Departamento de Inteligencia Artificial
Escuela Técnica Superior de Ingenieros Informáticos

PhD Thesis

Semantically-enabled Browsing of Large Multilingual Document Collections

Author: Carlos Badenes-Olmedo
Supervisors: Prof. Dr. Oscar Corcho

xxx, 2020

Tribunal nombrado por el Sr. Rector Magfco. de la Universidad Politécnica de Madrid, el día XX de xxx de 2020.

Presidente:	Dr. Xxx Xxx
Vocal:	Dra. Xxx Xxx
Vocal:	Dr. Xxx Xxx
Vocal:	Dra. Xxx Xxx
Secretario:	Dr. Xxx Xxx
Suplente:	Dra. Xxx Xxx
Suplente:	Dr. Xxx Xxx

Realizado el acto de defensa y lectura de la Tesis el día X de xxx de 2020 en la Escuela Técnica Superior de Ingenieros Informáticos

Calificación: ______________________________

EL PRESIDENTE VOCAL 1 VOCAL 2

VOCAL 3 EL SECRETARIO

A mis padres.
A Beatriz.
A Martín y Alonso.

Agradecimientos

xxxxx

Abstract

xxxxx

Resumen

xxxxxx

Contents

List of Figures

List of Tables

Acronyms

API: Application Programming Interface

CQ: Competency Question

GUI: Graphical User Interface

IDE: Integrated Development Environment

LD: Linked Data

LOD: Linked Open Data

UML: Unified Modeling Language

URI: Uniform Resource Identifier

URL: Uniform Resource Locator

WUI: Web User Interface

Chapter 1

Introduction

Huge amounts of textual documents are produced daily in digital format. Every second, more than two thousand blog entries are published, nine thousand tweets are written, and more than two million emails are sent on the internet[1]. The number of scientific publications per year has increased by 8-9% in the last decade (Rob Johnson and Mabe, 2018). More than one million papers, about two per minute, were submitted to the PubMed database, the leading database of references and abstracts on life sciences and biomedical research, during last year. Statistics on judicial activity are similar. More than 168,000 procedural documents and 3,000 judicial notices were published in the Official Journal of the European Union in 2019 [2]. Unlike the academic domain where articles are mostly published in English, legal documents are usually available in multiple languages. The Court of Justice of the European Union had to translate over 1 million texts into its 24 official languages, with 552 possible language combinations, in just one year. These numbers make it virtually impossible for an expert in an academic or legal domain to stay abreast by only reading a few articles nowadays. Navigating the growing torrent of textual data and exploring their content is not only necessary, but has become a crucial job that experts must add to their daily tasks.

Document retrieval techniques are being used nowadays to facilitate text review in such big collections. Major digital publishers specialized in scientific[3], technical[4], and

[1] https://www.internetlivestats.com/one-second

[2] `https://curia.europa.eu`

[3] `https://www.nature.com`

[4] `https://www.elsevier.com`

medical[5] content provide search engines to make it easier to browse their collections of scientific articles. Given a few keywords, a list of relevant papers is retrieved and offered for reading. Legal documents are also exploited through similar solutions. The Spanish[6], American[7] and European[8] intellectual property registration offices, for example, allow exploring their patent collections by search engines guided by keywords and/or categories. These categories are available because documents are manually categorized by their authors according to the International Patent Classification (IPC) system. It contains approximately 70,000 different codes for different technical areas. This label-based browsing[9] has been also adopted by several academic search engines[10] to organize papers by research areas, or even by evaluation tasks to browse state-of-the-art methods[11]. In the natural language processing domain, research papers are organized into 256 tasks such as 'knowledge representation', 'question-answering', 'machine translation' and so on. However, while there are initiatives to normalize research areas, the use of keywords by authors in the form of tags to categorize their scientific papers is still insufficient and some text processing tasks are required to set labels to articles following a uniform criteria. One of the main reasons that limits its widespread use is the difficulty that authors have in picking labels that describe their research work in sufficient detail.

But what happens when a new research field or a new category emerges?. Searches guided by keywords or tags are useful when the domain is known, but when new categories appear these mechanisms may be insufficient. Researchers use reviewing to keep abreast of new discoveries. It helps to spark new ideas and approaches for their projects. But the problem is not all researchers are asked to peer review and those who are might struggle to fit it into their busy schedules, especially when the article is outside their area of expertise. Exist academic reports that identify the hottest and emerging specialty areas in scientific research from the last years[12]. The methodology used in these reports assumes that cumulatively, and over time, citations in research

[5] `https://pubmed.ncbi.nlm.nih.gov`
[6] `https://www.oepm.es`
[7] `https://www.uspto.gov/`
[8] `https://www.epo.org`
[9] `https://patents.google.com`
[10] `https://academic.microsoft.com`
[11] `https://paperswithcode.com`
[12] `https://discover.clarivate.com/ResearchFronts2019_EN`

leave a trail that highlights progress and advancements across a range of fields. By regularly tracking these citations and unpacking the patterns and groupings of how papers are cited, in particular clusters of papers that are frequently cited together, new research areas take shape. After an in-depth analysis, reports describe the research lines covered by these groups of research articles. Such reports also exist on patent collections to discover technology trends[13].

Therefore, identifying relationships between documents is key to facilitating their exploration. A documentary exploration does not stop when a relevant article is found, but starts from its content shaping the area of interest through its relations. Most academic[14] and legal[15] search engines provide a list of related documents for each text and offer navigating through them. The relationship between two documents can be based on references, when documents are cited by others[16], or content, when documents share a thematic area, etc. The chains of articles derived from that related content can lead to more complex structures when cross-relations are considered. A document can be related to another that, in turn, is related to a third one that can be also related to the first article. This content-guided exploration helps to browse document collections by areas of interest not necessarily aligned with a list of predefined categories. A visual overview of an academic field, for example, can be provided by showing graphs of articles with similar content[17].

While these initiatives are valuable efforts to address access to huge amounts of documents, they are still insufficient to make the most out of the knowledge available within the textual collection. On an individual level, the knowledge derived from a text comes from the concepts evoked by its words (Griffiths et al., 2007). On a collective level, the knowledge derived from a document collection emerges from the relationships among its texts (Kenter and de Rijke, 2015). Two modes of knowledge inferences are needed, a *bottom-up* from documents to collection-aware units, and a *top-bottom* from document relationships to specific info in texts. A semantic awareness is desirable to interpret the relationships and guide the exploration. The focus is on why some texts are related and what concepts are key to those relationships. But analyzing

[13] https://www.wipo.int/tech_trends
[14] https://www.semanticscholar.org
[15] https://patents.google.com
[16] http://citationgecko.com
[17] https://www.connectedpapers.com

and comparing texts on a large scale requires addressing some challenges imposed by external conditions that have appeared in recent years:

- **Complexity**: The increasing variety of topics, subjects, in those ever growing collections, has forced a reconsideration of the way to compare them. The operations required to compute each comparison should be simplified as much as possible.

- **Efficiency**: The algorithms, besides being accurate enough, must be also efficient in order to be applied on a large scale. Brute-force techniques cannot be applied to compare all items in a huge corpus.

- **Explainability**: The associations between documents must be explained in such a way that the relationship itself provides knowledge about the content of the texts. It is not enough that one text is related to another, it is necessary to explain why it is so.

- **Multilinguality**: In addition, the increasing availability of texts written in different natural languages also makes it necessary to address comparison in multilingual collections. In these collections, external translation systems cannot been considered as the only applicable solution, since they increase processing costs and potentially introduce a bias in the relationships that are obtained.

In our work we aim at facilitating the exploration of huge collections of documents written in multiple languages. We address the problem of comparing them on a large scale while enabling a semantic-aware exploration through their content. Our proposal automatically discovers thematic associations between texts using probabilistic topic models and organizes document collections so that they can be efficiently and transparently browsed through the related content regardless of their language.

1.1 Contributions

The following contributions are presented in this thesis:

- **Scalable Framework for Topic Modeling**: We introduce a text processing model that supports the creation of probabilistic topic models. Based on this abstraction, we implement a framework that becomes the foundation of this thesis research, which is used as a tool for supporting performance analysis and algorithm design.

- **Lifecycle Management of Topic Models**: To facilitate the exploitation of trained topic models, we propose a unified form to publish and (re)use models based on Web services available from online repositories.
- **Hierarchical Thematic Annotations**: To study the problem of representativeness in high dimensional topic models, we exploit the relationships between texts derived from their topic distributions. We show how the distances vary between the same texts when the dimensions of the model change, and how less representative topics can influence their calculation. Our analytical and experimental results show that as more topics are available in the model, less representative are the distance measurements based on densities. During the study, we identified hierarchies in the topic distributions that maintain their representativeness regardless of the dimensions of the model, and without losing the ability to measure distances. We propose a method to annotate texts using topic hierarchies, and a distance metric based on this hierarchical representations.
- **Support for Large-scale Document Similarity Comparisons**: We present an efficient mechanism to index and retrieve related documents, described by hierarchical annotations, while allowing the exploration of the collection by the themes inferred from its texts.
- **Discovery of Cross-lingual Document Relations**: We introduce a technique to transform probabilistic topics from different languages into a single representation space based on shared concepts where texts can be thematically related regardless of the language used.

1.2 Thesis Structure

The thesis is structured as follows:

Chapter 2 describes the main concepts handled throughout the thesis, analyses the state of the art and identifies the main limitations. *Chapter 3* presents the research problems and hypotheses that guide our work, as well as assumptions and restrictions and details the methodology that has been followed. *Chapter 4* describes the software architecture proposed to analyze huge document collections and the format suggested to distribute and reuse the topic models built in this thesis. *Chapter 5* details the text annotation algorithm based on probabilistic topics. *Chapter 6* shows how to store and search documents efficiently from large collections when they are annotated with topic hierarchies. *Chapter 7* explains the method to relate texts written in different languages from their main topics without the need of supervision. *Chapter 8* describes real-world projects where contributions from this thesis have been used. Finally, *Chapter 9* introduces conclusions and future lines of work.

1.3 Publications

The following publications support the research work presented in this thesis:

- *Chapter 4- Creation and Publications of Probabilistic Topic Models*:
  - **Carlos Badenes-Olmedo**, José Luis Redondo-Garcia, and Oscar Corcho. Distributing Text Mining tasks with librAIry. Proceedings of the 17th ACM Symposium on Document Engineering (DocEng). Association for Computing Machinery, Valletta, Malta. 2017.
  - Victoria Kosa, Alyona Chugunenko, Eugene Yuschenko, **Carlos Badenes-Olmedo**, Vadim Ermolayev, and Aliaksandr Birukou. Semantic saturation in retrospective text document collections. Information and Communication Technologies in Education, Research, and Industrial Applications (ICTERI) PhD Symposium, vol. 1851, pages 1-8. CEUR-WS. 2017
  - Victoria Kosa, David Chaves-Fraga, Dmitriy Naumenko, Eugene Yuschenko, **Carlos Badenes-Olmedo**, Vadim Ermolayev, Aliaksandr Birukou, Nick Bassiliades, Hans-Georg Fill, Vitaliy Yakovyna, Heinrich C. Mayr, Mykola Nikitchenko, Grygoriy Zholtkevych, and Aleksander Spivakovsky. Cross-Evaluation of Automated Term Extraction Tools by Measuring Terminological Saturation. Information and Communication Technologies in Education, Research, and Industrial Applications, pages 135-163. Springer International Publishing. 2018

- *Chapter 5- Explainable Topic-based Associations*:
  - **Carlos Badenes-Olmedo**, José Luis Redondo-García, and Oscar Corcho. Efficient Clustering from Distributions over Topics. Proceedings of the 9th International Conference on Knowledge Capture (K-CAP), Article 17, 1–8. Association for Computing Machinery, Austin, TX, USA. 2017.
  - **Carlos Badenes-Olmedo**, Jose Luis Redondo-Garcia, and Oscar Corcho. An initial Analysis of Topic-based Similarity among Scientific Documents based on their Rhetorical Discourse Parts. Proceedings of the 1st Workshop on Enabling Open Semantic Science (SemSci) co-located with 16th International Semantic Web Conference (ISWC 2017), 15-22. Vienna, Austria. 2017.
- *Chapter 6- Large-scale Comparisons of Topic Distributions*:
  - **Carlos Badenes-Olmedo**, José Luis Redondo-García, and Oscar Corcho. Large-scale Semantic Exploration of Scientific Literature Using Topic-based Hashing Algorithms. Semantic Web, vol. 11, no. 5, pp. 735-750. 2020
- *Chapter 7- Cross-lingual Document Similarity*:
  - **Carlos Badenes-Olmedo**, José Luis Redondo-García, and Oscar Corcho. Scalable Cross-lingual Document Similarity through Language-specific Concept Hierarchies. Proceedings of the 10th International Conference on Knowledge Capture (K-CAP). Association for Computing Machinery, 147–153. Marina Del Rey, CA, USA. 2019
  - **Carlos Badenes-Olmedo**, José Luis Redondo-García, and Oscar Corcho. Legal document retrieval across languages: topic hierarchies based on synsets. arXiv e-prints, arXiv:1911.12637. 2019
  - Ahmet Soylu, Oscar Corcho, Brian Elvesaeter, **Carlos Badenes-Olmedo**, Francisco Yedro, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Dumitru Roman. Enhancing Public Procurement in the European Union through Constructing and Exploiting an Integrated Knowledge Graph. Proceedings of the 19th International Semantic Web Conference (ISWC). 2020

Chapter 2

State of the Art

2.1 Related Work

In this chapter we analyze the current state of the art and limitations to facilitate the exploration of large and multilingual document collections. First, the tasks derived from processing the texts and enabling a semantic-aware exploration of the corpus are described. Then, an overview of the existing methods that perform these tasks is introduced. And finally, for each research area involved, the limitations that must be addressed to achieve the ultimate goal of facilitating documentary exploration are presented. The concepts that will be used throughout the rest of the thesis are here introduced.

In order to browse large and multilingual document collections we need to process them in a way that is computationally affordable and provides enough knowledge to understand the relationships that arise. The annotation of human-readable documents is a well-known problem in the Artificial Intelligence (AI) domain in general and Information Retrieval (IR) and Natural Language Processing (NLP) fields in particular. Vector space models (VSM) (Salton, 1983) were proposed to represent texts as vectors where each entry corresponds to a different term and the number at that entry corresponds to how many times that term is present in the text. The objective was twofold, on the one hand to make document collections manageable since we move from having lots of terms for each text to only one vector per document with defined dimension, and on the other hand to have representations based on metric spaces where calculations can be made, for example comparisons by measuring vector distances. The definition and number of dimensions for each vector are key aspects in a VSM. Traditional retrieval tasks over large collections of textual documents highly rely on individual features like term frequencies (TF) (Hearst and Hall, 1999). A representational space is created where each term in the vocabulary is projected by a separate and orthogonal dimension. All terms in a document are treated as equally descriptive. To overcome this problem, Term-Frequency Inverse-Document Frequency (TF-IDF) (Christopher D. Manning and Schütze, 2008) relativizes the relevance of each term with respect to the entire corpus. TF-IDF calculates the importance of a term for a document, based on the number of times the term appears in the document itself (term frequency - TF) and the number of documents in the corpus, which contain the term (document frequency - DF). The absence of semantic information and the notion of word sequences, and the high-number of dimensions are the main drawbacks of these approaches that lead to the emergence of other techniques.

New ways of characterizing documents based on the automatic generation of models surfacing the main themes covered in the corpus are developed during recent years. Among them, text embedding proposes transforming texts into low-dimensional vectors by prediction methods based on (i) word sequences or (ii) bag-of-words. The first approach assumes words with similar meanings tend to occur in similar contexts. It considers word order relevant and is based on Neural Models (NM) that learn word vectors from pairs of target and context words, where context words are taken as words observed to surround a target word. Document vectors are usually created by taking the word vectors they contain or by considering them as target and context items. Skip-gram with negative sampling (Word2Vec) (Mikolov et al., 2013) and Global Vectors (GloVe) (Pennington et al., 2014) are indeed the most popular methods to learn word embeddings due to its training efficiency and robustness (Levy et al., 2015). The second approach does not consider the order of the words to be relevant, but their frequency is. It assumes words with similar meanings will occur in similar documents. Topic models (Blei et al., 2003; Deerwester et al., 1990; Hofmann, 2001) are the main methods based on this approach. This second approach is used in our work since *we are not only interested in representing words and documents, but we also seek internal structures that can provide knowledge about the collection as a whole.*

Probabilistic Topic Models (PTM) (Blei et al., 2003; Hofmann, 2001) are statistical methods based on bag-of-words that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, or how they change over time. PTM do not require any prior annotations or labeling of the documents. The topics emerge, as hidden structures, from the analysis of the original texts. These structures are topics distributions, per-resource topic distributions or per-resource per-word topic assignments. In turn, a topic is a distribution over terms that is biased around those words associated to a single theme. This interpretable hidden structure annotates each resource in the collection and these annotations can be used to perform deeper analysis about relationships between resources. Topic-based representations bring a lot of potential when applied over different IR tasks, as evidenced by recent works in different domains such as scholarly (Gatti et al., 2015), health (Lu et al., 2016; Tapi Nzali et al., 2017), legal (Greene and Cross, 2016; O'Neill et al., 2017), news (He et al., 2017) and social networks (Cheng et al., 2014). Topic modeling provides us an algorithmic solution to organize and annotate large collections of textual documents according to their topics.

The simplest generative topic model is *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). Along with *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) and *Probabilistic Latent Semantic Analysis* (pLSA) (Hofmann, 2001) are part of the field known as topic modeling. They are well-known latent variable models for high dimensional data, such as the bag-of-words representation for textual data or any other count-based data representation. They try to capture the intuition that documents can exhibit multiple themes. Each document exhibits each topic in different proportion, and each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. All the documents in a collection share the same set of topics, but each document exhibits these topics in a different proportion. Texts are described as a vector of counts with W components, where W is the number of words in the vocabulary. Each document in the corpus is modeled as a mixture over K topics, and each topic k is a distribution over the vocabulary of W words. Formally, a topic is a multinomial distribution over words of a fixed vocabulary representing some concept. Depending on the function used to describe that distribution there are different algorithms to create topic models. While LSA and pLSA propose a singular value decomposition, LDA, influenced by the generative Bayesian framework to avoid some of the over-fitting issues that were observed with pLSA, suggests the use of a Dirichlet function. It is a continuous multivariate probability distribution parameterized by a vector of positive reals whose elements sum to 1. It is continuous because the relative likelihood for a random variable to take on a given value is described by a probability density function, and is multivariate because it has a list of variables with unknown values. In fact, the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution and is responsible for, unlike LSA and pLSA, LDA can infer topic distributions in texts that have not been used during training.

Topic models are not restrictive clustering models where each document is assigned to one cluster, but allows documents to exhibit multiple topics. The topics covered in a set of documents are discovered from the own corpus and feature vectors are topic distributions expressed as vector of probabilities. Taking into account this premise, the similarity between two topic-based resources is based on the distance between their topic distributions, which can be also seen as two probability mass functions. A commonly used metric is the *Kullback-Liebler* (KL) divergence:

$$KL(P,Q) = \sum_{i=1}^{K} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{2.1}$$

However, it presents two major problems: (1) when a topic distribution is zero, KL divergence is not defined and (2) it is not symmetric, which does not fit well with semantic similarity measures that are usually symmetric (Rus et al., 2013).

Jensen-Shannon (JS) divergence (Lin, 1991; Rao, 1982) solves these problems considering the average of the distributions as below (Celikyilmaz et al., 2010):

$$JS(P,Q) = \sum_{i=1}^{K} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{K} q_i * \log \frac{2 * q_i}{q_i + p_i} \tag{2.2}$$

where K is the number of topics and p, q are the topics distributions

It can be transformed into a similarity measure as follows (Dagan et al., 1999) :

$$sim_{JS}(D_i, D_j) = 10^{-JS(p,q)} \tag{2.3}$$

where D_i, D_j are the documents and p, q the topic distributions of each of them.

Hellinger (He) distance is also symmetric and is used along with JS divergence in various fields where a comparison between two probability distributions is required (Blei and Lafferty, 2007; Boyd-Graber and Resnik, 2010; Hall et al., 2008):

$$He(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2)} \tag{2.4}$$

It can be transformed into a similarity measure by subtracting it from 1 (Rus et al., 2013) such that a zero distance means max. similarity score and vice versa:

$$sim_{He}(D_i, D_j) = 1 - He(p,q) \tag{2.5}$$

However, all these metrics are not well-defined distance metrics, that is, they do not satisfy triangle inequality (Charikar, 2002). This inequality considers $d(x,z) <= d(x,y) + d(y,z)$ for a metric d (Griffiths et al., 2007) and places strong constraints on distance measures and on the locations of points in a space given a set of distances. As a metric axiom, the triangle inequality must be satisfied in order to take advantage of the inferences that can be deduced from it. Thus, if similarity is assumed to be a monotonically decreasing function of distance, this inequality avoids the calculation of all pairs of similarities by considering that if x is similar to y and y is similar to z, then x must be similar to z.

S2JSD was introduced by (Endres and Schindelin, 2003) to satisfy the triangle inequality. It is the square root of two times the JS divergence:

$$S2JSD(P,Q) = \sqrt{2 * JS(P,Q)} \tag{2.6}$$

2.2 Research Areas

This thesis aims to enable a semantic-aware exploration of the knowledge arising from large and multilingual document collections by exploiting the capabilities of topic models and their metric spaces. There are several research areas of ongoing related work.

The first one is **topic creation and reuse**, key for understanding the steps needed to transform the unstructured data from a text into numerical values based on probabilistic topics. *The way in which topic models are created and reused is crucial to addressing large-scale analysis.*

The second area is **topic explainability**, which refers to the capacity of topics to capture and describe the content of a text. Topic explainability is important for *making understandable the relationships that are derived from topic distributions.*

The third area is **document similarity**, where the ability to measure the *semantic difference between texts from the distance between their topic distributions is addressed.*

Finally, the fourth area is **multilingual topics**, as we aim to explore collections of texts written in different languages through their topic-based relationships. A *strategy to relate the topics of each language is needed.*

These four areas are closely related to each other. Having efficient thematic representations of texts, distance metrics based on shared themes, and mechanisms to abstract the particularities of a language to represent the themes, may help to organize large multilingual document collections.

Each area and its limitations are described below. A summary can be found in Table 2.1.

Area	Scope	Limitation
large-scale topic creation	process texts and train topic models from large corpora	no scalable frameworks that integrates both tasks
topic reuse	calculate distributions from existing topic models	no unified models for exchanging topic models
topic explainability	describe and relate documents by topics	high dimensional models makes them difficult to interpret
document similarity	compare topic distributions by measuring their distances	unaffordable complexity in large collections
multilingual topics	topic distributions across languages	parallel or comparable training data required

Table 2.1: Research areas and limitations.

2.2.1 Topic Creation and Reuse

Textual content usually includes non-relevant information and keeping only what can bring value for the involved agents (general consumers, experts, companies, investors...) becomes a challenge. A necessary first step before leveraging documents for knowledge-intensive tasks is to preprocess them following different techniques. Recent studies (Westergaard et al., 2017) have shown that mining full-text articles gives consistently better results than only using sections or summaries. Given the size limitations and concise nature of summaries, they often omit descriptions or results that are considered to be less relevant but still are important for some IR tasks (Divoli et al., 2012). Since this behavior is present in many other domains, our interest is focused on processing full texts, not only summaries or parts of texts, as we will show in the remainder of this thesis.

Exists a broad set of algorithms able to analyze text for producing annotations at very different levels of granularity: from minimal units such as terms and entities, to descriptors at the level of the entire collection such as summaries or topics. Methods to perform Part-of-Speech (PoS) tagging, Named Entity Recognition (NER) tasks, or topic modeling following the LDA or any other approach. But their implementations have been designed to work in an isolated, non-collaborative way (Agerri et al., 2014; Manning et al., 2014). They have not paid special attention to facilitating their interoperability and use closed formats to manage their data which increase the technological dependence and limits their reuse and their expansion possibilities. For example, a topic model trained in Mallet[18] can only make inferences if it is used from Mallet itself or using its libraries, since ***there is no unified format for distributing topic models*** and each resource defines its own. In that example, the fact that Mallet is implemented in Java prevents reusing their models from Python or any other programming language. However, there are NLP tools (e.g. spaCy[19]) that have been provided through open services decoupled from their technical development (e.g Explosion[20]).

Some approaches have advanced in this direction and offer the creation and exploitation of topic models through an API based on libraries[21] or web services[22](Lisena

[18] `http://mallet.cs.umass.edu`
[19] `https://spacy.io`
[20] `https://github.com/explosion/spacy-services`
[21] `https://bab2min.github.io/tomotopy`
[22] `https://github.com/D2KLab/ToModAPI`

et al., 2020), but they are focused on the operations that can be performed on the model rather than abstracting the topic model as a resource. Others provide local[23] or remote[24] ecosystems where create and reuse learning models, but their format is not open and cannot be used out of the environment. To the best of our knowledge, the efforts made do not propose a unified model to exchange topic models, understood as an already accepted standards-based format. In this thesis we propose *reusable topic models and a scalable framework to create and use them.*

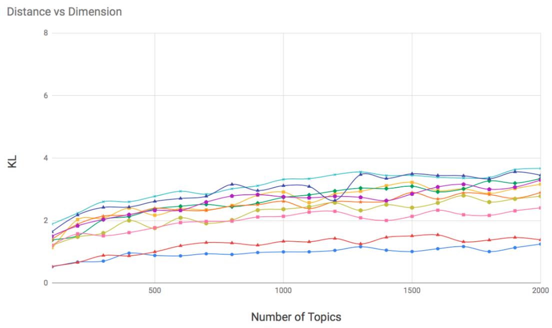
2.2.2 Topic Explainability

Even though distance metrics mentioned in Section 2.1 have been proposed and used in the SoA, making sense out of the similarity score based on compare topic distributions is not easy. As shown in figure 2.1, given a set of pairs of documents, their similarity scores vary according to the number of topics. So the distances between the same pairs fluctuate from being more to less distant when changing the number of topics, and are hence difficult to use for relate semantically documents. In this conditions, which of those distances is better representing the underlying collection?

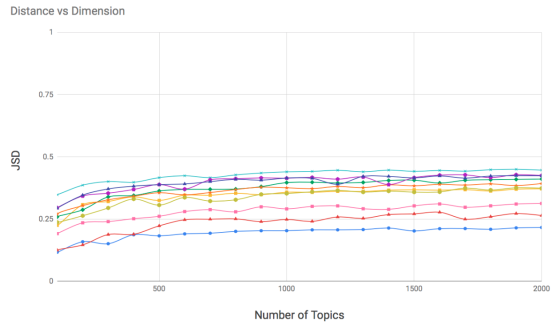
Distances between documents based on topic distributions, because it is a vector space, generally increase as the number of dimensions of the space increases. This is due to the fact that as the number of topics describing the model increases, the more specific the topics will be. Topics shared by a pair of documents can be broken down into more specific topics that are not shared by those documents. *Document similarity is then dependent on the model used to represent documents when considering this type of metrics.*

Each topic is drawn from a Dirichlet distribution with parameter β, while each document's mixture is sampled from a Dirichlet distribution with parameter α. These two priors, α and β, are also known as hyper-parameters of a topic model and they set the probability that a document or a word, respectively, contains more than one topic. We know that absolute distances between documents vary when we tune those hyper-parameters differently, but we also see that "relative distances" also change. Imagine that we have two documents, A and B, and one topic model, M1. The distance from the topic distribution of A to B is less than from A to C. However, in a

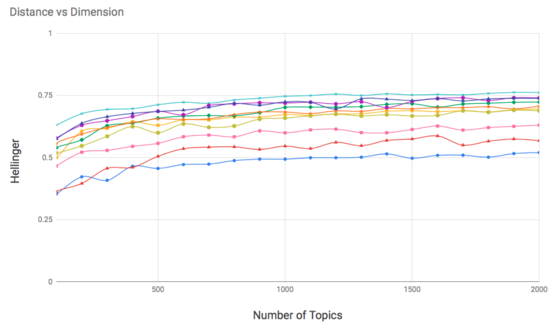
[23] `https://onnx.ai/`
[24] `https://vespa.ai`



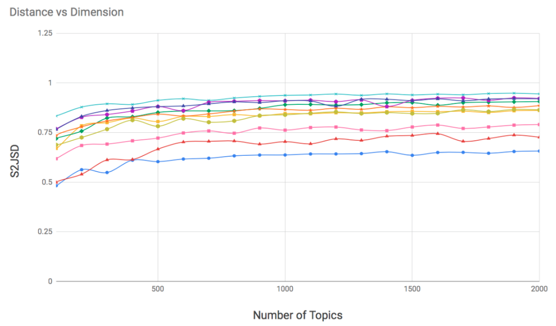
(a)



(b)



(c)



(d)

Figure 2.1: Distance values of 10 pairs of documents calculated in topic models with 100-to-2000 dimensions. The Kullback-Liebler(a), Jensen-Shannon Divergence(b), Hellinger(c) and S2JSD(c) metrics are considered.

second topic model, M2, trained with the same documents as M1 but with different hyper-parameters, the distance from the topic distribution of A to C is less than to B (cross-lines in fig 2.1). This behaviour highlights the ***difficulty of establishing absolute similarity thresholds and the complexity to measure distances taking into account all dimensions***. If we consider that documents are similar when their distance is lower than 0.2, a pair of documents may be similar when they are represented in low-dimensional topic models, and not similar when high-dimensional models are used to represent them. Distance thresholds should be model-dependent rather than general and metrics flexible enough to handle dimensional changes. In this thesis we have gone beyond the thematic and low-dimensional feature space created by topic models and propose a *hierarchical feature space suitable for big real-world data sets, where documents are only described by their most relevant topics.*

2.2.3 Document Similarity

In addition, document similarity comparisons are too costly to be performed in huge collections of data and require more efficient approaches than having to calculate all pairwise similarities. Using a naive approach creating a similarity matrix with all document comparisons takes $O(n^2)$ time (where n is the number of documents), so obtaining all possible pairs of similarities in a large collection of documents (e.g. a corpus of 32 million patents) can be unfeasible because of the exponential cost of comparing every pair of elements. Many different approaches have been proposed to reduce this complexity. For instance, computation can be approximated by a nearest neighbors (ANN) search problem (Indyk and Motwani, 1998). ANN search is an optimization problem that finds nearest neighbors of a given query in a metric space of n points.

Due to the low storage cost and fast retrieval speed, hashing is one of the most popular solutions for ANN search (Zhen et al., 2016). This technique transforms data points from the original feature space into a binary-code space, so that similar data points have larger probability of collision (i.e. having the same hash code). This type of formulation for the document similarity comparison problem has proven to yield good results in the metric space (Krstovski and Smith, 2011) due to the fact that ANN search has been designed to handle distance metrics (e.g. cosine, Euclidean, Manhattan). But distance metrics between topic distributions should be information-theoretically motivated metrics (e.g. Hellinger, Kullback-Leibler divergence, Jensen-Shannon divergence) since they compare density functions.

These challenges can be tackled by hashing methods based on clusters of topics to measure similarity, instead of directly using their weights. Hashing methods transform the data points from the original feature space into a binary-code Hamming space, where the similarities in the original space are preserved. They can learn hash functions (data-dependent) or use projections (data-independent) from the training data (Wang et al., 2016). Data-independent methods, unlike data-dependent ones do not need to be re-calculated when data changes, i.e. adding or removing documents to the collection. Taking large-scale scenarios into account (e.g. Document clustering, Content-based Recommendation, Duplicate Detection), data independency is a key feature along with the ability to infer hash codes individually (for each document) rather than on a set of documents. Data-independent hashing methods depend on two key elements: (1) data type and (2) distance metric. For vector-type data, as introduced in section 2.1, based on l_p distance with $p\epsilon[0, 2)$ lots of hashing methods have been proposed, such as p-stable Locality-Sensitive Hashing (LSH) (Datar et al., 2004), Leech lattice LSH (Andoni and Indyk, 2006), Spherical LSH (Terasawa and Tanaka, 2007), and Beyond LSH (Andoni et al., 2014). Based on the θ distance many methods have been developed such as Kernel LSH (Kulis and Grauman, 2012) and Hyperplane hashing Vijayanarasimhan et al. (2014). But only few methods handle density metrics in a simplex space, where topic distributions are projected. A first approach transformed the He divergence into an Euclidean distance so that existing ANN techniques, such as LSH and k-d tree, could be applied Krstovski et al. (2013). But this solution does not consider the special attributions of probability distributions, such as Non-negative and Sum-equal-one. Recently, a hashing schema (Mao et al., 2017) has been proposed taking into account the symmetry, non-negativity and triangle inequality features of the S2JSD metric for probability distributions. For set-type data, Jaccard Coefficient is the main metric used. Some examples are K-min Sketch (Li et al., 2012), Min-max hash (Ji et al., 2013), B-bit minwise hashing (Li and König, 2010) and Sim-min-hash (Zhao et al., 2013).

