Mining DEOPS Records: Big Data's Insights into Dictatorship

Conference Paper · December 2014

DOI: 10.1109/eScience.2014.34

Ronaldo Cristiano Prati
Universidade Federal do ABC (UFABC)

Mining DEOPS Records: Big Data's Insights into Dictatorship

Daniel de Moraes Navarro
Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC (UFABC)
Av. dos Estados, 5001, Bangú, 09210-580
Santo André, SP, Brazil
Email: daniel.phn@gmail.com

Ronaldo C. Prati
Centro de Matemática, Computação e Cognição (CMCC)
Universidade Federal do ABC (UFABC)
Av. dos Estados, 5001, Bangú, 09210-580
Santo André, SP, Brazil
Email: ronaldo.prati@ufabc.edu.br

***Abstract*—Historical data provide valuable information for the understanding of human interactions through time. However, mining this data is challenging as the available records are generally noise digitized handwritten, typewritten or press printed documents. In this research proposal, we plan to develop tools and techniques for pre-processing and extracting information from documents of the military dictatorship period that ruled Brazil from 1964 to 1985. The data to be analyzed consists of digitized images of records from DEOPS/SP (São Paulo State Department of Political and Social Order), an emblematic police agency which have monitored (and in some cases, harassed and tortured) hundreds of thousands Brazilian citizens during that period. The idea is to use state-of-the-art powerful artificial intelligence algorithms in conjunction with crowdsourcing techniques to preprocess and extract information from this important period of the Brazilian History.**

> "Those who cannot remember the past are doomed
> to repeat it"
> *George Santayana*

I. Introduction

Through the decades, humanity has collected huge amounts of data, spanning the total orbit of human and nonhuman activities. These datasets contain valuable information that can be used to help understand the behavior about the processes that generated them. Indeed, in the era of "Big Data," the time has come for processing and analyzing such large datasets.

Nevertheless, the data are useless until they are organized into conceptual frameworks able to answer useful questions. This is especially true for historical records, where data is generally available in physical formats or as digitized image collections of handwritten; press printed or typewritten documents.

Mining these datasets are very challenging as these documents are often image scan from original images corrupted by dirt, such as manual line strokes, oxidation and spilled ink [1]. Furthermore, even if it is possible to convert from digitized images to computer readable content, the data is stored in an unstructured form. Therefore, the data should be annotated to extract information and link content, such as persons, places, time and topic [2], [3].

The main objective of this project is to develop tools and techniques for pre-processing large collections of historical documents towards extracting information for data mining. The starting point will be records from São Paulo State Department of Political and Social Order (DEOPS/SP), a police agency which monitored hundreds of thousands of Brazilian citizens during the military dictatorship period from 1964 to 1985. Image digitized records from more than 250.000 people have recently been made publicly available on the Internet for research purposes in the São Paulo State Public Archive [4].

Although the data are partially annotated, the available annotations only cover a few entities (generally people) and a few records. Furthermore, automatic optical character recognition (OCR) tools have poor recognition performance. This is due to multiple reasons. OCR tools are generally tuned for English, press printed, modern fonts, and clean documents. However, DEOPS records are typewritten (from different typing machines) in Portuguese, and are very dirty due to poor storing conditions.

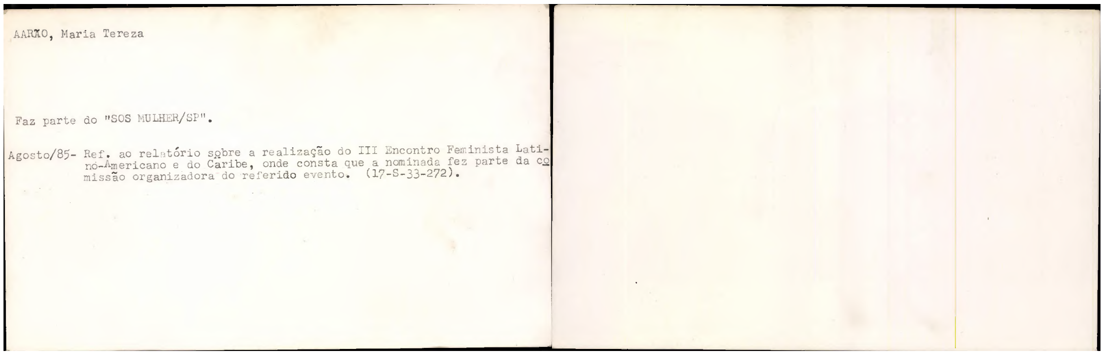
To overcome these problems, in this project we will use trainable state-of-the-art machine learning algorithms to extract text from scanned images [1], [5], allied to crowdsourcing to collect ground truth data. These algorithms can be tuned using the ground truth data to specific languages and font types, as well as use advanced image processing methods to recover contents from dirty documents.

