Laying bare EuroVoc by means of Latent Dirichlet Allocation

Giovanni Siragusa and Luigi Di Caro

Department of Computer Science
Via Pessinetto, 12
University of Turin, Italy
{dicaro,siragusa}@di.unito.it

Abstract. *EuroVoc* is a 23-languages thesaurus that covers the activity of the European Union. It contains more than 6000 descriptor terms that are organized in a 8-levels hierarchy. The main advantage of *EuroVoc* consists in the assignment of a set of descriptors to documents, which can be used to perform documentary search, document management and cross-language alignment. Unfortunately, descriptor terms are not in mutual exclusion with each others, sharing contextual similarities that may lead to skewed data, especially in document classification. In this paper, we present a corpus analysis via Latent Dirichlet Allocation in order to study whether and how *EuroVoc* descriptors capture documents' true content. In particular, we compared extracted topics with descriptors through a number of proposed measured on a bipartite graph.

Keywords: EuroVoc, Latent Dirichlet Allocation, topics

1 Introduction

The web-page [7] describes *EuroVoc* as a multi-language and multi-disciplinary thesaurus covering the activities of the European Union (EU). Managed by the Publication Office which moved forward to ontologies and semantic web, the current version of *EuroVoc*, 4.5, exists in 23 EU languages and aims to allow management of documents and documentary search. It contains large fields to cover both community and national points of view.

The main advantage of *EuroVoc* relies on the semantic web. It provides both XML and RDF schemas to structure and exchange structured meta-data and information between applications. In details, *Eurovoc* is composed by more than 6000 descriptor terms that are organized hierarchically into 8 levels. The terms are organized with 3 relationships: *broader terms* (BT), *narrower terms* (NT) and *related terms* (RT).

However, this structure has several disadvantages. Descriptor terms are not in mutual exclusion with each others, but they share contextual similarities. For instance, it can be difficult to recognize *gender identity* descriptor (uri c_558c1e00) from *gender mainstreaming* descriptor (uri c_b19d7503) in a document, because the information are not concentrated in a paragraph (or in a consecutive sequence of paragraphs), but they are dilute in entire documents. In other words, it is difficult to recognize to which category a document belongs. Additionally, documents can concern several arguments. For these reasons, *EuroVoc* allows to specify a set of descriptors to every document (see Listing 1.1 for details). But the choice of assigning multiple labels to a document can be problematic: corpora are, by their nature, finite, and some labels, especially when there is a large set of labels, may have few or, in the worst case, no associated documents. As reported in [4], this may lead to skewed data that are problematic for different application, like classification of legal texts. Moreover, Boella et al. [4] and Steinberger et al. [5] shows that each document contains from 2 to 17 descriptors with a pick near on 6. Further, only 1733 of more than 6000 descriptors are used only 5 times.

In this paper, we present our proposal: to apply *Latent Dirichlet Allocation* (LDA) to *EuroVoc*. Our assumption is there exists a one-to-one relation between a *EuroVoc* category and a topic. To validate our assumption, First we apply LDA on a large EuroVoc-based corpus to extract a set of topics equals to the number of *EuroVoc* categories at the third (mid) level contained in the corpus. Secondly, we compare topics (as Bag-of-Words) with *EuroVoc* categories (as Bag-of-Words) using cosine similarity in order to associate each *EuroVoc* category to the most similar topic. This allowed us to construct a bipartite graph, where topics and categories are the nodes whereas the associations between topics and categories are the edges. Finally, we evaluated the bipartite graph using 3 different metrics. The 3 metrics allowed us to understand how much *EuroVoc* categories reflect the true content of documents labeled with them. For example, a *EuroVoc* category can be associated to more than one topic. In such case, the category does not capture the true content expressed in the documents.

The remainder of this paper is organized as follows. In Section 2 we present a brief review about *Latent Dirichlet Allocation* (LDA) and *EuroVoc*. In Section 3 we describe the corpus where we applied LDA in order to analyse the *EuroVoc* categories. In details, we will describe where the corpus can be found, how many documents it contains, how it is structured and so forth. In Section 4 we describe the method we applied to study the *EuroVoc* vocabulary. Finally, in Section 5 and in Section 6 we describe the results and our conclusions.

