OBEGEF – Observatório de Economia e Gestão de Fraude

WORKING PAPERS

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

Conceição Rocha, Alípio Mário Jorge,
Márcia Oliveira, Paula Brito, João Gama,
Carlos Pimenta

OBEGEF
Observatório de Economia
e Gestão de Fraude

>> FICHA TÉCNICA

FROM ENTITY EXTRACTION TO NETWORK ANALYSIS: A METHOD AND AN APPLICATION TO A PORTUGUESE TEXTUAL SOURCE

WORKING PAPERS Nº 32/ 2014
OBEGEF – Observatório de Economia e Gestão de Fraude

Autores: Conceição Rocha [1,3], Alípio Mário Jorge[2,3], Márcia Oliveira[1,3], João Gama[1,3], Carlos Pimenta[1,4]

Editor: Edições Húmus
1ª Edição: Novembro de 2014
ISBN: 978-989-755-087-4

Localização web: *http://www.gestaodefraude.eu*
Preço: gratuito na edição electrónica, acesso por download.
Solicitação ao leitor: Transmita-nos a sua opinião sobre este trabalho.

Trabalho de pós-doutoramento, com bolsa da Faculdade de Economia da Universidade do Porto, inserido numa parceria entre o LIAAD/INESC e o OBEGEF.

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

>> ÍNDICE

>> **RESUMO**

Este artigo dá a conhecer os avanços conseguidos na extração de entidades (identificação de entidades referidas) num processo de mineração de texto cujo objetivo é revelar estruturas semânticas não triviais, tais como relações e interações entre as entidades ou comunidades. É proposto um método de três fases aplicável à língua Portuguesa e potencialmente a outras línguas. O método baseia-se em correspondência de padrões flexível, na marcação da categoria morfo-sintática de cada palavra, em regras lexicais e na distância entre os nomes das entidades. Todas as etapas são implementadas em software livre usando vários pacotes disponíveis. A avaliação da eficácia do método de extração de entidades é feita tendo por base uma parte de um livro escrito em português observando-se uma melhoria na medida F1. Para uma melhor compreensão e avaliação da utilidade do método proposto apresentamos um caso de um livro sobre Maçonaria. É também definida uma rede social das entidades referidas com base exclusivamente em citações do livro. Daí são extraídas informações estruturais que revelam conexões, relacionamentos e comunidades entre as entidades.

>> ABSTRACT

This paper reports advances in the entity extraction task (named entity identification) of a text mining process that aims at unveiling non-trivial semantic structures, such as relationships and interaction between entities or communities. We proposed a 3-phase method that is applicable to the Portuguese language and potentially applicable to other languages as well. The method relies on flexible pattern matching, part-of-speech tagging, lexical-based rules and distance-based entity name merging. All steps are implemented using free software and taking advantage of various existing packages. Evaluation of the efficacy of the entity extraction method on part of a book written in portuguese indicates improved F1 results. For further evaluation and illustration of the usefulness of the proposed method, it is applied to a book on Freemasonry and observe the differences in the entity word clouds produced. We also define a social network of named entities solely from information contained in the book and extract structural insights that reveal connections, relationships and communities between entities.

>> 1. INTRODUCTION

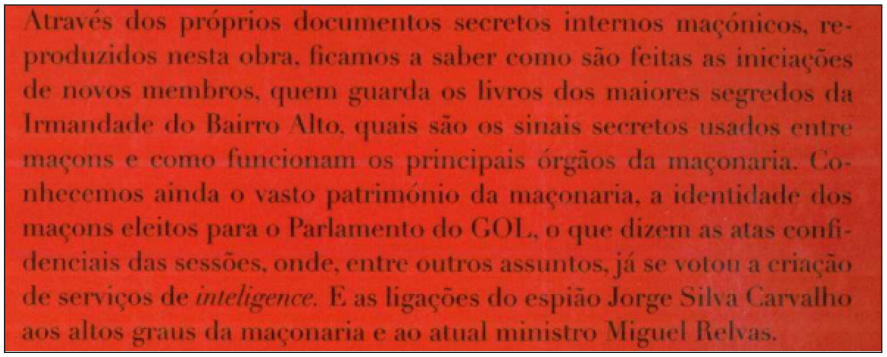
With the increasing amount of information that is produced and/or shared in all fields becomes more and more necessary to have automatic processes (such as text mining processes) to extract meaningful and useful information from unstructured texts [5, 4]. Entity extraction (named entity identification) is one of the tasks of those processes. In fact, named entity identification is also considered as an initial step to perform other tasks, such as relation extraction, classification or/and topic modelling [5].