The paper is organized as follow: Section II briefly describes the DEOPS records. Section III describes our attempts to extract information from these records. Section IV briefly discusses the crowdsourcing application we are developing to help in the character recognition task, and Section V concludes the paper.

II. DEOPS data set

São Paulo State Department of Political and Social Order (DEOPS/SP) was a police department created during 1920's with the aim to control and suppress political and social movements against the established government [6]. Several Brazilian states also had similar departments.

The agency has had strong activity throughout Brazilian History, especially during the military dictatorship years between 1964 to 1985. The authoritarian military dictatorship enacted restrictive Constitution, and stifled freedom of speech

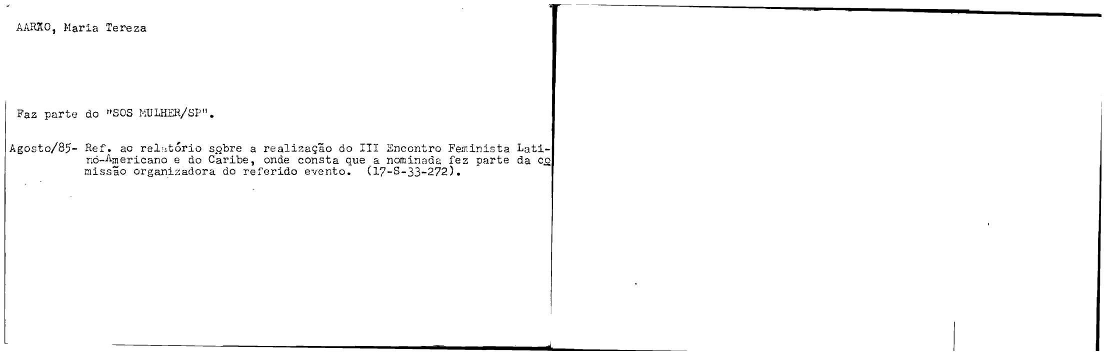


AARÃO, Maria Tereza

Faz parte do "SOS MULHER/SP".

Agosto/85- Ref. ao relatório sobre a realização do III Encontro Feminista Latino-Americano e do Caribe, onde consta que a nominada fez parte da comissão organizadora do referido evento. (17-S-33-272).

Fig. 1. An example of a DEOPS record



AARÃO, Maria Tereza

Faz parte do "SOS MULHER/SP".

Agosto/85- Ref. ao relatório sobre a realização do III Encontro Feminista Latino-Americano e do Caribe, onde consta que a nominada fez parte da comissão organizadora do referido evento. (17-S-33-272).

Fig. 2. An example of a DEOPS record after cleaning and color binarization

and political opposition. The regime also censored all media, tortured, banished and killed dissidents.

During this dictatorship period, DEOPS have monitored (and in some cases, harassed and tortured) hundreds of thousands Brazilian citizens. As the result of this action, DEOPS constructed a huge archive containing information about the people they monitored.

Although in some States these archives were burned or lost, the Brazilian Federal Police stored most of the records from DEOPS/SP after DEOPS extinction in 1985 (during the transition period between the dictatorship and democracy). In 1992, the archive was made available for consultation to the family of the tortured, killed or banished people. In 1994, the archive was made publicly available, and in 2013, digitized scans of more than 250.000 records were made publicly available on the Internet [4].

Figure 1 shows an example of a DEOPS record. As can be seen from the image, it is a digitized image of a typewritten record, with the name of the person (`AARÃO, Maria Tereza`), her affiliation (`Faz parte do "SOS Mulher/SP"`), meaning she participated in the woman-aid non-governmental organization "SOS/Mulher", and the activity they recorded (`Agosto/85 - Ref. ao relatório sobre a realização ...`), meaning she helped in the organization of a Caribbean and Latin-American meeting for women rights in 1985. The Figure also shows that the paper is oxidized and dirty (this is one of the best exemplars found in the collection).

III. Extracting information from DEOPS records

In order to extract information from DEOPS records, the digitized documents should be converted to text computer readable content. This task could be done using OCR tools. However, standard OCR tools have poor accuracy, as these tools are generally tuned for English, press printed, modern fonts, and clean documents. On the other hand, DEOPS records are typewritten (from different typing machines and print quality) in Portuguese, and are very dirty due to poor storing conditions.

In our initial attempt, we use OCROpus[1], a Python software package which implements image pre-processing tools and character recognition based on machine learning. The first step is to use a nonlinear, compute-intensive binarization method that works on degraded images to clean the image. The resulting image is shown in Figure 2.