2 Related Works

Our project relies on two State-of-the-Art aspects. On one hand, there exist works based on *EuroVoc*, such as the work presented by Boella et al.[4] and Steinberger et al. [6, 5]. On the other hand, there exist works based on *Latent Dirichlet Allocation* to model topics expressed in corpora and their evolution over time, like the works proposed by Cui et al. [2], Wang and McCallum [3] and Blei et al. [9].

Boella et al. proposed, in their work [4], a classifier based on Support Vector Machine (SVM) to assign *EuroVoc* labels to documents. Their method first transforms a multi-label text corpus into a mono-label one. Then, the mono-label corpus is used to train the classifier. In their work, they assume that an n-labelled document is a compact version of n different documents, each one related to a single label. Thus, their method virtually splits a document in a set of virtual documents and constructs a numerical vector for each one using the Pointwise Mutual Information and a function *sel* which maps each feature f into f itself or 0. Finally, the new set of virtual documents is passed in input to the SVM. In [5], Steinberger et al. proposed JEX, a system for the classification of texts into *Eurovoc* categories. The core of their approach is based on the construction of class profiles, that is sets of terms that define the related categories. The classification step is done by computing the cosine similarity of the documents with the class profiles. The main advantage of JEX is its application to all 23 languages. In [6], Steinberger et al. proposed a method that maps different text to an existing knowledge structure, in this case *EuroVoc*, which acts as a conceptual multilingual dictionary. The map allows to compute documents similarity and returns a list of documents that are similar to one given in input. First, they assign a set of *EuroVoc* descriptors to a document. The list of descriptors can be seen as a summary of the document. Then, they compute a similarity score between documents using the descriptors lists. For the similarity score, they use the Okapi formula weighted by a length factor.

Blei et al.'s *Latent Dirichlet Allocation* (LDA) [1] is a generative model that treats a document as a finite mixture of topics, where a topic is a distribution over words. In details, each topic captures word co-occurrences inside documents. Recent works have focused on using LDA to analyze corpora. Cui et al. in [2], treat a corpus as a set of related documents in which the content evolves over time. They assume each document contains both old and new information. In their paper, they proposed a framework to capture both topics distribution over time and critical events, such as birth, split, merge or death of topics. Furthermore, the model captures and represents word co-occurrences and co-occurrences frequency using threads. First the model defines main words computing a set of weights, then it represents co-occurrences through the wave bundle of the thread. The amplitude of the wave represents the number of co-occurrences between the main word and the other words inside a topic: high amplitudes represent elevate co-occurrences frequency. In [3] Wang and McCallum make a similar assumption to Cui et at. They assume that the corpus evolves over time and the time is responsible to generate both patterns and topics distribution. They proposed a modified LDA model, called *Topics Over Time*, where topic discovery is influenced both by word co-occurrences and temporal information. In their work, the authors model the time as a continue distribution, defined by a Beta distribution over a parameter Ψ, associated with each topic which is responsible to generate both patterns and topics distribution. They tested TOT on different corpus, extracting information about that ones with topics. Blei et al. proposed in their work [9] a model that apply a sequence of LDA models in which α and β parameters depends on the previous ones. Hence, their model is able to capture the evolution of topics inside a corpus, tracking the topic over time and how its structure change (words co-occurrence). However, their main limitation is to capture both birth and death of topics.

Differently from the above-mentioned works, we used LDA not for analyzing how the corpus evolves over time, but to study how to restructure *EuroVoc* categories using the semantic information provided by such unraveled topics. As reported in Section 1, our assumption is that a relation between LDA topics and *EuroVoc* categories should exist.

3 Dataset

In this section, we describe the corpus we used to study the *EuroVoc* classification scheme. *Acquis Communautaire* (AC) is a multi-lingual corpora that covers more than 20 languages, in which each language-specific corpus has more than 5000 documents. We downloaded the 2.2 English version of the corpus from the web-page [8], which consists of 7972 documents related to the EU legislation written between the 1950s and 2005 and covers a variety of domains, such as *Agriculture*, *Foodstuff* and *Business*. In detail, as reported in Section *Statistics* in the web-page [8], the corpus contains 7512013 total words and an average words number of 942.

