FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Constructing Taxonomies Using Results From Portuguese News Articles Topic Distillation

Rui Mário Seixas Teixeira

Master in Informatics and Computing Engineering

Supervisor: Luís Sarmento (PhD)

Supervisor: Eugénio Oliveira (PhD)

July, 2011

Constructing Taxonomies Using Results From Portuguese News Articles Topic Distillation

Rui Mário Seixas Teixeira

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

Chair: Name of the President (Title)
External Examiner: Name of the Examiner (Title)
Supervisor: Name of the Supervisor (Title)

31st July, 2011

Abstract

This work presents a method for establishing hierarchical relations between news article topics. This is accomplished using notions from Concept Analysis and Graph Theory and the application of the Knowledge Discovery in Data methodology (KDD).

A project called Verbatim produced methods to automatically classify Portuguese news articles with tags that had been extracted from the news media. This project produced a large and growing dataset, containing the news-tag associations, which was used for our experimental work.

Following the KDD methodology, preprocessing steps were taken to reduce the complexity of the dataset and remove inaccurate entries.

An algorithm to mine the relations between the news tags was designed. This algorithm looked for topic->subtopic associations and its application to the Verbatim dataset resulted in a set of topic->subtopic pairs that could be mapped in a news topic graph. The analysis of this graph revealed a large number of shallow, mostly independent subgraphs. Each of these subgraphs was considered to represent the topic->subtopic hierarchy of a single topic. By selecting the sources in each subgraph and reconstructing them as a tree, the taxonomy for each root topic was produced.

All of the topics that were found to be sources were considered super-topics studied to ascertain their similarities. This was done using an algorithm which redefined the root topics as sets of tags and compared them, calculating the overlap of these sets. Several graphs were constructed, by adjusting the similarity criterion, in which topics were linked.

The taxonomies of each topic, although shallow are usually accurate and the topic similarity graph provided a novel view and a deeper understanding of the relations between news topics.

Resumo

Este trabalho apresenta um método para estabelecer relações hierárquicas entre tópicos de notícias apoiando-se em noções de Análise de Conceitos e de Teoria dos Grafos e atraves da aplicação da metodologia *Knowledge Discovery in Databases* (KDD).

Como parte de um projecto chamado Verbatim, foi desenvolvido um método para classificar automaticamente notícias Portuguesas com *tags*. Este método produziu o conjunto de dados que foi usado para a componente experimental deste trabalho.

Seguindo a metodologia de KDD, foram tomados passos de pré-processamento para reduzir a complexidade do conjunto de dados e remover entradas incorrectas.

Um algoritmo para estudar as relações entre as *tags* das notícias foi projectado. Este algoritmo procurou associações entre tópicos e sub-tópico de notícias. A aplicação deste algoritmo ao conjunto de dados Verbatim resultou em um conjunto de pares (tópico -> sub-tópico) que podiam ser mapeados num grafo. A análise desse grafo revelou um grande número de sub-grafos independentes. Cada um destes sub-grafos representando a hierarquia entre de sub-tópicos de um único tópico. Ao seleccionar as fontes de cada sub-grafo e construindo-os como uma érvore, a taxonomia para cada tópico foi produzida.

Todos os tópicos que foram identificados como fontes foram considerados super-tópicos foram estudados para verificar suas semelhanças. Isso foi feito usando um algoritmo que redefiniu os super-tópicos conjuntos de *tags* e comparou-os, calculando a sobreposição desses conjuntos. Várias experiências foram realizadas, ajustando-se o critério de similaridade.

As taxonomias de cada tópico, embora pequenas são normalmente precisas e o grafo de semelhanças proporcionou uma nova visão e uma compreensão mais profunda das relações entre os tópicos de notícias.

"A little knowledge is a dangerous thing. "

Alexander Pope, (1688-1744)

Contents

List of Figures

Abbreviations

CRISP-MD	Cross Industry Standard Process for Data Mining
IPTC	International Press Telecommunications Council
KDD	Knowledge Discovery in Data
NN	Nearest Neighbor
RSS	RDF Site Summary
SVM	Support Vector Machine
Web	World Wide Web

Chapter 1

Introduction

In the information age, knowledge is hard to find. Our ability to generate and collect content has been increasing rapidly, well beyond our ability to process it [Han06]. As such, on the Internet, content and metadata are becoming inseparable from one another. Aiming to facilitate user access to content, metadata-based navigation has become a staple approach, as a means to cope with ubiquity of the Internet and it's overwhelming supply of content. Because of social media websites, such as YouTube.com, Gmail.com, linkedin.com, stackoverflow.com users now have high expectations regarding content, accessibility and navigation.

1.1 Context

In the Verbatim project, a SapoLabs [1] project now called voxx (Figure 1.1), a set of automatic software tools for information structuring and filtering was developed. Verbatim [2] is a system for acquiring information from live news feeds and extracting quotes and topics. This was created with the intent of providing a sort of a personal information "butler"[SNO09]. This system provides a quote extraction and browsing service aiming to ease the news articles information overflow. To facilitate the browsing and navigation of the extracted quotes, these quotes needed to be labeled with news article topic tags. As the news articles did not contain useful topic metadata this had to be extracted too.

The tag extraction process was achieved through the automatic mining of topic tags, using a lexical pattern present in some news articles - *<Topic Tag>:<News Title>*. This *<Topic Tag>* was taken as the adequate news tag. This pattern was not present in all articles and when present only one tag could be extracted. The articles for which this

[1] http://labs.sapo.pt/
[2] Available online at http://voxx.sapo.pt/

Figure 1.1: Homepage of voxx (ex-Verbatim), July 2011.

pattern was not present were given the same tag that similar articles had. This was done by classifiers trained using the original tagged articles. This process was further improved by a document clustering method - nearest neighbor. This was done to enhance the quality of tagging non-tagged documents and also to provide additional tags for the others [SNTO09].

From an information extraction perspective this resulted in a large and growing dataset of tagged news articles. This dataset, which provides multiple tag associations for each news item, is the basis for the work presented in this document. The dataset contains all the *tag - news* associations constructed by three methods - NN, SVM and Rocchio - as well as the score attributed to the association. In Figure 1.2 the *tag - news* pairs for a single news item are shown, as an example. The existence of multiple tags for each news item is a key point in this work. However the example shown reveals one of the major

problems that will be addressed throughout this work - how to determine the relevance of a given tag to the news content.

news_id	method	tag	value	news_title
772642	NN	Cavaco	0.0610563	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	Cavaco	0.06003	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	Justiça	0.0495682	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	Ensino particular	0.0474579	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	PSD	0.0447437	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	PS reage	0.0431733	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	BPN	0.0417574	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	PSD	0.0392989	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	BPN	0.038222	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	NN	Lacão	0.0367607	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Itália	-0.991309	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Real	-1.04528	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Futebol nacional	-1.05323	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Académica	-1.05534	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Liga Zon Sagres	-1.05913	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	SVM	Costa do Marfim	-1.10878	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	Rocchio	Liga Orangina	0.0188679	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"
772642	Rocchio	Liga Orangina	0.0188679	Cavaco Silva: "Para serem mais honestos do que eu têm que nascer duas vezes"

Figure 1.2: Example database entries for a single news item.

1.2 Problem Overview

Initially the challange was to find ways to improve the voxx news classification system exploring relations between tags. In a way, this meant trying to find out topics and subtopics as a result of a tag hierarchy, assuming that higher level tags in that hierarchy were broader topics, news wise. Eventually this could be used to better train the classifiers and ultimately to achieve an improved overall performance of the tagging system. While seeking this, the following question arouse:

How accurate is a tag classification?

Not all *tag - news* associations are equally valid or useful. In Figure 1.2 some tags are clearly not adequate, for instance, "Académica, Futebol Nacional, Liga Orangina, Liga Zon Sagres, Costa do Marfim, Itália". In order to take further advantage of the generated tags, it was necessary to have a better understanding of the *tag* and the *tag - news* associations, particularly how to distinguish correct and incorrect associations without manual supervision (detailed in Chapter 3). This leads to the next question:

How to deal with versions of the same tags?

Due to the free form nature of the tags available, multiple versions of the same tag appeared (at least 14% of tags, using a conservative point of view). These duplicates have a negative impact and should be reduced as much as possible. This was done without sacrificing genuine tags using a simple unsupervised spell correction method based on the

edit distance (detailed in Chapter 4). Having done this optimization, the work proceded to address the next question:

How can tag accuracy be predicted?

First a baseline had to be established. This was acomplished through manual validation (detailed in Chapter 3). Human evaluators were asked to evaluate random samples from the dataset. This process provived a detailed view on the behavior of the classifiers and ultimately permited to predict their accuracy (detailed in Chapter3) and consequently eliminate all the inadequate entries (73% of the original entries). The resulting dataset, although smaller, had a higher and known classification accuracy:

How can a taxonomy be constructed?

By defining a tag by the set of news for which it was correctly associated and considering that more general tags are those that are associated to a larger set of news, the notions of topic and subtopic are created as entities in a hierarchy. Through association of these entities, hierarchycal trees can be contructed. In these trees the root nodes can be considered broad topics, similar to the common language meaning of a topic (i.e Futebol, PEC, Líbia, etc)(detailed in Chapter4).

How can the constructed taxonomies be validated?

There are basically 4 ways of evaluating taxonomies:

- By comparing them to other pre-established same subject taxonomies (i.e. using a gold standard);
- Through manual evaluation performed by domain experts;
- By perfoming a statistical evaluation of the taxonomy through its structural features;
- By doing application-driven evaluation, in which a taxonomy is assessed on the basis of the improvement its use generates within an application.

The first alternative is generally the desired one, but in this case a suitable taxonomy to be used as a gold standard could not be found. As such, taxonomies were evaluated through manual validation, namely assessing the number of correctly connected nodes, that is, the number of nodes that correctly belong to their root nodes, the number of nodes that although incorrectly connected do belong to their root nodes. Several other structural aspects were evaluated, such as tree size, tree depth and news coverage for a given tree. The final method of evaluation is also the ultimate goal of this work, that is, improving the voxx news classification system.

1.3 Motivation

The Verbatim dataset presents the possibility of using the news classifier results in a novel way - using the tags (automatically generated and associated to news), to construct topic hierarchies.

While there is much work done in the construction of topic hierarchies from texts, the idea of constructing topic hierarchies from the automatically generated *tag - news* pairs for Portuguese news articles is a previously unexplored one.

This approach seeks to address several recurring issues in this field such as classifier evaluation, tag disambiguation, concept clustering, taxonomy construction and evaluation which constitutes very stimulating challanges. The possibilty to push the boundaries of any of these issues, even a little bit, is its own reward.

Chapter 2

Background and Related Work

This chapter introduces the academic background which sets the boundaries for the problems addressed in this work. The technical concepts and methodologies used on this approach are also presented. These definitions will be applied mostly within the scope of this work, and will be, simplified, working definitions.

2.1 Knowledge Discovery or Data Mining

There is some confusion around the term Data Mining, which is often used interchangeably as the task of discovering interesting patterns from large amounts of data or as the methodologies that incorporate this task.

To clarify, in the context of this work, Data Mining will be considered to be one of the steps in the knowledge discovery process models, such as in KDD or CRISP-DM among others. All these process models, while distinct, can be summarized into 3 underlying steps:

- Pre-processing
- Data mining
- Results validation

These 3 steps are also followed throughout this work providing a basis for the structure of this document.

2.2 News Topic Classification

Classifications are made to help the human or the program to structure or to increase the usefulness of information. A classification is a ready-made or evolving structure, much like a collection of "labeled boxes" in which the information is placed [vR03].

There are several distinct approaches currently in use for news topic classification.

One of these, is an effort to employ industry supplied standards *NewsCodes* - a facet based controlled vocabularies, used to categorize news content and in the IPTC news exchange formats. *NewsCodes* is a set of terms that expresses facets of news content. Facets could be the subject, the genre, the urgency, etc. However this is not widely used yet, particularly outside news providers.

Another, more recent, emerging aspect of news classification is the use of *social tags* (keyword annotations)[HRGM08]. These are a popular way to allow users to contribute with metadata to these large and dynamic corpora. *Social tags*, in variance with industries or academic standards, are free form, independent from any kind of preconceived categorization. This makes them appropriate for domains in which the content may change very rapidly, particularly web media. However social tags, as a recent phenomenon are not yet well understood. This gives rise to new research, seeking to comprehend social tag behavior, and related problems such as cold strapping (niche tags, that became popular overwhelm, obscuring other useful tag, increasing artificially the semantic complexity)[HRGM08].

2.3 Taxonomy

"Taxonomy is a hierarchy created according to data internal to the items in that hierarchy." [vR03]

Originally taxonomy referred to the orderly classification of plants and animals according to their presumed natural relationships. A clear example of taxonomy, in this sense, is the animal kingdom taxonomy. Kingdom "animals", "mammals", order "carnivores", genus "canis", species "canis lupus", which is the common gray wolf. Other members of the genus "canis" are the dog and the jackal. This is a taxonomy based on the presumed "is a kind of" relation.

A Taxonomy is a structure according to some relation between the entities. Taxonomic entities are classified in a hierarchical structure according to a specific relation between them. Then a taxonomic relation is a relation between entities in the taxonomy, for instance the relation between class and subclass or, more pertaining to this work the relation between news topic and new subtopic.

2.4 Taxonomy Construction

Taxonomy construction seeks to answer the need for some pre-existent knowledge between lexical terms when addressing semantic relations. Additionally even established taxonomies fail to be applicable outside their domains of creation.

As such, efforts to establish sets of rules between concepts, usually in a domain-specific knowledge space, are a common link between most of the work publish in the area. These rules are sought for creating networks of concepts, and ultimately taxonomies, which can then be used as background knowledge. This can be used as a stepping stone for approaching more specific problems (e.g. contracting multilayer hierarchical classifiers) or as a gold-standard in future work.

In one of the earliest works that helped define this domain Hearst [Hea92] established a method for the automatic acquisition of the hyponymy (IS-A, e.g. rabbit IS A mammal) lexical relation from text by identifying hyponymy relations through the use of lexico-syntactic patterns. Besides hyponymy, other lexical relations can also be acquired using the same method.

However, Sanderson [SC99] pointed out that, manual intervention was required for this discovery and the scope of the noun phrase pairs identified was limited. And also noted that other, simple, unsupervised methods such as term co-occurrence and document frequency could produce maps of related terms. In the same work a method of automatically deriving a hierarchical organization of concepts from a set of text documents without the use of training data or clustering techniques. In this words and phrases extracted from the documents are organized hierarchically using a type of co-occurrence known as subsumption. These methods were used to sort documents in a hierarchical structure according to their semantic context.

New methods used training sets of positively marked documents to constructed category classifiers, such as *rocchio* or *svm - support vector machines* [DC00]. Most of these approach resulted in flat categories, that would mean a complete independence between categories. As the number of categories tends to grow significantly large, this provides difficulty in browsing and searching the categories. This can be solved using hierarchical classification [SL01].

In "Hierarchical Text Classification and Evaluation" the objective was somewhat similar to the ultimate goal of this work. The objective was to construct a hierarchy of categories for improving the organization structure of text documents. The texts were organized in flat categories. These are categories that are predefined and are treated in isolation (i.e. there is no structure defining the relationships between them). As the number of categories increased, text browsing became a problem. By evaluating the similarities between categories they devised a method of structuring them in a hierarchy. Altough in a diferent context, the problem presented seemed to have considerable similarities to this work and

was important to the decision of persuing an hierarchycal approach. They proposed the use of additional classifiers to determine if a doccument should belong to a category or one of its subtrees. This method was tested by contructing a top-down level based hierarchical classification. And evolved by means of performance measures for evaluating semantic relationships and parent-child relationships among categories in a hierarchy [SL01].

Recently, with the advent of web2.0 and *social tags*, there has been an abundance of tagged data. This has led to the establishment of a new probabilistic data mining approach to these issues. These *social tags*, despite their free-form nature, tend to be used in limited sets for a particular object type [PL08]. A generative probabilistic mode for extracting taxonomies from social tags was proposed [PL08]. An example of using user specified tags from *Flickr* to construct taxonomies can be found in Plangprasopchok [Pla09]. These taxonomies, created from *social tags* can be evaluated through comparison to reference taxonomies, providing a fully automated taxonomy constructed method. However several issues were identified, this approach resulted in taxonomies which were: sparse, shallow, ambiguous, noisy, and inconsistent. These issues led to the refinement of the probabilistic approach, for instance, by including user specified hierarchical relations (e.g. folder and subfolders) and applying probabilistic affinity propagation for dealing with the user metadata attribute propagation [PLG11]. Another approach relied on enriching the data with metadata provided through the use of highly ranked web-search engine results[CC04].

2.5 Conclusions

While there are lexical resources that provide the semantic relations among words (e.g. WordNet [MBF+90], FrameNet, UNL) these are not complete or universally available to all languages. Moreover these resources become less valid as a domain grows in specificity and languages evolve [SM07]. Nonetheless their use as a gold standard for evaluation remains invaluable and extremely disseminated [HKR09] [Kor02].

Concerning the Portuguese language, altough there are several ongoing projects to increase the number of lexical resources, such as PAPEL [OSGS08], TeP [DdsM03], MultiWordNet [MAC+06], they have not yet matured to a WordNet, preventing their use as gold standards. Moreover *news tags* are mostly names, organizations and events, which are not usually found in theses resources.

This lack of a gold standard constitutes an issue that hampers the automatically evaluation of lexical-semantic systems.

It has been proved taxonomies can be constructed based solely upon metadata, such as *social tags*, user specified relations, etc.. This is particularly important because, in this work, news topic taxonomies are constructed using only news articles tag information without conducting any lexical analysis.

Chapter 3

Preliminary Dataset Analysis

This section describes the data level analysis processes that were conducted. This was done to establish a baseline level of understanding of the data. In a lesser degree, this was also done to improve the understanding of Portuguese news stories as a data domain.

When using KDD, an in-depth understanding of the Dataset is very useful. As noted in [Han06], background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction. As such, it becomes an even greater necessity to conduct this preliminary analysis. This way we can assert which data cleaning and integration techniques will be required.

As mentioned in Chapter 1, in Section 1.1 and Section 1.2 the dataset under consideration was generated from the work done in Verbatim and Voxx [SNO09] [SNTO09]. Verbatim, a system for acquiring information from live news feeds and extraction of quotes and topics, made use of several classification techniques in order to associate and propagate *tags* in an effort to classify news stories extracted hourly from Portuguese news RSS feeds.

Another objective of this Chapter, along with acquiring a better understanding the of data, is the collection of statistics. These statistics will be used to establish the baseline for the preprocessing steps that are to follow.

Some of the more important aspects of this analysis was the study of the *tag*, *value* attributes, and how these attributes relate to one another.

3.1 Dataset Overview

The dataset under study contains 175.814 entries corresponding to the data acquired from 10.804 news articles, selected from 2010-12-23 to 2011-04-10. Each entry represents the association between a news article, a *tag* assigned to that news article, a news article identification number and headline, and some details (method name and classification

score) about the method used to associate that news article with that tag. Some entries also have the news's original *tag*. It is important to note that these *original tags* were used to train the classifiers in Verbatim.

3.2 Tags

A *tag* follows a very loose definition, a *tag* is a string of characters which contains any of the following: alphanumerical characters, symbols, punctuation or white space.

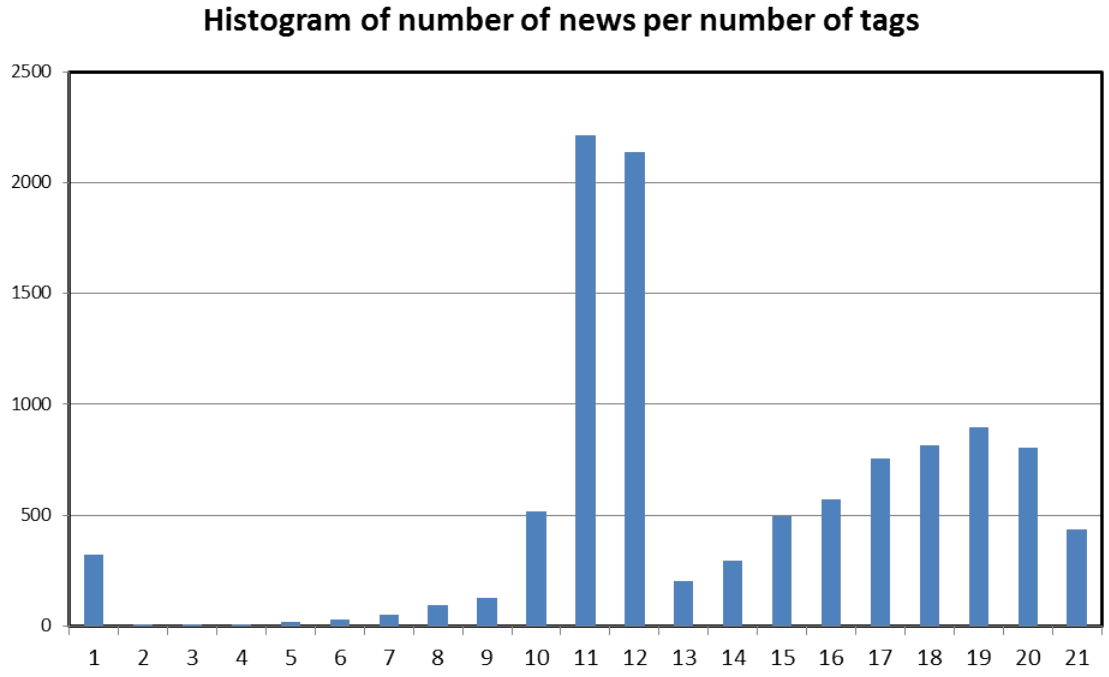


Figure 3.1: Histogram showing how many news there are for how many tags.

There are 4.260 distinct *tags* in the dataset, of these, less than 300 (around 6%) account for 75% of the 175.814 total occurrences.

The work done in verbatim suggests that minor spelling variations (e.g. misspellings, inconsistent spelling conventions) might account for a significant amount of distinct *tags*. *Tags* are extracted from professional news text, because of this these minor spelling variations should not occur frequently.

The distribution of *tags* (Figure 3.2), shows a great number of very uncommon *tags* (2.065 *tags* appear 3 times or less). These minor spelling variations should be partly responsible for the long tail which is found on the *tag* distribution.

The top 20 (Figure 3.3) most frequent *tags*(close to 0.5% the *tag* space) represent approximately 18,6% of all *tag* occurrences 3.2.

