Rebuilding the Story of a Hero: Information Extraction in Ancient Argentinian Texts

Eduardo Xamena[a,b]*, Walter Gabriel Marmanillo[a], and Ana Lidia Mechaca[a]

[a]Departamento de Informática (DI) - Facultad de Ciencias Exactas - UNSa
[b]Instituto de Investigaciones en Ciencias Sociales y Humanidades (ICSOH) - CONICET - UNSa
Universidad Nacional de Salta (UNSa)
Av. Bolivia 5150, Salta, Argentina

Abstract. Large amounts of ancient documents have become available in the last years, regarding Argentinian history. This fact turns possible to find interesting and useful aggregated information. This work proposes the application of Natural Language Processing, Text Mining and Visualization tools over Argentinian ancient document repositories. Conceptual maps and entity networks make up the first target of this preliminary paper. The first step is the normalization of OCR acquired books of General Güemes. Exploratory analyses reveal the presence of manifold spelling errors, due to the OCR acquisition process of the volumes. We propose smart automatic ways for overcoming this issue in the process of normalization. Besides, a first topic landscape of a subset of volumes is obtained and analysed, via Topic Modelling tools.

Keywords: Argentinian history, Natural Language Processing, Text Mining, Visualization, Big document repositories

1 Introduction

In the history of Argentina, many patriots have played key roles on the most salient events of this country. However, there are many blank spaces or opposite versions about relevant facts [12]. There are works that tried to determine issues of the identity of a country in terms of political, economics, social and even sports analyses [2]. For the purpose of making such tasks more efficient, the fields of Natural Language Processing (NLP), Text Mining (TM) and Visualization (VIS) provide tools that allow a faster and more comprehensive processing of the complete volume of texts available [1]. The work "Güemes Documentado" (GD) [7] consists of a compilation of documents and explanations about specific topics and moments on the life of General Martín Miguel de Güemes. The purpose of this project is to achieve interesting insights about the history of Güemes and other relevant figures of the Argentinian revolution period. This task is performed by means of first applying NLP, TM and VIS tools over GD volumes,

* Corresponding author: examena@di.unsa.edu.ar

and then extrapolating the resulting insights to other volumes from different sources.

The first step in the process of fully correct digital text acquisition has been performed by means of the Tesseract OCR engine application [14] over the GD available volumes. This step was necessary given the errors present in the raw digital versions, and the better versions attained with Tesseract open library. After this task, the text still had many spelling errors, due to the presence of noise in the physical volumes (spots, stains, scratches, etc.). This issue made an automated process of spelling correction necessary, given the high amount of errors and the usefulness of such tool for further spelling corrections in texts from other volumes. The automatic spelling errors correction task has been conducted in [13] for a general purpose spell correction project and in [9] for a specific context task of automatic spell correction. These works implement special networks of words for enhancing the process of correction, as explained in [5]. Another approach for this task is detailed in [6], by means of special Long-Short-Term Memory (LSTM) networks in a sequence-to-sequence (seq2seq) procedure.

Text Mining methods provide useful tools for the extraction of information from large volumes of documents. Particularly, the Topic Modeling task derives a landscape of topics made up by words, and associates every document with a topic in a probabilistic manner. Latent Dirichlet Allocation (LDA) [3] is a common method for Topic Modeling. Frequently, LDA is used as an exploratory data analysis in large volumes of texts. This work shows topic distributions and interpretations of subsets of documents from the GD volumes.

The present article is structured as follows: The Background and Related Work section details the techniques and methods used for this project. Then, preliminary results of the spelling correction and the first landscapes of topics are presented, accompanied by coherence values of the acquired models. Finally, conclusions and future work prospects are depicted.

2 Background and Related Work

For the purpose of performing different Text Mining tasks, such as Topic Modeling, a normalization phase is required. Such normalization includes the task of cleaning-up the documents in a corpus. As there are a plethora of spelling errors spread along the GD volumes, an automatic process of correction is required.

2.1 Automated Spelling Correction

A naive approach for the correction of spelling errors is the “brute-force” method. This consists in trying with the whole space of character combinations for finding nearby words in the corpus vocabulary when a spelling error is found. The concept “near” refers to the edition distance, which could be e.g. Levenshtein distance.This is excessively expensive in terms of computing efficiency. Besides, the words determined by the algorithm could be the non-appropriated ones in each found error.

