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 in 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, in the 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 second 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 with similar solutions. The Spanish[6], American[7] and European[8] patent and 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, for example, 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 necessary 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.

Along with searches based on keywords and categories, the third key aspect aimed at facilitating the exploration of documentary corpora is the provision of related texts. A documentary exploration does not stop when a relevant article is found, but starts from its content shaping the area of interest. Most academic[12] and legal[13] search engines provide a list of related documents for each text and offer navigating through them. The relationship can be based on references, when documents are cited by others, or content, when documents share a thematic area. 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

[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://www.semanticscholar.org`
[13] `https://patents.google.com`

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[14].

While these initiatives are valuable efforts to address access to huge amounts of documents, they are still insufficient to examine the content offered by their texts. 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). Exploring a textual corpus and therefore acquiring some knowledge of its texts requires understanding how its documents are organized through the concepts evoked by its words. It is necessary to focus on why some texts are related to others, and what concepts are key to those relationships. But analyzing and comparing texts on a large scale requires addressing some challenges imposed by external conditions that have appeared in recent years:

- **Complexity**: The huge number of documents has forced a reconsideration of the way to compare them in order to be able to deal with big collections. The time required to compute each comparison should be reduced 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.

[14] `https://www.connectedpapers.com`

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:

- **Large-scale Topic-creation Framework**: We introduce a text processing model that supports the creation of probabilistic topic models. Based on this model, 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.
- **Topic Model Publication and Exploitation**: To facilitate the reuse and exploitation of trained topic models, we propose a format to publish models as web services and provide an online repository with models available for use.
- **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 Massive Document Similarity Comparisons**: We present an efficient mechanism to index and retrieve related documents, described by multi-level 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 topic models on which the work presented in this thesis is built. *Chapter 5* details the text annotation algorithm from 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 for translation. The approach to relate multi-lingual texts from their representations based on topic hierarchies is evaluated in *Chapter 8*, where the results are explained in detail. *Chapter 9* describes real-world projects where contributions from this thesis have been used. Finally, *Chapter 10* introduces conclusions and future lines of work.

1.3 Publications

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

- *Chapter 4*:
  - **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*:
  - **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*:
  - **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*:

– **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, 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 loss of semantic information 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 subjects 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 averaging 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 that use word sequences 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 the feature vector is a topic distribution 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. 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):

$$JS(p,q) = \sum_{i=1}^{K} p_i * \log \frac{2 * p_i}{p_i + q_i} + \sum_{i=1}^{K} q_i * \log \frac{2 * q_i}{q_i + p_i} \tag{2.1}$$

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) :

$$sim_{JS}(D_i, D_j) = 10^{-JS(p,q)} \tag{2.2}$$

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):

$$He(p,q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2)} \tag{2.3}$$

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:

$$sim_{He}(D_i, D_j) = 1 - He(p, q) \tag{2.4}$$

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:

$$S2JSD(P, Q) = \sqrt{2 * JS(P, Q)} \tag{2.5}$$

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

Texts usually contain noisy, 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 using documents for knowledge-intensive tasks is to preprocess them following different techniques to leverage their

content. Recent studies (Westergaard et al., 2017) have shown that mining full-text articles give 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, so we have to take it into account during the whole process.

Exist 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[15] 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[16]) that have been provided through open services decoupled from their technical development (e.g Explosion[17]).

Some approaches have advanced in this direction and offer the creation and exploitation of topic models through an API based on libraries[18] or web services[19](Lisena 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 ecosystems[20] where multiple learning models can be created, but their format is not open and cannot be reused out of the environment. To the best of our knowledge, the ef-

[15] `http://mallet.cs.umass.edu`
[16] `https://spacy.io`
[17] `https://github.com/explosion/spacy-services`
[18] `https://bab2min.github.io/tomotopy`
[19] `https://github.com/D2KLab/ToModAPI`
[20] `https://onnx.ai/`