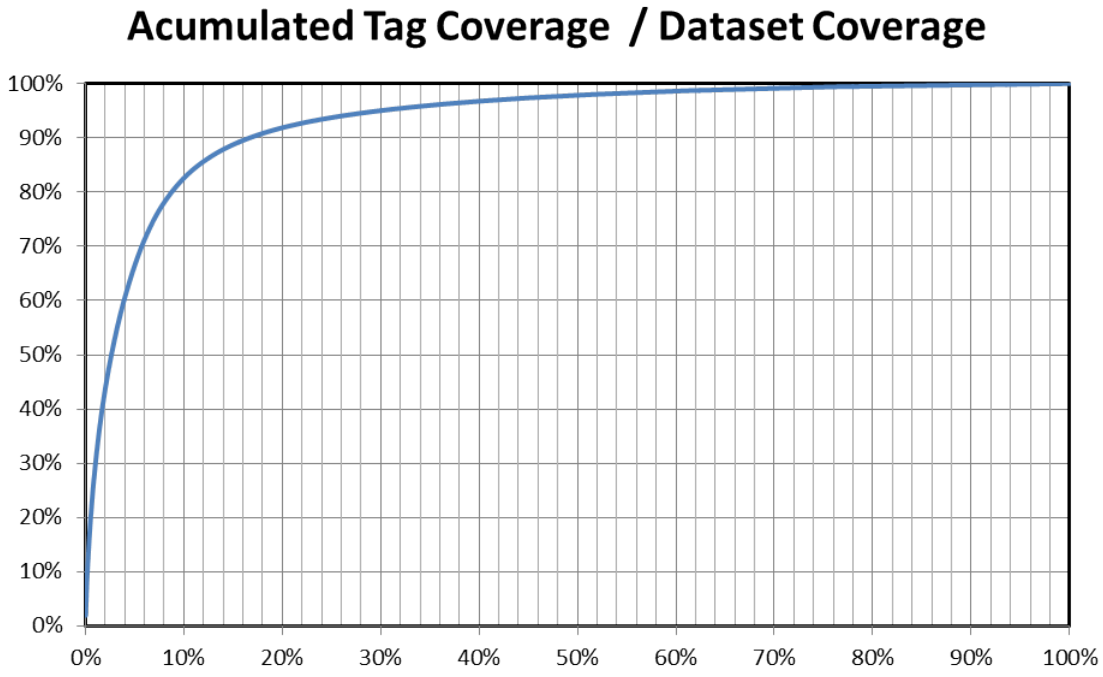


Figure 3.2: Accumulated histogram representing the distribution *tags* expressed in percentage of unique *tags* over dataset coverage.

3.3 Original Tag

The *original tag* is a particular case of *tag*. *Original tags* are assigned to news *a priori* [SNO09] and are considered to be. Having this in consideration, entries of the dataset in which the *tag* is identical to its *original tag* are considered duplicate entries and will not be used in the following steps.

3.4 News Classification

Each entry in the dataset is a single unique instance of news classification (i.e. each entry represent the classification of a news article with a single *tag*). As such, it is of the utmost importance to have a clear understanding of this aspect of the data. Associated to each classified *tag* there is a *method* label and a value obtained from the classifier. This value is expected to be an indicator of positive identifications. Because the dataset contains *tags* generated from distinct classification methods, the values accrued from this process must be evaluated separately.

The following is dedicated to understanding how the *values* behave in the dataset. The details of these methods will not be under study, as it does not appear to be relevant to the work at hand. The *method* is regarded simply as a label - as a modifier for the

#	Tag	Occurrences	#	Tag	Occurrences
1	Sporting	3451	11	Energia	1346
2	Futebol	2721	12	Marcelo	1335
3	Líbia	2427	13	Crise	1302
4	PEC	2261	14	Basquetebol	1280
5	Mercado	1898	15	Camarate	1279
6	Madeira	1757	16	Quiosque	1261
7	Governo/Demissão	1554	17	Académica	1129
8	PSD	1505	18	Portugal/Brasil	1114
9	Rio Ave	1454	19	Egito	1105
10	Brasil	1412	20	PSP	1096

Figure 3.3: The top 20 most popular tags, and the number of times each one occurs.

classification value. However, the *methods* have distinct *value* ranges and distribution which makes it necessary to divide the dataset accordingly.

The dataset contains results from the following 3 different methods. Descriptive statistics and histograms (Figure 3.4, Figure 3.5 and Figure 3.6) were generated for each subset as a means to help establish a baseline for optimization:

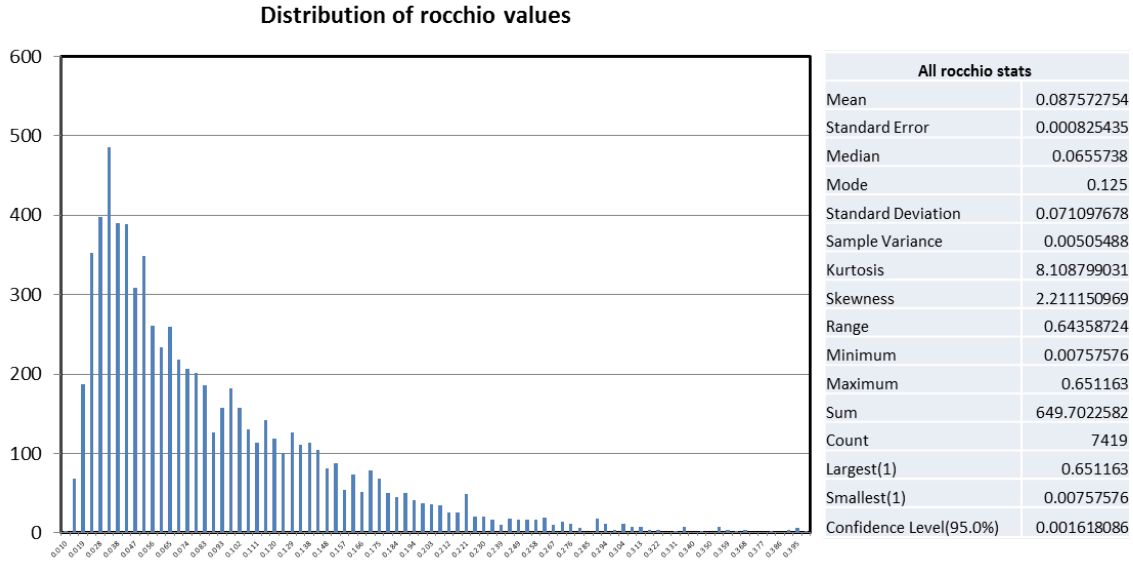
- *Rocchio classification* [Roc71], or *rocchio*: accounts for 12.307 (7%) of the 175.814 entries in the dataset;
- *Nearest Neighbor* [SNTO09], or *nn*: accounts for 62.767 (35.7%) of the 175.814 entries in the dataset;
- *Support Vector Machines* [Joa98], or *svm*: accounts for 100.740 (57.3%) of the 175.814 entries in the dataset.

As can be seen in the rocchio distribuition, it appears that some degree of truncation may have occurred.

In Figure 3.6 we can see a group of 4000 entries with *value* = 1. When evaluated manually these entries were considered to be all positive and, consequently, were given the same status of the *original tags*.

The *svm* (Figure 3.5) method provided the most entries and the smallest variance of the 3.

Methodologies for comparing classifiers are an important part of data mining research. The paper "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach" [Sal97] presents a set of methodological notes which were generally followed in this work. In order to compare the *methods* in the dataset, a simple criterion was chosen: the average accuracy of classification of each *method*. In other words, for a given method, what is the probability of selecting a correct *news* - *tag* classification if a random entry is selected?



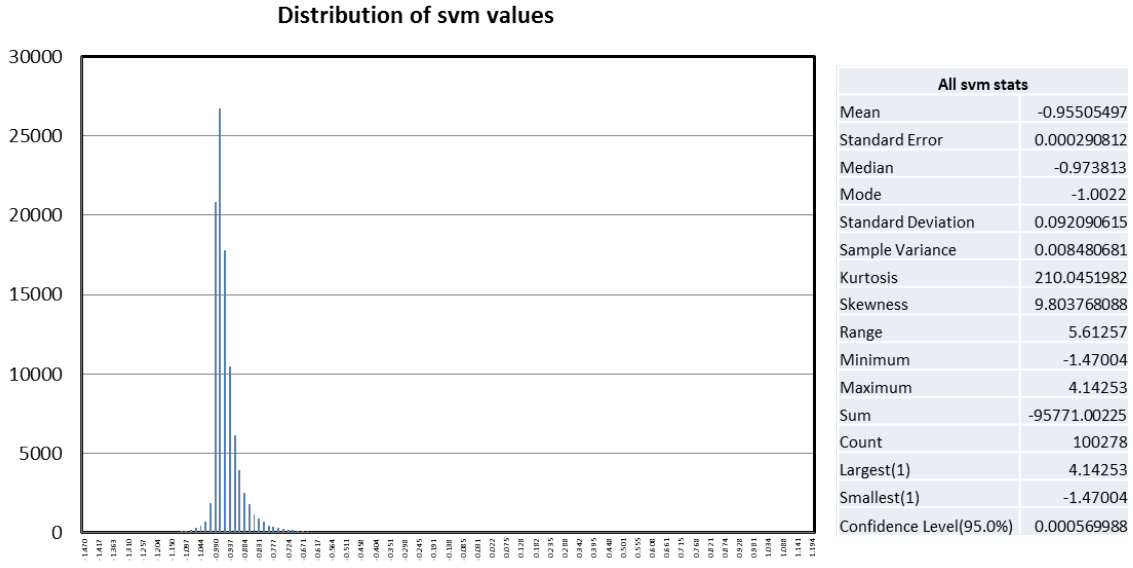
All rocchio stats	
Mean	0.087572754
Standard Error	0.000825435
Median	0.0655738
Mode	0.125
Standard Deviation	0.071097678
Sample Variance	0.00505488
Kurtosis	8.108799031
Skewness	2.211150969
Range	0.64358724
Minimum	0.00757576
Maximum	0.651163
Sum	649.7022582
Count	7419
Largest(1)	0.651163
Smallest(1)	0.00757576
Confidence Level(95.0%)	0.001618086

Figure 3.4: Histogram representing the *rocchio* values distribution and descriptive statistics.

One of the goals of this step is to estimate the average accuracy for each classifier. A test was designed based on the knowledge that the average accuracy is also the estimator of the proportion of *positive* observations. Because the observations are independent, this estimator follows a (scaled) binomial distribution. For a sufficiently large number of samples, the distribution will be closely approximated by a normal distribution with the same mean and variance as the average accuracy's. Using this approximation, it can be shown that around 95% of this distribution's probability lies within 2 standard deviations of the mean [NIS11].

In practical terms, this test consists in extracting random samples from the subsets and manual evaluate them as correct or incorrect classifications. This manual evaluation is subjective in nature as it depends on the knowledge of current events of the observers and on their subjective beliefs. In order to attempt to achieve accurate results the following guidelines were established:

- The value of the *tag-news* classification was hidden from the subjects (this was done in an effort to avoid confirmation bias);
- Subjects were asked to identify the *nature* of a news article by mean of analyzing its title;
- If the subjects could not positively identify the *nature* of the news through the use of its title they were given access to the news text;
- Subjects were then asked to decide, based on their previous experience, if given a certain news title, the associated *tag* could be used to locate that news article on the web;



All svm stats	
Mean	-0.95505497
Standard Error	0.000290812
Median	-0.973813
Mode	-1.0022
Standard Deviation	0.092090615
Sample Variance	0.008480681
Kurtosis	210.0451982
Skewness	9.803768088
Range	5.61257
Minimum	-1.47004
Maximum	4.14253
Sum	-95771.00225
Count	100278
Largest(1)	4.14253
Smallest(1)	-1.47004
Confidence Level(95.0%)	0.000569988

Figure 3.5: Histogram representing the *svm* values distribution and descriptive statistics.

- The samples were then evaluated as positive or negative accordingly.

For these tests, the sample size was set at 200 samples for each *method* subset. This sample size was chosen in an effort to avoid using a prohibitively large sample set and still achieve a high degree of confidence (approximate 92.5%).

Some examples from a sample test:

- *tag:* Ciência - *news title:* "Estacionar o carro e usar o metro e a Carris vai custar 49 euros por mês em Lisboa" - *manual evaluation:* false
- *tag:* Turismo - *news title:* "Aposta reforçada na promoção dos Açores" - *manual evaluation:* true

All manual validation tests, used to measure classifier accuracy, can be found in Apendix A - Human Classifier Validation A.

We can see from this analysis (Figure 3.7) that the *method rocchio* provides the highest average classification accuracy at 51% $\pm$ 5% followed by *method nn* with 42% $\pm$ 3% and *method svm* with 12% $\pm$ 1%. Given the low deviation of the svm method values its accuracy might be hard to improve.

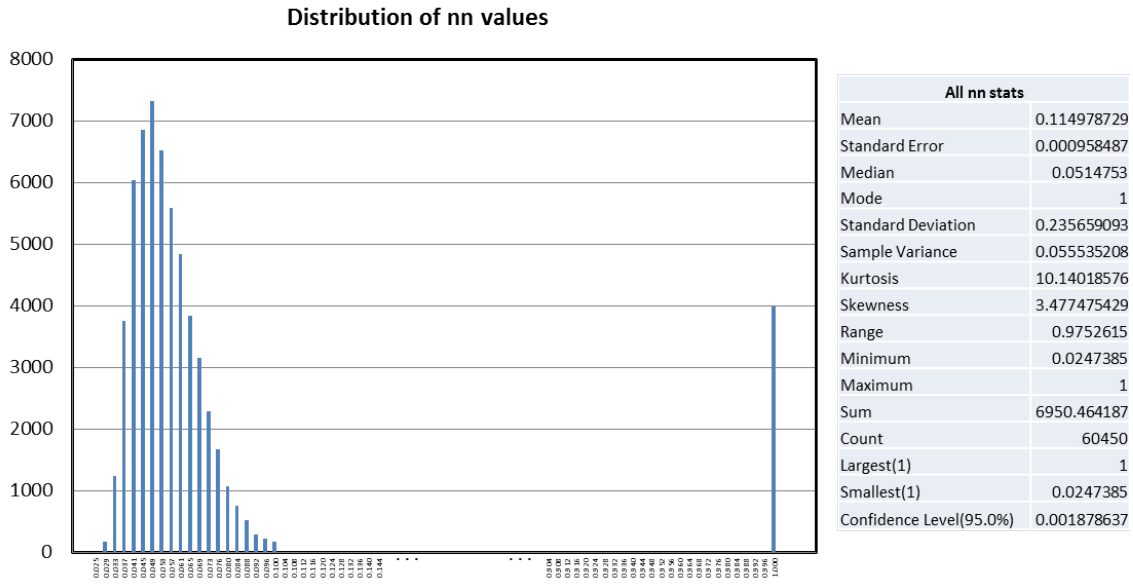


Figure 3.6: Histogram representing the distribution of *nn* values distribution and descriptive statistics.

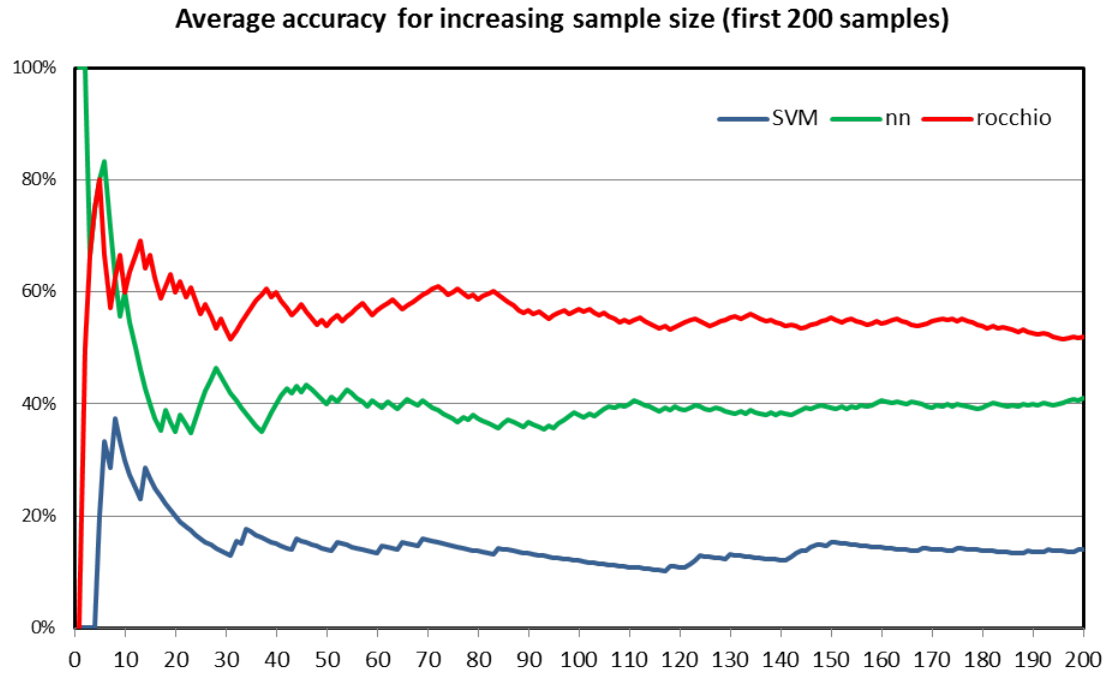


Figure 3.7: Average Accuracy of Classification of 200 Samples from svm, rocchio and nn subsets.

Chapter 4

Approach

This chapter starts by presenting a detailed view on the problem to be approached. This Chapter is divided in the following subsections:

- Problem Statement - Presents a detailed view on objectives, steps, issues and analysis methodology that are to be used throughout this approach;
- Data Preprocessing - In these subsections the necessary data cleaning and selection steps, which will be required in order to obtain analyzable results, are conducted;
- Taxonomy Construction - In this section the data acquired and refined in the previous sections is mined, extracting the relations which are used to construct a topic hierarchy;
- Results and Evaluation - In this section the methodology for evaluating the taxonomy is establish and the results from its application are analyzed.

4.1 Problem Statement

The objective of this work is to provide a method to construct an ordered network of *news topics* or taxonomy. This network will be constructed based on the relation *topic* -> *subtopic* in which, the *topic* is identifiable as a broader news *topic* than the *subtopic*. News *tags* will be used to represent the topics and sub topics.

News topics are constructed under the optics of concept analysis, in this sense, each *topic* is represented by a *tag* and defined as the set of *news* of all the *news* for which the *news* - *tag* association is accurate for that *tag*.

The *topic* -> *subtopic* association is based on the notion that *tags* which are accurate for more *news* will be *broader tags* (i.e. *tags* that can be used in different contexts or have different meanings). The association *topic* -> *subtopic* is defined as true if the *topic* is a super set of the *subtopic*. However, this definition is very strict and dependent on the accuracy of the *tag* - *news* relations.

The *topic* -> *subtopic* relation will be used to create concept trees. Each tree will represent a *broad news topic*. Each tree is constructed recursively, starting at the root node, which is a *topic* that is not a subset of any other *topic* and adding nodes for which the *tags* are subset of the root, in the *topic* -> *subtopic* relation.

4.2 Data Preprocessing

In this section the processes involved in cleaning, selection and transformation of the dataset are described and their results evaluated. More specifically, this will involve the following:

- removal of noise and inconsistent data;
- selection of the data from *nn, svm and rocchio* subsets.

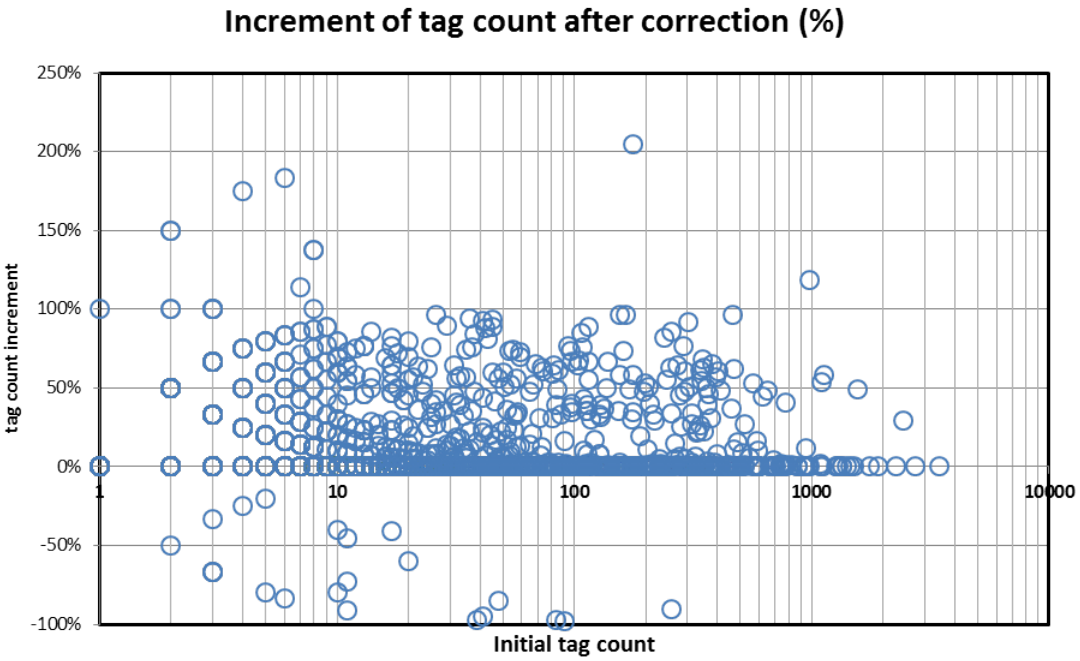


Figure 4.1: Percentual increment of news coverage per *tag* after the automatic correction.

4.2.1 Automatic Misspelling Correction

As observed in Chapter 3, one of the causes of the high number of distinct *tags* is minor spelling differences between *tags* (i.e. spelling mistakes and inconsistent terminology). Correcting these errors in preprocessing, ideally through the replacement of an incorrect *tag* with the correct version of that *tag*, would improved the quality of the dataset.

This would fall within the scope of spell checking and spelling suggestion problems, particularly automatic unsurpervised spellchecking. These are much studied problems and complex open fields, outside the scope of this work [TCSE07]. However, in this particular case an approach to automatic unsurpervised spellchecking can be achieved, assuming that the following hypotheses are true:

- for each incorrect *tag* there is one correct version of that tag in the dataset;
- for a pair of very *similar tags* there can be only one correct *tag*;
- the correct version of a *tag* is also the most popular version of that *tag* in the dataset.

The Levenshtein distance was chosen as the similarity criteria, this has been frequently used in the context of the identification of misspelled words [TCSE07]. The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion or substitution of a single character [Lev66]. A method was developed and implemented to calculate the highest (i.e. closest to 1) Levenshtein distance between all distinct pairs of *tags*. The minimum distance threshold for replacement was experimentally obtained and set to 0.8, this means that two *tags* with a Levenshtein distance above 0.8 are very similar. For very similar pairs of *tags*, the correct version was then defined as the *tag* which had no correct version - meaning no other very similar more popular *tag*. The correct *tag* was then propagated to all incorrect version of it.