A smarter approach is the construction of a Spelling Network (SpellNet) [5]. Such structure consists of a graph of words connected one with each other by weighted edges, and the weight of each edge is the edition distance between the words. In [13], for instance, the SpellNet is used for indexing the words for the search of candidates. Other approaches [6] use LSTM architectures for developing seq2seq mechanisms for the automatic correction of spelling errors. In fact, a special competence in the context of the ICDAR 2017 conference was carried on for this issue. In [4] a list of methods for automatic spelling correction is shown.

2.2 Topic Modeling with Latent Dirichlet Allocation

Topic Modeling is the task of identifying abstract topics or categories from a set of documents. Each topic is made up of words, and every document of the corpus has a membership proportion with each topic. LDA is a probabilistic method that provides generative models for a predefined number of topics [3]. This method has been used in recommendation systems for scientific articles [15] or in social networks analysis [8], among other applications.

The evaluation of a topic model achieved by LDA can be done with the Coherence Value [11]. This value measures degrees of interconnection of the words inside each topic, providing a proportion number. As such number gets closer to 1, the obtained model is better.

3 Proposed Framework

Given that GD volumes present many spelling errors in the first textual version, a Tesseract algorithm implementation was run on the original scanned volumes. As the selected Tesseract instance reprocesses the text pages priorly as images, the number of spelling errors decreased in some cases. The first step on automatic spelling correction is the construction of an exhaustive vocabulary of the well formed words. For this task, many lemmarios of ancient Spanish were used.

The normalization task of the framework requires the steps of word tokenization, remotion of punctuation signs and stopwords, and identifying phrasal structures such as bigrams, trigrams and so on. This stage, as required for the first task of Topic Modeling, is targeted for a bag of words representation, i.e. only accounting the frequency of each word in each document, regardless of the position of a word.

The data representation is one of the most important issues for tasks as Topic Modeling. In this case, a basic bag-of-words and a TF-IDF representation for documents have been adopted. It is known that the TF-IDF representation provides better topics. For the Topic Modeling task, the LDA method was employed.

Figure 1 illustrates the instances of the proposed workflow. The raw and Tesseract versions of the original documents are normalized. Then, the tokens that correspond to spelling errors are identified using ancient spanish lemmarios.

These lemmarios play the role of ground truth for the verification of automatic spelling correction algorithms. Besides, Bag of Words and TF-IDF representations are built upon the tokens resulting from the normalization phase. Such representations are employed for the task of Topic Modeling.

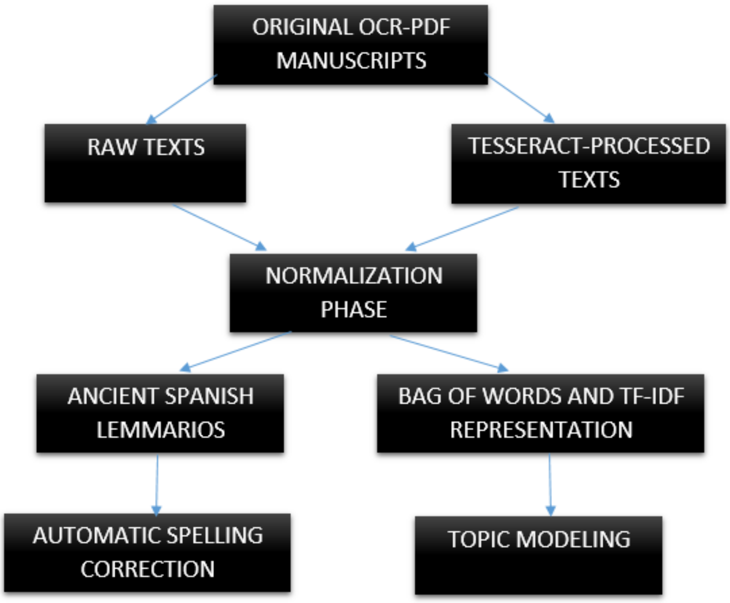
ORIGINAL OCR-PDF MANUSCRIPTS
RAW TEXTS
TESSERACT-PROCESSED TEXTS
NORMALIZATION PHASE
ANCIENT SPANISH LEMMARIOS
BAG OF WORDS AND TF-IDF REPRESENTATION
AUTOMATIC SPELLING CORRECTION
TOPIC MODELING

Fig. 1. Steps of the proposed workflow, for Automatic Spelling Correction and Topic Modeling tasks.