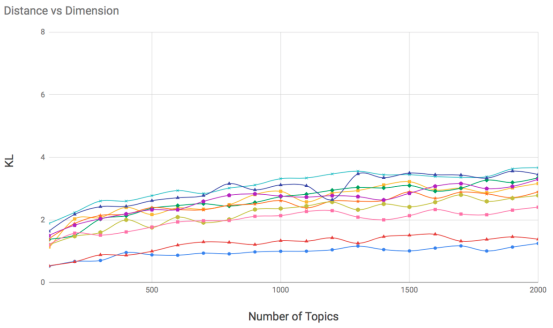
forts made do not propose a unified model to exchange topic models, understood as an already accepted standards-based format. In this thesis we propose *reusable topic models and a scalable framework to create and use them.*

2.2.2 Topic Explainability

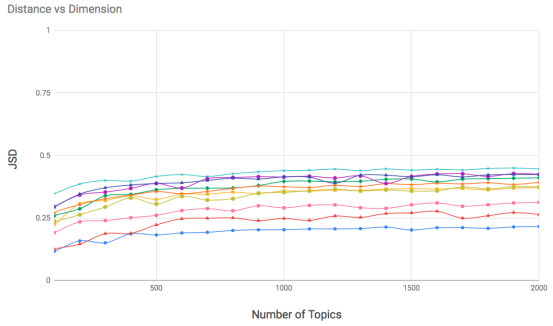
Even though distance metrics mentioned in Section 2.1 have been proposed and used in the SoA, making sense out of the similarity score based on compare topic distributions is not easy. As shown in figure 2.1, given a set of pairs of documents, their similarity scores vary according to the number of topics.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

Distances based on topic distributions between documents 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.*

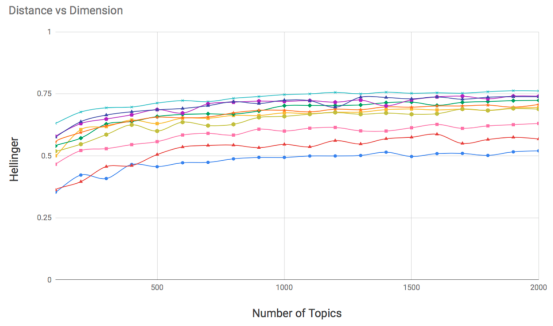
Each topic is drawn from a Dirichlet distribution with parameter β, while each document's mixture is sampled from a Dirichlet distribution with parameter α. These two priors, α and β, are also known as hyper-parameters of a topic model and they set the probability that a document or a word, respectively, 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 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***. Distance thresholds should be model-dependent rather than general and metrics flexible enough to handle dimensional changes. In this thesis we propose a *thematic and low-dimensional feature space suitable for big real-world data sets, where documents are only described by their most relevant topics.*



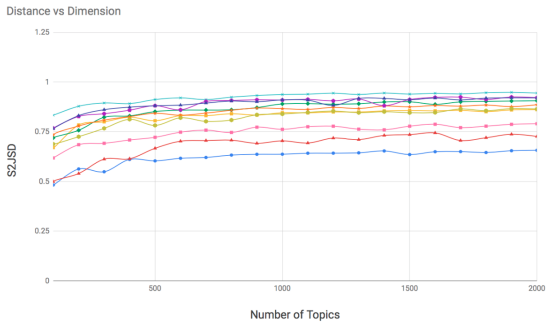
(a)



(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.

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), 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 **??**, 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 allows 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***. 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, giving the reasons 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 apply 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.

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* corpus, with sentence-aligned documents (e.g. Europarl[21] corpora), or *comparable* corpus, with theme-aligned documents (e.g. Wikipedia[22] 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 (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.