All of them have demonstrated efficiency in the search for similar documents, but none of them considers the search for documents (1) by thematic areas or (2) by similarity levels, nor they offer (3) an ***explanation about the similarity obtained beyond***

the vectors used to calculate it. In addition, binary-hash codes drop a very precious information: the topic relevance. This thesis proposes a *hash function-based approach that allows efficiently searching for related documents while maintaining topic-based annotation, preserving notion why two documents are related.*

2.2.4 Multilingual Topic Alignment

When the IR task is also cross-language, document retrieval must be independent of the language of the user's query. At execution time, the query in the source language is typically translated into a target language with the help of a dictionary or a machine-translation system. But for many languages we may not have access to translation dictionaries or a full translation system, or they can be expensive to execute, or expensive to train (lot of data required) in an online search system. In such situations it is useful to rely on smaller annotation units derived from the text so the full content does not need to be translated, for instance by finding correspondences with regard to the topics discussed in both the query and the documents being searched.

Some methods uses document-aligned corpora, where documents are grouped and constrained to the same topic distribution during training to align the different languages (De Smet and Moens, 2009a; Fukumasu et al., 2012; Mimno et al., 2009a; Ni et al., 2009; Zhang et al., 2013), or theme-aligned corpora, where similar themes and ideas appear in all languages(Boyd-Graber and Blei, 2009). Multilingual Probabilistic Topic Models (MuPTM) (Vulić et al., 2015) have emerged in this area as a group of language-independent generative machine learning models that can be used on theme-aligned multilingual texts. They are based on LDA, adding supervised associations between languages by using *parallel* corpora, with sentence-aligned documents (e.g. Europarl[25] corpus), or *comparable* comparable, with theme-aligned documents (e.g. Wikipedia[26] articles), in multiple languages. Once a MuPTM has been generated, documents can be represented by data points in a single feature space based on topics to detect similarities among them exploiting inference results and using distance metrics. Due to its generic language-independent nature and the power of inference on unseen documents, MuPTM's have found many interesting applications in many different cross-lingual tasks. They have been used on cross-lingual event clustering

[25]https://ec.europa.eu/jrc/en/language-technologies/dcep
[26]https://www.wikipedia.org/

(De Smet and Moens, 2009b), document classification (De Smet et al., 2011; Ni et al., 2011), semantic similarity of words (Mimno et al., 2009b; Vulić and Moens, 2012), information retrieval (Ganguly et al., 2012; Vulić and Moens, 2013), document matching (Platt et al., 2010; Zhu et al., 2013), and others.

Other methods are based on word alignments from bilingual dictionaries instead of aligned corpora. Topic models emerge as distributions over crosslingual equivalence classes of words (Hao and Paul, 2018; Jagarlamudi and Daumé, 2010; Zhang et al., 2010). A recent approach is placed between word and document alignments. It proposes crosslingual topic models using the language-independent categories assigned to each wikipedia article(Piccardi and West, 2020). Instead of using bags-of-words to represent texts, which would be language dependent, it explores the references of each article and represents them through bags-of-links, using the categories of each reference to represent the texts.

In short, ***the requirement of parallel/comparable corpora or dictionaries limits the usage of these models in many situations***. There are not many document collections that can be used for training since large parallel corpora are rare in most of the use cases, especially for languages with fewer resources. Moreover, in order to incorporate new languages or update the existing associations, these models must be re-trained with documents from all languages at the same time, making it difficult to scale to large corpora (Hao et al., 2018; Moritz and Ìˆchler, 2017). We take MuPTM a step further by making it cross-lingual through representations based on topic hierarchies. Documents from multi-language corpora are described by expressions of language independent concepts and can then be efficiently browsed and related with other languages without the need for translation or parallel or comparable corpora. In this thesis we propose to *automatically infer cross-lingual topics to browse multi-lingual document collections without the need for parallel or comparable corpus.*

Chapter 3

Methodology

The work presented in this thesis aims to facilitate the exploration of huge collections of multilingual documents through thematic associations inferred from their content. It assumes that, given a multilingual corpus, similar documents will contain the same topics. Each of the challenges arising from this objective defines a working dimension and guides the research carried out in this thesis.

The first dimension focuses on **scalability**, in order to create the text processing flows that are required to create or apply learning models. The workload required to process a corpus varies according to the number of documents, the length of texts and the kind of knowledge (annotations) that need to be inferred from the text. If the design of the workflow is scalable, there is no need to modify the processing logic when working with larger collections of documents, since adding a reasonable amount of computational resources is enough to perform it. These resources can be machines (i.e horizontal scaling) or processing units (e.g CPU, RAM) in an existing machine (i.e vertical scaling).

The second dimension covers the **representativeness** of the text annotations when projected into spaces where they are manipulated. The idea behind these spaces is to represent documents as points (or vectors in a vector space) that are close together when the texts are semantically similar, and far apart when they are semantically distant. The ability of these spaces to create meaningful representations is also studied in this work.

In the third dimension, we explore the usage of data structures that efficiently **cluster** texts from their representations based on probabilistic topics. Divisions of space into semantically-related regions are convenient to allow browsing large document collections. The *representativeness* covered in the previous dimension enables the interpretation of the relations and regions obtained.

And finally, the fourth dimension handles the **multilingualism** of collections that contain documents in several languages. On a multilingual space, documents are described and related across languages.

This chapter introduces our main hypothesis (section 3.1), and the associated research challenges (section 3.2), and presents the research methodology (section 3.3).

3.1 Research Hypotheses

We define our main hypothesis as follows:

Hypothesis 1 *Large multilingual document collections can be automatically analyzed to discover appropriate thematic relations that facilitate a semantically-enabled text browsing.*

Our hypothesis can be divided into four different sub-parts, which are related to the aforementioned scalability, representativeness, clustering, and multilinguality dimensions. First, by *distributing across different computation nodes both natural language processing tasks and representational models we can efficiently process huge collections of documents (**H1.1**).*

Second, it is possible to *semantically relate documents by comparing their most relevant topics (**H1.2**).* Furthermore, for this purpose we hypothesize that the use of *topic hierarchies (**H1.2.1**) and similarity metrics based on relevance levels (**H1.2.2**) help quantifying the semantic distance between texts.* Third, by *dividing the representational space into regions based on topics and relevance levels we can search for related documents without having to calculate all pairwise comparisons and without discarding the notion of topics for further processing (**H1.3**).*

And finally, *by abstracting the topic representations into concept-based descriptions across languages we can relate documents in various languages without having to translate them (**H1.5**).*

A summary of the hypotheses and how they tackle our research dimensions can be found in Table 3.1.

Hypothesis	Research Dimension
H1: Large multilingual document collections can be automatically analyzed to be semantically-browsed through thematic relations	D1: Scalability, D2: Representativeness, D3: Clustering, D4: Multilingualism
H1.1: it is possible to efficiently annotate documents on a large scale by distributing across different computation nodes natural language processing tasks and representation models	D1: Scalability
H1.2: it is possible to semantically relate texts from their most relevant topics	D2: Representativeness
H1.3: it is possible to find relevant documents with similar topic distributions without calculating all pairwise comparisons and without discarding the notion of topics from their representation	D3: Clustering
H1.4: it is possible to relate documents in different languages without having to translate them, by using language agnostic concepts from their main topics	D4: Multilingualism

Table 3.1: Hypotheses and research dimensions.

3.2 Research Challenges

Several research challenges emerge from these hypotheses. First, in order to facilitate reusing existing topic models by processing systems with different architectures and technological stacks, we need to define *topic-model programming interfaces*. Second, in order to describe and thematically relate documents, we must address how to produce *explainable topic-based associations*. Third, by working with huge collections of documents described by topics, we need to handle *large-scale comparisons of topic distributions*. Finally, in order to explore multilingual document collections from shared topic-based representational spaces, we have to provide *automatic cross-lingual topic alignment*. Each of these research challenges are described below and covered throughout this thesis.

3.2.1 Scalable Creation and Inference of Topics

Although some initiatives to facilitate the reuse of machine-learning models exist in the literature as discussed in section 2.2.1, there are still some restrictions that limit the use of topic models. Technical dependencies or closed data formats are the main reasons that prevent or make reproducibility of these models difficult by imposing conditions to work with them. It is common to provide the model only through its functionality, rather than its properties, which prevents it from being used for a different purpose than when it was created. *Reuse of topic models is limited by incompatibility problems **(RCInterface1)***.

The properties and functions offered by a topic model, instead of being uniform, vary according to the method used to build it. This limits the ability to reproduce the work done in this area and reuse topic models without losing information or knowledge. *There is no standard to specify the attributes and operations that a topic model can provide **(RCInterface2)***. Sometimes topics are described by the top ten or five most relevant words, and occasionally these word lists are not accompanied by weights, making a density-based analysis impossible. These differences in presenting the models can sometimes limit their reusability if they cannot infer new topic distributions even when the learning algorithm allows for it.

3.2.2 Explainable Topic-based Relations

In order to facilitate the exploration of document collections, vector space models are often used to semantically relate texts based on their word distributions. As described in Section 2.1, these models first create a dictionary with the words used in the collection, and then represent documents by vectors whose dimensions correspond to each word in the dictionary. In large collections, these models need to be adapted to make operations on vectors more manageable. As a result, a new abstraction method based on emerged topics that reduces the dimensions of vectors. Topics are described by word distributions over the entire vocabulary and documents by vectors containing topic distributions. Despite the extensive use of these representation models, *there is no common criteria for identifying the most representative topics in a document* ***(RCExplainable1)***.

In addition, since similarity metrics over this representation space are based on accumulating the difference in topic densities, *it is difficult to explain the distance between topic distributions* ***(RCExplainable2)***. And, unless a minimum distance threshold can be defined or a set of n-top topics agreed, *there is no common criterion for determining whether two documents are related* ***(RCExplainable3)***.

3.2.3 Large-scale Comparisons of Topic Distributions

There are many scenarios where we need to find related documents (e.g. a researcher doing literature review, or an R&D manager analyzing project proposals). Experts can benefit from discovering those connections to achieve these goals, but brute-force pairwise comparisons are not computationally adequate when the size of the corpus is too large. Some algorithms in the literature divide the search space into regions containing potentially similar documents, which are later processed separately from the rest in order to reduce the number of pairs compared. However, *there are no mechanisms that efficiently partition the topic-based search space without compromising the ability for thematic exploration* ***(RCComparison1)***.

In addition, documents from the same region should be compared and *there are no similarity metrics that compare partial distributions of topics* ***(RCComparison2)***.

3.2.4 Unsupervised Cross-lingual Topic Alignment

With the ongoing growth in the number of texts in different languages, we need annotation methods that enable browsing multilingual corpora. As discussed in section 2, multilingual probabilistic topic models have recently emerged as a group of semi-supervised machine learning models that can be used to perform thematic explorations on collections of texts in multiple languages. However, *there are no approaches that abstract the representation of probabilistic topics in language-independent spaces without translating texts or aligning documents* ***(RCCrossLingual1)***. Existing approaches require parallel or comparable training data to create a language-independent space.

A summary of the challenges covered in this work and how they map to the hypotheses is presented in table 3.2.

3.3 Research Methodology

The research presented in this thesis is based on four dimensions or research areas as discussed in section 3.2. Each one is motivated by different research problems that we need to solve in order to achieve our ultimate goal of making it easier to explore large multilingual document collections through their topics. Once a dimension is tackled, the next one is considered, and so on. This iterative and incremental methodology allows refining the research results by evaluating them with more experiments and addressing increasingly complex research problems.

Figure 3.1 shows the dimensions on which the research of this thesis has been built. The top of the pyramid is only reached once the lower dimensions are dealt with successfully. They are presented as a chain of four steps. The first step describes the motivation to perform a given task coming from real-world problems that we had to deal with, and is represented by a brown arrow. In the context of this task, the research problem arises and is framed by a pink arrow. For each of them a solution is proposed and evaluated according to a specific criterion. The proposed solution is represented by a green arrow and the evaluation with a blue arrow. Once a proposal has been validated, the next dimension of the pyramid is achievable and all the previous research problems are added to the new research problem as conditions to be taken into account.

Research Challenge	Hypotheses
RCInterface1: reuse of topic models is limited by incompatibility problems	H1.1: documents can be efficiently annotated on a large scale by distributing across different computation nodes both natural language processing tasks and representation models
RCInterface2: there is no standard that unifies the representation of topic models	H1.1: documents can be efficiently annotated on a large scale by distributing across different computation nodes both natural language processing tasks and representation models
RCExplainable1: there is no common criteria for identifying the most representative topics in a document	H1.2: texts can be semantically related from their most relevant topics, H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons and without losing the ability to explore them through their topics
RCExplainable2: it is difficult to understand the distance between topic distributions	H1.2: texts can be semantically related from their most relevant topics
RCExplainable3: there is no common criterion for determining whether documents are related	H1.2: texts can be semantically related from their most relevant topics
RCComparison1: there are no mechanisms that efficiently partition the topic-based search space without compromising the ability for thematic exploration	H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons
RCComparison2: there are no similarity metrics that compare partial distributions of topics	H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons
RCCrossLingual1: there are no approaches to abstract probabilistic topics in language-independent spaces without translating texts or aligning documents	H1.4: documents in different languages can be related without having to translate them using language agnostic concepts from their main topics

Table 3.2: Open Research Challenges and Hypotheses.

Technical objectives (i.e., develop a new resource) or research objectives (i.e., discover the solution to a problem) guide the solution proposal before moving on to the next dimension. They are presented below, organized by the research problem associated with each dimension.

3.3.1 Scalable Creation and Inference of Topics

This first dimension arose when we had to analyze a huge collection of documents describing research and innovation projects to discover which research areas are being addressed, measure their presence in the collection, and characterize them so their presence can be inferred in unseen documents. Such a high volume of data made difficult to process it manually, so we needed to automate the required processing to draw insights from it. Probabilistic topics allow describing research areas, so we defined a *distributed text-processing model for creating large probabilistic topic models* ***(RO1)*** and a *Web service template to distribute them* ***(RO2)***. In this way, the models themselves could be easily integrated into scalable text processing pipelines. As a result, we created a *scalable platform for topic modeling* ***(TO1)***, and produced a *model-as-a-service repository with pre-trained topic models* ***(T02)***. The efficiency of this solution was validated by processing a corpus of 100,000 documents collected from the CORDIS dataset[27], which contains descriptions of projects funded by the European Union under a framework programme since 1990 (Badenes-Olmedo et al., 2017b).

The main contributions under this dimension are described in Chapter 4 as follows:

- a software architecture to process big volumes of textual documents in a distributed and decoupled manner;
- the definition of a model-as-a-service template for probabilistic topic models;
- an implementation of the architecture, librAIry, following those design principles.

3.3.2 Explainable Topic-based Relations

In the second dimension we needed to browse scientific papers through their content-based relations. The problem of massively annotating documents with topic distributions came up. We had to *create annotations based on topic models in a way that*

[27] `https://data.europa.eu/euodp/es/data/dataset/cordisH2020projects`

*was computationally affordable and enabled a semantic-aware exploration of the knowledge inside them (**RO3**)*. Once documents were annotated, a *metric that compares documents and facilitates their interpretation from topic annotations (**RO4**)* was required. As a result, we integrated *the annotation method into the topic model service (**TO3**)* and implemented a text comparison metric based on partial representations of topics. These proposals were validated by classifying 500,000 scientific articles from the Open Research Corpus[28] in domains such as Computer Science, Neuroscience and Biomedicine (Badenes-Olmedo et al., 2017a, 2019a, 2017c).

The main contributions under this dimension are described in Chapter 5 as follows:

- a clustering algorithm based on probabilistic topic distributions;
- a hash function to transform topic distributions into topic hierarchies;
- a similarity metric based on topic sets.

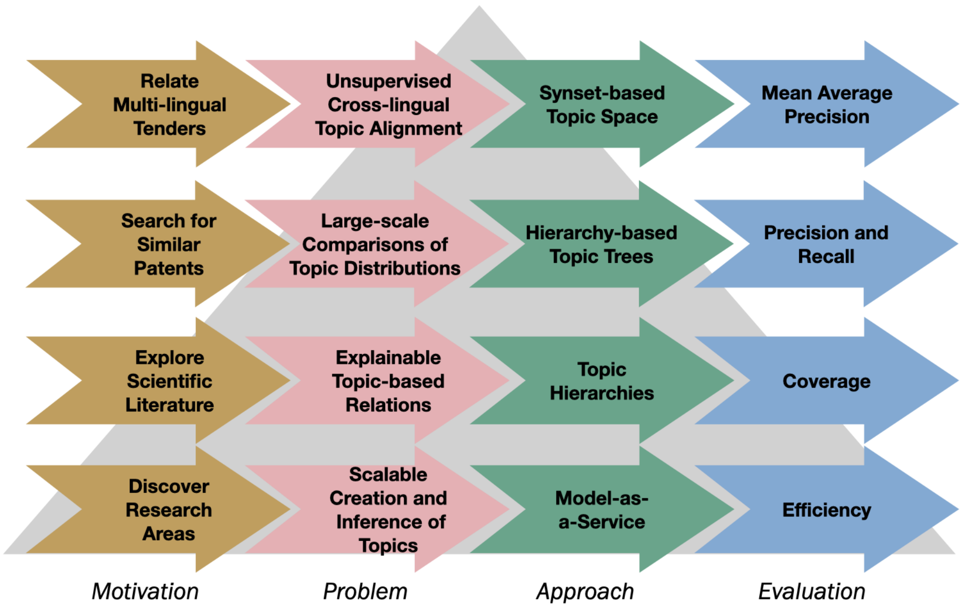


Figure 3.1: Research dimensions of the thesis. The first ones must be overcome before reaching higher dimensions.

[28] `https://allenai.org/data/open-research-corpus`

3.3.3 Large-scale Comparisons of Topic Distributions

This dimension covered the search for similar documents based on their most relevant topics. Thanks to having dealt with the above two dimensions, large collections of documents could be annotated with topic hierarchies and text distances could be measured from their annotations. Now, the aim was to find similar documents without losing the exploratory capacity offered by topics. Similarity comparisons were too costly to be performed in such huge collections of data and required more efficient approaches than having to calculate all pairwise similarities. We applied *techniques based on approximate nearest-neighbors to organize documents in regions with similar topic hierarchies (**RO5**)*. As a result, we developed *a system to automatically find similar documents (**TO4**)*. It was validated on a collection of one million texts retrieved from the United States patents corpus[29]. The relations between patents derived from their manual categorization were compared with those automatically obtained from their topic distributions (Badenes-Olmedo et al., 2019a, 2020).

The main contributions under this dimension are described in Chapter 6 as follows:

- a data structure to partition the search space and organize documents described by topic hierarchies;
- a corpus browser that leverages these representations to automatically relate documents.

3.3.4 Unsupervised Cross-lingual Topic Alignment

Finally, a new dimension on top of the previous ones emerged to relate texts coming from different languages. In particular, since document relations were based on their topics, this dimension was focused on aligning topics without supervision from models trained with texts in different languages. Since each language defined its own vocabulary, the topics were model-specific and could not be directly compared. We abstracted the *topic representations to create a single space out of the particularities of the language (**RO6**)*. This approach was validated on the English, Spanish, French, Italian and Portuguese editions of the JCR-Acquis[30] corpora and revealed promising results on

[29] `https://www.uspto.gov/ip-policy/economic-research/research-datasets`
[30] `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis`

classifying and sorting documents by similar content across languages (Badenes-Olmedo et al., 2019a,b).

The main contributions under this dimension are described in Chapter 7, as follows:

- an algorithm to represent probabilistic topics using concept sets;
- a repository of aligned topic models from the English, Spanish, French, Italian and Portuguese editions of the JRC-Acquis corpus.

Table 3.3 summarizes the research objectives (ROs), technical objectives (TOs) and connects them with the research challenges (RCs) from Table 3.2.

Research Objective	Research Challenge
RO1: Define a distributed text-processing model for creating large probabilistic topic models	RCInterface1
RO2: Define a template to package probabilistic topic models as web services	RCInterface2
RO3: Define annotations based on topics that enable a semantic-aware exploration of the knowledge inside a corpus	RCExplainable1
R04: Define a metric based on topic annotations that compares documents and facilitates their interpretation	RCExplainable2, RCExplainable3
RO5: Define nearest-neighbor techniques to organize documents in regions with similar topic hierarchies	RCComparison1, RCComparison2
R06: Define a transformation of the topic-based annotations to create a unique representational space out of the particularities from each language	RCCrossLingual1
TO1: Create a scalable platform for topic modeling	RCInterface1, RCInterface2
T02: Create a repository of Topic-based web services	RCInterface2
T03: Integrate the annotation method based on topic hierarchies into the topic model service	RCExplainable2, RCComparison2
T04: Create a system capable of finding similar documents automatically	RCExplainable2, RCExplainable3, RCComparison1, RCCrossLingual1

Table 3.3: Research and technical objectives and their related challenges.

Chapter 4

Creation and Publication of Probabilistic Topic Models

This chapter present ***librAIry*** (Badenes-Olmedo et al., 2017b), our framework to exploit probabilistic topic models through a service-oriented approach. In doing so, we reuse existing techniques and standards, which aim to make our results reusable and interoperable with other alternative approaches.

4.1 Topic Modeling Framework

Topic models have been successfully used in multiple domains in the last years (Greene and Cross, 2016; He et al., 2017; O'Neill et al., 2017; Tapi Nzali et al., 2017). Each of them has different characteristics. Some with longer texts, others with shorter texts, some with millions of documents, others with only a few hundred or thousands, some have only one computing unit to process them, others have multiple nodes distributed among several distributed machines. Adapting to such diversity, topic modeling algorithms have evolved to be more efficient in challenging situations (Liu et al., 2015). However, the methods are only focused on the learning process and a topic model life-cycle is broader than just the creation of the model (Fig. 4.1.) It covers a first stage of *document preparation*, where texts are divided first into phrases and then into words that are normalized before being counted to create bags-of-words (BoW). The next stage seeks patterns among word distributions, is the *training* stage, and as a result a topic model is created. The model is then packaged for distribution, this is the *publi-*

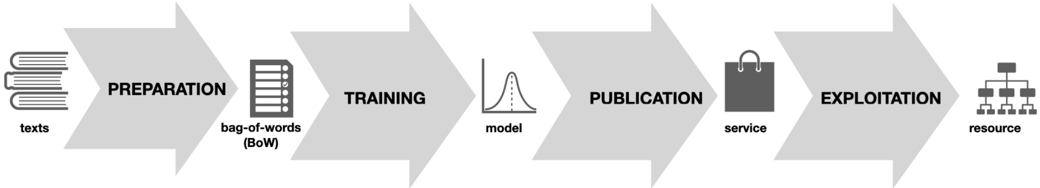


Figure 4.1: Topic model life-cycle. First texts are processed to retrieve tokens and build bags-of-words (BoW). They are then used to train a model that identifies patterns among words and builds topics. The model, which is also enabled to make inferences in unseen texts, is published as a service in an online repository. Finally, the service is used as resource in a particular solution, for example to categorize documents.

cation stage. And finally the model can be reused, it is the *exploitation* stage. In order to have a framework that covers the entire process of creating, publishing and reusing probabilistic topic models in both large- and small-scale, we have focused on adapting techniques and standards widely used in software engineering domain. In this section we cover the first two stages of the life-cycle: text preparation and model training. In Section 4.2.1 we describe how the models are published, and in section 4.2.2 how they can be exploited.

librAIry is a framework that covers the entire topic model life-cycle by combining learning algorithms with natural language processing methods and software distribution techniques. The main objective is to ***facilitate the creation of reusable probabilistic topic models by minimizing their technical dependencies***. Methods and algorithms proposed in this thesis have been implemented and evaluated in this framework, which therefore serves as the technological basis for our research.

Our design requirements, which have guided our development process, can be organized into three categories:

- **Corpora representation requirements**, which tackle the modeling of document collections and its metadata. This includes texts and their related annotations.

- **Task distribution requirements**, which refer to event management to notify changes in document collections. Coordination of this information is crucial for robust and reproducible results.

- **Process execution requirements**, which capture the operations involved in creating a topic model. The parallel task execution leads to the creation of models.

The rest of the section describes how we have adapted existing techniques and standards in *librAIry* to address each of the requirements categories described above. An open, distributed and scalable framework has been developed whose source code is publicly available for reuse[31].

4.1.1 Corpora Representation for Topic Modeling

Inspired by a Staged Event-Driven Architecture (SEDA) that exchanges messages and handle status changes, our framework is based on *resources* and *actions.* A *resource* can be a *document* that represents raw texts (e.g. a full-text research paper), or a *snippet* of text with a logical part (e.g. sections, summaries or even phrases grouped by their rhetoric), or a *domain* that contains a dataset of texts (e.g. a conference proceedings) or even an *annotation* made on them (e.g. review comments, named-entities, topics). *Actions* can be executed on resources to change their status (e.g. *create*, *update* or *delete*).

To better illustrate this model, take a sample of the research articles published at the latest K-CAP conference[32] (see Fig. 4.2). A *document* resource is created for each publication and contains the full text of the article. Each *document* is then associated with several *snippets*, one for each section of the article (e.g. abstract, introduction, conclusions, etc). Finally, a *domain* is created that groups all these *documents* under the same conference. This initial representation can be extended with *annotations*, that can provide more detailed information at different levels (e.g. named-entities, keywords or topics).

Resources, and *actions* are individually addressable and linkable (Turchi et al., 2012) following the Linked Data principles(Bizer et al., 2009). Each of them has: (1) a name, (2) a retrievable (or dereferenceable) HTTP URI so that it can be looked up, (3) a descriptive information provided by using standard notation (e.g. JavaScript Object Notation (JSON)) when it is looked up by URI, and (4) links to other URIs so that other resources can be discovered from it.

[31] `https://github.com/librairy`
[32] `http://www.k-cap.org/2019`

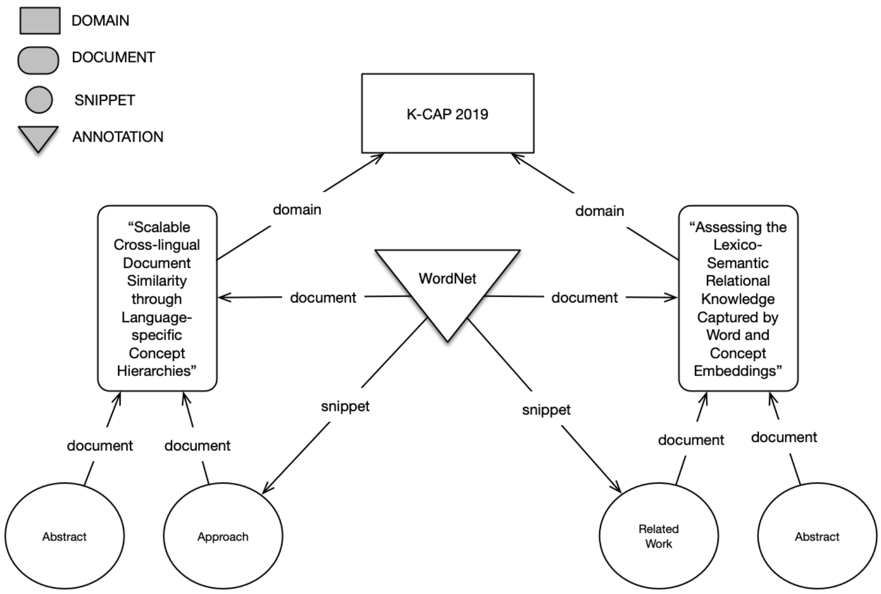


Figure 4.2: Representation of two scientific papers published at the International Conference on Knowledge Capture (K-CAP, 2019) that mention the same entity, *Wordnet*, in different sections.

More details about each of them is shown below.

4.1.1.1 Domain

A *domain* is an aggregation of *documents*. It is described as a group with parts separately described. A default *domain* is created and every *document* belongs, at least, to one *domain* (Fig 4.3).

A *domain* contains the following information:

- **uri**: identifier created from the resource type (i.e *domains*) and a Universally Unique Identifer (UUID) (e.g *domains/88b86fa6-11c8-11eb-adc1-0242ac120002*)
- **creation-time**: date when resource was created. It follows the ISO-8601[33].
- **name**: label associated to the resource.
- **description**: additional information about it.

[33] `https://www.iso.org/standard/40874.html`

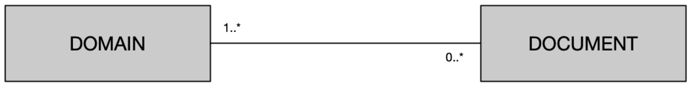


Figure 4.3: Relation between *domain* and *document*.

4.1.1.2 Document

A *document* is a resource consisting primarily of text. Examples include research papers, articles, books or patents. It follows the Open Archives Initiative for Object Reuse and Exchange[34] (OAI-ORE) and the Dublin Core Metadata Initiative[35] (DCMI).

A *document* contains the following information:

- **uri**: identifier created from the resource type (i.e *documents*) and a UUID (e.g *documents/809af686-11c8-11eb-adc1-0242ac120002*)
- **creation-time**: date when resource was created. It follows the ISO-8601.
- **publishedOn**: date when resource was published. It follows the ISO-8601.
- **publishedBy**: an entity responsible for making the document available. It can be a person, an organization or a service. It may be different from the entity that conceptually formed the resource (e.g. wrote the document), which is recorded as *authoredBy*. This entity should be identified by a valid Uniform Resource Identificator (URI) such as WebId[36], orcid[37] or internal URI.
- **authoredOn**: the time the *document* was conceptually formed. The author time should be present if different from *publishedOn*. It must be a formatted timestamp following ISO-8601.
- **authoredBy**: an entity primarily responsible for making the content of the *document*. It may be a list to indicate multiple authors. Each of them identified by a valid URI such as WebId, orcid or internal URI.

[34] `http://www.openarchives.org`
[35] `http://dublincore.org`
[36] `http://www.w3.org/wiki/WebID`
[37] `http://orcid.org`

Figure 4.4: Relation between *document* and *snippet*.

- **retrievedFrom**: a URI identifying the repository or source from which the document was derived. This property should be accompanied with *retrievedOn*.
- **retrievedOn**: the time the *document* was retrieved on. If this property is present, the *retrievedFrom* must also be present. It must be a formatted timestamp following ISO-8601.
- **format**: the physical or digital manifestation of the resource. Typically, it includes the media-type (i.e the IANA code[38]) of the *document*.
- **language**: the language(s) in which the document was written. It is defined by RFC-1766[39] with a two-letter language code followed, optionally, by a two-letter country code.
- **title**: a name given to the *document*. It is a name by which the *document* is formally known.
- **description**: it may include but is not limited to an abstract, or a free-text account of the content.
- **rights**: information about rights helds in and over the *document*.
- **content**: raw text from the *document*.

Furthermore, a *document* can contain zero or more *snippets* and a *snippet* can belong to one or more *documents*. Since *librAIry* can also discover relations among *documents*, a *document* may contain zero or more references to other *documents* (Fig. 4.4).

[38] `http://www.iana.org`

[39] `http://www.ietf.org/rfc/rfc1766.txt`

4.1.1.3 Snippet

A *snippet* is a resource that is included either physically or logically in a *document*. In a scientific *document*, for example, it may be the *abstract* section or a logical set of sentences sharing the same rhethorical class (e.g. approach, background, related-work, etc). As seen above (Fig. 4.4), a *snippet* can belong to one or more *documents*.

It contains the following information:

- **uri**: identifier created from the resource type (i.e *snippets*) and a UUID (e.g *snippets/7a5a46c8-11c8-11eb-adc1-0242ac120002*)

- **creation-time**: date when resource was created. It follows the ISO-8601.