After the binarization, the image is segmented in lines, and a recognizer based on Deep Learning neural network (more precisely, Long-Short Term Memory — LSTM [5]) is applied to recognize the characters. The first extracted line, and the corresponding recognized text, is shown in Figure 3. The full text extracted from the record sample is shown in Figure 4

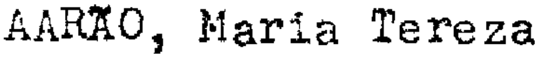
[1] https://code.google.com/p/ocropus/

```
AARN0, Haria Tereaa

Faa parte co ’GOG EULHEH/aP’’.


]Agosto/85- Ref, ao reltGrio sgbre a realiaacSo oo ncontro Feainista iati-]
ro-4mericano e do Caribe, onde consta que a nominada fez parte da 4S]
missSo organiaadora do referido evento. (17-S-33-272).
```

Fig. 4. Recognized characters from the sample DEOPS record


AARÃO, Maria Tereza

AARN0, Haria Tereaa

Fig. 3. First Line extracted from the record, and its corresponding OCR transcript

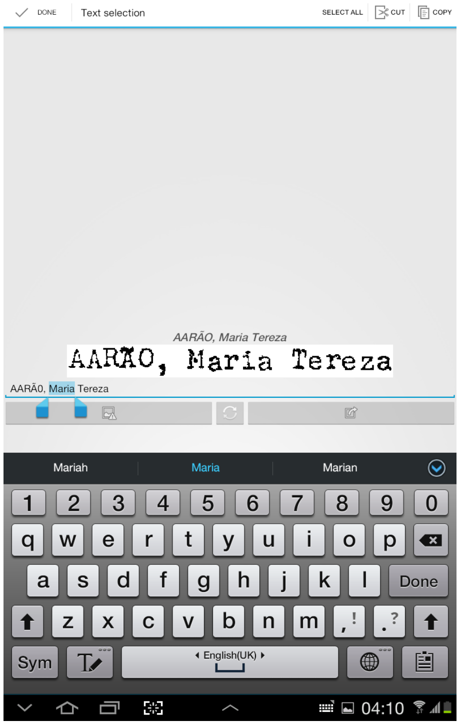
IV. A crowdsourcing application to support correct the text retrieved from scanned images

A large sample of ground truth data is required to train the techniques described in the previous section. To achieve this, we intend to use crowdsourcing [7]. Our idea is to implement an application for smartphones, tablets and social networks where the user receives a small strip of text from a record with (the imperfect) text transcript from the record. Then, the user can make the necessary corrections and submit it to a web server as a suggestion for the current strip of text retrieved from the web server.

The development of mobile applications is usually driven by the target-mobile operational system (OS), such as *Android OS*, *iOS*, *Windows Phone*, etc. Particularly in the case of smartphones, this scenario shows nowadays a 95% *Android-iOS* global duopoly with *Android* closing in on 80% by itself [8]. This is the main reason we chose *Android OS* to start developing our first mobile application. However, the technology and architecture adopted to build this first application allows the implementation of new applications for different mobile platforms or even for social networks that would easily be able to access and use the web server and databases.

The *Android* application prototype that we have been developing, initially named *AndroDEOPS*, is already capable of retrieving and displaying a small trip of text from a web database. The user may suggest any correction to the extracted text from a given scanned image and then submit the suggestion through the application. The suggestions are then stored into our database.

Figure 5 shows the main screen of the *AndroDEOPS* application. The main screen has basically four elements which can be described from the top of the screen as follows: the strip of text extracted from a given scanned image retrieved from the web database; the respective part of the processed scanned image (as described in Section III). Touching the image enables the full screen mode, which allows the user to zoom in and out the image to visually analyze specific parts of the image. Figure 6 shows the full screen mode enabled. Finally, there is a text box where the user can edit the extracted text. This edited text is the suggestion that would be submitted by the user.


AARÃO, Maria Tereza

AARÃO, Maria Tereza

Fig. 5. Main Screen of the Android application AndroDEOPS.

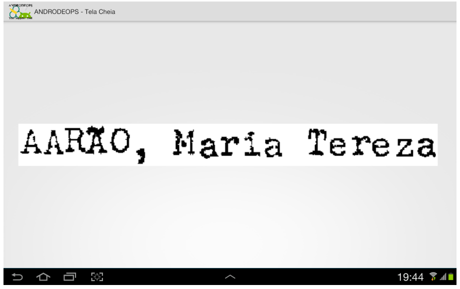


Fig. 6. Full Screen Mode of the Android application AndroDEOPS.