Each document is represented by an XML file which contains the document title, the *EuroVoc* code and the text. The *EuroVoc* code is not presented for all documents: documents older than 1955 and few new documents do not have *EuroVoc* categories. The remaining documents have been annotated using 2374 categories out of the total 6000 *EuroVoc* ones. The text is stored inside a set of <p> tags nested in <text> tag. Furthermore, each *p* tag contains the attribute *paragraph number* which is used to perform the paragraph alignment. Listing 1.1 shows the DTD schema of documents. We have taken the DTD schema from the web-page [8].

```
<TEI.2 id=”jrcCELEX–LG” n=”CELEX” lang=”LG”>
<teiHeader lang=”en” date.created=”DATE”>
<fileDesc>
    <titleStmt>
        <title>JRC–ACQUIS CELEX LANGUAGE</title>
        <title>Document Title</title>
    </titleStmt>
    <extent>
        nb_of_paragraphs paragraph segments
    </extent>
    <publicationStmt>
        <distributor>
            <xref url=”http://wt.jrc.it/lt/acquis/”>
                http://wt.jrc.it/lt/acquis/
            </xref>
        </distributor>
    </publicationStmt>
    <notesStmt>
         ....
    </notesStmt>
    <sourceDesc>
            <bibl>Downloaded from
                <xref url="Downloading_URL">
                   Downloading_URL
                </xref>
                on <date>Downloading_DATE</date>
            </bibl>
    </sourceDesc>
</fileDesc>
<profileDesc>
        <textClass>
                <classCode scheme="eurovoc">
                    Eurovoc_Code
                </classCode>
                    .....
        </textClass>
</profileDesc>
</teiHeader>
<text>
<body>
    <head n="1">Document Title</head>
    <div type="body">
        <p n="paragraph_number" >... TEXT...</p>
         .......
    </div>
    <div type="signature">
        <p n="paragraph_number">
            ... signature text...
        </p>
            ....
    </div>
    <div type="annex">
        <p n="paragraph_number">
            ... annex text...
        </p>
            ....
     </div>
</body>
</text>
</TEI.2>
```

Listing 1.1. Document DTD schema

4 Method

To evaluate the *EuroVoc* structure we generated topics using LDA, trying to associate to them each *EuroVoc* category. In this section, we describe the steps we performed to generate topics and to associate a label to one (or more than one) topic. In detail, we performed three steps: *document grouping* where we grouped documents according to the third-level *EuroVoc* categories presented inside documents; *document cleaning* in which we cleaned documents in order to define a Bag-Of-Words representation; and *category-topic linking*, where we associated a category with a topic using a similarity function.

To perform the *document grouping step*, we required the hierarchy of the *EuroVoc* categories in order to map each category to the third level. Thus, we constructed an XML file which reflects the *EuroVoc* hierarchy. The XML tree is defined as follows: conventionally defined the root of the XML at level 0, the level 1 contains the *EuroVoc* categories of the third level. The remain levels contain the *EuroVoc* categories from the 4th to the 8th. Since the XML file is structured as a direct acyclic graph, we assigned the categories to the first level we found during the visit. Figure 1 shows the number of categories at each level. As we can see from Figure 1, the 8th level contains only 6 categories, while the 4th level has 3754 categories and the 5th one has 2067 categories. If we allow to assign a category to multiple levels, we have 4202 categories for the 4th level and 2186 categories for the 5th level.

In the step called *document grouping*, first we filtered out all documents that do not have *EuroVoc* categories. Only 5157 of 7972 documents have at least an *EuroVoc* category. We mapped each *EuroVoc* category in the unfiltered documents with its corresponding third-level one. Then, we grouped documents according to their third-level categories, generating 452 groups[1]. In detail, each group represents a category and contains all documents labelled with it. Note that a document can belong to multiple groups. Figure 2 shows the top-5 set of documents having the most third-level categories. In the figure, the y-axis shows the number of categories, while the x-axis contains the document name.

In the *document cleaning* step, we cleaned the unfiltered documents generating a *Bag-Of-Words* for each one[2]. We used *Spacy*[3] to clean the documents, which allows to segment the text in sentences and to compute the Part-Of-Speech (POS) tags for each word in the sentence. The cleaning pipeline for each sentence is as follow:

1. we filter each word having POS tags different from *Noun*, *ADJ* (Adjective) and *ADV* (Adverb);
2. we filtered words belonging to the NLTK's[4] stopword list, to which we added alphabetic letters. Looking at the remaining words, we noticed that the POS

[1] Only 452 of 614 third level categories have been used to label documents in the corpus.