- **sense**: content-type. It refers to a section or any other criteria under which the following text makes sense.

- **content**: partial text retrieved from the full-text of a *document*.

4.1.1.4 Annotation

Annotations are data retrieved from resources that can be used to relate them. They are basically key-value data structures associated to *domains*, *documents* or *snippets*. Examples are entities mentioned in a text, or topics covered in a collection. Any resource can have zero or multiple annotations, which can be shared among several resources (Fig. 4.5)

It contains the following information:

- **uri**: identifier created from the resource type (i.e *annotations*) and a UUID (e.g *annotations/73671e68-11c8-11eb-adc1-0242ac120002*)

- **creation-time**: date when resource was created. It follows the ISO-8601.

- **key**: a category or type associated with the information it contains (e.g. entity, comment, topic, keywords, etc). Recommended best practice is to use a controlled vocabulary.

- **value**: a note about the resource in the form of free text.

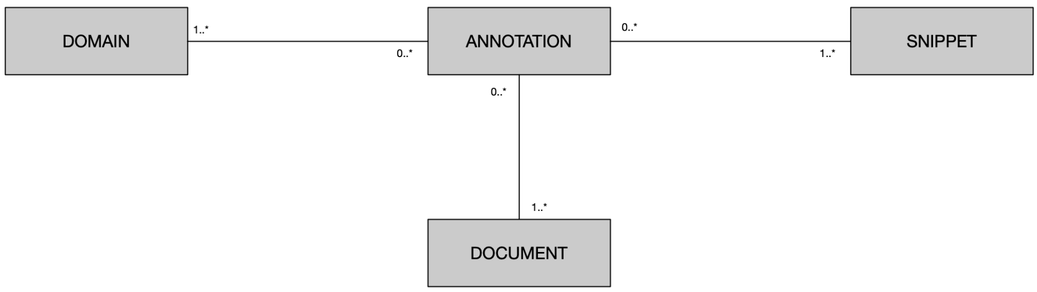


Figure 4.5: Relation between *annotations* and other resources.

4.1.2 Event-oriented Processing Workflow

Along with the resources mentioned above, there are two additional elements that provide a crucial behavior to the framework: *modules* and *events*. An *event* is a non-persistent time-based occurrence that describes a new action performed on a resource. *Modules* are responsible for carrying out operations on the resources (e.g. tokenize a *document* or create a topic model from a *domain*). *Events* are broadcasted so that any *module* is aware of the changes made to the resources and can perform actions on one or more resources in response to a new state reached by a given resource. These actions are paralleled since modules are replicated through distributed environments.

The framework follows a publisher/subscriber approach where *modules* can publish and read *events* to notify and to be notified about the state of a *resource* (Fig. 4.6). An *event* notifies a performed *action* (i.e. a resource and its new state), and follows the Representational State Transfer (REST) paradigm (Fielding and Taylor, 2002). It contains the resource type and the new state reached by a specific resource (i.e *created*, *deleted* or *updated*). For example, when a new *domain* is created, an *event* message is published to the channel: *domain.created*. A channel is a space where *events* are published and *modules* can be subscribed to read only some *events*. The actions performed by a module depend on the events to which it is subscribed. Therefore, the workflow of the framework is neither static nor explicitly defined. A distributed dynamic workflow emerges according to the *modules* subscribed to the *event* channels.

We used the Advanced Message Queuing Protocol (AMQP) as the messaging standard to avoid any technical dependency to the message broker (i.e the server that

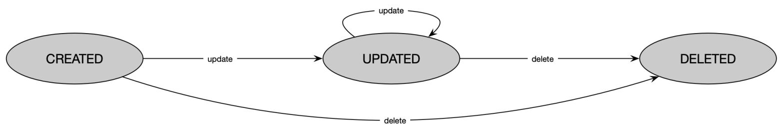


Figure 4.6: Resource states.

sends and receives messages). This protocol defines: *exchanges*, *queues*, *routing-keys* and *binding-keys* to communicate publishers (i.e message senders) and consumers (i.e message readers). *Exchanges* are like message inboxes, and *queues* are subscribed to them by specifying the message types they are interested in with a *binding-key*. A message sent by a publisher to an exchange is routed with a *routing-key* and consumers matching that *routing-key* with their *binding-key* (used to connect the *queue* to that *exchange*), will receive the message. This mechanism allows sending and receiving messages between consumers and producers by means of shared keys (i.e. *routing-keys* and *binding-keys*). A key follows the structure: *resource.status*. Since a wildcard-based definition can be used to set the key, this paradigm allow modules both listening to individual type events (e.g. *domains.created* for new *domains*), or multiple type events (e.g. *#.created* for all new resources).

4.1.3 Module-based Model Training

A microservice-oriented style has been used to define the framework architecture. Through multiple services the system analyzes texts, creates probabilistic topic models, publishes them as new services and uses them to annotate texts. A service is equivalent to a functionality, and each functionality is materialized by a *module* in the system. A *module* is then a cohesive and independent process (Dragoni et al., 2016) with a specific purpose (i.e functionality) based on the *events* to which it responds. These *events* correspond to the routing- and binding- keys attached to the module.

There are four types of *modules* (Fig. 4.7):

- **Harvester**: creates resources such as *documents*, *snippets* and *domains*, from local or remote repositories with textual files. We have developed harvesters to create scientific resources from Elsevier[40] or any other digital repository[41] that

[40] `https://github.com/librairy/harvester-elsevier`
[41] `https://github.com/librairy/harvester-research`

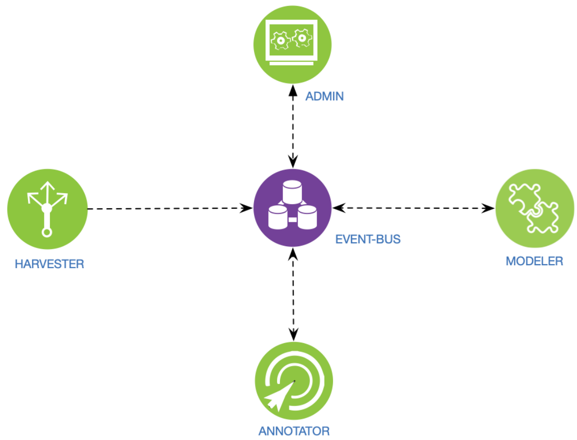


Figure 4.7: Modules (in green) publishing or receiving events from the messenger service (in purple).

follows the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[42]; as well as more general ones to retrieve public datasets from datos.gob.es[43] or resources located on local folders[44].

  - ***binding-queue*** (i.e listening for): nothing
  - ***routing-key*** (i.e publishing to): *document.created*, *snippet.created*, *domain.(created;updated)*

- **Annotator**: creates *annotations* (e.g named-entities, bag-of-words, topic distributions, etc) in *documents*, *snippets* or *domains*. We have developed an NLP Annotator[45] that creates bag-of-words from a given text by normalizing its content through natural language processing tasks such as named entity recognition, lemmatization or part-of-speech (PoS) tagging. The source code[46] is publicly available for reuse. Topic models are also annotators in this framework, as will be seen in section 4.2.1.

  - ***binding-key***: *document.(created;updated)*, *snippet.(created;updated)*
  - ***routing-key***: *annotation.(created;deleted)*

[42] `https://github.com/cbadenes/camel-oaipmh`
[43] `https://github.com/librairy/harvester-datosGobEs`
[44] `https://github.com/librairy/harvester`
[45] `http://librairy.linkeddata.es/nlp`
[46] `https://github.com/librairy/nlp`

- **Modeler**: creates probabilistic topic models for *domains* from the bag-of-word *annotations* of each *document*. We have developed a LDA Modeler[47] as well as a W2V Modeler[48] (the latter for testing purposes).
  - ***binding-key***: *domain.(created;updated)*, *annotation.created*
  - ***routing-key***: *annotation.(created;deleted)*
- **Administrator**: performs user-driven tasks such as reading/writing/updating resources or database queries. As with the other modules, the source code[49] is publicly available for reuse. Together with the API, we have also developed a web interface[50] to visualize documents, their relationships and the topics associated with each *domain*.
  - ***binding-key***: *#.#*
  - ***routing-key***: *domain.(created;updated;deleted)*, *document.(created;updated;deleted)*, *snippet.(created;updated;deleted)*, *annotation.(created;updated;deleted)*

Figure 4.8 shows a sequence diagram that illustrates how modules work collaboratively to create a topic model from documents added to the framework. Each module provides an Application Program Interface (API) over HTTP, that follows the web standards for the RESTful API development, and a Avro[51]-based interface over TCP for efficiency reasons. Among them the communication is done over TCP.

4.2 Reusable Topic Modeling

In order to use an existing topic model in our framework, which is micro-services-oriented, the model itself needs to be a service. This approach decouples the resources used to train a probabilistic topic model (e.g. data format or algorithm implementation), from the resources required to make inferences and thus avoids unexpected incompatibilities. In this way, we simultaneously facilitate the reuse of topic models and also their scalable execution.

[47] `https://github.com/librairy/modelerTopics-service`
[48] `https://github.com/librairy/modeler-w2v`
[49] `https://github.com/librairy/api`
[50] `https://github.com/librairy/explorer`
[51] `https://avro.apache.org`

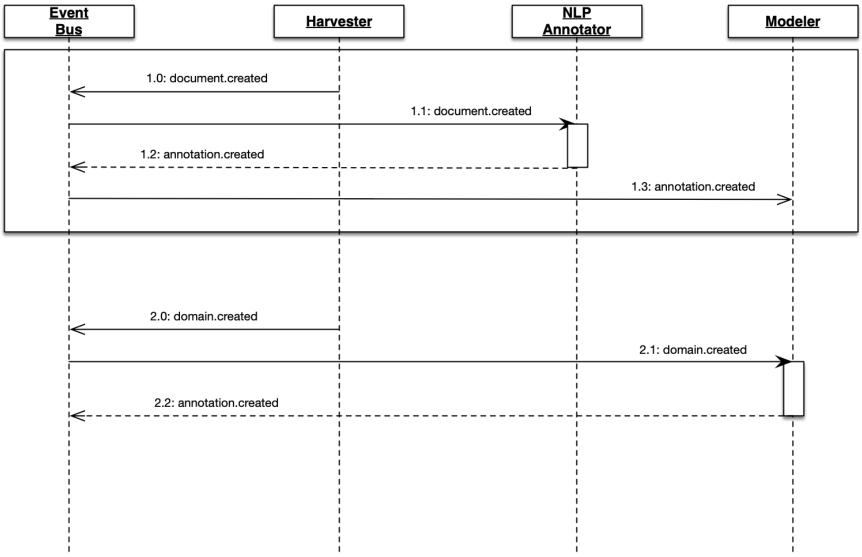


Figure 4.8: Sequence of *events* exchanged between modules to create a topic model from the *documents* added to a *domain*.

4.2.1 Topic Model Publication

We propose distribute topic models as services hosted in online repositories. Models are packaged as OS-level virtualization software that can run reliably applications from one computing environment to another. A model becomes a standalone and executable software package that includes everything needed to use it: code, data, runtime libraries, system tools and settings.

There are several technologies that can virtualize services. Among them, Docker[52] stands out as a de facto standard due to its wide adoption. It is a platform as a service (PaaS) environment that use operating service-level virtualization to deliver software in packages called *containers*. Containers are isolated from one another and bundle their own software, libraries and configuration files. All containers are run by a single operating system kernel and therefore use fewer resources than virtual machines.

Topic Models in *librAIry* are packaged as Docker *containers* and published in online repositories[53] so they can be easily downloaded and run on any machine or software

[52] `https://www.docker.com/`
[53] `https://hub.docker.com/repositories`

solution. Containers not only offers virtualization advantages, but also version and license control, since they handle some information that we use to characterize our models:

- **repository**: model name (e.g. dbpedia-model).
- **author**: model creator (e.g. librAIry)
- **version**: model version (e.g. 3.0)
- **license**: model license (e.g. Apache 2.0)

4.2.2 Topic Model Exploitation

Once a topic model has been packaged as a virtual service (i.e. Docker containers), either assembled as a *annotator* in librAIry or as an online service accessible via its HTTP (based on a Swagger interface, Fig. 4.9) or TCP (based on an Avro[54] interface) API, it should support all tasks that can be performed on a topic model. Taking into account the ***reproducibility*** of a model, the *conditions under which the model has been created*, in relation to the corpus (i.e. number of documents, vocabulary size, language,..), about NLP tasks (i.e stopwords, PoS filtering) and about the model itself (i.e. hyperparameters) should be collected and provided by the service. Taking into account the ***exploration*** of a model, the *topics and their word distributions* must also be available. The word distribution is usually omitted when publishing topic models providing only the top10 or top5 words per topic. This limits the capacity of the model to be exploited in tasks where that information is required, for example this thesis (as will be seen in chapter 4). And finally, taking into account the topic ***inference*** in new texts, which is the most extended ability of these models, the service must be able to calculate topic distributions from texts.

A summary of the tasks, and their purposes, provided by a topic model can be found in Table 4.1.

In order to cover these tasks by topic services, the API must offer operations that, individually or partially, allow them to be achieved. A topic model is: *reproducible* (T1) when its hyperparameters and the configuration of the training set are known,

[54] `https://github.com/librairy/modeler-service-facade/blob/master/src/main/avro/model.avpr`

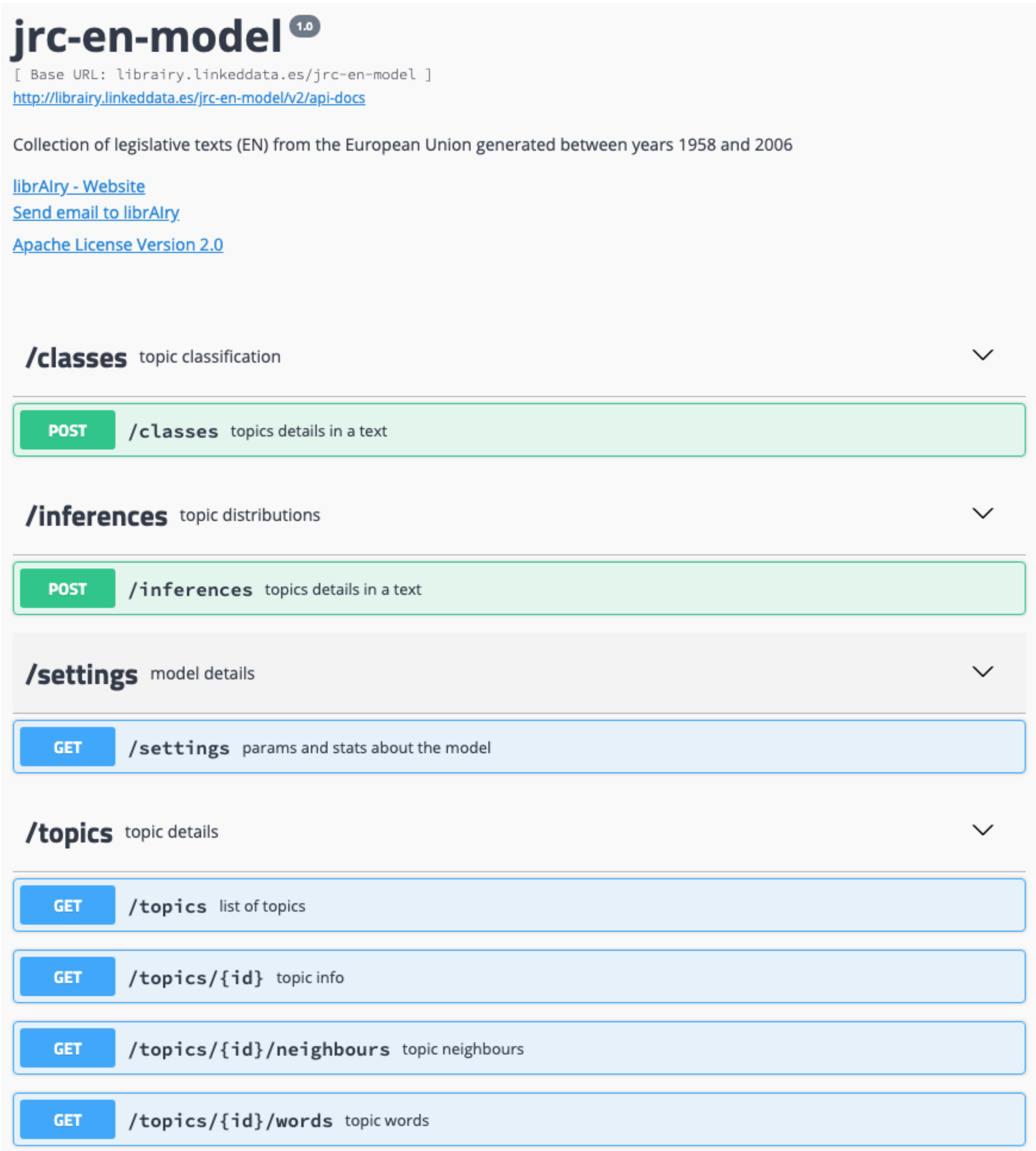


Figure 4.9: Swagger-based web interface of a probabilistic topic model service created with *librAIry*.

Task	Purpose
T1: Reproducibility	create a similar model
T2: Exploration	browse topics and words
T3: Inference	calculate topic distributions

Table 4.1: Tasks and scopes provided by a topic model.

explorable (T2) when the word distributions of each topic are known, and is *interpretable* (T3) when the presence of each topic can be measured in a text. Table 4.2 summarizes the operations offered by a topic model-as-a-service to support the tasks provided by a topic model.

Operations	Tasks
O1: reading of model hyperparameters (i.e alpha, beta, and number of topics)	T1: Reproducibility
O2: reading of pre-processing tasks (i.e stop-words, normalization, PoS filtering)	T1: Reproducibility
O3: reading of learning parameters (i.e number of iterations, seed, likelihood)	T1: Reproducibility
O4: reading of topics described by word distributions (i.e topic relevance)	T2: Exploration
O5: reading of words described by topic distributions (i.e word relevance)	T2: Exploration
O6: calculation of topic distributions in texts (i.e vector of topic distributions)	T3: Inference

Table 4.2: Operations offered by a Probabilistic Topic Model.

The rest of the section describes how the model implements these operations through service methods. A topic model-as-a-service is online available for review[55].

[55] `http://librairy.linkeddata.es/jrc-en-model`

4.2.2.1 Reproducibility Tasks

Through a single request to the model API, a list of parameters is provided to support the O1, O2, and O3 operations. The method is available by both a HTTP-GET request on `/settings` resource and by a TCP request on *getSettings* method. An online example is available here[56].

A detailed list of the parameters provided by the topic model is shown below:

- **algorithm**: method used to train the model (e.g. LDA, LLDA).
- **date**: when the model was created. It follows the ISO-8601.
- **params**: configuration of the learning process.
  - **seed**: numerical value to ensure consistent results (e.g. 1066).
  - **lowercase**: if true, text is converted to lowercase.
  - **topics**: number of topics.
  - **language**: language of texts.
  - **iterations**: number of sampling iterations.
  - **entities**: if true, NER tasks are performed.
  - **max-doc-ratio**: maximum word presence per document ratio (e.g. 0.9).
  - **min-freq**: minimum word presence per number of document (e.g. 5).
  - **alpha**: prior distribution over topic weights in each document (e.g. 0.1)
  - **beta**: prior distribution over word weights in each topic (e.g. 0.01)
  - **part-of-speech**: word classes used in the model (e.g. NOUN, VERB, ADJECTIVE)
  - **top-words**: number of words used to describe a topic (e.g. 10)
  - **stop-words**: list of words removed from the corpus (e.g. quantity, datum)
- **stats**: statistics after the learning process.
  - **loglikelihood**: how much the model fits with the training set.
  - **vocabulary**: number of unique words.

[56] `http://librairy.linkeddata.es/jrc-en-model/settings`

– **topic-coherence**: distance between topics from their top words (e.g min, max, mean, mode).

– **topic-distance**: distance between topics from their word distributions (e.g min, max, mean, mode).

– **corpus**: number of documents in the training set.

4.2.2.2 Exploration Tasks

The exploration task, with its respective operations of reading topics (O4) and words (O5), is broken down into four services:

- **Topic List**(O4). By means of the HTTP-GET */topics*[57] resource and the TCP *getTopics* method, all the topics of the model are listed. Each topic is described with an increasing unique *identifier* (from 0 to the maximum number of topics), a label or *name* in case it has been established, and a *description* with the most relevant words (based on their density distributions).

- **Topic Detail**(O4). By means of the HTTP-GET */topics/_id_*[58] resource and the TCP *getTopic* method, a topic identified by *id* is described by providing its *identifier*, *name*, *description* and *entropy* (i.e how different is with respect the other topics).

- **Topic Words**(O5). By means of the HTTP-GET */topics/_id_/words*[59] resource and the TCP *getTopicWords* method, a list of word distributions for the topic identified by *id* is provided. Every word has its weight (i.e relevance score) with respect to the topic.

- **Topic Neighbours**(O4). By means of the HTTP-GET */topics/_id_/neighbours*[60] resource and the TCP *getTopicNeighbours* method, a list of topics related to the topic identified by *id* is provided. The distance between topics is measured from their word distributions.

[57] http://librairy.linkeddata.es/jrc-en-model/topics
[58] http://librairy.linkeddata.es/jrc-en-model/topics/0
[59] http://librairy.linkeddata.es/jrc-en-model/topics/0/words
[60] http://librairy.linkeddata.es/jrc-en-model/topics/0/neighbours

4.2.2.3 Inference Tasks

A single request with the text to be analyzed returns a list with the presence of each topic (O5) in that text. The method is available by both a HTTP-POST request to `/inferences` or a TCP request to the *createInference* method, with a JSON message containing the text to be analyzed.

4.3 Summary

In Section 4.1 we have described *librAIry*, the framework to cover the entire topic model life-cycle in a scalable way. Algorithms and tools coming from different technologies work collaboratively to process and analyze huge collections of textual resources creating and using probabilistic topic models.

We tested and validated *librAIry* by using the framework in some real world scenarios such as DrInventor[61], where thousands of scientific publications were processed, TheyBuyForYou[62], where hundreds of thousands of public procurement texts were analyzed, or CorpusViewer[63], where millions of patents were automatically organized. Thus, *librAIry* has proven to be a text processing framework and addresses the first technical objective of this thesis (T01, *create a scalable platform for topic modeling*).

librAIry has been designed to represent corpora by organizing the data in three levels of detail: *snippets* to reflect parts or pieces of texts, *documents* to represent full texts, and *domains* to group documents. Transversally there are *annotations*, which allow providing more details to any of them. Figure 4.2 shows how two scientific articles published in the same conference and that mention a same resource in their papers can be represented. On this representation model based on SEDA architectures, actions over resources and status change notification events are introduced, which enable to distribute the processing of resources. Figure 4.7 shows the four modules involved in processing the resources. *Harvester* modules create new *documents*, *snippets* and *domains*. *Annotator* modules react to each new resource and introduce *annotations*. *Modeler* modules create new topic models for each new *domain*. And *Admin* modules perform administrative tasks and allow users to read the data. As shown in figure 4.8,

[61] `http://drinventor.eu`

[62] `https://theybuyforyou.eu`

[63] `https://www.plantl.gob.es/tecnologias-lenguaje/actividades/plataformas/Paginas/corpus-viewer.aspx`

modules coordinate their actions by reacting to the notifications. The actions are then executed in parallel through a distributed workflow and the first research objective of this thesis is addressed (RO1, *define a distributed text-processing model for creating large probabilistic topic models*).

In Section 4.2 we propose the publication of topic models as web services that can be used in a scalable way from external solutions or integrated into the *librAIry* framework. Regardless of their API, since they works both over HTTP and over TCP, there are three types of tasks that guide the definition of the service: *reproducibility*, *exploration*, and *inference*. Tables 4.1 and 4.2 detail the operations supported by the service to cover these tasks. The definition of a topic model-as-a-service covers the second research objective of this thesis (R02, *define a template to package probabilistic topic models as web services*).

And finally, in order to facilitate the reuse of the topic models published as web services, section 4.2.1 presents an online repository based on virtual services. This covers the second technical objective of the thesis (T02, *create a repository of topic-based web services*).

Chapter 5

Explainable Topic-based Associations

As stated in Chapter 3, one of our hypotheses aims to determine whether is possible to semantically relate texts from their most relevant topics (H1.2). In particular, our goal is to determine whether two documents can be related by identifying their most representative topics.

However, as seen in Section 2.2.2, interpreting how documents are related from their topic distributions is hard when using density-based measures. The same pair of documents may vary their distance from each other when using topic models with different dimensions to represent them, as shown in figure 2.1. High dimensional models create more specific topics than models with fewer dimensions, and this topic specificity influences the way in which topic distributions are related.

In order to better understand the relations derived from topic distributions, Section 5.1 compares scientific articles from their representations based on full-content, abstracts (i.e manual summaries), or summaries created automatically. Two types of metrics are considered: (i) *internal-representativeness*, focused on describe the content, and *external-representativeness*, focused on discover relations (Badenes-Olmedo et al., 2017a).

In Section 5.2, once we know how the topics are used to represent and relate texts, we propose a new topic-based annotation based only on the most representative topics and a new distance measure that takes advantage of these representations (Badenes-Olmedo et al., 2017c). The main goal is not only to cluster texts through their most relevant topics, but also to facilitate the interpretation of their relations.

5.1 Topic-based Relations

This section studies the ability of topic distributions to capture the representativeness of a text through the relations that can be derived from it. In particular, it examines the performance offered by topic-based representations to describe scientific articles from their full-texts compared to representations based on summaries (e.g. *abstract*). The objective is twofold, to analyze comparisons based on topic distributions on the one hand, and to identify strengths and weaknesses when using *abstracts* to compare scientific articles on the other hand . Two novel measures are proposed based on the capability of the summary to substitute the original paper (Figure 5.1): (1) *internal-representativeness*, which evaluates how well the summary represents the original full-text and (2) *external-representativeness*, which evaluates the summary according to how the summary is able to produce a set of related texts that are similar to what the original full-text has triggered.

Recent studies (Sciences and life, 2016; Westergaard et al., 2017) have shown that text mining of full research articles give consistently better results than using only their corresponding abstracts. Given the size limitations and concise nature of abstracts, they often omit descriptions or results that are considered to be less relevant but still are important in certain Information Retrieval tasks. Thus, when other researchers cite a particular paper, 20% of the keywords that they mention are not present in the abstract (Divoli et al., 2012).

An analysis about the *representativeness* of research article summaries based on topic distributions is presented, considering those based exclusively on abstracts and those based on their discursive structure (*approach*, *challenge*, *background*, *outcomes* and *future work*)(Simone Teufel, 2010). The *representativeness* of a summary with respect to the original full-text is assumed as the degree of relation with the original one (*internal-representativeness*), along with the capacity of mimicking the full text when finding related items (*external-representativeness*). In order to quantify this notions of internal-external representativeness, a probabilistic topic model is trained to have a vectorial representation of each text retrieved from a paper: full content-based and summary-based (Figure 5.1). The vectorial representations of full-papers

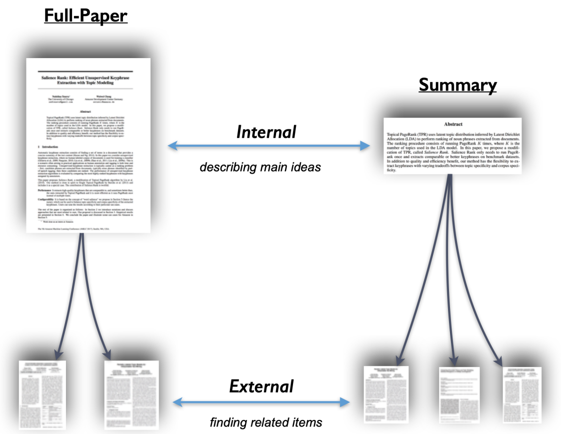


Figure 5.1: Internal and External Representativeness.

is used to measure the distance between them and those derived from abstract or summaries (*internal-representativeness*), and also to find similar documents (*external-representativeness*) based on the distance between their vectorial representations. An upper distance threshold is specified to filter less similar pairs and compose a set of related papers for each paper. Then, a comparison in terms of *precision* and *recall* is performed between sets obtained by only using the vectorial representation of full-papers, against sets produced by using other kind of summaries.

5.1.1 Research Articles Summaries

Some approaches have been proposed to summarize scientific articles (Cohan and Goharian, 2015) taking advantage of the citation context and the document discourse model. We have used the scientific discourse annotator proposed by (Ronzano and Saggion, 2015) to automatically create summaries from research papers by classifying each sentence as belonging to one of the following scientific discourse categories: *approach*, *challenge*, *background*, *outcomes* and *future work*. These categories were identified from the schemata proposed by (Teufel et al., 2009) with an original purpose of characterizing the content of Computer Graphics papers.

The annotator is based on a Support Vector Machine classifier that combines both lexical and syntactic features to model each sentence in a paper. The tool[64] was in-

[64]http://backingdata.org/dri/library/

tegrated in our *librAIry* framework through the *Rhetoric Module*[65] to automatically annotate research papers with their rhetorical content.

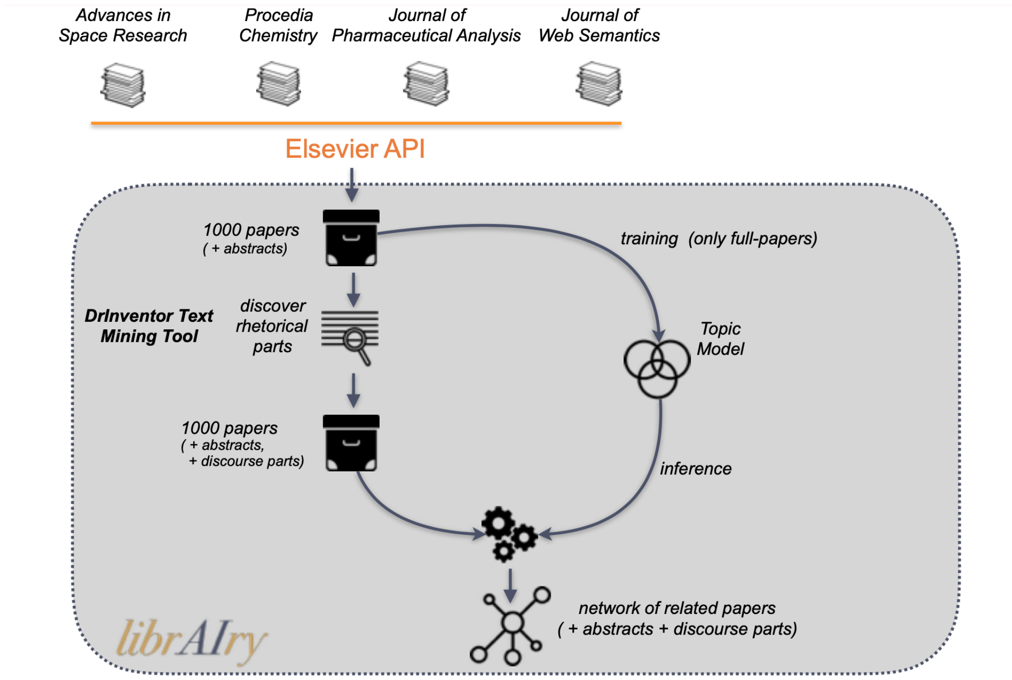


Figure 5.2: Experiment to analyze the ability of topic-based representations to create relations from summaries *vs* full-texts.

5.1.2 Feature Vectors

A representational model is required not only to measure distances between text fragments but, more importantly, to help to understand the differences in their content. As seen in Section 2.2.2, topic models are widely used to uncover the latent semantic structure from text corpora. In particular, Probabilistic Topic Models represent documents as a mixture of topics, where topics are probability distributions over words. Latent Dirichlet Allocation (LDA)(Blei et al., 2003) is the simplest *generative* topic model that makes it possible to characterize documents not previously used during the training task. This is a key feature for our evaluations because, although the model used for the experiments will be trained from the full-content of papers, it will be also used to describe the texts summaries.