In several fields, the network information — the set of actors and their relationships — is implicitly stored in unstructured natural-language documents [11]. 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 objective 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 is the extraction from texts of the necessary information to study the social context of possible economic and financial offences through social network analysis (SNA) since it enables the detection of clusters, or communities, and the identification of central actors in social networks. By harnessing the information extracted through named entity identification with the application of SNA techniques, it is possible to grasp insights about the semantic structure of textual data.

Although natural-language exhibits 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. 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 [14]. Most text mining tools developed are applicable to English literature. Hence, free text mining tools for Portuguese literature are still difficult to find.

In this work, we propose a 3-phase method to automatically extract named entities, from unstructured texts, applicable to the Portuguese language and potentially applicable to other languages as well. The method relies on flexible pattern matching, part-of-speech tagging, lexical-based rules and distance-based entity name merging. All steps are implemented using free software and taking advantage of various existing packages. The process is exemplified with a book about the Portuguese Freemasonry due

not only to its important connections to political organizations and individual actors [15] but also to explore the potential use of those available data mining tools to identify and target social networks or central network actors [13] 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.


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 (top panel) and the results after scanning with optical character recognition (bottom panel).

This paper is organized as follows. In Section 2 the entity extraction 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 presents some remarks and draws some conclusions.

>> 2. THE ENTITY EXTRACTION PROCESS

In this section we describe 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. The goal of this work is to extract named entities from Portuguese texts taking advantage of various text mining methods and techniques implemented and made available on free software. Hence, the main program was developed in **R** [12] and makes use of some specific text mining packages, as described later.

In our case, we have obtained the text from a book scanned with current OCR software 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 expected, a preprocessing step is necessary to remove extra lines, extra spaces between words or letters and page numbers, *i.e.*, it is necessary to remove some junk before proceeding to the information extraction process.

Preprocessing step.

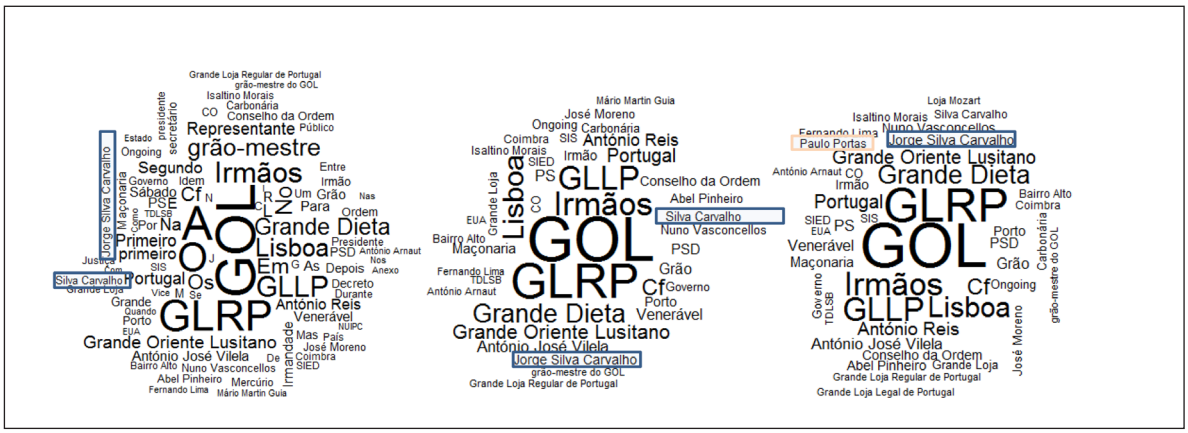
The R program was developed based on a pattern matching strategy and makes use of the following R packages: **tm** [7], **gdata** [16], **stringr** [17] and **memoise** [18].