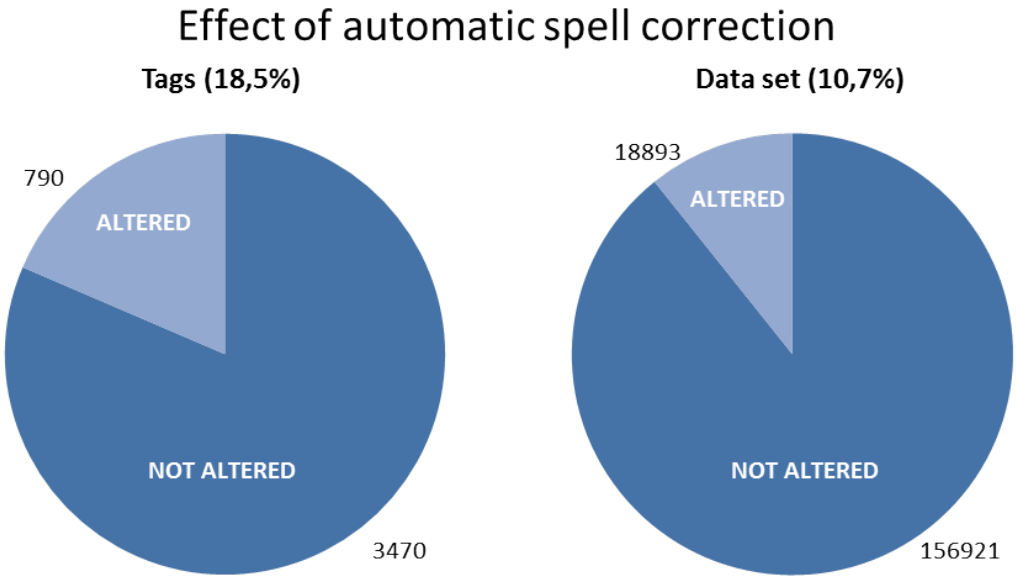


Figure 4.2: Impact of the automatic correcting on the number of tags and dataset entries.

This process affected 18.893 entries (10.7% of the dataset) (Figure 4.1 and Figure 4.2) and allowed the replacement of 790 *tags* for their more popular alternatives, reducing the number of distinct *tags* by 18.5% from 4260 to 3470.

4.2.2 Classification Analysis

As explored in the preliminary analysis in Chapter 3, the classification results have a low average accuracy which translates into a very high type 1 error rate, because of this, the data was considered too noisy. A method for distinguishing between accurate and inaccurate classifications had to be established.

The objective of this method is to determine the *value* that separates the low-accuracy results from the high-accuracy results, for each classifier *method*. This requires study of the relation between threshold *value* and average accuracy.

As before, the work done by Salzberg [Sal97] served as inspiration for establishing the steps for this process.

The process is described in the following steps:

- Extract 200 random samples from the dataset;
- The 200 randomly chosen samples are manually evaluate using the guidelines established in the previous chapter;
- Calculate the average accuracy from the set of 200 random samples;
- Analyze how the average accuracy changes when the threshold is increased (it is important to remember that any decrease in sample size translates into a decrease in precision);
- Based on the analysis of the previous step, set a new minimum *value*;
- Repeat this process, this time extracting a new set of samples with *value* greater than the threshold established in the previous step.

This process was conducted for each *method* subset, setting progressively higher threshold values, until there appeared to be no change in average accuracy or until the dataset became too small.

For the *nn* subset 3 set of 200 random samples were selected:

- *nn, value* > *0.029* - close to the set minimum below which no positive results were obtained through sampling;
- *nn, value* > *0.05* - this is the median, as *values* above the median appeared to score considerably better than the one bellow, in chapter 3 sampling, furthermore, this was chosen to divide the set into half;

- *nn, value > 0.06* - this *value* subdivides the set again, this was chosen to support the findings of the *value > 0.05* set.

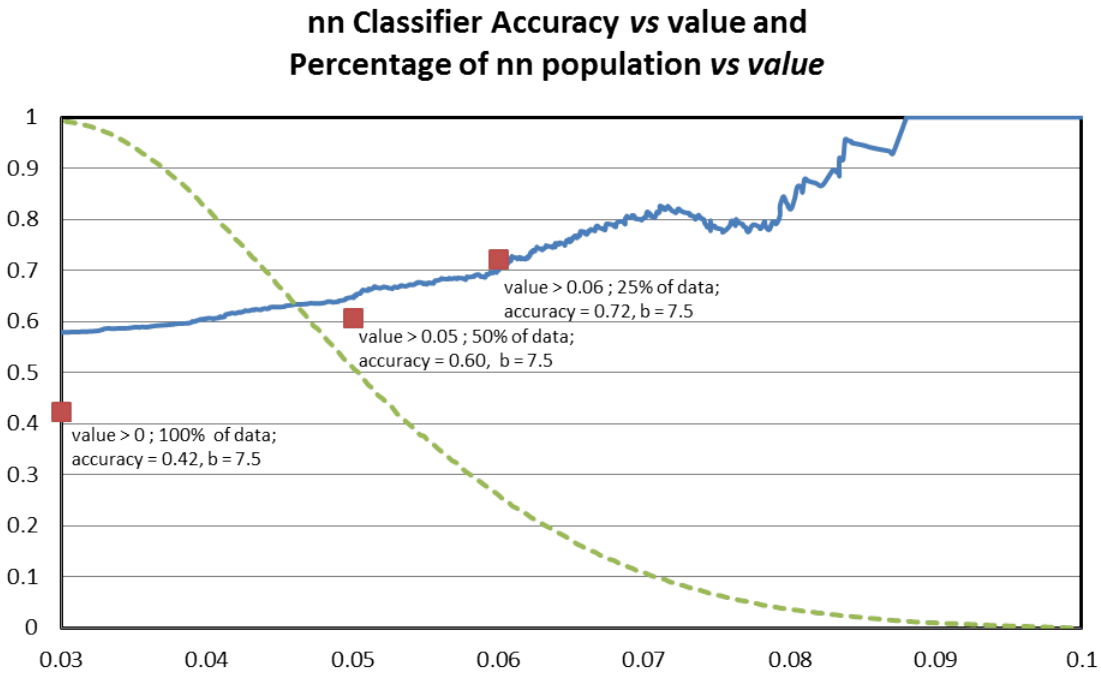


Figure 4.3: Distribution average classifier accuracy and percentage of population vs value threshold.

These samples were collectively mapped in Figure 4.3. Where the data points (in red, squares) show the mean accuracy obtained for each 200 random sample set. They increase in accuracy of classification from 0.42 (42%) to 0.72 (72%) as the threshold *value* increased.

The mean accuracy is also mapped across *value* increase. This is the mean accuracy of the threshold *value* for the 600 sample collectively. This was done in an effort to provide a deeper insight to the behavior of the classifier.

Finally the total percentage of subset was also mapped to show how the set size decreases with the increase in *value* threshold.

The previously described process was also performed for the *rocchio* subset. However for the *rocchio* subset only 2 sets of 200 random samples were extracted, these were chosen at specific minimum *values*:

- *rocchio, value > 0.008* - close to the set minimum: below this value no positive results were obtained through sampling;
- *rocchio, value > 0.2* - this value was suggested in the work previously done as an empirically chosen threshold [SE].

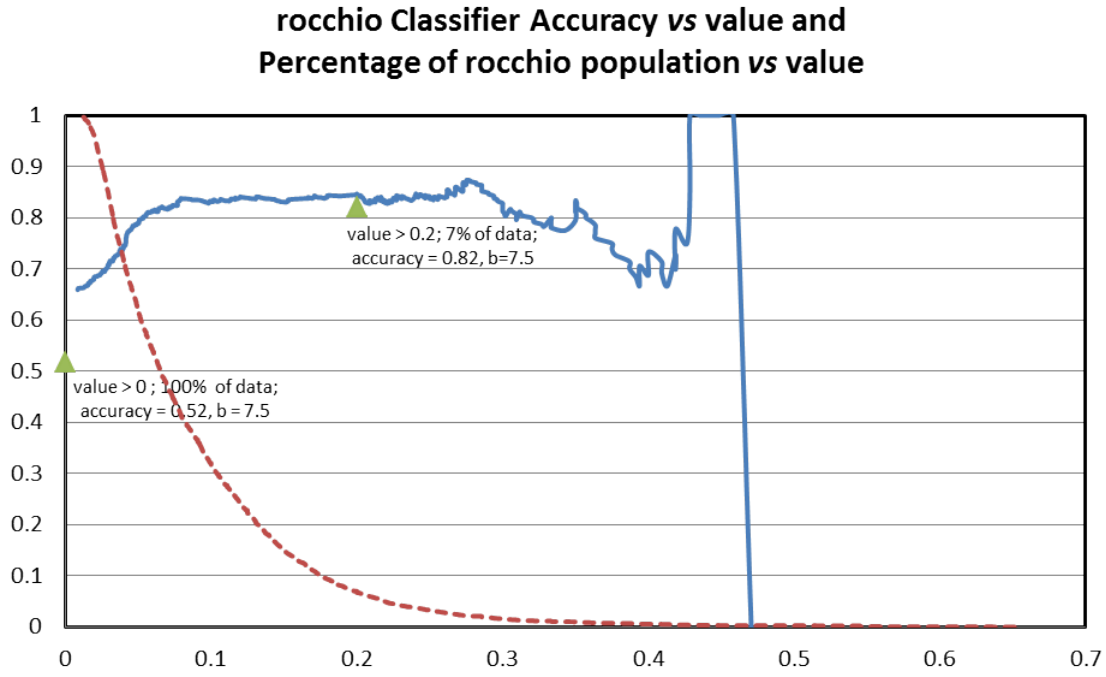


Figure 4.4: Distribution average classifier accuracy and percentage of population vs value threshold.

As before the sampled were collectively mapped, Figure 4.4. Where the data points (in red, triangles) show the mean accuracy obtained for each 200 random sample. They increase in accuracy of classification from 0.52 (52%) to 0.82 (82%) as the threshold *value* increases.

The distribution of mean accuracy was again mapped across *value* increase. The behavior appears to change drastically above $value > 0.3$ but this should simply be attributed to the very low number of samples from that *value* range and the lack of precision that entails.

Also, as before, the total percentage of the subset was mapped, showing how the set size decreases with the increase of *value*.

Finally this process was repeated to the *svm* subset. For this subset 3 sets of 200 random samples were chosen, as follows:

- *svm, value* > *-1.1* - close to the set minimum below this value no positive results were obtained through sampling;
- *svm, value* > *-0.09* - close to the median, as *values* above the median appear to score considerably better than the ones bellow, as seen in Chapter 3, furthermore, this was chosen to divide the set into half;
- *svm, value* > *-0.55* - suggested in the work previously done in as an empirically chosen threshold [SNO09].

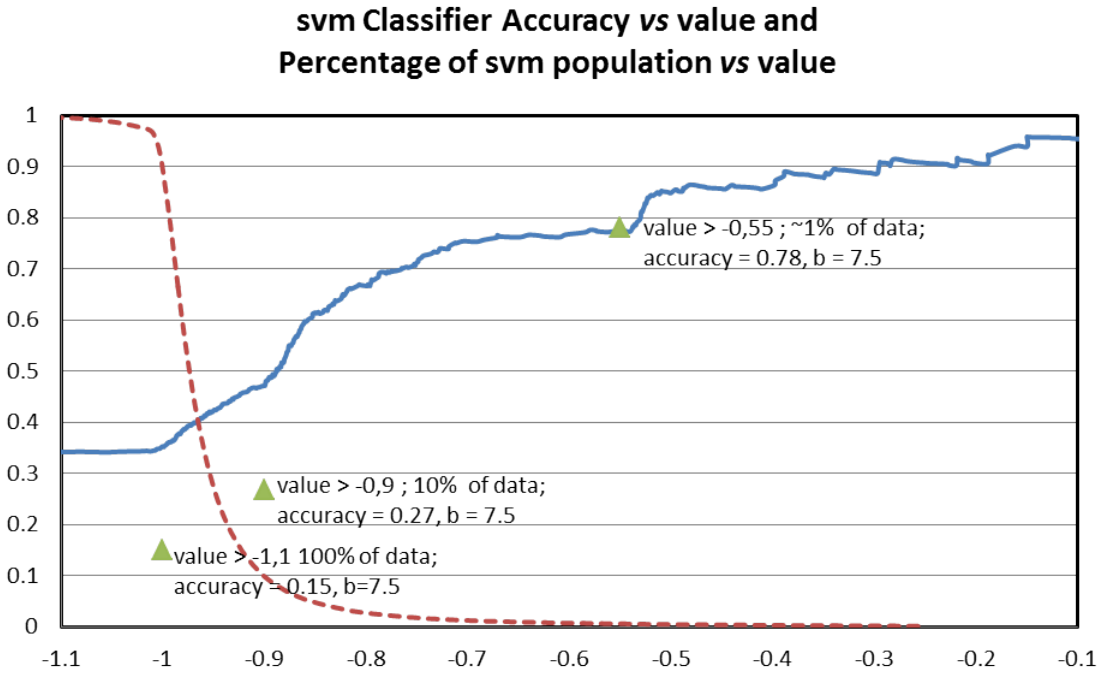


Figure 4.5: Distribution average classifier accuracy and percentage of population vs value threshold.

As before the samples were collectively mapped in Figure 4.5. The data points (in red, squares) show the mean accuracy obtained for each 200 random sample. They increase in accuracy of classification from 0.15 (15%) to 0.78 (78%) as the threshold *value* increases.

The distribution of mean accuracy was again also mapped across *value* increase. This is the mean accuracy of classification for threshold *value* for the 600 samples collectively.

Also as before, the total percentage of subset was mapped, showing how the set size decreases with the increase of *value*.

4.2.3 Dataset Selection

After the evaluation work in the previous section, it became possible, based on the *value*, to reach a balance in selecting the most accurate entries from the dataset with a certain degree of precision while maintaning as much of the data as possible.

The following rules, specificy the minimum *value* for each *method* and the number of dataset entries obtained by each rule.

- *original tags;* 2555 entries
- *nn, value = 1;* 3299 entries
- *nn, 1 > value > 0.065;* 9786 entries
- *rocchio, value > 0.1;* 4095 entries

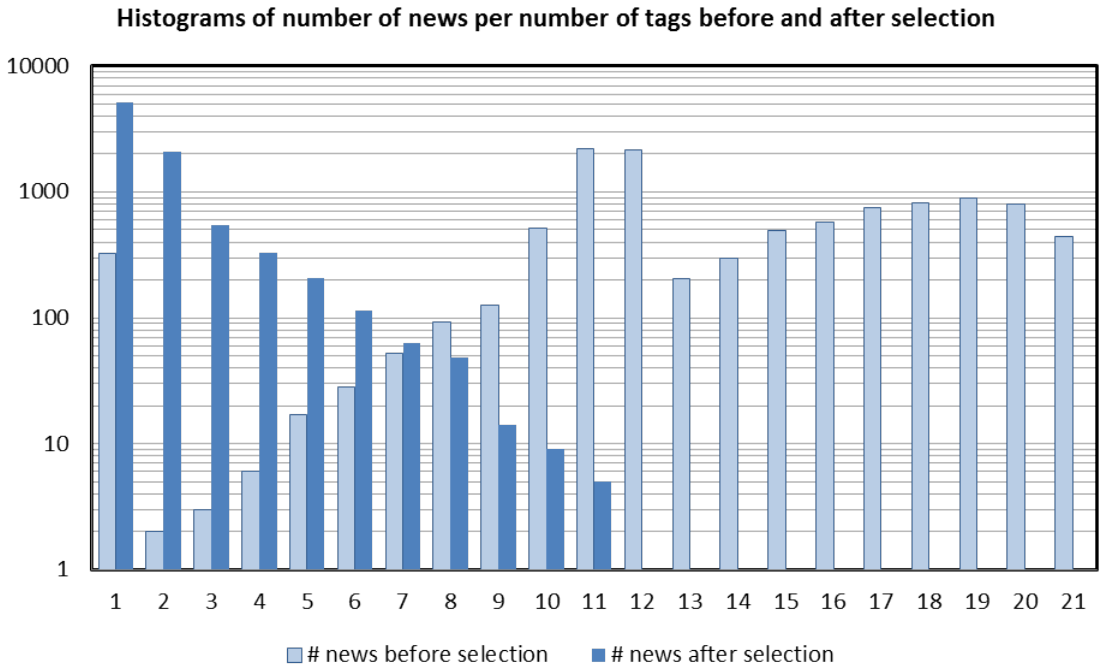


Figure 4.6: Histogram representing the distribution *tags* before and after automatic misspelling correction and data selection.

- *svm, value > -0.7;* 1375 entries

This selection results in a set with 21.092 entries and a mean expected accuracy of 83%. This value is the key that sets the baseline for the accuracy of the hierarchical relations to be extracted. The effect of these rules on the dataset can be in Figure 4.6.

4.3 Taxonomy Construction

The inicial approach resulted in a directed cyclic graph with the tags as nodes. An edge was added for each pair of tags in a news. Edge weight was defined as the number of times a given association appeared. Nodes were then given a score based on *tag* popularity.

This approach produced a very dense directed cyclic graph with just a few and very tightly packed clusters and also a very large amount of cycles and isolated nodes.

Besides its complexity, it was evident that in this graph two unrelated *tags* could be closely linked without the possibility of detecting that, foiling further attempts of transforming the graph into a taxonomical tree with language meaning. In practical terms, it seemed that this solution was diverging from the objective, so a new approach was devised from the lessons of this experience.

In order to produce a news topic taxonomy, the concept of *news topic* was first defined. A *topic* is a set of news which are all classified with the same *tag* (Figure 4.7.B).

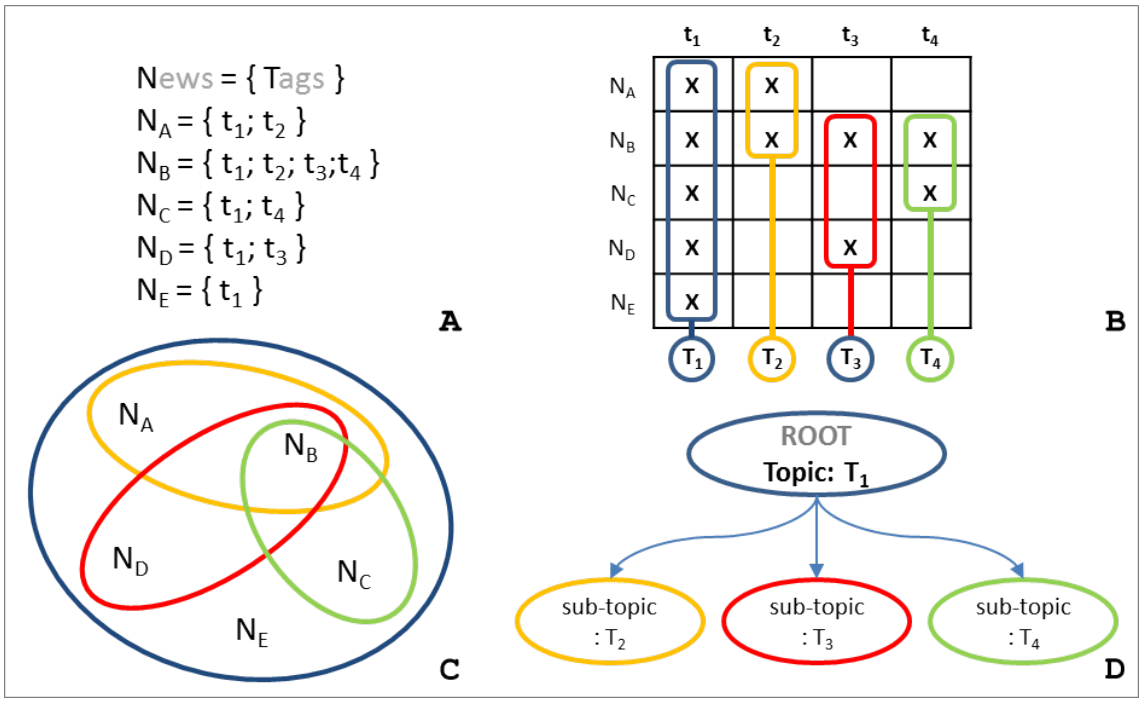


Figure 4.7: Detailed steps in the taxonomies trees construction algorithm.

A *subtopic* also has a very simple definition: *subtopic* is a subset of news of a topic (Figure 4.7.C). This implies that any node which is not a subset of any other node is special - a root node (Figure 4.7.D).

Taxonomical graphs can be constructed by starting at each root node and recursively applying these definitions in a top-down manner.

This method is much *cleaner* than the one in the initial attempt and resulted in a graph containing multiple disjoint directed acyclic sub graphs (Figure 4.8).

To achieve a taxonomical tree, the problem of having multiple paths that reach the same node must be solved. Because there is no relevant information available to choose between those multiple paths, an optimal branching algorithm could not be applied. Even if this was possible, this course of action could jeopardize information, eventually very important for taxonomical construction, from the news (language) point of view (example tree in Figure 4.9). As such, the approach taken was to preserve all possible paths, instead of choosing just one. So, when necessary, each node was replicated according to its parent node. This way multiple paths to the same node could exist in different trees according to distinct news topic meanings.

4.4 Topic Tree Results and Evaluation

One of the early considerations of this method was the following: whenever a certain *tag* was associated with a *news* which had only one single *tag*, that *tag* could never be used

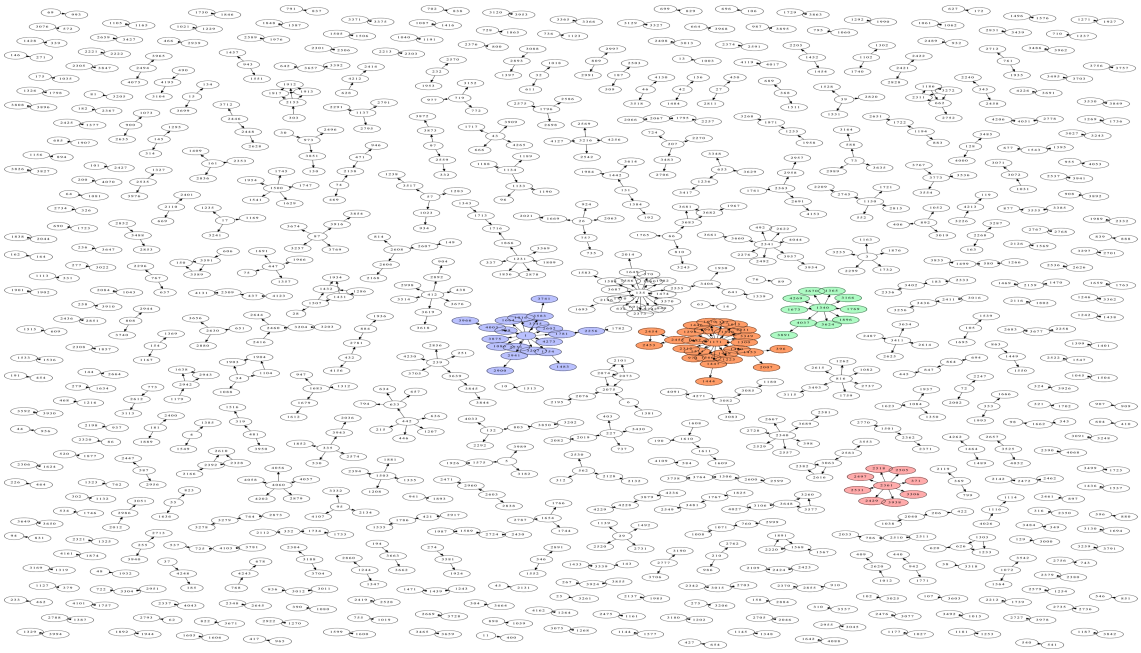


Figure 4.8: Graph containing multiples disjoint directed acyclic.