[65] https://github.com/librairy/annotator-rhetoric

Thus, we have used a LDA model to describe the inherent topic distribution of papers in the corpus. Some hyper-parameters need to be estimated: the *number of topics (k)*, the concentration parameter (α) for the prior placed on documents' distributions over topics and the concentration parameter (β) for the prior placed on topics' distributions over terms. Since the target of this experiment is not to evaluate the quality of the representational model, but to compare their topic distributions, we accepted as valid values those widely used in the literature: $\alpha = 0.1$, $\beta = 0.1$, and $k = 2 * \sqrt{n/2} = 44$ where n is the size of the corpus.

5.1.3 Similarity Measure

Feature vectors in Topic Models are topic distributions expressed as vectors of probabilities. Hence we opt for *Jensen-Shannon divergence* (JSD) (See Section 2.1) instead of the commonly used *Kullback-Liebler divergence* (KLD). The reason for this is that KLD is not defined when a topic distribution is zero and is not symmetric, what does not fit well with semantic similarity measures which in general are symmetric (Rus et al., 2013). JSD considers the average of the distributions as follows :

$$JSD(p,q) = \sum_{i=1}^{T} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{T} q_i * \log \frac{2 * q_i}{q_i + p_i} \tag{5.1}$$

where T is the number of topics and p, q are the topics distributions.

And the *similarity measure* used in our analysis is based on the JSD transformed into a similarity measure as follows (Dagan et al., 1999) :

$$similarity(D_i, D_j) = 10^{-JSD(p,q)} \tag{5.2}$$

where D_i, D_j are the documents and p, q the topics distributions of each of them.

5.1.4 Evaluation

The corpus used in the experiments was created by combining journals in different scientific domains such as *Advances in Space Research*, *Procedia Chemistry*, *Journal of Pharmaceutical Analysis* and *Journal of Web Semantics* (Figure 5.2). In total 1,000 papers were added, 250 from each journal. Both the abstract and the *full-content* of these documents were directly retrieved from the Elsevier API [66] by using our *Harvester module* [67]. The code used to perform the analysis along with the results obtained are publicly available[68].

Since the annotation process to automatically discovers the rhetorical parts of a research paper (Section 5.1.1) is sensitive to the structure of the phrases that are used when writing the text, only 20% of papers in the corpus could be fully annotated with all the fragments considered. In fact, these categories are not present in the same proportion in the corpus: *approach* (90%), *background* (78%), *outcome* (73%), *challenge* (57%) and *future work* (21%)

5.1.4.1 Internal Representativeness

The *internal-representativeness* of a text measures the similarity of a summary against the original full-text. This measure is based on the JSD between the topic distribution of each of them.

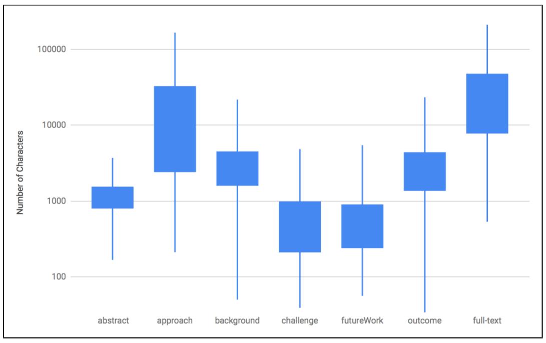


Figure 5.3: length of summaries

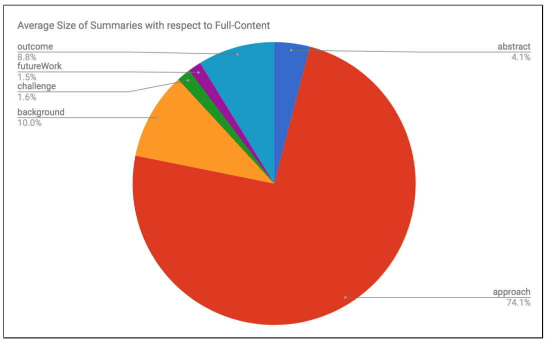


Figure 5.4: length of text parts

Since LDA considers documents as *bag-of-words*, the text length (e.g. full-content or summaries) affects the accuracy of the topic distributions inferred by the topic model described in Section 5.1.2. The occurrences of words in short texts are less discriminative than in long texts where the model has more word counts to know how words are related (Hong and Davison, 2010). In view of the above, the *approach*, the *background* and the *outcome* content of a paper generate more accurate topic distributions than those created from other approaches such as the abstract (Figure 5.3). Also, the relative

[66]https://dev.elsevier.com
[67]https://github.com/librairy/harvester-elsevier
[68]https://github.com/librairy/study-semantic-similarity

presence of each of them in a paper (figure 5.4) shows an unexpected result when compared to the IMRaD format (Nair and Nair, 2014). This style proposes to distribute the content of an abstract, and by extension the full-paper, as follows: *Introduction*(25%), *Methods*(25%), *Results*(35%) and *Discussion*(15%). However, the results (figure 5.4) show that *Method* section (*approach* content) is more extensive than *Results* section (*outcome* content) in our corpus.

All pairwise similarities between full-papers, abstracts and rhetorical-based summaries are calculated to measure the ***internal-representativeness*** of a summary with respect to the original text, i.e. the topic-based similarity value (equation 5.2) between the probability distributions of the full-text and each of the summaries. Results (table 5.1) suggest than summaries created from the *approach* content are more representative than others, i.e. the distribution of topics describing the text created from the *approach* content is the most similar to the one corresponding to the full-content of the paper.

	Min	Lower Quartile	Upper Quartile	Max	Dev	Median
abstract	0.0489	0.9109	0.9840	1.0000	0.1443	0.9741
approach	**0.0499**	**0.9969**	**1.0000**	**1.0000**	**0.0872**	**0.9998**
background	0.0463	0.8967	0.9937	0.9988	0.2037	0.9822
challenge	0.0426	0.7503	0.9517	0.9940	0.2224	0.8829
futureWork	0.0000	0.6003	0.9435	0.9948	0.2842	0.8814
outcome	0.0485	0.9267	0.9925	0.9990	0.1721	0.9835

Table 5.1: Internal-Representativeness

5.1.4.2 External-Representativeness

The *external-representativeness* metric tries to measure how different is the set of related documents obtained from summaries with respect to those derived from the original full-text. In terms of *precision*, *recall* and *f-measure*, a comparison has been performed to analyze the behavior of the summaries when trying to discover related content compared to use the full-text of the article.

By using the same topic model previously created, similarities among all pairs of documents were also calculated according to equation 5.2. Then, a minimum score or similarity threshold is required to define when a pair of papers are related. Each threshold is used to create a gold-standard which relates articles to others based on their similarity values. In order to discover that lower bound of similarity, a study about trends in the similarity scores (fig 5.5) as well as distributions of topics in the corpus (fig 5.6) was performed. We can see that topics are not equally balanced across papers. This fact generates separated groups of strongly related papers. We think this phenomena is due to our usage of a corpus created from journals where different domains are equally balanced. Then, we considered a similarity score equals to 0.99 (fig 5.5) as the threshold from which strong relations appear. However, to cover different interpretations of similarity, from those based on sharing general ideas or themes to those that imply to share a more specific content, the following list of thresholds was considered in the experiments: 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99.

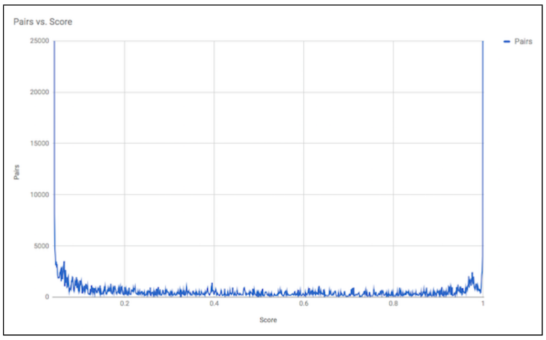


Figure 5.5: number of pairwises by similarity score (rounded up to two decimals)

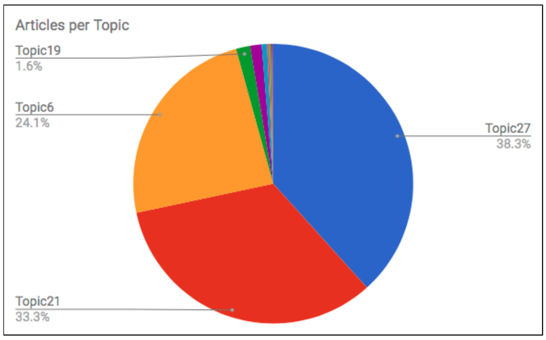


Figure 5.6: topics per article with value above 0.5

For each similarity threshold, a gold-standard was created based on considering as related those papers with a similarity value upper than the selected threshold. Results (figure 5.7) comparing the related papers inferred from the full-content with those inferred from the partial-content representation (i.e. abstract or rhetorical parts) suggest that strongly related papers are mainly discovered by using the summary created from the *approach* section. The reason for this may be based on the average size of this type of summaries or the particular content included in this part of a paper. While other summaries include more general-domain words, the *approach* content includes more specific words that describe the method or the final objective of the paper. So, for higher similarity thresholds, i.e. for strongly related papers, the recommendations discovered by using the *approach* are more precise than those discovered by using the abstract.

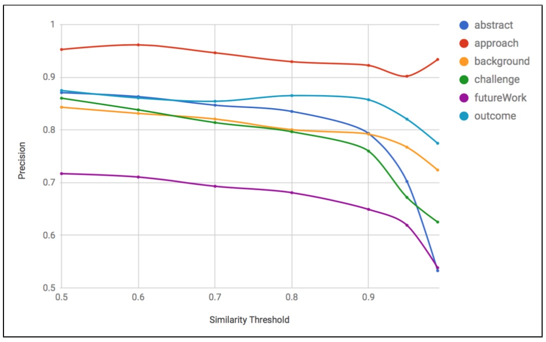


Figure 5.7: Precision at different similarity thresholds

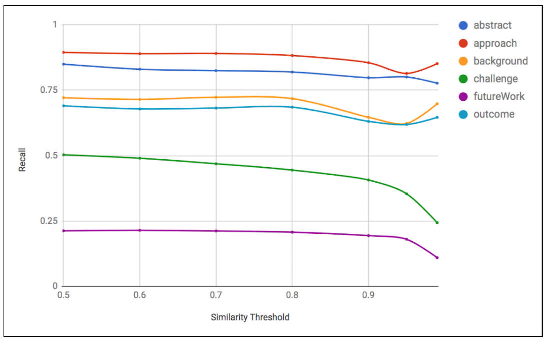


Figure 5.8: Recall at different similarity thresholds

In terms of *recall* (figure 5.8), the upward trend followed by the *approach*, the *outcome* and the *background* content remarks the assumption of summaries containing key words allow to discover more similar papers than others. Moreover, since *recall* overlooks false-negatives classifications, it suggests that these parts of a research paper share more words than others with strongly related papers but they may also present commonalities with highly related papers, except in case of *approach* which still exhibits higher *precision*.

As expected, only summaries created from the *approach*, the *outcome* and the *background* content maintain high accuracy values (fig 5.9) even for high similarity thresholds. Along with the results showed in figure 5.10, where the same three rhetorical classes present the lowest standard deviation over the *f-measure*, they can be considered as the most robust summaries containing the ideas that better characterize the paper compared to others.

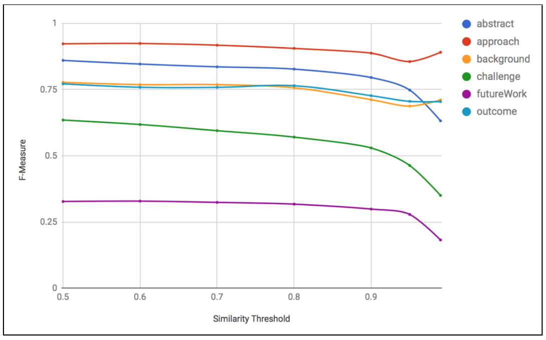


Figure 5.9: f-measure performance

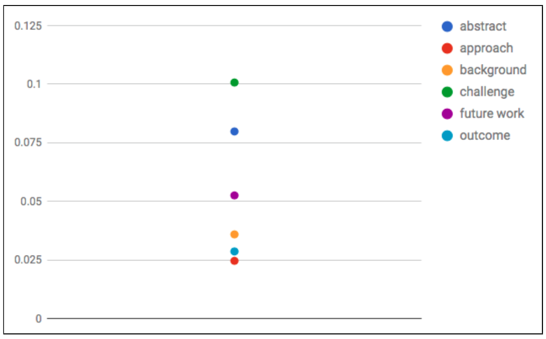


Figure 5.10: f-measure deviation

5.1.5 Conclusion

Topic-based relations have been studied among scientific documents based on their abstract sections with respect to summaries corresponding to their scientific discourse categories. For this purpose, two novel measures have been proposed: (1) *internal-representativeness* and (2) *external-representativeness.*

Results show that summaries created from the *approach*, *outcome* or *background* content of a paper describe more accurately its full-content in terms of overall ideas and related documents than abstracts. Although those summaries are more extensive in number of characters than other with similar *precision* such as the abstract content, they have proven to be particularly helpful discovering strongly related papers, i.e. papers with a similarity value close to 1.0.

5.2 Topic-based Clustering

Once we have a better understanding of the behavior of topic models to represent texts, and the relationships that can be derived, we examine their usefulness in browsing document collections. A way to explore the knowledge inside document collections is by moving from one information element to another based on certain criteria that relates them. This approach requires to calculate a similarity matrix with all possible comparisons between elements, so we can later select the most pertinent ones. Since computing a $n \times n$ matrix takes $O(n^2)$ time, obtaining all possible pairs of similarities in a large collection of documents can be unfeasible because of the exponential cost of comparing every pair of elements.

In order to reduce the complexity, some approaches have introduced mechanisms (mainly pre-election methods) to alleviate the problem of making this calculation over the whole set of pairs in the collection. However those methods are still quite costly.

In this section we propose a novel clustering technique based on topic model distributions that reduces the complexity to find relations between documents in a large corpus of textual documents, without compromising efficiency and providing additional information about relations. A detailed description of our algorithm is given in Section 5.2.1. The experiments to verify the efficiency and effectiveness of our clustering algorithms using real data, and demonstrate that our approach is competitive enough against both a centroid-based and a density-based clustering baselines are described

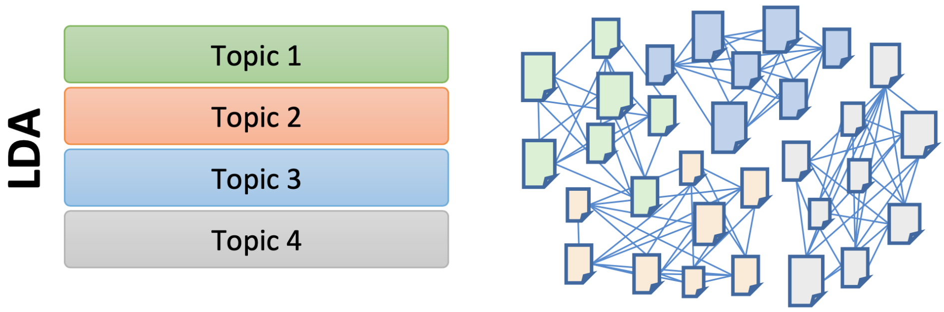


Figure 5.11: Probabilistic Topic Models, and in particular Latent Dirichlet Allocation (LDA), can efficiently divide the search space and speed up the process of finding relations among documents inside big collections.

in Section 5.2.2. The most relevant results and conclusions are finally presented in Section 5.2.3.

5.2.1 Most Relevant Topics

Our algorithm to identify the most relevant topics draw inspiration from other clustering techniques to divide the initial space of elements into smaller sub-groups where the complexity of calculating all possible distances is significantly reduced. Existing unsupervised approaches based on centroids or density measures require to make comparisons between elements to find groups of similar elements in the collection. They usually follow an iterative methodology to produce the final solution, based on calculating distances between the elements inside each intermediate state. A naïve approach would need to calculate all possible distances between elements, which takes $O(n^2)$ time for a $n \times n$ matrix. That makes it impossible to apply such techniques on large collections of documents, since the cost of comparing each element with the others escalates quickly. For those big volumes of data, a clustering task that only takes linear time to discover the clusters can significantly alleviate this problem. For example, a classification method that does not require any other data except the element information to assign the item to the corresponding cluster will take $O(n)$ time to compose those groups.

The classification method needs to take advantage of both the vectorial representations of the documents and the similarity measure used to relate them in a corpus.

Since the representational model considered is based on Probabilistic Topic Models (and more specifically on LDA), the classification method leverages on the particular behavior of Dirichlet distributions, which describes each document by a density vector where the sum of all the probability values must be equal to 1.0. Thus, analyzing the relations between the topics that compose a topic distribution becomes more important than comparing their probability values with another topic distribution.

Our hypothesis is that, *given a collection of topic distributions, an unsupervised classification with high precision and linear computing time can be performed by considering only the topic distribution of each document and without needing to further compare it with other document's distributions.*

All algorithms have been compared in terms of *cost*, *effectiveness* and *efficiency* (Halkidi et al., 2001). *Cost* is based on the number of pairwise similarity values. *Effectiveness* handles relevance measures such as *precision* and *recall*. And *efficiency* tries to measure the overall balance between *cost* and *effectiveness*. More details about those measures will be included in Section 5.2.2.

5.2.1.1 Trends-based Clustering Method

Topic distributions are formalized as probability distributions following a Dirichlet distribution, so their probability values sum to 1. In this way, the relevance of a topic is influenced and at the same time influences the relevance of the others items in the distribution. Our first approach named *Trends on Dirichlet distribution-based Clustering* (TDC) considers changes in the relevance, i.e. probability values of the topics instead of directly relying on the scores associated to a given topic distribution (Figure 5.12). It expresses the oscillations between topic weights considering a fixed order between them. The order can be any, as long as it remains constant in all distributions. Thus, a *probability-vector* composed by n density values is translated to a *trend-expression* made out of $n-1$ trend-values such as (1) upward, (2) downward and (0) sustained. This *trend-expression* will identify the cluster the distribution falls into, and therefore the corresponding item belongs to. TDC is defined as:

$$TDC(P) = T \tag{5.3}$$

where: $T_i = 1$, when $P_i < P_{i+1}$
$T_i = 2$, when $P_i > P_{i+1}$
$T_i = 0$, when $P_i = P_{i+1}$

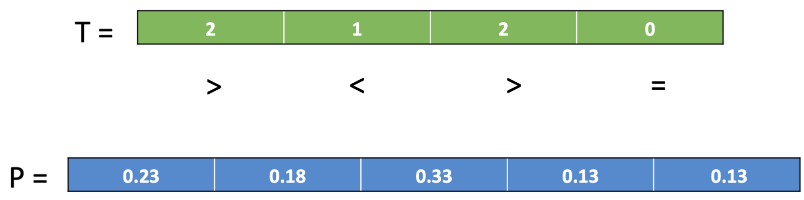


Figure 5.12: TDC considers variations across consecutive topics inside a document's topic distribution.

For example, given the distribution $P_1 = [0.23, 0.18, 0.33, 0.13, 0.13]$, the assigned cluster will be $T = 2120$. The first value is 2 because 0.23 is greater than 0.18 (same for other values).

5.2.1.2 Ranking-based Clustering Method

We propose a clustering technique named *Ranking on Dirichlet distribution-based Clustering* (RDC) that only considers the top n topics from the ranked list of probability distributions to classify similar topic distributions (Figure 5.13). It is based on the focal document selection proposed by (Towne et al., 2016) to validate LDA-based similarity algorithms against human perception of similarity. RDC is defined as:

$$RDC(P) = R \tag{5.4}$$

where $\forall i \in R, R_i >= R_{i+1}$ and $\forall j \in P, R_1 >= P_j$

This is based on the assumption that the highest weighted topics have a high influence in the rest of topics in terms of calculating distances, when comparing continuous multivariate probability distributions. Since similarity measures (Section 2.1) based on probability distributions are oriented to determine the uncertainty of the distribution, when a mixture of probability distributions is considered, as in the case of Topic Models, the top n distributions (i.e. the most relevant topics) should be sufficient to allow us grouping similar distributions. Taking into account the above considerations, the RDC algorithm classifies a topic distribution according to only n highest probability values. For instance, given the following topic distribution: $P_2 = [0.23, 0.18, 0.33, 0.13, 0.13]$, the assigned cluster is *3* from RDC-1 because that is the topic with the highest weight.

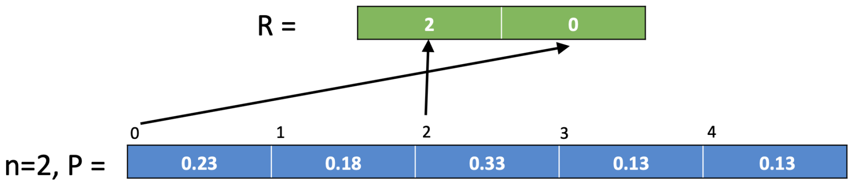


Figure 5.13: RDC only considers the top n topics from the ranked list of probability distributions.

5.2.1.3 Cumulative Ranking-based Clustering Method

A variant of the previous algorithm, named *Cumulative Ranking on Dirichlet distribution-based Clustering* (CRDC), also aims to discover the most representative topics that can help to group similar topic distributions. While RDC is based on a fixed number of topics, CRDC is based on the cumulative sum of the weights of the highest topics (Figure 5.14). The number of topics is now dynamically determined by a threshold, and once this threshold is reached no more topics are considered. CRDC is defined as:

$$CRDC(P) = C \tag{5.5}$$

where $\forall i \in C, C_i >= C_{i+1}$ and $\sum_{i=1}^{T} C_i >= w$ with T size of C, and w a cumulative weight threshold.

For instance, considering a CRDC algorithm with a cumulative weight threshold of 0.9, and the following topic distribution: $P_3 = [0.36, 0.58, 0.05, 0.01]$. The assigned cluster will be *21*. To come up with this cluster, a ranked list of topics based on their weights is first calculated, $R_p = 2|1|3|4$. Then, a sum of weights according to the order described by R_p is performed. When the accumulated sum is greater than the threshold, the topics taking part of the sum will be selected to "label" the cluster. In this case, the cumulative weight threshold is 0.9 therefore using only the first two topics we exceed the threshold: $w = 0.58 + 0.36 = 0.94$

5.2.2 Evaluation

In this section we present the experimental setup for evaluating our trends-based (TDC), ranking-based (RDC) and cumulative ranking-based (CRDC) clustering approaches, considering both JS divergence and He distance as similarity measures. We describe the datasets and baseline algorithms that will be used for comparison.

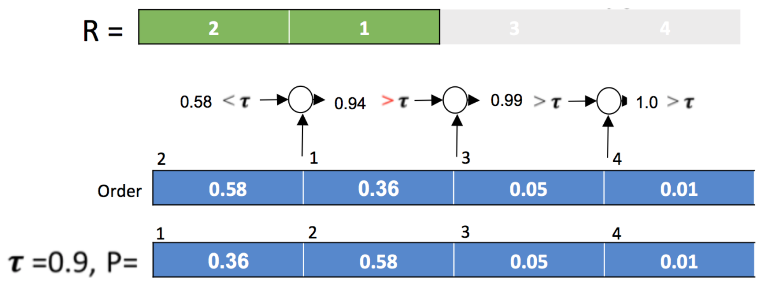


Figure 5.14: CRDC only considers the top n topics until the sum of the weights of the highest topics exceeded a given threshold.

5.2.2.1 Datasets

We used two datasets to evaluate the performance of the algorithms. The first dataset, DIRICHLET-RANDOM-MIXTURE (DRM), is synthetic (Badenes-Olmedo et al., 2017d). To generate the dataset, we sampled k probabilistic distributions from a *randomly k-dimensional selector* based on Dirichlet distributions. This implies that all probabilities must to sum to 1 for each sampled point. The number of sampled points from this mixture of Dirichlet distributions is $n = 1000$.

The second dataset has been created from a collection of research papers published in the *Advances in Engineering Software* (AIES) journal. They were retrieved from the Springer API by using the librAIry (Badenes-Olmedo et al., 2017b) framework and a Topic Model based on LDA was created from them. The sample is also composed by $n = 1000$ documents.

Topic models were trained from these datasets by using the criteria described by (Steyvers et al., 2006): $\alpha = 50/k$, $\beta = 0.01$ and $k = 2 * (\sqrt{(n/2)})$, where k is the number of topics and n is the number of documents. Since both datasets contain 1000 documents (n), the hyper-parameters α and β are assigned as follow: $\alpha = 1.136$, and $\beta = 0.01$, and the number of topics is fixed to $k = 44$. Further tuning of the settings is not crucial in this evaluation process, because we are not focusing on the quality of the model but on the efficiency when calculating similarities from their representational distributions.

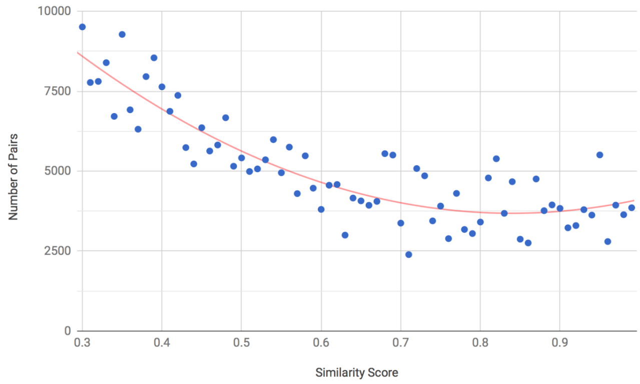


Figure 5.15: Similarity values grouped by frequency in AIES

5.2.2.2 Settings

Since there is no unified criteria to select a threshold inside the distance scores spectrum that allows us to determine when two documents are similar, we decided to study the distribution of similarity values calculated from all pairwise comparisons. In Figure 5.15, the result of grouping all similarities by the two most representative decimals, i.e. the first two decimals of the similarity value, is shown. Then, a polynomial function (red line) is approximated to describe the trend of these values. In this function, the similarity score 0.83 emerges as a global minimum and has been used for filtering out the non-similar document pairs.

5.2.2.3 Baselines

We compare the performance of TDC, RDC and CRDC algorithms against the following baselines:

- *K-Means* as a centroid-based clustering approach.
- *DBSCAN* as a density-based clustering approach.
- *Random*, which randomly selects R from the dataset

Initially, *K-Means* (Bahmani et al., 2012) randomly composes a set of centroids and assigns each point of the sample to its nearest cluster based on a distance measure. Then, a new set of centroids is calculated from the previous ones according to the assigned points. This process is repeated until the set of centroids does not change significantly between consecutive iterations or a maximum number of iterations is reached. The *scalable K-Means* approach used in our experiments is an improved version of *k-means* which obtains an initial set of centers ideally close to the optimum solution. The algorithm implemented at the Apache Commons Math library [69] was used in the experiments. Based on empirical results, the best configuration is: $k = number - of - topics = 44$ and $maxIterations = 50$

A widely known density-based algorithm is $DBSCAN$ (Ester et al., 1996), which compose clusters from the neighborhood of each point considering at least a minimum number of points and a given radius. Thus, it requires to specify the radius of the point's neighborhood, *Eps*, and the minimum number of points in the neighborhood *MinPts*. Based on empirical results, the best results were obtained with the following configuration: $eps = 0.1$ and $minPts = 50$

The *Random* algorithm takes as input a parameter m and randomly divides the dataset into m equal-sized groups of similar documents. For the evaluation, m was set to the number of topics, the dimension of the dataset.

With respect to the proposed algorithms and taking into account empirical results, the RDC algorithm is set to use the *top1* highest topics, and the cumulative weight threshold for the CRDC algorithm is set to 0.9.

5.2.2.4 Measures

A gold-standard is created for each dataset and distance metric considered. They are created by calculating all pairwise similarities from their documents. Since the $n \times n$ similarity matrix requires $O(n^2)$ time to be calculated, the selected size of datasets has not been too large $n = 1000$.

We considered three measures to evaluate our algorithms with respect to the baseline:

- ***cost***: based on the number of similarity score calculations required by the algorithm:

$$cost = (reqSim - minSim)/(totalSim - minSim) \tag{5.6}$$

  The *minSim* corresponds to the number of similar documents obtained from using the *threshold* score previously mentioned in section 5.2.2.2. The *totalSim*

[69] http://commons.apache.org/proper/commons-math/

corresponds to the Cartesian product of existing documents: $totalSim = n * n = 1,000,000$. And the $reqSim$ corresponds to the number of similarities calculated by the algorithm.

- ***effectiveness***: based on *precision* and *recall*. It expresses the quality of the algorithm:

$$effectiveness = \frac{precision^2 + recall^2}{2} \tag{5.7}$$

- ***efficiency***: based on the previous ones, it express a compromise between quality and performance:

$$efficiency = effectiveness - cost \tag{5.8}$$

5.2.2.5 Results

The code used to evaluate the algorithms along with the results obtained are publicy available(Badenes-Olmedo et al., 2017d).

In terms of ***effectiveness*** (Figures 5.16 and 5.17), the results highlight that *K-Means* and *CRDC* outperform the other algorithms. *K-Means* was expected to be a top performer because the algorithm itself performs comparisons to map clusters. The fact that *CRDC* has such good performance encourages us to think that, in fact, the most relevant topics when they altogether exceed a certain high weight threshold, are those that best represent the document and allow to group together similar documents. However, as shown in tables 5.2, 5.3, 5.5 and 5.4, considering a fixed number of more relevant topics (*RDC*) or considering the trend of their weights (*TDC*) does not seem to perform so well on aggregating similar documents, since their *precision* and *recall* values are very low in both cases. It is surprising that the *DBSCAN* has such low value. Taking a look at its *precision* and *recall* values, and also seeing the number of groups that each algorithm has created (Figure 5.18), we believe that having a corpus containing a very cohesive set of documents (all papers in corpus belong to the same journal) affects the performance of this algorithm since it divides the corpus into a lower number of groups. This way, it obtains high values of *recall* because most of the pair-wise distances are computed, but very low *precision*.

The results also show that the behavior of the algorithms does not differ significantly when using different similarity measures, for example JS divergence (Figure 5.16) and

He distance (Figure 5.17). This highlights the importance of the documents' topic distributions to successfully classify them into smaller groups of similar items, while other particular aspects such as the distance or similarity metric used to compare them are less influential.