Information extraction step.

In this work, entities are persons, places, institutions and organizations or associations. In this step we describe 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 designed 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 improving 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 an entity name.

- **Phase 1.** Since entity names frequently start 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 R packages: **tm** [7], **cwhmisc** [9] and **memoise** [18].
  For the paragraph in Figure 1, this phase identifies '*Irmandade do Bairro Ut O*', '*Conhecemos*', '*Parlamento do G*', '*L*', '*K*', '*Jorge Silva*', '*Ian*' and '*ministro Miguel Relvas*' as the candidate terms to named entities.
- **Phase 2.** In this phase, the list of identified terms (candidates to named entities) in 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 removes the first word from the ones having a 'prop' 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 nouns, and the second rule is to clean words that do not belong to the tags' structure of that entity name. This phase uses the following R packages: **tm** [7], **memoise** [18], **openNLP** [10], **Hmisc** [8].
  As a result of the application of this procedure to the previous list of terms, it is reduced to five terms, namely '*Irmandade do Bairro Ut O*', '*Parlamento do G*', '*Jorge Silva*', '*Ian*' and '*ministro Miguel Relvas*'. In fact, these five terms are the only named entities in the paragraph.
- **Phase 3.** This phase aims at associating different designations of the same entity, which are used throughout the text. For instance, '*Paulo Portas*' and *Portas* are two names used to refer to the portuguese personality Paulo Portas. In this phase a distance-based entity name merging is performed, by computing measures of similarity for each pair of entities names. These pairs are those present in the paragraph under analysis and in the corresponding adjacent paragraph, i.e., the previous and the next ones. The final criterion (described in Section 2.1) takes into consideration measures of the similarity between the names and the words present in the name. This phase uses the following R packages: **tm** [7], **stringr** [17], **Hmisc** [8], **Matrix** [1], **memoise** [18], **memisc** [6].



Phase 1 Phase 2 Phase 3

Figure 2: Entity word clouds depicting the terms that appear 25 or more times in the text. Each subfigure represents the terms returned at the end of each phase of the process.

2.1 Criterion

Given a pair of entities t_1 and t_2 represented by

$$t_1 = \left[W_{11} \ldots W_{1n_i} \right], n_i = number\ of\ words\ in\ t_1$$

$$t_2 = \left[W_{21} \ldots W_{2n_j} \right], n_j = number\ of\ words\ in\ t_2$$

we consider the feature vectors for each one defined as

$$v_1 = \begin{pmatrix} d_1(t_1,t_1) \\ d_2(t_1,t_1) \\ d_3(t_1,t_1) \\ d_1(t_1,t_2) \\ d_2(t_1,t_2) \\ d_3(t_1,t_2) \end{pmatrix} \quad v_2 = \begin{pmatrix} d_1(t_2,t_1) \\ d_2(t_2,t_1) \\ d_3(t_2,t_1) \\ d_1(t_2,t_2) \\ d_2(t_2,t_2) \\ d_3(t_2,t_2) \end{pmatrix}$$

where

$$d_1(t_1,t_2) = \frac{\sum_{i,j} I_{(W_{1i}=W_{2j})}}{\min(n_i,n_j)}, \quad (1)$$

$$d_2(t_1,t_2) = \frac{\sum_{i=1}^{n_i} \sum_{j=1}^{n_j} d_L(W_{1i},W_{2j})}{n_i n_j}, \quad (2)$$