[2] In this step we did not remove duplicated words.

[3] `https://spacy.io/`

[4] `http://www.nltk.org/`

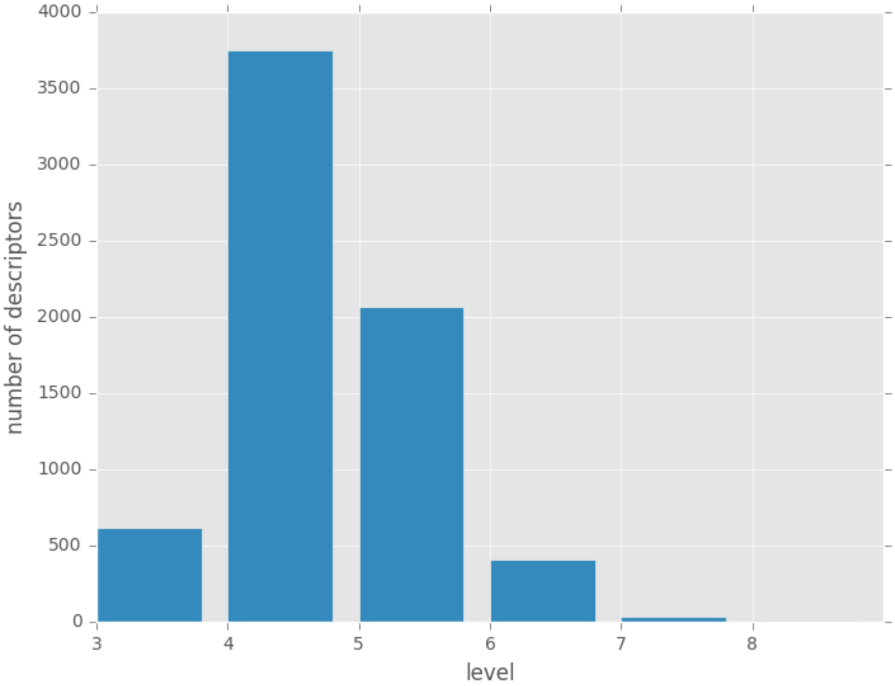

Fig. 1. The figure shows the number of categories for each level in the *EuroVoc* hierarchy.

tagger labelled some words, precisely numbers and words composed by letters and numbers, erroneously. Hence, it seems reasonable to check words using a regular expression and to filter the ones that do not match with the pattern. For this purpose, we used the regular expression $[A\text{-}Za\text{-}z]+$;

3. we mapped all unfiltered words to their lemma form.

In the *category-topic linking* step, we compute the LDA model passing in input the documents' Bag-Of-Words representation and the number of topics. We chose 452 (the number of *EuroVoc* categories found) as the number of topics according to our assumption. Next, we generated a Bag-Of-Words representation for each category merging the Bag-Of-Words of the documents belonging to it. In the creation of category-based Bag-Of-Words we removed duplicated words. Then, we treated topics and categories as nodes in a bipartite graph, where we assigned an edge between two nodes according to a similarity function. If the score of the similarity function is greater than a threshold, we link the two nodes.