4 Results

After the normalization phase, a spelling checker was run on the original and Tesseract resulting texts. During the spell-checking process, a big lemmario was built using the collected correct words from GD volumes and combining the mentioned web lemmarios. The total number of unique tokens and the number of correct and incorrect tokens for the original and Tesseract-processed versions are expressed in Table 1. As can be observed there, even though results are better in most Tesseract cases, sometimes the original version of the volume has less errors. However, as this is the case for unique tokens, further counts of the total frequency of errors should be computed. Regarding the spelling errors found in the acquired texts in both sources, most of them are single characters wrongly acquired. For instance, characters “ü” and “u” are recognized as two consecutive instances of character “i”, turning e.g. the word “Güemes” to “Giiemes”, or

“Jueves” to “Jiieves”. As most errors are of this type, largely the errors should be automatically rather than manually corrected.

Table 1. Number of unique tokens and spelling errors of original volumes and Tesseract processed volumes of GD.

GD volume	Original version		Tesseract version	
	Total of tokens	Spelling errors	Total of tokens	Spelling errors
1	20,477	4,992	25,469	1,087
2	18,220	496	18,716	1,385
3	19,522	4,666	24,188	4,336
4	15,918	2,677	18,595	2,376
5	15,944	1,485	17,429	3,312
6	19,565	6,144	25,709	4,337
7	17,910	299	18,209	2,680
8	17,365	2,676	20,041	4,216
9	17,176	5,954	23,130	697
10	18,349	7,577	25,926	2,285
11	19,490	2,977	22,467	3,303
12	15,405	2,420	17,825	2,954

With the normalized versions, Topic Modeling was performed over different GD document sets by means of LDA. Table 2 shows the results obtained for the different volumes, taking chapters as documents.

Table 2. Number of chapters, LDA Coherence value and number of topics for the best LDA configuration in each processed GD volume.

Volume	# Chapters	LDA Coherence value	Number of topics
2	7	0.38	3
3	15	0.28	5
4	28	0.39	4
5	28	0.46	3
6	408	0.36	3
7	6	0.34	5
8	15	0.51	4
9	16	0.41	5
10	13	0.37	3
11	18	0.36	5
12	12	0.51	6

Additional analyses were performed on the volume 2 of GD. For this volume, when taking chapters as independent documents, the best topic model was for 3 topics, with a coherence value of 0.38. The sets of the most representative words for each topic are the following:

- Topic 1: hacer, decir, dar, partir, enemigo, Salta, ciudad, ejército, mandar, señoría, poblar, provincia, pasar, oficiar, hallar, orden, día, nombrar, Jujuy, tomar, gobernar, recibir, acordar, excelencia, poner, general, tropa, noticiar
- Topic 2: hacer, Buenos Aires, decir, dar, presentar, gobernar, excelencia, servicio, Güemes, oficial, orden, año, enero, fecho, acordar, capital, pasar, cargar, oficiar, documento, Belgrano, general, Excelentísimo señor, causar, prisionero, mandar, informar
- Topic 3: hacer, partir, decir, enemigo, dar, Salta, provincia, poblar, oficiar, pasar, hallar, día, ciudad, gobernar, tomar, ejército, mandar, señoría, presentar, poner, nombrar, bien, Jujuy, noticiar, derecho, acordar, saber, Güemes

The corresponding interpretations for the topics could be:

- Topic 1: Running order of Argentinian Army and Action letter of General Güemes: Directive of moving the troops for different officers. Transfer of General Güemes to Buenos Aires due to a supposed love affair. Possible confrontations with enemy troops.
- Topic 2: Governor election: "Puesto de Marqués" governor election. Güemes appointment as governor of Salta intendancy. Information about Güemes' wife.
- Topic 3: Tracking of enemy: Enemy troops moves over north frontier. Informing General San Martín about possible confrontations.

As seen in the interpretation, the topic distribution seems to make sense when the chapters are read. However, when the documents inside each chapter (letters, service commissions, accounting states, etcetera) are taken into the LDA process, a more precise classification is attained. 65 documents were extracted from the volume 2 with the following lists of words describing each one of 5 obtained topics (with a coherence value of 0.44):

- Topic 1: excelencia, Güemes, Excelentísimo señor, señoría, servicio, teniente coronel, general, coronel, ejército, jefe, oficial, Buenos Aires, gobernar, militar, Martin, mandar, Miguel Güemes, capitán, teniente coronel don Martín, patrio, postillón, solicitud, marchar, presentar, hacer, testar, informar
- Topic 2: Güemes, merced, cargar, prisionero, Dios guarde, testar, abonar, enero, gobernar, capitán infantería, corriente, capital, agregar, dar, decir, orden, fecho, nuevo, general, presentar, oficial don Martín Miguel, líneo, conducir, interino, individuo, sueldo, Estado, acordar, resolución
- Topic 3: ciudad, hacer, oficial, decir, hombre, mandar, prisionero, Güemes, excelencia, coronel, enero, orden, Excelentísimo señor, Córdoba, don Pedro, Teniente, Capitán don, María, Comandante, enemigo, caminar, partir, don Juan, cargar, José, causar, conducir, capitán