$$d_3(t_1,t_2) = \frac{d_L(t_1,t_2)}{\min(nchar(t_1),nchar(t_2))}, \quad (3)$$

where $I_{(W1i = W2j)}$ is 1 if $W_{1i}=W_{2j}$ and zero otherwise, *nchar(t)* represents the length of *t*, and $d_L(x,y)$ represents the Levenshtein distance between the strings x and y, which is defined as the fewest number of insertions, substitutions, and deletions required to transform one string into another string. To compute the similarity between the feature vectors (v_1, v_2) we consider the cosine distance defined by Equation (4).

$$dcosine(t_1,t_2)=1-cos(v_1,v_2) \quad (4)$$

A pair of entities, t_1 and t_2, is considered as different designations of the same entity if the following criterion is met:

$$\begin{cases} dcosine(t_1,t_2)<0.293 \\ d_2(t_1,t_2)<0.65 \\ \frac{dcosine(t_1,t_2)}{0.293}+\frac{d_2(t_1,t_2)}{0.65}<1.5 \end{cases}, \quad (5)$$

where the thresholds considered are based on:

- a maximum angle of 45˚ between the vectors
- some words must belong to the two entities' names.

2.2 Analysis of results

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

From this book, a total of 12120 events are extracted by 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 recorded in 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 obtained 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.

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
F1	0.72	0.87

This means that the recall, given by Equation (6), decreases from 0.84 to 0.78 and the precision, given by Equation (7), increases from 0.63 to 0.97 in the second phase. We do not compute the recall and precision for phase 3 since this phase does not changes the number of terms or named entities, i.e., the named entities identified are the same of phase 2.

$$recall = \frac{number\ of\ named\ entities\ extracted}{number\ of\ named\ entities\ present\ on\ the\ text}, \quad (6)$$

$$precision = \frac{number\ of\ named\ entities\ extrated}{number\ of\ terms\ extracted\ from\ the\ text}, \quad (7)$$

Equation (8) defines another measure used to quantify the quality of the process, F1. The second phase of this process increases the value of F1 from 0.72 to 0.87.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \quad (8)$$

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 remark is that words such as *Venerável* and *Grão* are ambiguous.

>> 3. NETWORK ANALYSIS

In this section we use the named entities, previously extracted from the whole book, and we analyse the relations between entities using Social Network Analysis (SNA). 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, not only the named entity (or entities) that is most frequently cited with other named entities, but also to detect groups or communities of entities.

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 visualizing the sub-network of entities with degree at least 83. This filter enables the representation of the most central entities according to the degree measure (i.e., those vertices with larger number of neighbours), while removing those entities that are less representative of the book content. Some of the names visualized in this graph are in the entity word cloud as well, but they do not all have the same importance. (see Figure 3)

3.1 Communities found

In this work, communities are found using the implementation of the Louvain method made available in the Gephi software [2]. 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