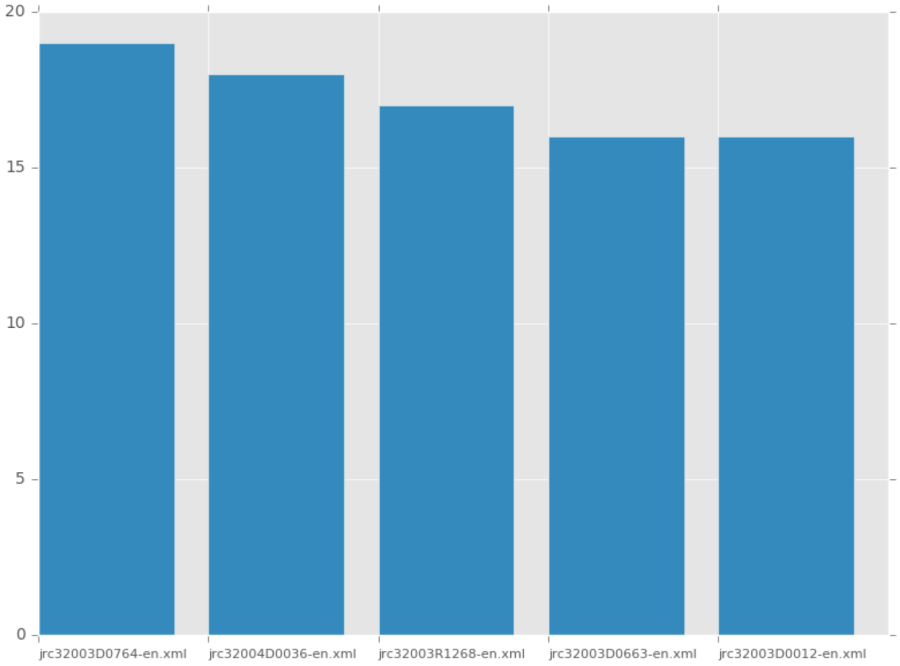

Fig. 2. The figure shows the top-5 documents having the most third level descriptors.

We used the cosine similarity computed over topics' list of words[5] and categories' Bag-Of-Words.

Finally, we computed two error measures and a convergence measure on the graph. Details about the cosine thresholds and the measures are discussed in the following section.

5 Analysis

Our assumption is that there exists a one-to-one relationship between topics and *EuroVoc* categories. In the previous section we showed the steps we performed to construct the bipartite graph; in this section we describe the analysis conducted on the graph to accept or refuse our assumption.

Since we represented topics, categories and their relations as a bipartite graph, we defined three evaluation measures based on node degree. Let n_i to be the i-th node in the graph, $\delta(n_i)$ the function that returns the degree of the i-th node, $|L|$ the number of categories and $\hat{L}$ the set of categories having a degree greater or equal to 1 and $I[\cdot]$ the indicator function, we can define the 3 measures as follows:

[5] The words list generated for each topic by the LDA model.

- $degree - error_1$ measure computes the ratio of the nodes with a degree greater than 1 and the number of total categories nodes:

$$degree - error_1 = \frac{\sum_{i=1}^{|L|} I[\delta(n_i) > 1]}{|L|} \tag{1}$$

- $degree - error_2$ measure computes the ratio of the nodes with a degree equals to 0 and the number of total categories nodes:

$$degree - error_2 = \frac{\sum_{i=1}^{|L|} I[\delta(n_i) == 0]}{|L|} \tag{2}$$

- *convergence* measure computes how much the categories with a degree at least equal to 1 are distributed between topics:

$$convergence = \frac{1}{|\hat{L}|} \sum_{i=1}^{|\hat{L}|} \frac{1}{\delta(n_i)} \tag{3}$$

All three measures have a co-domain that range from 0 to 1. In the convergence measure, the value 1 indicates maximum convergence (each category is associate with one and only one topic), while a value near to 0 indicates an high dispersion (categories are associated with multiple topics).

The three above-mentioned measures strictly depend on the threshold value used to link a label node with a topic one. Thus, we constructed and evaluated four graphs using different threshold values. We set the threshold to 0.1, 0.25, 0.35 and 0.5.

Figures 3, 4 and 5 show the results of the three measures applied on the four graphs. Figure 3 shows the error computed using Equation 1. In the figure, the error decreases with the increment of the threshold. Posing the threshold to 0.5 leads to an error equals to 0. However, this is not a good result: high threshold values tend to not associate topics to categories. This information is captured by Equation 2 and showed in Figure 4, in which we can see that the number of categories with no topic associated increases proportionally with the threshold, arriving to a value of 0.99. This means that only 4 categories have at least one topic associated.

Finally, Figure 5 shows the values computed using Equation 3 over the four graphs. The convergence measure growths slowly from 0 to 0.25 and has a pick on 0.35.

Generally, Figures 3, 4 and 5 report that categories are associated with more than one topic, especially when a large grain (low threshold values) is used.