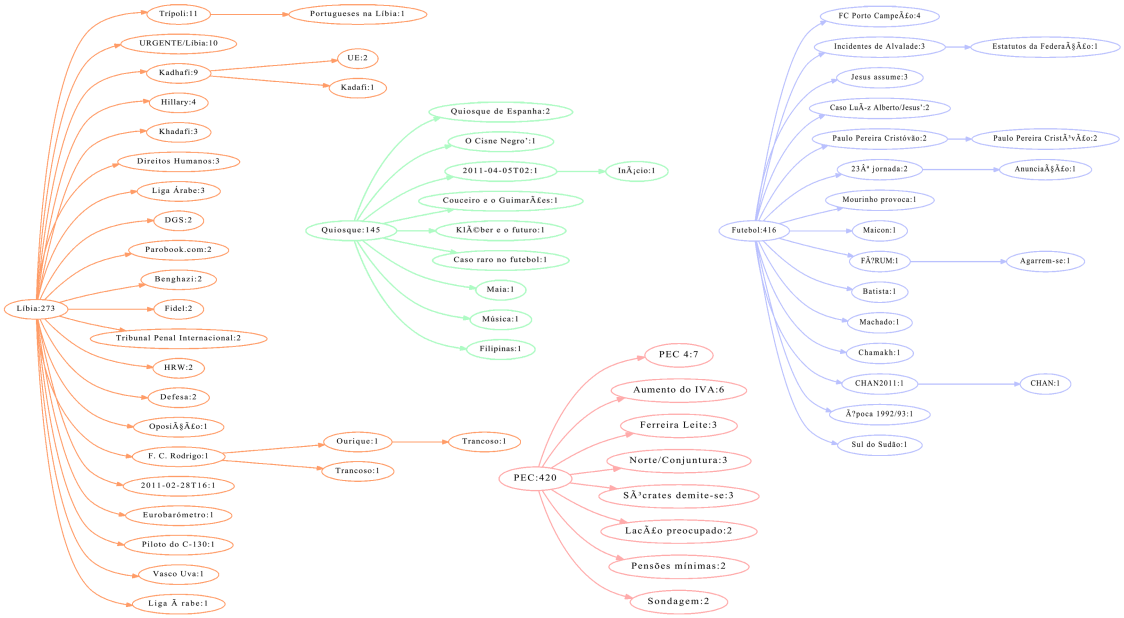


Figure 4.9: Detailed view of 4 topic trees.

as a *subtopic*.

Because of this, *news* that were associated to a single *tag* were considered to be *unhelpful* in constructing the taxonomy and were removed from the set before the next step.

This resulted in a set of 10.667 entries of *tag* - *news* pairs. This set covered 3.673 *news* and 1.379 *tags*. The majority of news had only 2 or 3 tags associated. The distribution of the number of *tags* associated to each *news* was already presented in Figure 4.6.

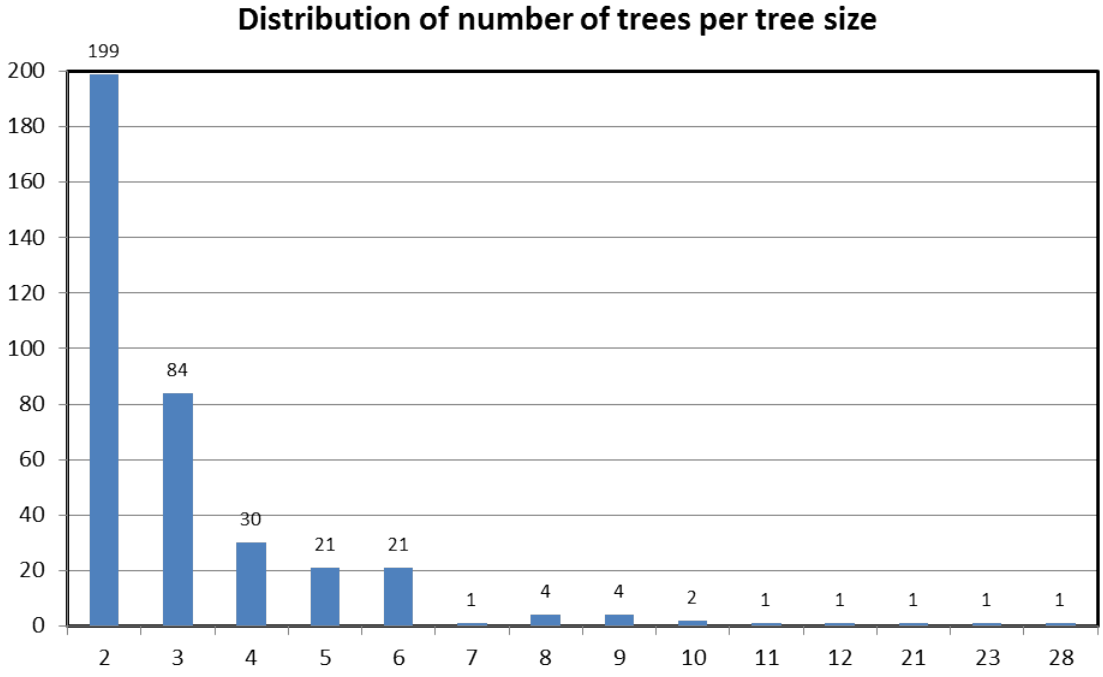


Figure 4.10: Distribution of the number of nodes in a tree.

The tree construction process (Section 4.3) was applied on this dataset and resulted in 371 distinct trees. The number of nodes in each tree, as well as the news coverage (i.e. the number of *news* which contain the *tags* in the tree) were collected and are represented in Figure 4.10 and Figure 4.12.

The size of the trees (i.e. the number of nodes in a tree) did not appear to be directly correlated to the news coverage of each tree (Figure 4.11). This appears to be an indicator that the trees are resistant to the popularity of the *tags*.

It is quite apparent from the Figure 4.8 that this method produces trees that are almost independent from one another.

Some trees were sampled and analyzed in detail. (shown in detail in Figure 4.9).

- Root node *Futebol*

  This tree has 20 nodes and covers 416 news. Under examination, most of the relations appeared accurate and could be positively identified as different news topics that had occurred in the period and were subtopics of *Futebol*. *Futebol* as a news topic, is a very broad topic. This becomes obvious simply by looking in detail to some of its subtopics. Most of them are either events or personalities somehow related with *Futebol*. Some of these are clearly identifiable, such as *FC Porto campeão*, *Incidentes de Alvalade*, *Mourinho provoca*. Others had to be confirmed by checking their respective news text, such as *Sul do Sudão*, *Machado*. Of all the nodes in this tree, only *FÓRUM* and *Agarrem-se* were found to be incorrect. The

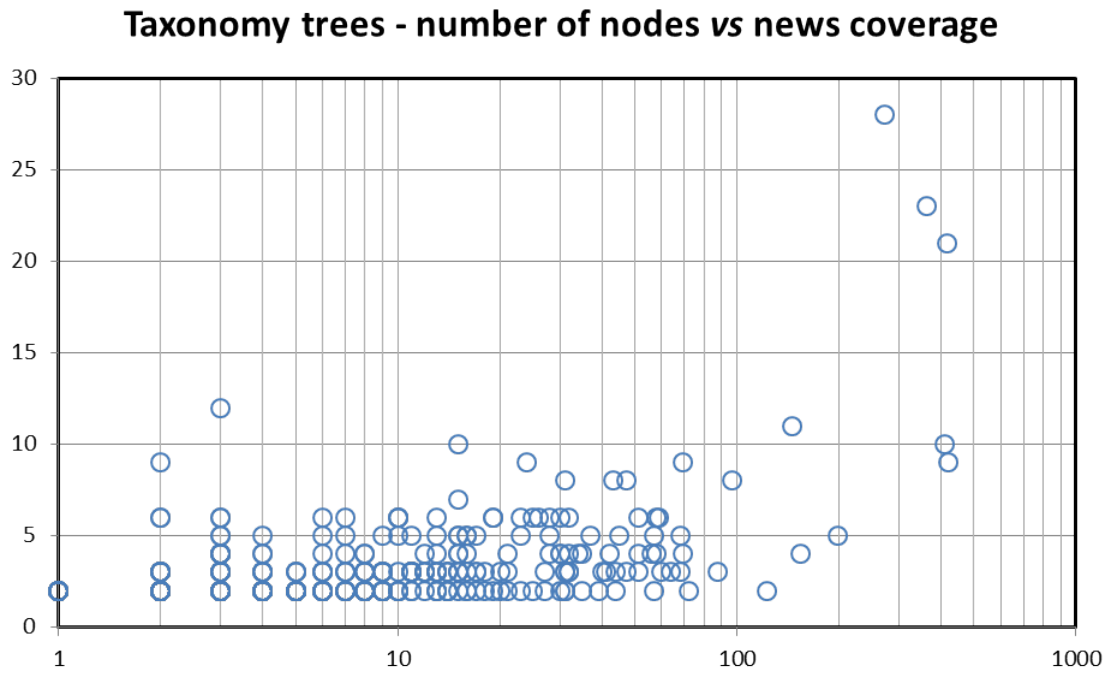


Figure 4.11: Distribution of the news coverage per number of nodes in a tree.

first is a common editorial name and not directly connected (i.e. there could be a *FÓRUM* about *Futebol*). The second is the only one which is most likely result of a misclassification, this was concluded after analysis of that *tag* news's text.

- Root topic *Líbia*

  This tree has 27 nodes and covers 217 news. As before, most of the relations appeared accurate and could be identified relating to the opinions that public figures and organizations had voiced about this topic, *Líbia*. Some topics, however were obviously synonyms (e.g. Kadhfi, Khadafi, Kadafi) the appearance of this topics is an indicator that the preprocessing step for removing duplicate misspelled entries could be further refined.

In general root trees seemed good candidates to be *News Topics* specially the large ones. Apart from that the influence of the *tags* wrongly classified can be seen at root node level (e.g. "Quiosque" is the name of a news column) and also at other tree levels (e.g. "Agarrem-se", "FC-Rodrigo", "Trancoso", "CHAN").

One of the types of errors can be traced to the selections of the *original tags*. Those errors derive from the use of tags that are meaningful only for their newspaper or magazine (e.g. the names of newspapers, sections, authors).

Overall, the general perception is that, despite having produced a large number of topics (371), the method produced topics that appeared not only usable but also very interesting from the language processing point of view. Another issue, is that some of

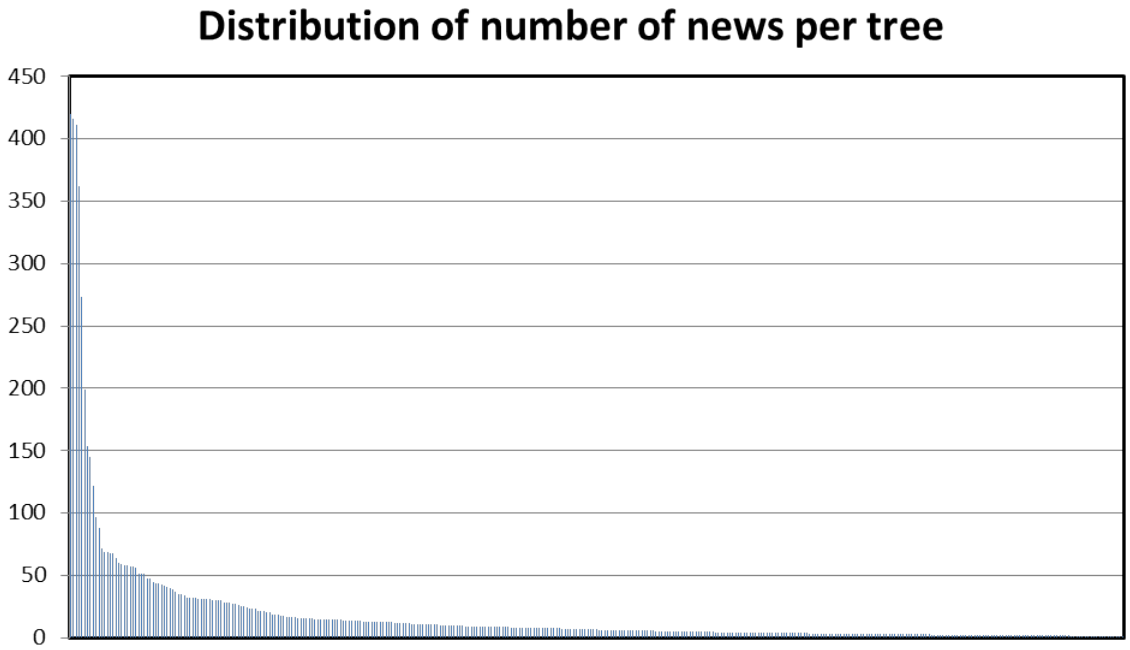


Figure 4.12: Distribution of the news coverage of all the trees.

the smaller trees could be expected to belong to some of the larger ones. This could be attributed to the discrete nature of the *topic* -> *subtopic* association. As such, it looks particularly promising to look at the topic trees as fragments of other much larger and more complex structures.

Although all the nodes of a taxonomical tree are topics, in the technical sense, it is obvious that their position in the tree bears a language significance. It is not clear, at this point, the full extent of the consequences in term of language processing. Nevertheless at least a kind of topic has its own importance - the root node.

4.5 Super-Topics

Trying to find out if there is a significant underlying structure relating the discovered 371 root topics, another step was taken.

In order to do that, the concept of super-topics needs to be defined. A super-topic is a set of tags which appear in all news of a root topic. All super-topics and related tags are listed in the Appendix D.

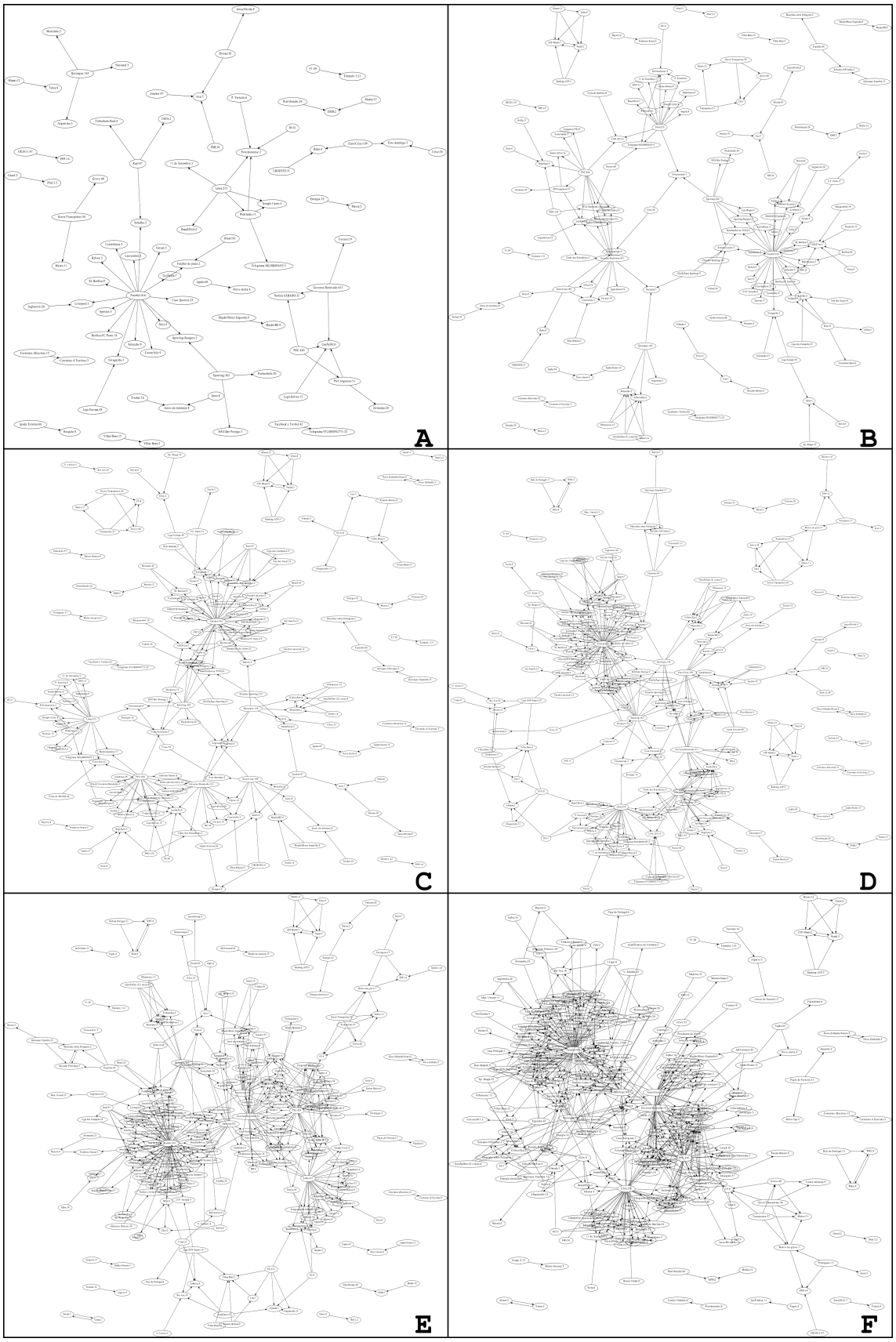


Figure 4.13: Composition of 6 super-topic graphs for diferent similarity values, A = 90%, B = 85%, C = 80%. D = 75%, E = 70%, F = 65%

The process described in Section 4.3, which was used to find the root topics, can now be used to find relations between super-topics. Although the process is, in essence, similar there is one notable methodological difference. The comparison criterion is not the strict inclusion (of one set in another) but a more *flexible* inclusion. This was done because almost no relations could be found when the strict inclusion criterion was used. In a way, as expected, this shows the unique nature of each super topic.

To uncover the relation between them, an experimental protocol was designed. This protocol consists in allowing a growing degree of relaxation of the comparison criterion. A parameter was defined - similarity - as the percentage of tags of a given super-topic that should match those of a larger super-topic (size was defined as the number of tags it has).

In Figure 4.13 results can be seen in the form of a direct graph for similarity = 90% to similarity = 65%. In Appendix E detailed views of each one can be found. In these graphs, each node represents a super-topic and each edge represents the connection between 2 super topics that met the similarity criterion at a given level. The edges are directed from the larger to the smaller super-topic.

When examining this graph, 3 diferent kinds of nodes were taken into consideration:

- Some nodes which are never successors of other nodes (i.e. there are no edges directed from other nodes to these ones) are called *sources*. These nodes represent *super topics* that are always larger than similar *super topics*;
- Some nodes which are never predecessor of other nodes (i.e. there are no edges directed from these ones to other nodes) are called *sinks*. These nodes represent *super topics* that are always smaller than similar *super topics*;
- The rest are predecessor to some nodes and successors to other.

Chapter 5

Conclusion

5.1 Brief Summary

The original dataset was composed by 175.814 entries regarding 10804 different Portuguese news from 23-12-2010 to 10-04-2011. Each entry contains a news title, a tag, and a classification.

There are 4 types of tags, original tags (attributed by the news producer) and 3 obtained by automatic classification methods - *nn* (62.767 entries), *svm* (100.740 entries), and *rocchio* (12.317).

Misspelling correction and duplicate entries elimination

From 4260 different tags this process permitted to reduce the number of tags to 3460. After duplicated entries were eliminated, the dataset was reduced to 151.387 entries.

Entries Selection

Thresholds for the acceptable classifier accuracy were empirically obtained for each classifier in order to secure a global estimated accuracy of 83%. The obtained dataset, contained 15.578 entries, 8.558 news. The dataset was further reduced by removing all the news with a single tag - 10.667 entries, 3773 news, and 1379 tags.

Taxonomy Construction

The concepts of topic, topic -> subtopic relation, and an algorithm to construct topic taxonomies were devised and applied to the dataset resulting in 371 trees, that is, in 371 topics.

Topic Relations

To investigate the underlying structure of the news produced in this period (23-12-2010 to 10-04-2011) the concept of super topic was introduced, the previous algorithm was adapted to extract the relations between super topics based on its tag-similarity.

Super Graph

Several graph were generated, using various degrees of similarity, to better understand the trade-off between complexity (noise) and information usability (semantic intuition).

5.2 Observations

As indicated in Taxonomy Construction 4.3 this process resulted in a large number of shallow trees. Upon observation this appear to be attributable to the restrictive nature of the rules used in establishing the *topic - subtopic* relations. The method used allows no leeway for incorrect *tag - news* associations. This is also believed to be the cause of there there being no two identical *topics* (i.e. topics defined by the exact same set of news).

Overall this methodology proved to be a powerful analysis tool allowing a deep insight to the structure of the Portuguese news from the perspective of taxonomical and similarity relations between its topics.

In a news sense, topics could be found automatically and ranked according to news coverage and tag relationships.

Although the feasibility of the process was demonstrated, several weaknesses imped its performance and demand further work. First of all, tags provide a better representation of the news content. There are many popular tags that are misleading, either because they are void of information or because they give clues in wrong directions. For instance the tag *quiosque*, which appear to be a common newspaper column name but was used in 145 of the final dataset (4%). This tag bears no news meaning, but nonetheless, due to its popularity it found its way to a relevant position in the super topic graph (it's a source with 12 direct successors).

5.3 Future Work

Impact of the classifiers accuracy on the final outcome

The dataset selection was based on the assumption that classification accuracy was important for the extraction of meaningful knowledge. However, it can also be argued that, to some extent, sample size could be more important than accuracy. This can be verified

through the conduction of several experiments with different threshold values (i.e. different datasets). In the same line of work, the use of original datasets could be improved by refining the rules which are used to extract the tags from headlines.

Classifier accuracy improvement

Based on the results from the topic taxonomical trees, better training sets could be found and used to train classifiers. Also, hierarchical classification models could be designed and used to improve the news classification process. Once this is done, the work presented here could be redone and evaluated accordingly. The repetition of these steps would be revelant as long as there was for room for improvement.