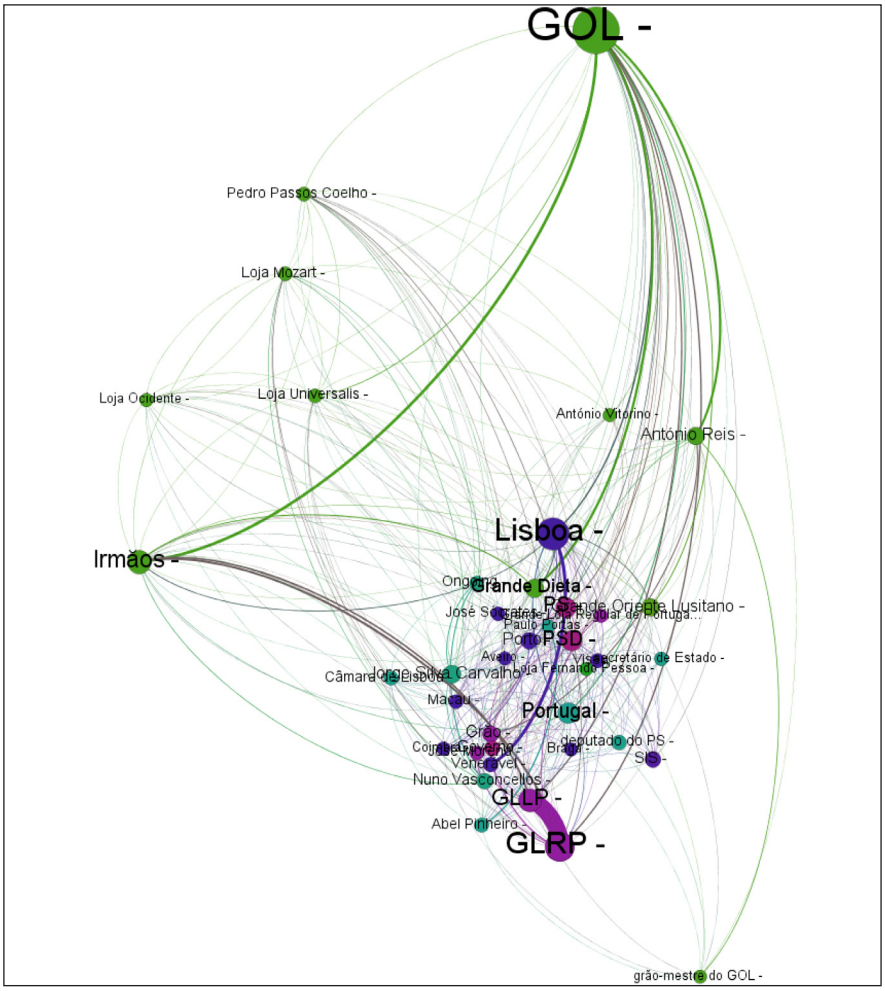


Figure 3: Social network graph filtered by degree - minimum 83.

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.

The entities are grouped in communities and the colours of each node are associated with the group it belongs to. Within the context of our application, this means that a community identifies 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 the six larger communities are represented by a square inside of which are their most central named entities. The thickness of the lines between the squares is proportional to the number of links between the 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*.

