From entity extraction to network analysis: a method and an application to a Portuguese textual source

Conceição Rocha*
Faculdade de Economia da Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
cnrocha@inesctec.pt

Alípio Mário Jorge*
Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre, s/n
4169-007 Porto, PORTUGAL
amjorge@fc.up.pt

Márcia Oliveira*
Faculdade de Economia da Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
mdbo@inesctec.pt

Paula Brito*
Faculdade de Economia da Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
mpbrito@fep.up.pt

João Gama*
Faculdade de Economia da Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
jgama@fep.up.pt

Carlos Pimenta†
Faculdade de Economia da Universidade do Porto
Rua Dr. Roberto Frias
4200-464 Porto, PORTUGAL
pimenta@fep.up.pt

ABSTRACT

This paper reports the challenges and the proposed solutions to extract named entities from a Portuguese book. Despite development on methods and techniques for text mining (TM) the truth is that there are few applications to extract information from unstructured text, and even less adequate for Portuguese language. The strategy adopted in this work consists on, whenever it is possible, use or combine different methods already implemented in free software exploring their capability to perform the goal. From this strategy results a text mining procedure that not only can be applied to any unstructured text written in Portuguese language but also can easily be adapted to any other Roman alphabetic writing system. In this work, only free software are used and all the packages are identified along with its application. Illustrative examples of the faced problem and of the finding solution are given for each step and the quality of each phase of the process is measures through the recall and precision. 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 long term aim is to exploit features of any topic from the vast collection of textual sources such as newspapers, magazines, books and web pages. Therefore, a social network analysis is performed on the named entities of a Portuguese book on Freemasonry to explore the potential use of this available data mining tools to identify and target social networks or central network actors in the Portuguese society.

Categories and Subject Descriptors

H.4 [**...Information Systems Applications**]: Miscellaneous; D.2.8 [**...Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

...Delphi theory

Keywords

Text mining, Information extraction

1. INTRODUCTION

Text mining (TM) techniques and methods were developed and applied to natural language text with the purpose of extracting meaningful and useful information from unstructured texts [6, 5]. Three important contributors to text mining are the fields of Natural Language Processing (NLP), Information Extraction (IE) and Information Retrieval (IR) [12, 19]. One important task of IE is Named Entity Recognition which is considered an initial step to perform other tasks, such as relation extraction, classification or/and topic modelling [6]. 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, 14]. 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 [12].

*All the five authors are also with the LIAAD/INESC TEC

†Carlos Pimenta is also with the OBEGEF

SAC'15 April 13-17, 2015, Salamanca, Spain.

http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

Although natural language offers some patterns that could and should be used to extract information from unstructured texts, text mining still faces many challenges due to the ambiguous features of natural language. 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 subject under study [18].

In several fields, the network information — the set of actors and their relationships — is implicitly stored in unstructured natural-language documents [13]. Hence, text mining techniques, 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 understand books more easily and even have the opportunity to link what they are reading with other information sources. Another potential application for this is to provide the information to study the social context of possible economic and financial offences through social network analysis (SNA) since it detects clusters or communities and identify central actors in social networks.

In this work, we propose a novel method to automatically extract named entities from unstructured texts using free software. All the process is exemplified with a book about the Portuguese Freemasonry due to its important connections to political organizations and individual actors [20]. The book has been fully scanned with optical character recognition and then processed using text mining techniques and social network analysis. The aim is to explore the potential use of those available data mining tools to identify and target social networks or central network actors [17] 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.

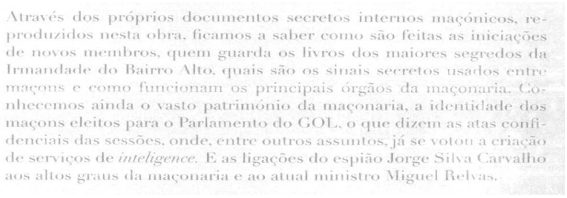
This article is organized as follows. In Section 2 the text mining process, including challenges and adopted strategies, is fully described and some quality measures are estimated. Section 3 presents the social network analysis. Finally, Section 4 draws some remarks and conclusions.