The button on the left hand side sets a flag indicating that the current image is illegible. This option is justified by the fact that the image processing and text extraction is intended to be an automated process, or at least semi-automated, in the sense that the resulted processed image and the respective extracted text will be directly available to suggestions submitted through the *Android* application, possibly resulting in some images of poor quality. The refresh button in the middle retrieves a new strip from the web server without submitting any suggestion. The bottom line button submits the edited text as a suggestion for the currently displayed strip of text. A new strip of text is randomly selected and presented to the user after selecting any of these three options.

Once submitted, the users' correction suggestions are received by a web server (implemented as a servlet), which stores the suggested texts in a MySQL database table together with the document id and the strip id. The idea is to use some sort of consensus aggregation to correct the mistakes made in the recognition phase. As the *Android* application AndroDEOPS has not yet been publicly available (we plan to do this in the near future), we do not have enough data to experiment with possible aggregation approaches. A possible option includes majority voting, although other alternative options are available [9].

The corrected texts will then be used as ground truth to tune the recognition algorithms. The idea is similar to reCAPTCHA, a web service used to protect websites from robots attempting to access restricted areas and helps digitize texts from books and old newspapers[2].

The motivation of using crowdsourcing is twofold: first, we believe the application has a great public appeal, as this period of the Brazilian History was marked by repression and authoritarianism from the state[3]. Second, crowdsourcing is an effective and cheap way to process large collection of data that requires human intervention.

Once converted to computer-readable formats, our intent is to develop and apply a methodology based on Natural Language Processing, Text Mining and Pattern Analysis for large-scale quantitative narrative analysis [10], in order to extract information about people, location, dates and activities. This will make it possible to convert the corpus to a network by linking key actors and objects. This network can them be used to extract knowledge about the actors, making a valuable source of research of information for historians and social scientists. The resulting processed texts and extracted information will be freely released on the Internet, to foment further research.

V. Conclusion

This paper describes our ongoing work to extract information from São Paulo State Department of Political and Social Order (DEOPS/SP) records, a large collection of records of Brazilian citizens collect by this police agency during the 1964-1985 Brazilian military dictatorship period in São Paulo State. The idea is to combine machine intelligence techniques to crowdsourcing, to extract information from the image digitized records.

We have implemented an *Android* prototype tool, in final test phase for public release in the near future. This tool will help us to collect ground truth data to fine tune the image processing and character recognition tools so that the records will be available in computer readable format for further processing. We believe this is an important step to help understanding this period of Brazilian History.

[2] reCAPTCHA's FAQ reports the displayment of 100 million CAPTCHAs every day, which help the digitizing of 13 million articles (equivalent to thirty years) of the New York Times, as well as millions of ancient books.

[3] There is a national movement to investigate human right violations during this period, named National Truth Commission (in Portuguese: Comissão Nacional da Verdade), approved by the National Congress and signed by the President Dilma Rouseff.

References

[1] Z. Dai and J. Lucke, "Autonomous document cleaning — a generative approach to reconstruct strongly corrupted scanned texts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. PrePrints, 2014.

[2] P. Manning, *Big Data in History*. London: Palgrave Pivot, 2013.

[3] K. Leetaru, *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content*. New York: Routledge, 2012.

[4] Arquivo Público do Estado de São Paulo. (2014). [Online]. Available: http://www.arquivoestado.sp.gov.br

[5] M. I. A. A. Azawi, M. Liwicki, and T. M. Breuel, "WFST-based ground truth alignment for difficult historical documents with text modification and layout variations," in *DRR*, ser. SPIE Proceedings, R. Zanibbi and B. Coüasnon, Eds., vol. 8658. SPIE, 2013.

[6] L. R. Corrêa. O Departamento Estadual de Ordem Política e Social de São Paulo: as atividades da polícia política e a intrincada organização de seu acervo. http://www.historica.arquivoestado.sp.gov.br/materias/anteriores/edicao33/materia04/.

[7] D. C. Brabham, *Crowdsourcing*. Cambridge: MIT Press, 2013.

[8] H. Kaur and M. Abodallahian, "Analytical study of global mobile market: Forecasting and substitution," in *Information Technology: New Generations (ITNG), 2014 11th International Conference on*. IEEE, 2014, pp. 485–489.

[9] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, ser. AAMAS '12. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 467–474. [Online]. Available: http://dl.acm.org/citation.cfm?id=2343576.2343643

[10] S. Sudhahar and N. Cristianini, "Automated analysis of narrative content for digital humanities," *International Journal of Advanced Computer Science*, vol. 3, no. 9, pp. 440–447, 2013.