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

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, to make them cross-lingual through representations based on topic hierarchies. Documents from multi-language corpora are described by expressions of multi-lingual concepts and can then be efficiently browsed and related without the need for translation or parallel or comparable corpora. In this thesis we propose to *automatically learn 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. 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, data structures that efficiently **sort** texts from their representations based on probabilistic topics are studied. Divisions of space into semantically-related regions are convenient to allow browsing large document collections. The *rep-*

resentativeness 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, sorting, and multilinguality dimensions respectively. First, by *distributing both natural language processing tasks and representational models we can efficiently process huge collections of documents (**H1.1**).*

Second, we can *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**) can 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 losing the ability to rely on 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: Sorting, D4: Multilingualism
H1.1: it is possible to efficiently annotate documents on a large scale by distributing 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 documents with similar topic distributions without calculating all pairwise comparisons and without losing the ability to explore them through their topics	D3: Sorting
H1.4: it is possible to relate documents in different languages without having to translate them 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 Topic-model Programming Interface

Although some initiatives to standardize the format of machine-learning models and to provide tools that facilitate their transformation among the most widespread proprietary formats already exist in the literature, there are still some software restrictions that can limit their reuse. These models may hold certain software dependencies that e.g. force using a specific version of a programming language (python2 vs python3[23]) or an operating system (e.g., linux kernel vs on-cloud environments[24]) to load them or to launch the service that deploys them (e.g., ONNX[25]). This limits their ability to be reused in domains that are not familiar with these technological stacks. *Integrating pre-trained topic models into general-purpose systems is not easy* ***(RCInterface1)***.

Topic models, as many other machine learning models, may be distributed in a proprietary or standard format with software dependencies or by directly providing the data. However, *there is no standard way to specify the topics and the operations that can be performed on them* ***(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.

[23] `https://www.python.org`
[24] `https://vespa.ai`
[25] `https://onnx.ai`

3.2.2 Explainable Topic-based Associations

In order to facilitate the exploration of document collections, vector space models are often used to semantically relate texts based on their word distributions. 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 topics emerged 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 is defined or a 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 finding related documents in a huge corpus is desirable (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 multi-lingual 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: integrating pre-trained topic models into general-purpose systems is not easy	H1.1: documents can be efficiently annotated on a large scale by distributing natural language processing tasks and representation models
RCInterface2: there is no standard presentation of topics that facilitates their reuse	H1.1: documents can be efficiently annotated on a large scale by distributing 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 *platform for large-scale text analysis (**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[26], 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 Associations

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*

[26] `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[27] 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.

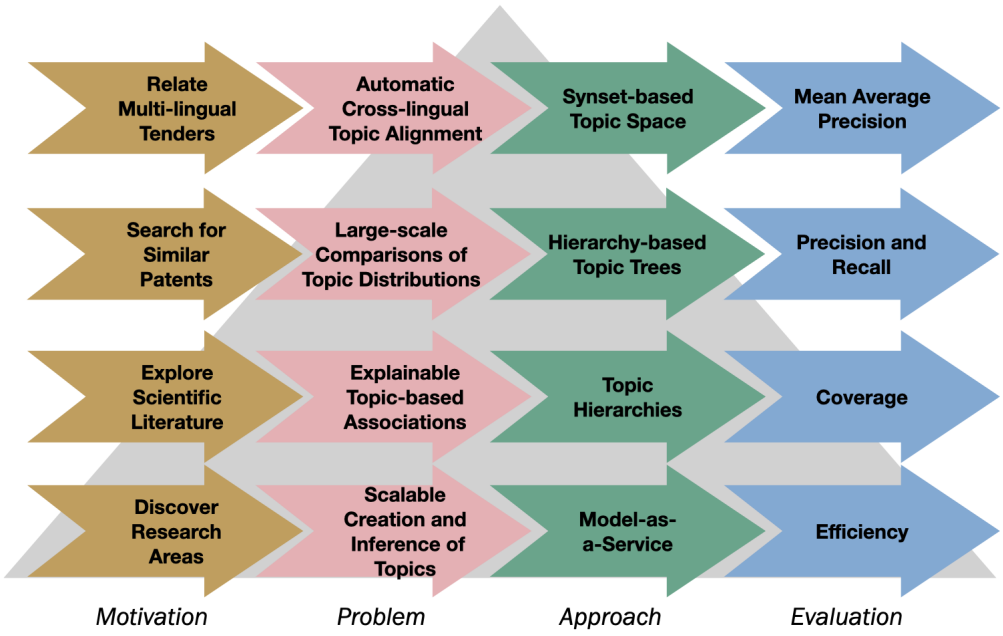


Figure 3.1: Research dimensions of the thesis. The first ones must be overcome before reaching higher dimensions.