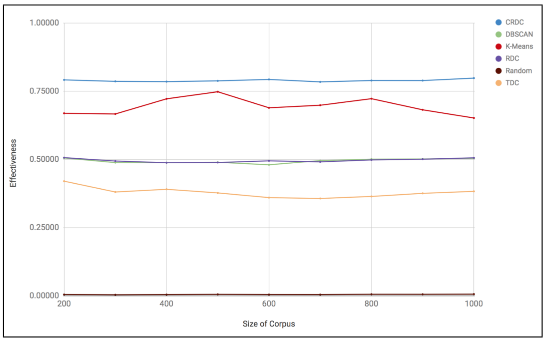


Figure 5.16: Effectiveness (JS-based) in AIES

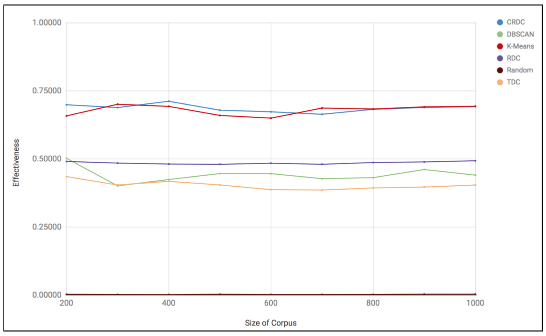


Figure 5.17: Effectiveness (He-based) in AIES

Size	CRDC	DBSCAN	K-Means	RDC	TDC	Random
200	0.94	0.10	**0.96**	0.31	0.42	0.12
300	0.93	0.15	**0.94**	0.30	0.39	0.08
400	**0.93**	0.15	0.89	0.29	0.39	0.09
500	**0.92**	0.30	0.90	0.28	0.38	0.09
600	**0.92**	0.19	0.88	0.28	0.38	0.08
700	**0.92**	0.20	0.91	0.28	0.38	0.09
800	**0.92**	0.12	0.89	0.30	0.39	0.10
900	**0.92**	0.13	0.87	0.30	0.40	0.10
1000	**0.93**	0.13	0.90	0.30	0.40	0.10

Table 5.2: Precision (JS-based) in AIES

In terms of ***cost*** (Figures 5.19 and 5.20), the best clustering algorithm, as expected, is based on *random* selection. This is due to the fact that the number of pairs compared by this algorithm is always the minimum, given the dataset is simply randomly divided into m equal-sized groups, where m is equals to the number of topics, i.e. dimension of the dataset. Since *K-Means* and *DBSCAN* make comparisons between documents until their internal condition is satisfied, they are the most inefficient approaches. *K-Means* involves the highest cost because it compares all the documents with the 44 centroids

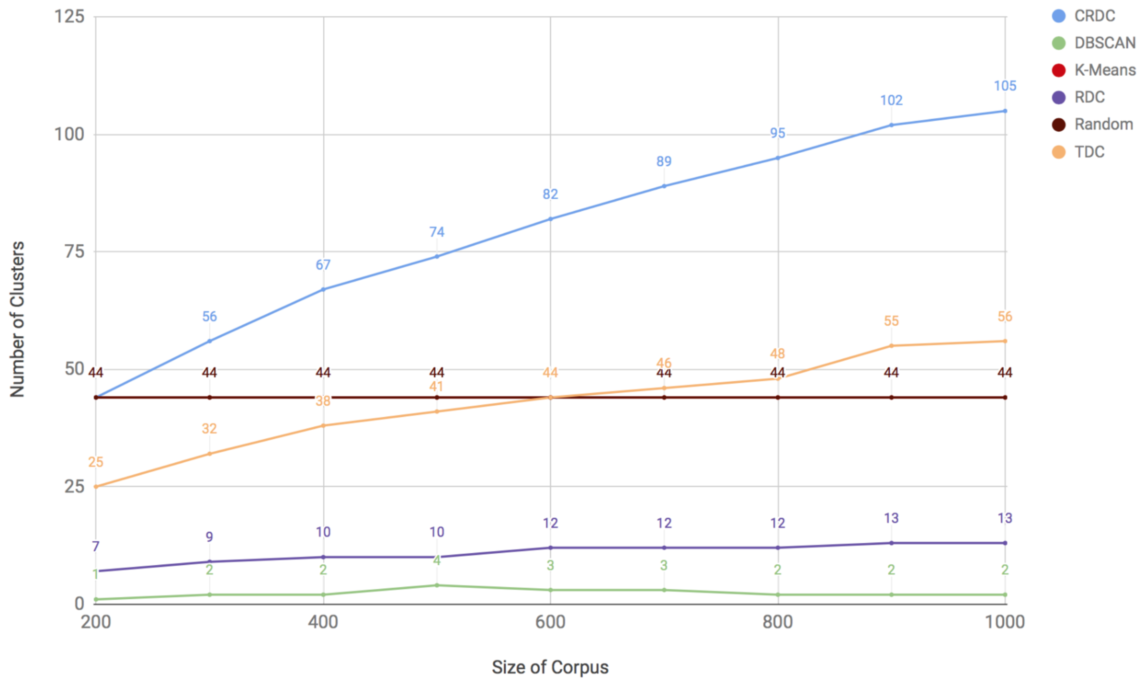


Figure 5.18: Clusters in AIES

Size	CRDC	DBSCAN	K-Means	RDC	TDC	Random
200	0.75	0.07	**0.84**	0.23	0.08	0.33
300	0.74	0.08	**0.83**	0.23	0.06	0.32
400	**0.76**	**0.09**	0.76	0.22	0.06	0.32
500	0.73	0.08	**0.74**	0.21	0.08	0.31
600	0.72	0.08	**0.73**	0.21	0.06	0.30
700	0.71	0.10	**0.76**	0.21	0.06	0.30
800	0.73	0.11	**0.78**	0.22	0.07	0.31
900	0.73	0.12	**0.80**	0.22	0.08	0.32
1000	0.74	0.15	**0.77**	0.23	0.08	0.32

Table 5.3: Precision (He-based) in AIES

in each iteration.

Among our proposals, the main reason for an algorithm to present a higher cost is due to the number of groups the corpus is divided into (see Figure 5.18). The greater the number of groups, the fewer the number of later comparisons that have to be made

Size	CRDC	DBSCAN	K-Means	RDC	TDC	Random
200	0.92	**1.00**	0.79	0.96	0.02	0.87
300	0.91	0.89	0.84	**0.96**	0.02	0.84
400	0.92	0.92	0.90	**0.96**	0.02	0.86
500	0.91	0.94	0.88	**0.96**	0.03	0.85
600	0.91	0.94	0.87	**0.96**	0.02	0.83
700	0.91	0.92	0.90	**0.96**	0.02	0.83
800	0.92	0.92	0.88	**0.96**	0.02	0.83
900	0.92	0.95	0.86	**0.96**	0.02	0.83
1000	0.92	0.93	0.89	**0.97**	0.02	0.84

Table 5.4: Recall (JS-based) in AIES

Size‘	CRDC	DBSCAN	K-Means	RDC	TDC	Random
200	0.84	**1.00**	0.65	0.96	0.02	0.82
300	0.84	**0.98**	0.76	0.95	0.02	0.78
400	0.84	**0.98**	0.79	0.94	0.02	0.79
500	0.85	0.94	0.87	**0.95**	0.02	0.78
600	0.86	**0.96**	0.80	0.95	0.02	0.76
700	0.85	**0.98**	0.80	0.95	0.02	0.76
800	0.85	**0.99**	0.81	0.95	0.02	0.76
900	0.85	**0.99**	0.75	0.95	0.02	0.77
1000	0.86	**1.00**	0.74	0.96	0.02	0.78

Table 5.5: Recall (He-based) in AIES

and, therefore, the lower the cost of the algorithm.

The behavior of the *DBSCAN* algorithm depends remarkably on the similarity metric used. We think that this may be due to the way in which both measures satisfy the triangle inequality condition, since one is based on divergence (JS) and the other on distance (He). This property, which defines $distance(a, b) \leq distance(a, c) + distance(c, b)$, is very important in the calculations that *DBSCAN* makes to discover the groups, since it only calculates the distances between near points.

Finally, in terms of ***efficiency*** (Figures 5.21, 5.22), regardless of the similarity measure used, the algorithm that yields the best performance according to the results obtained is *CRDC*. Overall, *CRDC* demonstrates a high accuracy classification and a

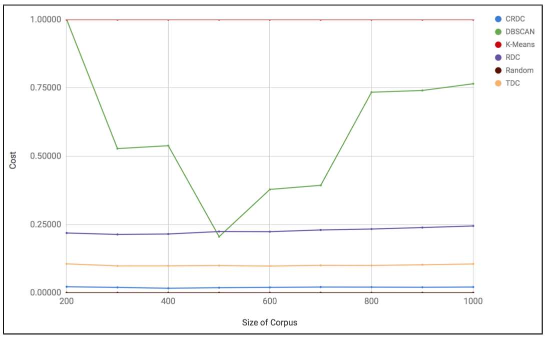


Figure 5.19: Cost (JS-based) in AIES

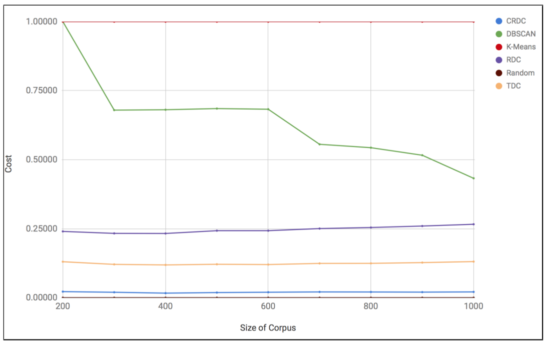


Figure 5.20: Cost (He-based) in AIES

lower cost by improving the performance offered by centroid-based or density-based approaches.

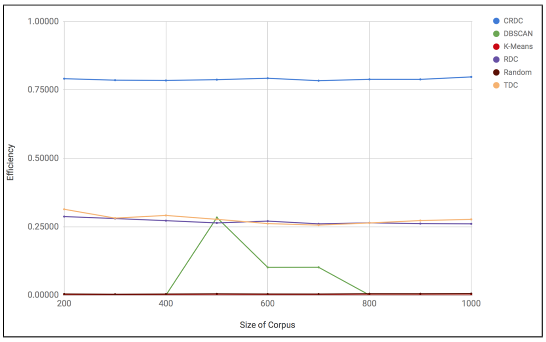


Figure 5.21: Efficiency (JS-based) in AIES

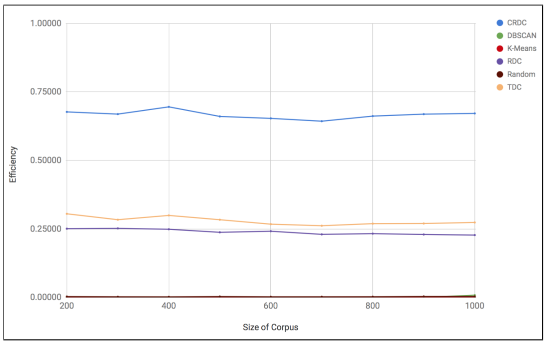


Figure 5.22: Efficiency (He-based) in AIES

We have also created a synthetic dataset, DRM (Section 5.2.2.1), composed of 1000 Dirichlet distributions with the same dimensions than topics in AIES: $k = 44$. Unlike AIES, topic distributions have been randomly generated which imply that the similarity values are not so high: $min = 0.06$, $mean = 0.18$ and $max = 0.61$. Following the same criteria than before (Section 5.2.2.2), the similarity threshold is now fixed to 0.34 (Figure 5.23). Results **in terms of *effectiveness*** (Figure 5.24) show a poor performance of the RDC and CRDC algorithms. The reason is that both are based on the fact that the highest weighted topics are shared between similar distributions. However, this condition is not satisfied when the similarity value between them is low.

To confirm this behavior, we created a third dataset (DRM2) with the same size but with only 4 dimensions (4 topics). The goal is to achieve more similar distributions

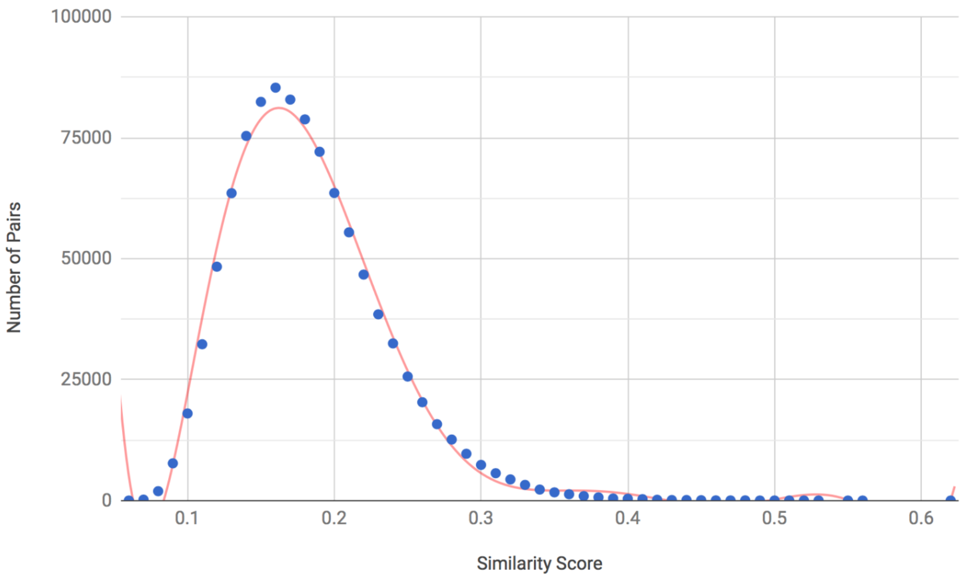


Figure 5.23: Similarity values grouped by frequency in DRM

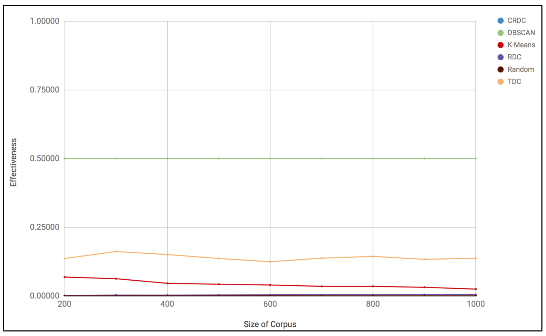


Figure 5.24: Effectiveness (JS based) in DRM

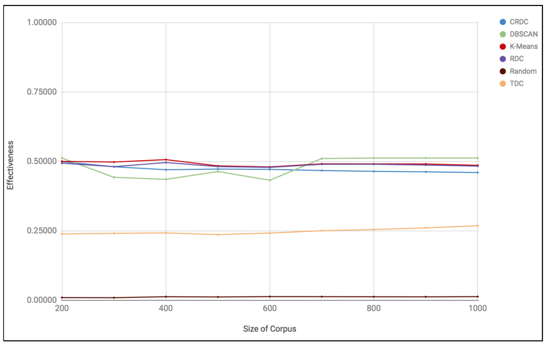


Figure 5.25: Effectiveness (JS based) in DRM2

than in DRM even though they are also randomly generated. Since the similarity values range from $min = 0.04$, $mean = 0.34$ to $max = 0.99$, the similarity threshold is now fixed to 0.66 (more details in section 5.2.2.2). The results (Figure 5.25) show an improvement in the accuracy of both the RDC and CRDC algorithms. Although scores are still not as high as for the AIES dataset, the increase compared to the DRM dataset shows that their *precision* and *recall* improve when the similarity threshold is higher. On the other hand, both the DBSCAN and TDC algorithms show similar behavior in both datasets, which means that their performance is not affected by the similarity threshold.

5.2.3 Conclusion

Processing a continuously growing collection of human generated documents requires techniques that divide the space into smaller regions containing potentially similar documents. Some algorithms in the literature tackle this problem from an unsupervised point of view, but they incur in high temporal costs and may not be suited for the domain being studied.

Three novel unsupervised clustering algorithms, *TDC*, *RDC* and *CRDC*, are described in this paper relying on the distributions inferred from a topic modeling algorithm (LDA). They are presented as a means to identify a smaller set of documents where only the similarity function has to be computed. They leverage on the particular behavior of Dirichlet distributions describing topic distributions, where the highest weighted topics have a high influence on the rest of topics. This also means that given a topic distribution, the relations between their topic weights such as order or trends between them, are more important than the density values.

Although we initially thought that using only a fixed number of topics with higher weights of a topic distribution (*RDC*), or taking into account only the trend changes between the weights of consecutive topics (*TDC*), could be enough to classify similar topic distributions, the results obtained have shown that these properties are not sufficient. Results in terms of *efficiency*, *effectiveness* and *cost* have been shown comparing the proposed algorithms with existing centroid-based and density-based clustering techniques. They reveal that obtaining the most representative topics of a topic distribution by comparing the sum of their weights with respect to the rest (*CRDC*) is a promising approach, which improves the *efficiency* obtained by other centroid-based and density-based approaches. While *K-Means* takes $O(n^k * \log n)$ and *DBSCAN* takes $O(n * \log n)$ time to classify n documents in a collection, the proposed algorithms only take linear time ($O(n)$) because they do not require any other data except their own topic distribution to assign it to a cluster.

5.3 Summary

In Section 5.1, we have analyzed the representativeness of topics to describe texts. In the particular case of scientific articles, it is concluded that the abstracts are not sufficiently representative to describe, by means of topics, the content of a paper. This behavior suggests that texts with greater vocabulary that emphasize key terms through repetition, favor topic-based representation.

Taking into account the relevance of topics to describe texts, we analyze in Section 5.2 the behavior of topic distributions to calculate distances between documents using topic models with different dimensions. By using clustering techniques at the topic level, the most representative topics of a topic distribution are identified regardless of the number of dimensions that the model has. A topic-based representation is then proposed that covers the third research objective of this thesis (R03, *define annotations based on topics that enable a semantic-aware exploration of the knowledge inside a corpus*).

A new distance metric is also proposed that takes advantage of such representation to compare documents. Its performance is analyzed by automatically clustering the JRC-Acquis corpus according to EUROVOC categories. Tables XX and XY show results with high precision and recall in unsupervised classification tasks. This new way of relating documents from their most representative topics covers the fourth research objective of this thesis (R04, *define a metric based on topic annotations that compares documents and facilitates their interpretation*).

In order to perform the experiments, both the representation based on the most relevant topics and the distance metric based on these representations have been implemented in *librAIry*. This partially covers the third and fourth technical objectives (T03, *integrate the annotation method base on topic hierarchies into the topic model service*) (T04, *create a system capable of finding similar document automatically*).

Chapter 6

Large-scale Comparisons of Topic Distributions

As we showed in Section 5.2, grouping topics by a cumulative ranking is a useful mechanism for simplifying representations based on topic distributions. Relevant topics emerge as those whose accumulated weight exceeds a threshold, after ordering all topics and starting from the top. This technique has shown a promising performance to cluster documents (Section 5.2.2.5) and suggests that *similar documents share the most relevant topics*. However, the approach still has limitations: it depends on the manual tuning of a parameter, the threshold; and it does not measure degrees of similarity since it only establishes whether or not two documents are similar. As shown in Chapter 3, we hypothesize that is possible to find relevant documents with similar topic distributions without calculating all pairwise comparisons and without discarding the notion of topics from their representation. In this chapter we introduce relevance levels between topics and present our approach to compare documents from huge collections through hierarchical representations of their topic distributions (Badenes-Olmedo et al., 2019b).

6.1 Hashing Topic Distributions

One of the greatest advantages of using Probabilistic Topic Models (PTM) in large document collections is the ability to represent documents as probability distributions over a small number of topics, thereby mapping documents into a low-dimensional latent space (the K-dimensional probability simplex, where K is the number of topics).

A document, represented as a point in this simplex, is said to have a particular topic distribution. As seen in Section 5.2, the low-dimensional feature space created by topic models could also be suitable for document similarity tasks, especially on big real-world data sets, since topic distributions are continuous and not as sparse as discrete-term feature vectors and can be explained in terms of relevance.

Hashing methods transform the data points from the original feature space into a binary-code Hamming space, where the similarities in the original space are preserved. They can learn hash functions (data-dependent) or use projections (data-independent) from the training data Wang et al. (2016). Data-independent methods unlike data-dependent ones do not need to be re-calculated when data changes, i.e. adding or removing documents to the collection. Taking large-scale scenarios into account (e.g. Document clustering, Content-based Recommendation, Duplicate Detection), this is a key feature along with the ability to infer hash codes individually (for each document) rather than on a set of documents.

Data-independent hashing methods depend on two key elements: (1) data type and (2) distance metric. For vector-type data, as introduced in Section 2.1, based on l_p distance with $p\epsilon[0, 2)$ lots of hashing methods have been proposed, such as p-stable Locality-Sensitive Hashing (LSH) (Datar et al., 2004), Leech lattice LSH (Andoni and Indyk, 2006), Spherical LSH (Terasawa and Tanaka, 2007), and Beyond LSH (Andoni et al., 2014). Based on the θ distance many methods have been developed such as Kernel LSH (Kulis and Grauman, 2012) and Hyperplane hashing (Vijayanarasimhan et al., 2014). But only few methods handle density metrics in a simplex space. A first approach transformed the He divergence into an Euclidean distance so that existing ANN techniques, such as LSH and k-d tree, could be applied (Krstovski et al., 2013). But this solution does not consider the special attributions of probability distributions, such as Non-negative and Sum-equal-one. Recently, a hashing schema (Mao et al., 2017) taking into account the symmetry has been proposed, non-negativity and triangle inequality features of the S2JSD metric for probability distributions. For set-type data, Jaccard Coefficient is the main metric used. Some examples are K-min Sketch (Li et al., 2012), Min-max hash (Ji et al., 2013), B-bit minwise hashing (Li and König, 2010) and Sim-min-hash (Zhao et al., 2013).

All of them have demonstrated efficiency in the search for similar documents, but none of them allows the search for documents (1) by thematic areas or (2) by similarity

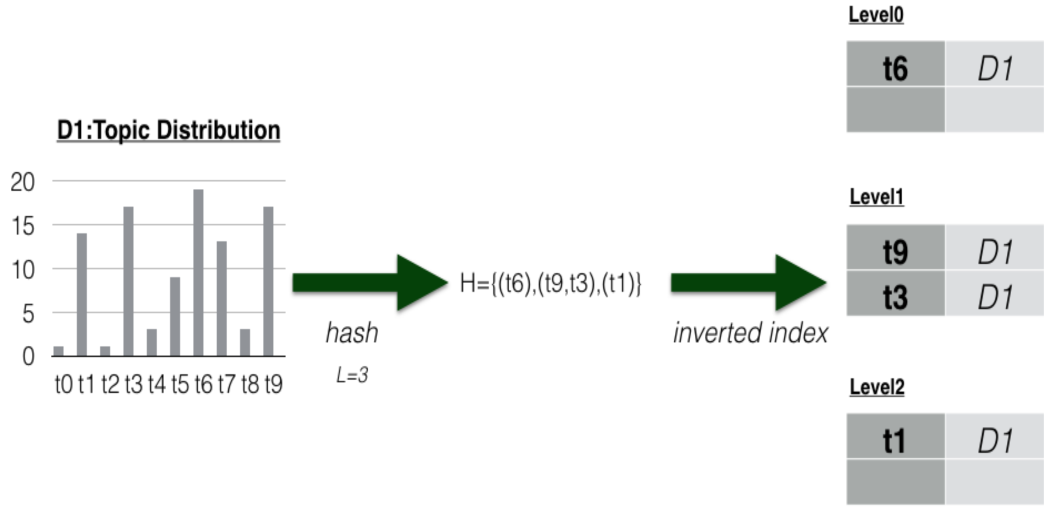


Figure 6.1: Hash method based on hierarchical set of topics from a given topic distribution

levels, nor they offer (3) an explanation about the similarity obtained beyond the vectors used to calculate it. Binary-hash codes drop a very precious information: the topic relevance.

A new hierarchical set-type data is proposed (Figure 6.1). Each level of the hierarchy indicates the importance of the topic according to its distribution. Level 0 contains the topics with the highest score. Level 1 contains the topics with highest score once the first ones have been eliminated, and so on. From a vector of components, where each of the components is the score of topic t, a vector containing set of topics is proposed, where each of the dimensions means a topic relevance. Thus, for the topic distribution $q = [0.02, 0.14, 0.02, 0.16, 0.04, 0.09, 0.19, 0.12, 0.04, 0.17]$, a hierarchical set of topics may be $h = \{(t6), (t9, t3), (t1)\}$. It means that topic $t6$ (0.19) is the most relevant, then topics $t9$ (0.17) and $t3$ (0.16) and, finally, topic $t1$ (0.14). This is just an example about the data structure that will support the different hashing strategies. In Section 6.1.3 some approaches to create hash codes based on this data structure are described.

6.1.1 Data Type

As seen in Section 2.1, a traditional approach to text representation usually requires encoding of documents into numerical vectors. Words are extracted from a corpus as feature candidates and based on a certain criterion they are assigned values to describe the documents: term-frequency, TF-IDF, information gain, and chi-square are typical measures. But this causes two main problems: huge number of dimensions and sparse distribution. The use of topics as feature space has been extended to mapping documents into low-dimensional vectors. However, as shown in Figure 2.1, the distance metrics based on probability densities vary according to the dimensions of the model and reveal the difficulty of calculating the similarity values using the vectors with the topic distributions.

Since hashing techniques can transform both vector and set-based data (Ji et al., 2013; Mao et al., 2017) into a new space where the similarity (i.e. closeness of points) in the original feature space is preserved, a new set-based data structure is proposed in this paper. It is created from clusters of topics organized by relevance levels and it aims to extend the ability of building queries with topic-based restrictions over the searching space while maintaining high level of accuracy.

The new hierarchical set-type data describes each document as a sequence of sets of topics sorted by relevance. Each level of the hierarchy expresses how important those topics are in that document. In the first level (i.e level 0) are the topics with the highest score. In the second level (i.e level 1) are the topics with the highest score once the first ones have been removed, and so on. In this work, several clustering approaches have been considered to assign topics to each level.

In a feature space created from a PTM with eight topics, for example, each data point p is described by a eight-dimensional vector with the topic distributions: $vp = [t0, t1, t2, t3, t4, t5, t6, t7]$. Then, given a point $q1 = [0.18, 0.15, 0.2, 0.05, 0.14, 0.11, 0.09, 0.08]$, the three-level hierarchical set of topics may be $h = [\{t2\}, \{t0\}, \{t1, t4\}]$. It means that $t2$ is the most relevant topic, then topic $t0$ and finally topics $t1$ and $t4$. This is just an example about the data structure that will support the hashing strategies. In section 6.1.3 some approaches to create hash codes based on this data structure are described.

Domain-specific features such as vocabulary, writing style, or speech type, have a major influence on the topic models, but not in the hashing algorithms described in this article. The methods for creating hash codes are agnostic of these particularities since they are only based on the topic distributions generated by the models.

6.1.2 Distance Metric

Since documents are described by set-type data, the proposed distance metric is based on the Jaccard coefficient. This metric computes the similarity of sets by looking at the relative size of their intersection as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

where A and B are set of topics.

More specifically, d_J is based on the Jaccard distance, which is obtained by subtracting the Jaccard coefficient J from 1:

$$d_J(A, B) = 1 - J(A, B) \tag{6.2}$$

The proposed distance measure d_H used to compare hash codes created from set of topics is the sum of the Jaccard distances d_j for each hierarchy level, i.e. for each set of topics:

$$d_H(H_1, H_2) = \sum_{l=1}^{L} \Big(d_J(H_1(x_l), H_2(x_l)) \Big) \tag{6.3}$$

where H_1 and H_2 are hash codes, $H_1(x_l)$ and $H_2(x_l)$ are the set of topics up to level l for each hash code H and L is the maximum hierarchy level. A corner case is $L = T$, where T is the number of topics in the model.

6.1.3 Hash Function

The hash function clusters topics based on relevance levels. Three approaches are proposed depending on the criteria used to group topics: threshold-based, centroid-based and density-based.

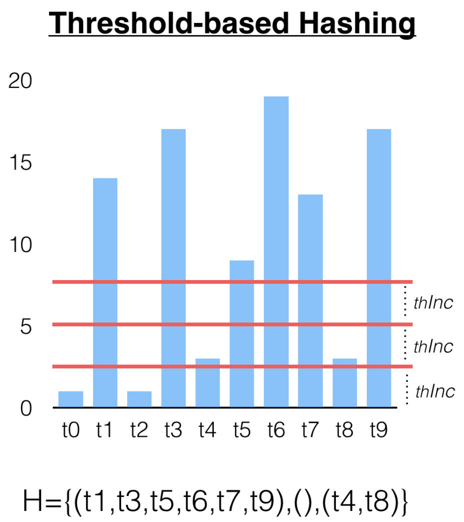


Figure 6.2: Threshold-based Hierarchical Hash (L=3)

6.1.3.1 Threshold-based Hierarchical Hashing Method

This approach is just an initial and naive way of grouping topics by threshold values into each relevance level. They can be manually defined or automatically generated by thresholds dividing the topic distributions as follows:

$$th_{inc} = \frac{1}{(L+1) \cdot T} \tag{6.4}$$

where L is the number of hierarchy levels, and T the number of topics.

If $L = 3$ and $T = 10$ for a topic distribution td defined as follows:

$$\begin{aligned} td = [&0.017, 0.141, 0.010, 0.172, 0.030, \\ &0.090, 0.199, 0.133, 0.031, 0.171] \end{aligned} \tag{6.5}$$

Then, a threshold-based hierarchical hash H_T, with an automatically created threshold defined by equation 6.4, is equals to $H_T = \{(t1, t3, t5, t6, t7, t9), (), (t4, t8)\}$ with $th_{inc} = 0.025$ (Fig 6.2).

6.1.3.2 Centroid-based Hierarchical Hashing Method

This approach assumes topic distributions can be partitioned into k clusters where each topic belongs to the cluster with the nearest mean score. It is based on the k-Means clustering algorithm, where k is obtained by adding 1 to the number of hierarchy levels.

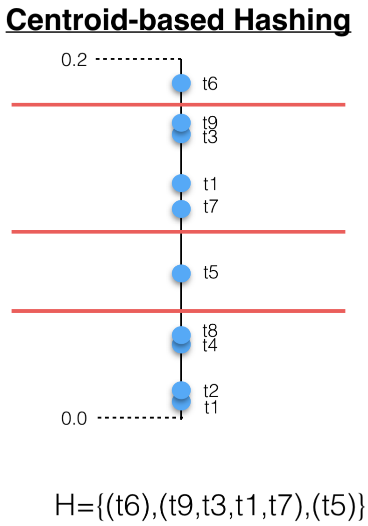


Figure 6.3: Centroid-based Hierarchical Hash (L=3)

Unlike the previous method, threshold values used to define the hierarchy levels may vary between documents, i.e. for each topic distribution, since they are calculated for each distribution separately.

Following the previous example, if $L = 3$ and $T = 10$ for a topic distribution td defined in equation 6.5, then a centroid-based hieararchical hash H_C equals to $H_C = \{(t6), (t9, t7, t3, t1), (t5)\}$ (Fig 6.3).

6.1.3.3 Density-based Hierarchical Hashing Method

This approach also considers relative hierarchical thresholds for each relevance level. Now, a topic distribution is described by points in a single dimension. In this space, topics closely packed together are grouped together. This approach does not require a fixed number of groups. It only requires a maximum distance (eps) to consider two points close and grouped together. This value can be estimated from the own distribution of topics (e.g. variance).

Following the above example, if $L = 3$ and td is the topic distribution defined in equation 6.5, then a density-based hierarchical hash H_D is equals to $H_D = \{(t6), (t9, t3), (t1)\}$ when eps equals to the variance of the topic distribution (Fig 6.4).

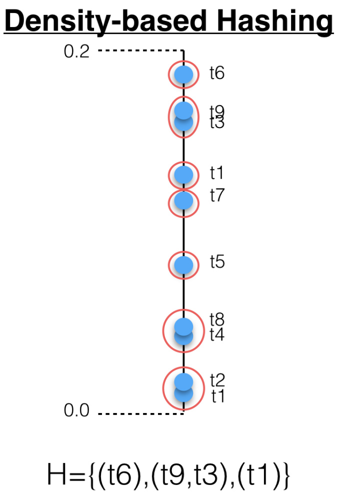


Figure 6.4: Density-based Hierarchical Hash (L=3)

6.1.4 Online-mode Hashing

Hashing methods are batch-mode learning models that require huge data for learning an optimal model and cannot handle unseen data. Recent work address online mode by learning algorithms (Huang et al., 2018) that get hashing model accommodate to each new pair of data. But these approaches require the hashing model to be updated during each round based on the new pairs of data.

Our methods rely on topic models to build hash codes. These models do not require to be updated to make inferences about data not seen during training. In this way, the proposed hashing algorithms can work on large-scale and real-time data, as the size and the novelty of the collection does not influence the annotation process.

6.2 Evaluation

As mentioned in Section 2.2.2, it is difficult to interpret the similarity score calculated by metrics in a probability space. Since all of them are based on adding the distance between each dimension of the model (eq. 2.1, 2.2 and 2.4), distributions that share a fair amount of the less representative topics may still get higher similarity values than those that share the most representative ones specially if the model has a high number of dimensions.