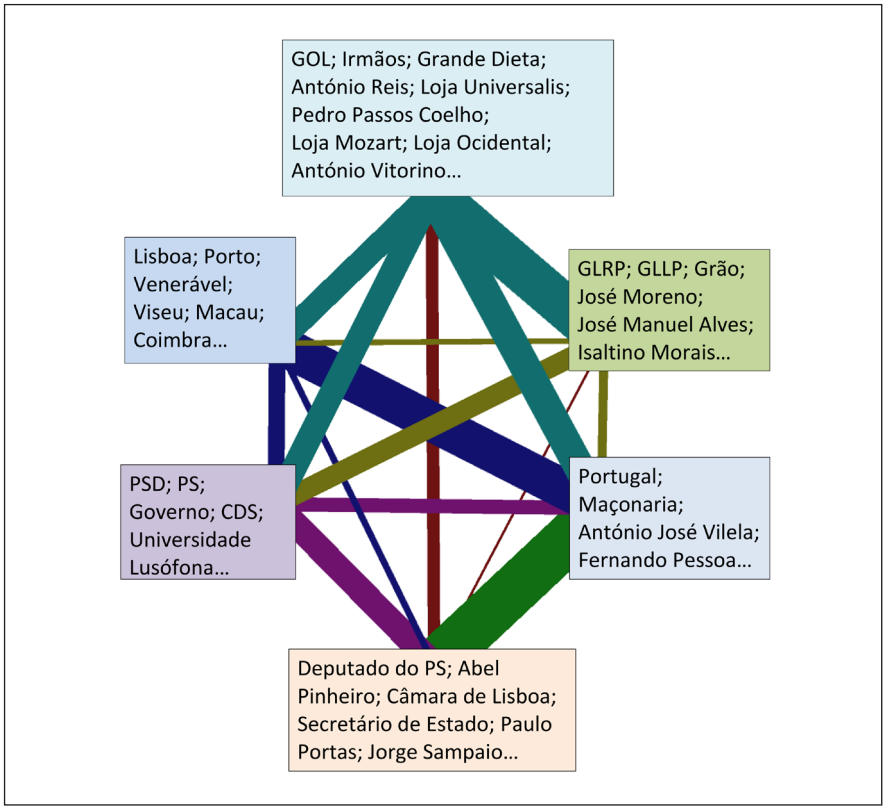


Figure 4: Higher-level representation of the previous graph highlighting identified communities and the corresponding entities; the thickness of the edges is proportional to the strength of the connection between communities.

3.2 Important nodes

After the detection of communities of named entities, we proceed to identify which are the most important entities in the book. This importance is measured with respect to two different types of centrality: betweenness centrality and eigenvector centrality.

Betweenness is defined as the number of shortest paths between two arbitrary nodes that pass through the node under analysis, measuring the extent to which a node lies between other nodes in a network. 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. Unlike

degree centrality, the eigenvector centrality value measures how well a given node is connected to other well-connected nodes in the network, by taking into account both direct and indirect influences. Hence, according to the betweenness and the eigenvector centrality values, *GOL (Grande Oriente Lusitano)*, *Lisboa* (location) 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* and *PSD* (political parties), *SIS* (secret services), *Nuno Vasconcelos* and *António Reis* (politicians), *Jorge Silva Carvalho* (spy), *Portugal* and *Porto* (locations).

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

3.3 Discussion

Some statistical measures that characterize the network are summarized in Table 3.2. 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 [3].

Figure 5 depicts the network graph for the main community, i.e. the community comprising the larger number 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.

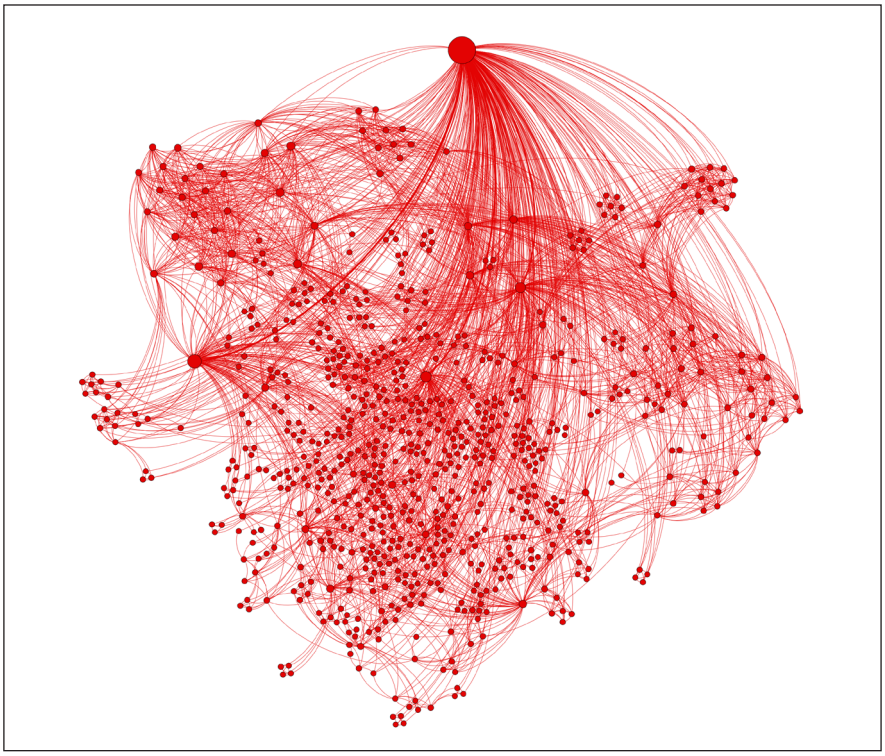


Figure 5: Social network graph for the larger cluster - the main community.