6 Conclusion

The aim of the paper was to investigate the coherence of *EuroVoc* categories using Latent Dirichlet Allocation. Our assumption is that a category is able to

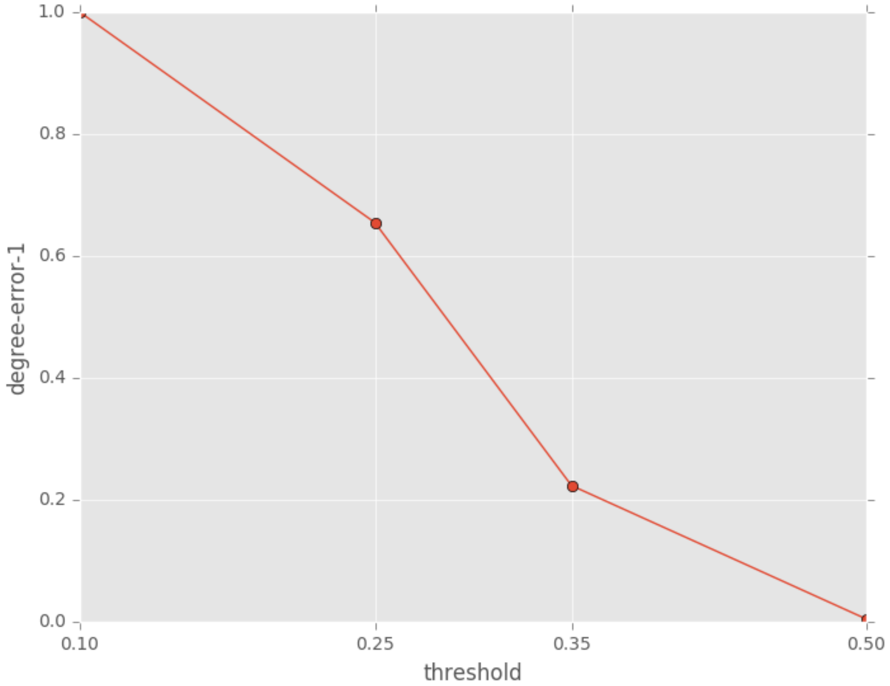

Fig. 3. The figure shows the error computed using Equation 1.

capture the content expressed in the documents. In Section 4, we showed several statistics regarding the corpus and how we processed it for our experiments. In Section 5, we evaluated our assumption using 3 measures on 4 bipartite graphs, each one generated using a different threshold value. As shown in the results, the *EuroVoc* structure seems to be not able to capture the actual documents content. Thus, we rejected our initial assumption.

This paper describes the fist step of an on-going work: to use topic modeling to restructure the *EuroVoc* classification scheme. A different *EuroVoc* structure would lead to less tags and better annotations of documents. In our future work we will use topics to hand-label a small set of documents, comparing the classifier trained using the topic annotated data with the classifier trained using the *EuroVoc* categories. This experiment will highlight if topics can improve the performance of the classifier. Moreover, not all topics may capture documents content. Some topics are meaningless for humans because they contain a set of semantic unrelated words that frequently occur together. Thus, we are interested also in selecting topics that can enhance both the annotation and the EuroVoc original structure.

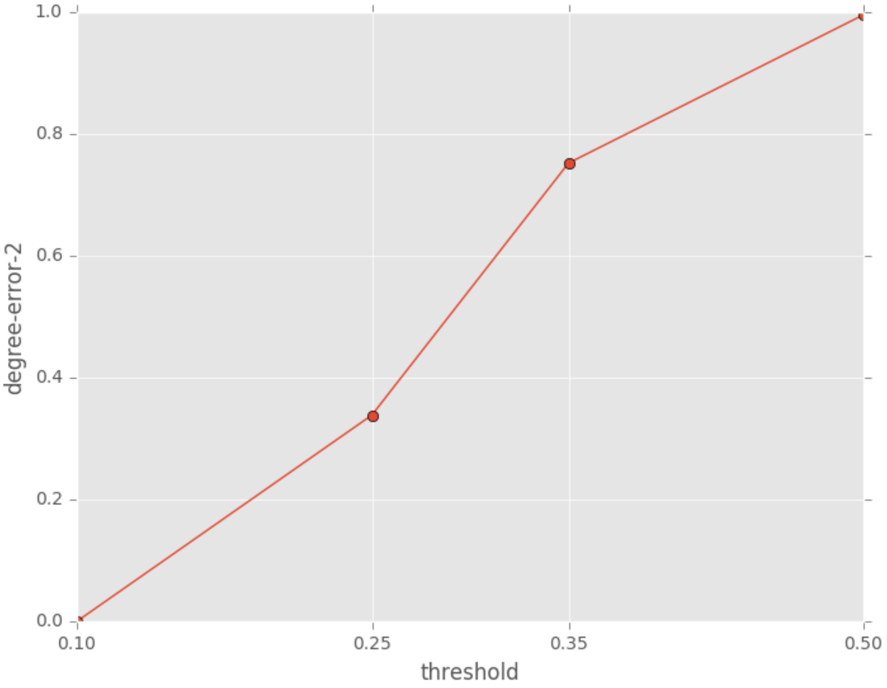

Fig. 4. The figure shows, for each threshold value, a score that represents how many categories are associated with no topic. The values are computed using Equation 2.