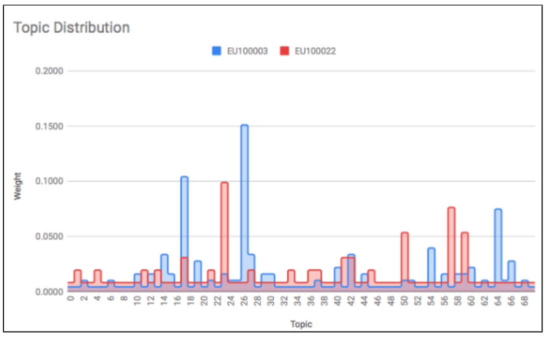


Figure 6.5: Topic Distribution of two documents with similarity score, based on JSD, equals to 0.74

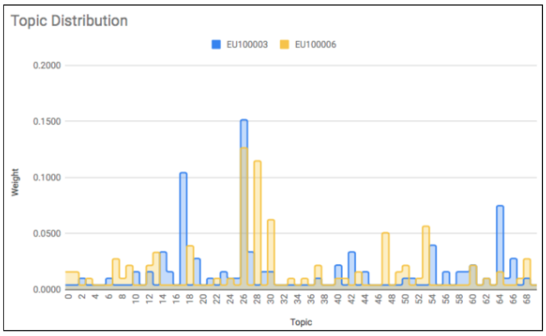


Figure 6.6: Topic Distribution of two documents with similarity score, based on JSD, equals to 0.71

Figures 6.5 and 6.6 show overlapped topic distributions of two pairs of documents. In the first case (fig 6.5), none of the most representative topics of each document is shared between them. However, the similarity score calculated from divergence-based metrics (eq 2.2) is higher than in the second case (fig 6.6), where the most representative topic is shared (topic 26). This behavior is due to the sum of the distances between the less representative topics (i.e. topics with a low weight value) being greater than the sum of the distances between the most representative ones (i.e. topic with a high weight value). In high-dimensional models, that sum may be more representative than the one obtained with the most relevant topics, which are fewer in number than the less relevant ones.

The following experiments aim to validate that ***hash codes based on hierarchical set of topics not only make it possible to search for similar documents with high accuracy, but also to extend queries with new restrictions and to offer information that helps explaining why two documents are similar***.

6.2.1 Datasets and Evaluation Metrics

Three datasets (Badenes-Olmedo et al., 2019c) are used to validate the proposed approach. The OPEN-RESEARCH[70] dataset consist of 500k research papers in Computer Science, Neuroscience, and Biomedical randomly selected from the Open Research Corpus (Waleed et al., 2018). The CORDIS[71] dataset contains 100k documents describing

[70]https://labs.semanticscholar.org/corpus/
[71]https://data.europa.eu/euodp/data/dataset/cordisref-data

research and innovation projects funded by the European Union under a framework programme since 1990. The PATENTS dataset consists of 1M patents randomly selected from the USPTO[72] collection. For each dataset, documents are mapped to two latent topic spaces with different dimensions using LDA. We perform parameter estimation using collapsed Gibbs sampling for LDA (Griffiths and Steyvers, 2004) from our librAIry framework. The number of topics varies to study their influence on the performance of the algorithm (i.e. CORDIS-70 indicates a latent space created with 70 topics).

Experiments use JS divergence as an information-theoretically motivated metric in the probabilistic space created by topic models. Since it is a smoothed and symmetric alternative to the KL divergence, which is a standard measure for comparing distributions (Cha, 2007), it has been extensively used as state-of-the-art metric over topic distributions in literature (Aletras et al., 2017; Mao et al., 2017; Towne et al., 2016). Our upper bound is created from the brute-force comparison of the reference documents with all documents in the collection to obtain the list of similar documents.

In this scenario the goal is to minimize the accuracy loss introduced by hashing algorithms. Since this is a large-scale problem and an accuracy-oriented task, recall is not a good measure to be considered and precision is only relevant for sets much smaller than the total size of data (between 3-5 candidates).

All the experimental results are averaged over random training/set partitions. For each topic space, 100 documents are selected as references, and the remaining documents as search space. As noted above, only p@5 will be used to report the results of the experiments.

6.2.2 Retrieving Similar Documents

It is challenging to create an exhaustive gold standard, given the significant amount of human labour that is required to get a comprehensive view of the subjects being covered in it. In order to overcome this problem, the list of similar documents to a given one is obtained after comparing the document with all the documents of the repository and sorting the result. We have observed that different distance functions perform similarly in this scenario (Fig. 2.1), so we have decided to use only the JS divergence (eq. 2.2) in our experiments.

[72] https://www.uspto.gov/learning-and-resources/ip-policy/economic-research/research-datasets

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.22	0.20	**0.86**	1.00	0.66	0.80
3	0.23	0.20	**0.87**	1.00	0.81	1.00
4	0.27	0.20	**0.89**	1.00	0.86	1.00
5	0.27	0.20	**0.92**	1.00	0.89	1.00
6	0.27	0.20	**0.94**	1.00	0.92	1.00

OPEN-RES-100 (p@5)

Table 6.1: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Open Research dataset using a model with 100 topics. LEVEL column indicates the number of hierarchies used.

Only the top N documents obtained from this method are used as reference set to measure the performance of the algorithms proposed in this paper. The value of N is equals to 0.5% of the corpus size (i.e. if the corpus size is equal to 1000 elements, only the top 5 most similar documents are considered relevant for a given document). This value has been considered after reviewing datasets used in similar experiments (Krstovski et al., 2013; Mao et al., 2017). In those experiments, the reference data is obtained from existing categories, and the minimum average between corpus size and categorized documents is around 0.5%.

Once the reference list of documents similar to a given one is defined, the most similar documents through the proposed methods (i.e. threshold-based hierarchical hashing method (thhm), centroid-based hierarchical hashing method (chhm) and density-based hierarchical hashing method (dhhm)) are also obtained. An inverted index has been implemented by using Apache Lucene[73] as document repository. The source code of both the algorithms and tests is publicly available (Badenes-Olmedo et al., 2019c).

Let's look at an example to better understand the procedure. We want to measure the accuracy and data size ratio used to identify the top5 similar documents to a new document d_1 from a corpus of 1000 documents . The similarity between $d1$ and all the documents in the corpus is calculated based on JS divergence. The top50 (0.5%) documents with the highest values will be the set of documents considered as similar to d_1. As we are going to use an ANN-based approach, we need the hash expressions of all documents to measure similarity. The data structure proposed in this work is

[73]http://lucene.apache.org

OPEN-RES-500 (p@5)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.23	0.20	**0.76**	0.80	0.67	0.80
3	0.24	0.20	**0.80**	1.00	0.71	0.80
4	0.25	0.20	**0.83**	1.00	0.74	0.80
5	0.25	0.20	**0.86**	1.00	0.81	1.00
6	0.24	0.20	**0.89**	1.00	0.86	1.00

Table 6.2: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Open Research dataset using a model with 500 topics. LEVEL column indicates the number of hierarchies used.

CORDIS-70 (p@5)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.18	0.20	**0.92**	1.00	0.66	0.70
3	0.20	0.20	**0.92**	1.00	0.80	0.80
4	0.22	0.20	**0.94**	1.00	0.86	1.00
5	0.23	0.20	**0.91**	1.00	0.89	1.00
6	0.19	0.20	**0.92**	1.00	0.91	1.00

Table 6.3: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on CORDIS dataset using a model with 70 topics. LEVEL column indicates the number of hierarchies used.

CORDIS-150 (p@5)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.19	0.20	**0.88**	1.00	0.78	0.80
3	0.19	0.20	**0.92**	1.00	0.80	1.00
4	0.25	0.20	**0.91**	1.00	0.82	1.00
5	0.25	0.20	**0.91**	1.00	0.83	1.00
6	0.27	0.20	**0.91**	1.00	0.86	1.00

Table 6.4: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on CORDIS dataset using a model with 150 topics. LEVEL column indicates the number of hierarchies used.

a hierarchy of sets of topics, so that the most similar documents are those that share most of the topics at the highest levels of the hierarchy.

The representational model for this example only considers 8 topics, that is, a document is described by a vector with 8 dimensions where each dimension corresponds to a topic (i.e $[t0, t1, t2, t3, t4, t5, t6, t7]$) and its value will be the weight of that topic in the document, for example $d_1 = [0.18, 0.15, 0.2, 0.05, 0.14, 0.11, 0.09, 0.08]$. The hierarchy level ($L$) will be equal to 2, i.e. the hash expression has two hierarchical sets of topics: $h = \{h_0, h_1\}$.

According to methods described at Section 6.1.3, there are 3 ways to create the hierarchical hash codes for documents:

1. threshold-based (*thhm*): 2 thresholds are defined as described in section 6.1.3.1, for example 0.15 and 0.1 . h_0 includes the topics with a weight greater than 0.15, and h_1 the remaining topics with a weight greater than 0.1. Then $h_0 = \{t0, t1, t2\}$ and $h_1 = \{t4, t5\}$. Based on the hash expression $h = \{(t0, t1, t2), (t4, t5)\}$, the documents that share more topics in those levels (i.e $h0 = (t0 \text{ OR } t1 \text{ OR } t2)$, $h1 = (t4 \text{ OR } t5)$) or in other levels but with less relevance are ordered. Since there are many topics in the expression, potentially many documents are similar when sharing at least one of them. This increases the data ratio. Accuracy is also affected, as the algorithm is not able to bring under the same bucket similar documents. In short, the hash expression is not representative of the document, for the given exploratory task.

2. centroid-based (*chhm*): sets of topics are created using a clustering algorithm based on centroids as described in section 6.1.3.2. The cardinalities of the hierarchical groups are generally more uniform with this method. Since $k = L + 1 = 3$ in this example, $h_0 = \{t0, t2\}$ and $h_1 = \{t1, t4\}$. The number of representative topics at each level of the hierarchy is usually lower, and this causes the data ratio used to discover similar documents to decrease as well. This approach increases the precision because now the hierarchy is more selective to distinguish similar documents. However, the size of region of similar candidates is still high.

3. density-based (*dhhm*): now the clustering algorithm is based on how dense certain regions in the topic relevance dimensions are as described in section 6.1.3.3. It can group topics that have unbalanced distributions and, therefore, generates

PATENTS-250 (p@5)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.03	0.00	**0.71**	0.80	0.67	0.80
3	0.08	0.00	**0.91**	1.00	0.90	1.00
4	0.11	0.00	**0.95**	1.00	**0.95**	1.00
5	0.12	0.00	0.95	1.00	**0.96**	1.00
6	0.11	0.00	**0.97**	1.00	**0.97**	1.00

Table 6.5: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Patents dataset using a model with 250 topics. LEVEL column indicates the number of hierarchies used.

PATENTS-750 (p@5)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	0.02	0.00	**0.77**	0.80	0.76	0.80
3	0.04	0.00	0.94	1.00	**0.95**	1.00
4	0.06	0.00	**0.97**	1.00	**0.97**	1.00
5	0.08	0.00	**0.97**	1.00	**0.97**	1.00
6	0.06	0.00	**0.97**	1.00	**0.97**	1.00

Table 6.6: Precision at 5 (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Patents dataset using a model with 750 topics. LEVEL column indicates the number of hierarchies used.

more discriminating hash expressions than with the previous algorithm. In the example, we would have a hash expression like this: $h_0 = \{t2\}$ and $h_1 = \{t0\}$. This significantly reduces the data ratio used to discover similar documents and does not excessively penalize accuracy. Obviously, increasing L (i.e. number of hierarchies) increases precision, but with $L > 3$ that gain is not so significant.

As it can be seen in tables 6.1 to 6.6, the mean and median of precision are calculated to compare the performance of the methods. In this assessment environment, the variance is not robust-enough because score values don't follow a normal distribution. We consider the result obtained as significant, based on the fact that mean and median values are fairly close. The centroid-based method (chhm) and the density-based method (dhhm) show a similar behaviour to the one offered by the use of brute force by means of JS divergence.

In terms of efficiency, we consider the times to compare pairs of topic distributions

OPEN-RES-100 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	99.8	99.9	45.2	45.9	**4.9**	2.5
3	99.9	99.9	74.4	77.6	**13.4**	10.7
4	99.9	99.9	87.4	90.2	**27.2**	22.8
5	99.9	99.9	95.4	96.3	**49.9**	42.6
6	99.9	99.9	97.9	98.7	**72.2**	65.8

Table 6.7: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Open Research dataset and 100 topics.

OPEN-RES-500 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	95.9	96.3	22.2	22.1	**1.4**	0.3
3	99.1	99.2	43.9	43.7	**5.1**	4.1
4	99.6	99.6	57.1	57.3	**11.7**	10.3
5	99.6	99.6	70.7	70.7	**28.8**	22.0
6	99.9	99.9	81.5	80.6	**50.3**	40.1

Table 6.8: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Open Research dataset and 500 topics.

CORDIS-70 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	99.9	99.9	51.3	56.3	**5.1**	5.0
3	99.9	99.9	84.8	89.5	**10.5**	10.6
4	99.9	99.9	96.1	97.6	**20.8**	19.5
5	99.9	99.9	98.9	99.4	**35.0**	32.7
6	99.9	99.9	99.7	99.8	**53.1**	51.2

Table 6.9: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on CORDIS dataset and 70 topics.

CORDIS-150 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	99.9	99.9	40.9	41.2	**3.1**	2.9
3	99.9	99.9	75.3	76.7	**6.2**	6.1
4	99.9	99.9	90.0	92.1	**12.1**	11.8
5	99.9	99.9	96.4	96.9	**21.6**	20.6
6	99.9	99.9	98.1	98.9	**36.5**	33.9

Table 6.10: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on CORDIS dataset and 150 topics.

constant, and we focus on the number of comparisons needed. Thus, algorithms with larger candidate spaces will be less efficient than others when the accuracy in both is the same. Tables 6.7-6.12 show the percentage of the corpus used by each of the algorithms to discover similar documents. Tables 6.1-6.6 show the accuracy of each algorithm for each of these scenarios. Density-based algorithm (dhhm) shows better balance between accuracy and volume of information (efficiency). It uses smaller samples (i.e lower ratio size) than others in all tests and even when it only uses a subset that is a 6.2% (Table 6.10) of the entire corpus, it obtains an accuracy of 0.808 (Table 6.4).

The precision achieved by the algorithm based on density (dhhm), which is much more restrictive than the others, suggests that few topics are required to represent a document in order to obtain similar ones. In addition, the number of topics does not seem to influence the performance of the algorithms, since their precision values are similar among the datasets of the same corpus. This shows that hashing methods based on hierarchical set of topics are robust to models with different dimensions.

PATENTS-250 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	99.9	99.9	43.2	32.7	**35.1**	23.0
3	99.9	100.0	82.4	100.0	**78.2**	100.0
4	99.9	100.0	96.5	100.0	**95.1**	100.0
5	99.9	99.9	99.2	100.0	**98.9**	100.0
6	100.0	100.0	99.8	100.0	**99.7**	100.0

Table 6.11: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Patents dataset and 250 topics.

PATENTS-750 (data-ratio)

LEVEL	THHM		CHHM		DHHM	
	mean	*median*	*mean*	*median*	*mean*	*median*
2	99.9	100.0	35.2	23.6	**31.8**	19.9
3	99.9	99.9	81.4	99.8	**79.6**	98.8
4	99.9	99.9	96.5	99.9	**95.5**	99.5
5	97.7	96.6	99.0	99.9	**98.6**	99.7
6	99.1	98.6	99.7	99.9	**99.5**	99.8

Table 6.12: Data size ratio used (*mean* and *median*) of threshold-based (THHM), centroid-based (CHHM) and density-based (DHHM) hierarchical hashing methods on Patents dataset and 750 topics.

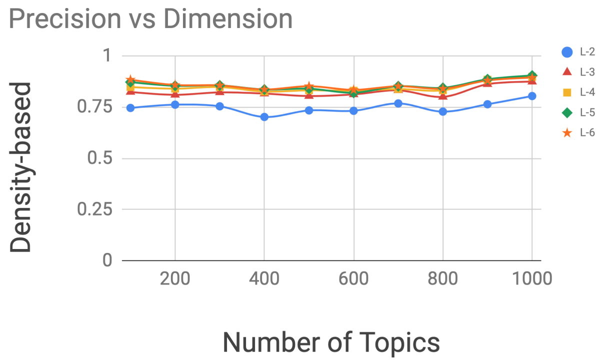


Figure 6.7: Precision at 5 (*mean*) of threshold-based hashing method when number of topics varies in CORDIS dataset.

The behavior of the algorithms have also been analyzed when the number of topics in the model varies. Models with 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 topics were created from the CORDIS corpus. For each model, the p@5 of the hashing methods is calculated taking into account the hierarchy levels: 2, 3, 4, 5 and 6. Figures 6.7 to 6.9 show the results obtained for each algorithm. It can be seen how the performance, i.e precision, of each of the algorithms is not influenced by the dimensions of the model.

6.2.3 Exploration

In a certain domain, we may want to retrieve similar documents to one given. For example, searching for articles in the Biomedical domain that are similar to an article about Semantic Web. In terms of topics this kind of search requires to narrow down the initial search space to a subset with only documents that contain the topics that better describe the queried domain.

Existing hashing techniques based on a binary-code Hamming space do not allow to customize the search query beyond the reference document itself. However, the algorithms proposed in this work allow adding new restrictions to the initial query based on the reference document, since they use a hierarchy of set of topics as hash codes.

Through the following example we describe the workflow to enable such retrieval operations. For simplicity we consider hash expressions with only two hierarchy levels. The reference document d_1 has the following hash expression: $h = \{h_0, h_1\} =$

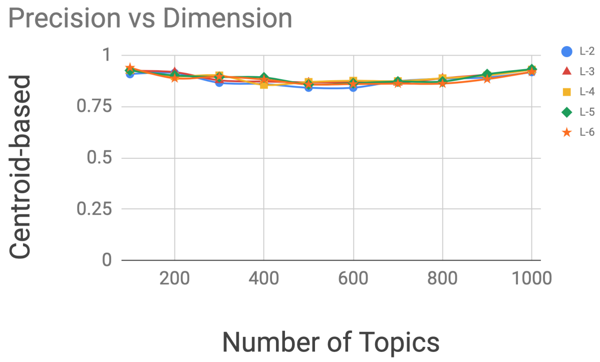


Figure 6.8: Precision at 5 (*mean*) of centroid-based hashing method when number of topics varies in CORDIS dataset.

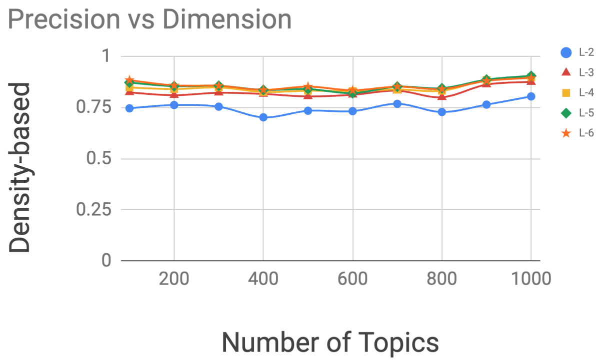


Figure 6.9: Precision at 5 (*mean*) of density-based hashing method when number of topics varies in CORDIS dataset.

$\{(t10), (t18)\}$.

The first query, $Q1$, searches for documents similar to the reference document d_1 among all documents in the corpus. One of the ways to formalise this query looks like this: $Q1 = h_0 : t10\hat{}100$ or $h_0 : t18\hat{}50$ or $h_1 : t10\hat{}50$ or $h_1 : t18\hat{}100$. It sets a maximum boost (100) when the same restrictions as the reference document ($t10$ in h_0 and $t18$ in h_1) are fulfilled, and a lower boost (50) for the others ($t18$ in h_0 and $t10$ in h_1). In the specific case of applying this query to the CORDIS dataset, we observed that most of the retrieved documents included topic t18 (fig 6.10).

But if we were only interested in similar documents to d_1 that have topic $t10$, we could restrict the previous query $Q1$ to express this condition in the following way: $Q2 = (h_0 : t10\hat{}100$ or $h_1 : t10\hat{}50)$ and $(h_1 : t10\hat{}50$ or $h_1 : t18\hat{}100)$. The result obtained by $Q2$ (fig 6.10) shows that the condition has been considered since there is

OPEN-RESEARCH-100			
hash	**q1**	**q2**	**ratio**
thhm	499,755	160,660	67.8
chhm	356,111	1,976	99.44
dhhm	49,068	766	98.43

Table 6.13: Number of documents similar to a given one (q1) and also in a specific domain (q2) for threshold-based (thhm), centroid-based (chhm) and density-based (dhhm) hierarchical hashing methods.

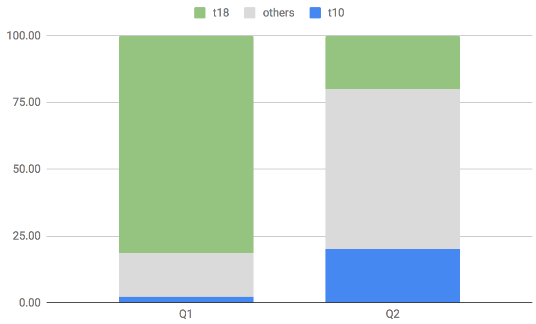


Figure 6.10: Most relevant topics in similar documents from using a document as query (Q1) and setting topic t10 as mandatory (Q2).

a balance between topics $t10$ and $t18$ among the documents similar to d_1.

This type of restrictions based on the semantics offered by topics in the hash expression get enabled thanks to the methods proposed in this work.

6.3 Summary

The usefulness of topics created by probabilistic models when exploring document collections on large-scale has been widely studied in the literature. Each document in the corpus is described by probability distributions that measure the presence of those topics in their content. These vectors can also be used to measure the similarity between documents by using metrics such as Jensen-Shannon divergence (see Section 2.1). But with large amounts of items in the collection, discovering the entire set of nearest neighbors to a given document would be infeasible. Due to the low storage cost and fast retrieval speed, hashing is one of the popular solutions for approximate nearest neighbors. However, existing hashing methods for probability distributions only focus on the efficiency of searches from a given document, without handling complex queries or offering hints about why one document is considered more similar than another.

In this chapter we have introduced a new data structure to represent hash codes, based on topic hierarchies created from the topic distributions. This approach has proven to obtain high-precision results and can accommodate additional query restriction. In doing so, we have showed a new way to annotate documents by topic inferences made from topic models. This addresses the technical objective of this thesis T03 (*integrate the annotation method based on topic hierarchies into the topic model service*).

This way of encoding documents can also help to understand why two documents are similar, based on the intersection of topics at hierarchies of relevance. We have proposed a method to compare and organize huge document collections based on similar topic-based annotations, thus addressing the research objective R05 (*define nearest-neighbor techniques to organize documents in regions with similar topic hierarchies*).

In addition, we have implemented the technique to compare documents in our librAIry framework, encouraging the achievement of the T04 technical objective (create a system capable of finding similar documents automatically).

Chapter 7

Cross-lingual Document Similarity

As stated in Chapter 3, the last of our hypotheses aims to determine whether documents in different languages can be related without having to translate them, by using language agnostic concepts from their main topics (H1.4). In particular, our goal is to find abstractions that capture the content of documents, independently from the language used, in order to draw relations between them. A way is by creating multilingual topics from comparable or parallel corpora and relating documents from their topic distributions (See Section 2.2.4). A parallel corpora contains sentence-aligned documents (e.g. Europarl[74] corpora), and a comparable corpora contains theme-aligned documents (e.g. Wikipedia[75] articles). Other types of abstractions may be obtained using multilingual dictionaries to translate documents in a common language from which they can be related.

But these approaches based on aligned corpora or document translations require prior knowledge. Connections at document-level (by parallel or comparable corpora) or at word-level (by dictionaries) are necessary to create topic models that represent documents in a common, language-independent space. In this way, the pre-established language relations condition the creation of the topics (supervised method), instead of being inferred from the topics themselves as a posteriori knowledge (non-supervised method). We propose a completely unsupervised way of building cross-lingual topic

[74]https://ec.europa.eu/jrc/en/language-technologies/dcep
[75]https://www.wikipedia.org/

models that uses sets of cognitive synonyms (*synsets*) to establish relations between language-specific topics once the model is created and does not require parallel or comparable data for training (Badenes-Olmedo et al., 2019a,b). These models can be used for large-scale multi-lingual document classification and information retrieval tasks.

In Section 7.1, we propose language independent conceptual abstractions for topic models. Topics are no longer described by words, but by concepts. Cross-lingual models are then created from the synset-based representations of topics and documents, described with these models, are related to perform evaluations. One analysis consists of cross-lingual document classification, while the other one performs cross-lingual information retrieval.

7.1 Synset-based Representational Space

Each topic is annotated with a list of synset (Bond and Foster, 2013) retrieved from WordNet[76](Miller, 1995) based on its top$n$ words (Fig 7.1). Word by word are queried in WordNet to retrieve its synsets. The final set of synsets for a topic is the union of the synsets from the individual top-words of a topics. Based on empirical evidence from different executions of the algorithm, n=5 is the configuration that offered the best performance in our tests. Let's look at an example to clarify how it works. Given the topics of Table 7.1, the EN-Topic (*"communications systems"*) is annotated with the following synset list: *radio.a.01, radio.v.01, radio.n.03, radio.n.01, radio_receiver.n.01, equipment.n.01, network.n.02, network.n.04, network.v.01, network.n.05, network.n.01, net.n.06, communication.n.02, communication.n.03, communication.n.01, regulative.s.01.* The list of synset for the ES-Topic (*"sistema de comunicación"*) is: *kit.n.02,team.n.01, equipment.n.01, net.n.02, net.n.05, network.n.05, web.n.06, network.n.01, web.n.02, communication.n.02, communication.n.01, announcement.n.02, spectrum.n.02, spectrum.n.01, creep.n.01, ghost.n.01, apparition.n.02, electromagnetic.a.01.* And the list for FR-Topic (*"systeme de communication"*) is: *access.n.02, approach.n.07, approach.n.02, access.n.06, access.n.03, access.n.05, assault.n.03, bout.n.02, approach.n.01, entree.n.02, entry.n.01, entrance.n.01, entry.n.03, admission.n.01, submission.n.01, introduction.n.01.*

[76]https://wordnet.princeton.edu/

The librAIry NLP service[77] was used to identify the list of synsets from a topic description based on top words. It is based on the Open Multilingual WordNet[78] (Bond and Paik, 2012).

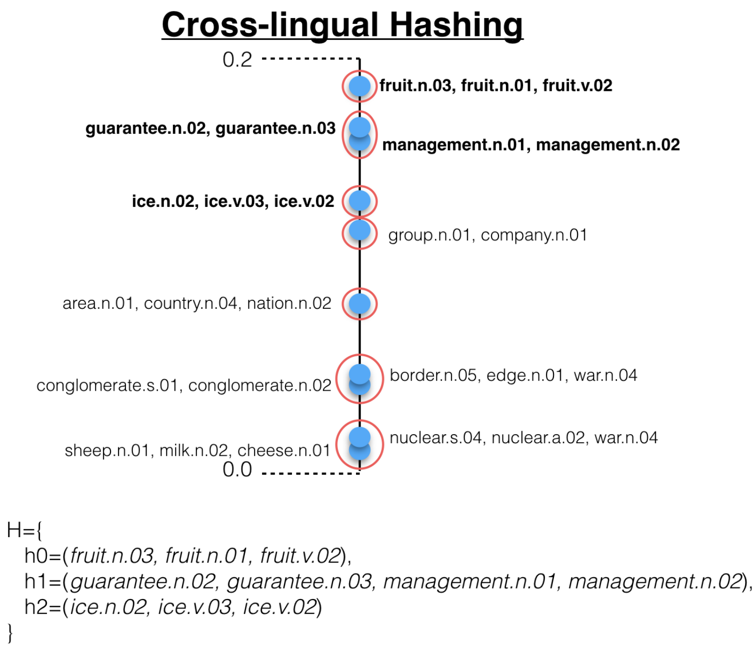


Figure 7.1: Cross-lingual hash-expression (H) of a document based on WordNet-synset annotations created from the top words of each topic distribution. The most relevant topics are grouped according to their importance in three levels (h0, h1 and h2)

7.1.1 Document representation

Documents (i.e seen as data points in the generated topic-based space) are transformed from the original feature space based on mono-lingual topic distributions into a hierarchical-code space, so that similar data points share relevant cross-lingual concepts. Since topic models create latent themes from word co-occurrence statistics in a corpus, a cross-lingual concept specifies the knowledge about the word-word relations it contains for each language. This abstraction can be extended to cover the knowledge derived from sets of topics. The topics are obtained via state-of-the art

[77]http://librairy.linkeddata.es/nlp
[78]http://compling.hss.ntu.edu.sg/omw/

methods, collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for LDA, and hierarchically divided into groups with different degrees of semantic specificity in a document (See Chapter 6). Documents represented as a weighted mixture of latent topics (per-document topic distributions) are then annotated in these feature spaces with the relation between topics inside each hierarchy level. Regardless of their language, they are then described by cross-lingual concepts (based on WordNet-synset annotations) and hash codes are calculated to summarize their content. The hash expression sets a 3-level hierarchy of cross-lingual concepts. Topics with similar presence in a document (i.e. relevance) are grouped together in the same hierarchical level (Fig 7.1). Each level of the hierarchy indicates the importance of the topic according to its distribution. *Level 0* describes the topics with the highest score. *Level 1* describes the topics with highest score once the first ones have been eliminated, and so on. Documents are described by vectors containing set of topics (i.e. set of synsets), where each dimension means a topic relevance. Given a document d with a topic distribution $q = [t0 = 0.28, t1 = 0.05, t2 = 0.44, t3 = 0.23]$, the hash expression may be $H_d = (ts2), (ts0, ts3), (ts1)$. It means that topic $t2$ described by the synset $ts2$ is the most relevant (i.e 0.44 score), then topics $t0$ and $t3$ described by synsets $ts0$ and $ts3$ (i.e 0.28 and 0.23 scores) and, finally, topic $t1$ described by synset $ts1$ (i.e 0.05).

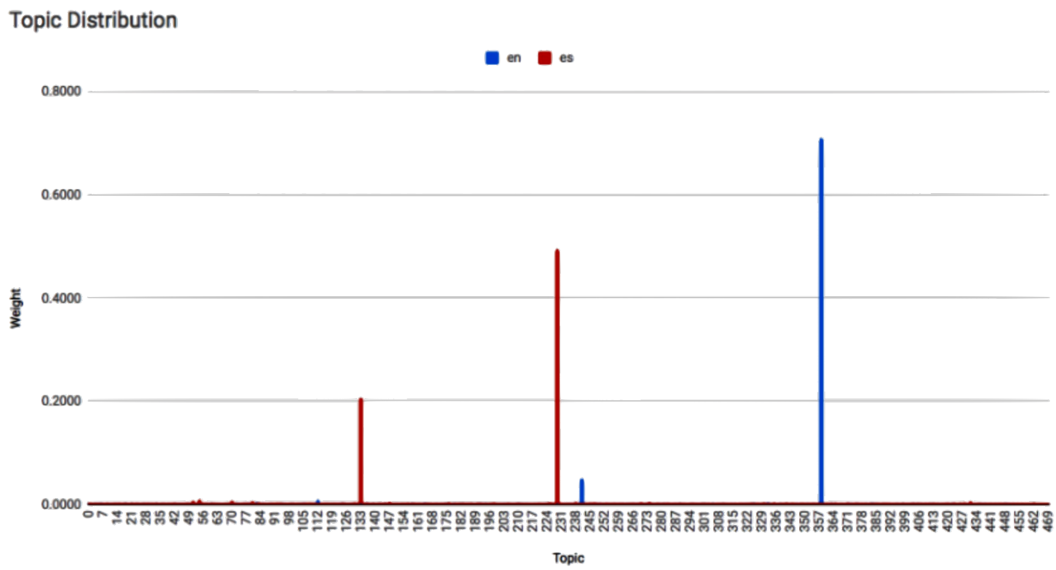


Figure 7.2: topic distributions from the same document in English ($h_{EN} = \{(t3062), (t335), (t8278)\}$) and Spanish ($h_{ES} = \{(t335), (t4060), (t5769)\}$).

7.1.2 Similarity metric

In this workspace based on hierarchical representations of topics we use the distance metric proposed in Section 6.1.2 based on the Jaccard coefficient. This metric is mainly used for set-type data (Ji et al., 2013; Li and König, 2010; Li et al., 2012; Zhao et al., 2013) and computes the similarity of sets by looking at the relative size of their intersection (See Eq. 6.1). Thus, it allows us to measure the intersection of cross-lingual topics described by hierarchical hash-sets:

$$d_H(H_A, H_B) = \sum_{l=1}^{L} \Big(d_J(H_A(h_l), H_B(h_l)) \Big) = \sum_{l=1}^{L} \Big(1 - \frac{H_A(h_l) \cap H_B(h_l)}{H_A(h_l) \cup H_B(h_l)} \Big) \tag{7.1}$$

where H_A and H_B are hash codes, $H_A(h_l)$ and $H_B(h_l)$ are the set of topics up to level l for each hash code H, and L is the maximum hierarchy level. A corner case is $L = T$, where T is the number of topics in the model.