2. THE ENTITY EXTRATION PROCESS

In this section is described the process to extract the named entities from Portuguese unstructured texts. Examples selected from a Portuguese book are used to illustrate the difficulties and the solutions adopted along the process. As the goal of this work is to extract named entities from Portuguese texts using methods and techniques of text mining implemented and make available on free software. Hence, the main program was developed in `R` [15] and makes use of some specific text mining packages, as described later.

In our case, we have obtained the text from a scanned book which poses additional difficulties. Figure 1 depicts some differences between the original text, represented in the top panel, and the text to be submitted to text mining tools, represented in the bottom panel.

As it is expected a preprocessing step is necessary to remove at least extra lines, extra spaces between words or letters and page numbers, *i.e.*, it is necessary to remove some


Através dos próprios documentos secretos internos maçónicos, reproduzidos nesta obra, ficamos a saber como são feitas as iniciações de novos membros, quem guarda os livros dos maiores segredos da Irmandade do Bairro Alto, quais são os sinais secretos usados entre maçons e como funcionam os principais órgãos da maçonaria. Conhecemos ainda o vasto património da maçonaria, a identidade dos maçons eleitos para o Parlamento do GOL, o que dizem as atas confidenciais das sessões, onde, entre outros assuntos, já se votou a criação de serviços de *inteligence*. E as ligações do espião Jorge Silva Carvalho aos altos graus da maçonaria e ao atual ministro Miguel Relvas.

\ traves dos próprios documentos secretos internos maçónicos, reproduzidos nesta obra. ficamos a saber como são feitas as iniciações de novos membros, quem guarda os livros dos maiores segredos da Irmandade do Bairro Ut O, quais são os sinais secretos usados entre maçons e como funcionam os principais órgãos da maçonaria. Conhecemos ainda o vasto património da maçonaria, a identidade dos maçons eleitos para o Parlamento do G( )L. o que dizem as alas confidenciais das sessões, onde. entre outros assuntos,já se votou a criação de sen iços de inteligence. K as ligações do espião Jorge Silva < Ian alho aos altos graus da maçonaria e ao aluai ministro Miguel Relvas.

Figure 1: One example of a paragraph of the book in the top panel and the results after been scanned with optical character recognition on the bottom panel.

junk before proceeding to the information extraction process.

Preprocessing step.

The program was constructed based on pattern matching strategy and make use of the following packages: **tm** [8], **gdata** [21], **stringr** [22] and **memoise** [23].

Information extraction step.

In this work entities are persons, places, institutions and organizations or associations. This step describes how the list of named entities present in the 2508 sentences of the book is extracted. To identify the entities, an information extraction procedure was constructed using regular expressions and other pattern matching strategies along with part-of-speech tagging, *i.e.*, using a POS tagger tool.

The process is divided in three phases and the inclusion of phase 1 before the use of the POS tagger tool is relevant for the quality of the process. For instance, in the entity name *'Grande Oriente Lusitano'* only *Lusitano* is identified by the POS tagger. Another example is the entity name *'Fernando Pessoa'* where only *Fernando* is identified as entity name.

- **Phase 1** Since entity names start frequently with capital letters, a regular expression based on that pattern is used to extract the named entities. To include the named entities that often appear written in lower case a list of name expressions — *e.g. presidente*, *câmara*, *deputado*— is added to the regular expression used to extract the terms (candidates to named entities). This phase uses the following packages: **tm** [8], **cwhmisc** [10] and **memoise** [23].

  From this phase and for the paragraph in Figure 1 result *'Irmandade do Bairro Ut O'; 'Conhecemos'; 'Parlamento do G'; 'L'; 'K'; 'Jorge Silva'; 'Ian'* and *'ministro Miguel Relvas'* as the terms candidates to named entities.