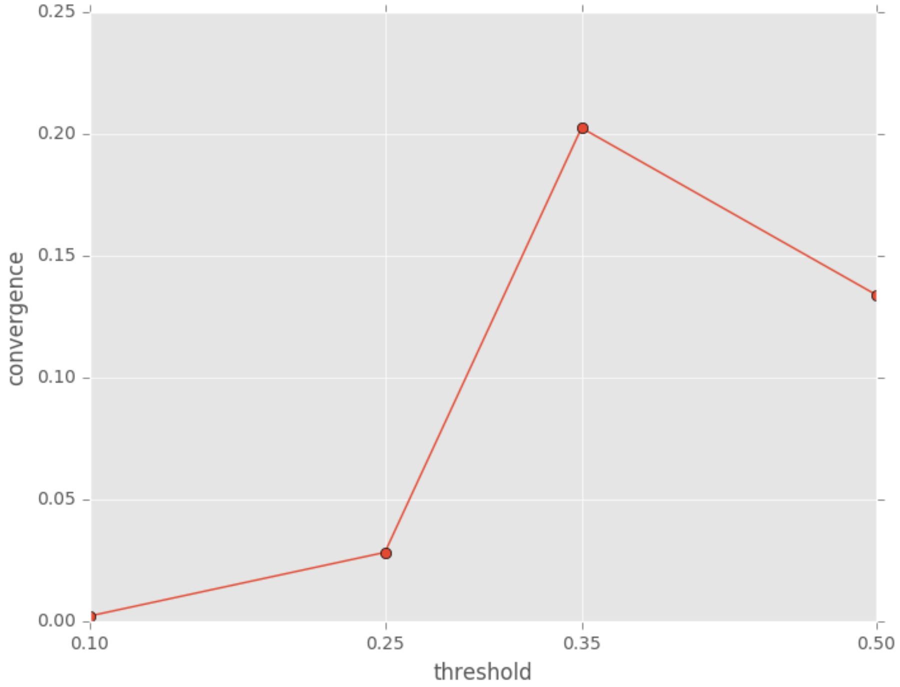

Fig. 5. The figure shows the convergence value for the 4 different thresholds.

References

1. Blei, David M and Ng, Andrew Y and Jordan, Michael I: Latent dirichlet allocation. the Journal of machine Learning research, vol 3, pp. 993–1022, JMLR.org (2003).
2. Cui, Weiwei and Liu, Shixia and Tan, Li and Shi, Conglei and Song, Yangqiu and Gao, Zekai J and Qu, Huamin and Tong, Xin: Textflow: Towards better understanding of evolving topics in text. Visualization and Computer Graphics, IEEE Transactions on, vol. 17 (12), pp. 2412–2421, IEEE (2011).
3. Wang, Xuerui and McCallum, Andrew: Topics over time: a non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424–433, ACM (2006).
4. Boella, Guido and Di Caro, Luigi and Lesmo, Leonardo and Rispoli, Daniele and Robaldo, Livio: Multi-label Classification of Legislative Text into EuroVoc. Legal Knowledge and Information Systems: JURIX 2012, the Twenty-fifth Annual Conference, vol. 250, IOS Press (2012).
5. Steinberger, Ralf and Ebrahim, Mohamed and Turchi, Marco: JRC EuroVoc Indexer JEX-A freely available multi-label categorisation tool. arXiv preprint arXiv:1309.5223 (2013).
6. Steinberger, Ralf and Pouliquen, Bruno and Hagman, Johan: Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. International Conference on Intelligent Text Processing and Computational Linguistics, pp. 415–424. Springer (2002).
7. http://eurovoc.europa.eu/drupal/?q=node
8. http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html
9. Blei, David M and Lafferty, John D: Dynamic topic models. Proceedings of the 23rd international conference on Machine learning, pp. 113–120, ACM (2006).