Time-based analysis

Another way of looking at the dataset would be from a temporal standpoint, altough disregarded, together with data sequence, in this work. The results reflect a specific time period of about 15 weeks, during which, the Portuguese news panorama changed significantly. Some news topics emerged while other disappeared. In future endeavors this or other similar datasets could be considered as a temporal window from which to look at the way news change over time. This approach could be capable of proving means to analyze *current events*, to identify trends or cycles in the news, to link data to seasonal variations, or even to predict variations in the dataset.

Conclusion

References

[CC04] S.L. Chuang and L.F. Chien. A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 127–136. ACM, 2004.

[DC00] Susan Dumais and Hao Chen. Hierarchical classification of Web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00*, pages 256–263, 2000.

[DdsM03] Bento Carlos Dias-da silva and Hélio Roberto De Moraes. A CONSTRUÇÃO D E UM THESAURUS ELETRÔNICO PARA O PORTUGUÊS DO BRASIL Introdução. 47(2):101–115, 2003.

[Han06] Jiawei Han. *Data Mining*. Morgan Kaufmann, San Diego, 2006.

[Hea92] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[HKR09] Eduard Hovy, Zornitsa Kozareva, and Ellen Riloff. Toward completeness in concept extraction and classification. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, page 948, 2009.

[HRGM08] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*, page 531, 2008.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.

[Kor02] Anna Korhonen. Subcategorization acquisition. (530), 2002.

[Lev66] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–+, February 1966.

[MAC+06] Palmira Marrafa, Raquel Amaro, R.P. Chaves, S. Lourosa, Catarina Martins, and S. Mendes. WordNet. PT–Uma rede léxico-conceptual do Português on-line. *clul.ul.pt*, 2006.

[MBF+90] George a. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database *. *International Journal of Lexicography*, 3(4):235–244, 1990.

[NIS11] NIST/SEMATECH. e-handbook of statistical methods @ONLINE, June 2011.

[OSGS08] H. Oliveira, D. Santos, P. Gomes, and N. Seco. PAPEL: a dictionary-based lexical ontology for Portuguese. *Computational Processing of the Portuguese Language*, pages 31–40, 2008.

[PL08] Anon Plangprasopchok and Kristina Lerman. On constructing shallow taxonomies from social annotations. *Information Sciences*, 2008.

[Pla09] A Plangprasopchok. Constructing folksonomies from user-specified relations on flickr. *Proceedings of the 18th*, 2009.

[PLG11] A. Plangprasopchok, Kristina Lerman, and Lise Getoor. A probabilistic approach for learning folksonomies from structured data. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 555–564. ACM, 2011.

[Roc71] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[Sal97] S.L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and knowledge discovery*, 1(3):317–328, 1997.

[SC99] Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, pages 206–213, 1999.

[SE] Luís Sarmento and Faculdade De Engenharia. BACO – A large database of text and co-occurrences. *Technology*.

[SL01] Aixin Sun and Ee-peng Lim. Hierarchical text classification and evaluation. *Proceedings 2001 IEEE International Conference on Data Mining*, (November):521–528, 2001.

[SM07] Win De Smet and Marie-Francine Moens. Generating a Topic Hierarchy from Dialect Texts. *18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, pages 249–253, September 2007.

[SNO09] L. Sarmento, S. Nunes, and E. Oliveira. Automatic extraction of quotes and topics from news feeds. In *4th Doctoral Symposium on Informatics Engineering*, 2009.

[SNTO09] L Sarmento, S Nunes, Jorge Teixeira, and E. Oliveira. Propagating Fine-Grained Topic Labels in News Snippets. In *Proceedings of the 2009*

IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03, number April, pages 515–518. IEEE Computer Society, 2009.

[TCSE07] Alexander Turchin, Julia T Chu, Maria Shubina, and Jonathan S Einbinder. Identification of misspelled words without a comprehensive dictionary using prevalence analysis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 751–5, January 2007.

[vR03] R. van Rees. Clarity in the usage of the terms ontology, taxonomy and classification. *CIB REPORT*, 284:432, 2003.

REFERENCES

Appendix A

Classifier Evaluation

This Appendix contains the first 100 results from the manual evalution of *tag - news* association random sample for the *rocchio* set.

Classifier Evaluation

News Id	News Title	Score	Tag	Evaluation
820197	Estilista John Galliano diz adorar Hitler - video	0.0419	Líbia	no
915256	PSD quer mobilidade forçada. Sindicatos dizem não	0.0735	Governo/Demissão	no
898793	Pingo Doce nega bactérias em camarão congelado	0.0667	Alimentação	yes
806562	NATO não vai intervir na crise da Líbia	0.0702	Líbia	yes
919784	Bagão Félix não retira acusações de mentiroso a Sócrates	0.1200	Governo/Demissão	yes
917259	Preferiria mil vezes mexer no IVA a ter que cortar pensões e reformas	0.0579	Governo/Demissão	yes
826658	Vargas Llosa garante que vai falar na Feira do Livro de Buenos Aires	0.0373	Líbia	no
827044	Governo aprova proibição de estágios profissionais não remunerados	0.0556	Juros	no
919262	água de Luso aponta para crescimento de 6% a 8% este ano	0.0506	Dívida pública	no
797913	EMEL satisfeita com um ano de Smartpark e prepara pagamentos por telemóvel	0.0429	Sub-20	no
914734	E a tua universidade, é verde?	0.0372	EUA	no
908949	Actual Governo deve assegurar condições de liquidez	0.0783	PEC	yes
787396	"Estamos mais fortes do que há um ano"	0.0299	Futebol	yes
779198	Sé de Braga entra na "Rota das Catedrais"	0.0355	Egito	no
905026	Sonae Sierra entra em áfrica	0.0342	Reino Unido	no
794198	Homem que acompanhava presidente da Junta feridos	0.4286	Portalegre	no
784954	Encontradas espécies invasoras em madeira importada à venda nos Açores	0.0201	Reino Unido	no
821521	Mais de 9 mil reformas antecipadas com corte médio de 14%	0.0293	CP	no
889795	Vieira da Silva diz que 5 de Junho é a data mais adequada para eleições	0.2128	PEC	no
887490	Pacheco Pereira defende acordo pré-eleitoral entre PSD e PS	0.1200	PEC	no
888219	Carlos Queiroz corrosivo na resposta a Pepe	0.1029	Caso Queiroz	yes
781752	Diabetes está a crescer em Portugal	0.0197	Alemanha	no
788387	Roubini prefere Mario Draghi para liderar BCE	0.0326	Alemanha	no
783532	Francisco Louçã diz que limite ao défice na Constituição torna economia estúpida e incompetente	0.0714	Egipto	no
789889	Manuel Fernandes assume favoritismo frente à Naval	0.1167	V. Setúbal	yes
920589	Assis admite que Portugal possa pedir ajuda	0.1277	Governo/Demissão	yes
826046	Qualidade que marca pontos	0.0290	Dívida	no
785508	Euribor voltam a subir em todos os prazos	0.0484	Dívida	yes
779362	Projectos de Design do Politécnico de Viana do Castelo expostos na Interdecoração	0.0273	Ecumenismo	no
889371	Austeridade em Portugal é um "delírio"	0.2273	JN	no
904187	Ministro da Holanda diz que está tudo pronto caso Portugal decida pedir ajuda	0.1798	Juncker	no
807802	Embaixador líbio em Brasília diz que Al-Qaida quer usar a Líbia para atacar a Europa	0.1169	Líbia	yes
899775	Ministro da Economia acusa Telmo Correia de demagogia	0.1333	Governo	yes
918479	Empresários suspeitos de apedrejarem veículo pesado absolvidos em Tribunal	0.0282	Paquistão	no
...	...	...	...	...

Table A.1: An example of table

News Id	News Title	Score	Tag	Evaluation
...	...	...	...	...
819805	Empregado da Apple admite ter vendido informação da companhia	0.1167	Estudo	no
888336	Dionísio Castro promete 'forcing' para colocar Vagner Love em Alvalade	0.0963	Sporting/Eleições	yes
785532	Parlamento vai ouvir ministro da Presidência sobre problemas com cartão do cidadão	0.0417	Eleições	no
783717	Chefes do partido no poder no Egipto abandonam cargos	0.0982	Egipto	yes
890077	Passos Coelho diz que lutará por uma maioria absoluta do PSD	0.2143	Governo/Demissão	yes
783480	Paulo Portas diz que não deixará que PSD extinga Ministério da Agricultura	0.0370	Dívida	no
887178	"Sporting está obrigado a lutar por títulos todos os anos" - Carlos Freitas	0.0599	Sporting	yes
783332	Leonardo Jardim admite "desinvestimento" no plantel sem perda de ambição	0.1679	Beira-Mar	yes
815527	Saldos de Inverno terminam hoje com quebras entre 10 e 20%	0.0476	Líbia	no
909826	Cavaco diz que corte de rating "é exagero muito grande"	0.1731	GovernoDemissão	yes
778618	Médicos contratados por empresas são mais caros, mas não melhores	0.0273	Egito	no
810113	Rui Vitória promete equipa personalizada para receção ao V. Setúbal	0.0867	Futebol	yes
902685	Mais de 2,5 milhões de pessoas recenseadas através da Internet	0.1513	Censos2011	yes
886670	Miccoli não acredita no regresso à seleção italiana	0.0702	Itália	yes
797517	Grande Prémio do Bahrein cancelado	0.1304	Bahrein	yes
816986	Novos testes de stress seriam "tiro no pé", alerta Ulrich	0.0526	Trichet	no
816484	Principais bancos criam fundo de capitalização para financiar melhores PME	0.0505	Seguros	no
785832	O Português João Damas é um dos guardiães da Internet	0.0307	Internet	yes
820375	Mikhail Gorbachev anuncia prémio com o seu nome	0.0400	óscares	no
908080	Projecto científico procura desbravar os fundos submarinos dos Açores	0.0151	Artur Agostinho ao SOL	no
894182	V. Guimarães na final da Taça de Portugal	0.0978	Taça de Portugal	yes
782437	Montepio alheio às investigações sobre acções do Finibanco	0.0252	OFICIAL	no
888576	Eleições a 5 Junho são "a melhor saída", afirma Jerónimo	0.2373	PCP	yes
884157	Presidente da CAP diz que demissão do Governo traz nova esperança	0.1719	Governo/Demissão	yes
911693	Fundo Europeu preparado a ajudar Portugal	0.1163	Juncker	yes
899050	Lula da Silva pede uma nova governação mundial	0.0622	Portugal/Brasil	yes
780411	Louçã propõe redução da taxa de contribuição dos recibos verdes	0.0403	Crise	yes
809356	BE acusa PS e Governo de não cumprir compromisso e de tentar esconder precariedade dos professores	0.0269	BE	yes
929020	PT reforça linha de crédito para 1.050 milhões de euros	0.2821	Telecom	yes
888308	A melhor saída para crise política é através de eleições	0.1310	PEC	no
916713	Necessidades de financiamento estão a condicionar as nossas políticas	0.0621	Euro/Crise	yes
781637	Centros de saúde do Nordeste deixam de fazer atendimento depois das 22h00	0.0172	Egito	no
893203	Governo está a fazer todos os esforços para regularizar dívida com a Madeira Medical Center	0.0962	Madeira	yes
...	...	...	...	...

Classifier Evaluation

Table A.2: An example of table

Classifier Evaluation

News Id	News Title	Score	Tag	Evaluation
...	...	...	...	...
817011	Carris pede ao Governo para manter pagamento de remunerações variáveis	0.0619	Cortes salariais	yes
829139	Soldado Manning é "herói sem igual" - Assange	0.1084	WikiLeaks	yes
786357	BE anuncia que vai apresentar moção de censura ao governo	0.0923	AR/Censura	yes
777831	Stanley Ho acusa 3a mulher e 5 filhos de "apropriação fraudulenta" de acções	0.0157	Rio Ave	no
827344	HP e Microsoft querem levar as 50 maiores empresas portuguesas para a nuvem	0.0459	Exportações	no
788081	Projeto do "Enfermeiro de Família" permitiu 4.300 consultas no Centro de Saúde de Vila Franca do Campo	0.0412	Açores	no
811294	Valter Lemos diz que não é "responsável pela incompetência" do PSD	0.0769	Emprego	no
887604	"Rating voltará a subir com medidas do PSD"	0.1092	PSD	yes
827050	Carlos Barbosa defende "limpeza" na arbitragem	0.1667	Sporting	yes
797106	Trabalhadores dizem que reduções afectam segurança do metro de Lisboa	0.0841	Greve/Transportes	yes
885262	Casillas pede "mais educação" com Mourinho	0.2088	Quiosque	no
785196	Panda Security revela principais ameaças e dá dicas de segurança	0.0947	Internet	yes
807596	Ministro Rui Pereira diz que número de eleitor é anacronismo, oposição condena proposta precipitada	0.0425	Segurança	no
794064	Portugal renova acordo petrolífero com a Venezuela	0.1777	Venezuela	yes
903567	Japão. Autoridades vão desactivar quatro reactores de Fukushima	0.1231	Japão	yes
900448	O Mobi.E pode ser comparado ao Facebook	0.0769	Governo/Demissão	no
901689	PSD apresentou pacote legislativo para promover "transparência das contas públicas"	0.0663	PEC	no
916348	Novos reformados trabalharam entre 25 e 34 anos	0.1136	Portugal	no
918581	Barco de pesca encalha em Peniche	0.0417	Greve/Transportes	no
913236	Luís Procuna dirige escola de toureio da Moita	0.0440	Sporting	no
791024	Vaga de contestação continuar a agitar mundo muçulmano	0.0701	Iémen	yes
905561	PSD recua e vota contra prescrição de medicamentos por substância activa	0.1707	Educação	no
891372	Sporting/Eleições - José Roquette alerta para risco de o clube cair nas mãos de um aventureiro	0.1475	Sporting	yes
828950	Siza Vieira diz que Casa da Arquitectura é necessária para preservar espólio nacional	0.0379	I Liga	no
930412	Governo deve criar fundo social de emergência, propõe Cáritas	0.0864	Crise	yes
778080	Rui Pereira vai "insistir" na aprovação da videovigilância em Lisboa	0.0463	FPF	no
784487	Farmacêuticos vão poder trocar remédios mesmo quando médico não autoriza, diz sindicato	0.0391	Saúde	yes
817563	Roff filial na Suécia	0.0301	Telecom	no
822401	"PS está a fazer o trabalho sujo do PSD"	0.0672	AR/Censura	yes
826887	Mais de 200 mil fogem de conflitos na Costa do Marfim	0.1692	Costa do Marfim	yes
901948	Açores recusam suspender processo de avaliação dos professores	0.1042	Professores	yes
901953	Japão desactivará quatro reactores de central danificada	0.2410	Japão	yes
779545	Três dias de luta foram extremamente positivos	0.0196	Taça de Portugal	no
897420	Metro de Lisboa faz hoje quarta greve parcial desde Fevereiro	0.1871	Greve	yes
779829	Autoeuropa suspende sábados de produção extraordinária em Fevereiro	0.0135	Rio Ave	no

Table A.3: An example of table

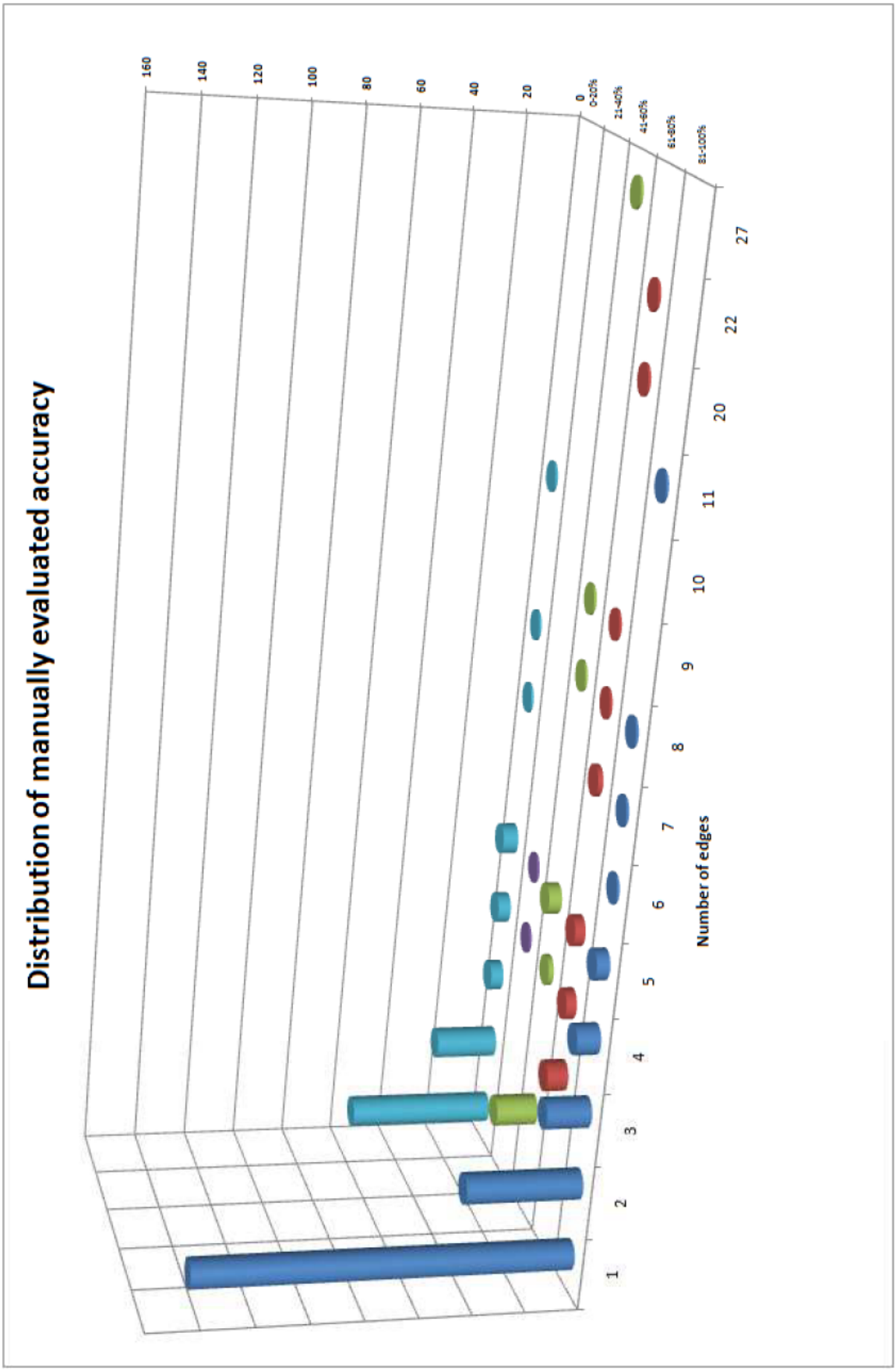
Appendix B

Taxonomy Evaluation

This Appendix contains the results of the manual evalution of the taxonomical trees constructed.

Each tree was evaluated under the following format:

- root Root node tag;
- count Number of edges in the tree;
- T ok Total number of adequate nodes (i.e. nodes obeying the topic -> subtopic relation);
- T wrong Total number of inadequate nodes (i.e. nodes not obeying the topic -> subtopic relation);
- T ??? Total number of nonsensical nodes (e.g. misspelled synonyms);
- % ok Percentage of adequate nodes (i.e. nodes obeying the topic -> subtopic relation);
- % wrong Percentage of inadequate nodes (i.e. nodes not obeying the topic -> subtopic relation);
- % ??? Percentage of nonsensical nodes (e.g. misspelled synonyms).


Distribution of manually evaluated accuracy
160
140
120
100
80
60
40
20
0
0-20%
21-40%
41-60%
61-80%
81-100%
1
2
3
4
5
6
7
8
9
10
11
20
22
27
Number of edges

root	count	T ok	T wrong	T ???	% ok	% wrong	% ???
Líbia	27	15	5	7	55.6%	18.5%	25.9%
Sporting	22	16	3	3	72.7%	13.6%	13.6%
Futebol	20	13	2	5	65.0%	10.0%	25.0%
Roger Waters em Lisboa	11	9	0	2	81.8%	0.0%	18.2%
Quiosque	10	0	0	10	0.0%	0.0%	100.0%
Governo/Demissão	9	6	1	2	66.7%	11.1%	22.2%
Man. United	9	4	0	5	44.4%	0.0%	55.6%
Egito	8	4	0	4	50.0%	0.0%	50.0%
PEC	8	8	0	0	100.0%	0.0%	0.0%
Portimonense	8	0	8	0	0.0%	100.0%	0.0%
SUB-21	8	5	1	2	62.5%	12.5%	25.0%
Brasil	7	5	1	1	71.4%	14.3%	14.3%
JN	7	0	0	7	0.0%	0.0%	100.0%
Juncker	7	5	2	0	71.4%	28.6%	0.0%
OE2011	7	7	0	0	100.0%	0.0%	0.0%
V. Setúbal	6	5	0	1	83.3%	0.0%	16.7%
«Discurso Directo»	5	0	0	5	0.0%	0.0%	100.0%
125cc	5	0	0	5	0.0%	0.0%	100.0%
Arquitetura	5	3	2	0	60.0%	40.0%	0.0%
Basquetebol	5	3	0	2	60.0%	0.0%	40.0%
Bayern	5	5	0	0	100.0%	0.0%	0.0%
Congresso PS	5	4	0	1	80.0%	0.0%	20.0%
Crise	5	5	0	0	100.0%	0.0%	0.0%
Distribuição	5	3	0	2	60.0%	0.0%	40.0%
Dívida	5	3	0	2	60.0%	0.0%	40.0%
Este domingo	5	0	0	5	0.0%	0.0%	100.0%
EUA	5	4	0	1	80.0%	0.0%	20.0%
Faro	5	2	1	2	40.0%	20.0%	40.0%
Flash/ Governo/Demissão	5	0	5	0	0.0%	100.0%	0.0%
iPad 2	5	3	0	2	60.0%	0.0%	40.0%
Madeira	5	4	0	1	80.0%	0.0%	20.0%
Mourinho sobre Pellegrini	5	5	0	0	100.0%	0.0%	0.0%
Petróleo	5	5	0	0	100.0%	0.0%	0.0%
Rio Ave	5	5	0	0	100.0%	0.0%	0.0%
Telecom	5	4	0	1	80.0%	0.0%	20.0%
Tonel	5	0	5	0	0.0%	100.0%	0.0%
Van der Vaart	5	1	0	4	20.0%	0.0%	80.0%
21Âª Jornada	4	4	0	0	100.0%	0.0%	0.0%
Â«Discurso DirectoÂ»	4	0	0	4	0.0%	0.0%	100.0%
Académica	4	3	0	1	75.0%	0.0%	25.0%
Alemanha	4	4	0	0	100.0%	0.0%	0.0%
Axa	4	3	1	0	75.0%	25.0%	0.0%
Cabral	4	4	0	0	100.0%	0.0%	0.0%
City	4	0	4	0	0.0%	100.0%	0.0%
Educação	4	3	0	1	75.0%	0.0%	25.0%
Espanha	4	0	0	4	0.0%	0.0%	100.0%
Euro/Crise	4	4	0	0	100.0%	0.0%	0.0%
F.C. Porto	4	1	2	1	25.0%	50.0%	25.0%
Gaia	4	0	0	4	0.0%	0.0%	100.0%