[27] `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[28]. 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 Automatic 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[29] corpora and revealed promising results on

[28] `https://www.uspto.gov/ip-policy/economic-research/research-datasets`
[29] `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 platform for large scale text processing	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 document 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 create, publish and 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 Distributed Topic Modeling

There are numerous and varied domains where probabilistic topic models have been successfully used in recent years (Greene and Cross, 2016; He et al., 2017; O'Neill et al., 2017; Tapi Nzali et al., 2017). Each one of them with its particularities. Some with only a few documents, and others with thousands, and even millions, texts. There are environments with only one processor, or environments with multiple processors spread over one or several distributed machines. Taking into account such diversity, topic modeling algorithms have evolved to improve their efficiency in challenging situations, but they only cover the training process of the model. The probabilistic topic model life-cycle begins with text pre-processing, continues with model training, follows with model distribution, and ends with model exploitation. In order to have a framework that covers the entire process of creating probabilistic topic models in both large- and small-scale, we have focused on adapting and reusing techniques and standards widely used in software engineering domain.

librAIry is a framework to manage probabilistic topic models that combines training algorithms with natural language processing tools to create and distribute models suitable for stand-alone use. The main objective is to facilitate the creation of reusable 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[30].

4.1.1 Representing Corpora 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).

[30] `https://github.com/librairy`

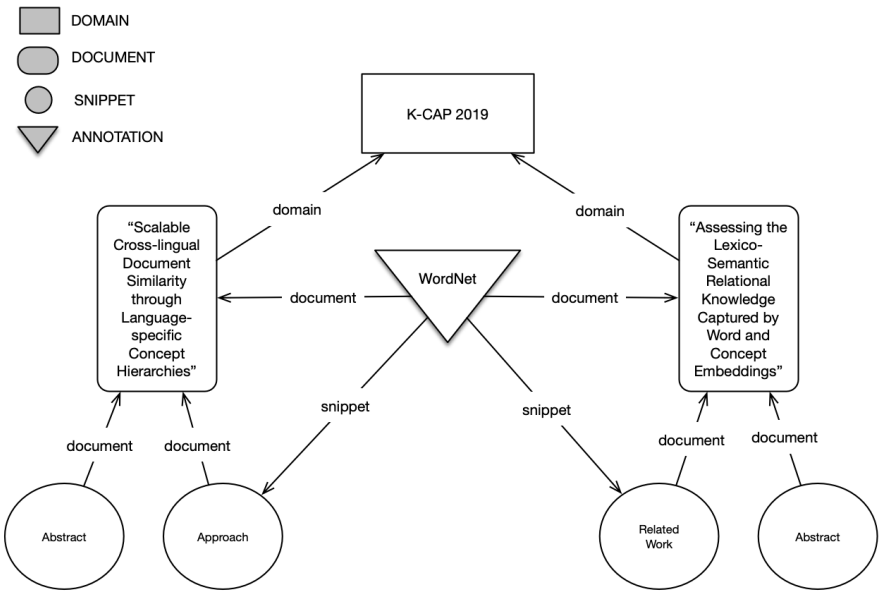


Figure 4.1: Representation of two scientific papers published at the International Conference on Knowledge Capture (K-CAP, 2019) that mention the same entity, Wordnet, in different sections.

Actions can be executed on these resources (e.g. *create*, *update* or *delete*), to change their status (e.g. *created*, *updated* or *deleted*).

To better illustrate this model, take the research articles published at the latest K-CAP conference[31] (See fig. 4.1). A new *document* is created for each publication with the full text of the article. Each *document* is then associated with several *snippets*, one for each section of the article. 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, topics, or keywords).