- **Phase 2** In this phase the terms list, of the previous phase, is part-of-speech tagged. After that, the program removes all the terms that do not have at least one tag 'prop'(POS tag meaning a proper noun) and remove the first word from the ones having a 'prop'

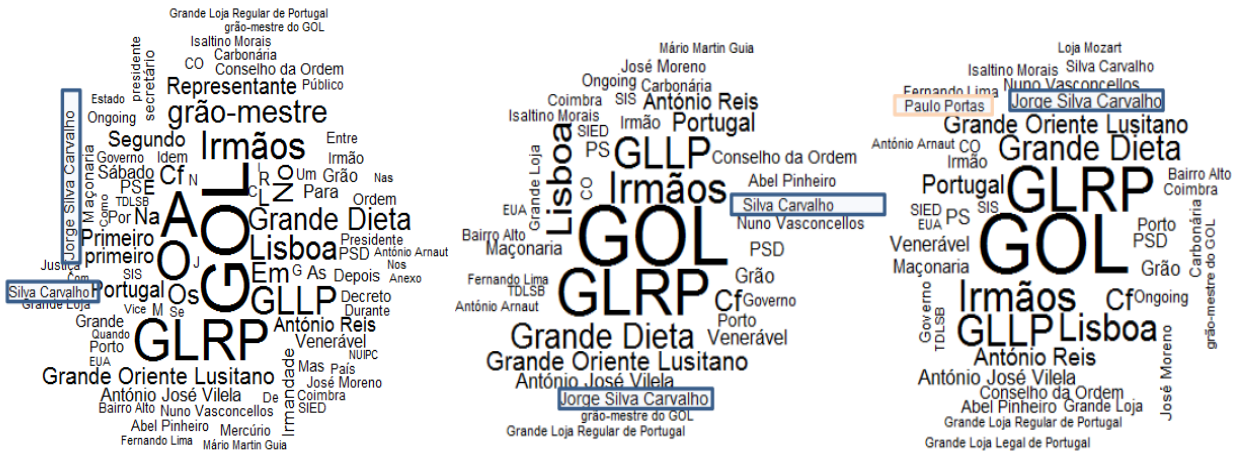


Figure 2: Word cloud representation of the terms (candidates do named entities) that appear 25 or more times in the text after each one of the three phases of the process.

tag but starting by 'pron-det' (POS tag meaning determiner pronoun). Finally, some stop words are removed and the terms present in the sentences of the text, are identified. The two rules used in this phase are based on the tags characteristics of the named entities. Since the entities considered in this work are persons, associations, institutions and places, the first rule is to remove terms which are not tagged as proper noun and the second rule cleans words that do not belong to the tags structure of that entity name. This phase uses the following packages: **tm** [8],**memoise** [23], **openNLP** [9], **Hmisc** [11].

As a result of the application of this procedure to the previous list of terms is the reduction to five terms, *'Irmandade do Bairro Ut O'; 'Parlamento do G'; 'Jorge Silva'; 'Ian'* and *'ministro Miguel Relvas'*. In fact, all the five terms are the only named entities in the paragraph.

- **Phase 3** This phase aims to associate different designations, of the same entity, used on the text to these entity. For instance, *'Paulo Portas'* and *Portas* are two names used to refer to the portuguese minister Paulo Portas. This phase uses the following packages: **tm** [8], **stringr** [22], **Hmisc** [11], **base** [16], **Matrix** [2], **memoise** [23], **memisc** [7]. To perform this phase measures of similarity are computed for each par of entities named present in the three paragraph, the one in analyse, the previous and the next paragraph. The final criterion takes in consideration measures of the similarity between the names and the words present in the name.