Camarate	2	1	1	0	50.0%	50.0%	0.0%
Cancro	2	2	0	0	100.0%	0.0%	0.0%
Cartão Cidadão	2	2	0	0	100.0%	0.0%	0.0%
Censos2011	2	1	0	1	50.0%	0.0%	50.0%
China	2	1	1	0	50.0%	50.0%	0.0%
Ciclismo	2	2	0	0	100.0%	0.0%	0.0%
Cinema	2	0	0	2	0.0%	0.0%	100.0%
CM	2	0	2	0	0.0%	100.0%	0.0%
Combustíveis	2	1	1	0	50.0%	50.0%	0.0%
Correios	2	0	0	2	0.0%	0.0%	100.0%
Correntes dEscritas	2	0	0	2	0.0%	0.0%	100.0%
Cortes salariais	2	2	0	0	100.0%	0.0%	0.0%
Costa do Marfim	2	2	0	0	100.0%	0.0%	0.0%
Eleições Sporting	2	2	0	0	100.0%	0.0%	0.0%
Ensino Básico	2	1	0	1	50.0%	0.0%	50.0%
Estágios	2	0	0	2	0.0%	0.0%	100.0%
FC Porto	2	2	0	0	100.0%	0.0%	0.0%
FinlÃ¢ndia	2	2	0	0	100.0%	0.0%	0.0%
França	2	1	0	1	50.0%	0.0%	50.0%
Futebol nacional	2	0	0	2	0.0%	0.0%	100.0%
Geração (Ã) rasca	2	0	0	2	0.0%	0.0%	100.0%
Governo	2	2	0	0	100.0%	0.0%	0.0%
Greve	2	2	0	0	100.0%	0.0%	0.0%
I Liga	2	2	0	0	100.0%	0.0%	0.0%
II Divisão	2	0	0	2	0.0%	0.0%	100.0%
Juros	2	2	0	0	100.0%	0.0%	0.0%
Kardec	2	2	0	0	100.0%	0.0%	0.0%
Legislativas	2	2	0	0	100.0%	0.0%	0.0%
Libertadores	2	1	0	1	50.0%	0.0%	50.0%
Liedson	2	0	2	0	0.0%	100.0%	0.0%
Liga ZON Sagres	2	2	0	0	100.0%	0.0%	0.0%
Liverpool	2	2	0	0	100.0%	0.0%	0.0%
Media	2	2	0	0	100.0%	0.0%	0.0%
Medicamentos	2	2	0	0	100.0%	0.0%	0.0%
Médio Oriente	2	2	0	0	100.0%	0.0%	0.0%
Moçambique	2	0	0	2	0.0%	0.0%	100.0%
Moutinho	2	2	0	0	100.0%	0.0%	0.0%
Nacional	2	0	0	2	0.0%	0.0%	100.0%
Nuclear	2	2	0	0	100.0%	0.0%	0.0%
OCDE	2	0	0	2	0.0%	0.0%	100.0%
Olhanense	2	2	0	0	100.0%	0.0%	0.0%
ONU	2	2	0	0	100.0%	0.0%	0.0%
Pandiani	2	2	0	0	100.0%	0.0%	0.0%
Passos	2	2	0	0	100.0%	0.0%	0.0%
Portagens	2	2	0	0	100.0%	0.0%	0.0%
Portugal	2	0	0	2	0.0%	0.0%	100.0%
Presidenciais	2	1	1	0	50.0%	50.0%	0.0%
PS	2	1	1	0	50.0%	50.0%	0.0%
PSP	2	0	0	2	0.0%	0.0%	100.0%
Quiosque Espanha	2	0	0	2	0.0%	0.0%	100.0%

Rali de Portugal	2	2	0	0	100.0%	0.0%	0.0%
Ralis	2	1	0	1	50.0%	0.0%	50.0%
Real	2	2	0	0	100.0%	0.0%	0.0%
Real Madrid	2	2	0	0	100.0%	0.0%	0.0%
Resgate	2	2	0	0	100.0%	0.0%	0.0%
Salvador PÃ©ndon	2	2	0	0	100.0%	0.0%	0.0%
Selecção	2	2	0	0	100.0%	0.0%	0.0%
Selecção	2	2	0	0	100.0%	0.0%	0.0%
Selecção Sub-21	2	1	1	0	50.0%	50.0%	0.0%
Sílvio	2	0	0	2	0.0%	0.0%	100.0%
Taça de Portugal	2	2	0	0	100.0%	0.0%	0.0%
Telegrama 07LISBON2771	2	0	0	2	0.0%	0.0%	100.0%
Ténis	2	2	0	0	100.0%	0.0%	0.0%
Trabalho	2	2	0	0	100.0%	0.0%	0.0%
Turismo	2	1	0	1	50.0%	0.0%	50.0%
Turquia	2	0	0	2	0.0%	0.0%	100.0%
UcrÃ¢nia	2	0	2	0	0.0%	100.0%	0.0%
Ukra	2	0	0	2	0.0%	0.0%	100.0%
Valdomiro	2	1	1	0	50.0%	50.0%	0.0%
Villas-Boas	2	1	1	0	50.0%	50.0%	0.0%
Vitória responde a Domingos	2	2	0	0	100.0%	0.0%	0.0%
Wikileaks	2	2	0	0	100.0%	0.0%	0.0%
WTCC	2	2	0	0	100.0%	0.0%	0.0%
«Angles»	1	0	0	1	0.0%	0.0%	100.0%
«Jesus de Nazaré»	1	0	0	1	0.0%	0.0%	100.0%
11 de Setembro	1	1	0	0	100.0%	0.0%	0.0%
AcadÃ©mica de Coimbra	1	1	0	0	100.0%	0.0%	0.0%
AçÃºcar	1	1	0	0	100.0%	0.0%	0.0%
Afeganistão	1	1	0	0	100.0%	0.0%	0.0%
Agricultura	1	1	0	0	100.0%	0.0%	0.0%
Alcobaça/Petróleo	1	1	0	0	100.0%	0.0%	0.0%
Álcool	1	0	0	1	0.0%	0.0%	100.0%
Algarve	1	1	0	0	100.0%	0.0%	0.0%
Almeida Santos	1	1	0	0	100.0%	0.0%	0.0%
Andebol/T.Challenge	1	1	0	0	100.0%	0.0%	0.0%
Angola	1	0	0	1	0.0%	0.0%	100.0%
Antevisão da Pública	1	1	0	0	100.0%	0.0%	0.0%
AR/Censura	1	1	0	0	100.0%	0.0%	0.0%
Árbitros	1	1	0	0	100.0%	0.0%	0.0%
Artur Agostinho ao SOL	1	0	1	0	0.0%	100.0%	0.0%
ATP Miami	1	1	0	0	100.0%	0.0%	0.0%
Áudio	1	0	0	1	0.0%	0.0%	100.0%
Austeridade	1	0	1	0	0.0%	100.0%	0.0%
Autismo	1	1	0	0	100.0%	0.0%	0.0%
Bahrein	1	1	0	0	100.0%	0.0%	0.0%
Banco Mundial	1	1	0	0	100.0%	0.0%	0.0%
Barcelona	1	1	0	0	100.0%	0.0%	0.0%
Bebidas	1	0	1	0	0.0%	100.0%	0.0%
Belenenses	1	1	0	0	100.0%	0.0%	0.0%
BPI	1	1	0	0	100.0%	0.0%	0.0%

Braga	1	1	0	0	100.0%	0.0%	0.0%
Bruxelas	1	1	0	0	100.0%	0.0%	0.0%
BTT atravessa o Saara	1	1	0	0	100.0%	0.0%	0.0%
Bwin Cup	1	0	0	1	0.0%	0.0%	100.0%
Caldas da Rainha	1	1	0	0	100.0%	0.0%	0.0%
CAN 2012	1	0	0	1	0.0%	0.0%	100.0%
Carcavelos	1	1	0	0	100.0%	0.0%	0.0%
Cardozo	1	0	1	0	0.0%	100.0%	0.0%
Caso Portucale	1	1	0	0	100.0%	0.0%	0.0%
Caso Queiroz	1	1	0	0	100.0%	0.0%	0.0%
Cavaco	1	1	0	0	100.0%	0.0%	0.0%
Clube dos Pensadores	1	0	0	1	0.0%	0.0%	100.0%
Coentrão	1	1	0	0	100.0%	0.0%	0.0%
Conjuntura	1	0	0	1	0.0%	0.0%	100.0%
Conselho das Finanças Públicas	1	1	0	0	100.0%	0.0%	0.0%
Constituição	1	1	0	0	100.0%	0.0%	0.0%
Corinthians	1	1	0	0	100.0%	0.0%	0.0%
Correntes d´Escritas	1	1	0	0	100.0%	0.0%	0.0%
CP	1	1	0	0	100.0%	0.0%	0.0%
Cristianismo	1	0	1	0	0.0%	100.0%	0.0%
DE	1	0	0	1	0.0%	0.0%	100.0%
Demissão/Governo	1	1	0	0	100.0%	0.0%	0.0%
Desert Challenge	1	1	0	0	100.0%	0.0%	0.0%
Desporto	1	0	1	0	0.0%	100.0%	0.0%
Diplomacia	1	1	0	0	100.0%	0.0%	0.0%
Eleições Sporting	1	1	0	0	100.0%	0.0%	0.0%
Energia eléctrica	1	1	0	0	100.0%	0.0%	0.0%
Escutas	1	1	0	0	100.0%	0.0%	0.0%
ESIB	1	0	0	1	0.0%	0.0%	100.0%
Estudante-voluntário	1	0	0	1	0.0%	0.0%	100.0%
Estugarda	1	1	0	0	100.0%	0.0%	0.0%
Euro2012	1	1	0	0	100.0%	0.0%	0.0%
Euro2013 (sub21)	1	1	0	0	100.0%	0.0%	0.0%
Face Oculta	1	1	0	0	100.0%	0.0%	0.0%
Fogos	1	1	0	0	100.0%	0.0%	0.0%
Fórum	1	0	0	1	0.0%	0.0%	100.0%
FPF	1	1	0	0	100.0%	0.0%	0.0%
Frederico Sousa	1	0	0	1	0.0%	0.0%	100.0%
Fucile	1	1	0	0	100.0%	0.0%	0.0%
Fukushima	1	1	0	0	100.0%	0.0%	0.0%
Fundações	1	1	0	0	100.0%	0.0%	0.0%
Futebol de praia	1	1	0	0	100.0%	0.0%	0.0%
Futebol/Particular	1	0	0	1	0.0%	0.0%	100.0%
Futsal	1	1	0	0	100.0%	0.0%	0.0%
Gasol	1	1	0	0	100.0%	0.0%	0.0%
Golfe	1	0	1	0	0.0%	100.0%	0.0%
Golo de Matías	1	1	0	0	100.0%	0.0%	0.0%
Google Carry	1	0	0	1	0.0%	0.0%	100.0%
Guarín	1	1	0	0	100.0%	0.0%	0.0%
Guimarães	1	1	0	0	100.0%	0.0%	0.0%

Guiné-Bissau	1	0	1	0	0.0%	100.0%	0.0%
HÃ©lder Postiga	1	1	0	0	100.0%	0.0%	0.0%
Haiti/Eleições	1	1	0	0	100.0%	0.0%	0.0%
Helton	1	1	0	0	100.0%	0.0%	0.0%
História	1	1	0	0	100.0%	0.0%	0.0%
Holanda	1	0	0	1	0.0%	0.0%	100.0%
Ibrahimovic	1	1	0	0	100.0%	0.0%	0.0%
Iémen	1	1	0	0	100.0%	0.0%	0.0%
Igreja/Espiritualidade	1	1	0	0	100.0%	0.0%	0.0%
Imobiliário	1	0	1	0	0.0%	100.0%	0.0%
IncÃªndios	1	1	0	0	100.0%	0.0%	0.0%
Indústrias criativas	1	1	0	0	100.0%	0.0%	0.0%
Inflação	1	1	0	0	100.0%	0.0%	0.0%
Insólito	1	0	0	1	0.0%	0.0%	100.0%
Irão	1	1	0	0	100.0%	0.0%	0.0%
Irlanda/Eleições	1	1	0	0	100.0%	0.0%	0.0%
Jara	1	1	0	0	100.0%	0.0%	0.0%
Javi García	1	1	0	0	100.0%	0.0%	0.0%
Jerónimo	1	1	0	0	100.0%	0.0%	0.0%
Juros em mínimos	1	1	0	0	100.0%	0.0%	0.0%
Laurentino	1	1	0	0	100.0%	0.0%	0.0%
Liga dos Campeões	1	1	0	0	100.0%	0.0%	0.0%
Louçã	1	1	0	0	100.0%	0.0%	0.0%
Louçã	1	1	0	0	100.0%	0.0%	0.0%
Luís Duque	1	1	0	0	100.0%	0.0%	0.0%
Maisfutebol na TVI24	1	0	0	1	0.0%	0.0%	100.0%
Marítimo	1	1	0	0	100.0%	0.0%	0.0%
Metro	1	0	0	1	0.0%	0.0%	100.0%
Mexia	1	0	0	1	0.0%	0.0%	100.0%
Ministro	1	0	0	1	0.0%	0.0%	100.0%
Misericórdias	1	1	0	0	100.0%	0.0%	0.0%
Moção	1	0	0	1	0.0%	0.0%	100.0%
Moção de censura	1	1	0	0	100.0%	0.0%	0.0%
Moção/BE	1	1	0	0	100.0%	0.0%	0.0%
Moção/Bloco Esquerda	1	1	0	0	100.0%	0.0%	0.0%
Mogadouro	1	0	0	1	0.0%	0.0%	100.0%
Moratti	1	1	0	0	100.0%	0.0%	0.0%
Motociclismo	1	1	0	0	100.0%	0.0%	0.0%
Mundialito de clubes	1	1	0	0	100.0%	0.0%	0.0%
Nadal	1	1	0	0	100.0%	0.0%	0.0%
Naval 1Âº Maio	1	1	0	0	100.0%	0.0%	0.0%
NBA	1	1	0	0	100.0%	0.0%	0.0%
NigÃ©ria/Eleições	1	1	0	0	100.0%	0.0%	0.0%
Nova Zelândia	1	1	0	0	100.0%	0.0%	0.0%
Nova Zelândia/Sismo	1	0	0	1	0.0%	0.0%	100.0%
Novo alerta	1	1	0	0	100.0%	0.0%	0.0%
Obama	1	1	0	0	100.0%	0.0%	0.0%
Oeiras	1	0	0	1	0.0%	0.0%	100.0%
P. Ferreira	1	1	0	0	100.0%	0.0%	0.0%
Paços 2-0 SetÃºbal	1	1	0	0	100.0%	0.0%	0.0%

Paços de Ferreira	1	1	0	0	100.0%	0.0%	0.0%
Palermo	1	0	0	1	0.0%	0.0%	100.0%
Parlamento	1	1	0	0	100.0%	0.0%	0.0%
Penafiel	1	1	0	0	100.0%	0.0%	0.0%
Pilar Ribeiro	1	0	0	1	0.0%	0.0%	100.0%
PJ	1	1	0	0	100.0%	0.0%	0.0%
Portalegre	1	1	0	0	100.0%	0.0%	0.0%
Portugal-Argentina	1	1	0	0	100.0%	0.0%	0.0%
PPP	1	1	0	0	100.0%	0.0%	0.0%
Presidente do IFO	1	1	0	0	100.0%	0.0%	0.0%
Presidente GM	1	0	1	0	0.0%	100.0%	0.0%
Professores	1	1	0	0	100.0%	0.0%	0.0%
PSG-Benfica	1	1	0	0	100.0%	0.0%	0.0%
Racismo	1	0	0	1	0.0%	0.0%	100.0%
Ranking ATP	1	0	1	0	0.0%	100.0%	0.0%
Redknapp	1	1	0	0	100.0%	0.0%	0.0%
Refugiados	1	0	0	1	0.0%	0.0%	100.0%
Regiões	1	1	0	0	100.0%	0.0%	0.0%
Rehn	1	0	0	1	0.0%	0.0%	100.0%
Reino Unido	1	1	0	0	100.0%	0.0%	0.0%
Religião	1	1	0	0	100.0%	0.0%	0.0%
Remo	1	0	1	0	0.0%	100.0%	0.0%
Ricardo Batista	1	1	0	0	100.0%	0.0%	0.0%
Roma	1	1	0	0	100.0%	0.0%	0.0%
Ronaldo	1	1	0	0	100.0%	0.0%	0.0%
Rússia	1	1	0	0	100.0%	0.0%	0.0%
Santana Lopes	1	1	0	0	100.0%	0.0%	0.0%
Santos Silva	1	1	0	0	100.0%	0.0%	0.0%
São TomÃ© e Príncipe	1	1	0	0	100.0%	0.0%	0.0%
Schalke	1	1	0	0	100.0%	0.0%	0.0%
Scolari	1	1	0	0	100.0%	0.0%	0.0%
Scut	1	1	0	0	100.0%	0.0%	0.0%
Seguros	1	1	0	0	100.0%	0.0%	0.0%
Seixal	1	0	0	1	0.0%	0.0%	100.0%
Sidnei	1	1	0	0	100.0%	0.0%	0.0%
SL Benfica	1	1	0	0	100.0%	0.0%	0.0%
Slayer sofrem nova baixa	1	1	0	0	100.0%	0.0%	0.0%
Sp. Braga	1	1	0	0	100.0%	0.0%	0.0%
Sporting-Rangers	1	1	0	0	100.0%	0.0%	0.0%
Submarino	1	1	0	0	100.0%	0.0%	0.0%
Surf	1	1	0	0	100.0%	0.0%	0.0%
Taça	1	1	0	0	100.0%	0.0%	0.0%
Taça Challenge	1	1	0	0	100.0%	0.0%	0.0%
Taça do Rei	1	0	0	1	0.0%	0.0%	100.0%
Taça Inglaterra	1	1	0	0	100.0%	0.0%	0.0%
Taça Portugal	1	1	0	0	100.0%	0.0%	0.0%
Telegrama 08LISBON433	1	0	0	1	0.0%	0.0%	100.0%
Timor-Leste	1	1	0	0	100.0%	0.0%	0.0%
Torneio do Algarve Sub-17	1	1	0	0	100.0%	0.0%	0.0%
Tottenham-Real	1	1	0	0	100.0%	0.0%	0.0%

Transportes	1	1	0	0	100.0%	0.0%	0.0%
Trichet	1	1	0	0	100.0%	0.0%	0.0%
UEFA	1	1	0	0	100.0%	0.0%	0.0%
URGENTE	1	0	0	1	0.0%	0.0%	100.0%
URGENTE Açores/pesqueiro	1	1	0	0	100.0%	0.0%	0.0%
URGENTE/Futebol	1	0	0	1	0.0%	0.0%	100.0%
Valdano	1	0	0	1	0.0%	0.0%	100.0%
Valência	1	1	0	0	100.0%	0.0%	0.0%
Vela	1	0	0	1	0.0%	0.0%	100.0%
Venezuela	1	1	0	0	100.0%	0.0%	0.0%
Viana	1	0	0	1	0.0%	0.0%	100.0%
Viana do Castelo	1	0	0	1	0.0%	0.0%	100.0%
Vieira	1	0	0	1	0.0%	0.0%	100.0%
Vila do Conde	1	1	0	0	100.0%	0.0%	0.0%
Villas Boas	1	0	0	1	0.0%	0.0%	100.0%
Viseu	1	0	0	1	0.0%	0.0%	100.0%
Vitória de Guimarães	1	0	1	0	0.0%	100.0%	0.0%
Vitória de Setúbal	1	1	0	0	100.0%	0.0%	0.0%
Voluntariado	1	1	0	0	100.0%	0.0%	0.0%
WRC	1	1	0	0	100.0%	0.0%	0.0%
Xavi	1	1	0	0	100.0%	0.0%	0.0%
Zapatero	1	1	0	0	100.0%	0.0%	0.0%

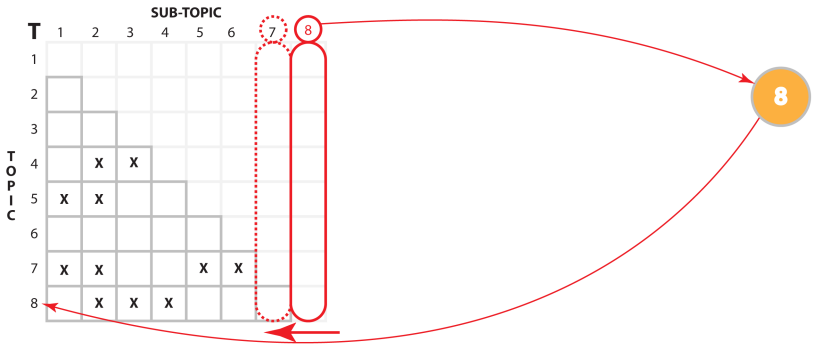
Appendix C

Tree Construction Example

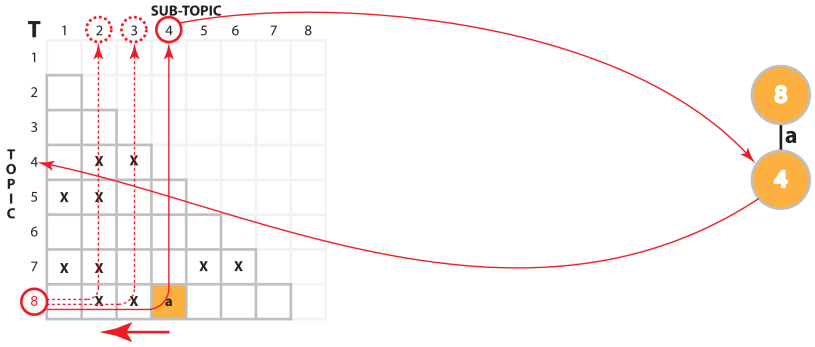
This section details and exemplifies the construction of hierarchical topic tree, according to this approach.

Tree Construction Example

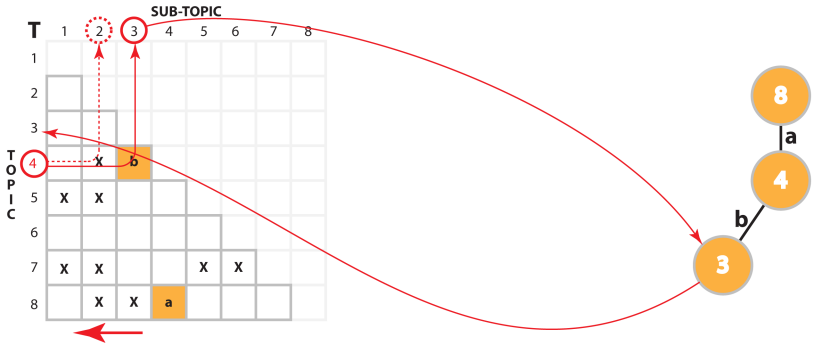
#1
Find a *root* (i.e. a *sub-topic* that doesn't belong to any *topic* - empty column).