- Topic 4: servicio, excelencia, patrio, teniente, Buenos Aires, orden, Santiago Estero, decir, servir, Bautista, Güemes, hallar, agostar, solicitud, Tucumán, López, oficial, general, partir, destinar, ciudad, informar, noviembre, presentar, capital, prisionero, Excelentísimo señor, ejército, mayor
- Topic 5: Vuestra Merced, orden, excelencia, dar, señor, patrio, causar, honor, Martin, presentar, ciudad, mayor, servicio, Ejército, esperar, virtud, oficiar, señoría, cargar, capital, Santiago, Excelentísimo señor, salir, Dios guarde excelencia años, noticiar, dentro, gobernador, granadero, formar, decir

The interpretation for the previous topics could be:

- Topic 1: Payments due to General Güemes by his services to the army, because of a financial penalty.
- Topic 2: Letters of Francisco Fernández de la Cruz, asking for Güemes return to work, and attesting Güemes good behavior.
- Topic 3: Letters about military actions directed by Güemes.
- Topic 4: Transfer of Güemes and prisoners to Santiago del Estero.
- Topic 5: Güemes returning to the service, and apologies from General Belgrano.

In this particular case, the granularity level of documents instead of chapters lends a more accurate topic distribution according to coherence values. However, the precision of topical models could be different regarding other text structures. The chapters of GD volumes consist of documents accompanied by narrations of the author, describing the facts of such documents. This structure can be adding some noise to the topic models. The case of only extracting documents without narrations can be free of such noise, thus acquiring higher accuracy.

In Figure 2 two graphic schemes of the topics of GD volume 2 are shown. The image at the left side corresponds to the 3 topics associated with the model with the best coherence value, for the distribution by chapters. At the right side, the image is associated with the best topic model but for the distribution by the 65 documents inside chapters. The position of the circles representing the topics of each model reflect the relative distance between every pair of topics. The overlapping of some topics in the two cases means that such topics could be merged. Besides, the diameter of a circle is proportionally related to the number of documents that belong to it. Some topic circles are quite larger than others, meaning that there is not an even distribution over the number of documents belonging to each one of them.

Coherence values were computed for several LDA models both for chapters and extracted documents. Figure 3 shows two charts, the first for chapter distribution and the second for documents. The first chart shows an important drop between 3 and 4 topics. As the number of documents is low for the case of chapters, the best topic model should have 3 topics. However, in the second chart, 5 topics configuration shows a peak in coherence value. The subsequent values are not considerably higher. Hence, 5 is a good number of topics for the corresponding model.

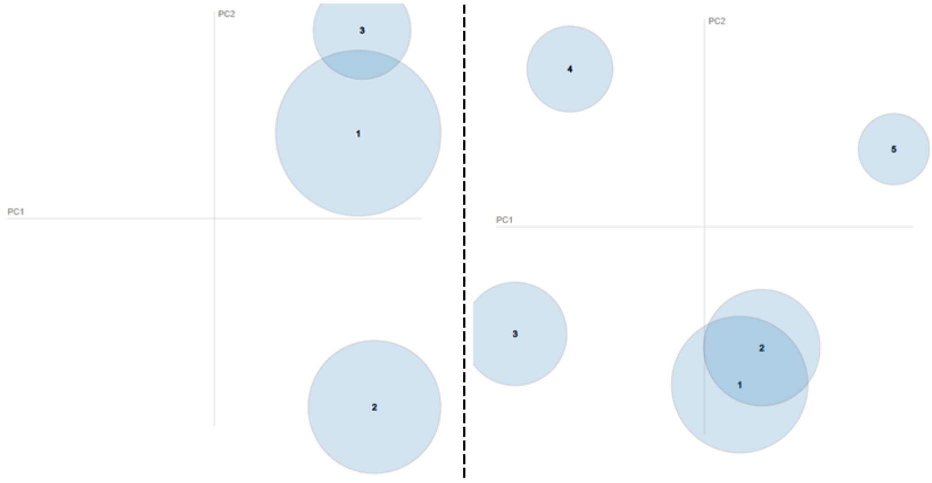

Fig. 2. Topic schemes for LDA over volume 2 of GD. At the left side, LDA topics over the complete chapters. At the right side, LDA over 65 documents extracted from the volume.