To give some idea of the improvement introduced by each one of the phases we represent the entities, along with their frequency, in a word cloud representation, where words with higher frequency have larger font size. As can be observed in Figure 2, after the phase 1 some words that do not refer to entities such as *'Idem', 'Entre'* and *'Nas'* are represented in the cloud. And, as it was expected, after the phase 2 that words disappear from the cloud. The phase 3, as can be observed in Figure 2 by the examples marked with squares, joint the terms *'Silva Carvalho'* and *'Jorge Silva Carvalho'* into only one named entity *'Jorge Silva Carvalho'*. And, finally, a name that do not appears before phase 3, *'Paulo Portas'* is visible in a square on the cloud representation after phase 3. As it was already referred some times this entity is cited as *'Portas'* and some times as *'Paulo Portas'* and the phase 3 of the process associates the two names to only one entity increasing the frequency of that named entity in the text.

Table 1: Measures of the process quality

Variable	Phase 1	Phase 2
extracted terms	5089	3075
named entities on the extracted terms	3205	2982
recall	0.84	0.78
precision	0.63	0.97

From this book, a total of 12120 events are extracted from the text mining process corresponding to 5159 unique terms. To quantify the quality of this process a total of 125 pages of the book (the first ones) were manually labelled (1/3 of the text book) and the computed measures are record on Table 1. This part of the book contains 3836 named entities. The recall and precision are estimated for the first two phases based on the results obtain on the 125 pages of the book. A total of 5089 terms were labelled as entities in the first phase and 3075 in the second phase. The true positives were 3205 in the first phase and 2982 in the second phase. This means that the *recall*, given by Equation 1, decreases from 0.84 to 0.78 and the *precision*, given by Equation 2, increases from 0.63 to 0.97 in the second phase.

$$recall = \frac{\text{number of named entities extracted}}{\text{number of named entities present on the text}} \quad (1)$$

$$precision = \frac{\text{number of entities' names extracted}}{\text{number of terms extracted from the text}} \quad (2)$$

Equation 3 defines another measure used to quantify the quality of process, the $F - measure$. The second phase of this process increases the $F - measure$ from 0.72 to 0.865.

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

Two remarks regarding Figure 2: the words *GOL* (a Portuguese Freemasonry lodge) and *Grande Oriente Lusitano* appear as two different entities whereas in fact they refer to the same entity; the second one is that words such as *Venerável* and *Grão* are ambiguous.

3. NETWORK ANALYSIS

In this section we use the named entities, previously extracted from all the book, and we analyse the relations between entities using networks. The entities that co-occur in the same sentence of the book are considered to be connected and each combination of two of them forms a pair of linked nodes. The weight of the relation is then the number of times that both entities appear in the same sentence in the whole book. The pairs of entities and the number of times they appear together in the same sentence form the social network under consideration.

Note that in this work, we do not perform a deep analysis of the connections between entities. Nevertheless, it is interesting to analyse, note only, the named entity (or entities) that is citied, simultaneously, more times and with more named entities, but also, to detect groups or communities of entities formed

The whole social network has 4364 nodes and 23875 edges. It is an undirected and weighted graph which depicts the size and complexity of the underlying network. Nodes without links do not appear. To allow for a detailed visualization of the most connected entities a degree-based filter is applied. The resulting representation is shown in Figure 3, which allows to visualize the sub-network of entities with minimum degree 83. This allows to reduce the number of nodes presents in the representation and to identify some of them. Some of the names visualized in this graph are in the word cloud as well, but not all of them have the same importance. In this representation, the vertices with greater number of neighbours in the network may be identified.

In this work and in order to detect communities it is used the implementation of the Louvain method made available in the Gephi software [3]. This is a modularity-maximization algorithm that produces good quality partitions of the network in a very fast way. The method comprises two phases. The first phase optimizes modularity in a local way by looking for positive gains in modularity when moving a node to a neighbouring community. The second phase is similar to the first one, with the difference that now we deal with a modified network, where each vertex is a super-vertex, which represents the previously found communities. Considering this higher-level setting, the steps of the first phase are repeated iteratively until a maximum of modularity is attained and new hierarchical levels and super-graphs are yielded. The algorithm stops when modularity converges to a value where no more gains are possible.