7.1.3 Cross-lingual Models

Our approach considers that cross-lingual models can be built from non-parallel or even non-comparable collections of multilingual documents. It first creates a probabilistic topic model for each language separately, and then annotates the topics with cross-lingual labels (Fig 7.3). In the same way, the topic distribution of documents expressed through weighted vectors are first transformed into hierarchies of topics according to their relevance. And then documents are described by a 3-level hierarchy of cross-lingual concepts.

In order to be able to compare the performance of our unsupervised algorithm with a semi-supervised algorithm (MuPTM-based) it is necessary to use theme-aligned corpora that map topics across languages. We used the JRC-Acquis[79] corpora (Steinberger et al., 2006). It is a collection of legislative texts written in 23 languages, although we only use English, Spanish, French, Italian and Portuguese editions for the tests. Most texts have been manually classified into subject domains according to the EUROVOC[80] thesaurus (Eurovoc, 1995), which exists in one-to-one translations into approximately twenty languages and distinguishes about 6,000 hierarchically organised descriptors

[79] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[80] http://eurovoc.europa.eu/

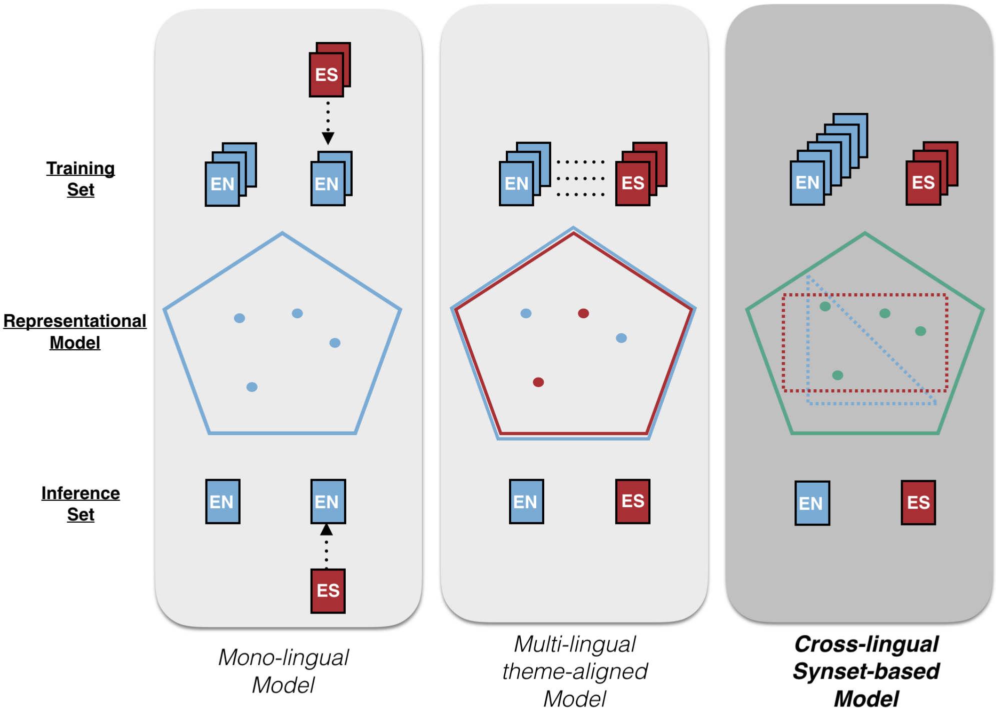


Figure 7.3: Graphical representation of the model that relies on the latent layer of cross-lingual topics obtained by LDA and hash functions through hierarchies of synsets. Mono-lingual approaches force to translate the documents to the same language to represent them in a unique feature space. Multi-language approaches require previously aligned topics from different languages so that documents can be represented in an equivalent feature space. Cross-lingual Synset-based approach creates a new space by combining the feature spaces of each language (i.e synsets from topn topic words). Documents are then represented in this unique space.

(subject domains). More than 20k documents were used for each language-specific model, a total of 112,569 texts are included in the training-test package, which is publicly available[81] for reuse.

The JRC-Acquis corpus is annotated with EUROVOC categories. These categories are shared among languages and will serve as support for building the topic models. Moreover, the topic independence assumption (Blei et al., 2003) of LDA models should be also satisfied, so the categories must first be moved to their base concepts and therefore disjointed categories. The EUROVOC taxonomy has 7,193 concepts/labels from 21 domain areas such as politics, international relations, european union, law,

[81] http://librairy.linkeddata.es/data/jrc/select?q=*:*

economics, etc. There are 4,904 reciprocal hierarchical relationships (no polyhierarchy) and 6,992 reciprocal associative relationships. Using hierarchical relations, we identified the root concepts from which all other categories derive. The initial 7,193 labels were then reduced to 452 labels, which are independent (topic independence assumption from LDA is satisfied), and can be used to train the topic models.

EN-Topic 3 *"communications systems"*	**ES-Topic 3** *"sistema de comunicación"*	**FR-Topic 26** *"systeme de communication"*	**PT-Topic 10** *"meios de comunicação"*	**IT-Topic 3** *"strutture di comunicazione"*
radio	equipo	communications	rede	rete
equipment	red	reseaux	comunicação	comunicazione
network	comunicación	electroniques	electronico	apparecchiatura
communication	espectro	acces	acesso	radio
regulatory	electromagnético	telecommunications	utilizador	regolamentazione
spectrum	electrónico	service	operador	spettro
electronic	reglamentación	universel	regulador	elettronico
access	banda	reglamentaires	universal	armonizzare
standard	etsir	nationales	garantir	mobile
mobile	compatibilidad	fourniture	regulamentar	banda

Table 7.1: Randonly selected theme-aligned topics described by top 10 words based on EUROVOC annotations from JRC-Acquis dataset

A pre-processing of the documents was required to clean texts and to build a suitable data set for the model. We assume that terms with high frequency are not specific to a particular topic, so words present in more than 90% of the corpus are considered stopwords and removed from the model. Also, rare terms that occur infrequently are considered not representative of a single topic since they do not appear enough to infer that it is salient for a topic. Then, words present in less than 0.5% of the corpus are also removed from the model. Lemmatized expressions of names, verbs and adjectives were used to create the bag-of-words, and documents with less than 100 characters were discarded since LDA has proven to has lower performance with these type of texts (Cheng et al., 2014).

Then, we set the number of topics $K = 500$ (several configurations were evaluated, but this was the closest to the performance obtained with the supervised model based on categories). We run the Gibbs samplers for 1000 training iterations on LDA from our librAIry framework. The Dirichlet priors $\alpha = 0.1$ and $\beta = 0.01$ were set following

(Hu et al., 2014). Once the word distributions for each topic is available, the list of synsets related with the top5 words for each topic are identified (this number is set to offer better performance after trying several alternatives). Finally, the 3-level hierarchy of topics per document is replaced by a 3-level hierarchy of synsets. Probabilistic topic models in Spanish[82], English [83], French[84], Italian[85] and Portuguese[86] were created independently without previously establishing any type of alignment between their topics.

In order to compare the performance of this non-supervised approach with approaches based on aligned topics, we need to use a variant of LDA to force the correspondence between the 452 root categories identified in the EUROVOC thesaurus and the latent topics of the model. Thus, LabeledLDA (Ramage et al., 2009), a supervised version of LDA, was used to perform parameter estimation. Theme-aligned probabilistic topic models in Spanish[87], English [88], French[89], Italian [90] and Portuguese [91] were created sharing the topics but not its definitions (i.e. vocabulary) (see table 7.1).

A simple way of looking at the output quality of the topic models is by simply inspecting top words associated with a particular topic learned during training. A latent topic is semantically coherent if it assigns high probability scores to words that are semantically related (Gliozzo et al., 2007; Mimno et al., 2011; Newman et al., 2010). It is much easier for humans to judge semantic coherence of cross-lingual topics and their alignment across languages when observing the actual words constituting a topic. These words provide a shallow qualitative representation of the latent topic space, and could be seen as direct and comprehensive word-based summaries of a large document collection.

Samples of cross-lingual topics are provided in Table 7.1. We may consider this visual inspection of the top words associated with each topic as an initial qualitative evaluation, suitable for human judges. Documents present similar topic distributions

[82]http://librairy.linkeddata.es/jrc-es-model-unsupervised
[83]http://librairy.linkeddata.es/jrc-en-model-unsupervised
[84]http://librairy.linkeddata.es/jrc-fr-model-unsupervised
[85]http://librairy.linkeddata.es/jrc-it-model-unsupervised
[86]http://librairy.linkeddata.es/jrc-pt-model-unsupervised
[87]http://librairy.linkeddata.es/jrc-es-model
[88]http://librairy.linkeddata.es/jrc-en-model
[89]http://librairy.linkeddata.es/jrc-fr-model
[90]http://librairy.linkeddata.es/jrc-it-model
[91]http://librairy.linkeddata.es/jrc-pt-model

when projecting their content on topics according to their language as can be seen in fig 7.2. Since the topic identifiers are not aligned, the graphs appear displaced.

7.2 Evaluation

A way to evaluate our cross-lingual document similarity algorithm is to test how well it performs in practice for different real-life tasks: document classification and information retrieval. Evaluation is done using the B-Cubed metrics (Bagga and Baldwin, 1998) to estimate the fit between two clusters, the one obtained from a supervised category-based topic alignment algorithm and the one obtained from our unsupervised synset-based topic alignment algorithm.

JRC-Acquis Corpora											
		en		**es**		**fr**		**pt**		**it**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
prec	*min*	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	max	1.00	0.95	1.00	0.87	1.00	0.87	1.00	0.83	1.00	0.85
	mean	**0.58**	0.48	**0.55**	0.48	**0.55**	0.41	**0.53**	0.42	**0.54**	0.45
	dev	0.27	0.23	0.27	0.22	0.26	0.20	0.24	0.21	0.25	0.22
rec	*min*	0.01	0.03	0.01	0.04	0.01	0.05	0.01	0.04	0.01	0.05
	max	0.96	1.00	0.93	1.00	0.95	1.00	0.92	1.00	0.94	1.00
	mean	0.39	**0.52**	0.36	**0.49**	0.42	**0.51**	0.40	**0.47**	0.39	**0.48**
	dev	0.24	0.20	0.23	0.20	0.23	0.23	0.23	0.21	0.23	0.20
f1	*min*	0.02	0.03	0.01	0.02	0.02	0.03	0.02	0.02	0.01	0.02
	max	0.70	0.75	0.70	0.71	0.70	0.73	0.70	0.71	0.70	0.72
	mean	0.35	**0.42**	0.32	**0.41**	0.37	**0.39**	0.35	**0.38**	0.31	**0.40**
	dev	0.16	0.15	0.15	0.15	0.17	0.17	0.16	0.16	0.16	0.15

Table 7.2: Document classification performance (precision-'prec', recall-'rec' and fMeasure-'f1') of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in monolingual document collections

Let CL_i be the cluster that document t_i gets clustered in, and G_i its correct cluster from the ground truth. The B-Cubed metric then calculates $precision = \frac{|CL_i \cap G_i|}{|CL_i|}$ and $recall = \frac{|CL_i \cap G_i|}{|G_i|}$. The total precision and recall of the clustering are taken as the average of the precision and recall scores over all documents. Results are also presented in terms of the F_1 measure to balance between precision and recall: $F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$.

The aim is to measure the performance of the algorithm taking into account documents with manual category assignments.

7.2.1 Cross-lingual Document Classification

A random group of 1k documents from the JRC-Acquis corpora, which have not been used to train the models, is considered for evaluation as they are manually tagged with EUROVOC categories. For each document, the cluster to which it belongs is identified from its categories. This cluster is then compared (B-Cubed metrics) with the one obtained from the labels generated from its most representative topics (*cat*) and with the one obtained from the labels generated with the WordNet-Synsets of those topics (*syn*). Algorithm performance is evaluated in monolingual, bilingual, and multilingual document collections (tables 7.2 and 7.3) .

JRC-Acquis Corpora									
		en-es		**en-es-fr**		**en-es-fr-pt**		**en-es-fr-pt-it**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
prec	*min*	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	max	1.00	0.97	1.00	0.98	1.00	0.97	1.00	0.98
	mean	**0.62**	0.55	**0.59**	0.52	**0.56**	0.50	**0.57**	0.52
	dev	0.26	0.23	0.26	0.25	0.26	0.26	0.26	0.26
rec	*min*	0.01	0.09	0.01	0.07	0.01	0.06	0.01	0.04
	max	1.00	1.00	0.86	0.93	0.83	0.91	0.80	0.89
	mean	0.33	**0.57**	0.25	**0.39**	0.21	**0.36**	0.23	**0.37**
	dev	0.16	0.23	0.13	0.15	0.12	0.13	0.12	0.15
f1	*min*	0.02	0.02	0.02	0.05	0.02	0.06	0.02	0.08
	max	0.75	0.81	0.62	0.66	0.61	0.64	0.59	0.62
	mean	0.36	**0.49**	0.30	**0.38**	0.29	**0.35**	0.30	**0.36**
	dev	0.16	0.18	0.11	0.12	0.11	0.11	0.11	0.12

Table 7.3: Document classification performance (precision-'prec', recall-'rec' and fMeasure-'f1') of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in multi-lingual document collections

The results show a higher performance of the semi-supervised algorithm (categories-based topic alignment) in terms of precision, and of the unsupervised algorithm (synset-based topic alignment) in terms of coverage. The cause lies in the set of synonyms generated by WordNet, being able to share the same synset for two different topics.

From a more general point of view (fMeasure), the benefit obtained by the increase in coverage (recall) is greater than by the loss of accuracy (precision).

7.2.2 Cross-lingual Information Retrieval

Given a set of documents and a text, the task is to rank the documents according to their relevance to the query text regardless of the language used. The JRC-Acquis corpus is used because by having texts tagged with EUROVOC categories we can build a ground-truth set grouping the documents that share the same codes as those used in the query document. A collection of 1k randomly selected documents (monolingual, bi-lingual and multi-lingual) are annotated by the category-based and synset-based topic alignment algorithms. Then, we randomly take articles to search in D for documents that share the same categories than the query document (i.e the ground-truth set). Next, the query text is used to search in D for similar documents using category-based annotations and synset-based annotations. We evaluate the performance of the algorithms in terms of precision@3, precision@5 and precision@10 (tables 7.4 and 7.5)
.

JRC-Acquis Corpora											
		en		**es**		**fr**		**pt**		**it**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
p@3	*mean*	**0.84**	0.83	**0.81**	0.78	**0.83**	0.74	**0.79**	0.78	**0.80**	0.75
	dev	0.26	0.26	0.27	0.29	0.26	0.32	0.27	0.29	0.27	0.29
p@5	*mean*	**0.82**	0.80	**0.79**	0.75	**0.80**	0.72	**0.77**	0.75	**0.78**	0.72
	dev	0.25	0.25	0.25	0.27	0.25	0.29	0.25	0.26	0.26	0.28
p@10	*mean*	**0.77**	0.76	**0.75**	0.73	**0.77**	0.68	**0.72**	0.71	**0.74**	0.68
	dev	0.23	0.25	0.25	0.27	0.24	0.27	0.25	0.27	0.25	0.26

Table 7.4: Information retrieval performance (precision@3, precision@5 and precision@10) of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in monolingual document collections

Although the precision values are lower than those obtained by semi-supervised approximation, they are sufficiently promising (around 0.75) to think that introducing improvements in the lemmatization process would increase the quality of the WordNet-synset annotations derived from the most representative words of each topic (precision values close to 0.8 in the English corpus).

JRC-Acquis Corpora									
		en-es		**en-es-fr**		**en-es-fr-pt**		**en-es-fr-pt-it**	
		cat	*syn*	*cat*	*syn*	*cat*	*syn*	*cat*	*syn*
p@3	*mean*	**0.84**	0.79	**0.85**	0.75	**0.81**	0.69	**0.82**	0.71
	dev	0.25	0.28	0.24	0.31	0.23	0.29	0.25	0.29
p@5	*mean*	**0.82**	0.76	**0.81**	0.72	**0.78**	0.67	**0.79**	0.69
	dev	0.24	0.26	0.23	0.27	0.24	0.25	0.21	0.26
p@10	*mean*	**0.78**	0.73	**0.76**	0.67	**0.73**	0.62	**0.74**	0.63
	dev	0.22	0.24	0.23	0.26	0.22	0.24	0.23	0.24

Table 7.5: Information retrieval performance (precision@3, precision@5 and precision@10) of the categories-based (*cat*) and synset-based (*syn*) topic alignment algorithms in multi-lingual document collections

7.2.3 Text Length on Cross-lingual Representations

The ability of topics to materialize the underlying knowledge of the documents depend on the texts used to train models. We have studied the impact that the length of texts has, since they determine the space where words can co-occur, to semantically relate multilingual documents described in a probabilistic topic space and, therefore, to capture the knowledge derived from their relationships (Álvarez, 2020).

The aim is to know how text length influences the cross-lingual topics created to semantically relate multilingual documents through state-of-the-art similarity metrics (See Chapter 2). We have designed several document retrieval tasks based on the models described in Section 7.1.3, to compare the clusters of documents created from their manual annotations (i.e. EUROVOC tags) with those created automatically from their topic distributions (Fig.7.4).

The JRC-Acquis dataset used in the previous experiments (Section 7.2.1 and 7.2.2), was extended with the DGT-AcquisSteinberger et al. (2014) collection to increase the total number of documents and the diversity of text length. It contains documents from the Official Journal of the European Union from 2004 to 2011. Given that both datasets are constructed from the same domain with no overlap for data since 2007, we decided to merge both of them into a single collection, the Acquis corpus, formed with all the JRC-Acquis dataset and the documents of DGT-Acquis from that year. *English* and *Spanish* versions of the Acquis corpus were used to validate the results

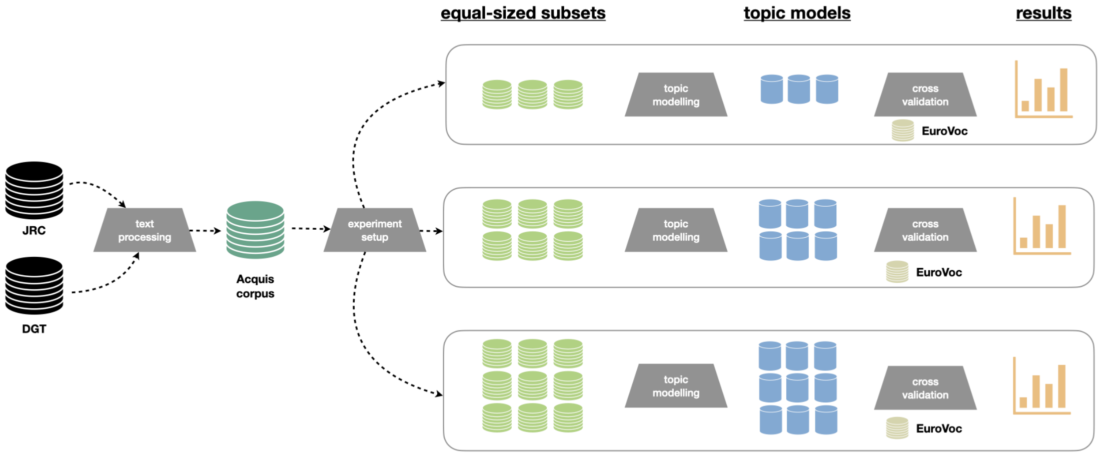


Figure 7.4: Preparation of experiments by creating topic models for each subset of the original corpus and cross-validated with EuroVoc thesaurus

across languages (see table 7.6 for a summary of the data used).

		English			Spanish		
		DGT	JRC	Acquis	DGT	JRC	Acquis
Documents		51521	16260	67781	51585	16470	68055
Tokens	Median	135	197	152	129	204	150
	Mean	185.8762	261.9931	204.1359	181.9172	271.2842	203.5449
	Variance	34806.26	35716.91	36080.66	34624.02	38700.03	37074.97
	Min	7	7	7	6	6	6
	Max	1360	1063	1360	1411	1110	1411

Table 7.6: Number of documents and tokens by dataset

Texts were pre-processed before the models were trained. We removed stopwords, including general NLP and domain-specific ones based on topic distributions. Rare terms with extremely low total document frequency were also removed. Words were lemmatized and changed to lower-case. A lower and an upper limit on the number of words (i.e. tokens) were defined to discard texts. These bounds were inferred from the interquartile range (Fig. 7.7).

We then divided the original corpora into several subsets (3, 6 and 9) of the same size with texts of similar length. The aim is to compare how similarity metrics based on topic distributions behave for each dataset in document retrieval tasks and to un-

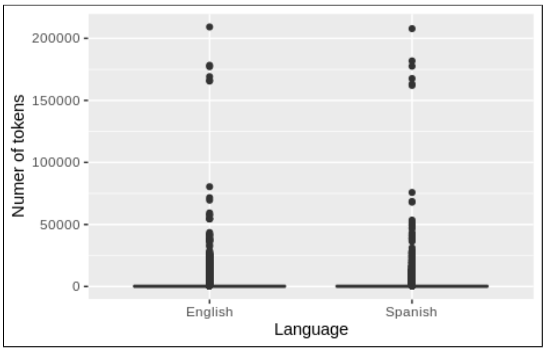


Figure 7.5: Before pre-processing

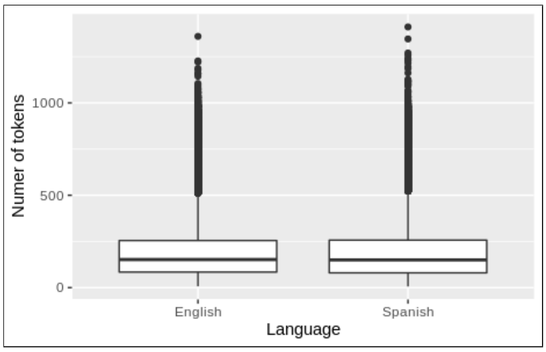


Figure 7.6: After pre-processing

Figure 7.7: Distribution of articles by number of tokens in corpora

derstand the influence of text lengths in the quality of topic models. For each data set, we reserved a sample (5%) for testing and the rest (95%) was used to train a topic model. The test set was projected in the topic space according to the trained model and then the similarity metrics were used to compare those documents with all documents in the corpus and obtain the most similar ones. The list with the top10 most similar documents is compared in terms of Mean Average Precision (MAP) with the one obtained when comparing them from the EuroVoc labels. This metric allows evaluating on average how good the first 10 results of a query are by taking the mean of all average precision for the first 10 results when comparing a list of retrieved documents and the ground truth.

Each document was manually annotated with one or more EuroVoc categories. Documents that share the same categories were therefore considered semantically related. For each document in the test set, its topic-based similarity to the others is calculated according to density-based metrics (2.1) such as Jensen-Shannon divergence (JSD) and Hellinger distance (HE); and our hierarchy-based metric (6.1.2) that we named Weighted Jaccard Levels (WJL) in experiments. The list of related documents from the EuroVoc categories is then compared to the list of related documents by topics expressed by density or by hierarchies. Since several topic models have been created for each dataset (with 50,100,300 and 500 topics), the precision results for each model were averaged following the mean average precision (MAP) metric. Thus, results reflect the capacity of each topic model to express the knowledge required to relate texts from their content without supervision. Table 7.7 shows results for each trained model and each similarity metric. A distinction is made between texts written in English and Spanish.

Acquis (MAP@10)

Lang	Topics	JSD	HE	WJL
Spanish	50	**0.80060**	0.79665	0.70583
	100	**0.82741**	0.77930	0.75555
	300	**0.84261**	0.58531	0.79036
	500	**0.81238**	0.68482	0.79336
English	50	**0.81421**	0.80150	0.73367
	100	**0.85510**	0.74060	0.80315
	300	**0.84005**	0.52082	0.83277
	500	0.78874	0.43636	**0.84555**

Table 7.7: Aggregated MAP results by metric and model

The results suggest that PTM highly capture the knowledge required to relate semantically documents, since all models tested had at least one metric above 0.8 in precision. Among the metrics used to relate documents, the Jensen-Shannon divergence (JSD) offers a better performance in general terms compared to the other approaches. Although a downward trend is suggested in density-based metrics when increasing the number of topics, compared to hierarchical metrics that improve their performance when increasing the number of topics. This happens because the sum of distances of the less representative topics for JSD gets bigger as the number of topics diverge from its optimum while activated topics (i.e selected topics at one of the hierarchy levels) get more discriminative which lead to an increase of WJL. Another way to think about the number of topics is the level of detail they capture. Models with low number of topics will present general themes shared by all documents. On the contrary, with more dimensions topics are able to discriminate particular thematics only shared a subset of the document in the collection which is analogous to how EUROVOC and other theasurus works. Hierarchical metrics work best on high-dimensional topic representations because their calculations are based only on the most relevant topics. Therefore, we can conclude that automatically generated annotations from topic models offer a knowledge close to that offered by those manually assigned from the EuroVoc thesaurus in the Acquis legal corpus to relate texts. In the case of large and very heterogeneous collections, i.e. with a high number of different topics, it would be more appropriate to annotate documents by topic hierarchies. In view of these results, the knowledge offered by topics allows automatically discovering what is being treated in a collection of documents, and the knowledge offered by its hierarchical representation allows understanding why documents are related in a similar way as it would be done with manually assigned labels.

Acquis-3 (MAP@10)

			Training Set					
			1		2		3	
			es	*en*	*es*	*en*	*es*	*en*
Test Set	1	*JSD*	0.85	0.83	0.86	0.85	**0.87**	**0.87**
		WJL	0.85	0.85	0.86	0.86	0.85	0.86
	2	*JSD*	0.80	0.75	0.77	0.75	**0.82**	0.80
		WJL	0.73	0.77	0.81	0.83	0.82	**0.83**
	3	*JSD*	0.72	0.62	0.68	0.65	0.69	0.68
		WJL	0.55	0.65	0.67	0.72	**0.73**	**0.77**

Table 7.8: MAP results by dividing the corpus into three equal subsets to train and evaluate the models in English(*en*) and Spanish(*es*)

To better understand how the length of the texts affects the creation of probabilistic topics, we have prepared three different scenarios that divided the original corpora into subsets with similar text sizes. In the first scenario we have created three equal sets (Table 7.8), in the second scenario there were six subsets (Table 7.9), and in the third scenario a total of nine subsets were created (Table 7.10). We have only considered the similarities calculated from JSD and WJL, as they offered the best performance for each approach.

Models created from texts (training set) with greater or equal length to the texts used in the inferences (test set) offered better performance in document retrieval tasks regardless of the language used. This is evidenced by the fact that those models performed better for almost all sets. Although for some evaluations of small documents models trained with large texts didn't yield the best performances they were not significantly different from the best models. For small documents both metrics performed similarly.

Acquis-6 (MAP@10)

			Training Set											
			1		2		3		4		5		6	
			es	*en*	*es*	*en*	*es*	*en*	*es*	*en*	*es*	*en*	*es*	*en*
Test Set	1	*jsd*	0.79	0.76	0.79	0.77	0.79	**0.78**	**0.79**	0.77	0.79	0.77	0.77	0.73
		wjl	0.78	0.74	0.78	0.76	0.77	0.77	0.78	0.76	0.78	0.76	0.74	0.69
	2	*jsd*	0.82	0.80	0.81	0.77	0.80	0.76	0.81	0.79	0.85	0.82	0.84	0.84
		wjl	0.81	0.82	**0.85**	0.86	0.84	**0.86**	0.85	0.85	0.85	0.85	0.83	0.84
	3	*jsd*	0.76	0.73	0.78	0.73	0.72	0.68	0.78	0.71	0.81	0.75	0.81	0.76
		wjl	0.73	0.70	0.72	0.78	0.81	0.79	0.81	0.80	**0.82**	**0.80**	0.79	0.80
	4	*jsd*	0.69	0.68	0.72	0.67	0.71	0.68	0.68	0.63	0.73	0.69	0.74	0.71
		wjl	0.63	0.69	0.66	0.72	0.74	0.76	0.77	0.78	**0.77**	**0.79**	0.75	**0.79**
	5	*jsd*	0.62	0.57	0.69	0.61	0.66	0.62	0.67	0.59	0.63	0.59	0.70	0.65
		wjl	0.60	0.63	0.57	0.64	0.65	0.69	0.70	0.71	**0.73**	0.74	0.72	**0.75**
	6	*jsd*	0.55	0.52	0.67	0.56	0.61	0.56	0.63	0.56	0.63	0.56	0.59	0.55
		wjl	0.51	0.57	0.51	0.60	0.58	0.63	0.63	0.67	0.66	0.70	**0.69**	**0.71**

Table 7.9: MAP results by dividing the corpus into six equal subsets to train and evaluate the models in English(*en*) and Spanish(*es*)

On the other hand, as the document in the test set get bigger WJL significantly outperformed JSD. As an extreme example look at the Acquis division in 9 groups. The results for the evaluation of the 9^{th} set (group with biggest document) with the 9^{th} model (trained with the biggest document set) were 13% better in the English case and 21% better, suggesting that, with enough text data, PTM models produce sparser topic proportion vectors from which WJL metric benefits.

This perception of the efficiency of hierarchy-based metrics when the topic models are large was analyzed by capturing the computational time (in seconds) that each metric used to compare the documents (Fig.7.8). For almost all number of topics, hierarchical-based metrics are so much faster than probabilistic ones. However, for small number of dimensions (i.e. topics) the topic proportion vector is not sparse enough to identify any relevant topics from the uninformative ones, resulting in hierarchies containing all topics for every document. In other words, all documents share at least one topic. Increasing the number of topics alleviates this problem to the point of archiving an almost constant time for more than 35 topics. Although density-based metrics (e.g JSD and HE) increased their computational time linearly with the representation size, JSD calculation requires computing two logarithms for each dimension in the document representation, which is way more time consuming than the HE met-

Acquis-9 (MAP@10)

Test Set		1 *es*	1 *en*	2 *es*	2 *en*	3 *es*	3 *en*	4 *es*	4 *en*	5 *es*	5 *en*	6 *es*	6 *en*	7 *es*	7 *en*	8 *es*	8 *en*	9 *es*	9 *en*
1	*jsd*	0.88	0.85	0.87	0.87	0.88	0.88	0.88	0.8	0.89	0.89	0.88	0.88	0.89	**0.89**	**0.89**	0.88	0.88	0.82
	wjl	0.89	0.79	0.89	0.82	0.87	0.86	0.88	0.86	0.89	0.87	0.88	0.87	0.89	0.87	0.89	0.87	0.87	0.77
2	*jsd*	0.70	0.66	0.70	0.63	0.71	0.64	0.69	0.63	0.71	0.66	0.72	0.68	0.73	0.70	**0.74**	**0.73**	0.71	0.71
	wjl	0.64	0.59	0.69	0.69	0.69	0.70	0.71	0.68	0.69	0.69	0.71	0.69	0.71	0.70	0.72	0.71	0.67	0.68
3	*jsd*	0.83	0.82	0.86	0.80	0.80	0.75	0.81	0.75	0.83	0.77	0.83	0.79	0.85	0.81	0.86	0.81	0.84	0.82
	wjl	0.80	0.78	0.84	0.83	0.87	0.86	**0.88**	**0.86**	0.88	0.85	0.87	0.85	0.86	0.85	0.87	0.84	0.84	0.83
4	*jsd*	0.74	0.72	0.77	0.70	0.72	0.67	0.65	0.63	0.69	0.63	0.73	0.66	0.76	0.70	0.78	0.72	0.77	0.73
	wjl	0.68	0.67	0.73	0.73	0.76	0.77	0.78	**0.80**	0.79	0.78	**0.80**	0.79	0.80	0.79	0.80	0.77	0.77	0.76
5	*jsd*	0.68	0.68	0.73	0.67	0.70	0.66	0.68	0.64	0.62	0.59	0.69	0.62	0.71	0.67	0.73	0.67	0.74	0.70
	wjl	0.60	0.65	0.64	0.72	0.67	0.73	0.72	0.76	0.75	0.77	**0.77**	**0.78**	0.73	0.78	0.75	0.77	0.74	0.77
6	*jsd*	0.61	0.61	0.68	0.59	0.64	0.58	0.63	0.59	0.63	0.56	0.57	0.54	0.65	0.59	0.68	0.60	0.68	0.62
	wjl	0.53	0.58	0.61	0.65	0.60	0.65	0.67	0.70	0.69	0.71	0.69	0.73	0.71	0.73	**0.71**	**0.74**	0.69	0.71
7	*jsd*	0.53	0.57	0.62	0.52	0.59	0.53	0.56	0.54	0.58	0.52	0.57	0.52	0.52	0.50	0.59	0.53	0.63	0.55
	wjl	0.47	0.55	0.53	0.63	0.52	0.64	0.58	0.66	0.62	0.66	0.65	0.69	0.65	0.70	**0.66**	**0.71**	0.63	0.68
8	*jsd*	0.52	0.48	0.60	0.47	0.59	0.48	0.56	0.47	0.56	0.47	0.57	0.48	0.58	0.47	0.53	0.45	0.60	0.50
	wjl	0.47	0.49	0.53	0.56	0.47	0.56	0.55	0.57	0.56	0.59	0.61	0.62	0.62	0.63	0.64	0.66	**0.64**	**0.66**
9	*jsd*	0.54	0.48	0.62	0.47	0.62	0.50	0.58	0.49	0.59	0.48	0.59	0.49	0.60	0.51	0.60	0.50	0.54	0.45
	wjl	0.51	0.48	0.55	0.55	0.54	0.57	0.59	0.59	0.61	0.59	0.64	0.62	0.63	0.65	0.66	0.65	**0.67**	**0.67**

Table 7.10: MAP results by dividing the corpus into nine equal subsets to train and evaluate the models in English(en) and Spanish(es)

ric square-roots. For the same reason, with a small number of dimensions, pairwise comparison is faster using the probabilistic metrics than the hierarchical metrics.