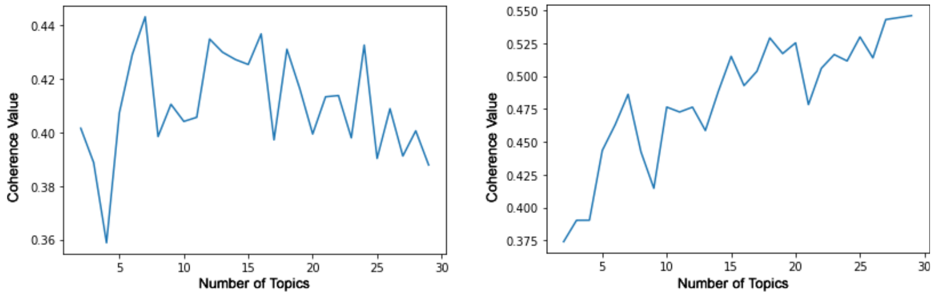

Fig. 3. Charts of Coherence values for LDA models of GD volume 2. At the left side, values for chapter distribution (7 documents). At the right side, values for document distribution (65 documents).

Apart from the quantitative criterion of the coherence value, the chosen topic model should provide a clear interpretation of what each topic means, avoiding overlapping as much as possible. Besides, for the case of chapter distribution of GD 2, it is important to note that even when there are higher coherence values for more than 3 topics, the total number of documents in the dataset is 7. Hence the peaks beyond 5 topics in the chart correspond to models that do not make sense, as there would be the same number or more topics than documents.

An enhancement on the LDA algorithm was proposed by Andrew McCallum [10]. This enhancement was implemented on the GD volumes. Specifically for volume 2, a coherence value of 0.31 was obtained for the corpus of chapters, when 5 topics were modelled with the basic LDA algorithm. The coherence value for the same configuration but using LDA Mallet algorithm was increased to 0.43. This suggests that the topics acquired with LDA Mallet could depict better distributions, as similar situations were observed for other models.

5 Conclusions and Future Work

As part of a broader project of information extraction from ancient documents of the Argentinian history, a framework for Text Normalization and Topic Modeling has been presented. This framework has been applied over a set of volumes of texts regarding the life of General Martín Miguel de Güemes, an Argentinian patriot that played a key role on the independence of this country. The results of this task are very promising. The next step for the project will be related to the recognition of named entities and the construction of conceptual maps over them. The purpose of such task is finding interactions among history figures that allow the discovery of interesting facts not easily findable.

Besides, new automatic spelling correction mechanisms will be studied on top of other related works. Such task can be enhanced by means of language models, as e.g. Google n-grams. Another related task for considering as future work is the detection and correction of malformed sentences, given that every page break can cause a bad concatenation of words when there are footnotes or similar structures in the text. A first proposal for overcoming this issue is the use of a seq2seq approach using LSTM architectures.

Acknowledgments

This work has been supported by Universidad Nacional de Salta (Proyecto CIUNSa C 2659).

References

1. Alencar, A.B., de Oliveira, M.C.F., Paulovich, F.V.: Seeing beyond reading: a survey on visual text analytics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6), 476–492 (2012)

2. Andrews, D.L., Jackson, S.J.: Sport stars: The cultural politics of sporting celebrity. Routledge (2002)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
4. Chiron, G., Doucet, A., Coustaty, M., Moreux, J.P.: Icdar2017 competition on post-ocr text correction. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1423–1428. IEEE (2017)
5. Choudhury, M., Thomas, M., Mukherjee, A., Basu, A., Ganguly, N.: How difficult is it to develop a perfect spell-checker? a cross-linguistic analysis through complex network approach. arXiv preprint physics/0703198 (2007)
6. Etoori, P., Chinnakotla, M., Mamidi, R.: Automatic spelling correction for resource-scarce languages using deep learning. In: Proceedings of ACL 2018, Student Research Workshop. pp. 146–152 (2018)
7. GUEMES, L.: Guemes documentado (1979)
8. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. pp. 80–88. acm (2010)
9. Huang, Y., Murphey, Y.L., Ge, Y.: Automotive diagnosis typo correction using domain knowledge and machine learning. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). pp. 267–274. IEEE (2013)
10. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), `http://mallet.cs.umass.edu`
11. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. Association for Computational Linguistics (2010)
12. O'Donnell, P.: Los héroes malditos. DEBOLS! LLO (2017)
13. Reynaert, M.: Ticclops: Text-induced corpus clean-up as online processing system. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. pp. 52–56 (2014)
14. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
15. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 448–456. ACM (2011)