#2
Find the first *sub-topic* of that *root* (i. e. the largest tag that belongs to that *topic*).



#3
Taking the previous *sub-topic* as a *topic* find the first *sub-topic* (i. e. apply the same process recursively).



#4
If there is no *sub-topic* of a given *topic* retreat one level and repeat the process.

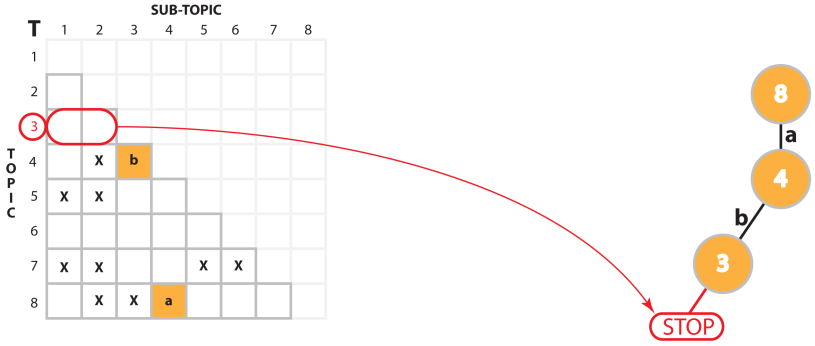
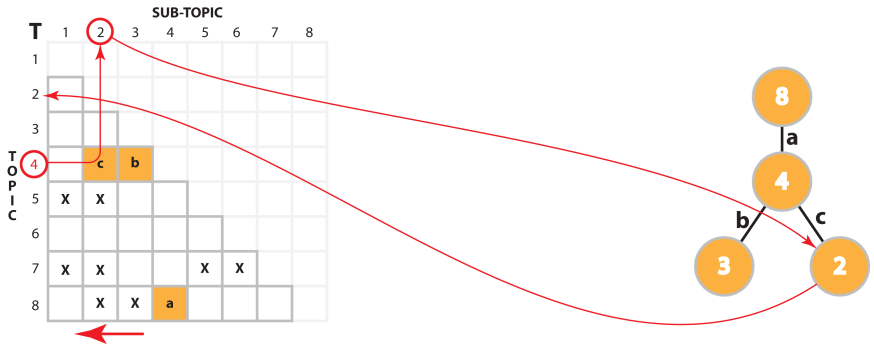
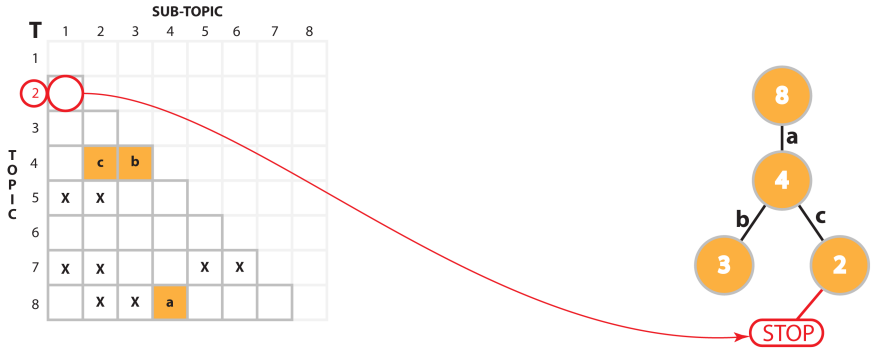


Figure C.1: Depiction of topic hierarchical trees construction process, steps 1 to 4 of the tree construction example.

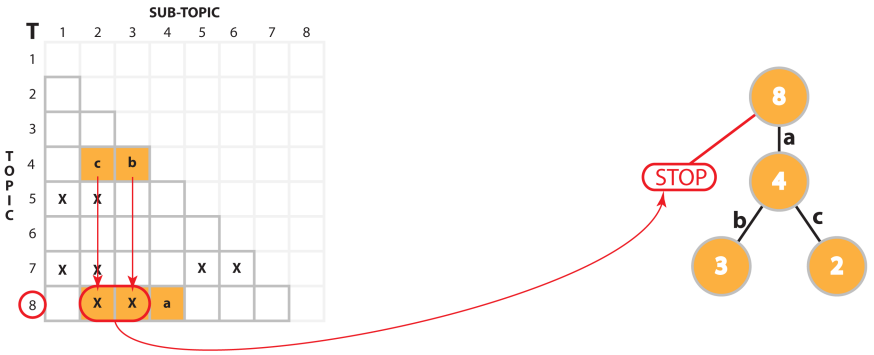
#5
Find all *sub-topics* of a given topic repeating the same process recursively.



#6
As in step #4 retreat one level if there is no *sub-topic* left.



#7
If the root has no more *sub-topics* or if the remainging *sub-topics* had already been used elsewhere in the tree, the tree is complete.



#8
Find another *root* as in step #1.

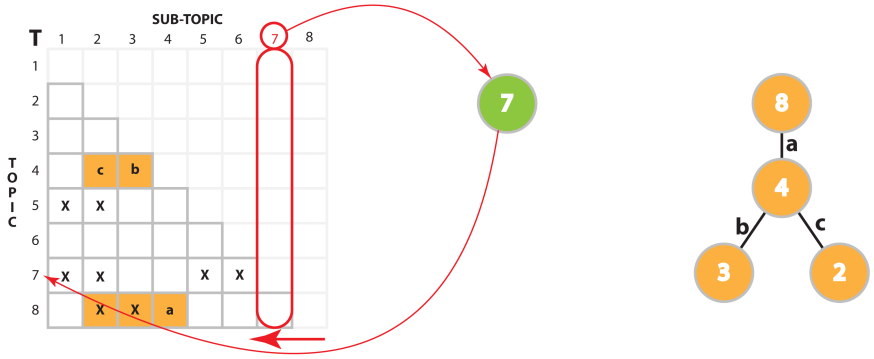
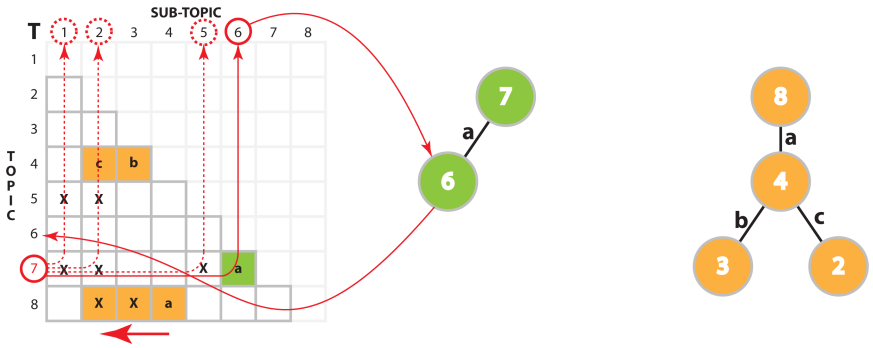
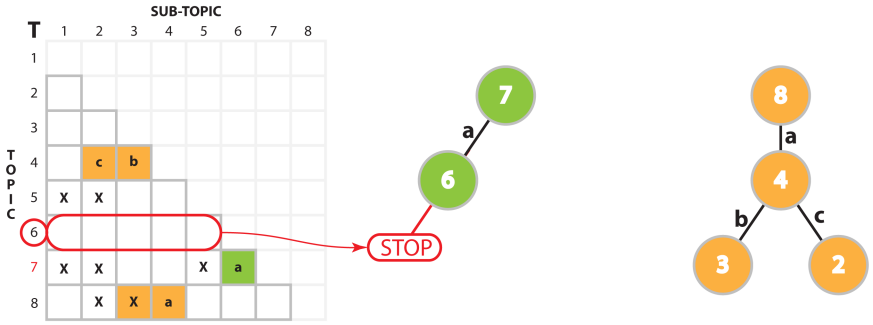


Figure C.2: Depiction of topic hierarchical trees construction process, steps 5 to 8 of the tree construction example.

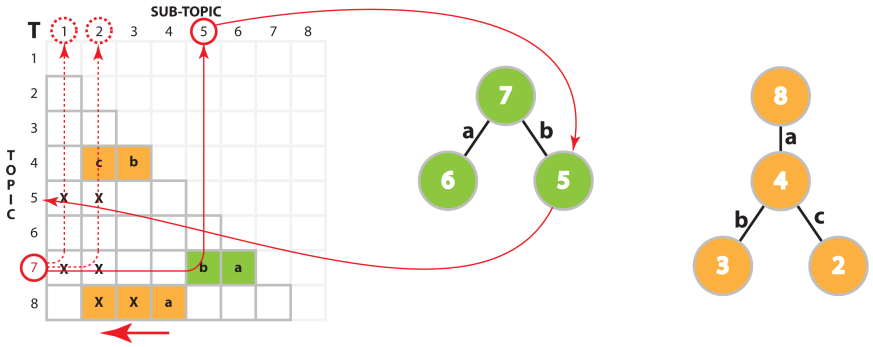
#9
This *root* (7) has four candidate *sub-topics* (6, 5, 2, 1).



#10
Starting with the largest (6) try to find any *sub-topics* of *tag* 6. As there is none go back to tag 7 (i. e. one level up).



#11
Repeat the process for *tag* 5.



#12
Tag 5 has two candidate *sub-topics* (2, 1). The process is repeated starting with the largest *tag* candidate (2).

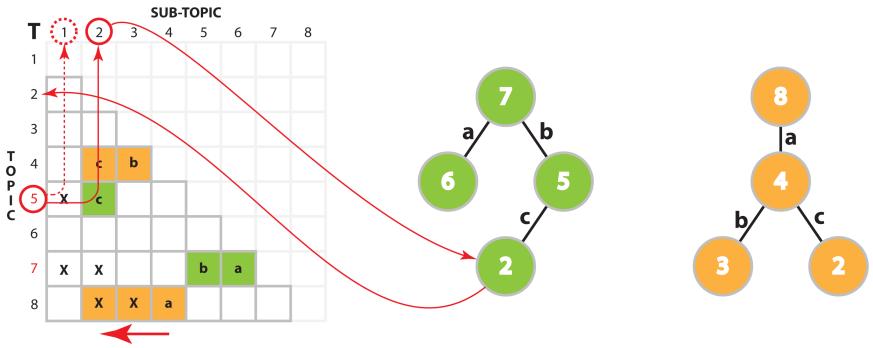
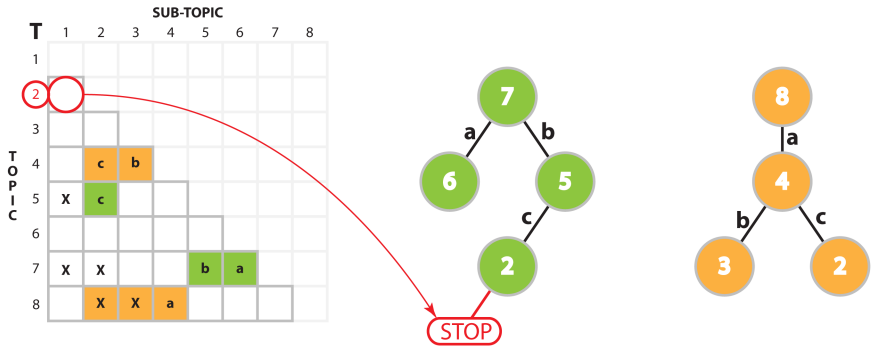
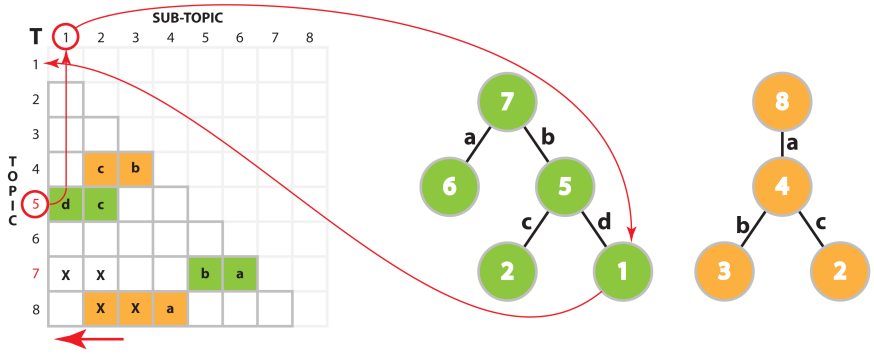


Figure C.3: Depiction of topic hierarchical trees construction process, steps 9 to 12 of the tree construction example.

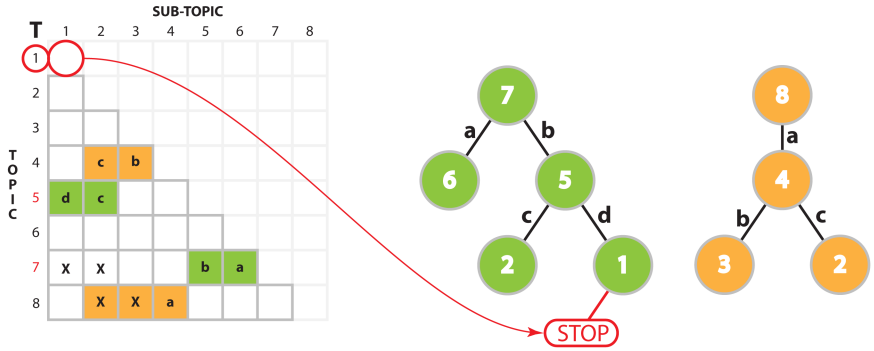
#13
Tag 2 has no *sub-topic* so go back one level (i. e. *tag* 5).



#14
Tag 5 has one *sub-topic* more (*tag* 1).



#15
Tag 1 has no *sub-topics*, so go back one level to *tag* 5. All *sub-topics* of *tag* 5 have been processed so go back one level to *tag* 7 (in this example, the *root*).



#16
Although there are still two *tags* left, they have already been used in the same tree (green tree). So this tree is complete.

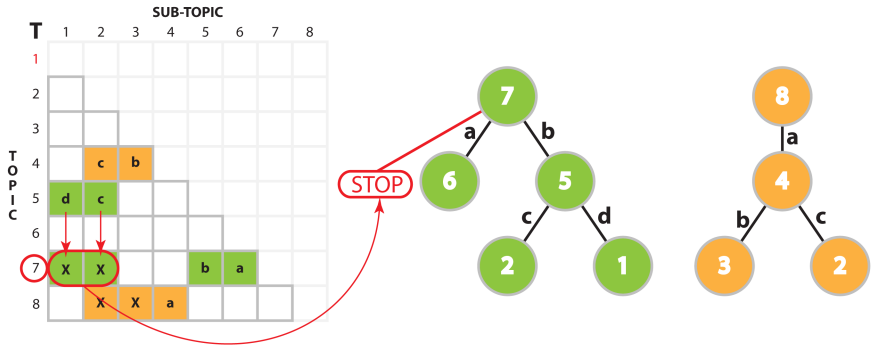


Figure C.4: Depiction of topic hierarchical trees construction process, steps 13 to 16 of the tree construction example.

Appendix D

All Super-topics and Related Tags

PEC:

Grupo E	1	15	1615
Gaia	1	6	1615
Norte/Conjuntura	3	3	1615
Silva Pereira	4	7	1615
EleiÃ§Ãµes legislativas	2	3	1615
NBA	1	6	1615
Ensino BÃ¡sico	1	8	1615
JerÃ³nimo	11	20	1615
Liga Europa EM DIRECTO	1	1	1615
URGENTE	3	31	1615
TGV	3	4	1615
Ã‚ngelo PaupÃ©rio	11	29	1615
Eurogrupo	3	4	1615
OCDE	1	7	1615
CAN 2012	1	4	1615
Conjuntura	15	31	1615
Aumento do IVA	6	6	1615
BE/ConvenÃ§Ã£o	1	1	1615
PS/Congresso	24	51	1615
Real	1	47	1615
Crise	19	58	1615
AR/Censura	4	44	1615
Assis	11	13	1615
MoÃ§Ã£o de censura	2	15	1615
Sporting	8	362	1615
UE/Cimeira	22	60	1615
Kardec	1	13	1615
Sporting-U. Leiria (antevisÃ£o)	1	1	1615
Flash/ Governo/DemissÃ£o	1	3	1615
PetrÃ³leo	5	19	1615
Passos	18	60	1615
GrÃ©cia	3	14	1615
PrivatizaÃƒÂ§ÃƒÂµes	1	1	1615
JustiÃ§a	1	15	1615
Ã‚Â«Discurso DirectoÃ‚Â»	7	16	1615
JapÃ£o	17	69	1615
Macau	2	2	1615
Benzema	2	21	1615
Teixeira dos Santos	3	7	1615
LouÃƒÂ§ÃƒÂ£	4	6	1615
DÃ-vida	8	59	1615
EleiÃ§Ãµes do PS	1	2	1615
Ferreira Leite	3	3	1615
Marcelo	19	53	1615
RegiÃƒÂµes	1	5	1615
Media	1	13	1615
Benfica	2	64	1615
GuarÃƒÂ-n	2	8	1615

Fundo do euro	1	1	1615
PCP	37	83	1615
Medicamentos	10	17	1615
Congresso PS	10	23	1615
URGENTE/EleiÃ§Ãµes	2	3	1615
AvaliaÃ§Ã£o dos Professores	1	2	1615
Relvas	1	1	1615
SÃƒÂ³crates demite-se	3	3	1615
Super-Lua cheia	1	1	1615
Ajuda	4	9	1615
Governo/DemissÃ£o	152	411	1615
DemissÃ£o/Governo	2	16	1615
Caerlos CÃƒÂ©sar	1	2	1615
125cc	1	25	1615
ConcertaÃ§Ã£o Social	3	13	1615
U. Leiria	1	11	1615
CovilhÃ£	2	4	1615
RegionalizaÃ§Ã£o	1	5	1615
JosÃƒÂ© Costa	4	7	1615
Telegrama 07LISBON2771	3	32	1615
Drenthe	1	6	1615
Valdomiro	1	9	1615
TaÃ§a da Liga	1	9	1615
Agricultura	6	18	1615
WTA Monterrey (MÃ©xico)	1	3	1615
Sp. Braga	1	35	1615
Novos cortes	5	5	1615
SchÃ¤uble	3	6	1615
ArmÃ©nia	1	1	1615
EUA	2	51	1615
Zapatero	1	3	1615
NotÃ-cia SÃBADO	15	32	1615
EleiÃƒÂ§ÃƒÂµes	7	48	1615
Facebook e Twitter	4	42	1615
BE	13	47	1615
ValÃªncia	1	3	1615
Costa do Marfim	1	44	1615
Crise polÃ-tica	8	11	1615
EleiÃ§Ãµes na EstÃ³nia	2	4	1615
MoÃ§ambique	5	9	1615
Portugal/Brasil	6	75	1615
Futuros	4	9	1615
BE e PCP	1	1	1615
ResoluÃ§Ãµes contra o PEC	1	1	1615
Entrevista a Dias Ferreira	1	2	1615
Governo	11	32	1615
Madeira	10	32	1615
PCP/90 anos	11	11	1615
TDT	3	4	1615
Amado	3	7	1615

GP Qatar	1	7	1615
CPTT	1	2	1615
CGTP/Manif	5	6	1615
Viseu	1	4	1615
Carrilho	1	2	1615
CHAT	2	9	1615
CombustÃ-veis	8	27	1615
Portas	2	5	1615
Espanha	9	68	1615
LacÃƒÂ£o preocupado	2	2	1615
Barclays	1	1	1615
Marcelo R. de Sousa	1	2	1615
Greve/Transportes	1	56	1615
Cavaco Silva esclarece	1	9	1615
Caso Portucale	2	3	1615
Sondagem TVI	7	9	1615
Telegrama 07LISBON2806	1	2	1615
PCP e BE	1	1	1615
Ministro	20	57	1615
FMI	2	16	1615
Congresso PS ao minuto	2	3	1615
PetrÃ³leo sobe	1	2	1615
Transportes	4	27	1615
Imprensa venezuelana	1	1	1615
Ramires	1	2	1615
Barcelona	3	39	1615
Banca	3	33	1615
Quiosque	4	145	1615
CDS-PP	12	25	1615
Professores	13	30	1615
Almeida Santos	1	6	1615
Sondagem	2	2	1615
Soares	2	5	1615
Portugal	8	16	1615
PEC 4	7	7	1615
Euro/Crise	33	199	1615
URGENTE/LÃ-bia	1	10	1615
I	3	9	1615
LÃ-bia	38	273	1615
JN	1	31	1615
Dia D	6	7	1615
CDS	1	6	1615
Desporto	7	72	1615
Passos Coelho	9	32	1615
Cavaco	2	19	1615
CP	1	8	1615
Bruxelas	1	5	1615
Abono de famÃ-lia	2	2	1615
Energia	5	35	1615
Igreja/Sociedade	1	1	1615

Couceiro	2	28	1615
SCUT	1	5	1615
Austeridade	6	7	1615
PÃºblico	3	9	1615
PSD	83	154	1615
Legislativas	8	12	1615
PS	32	88	1615
MaurÃƒÂ-cio	2	5	1615
Cabo Verde	1	9	1615
Esqui Alpino	1	1	1615
Urgente/Governo	2	8	1615
Futebol	1	416	1615
Ao minuto	5	9	1615
PolÃ-tica	4	4	1615
Telegrama 07LISBON821	1	1	1615
Assis para SÃƒÂ³crates	3	5	1615
Golfe	4	13	1615
OE2011	1	47	1615
PEC	420	420	1615
EducaÃ§Ã£o	20	57	1615
Greve	1	40	1615
Vaclav Havel	2	2	1615
Ministra espanhola	4	4	1615
Wenger	1	14	1615
DÃ©fice 2010	2	3	1615
Eurostat	2	5	1615
MilÃ£o-Sanremo	1	4	1615
PR/Posse	29	31	1615
Telegrama 09LISBON136	2	4	1615
BPN	3	5	1615
LogÃ-stica	1	1	1615
EleiÃ§Ãµes	23	70	1615
Santos Silva	7	14	1615
BÃ¡sico	1	1	1615
Bank of America	2	2	1615
DE	2	10	1615
Â«Os VerdesÂ»	1	1	1615
SaÃºde	4	15	1615
PensÃµes mÃ-nimas	2	2	1615
2011-04-09T15	2	5	1615
Lula	7	62	1615
Juncker	16	97	1615
Ajuda externa	19	68	1615
Belluschi	3	11	1615
Battle	1	2	1615

Appendix E

Super-topic Graphs

This section shows the Super-topic Graphs for the all the super-topics, for 90%, 85%, 80%, 75%, 70% and 65% topic similarity.

