that the grantees did not published any paper between being awarded the grant and 12 years after the award. Other patterns are also interesting: the clustering algorithm identified six patterns where the grantee did not publish on the first year (also expected), published one or more papers on the second year and fewer and fewer on subsequent years.

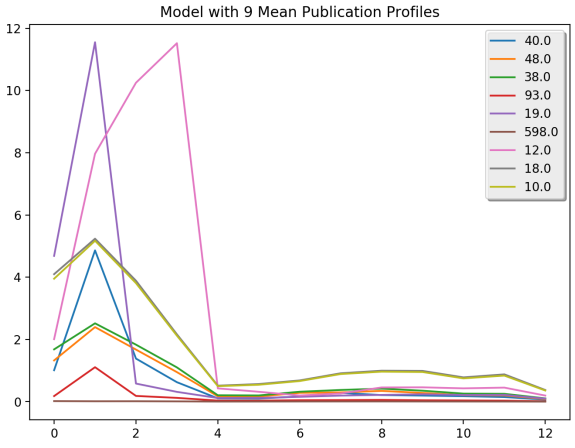

Figure 8: Model with all data classified in nine centroids.

Some of the patterns shown in Figure 8 indicates that the grantees kept publishing for years after being awarded the grant.

It is possible to see a large cluster corresponding to grantees that had no publications over the years consider in this example – the one pattern we know that that was in our data. To keep with the Exploratory Data Analysis approach we eliminated from the data frame (again, a simple operation since the Lattes data was converted to a data frame) those grantees which have never published, and ran the Fuzzy C-Means algorithm with eight clusters. Resulting centroids are shown in Figure 9.

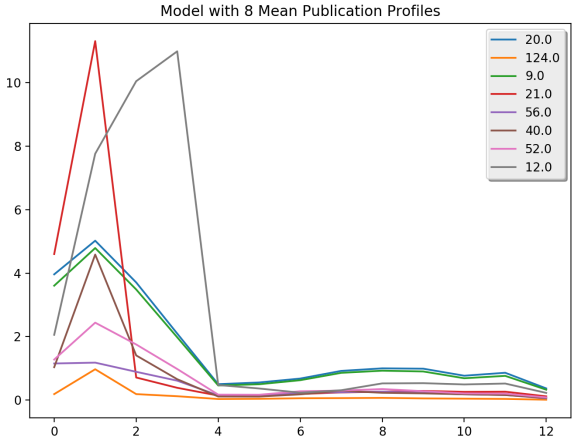

Figure 9: Model with all non-zero data classified in eight centroids.

As expected, profiles (centroids) in Figures 8 and 9 are very similar, indicating that the removal of the profile with zero publications did not change the clustering results much. Further analysis could be performed to better characterize the remaining profiles, or to investigate different profiles that may appear when we consider only publications after four years of being awarded the grant.

5 Conclusion and Future Work

As shown in the previous section, the LattesLab library has the capability to generate a data frame containing most of the metadata and quantitative data (counts for categories and years) of a local, thematic collection of Lattes CVs. Different types of analysis can be easily done when combining the library with other Python libraries in a standalone application or Jupyter notebook. Other interesting reports and analysis that can be done with this kind of data are:

- List CVs from the local thematic collection that must be downloaded again (based on the age of the CV).
- Cluster CVs by different criteria using different methods more suited to Exploratory

Data Analysis, such as the Kohonen Self-Organizing Networks for visualization (Morais et al., 2015).

- Generate a histogram of all the publications, per category and year, of the researchers on a specific group.
- Based on the previous task, create simulations that include or exclude certain members of the groups, to evaluate *what if* scenarios of researchers leaving groups or departments.
- Again based on the previous tasks, compare two subsets of Lattes CVs by yearly production, averaged by the number of researchers in each group, for evaluation of publications between departments or universities.

It must be pointed out that these are actual requests from some coordinators of graduate programs and head of departments that are acting as beta testers/evaluators of the tool.

Lattes CVs can be considered social networks in the sense that co-publications, co-orientations and participation in events as organizers or in committees can be extracted from the data, since coauthors are listed and sometimes identified by a unique ID used in the Lattes database. Future versions of the library will have the capability of extracting a co-occurrence matrix of IDs that identify categories and times of collaborations between members of a group. This could be used to explore the social network aspect of the Lattes CVs (to find cliques, temporal changes between groups, etc.)

Another improvement being considered is the creation of another type of data frame that represents the textual information associated with each researcher publication – papers titles, names of conferences, etc. This could be used to identify areas of interest and keywords through text mining techinques, making it possible to explore other ways to consider similarity between researchers.

References

Alexandre D Alves, Horacio H Yanasse, and Nei Y Soma. 2011a. Lattesminer: a multilingual dsl for information extraction from lattes platform. In *Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11*. ACM, pages 85–92.

Alexandre Donizeti Alves, Horacio Hideki Yanasse, and Nei Yoshihiro Soma. 2011b. Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*. IEEE, pages 1–6.

Wonder AL Alves, Saulo D Santos, and Pedro HT Schimit. 2016. Hierarchical clustering based on reports generated by scriptlattes. In *IFIP International Conference on Advances in Production Management Systems*. Springer, pages 28–35.

Eduardo B Araújo, André A Moreira, Vasco Furtado, Tarcisio HC Pequeno, and José S Andrade Jr. 2014. Collaboration networks from a large cv database: dynamics, topology and bonus impact. *PloS one* 9(3):e90537.

James C Bezdek, Robert Ehrlich, and William Full. 1984. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10(2-3):191–203.

L Digiampietri, J Mena-Chalco, J de Jésus Pérez-Alcázar, Esteban F Tuesta, K Delgado, and Rogério Mugnaini. 2012. Minerando e caracterizando dados de currıculos lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Gustavo de O Fernandes, Jonice de O Sampaio, and JM Souza. 2011. Xmlattes a tool for importing and exporting curricula data. In *International Conference on Information and Knowledge Engineering*.

Ross Ihaka and Robert Gentleman. 1996. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* 5(3):299–314.

Anthony Kay. 2007. Tesseract: an open-source optical character recognition engine. *Linux Journal* 2007(159):2.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. 2016. Jupyter notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* page 87.

Jesús Pascual Mena-Chalco, Luciano Antonio Digiampietri, Fabrício Martins Lopes, and Roberto Marcondes Cesar. 2014. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology* 65(7):1424–1445.

Jesús Pascual Mena-Chalco, Cesar Junior, and Roberto Marcondes. 2009. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society* 15(4):31–39.

Alessandra Marli M Morais, Rafael DC Santos, and M Jordan Raddick. 2015. Visualization of citizen science volunteers' behaviors with data from usage logs. *Computing in Science & Engineering* 17(4):42–50.

R. D. Peng. 2011. Reproducible research in computational science. *Science* 334(6060):1226–1227. https://doi.org/10.1126/science.1213847.

Evelyn Perez-Cervantes, Jesús P Mena-Chalco, and Roberto M Cesar. 2012. Towards a quantitative academic internationalization assessment of brazilian research groups. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*. IEEE, pages 1–8.

Rachel Schutt and Cathy O'Neil. 2013. *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc.

Guido VanRossum and Fred L Drake. 2010. *The python language reference*. Python Software Foundation Amsterdam, Netherlands.