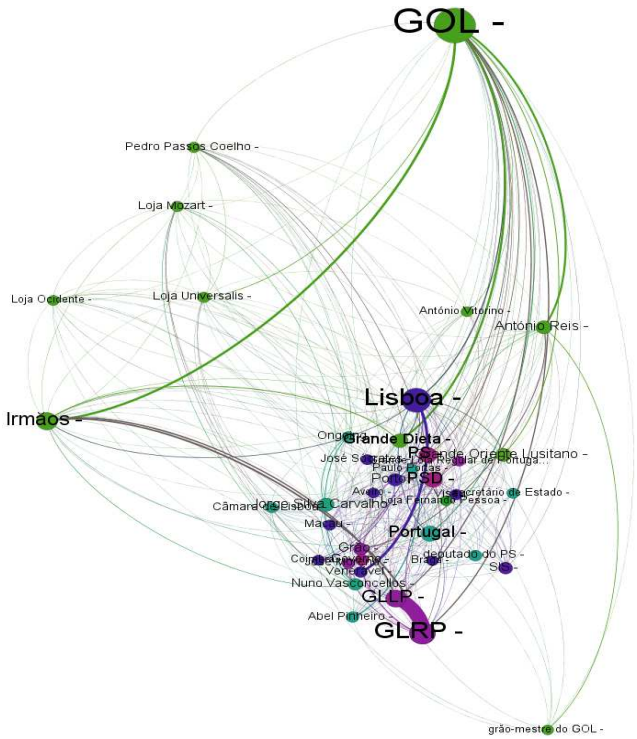


Figure 3: Social network graph filtered by degree - minimum 83.

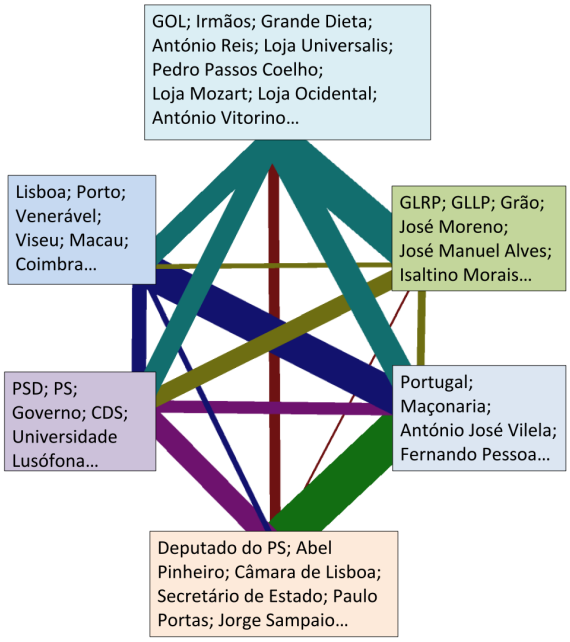


Figure 4: Higher-level representation of the previous graph highlighting identified communities and the corresponding entities; the thickness of the edges is proportional to the strength of the connection between communities.

Table 2: Network-level metrics

Network	global clustering coefficient	0.855
	average path length	3.37
	average degree	10.94
	average weighted degree	12.245
	network diameter	12
	network radius	1
	graph density	0.003
	modularity	0.68
	Nº of communities	223
	Nº weakly connected components	191

The entities are grouped in communities and the colours of each node are associated to a group. Within the context of our application, this means that a community identify sets of entities which co-occur more frequently with each other than with other entities, when considering the whole book. A further representation of this information is added in Figure 4 where entities are grouped by communities. From Figure 4 it is clear that the Portuguese Freemasonry has two great factions, *GOL* and *GLLP/GLRP* (two references to another lodge), and that they are linked to entities from the two major Portuguese political parties, *PS* and *PSD*.

