Text mining a Portuguese book on Freemasonry: Disclosing network communities' features

Conceição Rocha[1,3], Alípio Mário Jorge[2,3], Márcia Oliveira[1,3], Paula Brito[1,3], João Gama[1,3], and Carlos Pimenta[1,4]

[1] FEP
Rua Dr Roberto Frias, 4200-464 Porto, Portugal
[2] FCUP
Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal
[3] LIAAD/INESC TEC
Campus da FEUP, Rua Dr Roberto Frias, 4200-465 Porto, Portugal
[4] OBEGEF
FEP - Gabinete 519, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
{cnrocha,mdbo}@inesctec.pt,
{amjorge}@fc.up.pt,
{mpbrito,jgama,pimenta}@fep.up.pt

Abstract. This work presents a social network analysis based on entities extracted from a Portuguese book on Freemasonry. The book is processed with text mining techniques. Named entities are identified and extracted and a network model comprised of entities, represented by nodes, and their co-occurences in the same sentences, represented by weighted links is studied. Some network patterns, such as the internal structure of a network community, are explored and visualized through word cloud and graph representations. The results indicate that the applied text mining procedure is able to extract most entities from the text book despite some current limitations that still need to be improved. The social network analysis identifies reasonable connections between some well known entities.

Keywords: Social network analysis, Network targeting, Text mining, Information extraction

1 Introduction

Social network analysis (SNA) is used not only to detect clusters or communities but also to identify central actors in social networks. In several fields, the network information — the set of actors and their relationships — is implicitly stored in unstructured natural-language documents [5]. Therefore, text mining techniques (TM), such as information extraction, are required to pre-process the texts in order to extract the social entities and the relations between them. One potential application for this is to provide a book's second screen and to help readers understanding books more easily and even have the oportunity to link what they are reading with other information sources.

Methods and techniques from text mining have been developed and applied to natural language text with the purpose of extracting meaningful and useful information from these unstructured texts [2, 3]. Most text mining tools developed are applicable to English literature. Hence, free text mining tools for Portuguese literature are still difficult to find. Furthermore, when employing information extraction technology to discover knowledge in text, the text language is important, since the text mining tools must be adapted to the problem and to the particular network under study [7].

Three important contributors to text mining are the fields of Natural Language Processing (NLP), Information Extraction (IE) and Information Retrieval (IR) [4, 9]. One example of IE is the Named Entity Recognition which is considered an initial step to perform other tasks, such as relation extraction, classification or/and topic modeling [2]. The IE techniques are mostly based on pattern matching but an alternative approach using semantic relations has already been studied by the TM community [1, 6]. On the other hand, techniques as tokenization, morphological or lexical analysis, syntactic analysis and semantical analysis are considered components of Natural Language Processing (NLP), whose complexity becomes apparent when it is possible to tag a word with more than one part of speech [4].

Although the natural language offers some patterns that could and should be used to extract information from unstructured texts, text mining still poses many challenges due to the ambiguous features of natural language.

In this work, we process a book about the Portuguese Freemasonry and its important connections to political organizations and individual actors [10]. The book has been fully digitalized with optical character recognition and then processed using text mining techniques and social network analysis. The aim is to explore the potential to use available data mining tools to identify and target social networks or central network actors [8] in the Portuguese society. The long term aim is to exploit features from the vast collection of textual sources in the subject, such as newspapers, magazines, books and web pages.

Acknowledgments. This work is partially funded by FCT/MEC through PIDDAC and ERDF/ON2 within project NORTE-07-0124-FEDER-000059 and through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281. Márcia Oliveira acknowledges funding from FCT, through Ph.D. grant SFRH/BD/81339/2011.

References

1. K. Barkschat. Semantic Information Extraction on Domain Specific Data Sheets. *The Semantic Web: Trends and Challenges, Springer International Publishing*, 8465:864–873, 2014.

2. D. Campos, S. Matos, and J. L. Oliveira. Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. *Theory and Applications for Advanced Text Mining, S. Sakurai, Ed. InTech*, pp.175–195, 2012.
3. M. S. Conrado, A. D. Filippo, T. A. S. Pardo, and S. O. Rezende. A Survey of Automatic Term extraction for Brazilian Portuguese. *Journal of the Brazilian Computer Society*, 20(12), 2014.
4. S. Jusoh, and H. M. Alfawareh. Techniques, Applications and Challenging Issue in Text Mining. *International Journal of Computer Science Issues*, 9(6) no.2: 431–436, November 2012.
5. R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, June 2005.
6. A. Moschitti, P. Morarescu, and S. M. Harabagiu. Open Domain Information Extraction via Automatic Semantic Labeling. *American Association for Artificial Intelligence*, 20013.
7. R. Sagayam, S. Srinivasan, and S. Roshni, A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques. *International Journal of Computational Engineering Research*, 3(5):1443–1446, 2012.
8. D. M. Schwartz and T. Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009.
9. S. Umajancy, and Dr. A. S. Thanamasi. An analysis on text mining Text retrieval and text extraction. *International Journal of Advanced Research in computer and communication engineering*, 2(8), August 2013.
10. A. Vilela. *Segredos da maçonaria portuguesa.* A Esfera dos Livros, 2013.