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.

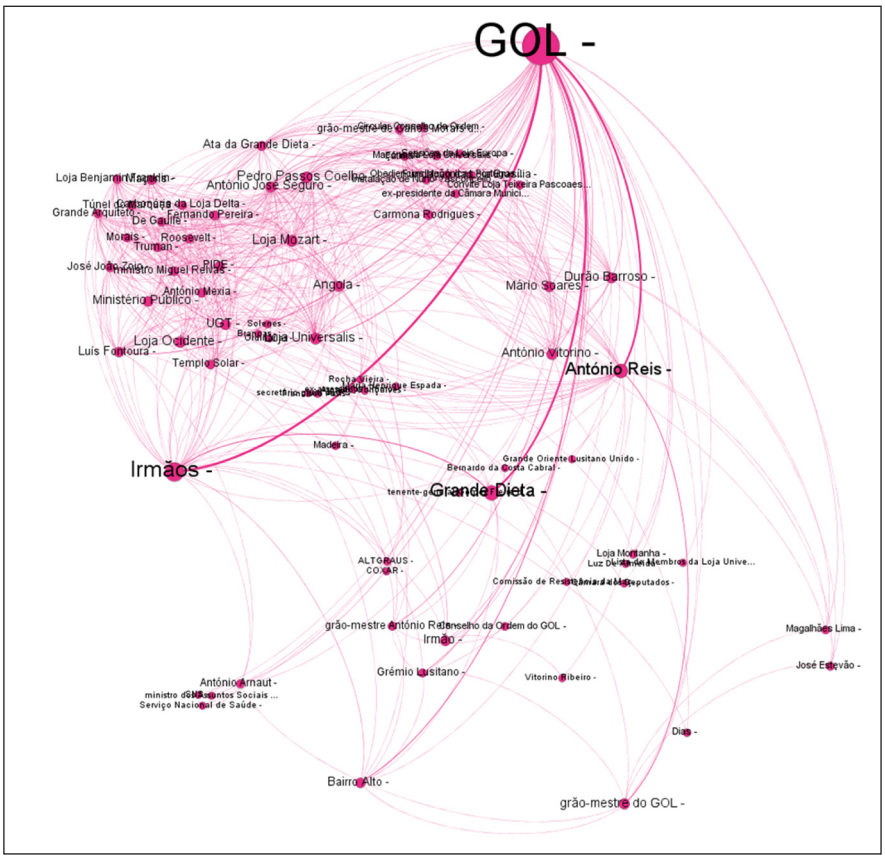


Figure 6: Social network graph for 10% of the larger cluster. Nodes' size is proportional to their degree centrality.

>> 4. REMARKS AND CONCLUSIONS

The proposed 3-phase method to extract named entities from unstructured text was successfully implemented using free software. The preprocessing step was applied to overcome limitations of the OCR process and may not be necessary when processing text from other sources. The inclusion of the POS tagging into the process allows improving the quality of the extraction process since F1 increases from 0.72 to 0.87. Although phase 3 does not affect F1, the recall and the precision measures, the inclusion of this phase allows improving the quality of the extracted information as well as the social network analysis. 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 uncover 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 disambiguation step, as well as by adjusting the network model so that entities are linked based on the verbs that occur in the sentences.

REFERENCES

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

[02] 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.

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

[04] 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.

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

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

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

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

[09] C. W. Hoffmann. *cwhmisc: Miscellaneous Functions for math, plotting, printing, statistics, strings, and tools*, 2013. R package version 4.0.

[10] K. Hornik. openNLP: *Apache OpenNLP Tools Interfac*e, 2014. R package version 0.2-3.

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

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

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

[14] 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.

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

[16] 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.

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

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