These results reinforce us in the use of probabilistic topic models to facilitate the exploration of large collections of multilingual documents. The knowledge inferred by these models to automatically group semantically related documents is highly sensitive to the texts used in their training. Their ability to generalize such knowledge only seems to make sense in one direction: with texts whose length is equal to or longer than those used during training. This allows us to conclude that, for example, the knowledge extracted from the topics inferred from a collection of tweets (texts of no more than 260 characters), cannot be extended to automatically classify, for example, blog posts (more than 300 characters). If we assume that the complexity of a text increases as its length increases, the logic used to infer topics is unable to capture more complex knowledge than was proposed during training.

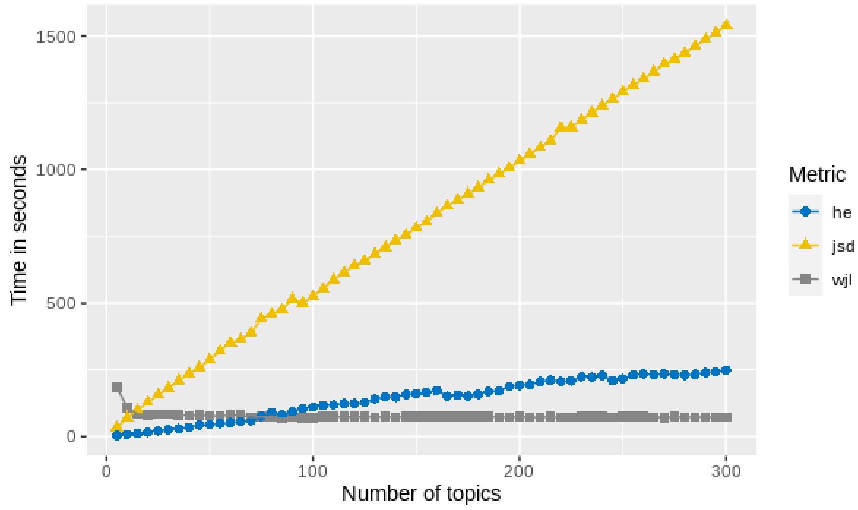


Figure 7.8: Time needed (in milliseconds) to perform the document similarity task using similarity metrics in a corpus of 10^4 synthetic documents

7.3 Summary

In this chapter we have described documents based on topic models that create a unique space of representation between different languages. Topics are created independently for each language, and are projected on concepts instead of words. On concept-based representations, documents in different languages coexist together and can be related. This addresses the last research objective of this thesis (R06, *define a transformation of the topic-based annotations to create a unique representational space out of the particularities from each language*). Representations are analyzed in classification and information retrieval tasks on multilingual document collections. As expected, the performance in terms of accuracy is not as good as that of the approach based on prior knowledge (i.e. topics previously aligned by documents annotated with categories). However, in terms of coverage, the performance of the unsupervised approach is much greater than that offered by the semi-supervised approach, to the point of offering better overall performance (i.e f1) in classification tasks. In addition, the algorithm has proved to perform close to the semi-supervised algorithm in information retrieval task, which makes us think that the process of topic annotation by set of synonyms should be improved to filter those elements that are not sufficiently representative.

In order to perform the evaluations, the new representation system was implemented in our libAIry framework. This extension, together with those described in chapters 5 and 6, cover the last technical objective of this thesis (T04, *create a system capable of finding similar documents automatically*).

Chapter 8

Experiments

The hypotheses raised in this thesis have been further validated in real projects. The implementation of the proposed algorithms and their execution in systems used by third parties, has allowed us to verify the results obtained during the evaluations. Some of these projects are framed in the health field, specifically in the area of HIV treatment (Section 8.1) and, recently, in the field of COVID-19 treatment (Section 8.5). In both cases our algorithms have helped to analyze the use of drugs to treat diseases. There are also projects aimed at measuring the impact of scientific research, at a national level through the creation of patents and research collaborations (Section 8.3), and by research area from a point of view of originality and creativity of research (Section 8.2). Finally, the results of this thesis have also been used to facilitate the exploration of public procurement data in Europe (Section 8.4). Tenders performed by public administrations across Europe were related in an automatic and language-independent way to facilitate their exploration and allow local administrations to know how similar processes are managed in other countries.

8.1 Polypharmacy and Drug-drug Interactions

Does the number of concomitant drugs in people living with Human Immunodeficiency Virus (HIV) increase with age and is it greater than in non-HIV-infected persons? This research question (López-Centeno et al., 2019) seems to be far from the scope of this thesis. However, if we take into account that concomitant drugs are the drugs that a patient also uses to treat other diseases, we can draw an analogy to bring it closer to our domain. It aims to analyze patients from the medicines they receive, and our algorithm organizes documents described by topics. A patient, seen as a document, is described by the drugs he/she receives to treat HIV, which can be represented as topics, and the drugs he/she receives to treat other diseases (i.e concomitant drugs) , which would be used to set the relevance of each topic from their interactions. The relationships between patients depend on the drugs they share, and can be measured by the topics they share when they are represented as documents.

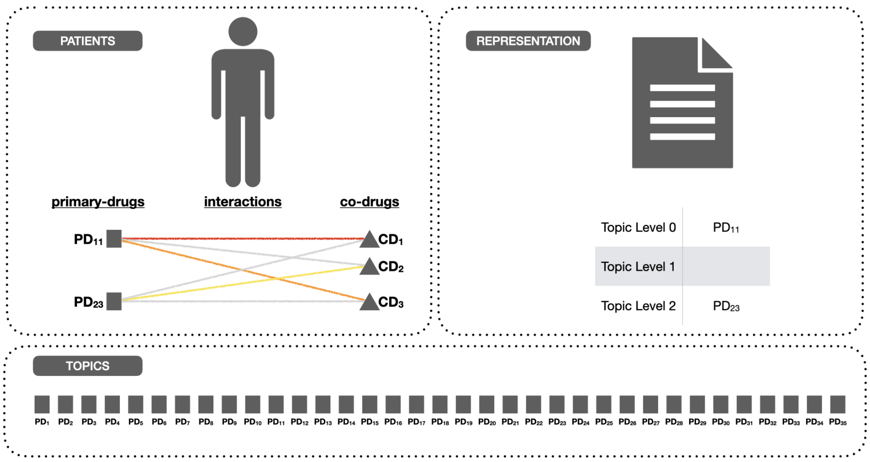


Figure 8.1: Representation of patients through topic hierarchies based on the interactions between primary-drugs and co-drugs

Specifically, in this work we were able to put into practice our approach to make comparisons (Chapter 6) when discovering drug–drug interactions (DDIs) in HIV patients. The anonymized registries from the Madrid Health Service (SERMAS) database were used to build a working database with information about all patients who picked up antiretrovirals (ARVs) or non-HIV medications during the study period. A total of 22,945 people living with HIV and 6,613,506 individuals without HIV had received medications. Medications from the SERMAS database were separated into ARVs and non-HIV drugs: 35 primary-drugs (i.e ARVs medications) and 1,058 co-drugs (i.e non-HIV medications). All drug pairs between primary- and co-drugs were interrogated using the University of Liverpool (UoL) Drug Interactions Database[92], with more than

[92]`https://www.hiv-druginteractions.org`

24,000 HIV DDIs between ARVs and non-HIV medications, to generate a comprehensive list of potential DDIs. The Liverpool flag classification categorizes the severity of DDIs as follows: a red flag indicates medications that should not be coadministered as they might lead to serious adverse events or profoundly affect antiretroviral therapy efficacy; an orange flag indicates a potential interaction that might require dosage modification or close monitoring to minimize clinical consequences; a yellow flag indicates a potential interaction of weak relevance not requiring additional monitoring or dosage adjustment; a green flag indicates no anticipated risk of inter-action; and a gray flag indicates no clear data are available to assess whether a DDI will occur.

We built a representation system where each primary-drug was represented as a topic, and each patient was described as a document containing those topics with different weights (Figure 8.1). This value depends on the DDIs between the primary-drugs and the co-drugs of the patients. Red-flag interactions correspond to the most relevant topics (i.e. level 0), orange-flag interactions correspond to the following topics (i.e. level 1), and yellow-flag interactions correspond to the least relevant topics (i.e. level 2). Primary-drugs are assigned to the most relevant level. Patients are thus described by DDIs, making it easy to identify sets of risks according to the severity of the interactions.

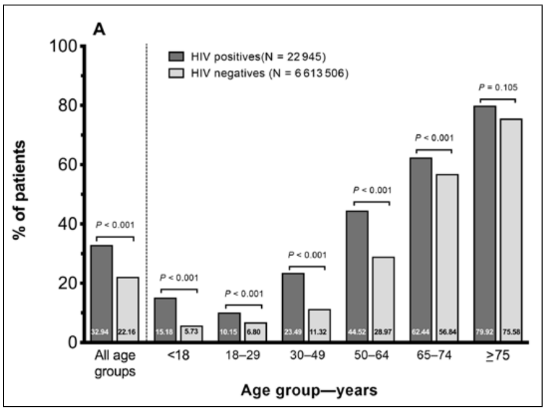


Figure 8.2: Patients grouped by age

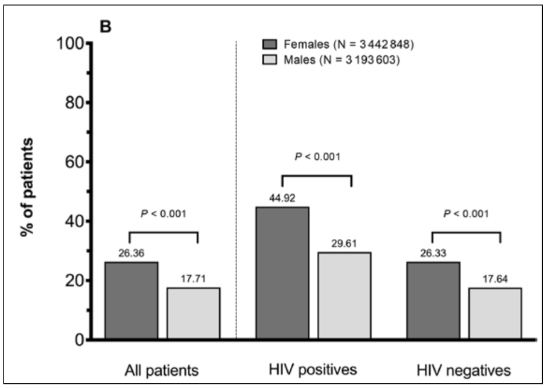


Figure 8.3: Patients grouped by gender

Figure 8.4: Distribution of polypharmacy among people living with and without HIV according to age and gender (López-Centeno et al., 2019).

Our contribution facilitates the analysis of patients and drugs (interactions) without having to make all the comparisons between them (Figure 8.4). Patients are grouped by interactions on the same primary medications. The source code we have developed to discover drug interactions is publicly available for reuse[93], but for privacy reasons we cannot publish the code to manage patients.

8.2 DrInventor

Scientific creativity and innovation are key concepts at a time of rapid technological change. Technologies have great potential to supplement human ingenuity in science by overcoming the limitations that people suffer in pursuing scientific discovery. DrInventor[94] proposes an original system to provide inspiration for scientific creativity by utilizing the rich presence of web-based research resources (**?**). It is like a personal research assistant that informs researchers of a broad spectrum of relevant research concepts and approaches, by assessing the novelty of research ideas, and by offering suggestions of new concepts and workflows with unexpected features for new scientific discovery.

Our topic modeling framework, librAIry (more details in Chapter 4), and the topic-based characterization to measure the similarity between documents (described in Section 5.1.2) powered the DrInventor platform to automatically relate scientific publications from their content. We created a harvester module[95] (Section 4.1.3) that was able to ingest and index research resources from external sources based on the Open Archive Initiative Protocol for Metadata Harvesting[96]. The resources were processed at different levels of granularity: from the entire documents, to their individual items, parts or even individual words contained in them. On top of those resources DRInventor attaches different annotations that further described the instances and gave support to different operations leveraging on them. The model (Figure XXX) provided a standard way of representing research documents, and was flexible enough to give support to a great variety of analysis techniques bringing value to the information stored in it.

The main types of resources (Figure 8.5) that were considered in DRInventor, from the most fine/grained to the most general ones, were:

[93] `https://github.com/cbadenes/hiv-ichart-client`
[94] `http://drinventor.eu`
[95] `https://github.com/cbadenes/camel-oaipmh`
[96] `https://www.openarchives.org/pmh/`

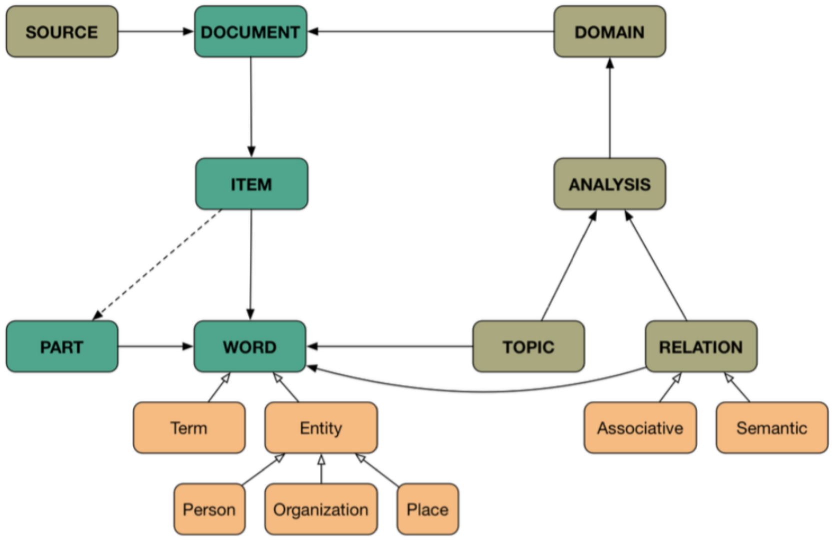


Figure 8.5: Overview of Resources in DRInventor Platform

- **Word**: a meaningful element of writing inside a document, formed by a sequence of characters with no blanks.
- **Part**: logical division of a document, based on categories of the research discourse such as abstract, introduction, methods, results, conclusions, etc., including also the types of rhetorical sentences (Ronzano and Saggion, 2015) (i.e. approach, background, challenge, future work or outcomes).
- **Item**: each of the elements that make up a research object (i.e. a document) such as a paper, programming-code, an image, a workflow, and so on.
- **Document**: meta-information retrieved from a research object. A document is composed by a set of *items*.
- **Source**: indicates the repository where the raw research objects were collected from a link where the platform can look at in order to retrieve them.
- **Domain**: represents the collection of the resources inside DRInventor generated after ingesting the research objects from the repository specified by the *source*.
- **Analysis**: it represents the execution of an algorithm over a particular *domain* in the platform. It is responsible for the creation of annotations, such as *topics*

and *relations*.

- **Term**: represents and abstracts concepts, such as *entities* (persons, locations, organizations...) as results of the execution of different Natural Language Processing algorithms.
- **Topic**: helps to materialize the main subjects that the corpus is elaborating on, such as the research areas or trending issues in the scientific domain.
- **Relation**: is an associative or semantic connection between resources in a domain. It can model for example the high degree of similarity between two *parts* belonging to different research objects.

The DrInventor model served us to refine the model proposed in this thesis (see Section 4.1.1). Some resources were eliminated (e.g. *Source*) to minimize the representation elements to those that can be used to create probabilistic topic models and to calculate similarities between documents (e.g. document, domain, annotation). It has also allowed us to generalize our model by adding a *snippet* resource to represent any element (e.g. part, sentence, entity) that may appears in a document.

8.3 Corpus Viewer

...

8.4 TheyBuyForYou

...

8.5 Drugs4Covid

...

Chapter 9

Conclusions

9.1 Assumptions and Restrictions

...

9.2 Contributions

...

9.3 Impact

...

9.4 Limitations

...

9.5 Future Work

...

Bibliography

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 26–31, Reykjavik, Iceland. 15

Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. (2017). Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167. 90

Álvarez, B. L. (2020). The influence of text length for probabilistic topic models and their hierarchical representation. *Master Thesis*. 114

Andoni, A. and Indyk, P. (2006). Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468. IEEE. 19, 82

Andoni, A., Indyk, P., Nguyá≫. . . n, H. L., and Razenshteyn, I. (2014). Beyond Locality-Sensitive Hashing. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1018–1028. 19, 82

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2017a). An Initial Analysis of Topic-based Similarity among Scientific Documents Based on their Rhetorical Discourse Parts. In Garijo, D., van Hage, W., Kauppinen, T.and Kuhn, T., and Zhao, J., editors, *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic Web Conference (ISWC)*, volume 1931 of *CEUR Workshop Proceedings*, pages 15–22. CEUR-WS.org. 31, 55

Badenes-Olmedo, C., Redondo-Garcia, J., and Corcho, O. (2017b). Distributing Text Mining tasks with librAIry. In *17th ACM Symposium on Document Engineering (DocEng)*. 30, 35, 69

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2019a). Legal document retrieval across languages: topic hierarchies based on synsets. *arXiv e-prints*, page arXiv:1911.12637. 31, 32, 33, 104

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2019b). Scalable cross-lingual document similarity through language-specific concept hierarchies. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 147–153. 33, 81, 104

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2020). Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms. *Semantic Web*. 32

Badenes-Olmedo, C., Redondo-Garcia, J. L., and Corcho, O. (2017c). Efficient Clustering from Distributions over Topics. In *9th International Conference on Knowledge Capture (K-CAP)*, page 8. 31, 55

Badenes-Olmedo, C., Redondo-Garcia, J. L., and Corcho, O. (2017d). librAIry/eval-similarity-calculation. 69, 72

Badenes-Olmedo, C., Redondo-Garcia, J. L., and Corcho, O. (2019c). Large-scale Topic-based Search Java Algorithm. 89, 91

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the 1st international conference on language resources and evaluation workshop on linguistics coreference*, pages 563–566. 111

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable K-Means ++. *Proceedings of the VLDB Endowment (PVLDB)*, 5:622–633. 70

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on Semantic Web and Information Systems*, 5(3):1–22. 37

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35. 13

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022. 10, 11, 58, 108

Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362. 104

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71. 105

Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 75–82, Arlington, Virginia, USA. AUAI Press. 20

Boyd-Graber, J. and Resnik, P. (2010). Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (October):45–55. 13

Celikyilmaz, a., Hakkani-Tur, D., and Tur, G. (2010). LDA Based Similarity Modeling for Question Answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 1–9. 12

Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):1–8. 90

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing - STOC '02*, page 380. ACM Press. 13

Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). BTM : Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941. 11, 109

Christopher D. Manning, P. R. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, UK. 10

Cohan, A. and Goharian, N. (2015). Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. In *Conference on Empirical Methods in Natural Language Processing*, pages 390–400. 57

Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-Based Models of Word Cooccurrence Probabilities. *Machine Learning*, 34(1-3):43–69. 12, 59

Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry - SCG '04*, page 253. ACM Press. 19, 82

De Smet, W. and Moens, M.-F. (2009a). Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining*, SWSM '09, page 57–64, New York, NY, USA. Association for Computing Machinery. 20

De Smet, W. and Moens, M.-F. (2009b). Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, page 57. 21

De Smet, W., Tang, J., and Moens, M.-F. (2011). Knowledge Transfer across Multilingual Corpora via Latent Topics. In *Advances in Knowledge Discovery and Data Mining*, pages 549–560. 21

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407. 10, 11

Divoli, A., Nakov, P., and Hearst, M. A. (2012). Do peers see more in a paper than its authors? *Advances in Bioinformatics*. 15, 56

Dong, F., O'Donoghue, D., Ersotelos, N., Wu, S., Saggion, H., Ronzano, F., Corcho, Ó., Hurley, D., Abgaz, Y. M., Zhang, J., Chaudhry, E., Yang, X., Wei, H., Deng, Z., Mahdian, B., and Careil, J. (2017). Dr. inventor, promoting scientific creativity by utilising web-based research objects. Number 2, pages 40–44.

Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., and Safina, L. (2016). Microservices: yesterday, today, and tomorrow. *CoRR*, abs/1606.0:1–17. 43

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860. 13

Ester, M., Kriegel, H.-p., S, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. 71

Eurovoc (1995). Thesaurus EUROVOC - Volume 2: Subject-Oriented Version. Ed. 3/English Language. Annex to the index of the Official Journal of the EC. In *Luxembourg, Office for Official Publications of the European Communities*. 107

Fielding, R. T. and Taylor, R. N. (2002). Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology*, 2(2):407–416. 42

Fukumasu, K., Eguchi, K., and Xing, E. P. (2012). Symmetric correspondence topic models for multilingual text analysis. In *Proceedings of the 25th annual conference on advances in neural information processing systems (NIPS)*, pages 1295–1303. 20

Ganguly, D., Leveling, J., and Jones, G. (2012). Cross-Lingual Topical Relevance Models. In *Proceedings of COLING 2012*, pages 927–942. 21

Gatti, C. J., Brooks, J. D., and Nurre, S. G. (2015). A Historical Analysis of the Field of OR/MS using Topic Models. *CoRR*, abs/1510.0. 11

Gliozzo, A. M., Pennacchiotti, M., and Pantel, P. (2007). The domain restriction hypothesis: Relating term similarity and semantic consistency. *In. Proceedings of NAACL HLT*, (April):131–138. 110

Greene, D. and Cross, J. P. (2016). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1):77–94. 11, 35

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl:5228–35. 90, 106

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244. 3, 13

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145. 66

Hall, D., Jurafsky, D., and Manning, C. D. (2008). *Studying the History of Ideas Using Topic Models.* 13

Hao, S., Boyd-Graber, J. L., and Paul., M. J. (2018). Lessons from the Bible on Modern Topics: Adapting Topic Model Evaluation to Multilingual and Low-Resource Settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies, NAACL-HLT*, pages 1090–1100. 21

Hao, S. and Paul, M. J. (2018). Learning multilingual topics from incomparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2595–2609, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 21

He, J., Li, L., and Wu, X. (2017). A self-adaptive sliding window based topic model for non-uniform texts. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, volume 2017-Novem, pages 147–156. 11, 35

Hearst, M. a. and Hall, S. (1999). Untangling Text Data Mining. In *the 37th Annual Meeting of the Association for Computational Linguistics*, pages 1–13. 10

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196. 10, 11

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 80–88. 60

Hu, Y., Zhai, K., Eidelman, V., and Boyd-Graber., J. L. (2014). Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1166–1176. 110

Huang, L. K., Yang, Q., and Zheng, W. S. (2018). Online Hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6):2309–2322. 88

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 604–613, New York, NY, USA. Association for Computing Machinery. 18

Jagarlamudi, J. and Daumé, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'2010, page 444–456, Berlin, Heidelberg. Springer-Verlag. 21

Ji, J., Li, J., Yan, S., Tian, Q., and Zhang, B. (2013). Min-Max Hash for Jaccard Similarity. In *2013 IEEE 13th International Conference on Data Mining*, pages 301–309. IEEE. 19, 82, 84, 107

Kenter, T. and de Rijke, M. (2015). Short Text Similarity with Word Embeddings Categories and Subject Descriptors. *Proceedings of 24th ACM International Conference on Information and Knowledge Management*, pages 1411–1420. 3

Krstovski, K. and Smith, D. A. (2011). A Minimally Supervised Approach for Detecting and Ranking Document Translation Pairs. In *Workshop on Statistical MT*. 18

Krstovski, K., Smith, D. A., Wallach, H. M., and McGregor, A. (2013). Efficient Nearest-Neighbor Search in the Probability Simplex. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval - ICTIR '13*, pages 101–108, New York, New York, USA. ACM Press. 19, 82, 91

Kulis, B. and Grauman, K. (2012). Kernelized Locality-Sensitive Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1092–1104. 19, 82

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225. 10

Li, P. and König, C. (2010). b-Bit minwise hashing. In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 671. ACM Press. 19, 82, 107

Li, P., Owen, A. B., and Zhang, C.-H. (2012). One Permutation Hashing. *Advances in Neural Information Processing.* 19, 82, 107

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. 12

Lisena, P., Harrando, I., Kandakji, O., and Troncy, R. (2020). ToModAPI: A Topic Modeling API to Train, Use and Compare Topic Models. In *2^{nd} International Workshop for Natural Language Processing Open Source Software (NLP-OSS).* 15

Liu, X., Zeng, J., Yang, X., Yan, J., and Yang, Q. (2015). Scalable parallel em algorithms for latent dirichlet allocation in multi-core systems. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 669–679, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. 35

López-Centeno, B., Badenes-Olmedo, C., Mataix-Sanjuan, Á., McAllister, K., Bellón, J. M., Gibbons, S., Balsalobre, P., Pérez-Latorre, L., Benedí, J., Marzolini, C., Aranguren-Oyarzábal, A., Khoo, S., Calvo-Alcántara, M. J., and Berenguer, J. (2019). Polypharmacy and Drug-Drug Interactions in People Living With Human Immunodeficiency Virus in the Region of Madrid, Spain: A Population-Based Study. *Clinical Infectious Diseases*, pages 352–362. xx, 123, 125

Lu, H. M., Wei, C. P., and Hsiao, F. Y. (2016). Modeling healthcare data using multiple-channel latent Dirichlet allocation. *Journal of Biomedical Informatics*, 60:210–223. 11

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings* *of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 15

Mao, X., Feng, B.-S., Hao, Y.-J., Nie, L., Huang, H., and Wen, G. (2017). S2JSD-LSH: A Locality-Sensitive Hashing Schema for Probability Distributions. In *AAAI*. 19, 82, 84, 90, 91

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2:3111–3119. 10

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41. 104

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics. 110

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009a). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore. Association for Computational Linguistics. 20

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009b). Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics. 21

Moritz, M. and Ìˆchler, M. B. (2017). Ambiguity in Semantically Related Word Substitutions: an Investigation in Historical Bible Translations. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 18–23. 21

Nair, P. R. and Nair, V. D. (2014). Organization of a Research Paper: The IMRAD Format. In *Scientific Writing and Communication in Agriculture and Natural Resources*, page 150. 61

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North {A}merican Chapter of the Association for Computational Linguistics*, pages 100–108. 110

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2009). Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 1155–1156, New York, NY, USA. Association for Computing Machinery. 20

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2011). Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 375–384. 21

O'Neill, J., Robin, C., O'Brien, L., and Buitelaar, P. (2017). An analysis of topic modelling for legislative texts. *CEUR Workshop Proceedings*, 2143. 11, 35

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. 10

Piccardi, T. and West, R. (2020). Crosslingual Topic Modeling with WikiPDA. *arXiv e-prints*. 21

Platt, J. C., Toutanova, K., and Yih, W.-t. (2010). Translingual Document Representations from Discriminative Projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 251–261, Stroudsburg, PA, USA. Association for Computational Linguistics. 21

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, number 1, pages 248–256. 110

Rao, C. R. (1982). Diversity: Its Measurement, Decomposition, Apportionment and Analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 44(1):1–22. 12

Rob Johnson, A. W. and Mabe, M. (2018). The stm report: An overview of scientific and scholarly journal publishing fifth edition. 1

Ronzano, F. and Saggion, H. (2015). Dr. Inventor Framework: Extracting Structured Information from Scientific Publications. In *Discovery Science: 18th International Conference*, pages 209–220. 57, 127

Rus, V., Niraula, N., and Banjade, R. (2013). Similarity Measures Based on Latent Dirichlet Allocation. In *Computational Linguistics and Intelligent Text Processing*, pages 459–470. 12, 13, 59

Salton, G. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA. 9

Sciences, E. R. S. f. P. and life (2016). Harnessing the power of content - Extracting value from scientific literature: the power of mining full-text articles for pathway analysis Harnessing the Power of content. 56

Simone Teufel (2010). The Structure of Scientific Articles - Applications to Citation Indexing and Summarization. In *CSLI Studies in Computational Linguistics*. 56

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707. 114

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, volume 4, pages 2142–2147. 107

Steyvers, M., Ths, G., and T (2006). Probabilistic topic models. *In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum. Tang, Z. and MacLennan, J.* 69

Tapi Nzali, M. D., Bringay, S., Lavergne, C., Mollevi, C., and Opitz, T. (2017). What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer. *JMIR medical informatics*, 5(3):e23. 11, 35

Terasawa, K. and Tanaka, Y. (2007). Spherical LSH for Approximate Nearest Neighbor Search on Unit Hypersphere. In *Algorithms and Data Structures*, pages 27–38. 19, 82

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent Argumentative Zoning: Evidence from chemistry and computational linguistics. *Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502. 57

Towne, W. B., Rosé, C. P., and Herbsleb, J. D. (2016). Measuring similarity similarly: Lda and human perception. *ACM Transactions on Intelligent Systems and Technology*, 8(1):1–25. 67, 90

Turchi, S., Ciofi, L., Paganelli, F., Pirri, F., and Giuli, D. (2012). Designing EPCIS through Linked Data and REST principles. *Software, Telecommunications and Computer Networks ({SoftCOM}), 2012 20th International Conference on*, pages 1–6. 37

Vijayanarasimhan, S., Jain, P., and Grauman, K. (2014). Hashing Hyperplane Queries to Near Points with Applications to Large-Scale Active Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):276–288. 19, 82

Vulić, I., De Smet, W., Tang, J., and Moens, M. F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management*, 51(1):111–147. 20

Vulić, I. and Moens, M.-F. (2012). Detecting Highly Confident Word Translations from Comparable Corpora Without Any Prior Knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459. 21

Vulić, I. and Moens, M.-F. (2013). A Unified Framework for Monolingual and Cross-Lingual Relevance Modeling Based on Probabilistic Topic Models. In *Advances in Information Retrieval*, pages 98–109. 21

Waleed, A., Dirk, G., Chandra, B., Iz, B., Miles, C., Doug, D., Jason, D., Ahmed, E., Sergey, F., Vu, H., Rodney, K., Sebastian, K., Kyle, L., Tyler, M., Hsu-Han, O., Matthew, P., Joanna, P., Sam, S., Lucy, Lu, W., Chris, W., Zheng, Y., Madeleine, v. Z., and Oren, E. (2018). Construction of the Literature Graph in Semantic Scholar. In *NAACL*. 89

Wang, J., Liu, W., Kumar, S., and Chang, S.-F. (2016). Learning to Hash for Indexing Big Data-A Survey. *Proceedings of the IEEE*, 104(1):34–57. 19, 82

Westergaard, D., Stærfeldt, H.-h., Tønsberg, C., Jensen, L. J., and Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*. 15, 56

Zhang, D., Mei, Q., and Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137, Uppsala, Sweden. Association for Computational Linguistics. 21

Zhang, T., Liu, K., and Zhao, J. (2013). Cross lingual entity linking with bilingual topic model. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 2218–2224. AAAI Press. 20

Zhao, W.-L., Jégou, H., and Gravier, G. (2013). Sim-min-hash. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 577–580. 19, 82, 107

Zhen, Y., Gao, Y., Yeung, D.-Y., Zha, H., and Li, X. (2016). Spectral Multimodal Hashing and Its Application to Multimedia Retrieval. *IEEE Transactions on Cybernetics*, 46(1):27–38. 18

Zhu, Z., Li, M., Chen, L., and Yang, Z. (2013). Building Comparable Corpora Based on Bilingual {LDA} Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–282. 21