Betweenness is defined as the number of shortest paths between two arbitrary nodes that pass through the node under analysis and the betweenness value measures the extent to which a node lies between other nodes in a network. Hence, according to the betweenness and the eigenvector centrality values, *GOL (Grande Oriente Lusitano)*, *Lisboa* and *GLRP/GLLP* are the three most central entities. Apart from these three entities, the list of the twenty most central entities contains, among others, the following eight names: *PS*, *SIS*, *Nuno Vasconcelos*, *António Reis*, *Portugal*, *Porto*, *PSD* and *Jorge Silva Carvalho*. The idea behind **Eigenvector centrality** is that a node that is connected to nodes that are themselves well connected should be considered more central than a node that is connected to an equal number of less connected nodes and unlike degree centrality, the eigenvector centrality value measures how well a given node is connected to other well-connected nodes in the network,the eigenvector concept takes into account both direct and indirect influences.

Some statistical measures that characterize the network are summarized in Table 1. From the graph density value, *i.e.*, the proportion of the number of links present in the network relative to the maximum number of possible links, we can see that the network is sparse. Nevertheless, the local node group cohesiveness is high, as indicated by the large value of the global clustering coefficient — fraction of triangles in the network relative to the maximum number of possible triangles. It measures the overall cohesion of the node's neighbourhood in a network. The existence of such tightly knit neighbourhoods can be explained by the network model and the assumption that entities occurring in the same sentence are all linked to each other. The number of communities and of weakly connected components are also high, 223 and 191, respectively. The communities detected in this network are considered meaningful since the modularity value is large and above 0.3 [4].

Figure 5 depicts the network graph for the main community, i.e. the community comprising the majority of nodes. A closer look at this figure reveals the heterogeneity of the main community, which exhibits a rich internal structure. This structure can be further examined and decomposed into more refined communities.

Figure 5: Social network graph for the larger cluster - The main community.

Once again, for enhancing the visualization of the constituent entities, only 10% of the community is considered. Its graph is represented in Figure 6. Names like *GOL*, *Grande Dieta*, *António Reis* and *Irmãos* are present in this graph and were also emphasized in all the previous figures. Names like *GLRP*, *GLLP*, *Lisboa* and *Portugal* do not appear in this graph since they belong to another community.

4. REMARKS AND CONCLUSIONS

Considering that the text mining procedure used to identify and extract entity names is not finished, and that the relation between entities is given by their co-occurrence in the same sentence, the results are quite meaningful. In fact, the text mining procedure was able to both extract the book's entities and to identify the lodge *GOL (Grande Oriente Lusitano)*, the location *Lisboa* and the lodge *GLRP/GLLP* as the three most central entities, even without a disambiguate step. The inclusion of the second phase on the process allows to improve the quality of the extraction process since the $F-measure$ increases from 0.72 to 0.865. We can also see the relevant connections in terms of some political organizations, politicians and other public figures. Nevertheless, the objective here was not to analyse the network actors but rather to bring out the overall network community structure and show that this approach can be used for providing a diagrammatic synthesis of the book. The results obtained so far may also be considered a step towards the creation of a text intelligence system to be used in the study of the social context of possible economic and financial offences. Moreover, our exploratory analysis of the main community highlights some heterogeneity in its structure that could be further explored. As future work the authors are planning to improve the text mining procedure, by including a disambiguate step, as well as by adjusting the network model so that entities are linked based on the verbs present