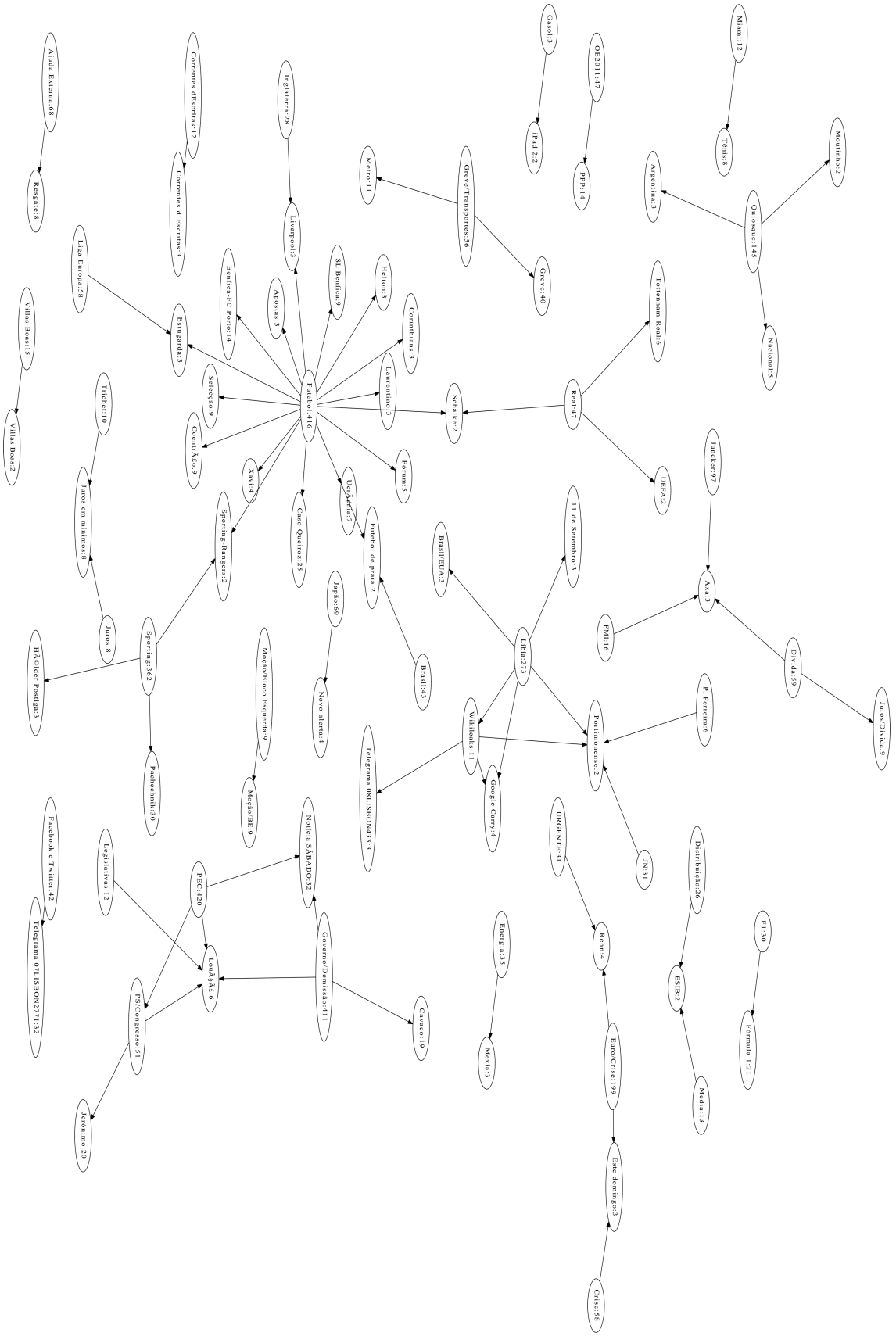


Figure E.1: Graph of super-topics with similarity greater than 90%.

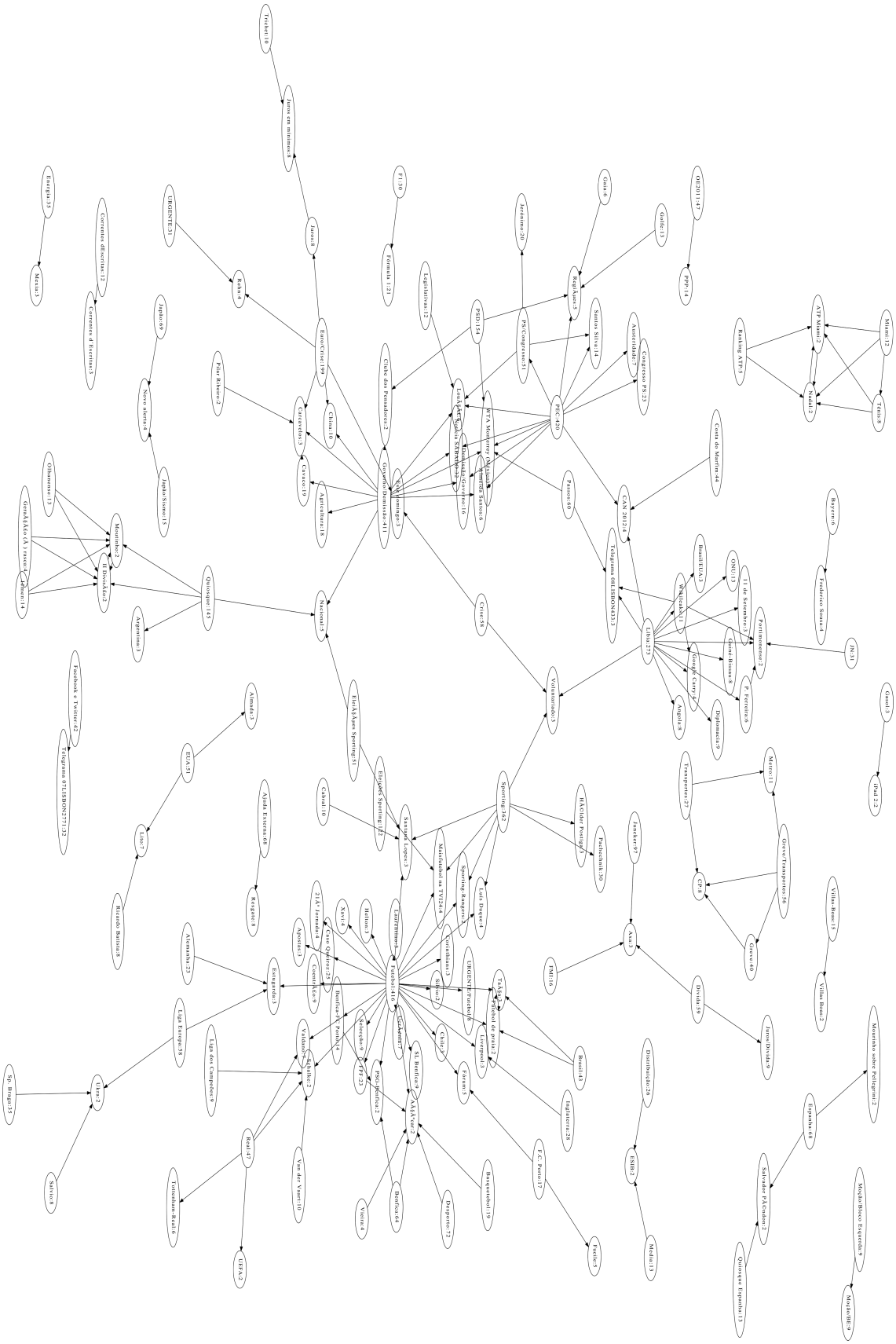


Figure E.2: Graph of super-topics with similarity greater than 85%.

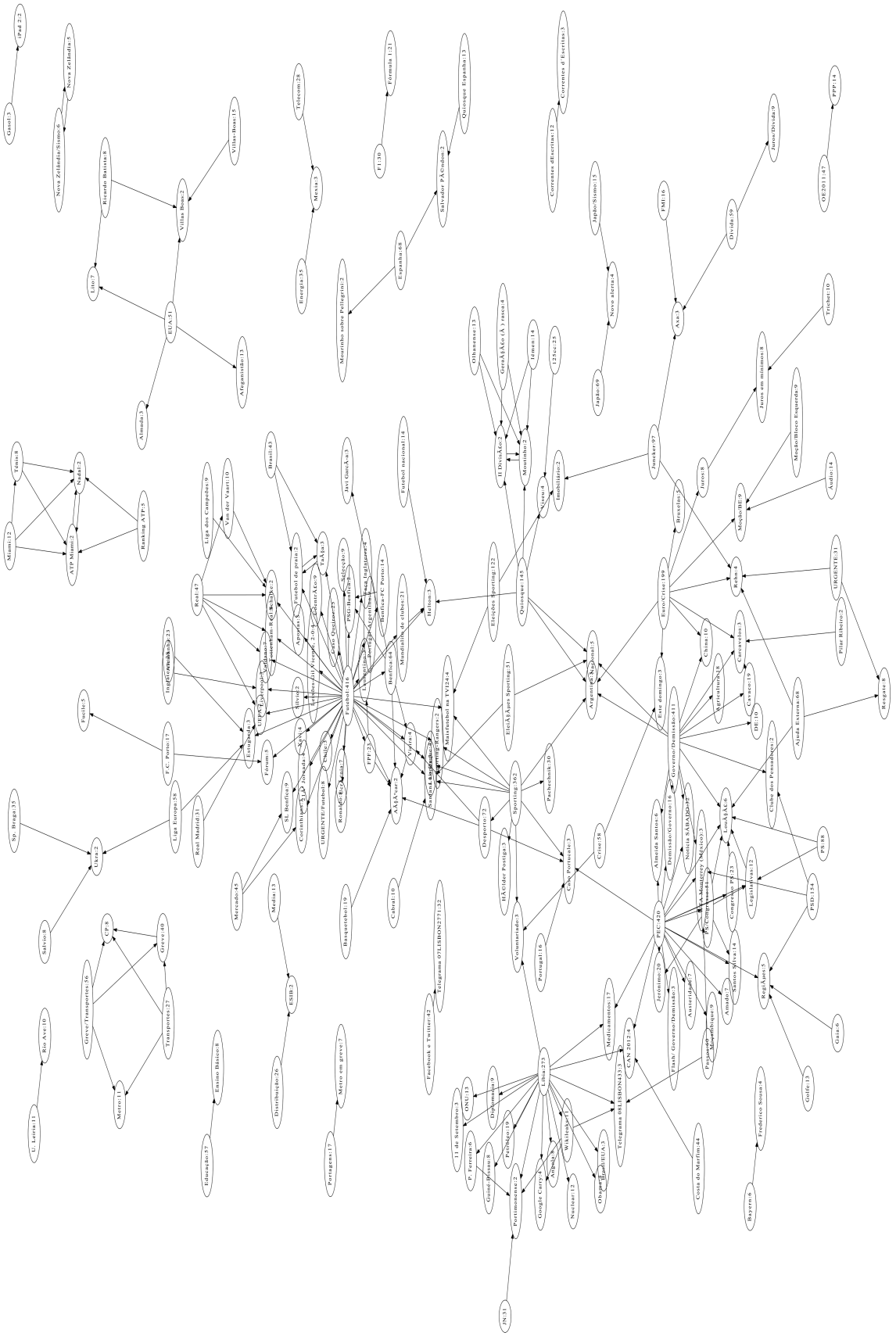


Figure E.3: Graph of super-topics with similarity greater than 80%.

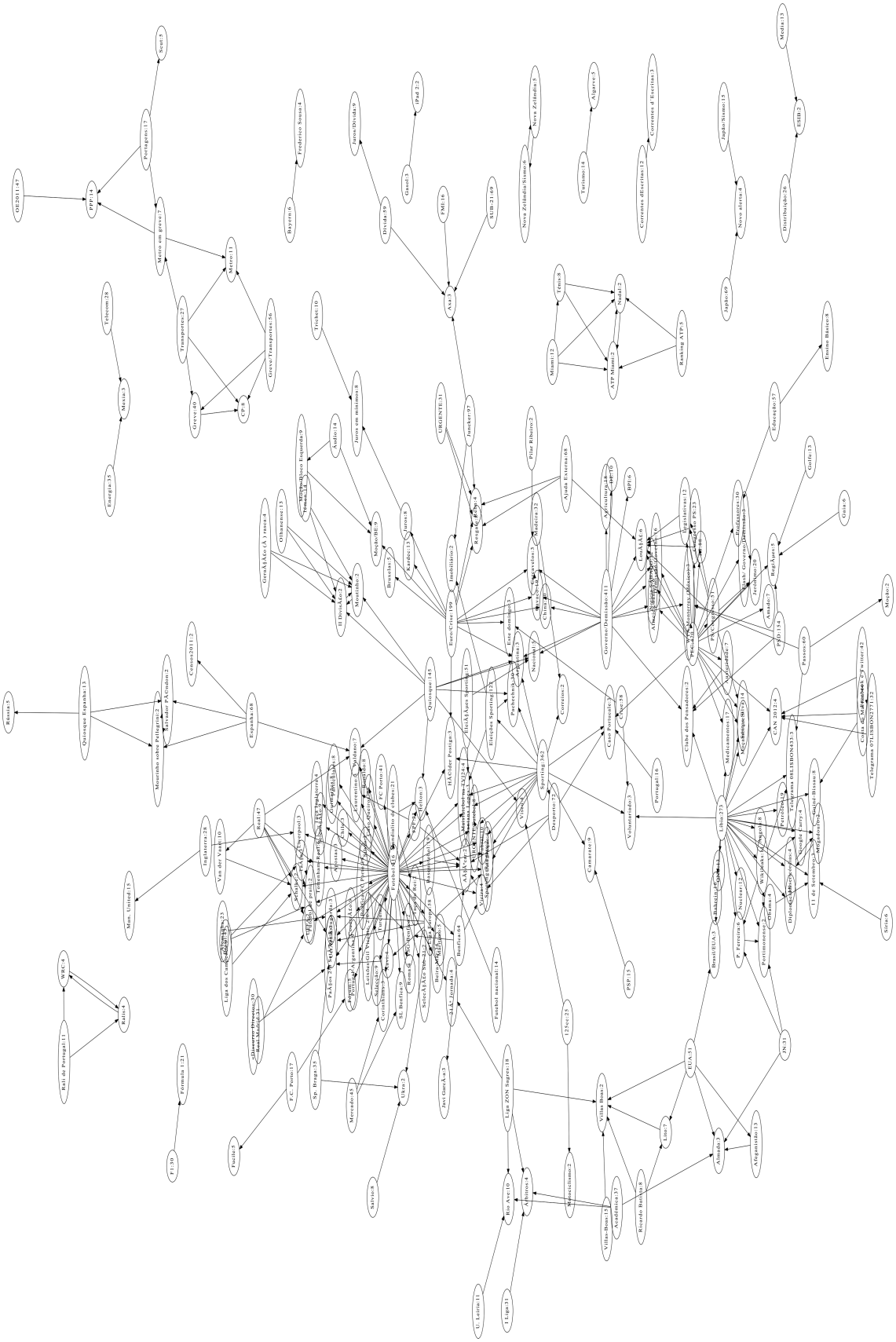


Figure E.4: Graph of super-topics with similarity greater than 75%.

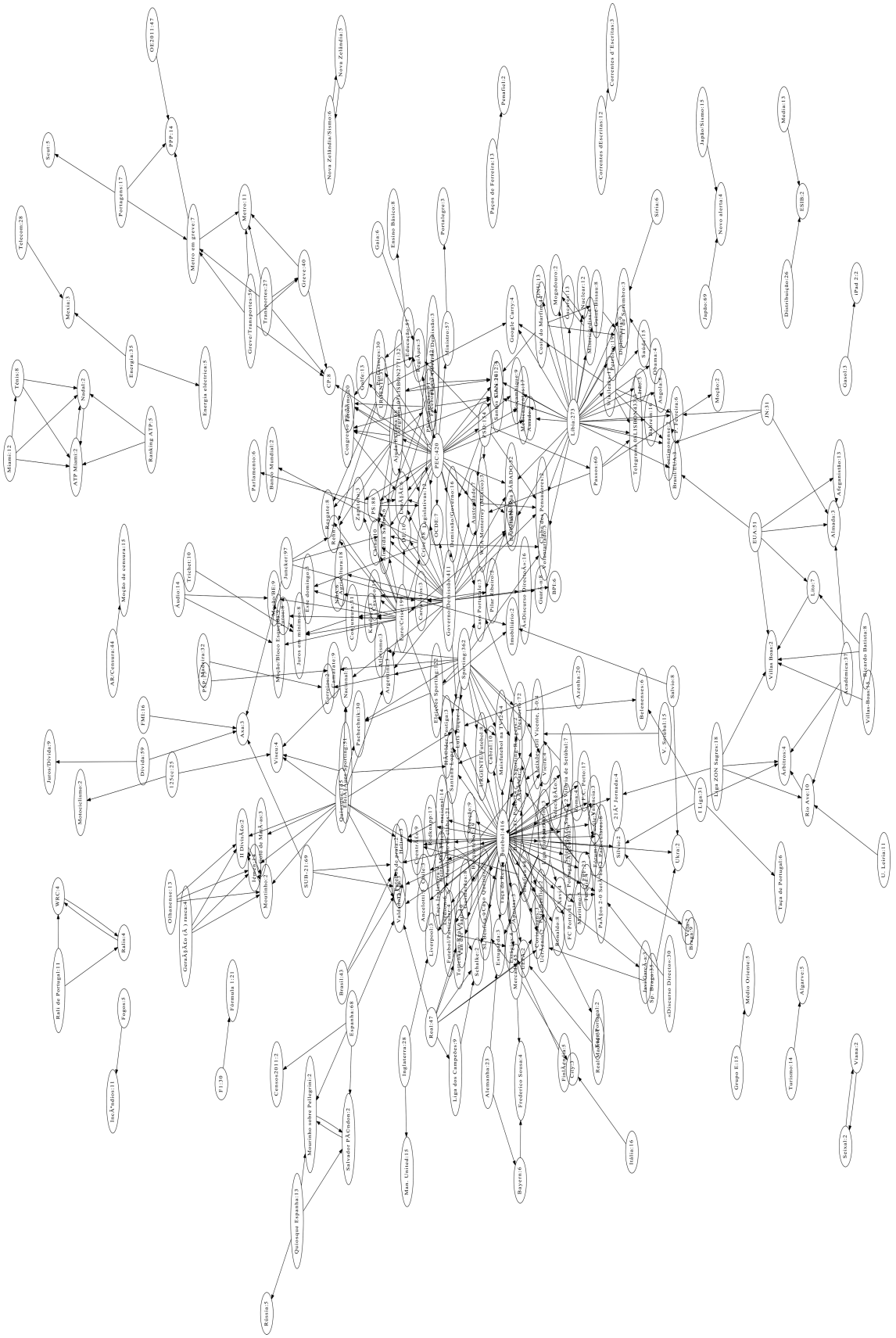


Figure E.5: Graph of super-topics with similarity greater than 70%.

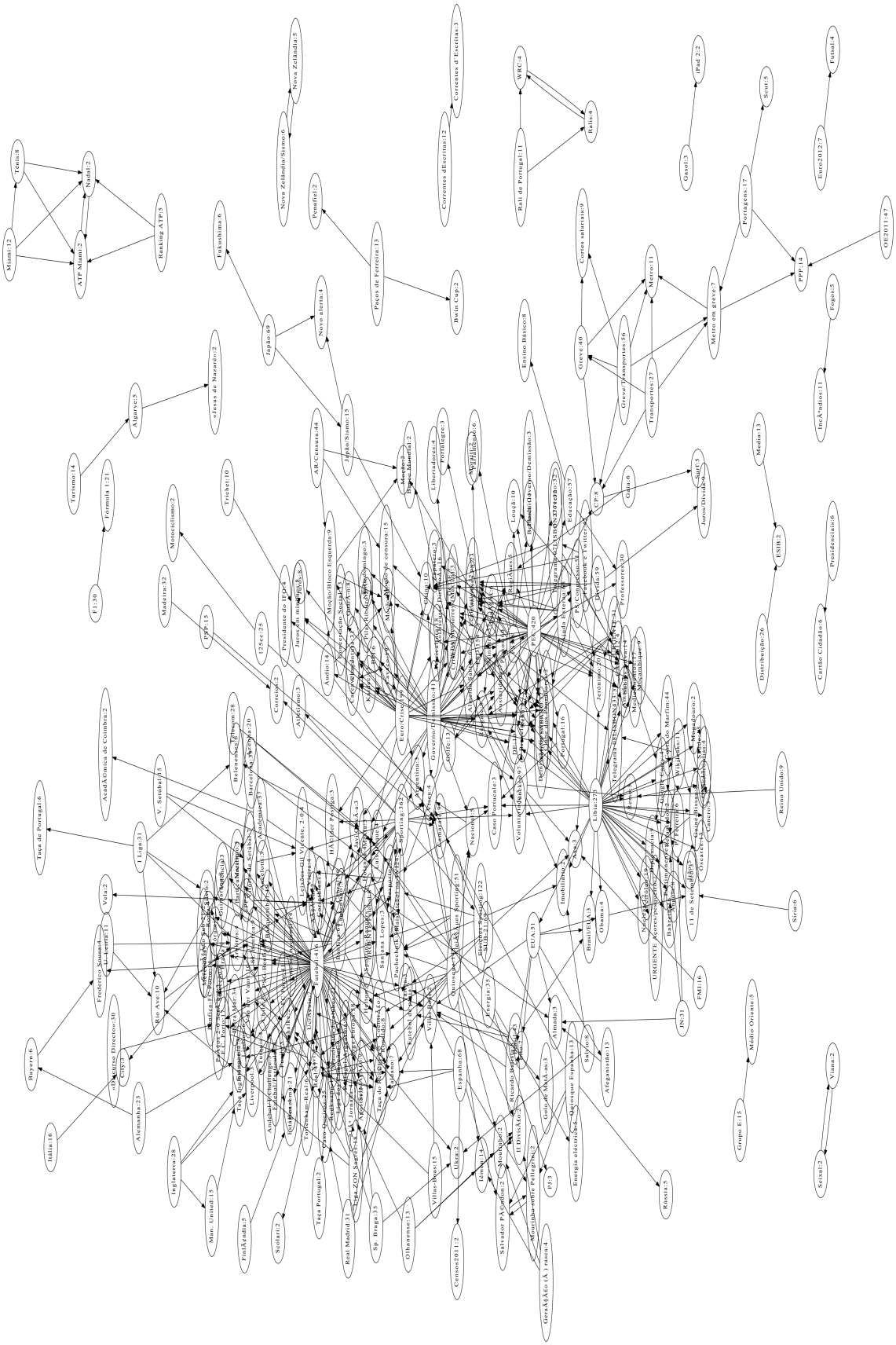


Figure E.6: Graph of super-topics with similarity greater than 65%.