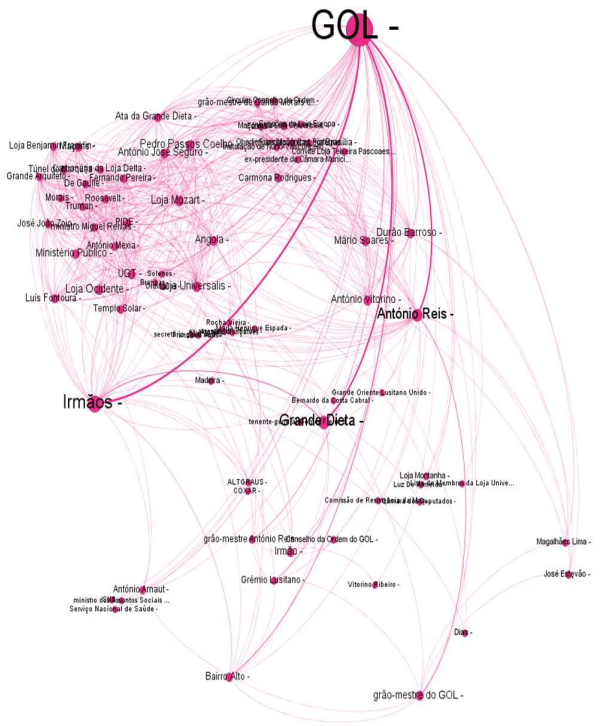


Figure 6: Social network graph for 10% of the larger cluster. Nodes' size is proportional to their degree centrality.

on the sentences.

5. 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.

6. REFERENCES

[1] K. Barkschat. Semantic information extraction on domain specific data sheets. In V. Presutti, C. dâĂŹAmato, F. Gandon, M. dâĂŹAquin, S. Staab, and A. Tordai, editors, *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 864–873. Springer International Publishing, 2014.

[2] D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2014. R package version 1.1-4.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004.

[5] M. da Silva Conrado, A. Felippo, T. Salgueiro Pardo, and S. Rezende. A survey of automatic term extraction for brazilian portuguese. *Journal of the Brazilian Computer Society*, 20(1):12, 2014.

[6] S. M. David Campos and J. L. Oliveira. *Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools*. S. Sakurai, Ed. InTech, 2012.

[7] M. Elff. *memisc: Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation*, 2013. R package version 0.96-9.

[8] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, 3 2008.

[9] K. Hornik. *openNLP: Apache OpenNLP Tools Interface*, 2014. R package version 0.2-3.

[10] C. W. H. <http://www.echoffmann.ch>. *cwhmisc: Miscellaneous Functions for math, plotting, printing, statistics, strings, and tools*, 2013. R package version 4.0.

[11] F. E. H. Jr, with contributions from Charles Dupont, and many others. *Hmisc: Harrell Miscellaneous*, 2014. R package version 3.14-4.

[12] S. Jusoh and H. M. Alfawareh. Techniques, applications and challenging issue in text mining. *International Journal of Computer Science Issues*, 9(6):431–436, 11 2012.

[13] R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, June 2005.

[14] A. Moschitti, P. Morarescu, and S. M. Harabagiu. Open domain information extraction via automatic semantic labeling. In *FLAIRS Conference'03*, pages 397–401, 2003.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[17] D. Schwartz and T. (D.A.) Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009.

[18] D. M. Schwartz and T. Rouselle. A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research*, 2(5):1443 –1446, 2012.

[19] S. Umajancy and D. A. S. Thanamasi. An analysis on text mining âĂŞ text retrieval and text extraction. *International Journal of Advanced Research in computer and communication engineering*, 2(8):3125–3129, 2013.

[20] A. Vilela. *Segredos da maçonaria portuguesa*. A Esfera dos Livros, 2013.

[21] G. R. Warnes, B. Bolker, G. Gorjanc, G. Grothendieck, A. Korosec, T. Lumley, D. MacQueen, A. Magnusson, J. Rogers, and others. *gdata: Various R programming tools for data manipulation*, 2014. R package version 2.13.3.

[22] H. Wickham. *stringr: Make it easier to work with strings.*, 2012. R package version 0.6.2.

[23] H. Wickham. *memoise: Memoise functions*, 2014. R package version 0.2.1.