Resources, *actions* and *states* 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] `http://www.k-cap.org/2019`

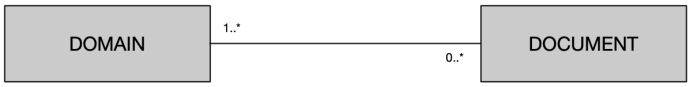


Figure 4.2: Relation between *domain* and *document*.

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 4.2).

A *domain* can contain 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[32].
- **name**: label associated to the resource.
- **description**: additional information about it.

4.1.1.2 Document

A *document* is a resource consisting primarily of words for reading. Examples include research papers, articles, books or patents. It follows the Open Archives Initiative for Object Reuse and Exchange[33] (OAI-ORE) and the Dublin Core Metadata Initiative[34].

A *document* can contain 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.

[32] `https://www.iso.org/standard/40874.html`
[33] `http://www.openarchives.org`
[34] `http://dublincore.org`

- **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[35], orcid[36] 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.
- **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[37]) of the *document*.
- **language**: the language(s) in which the document was written. It is defined by RFC-1766[38] 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.

[35] `http://www.w3.org/wiki/WebID`
[36] `http://orcid.org`
[37] `http://www.iana.org`
[38] `http://www.ietf.org/rfc/rfc1766.txt`

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

- **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 analogies among *documents*, a *document* may contain zero or more references to other *documents* (Fig. 4.3).

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. 4.3), 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[39].
- **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 the *document*.

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

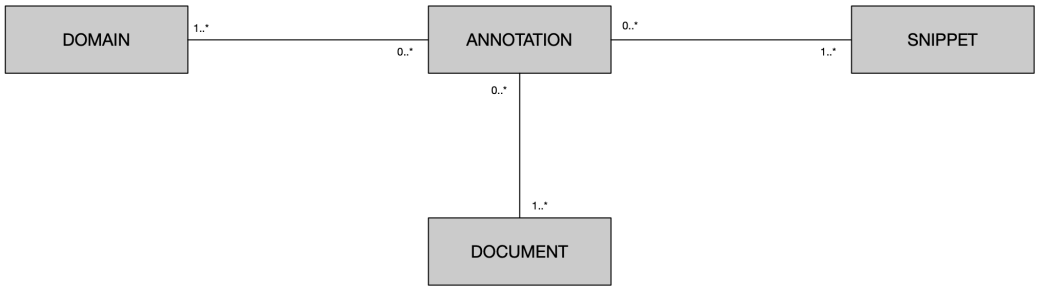


Figure 4.4: Relation between *annotations* and other resources.

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. 4.4)

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[40].
- **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..

4.1.2 Event-oriented Processing Workflow

Along with the resources mentioned above, there are two additional elements that provide a special behavior to the system: *modules* and *events*. An *event* is a non-persistent time-based occurrence that describes a new action performed on a resource.

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

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. 4.5). An *event* notifies a performed action (i.e. a resource and its new state), and follows the Representational State Transfer (REST)Fielding and Taylor (2002) paradigm. 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 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 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).

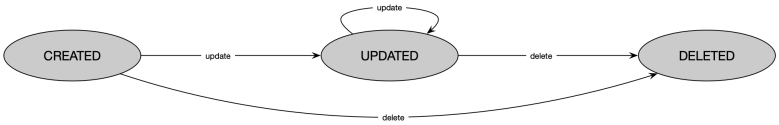


Figure 4.5: Resource states.

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. 4.6):

- **Harvester**: creates resources such as *documents*, *snippets* and *domains*, from local or remote repositories with textual files.
  - *binding-queue* (i.e listening for):
    nothing
  - *routing-key* (i.e publishing to):
    document.created, *snippet.created*,
    domain.(created;updated)
- **Annotator**: makes NLP annotations (i.e named-entities, bag-of-words, etc) and infers topics in *documents* and *snippets*.
  - *binding-key*: *document.(created;updated)*,
    snippet.(created;updated)
  - *routing-key*: *annotation.(created;deleted)*
- **Modeler**: creates a probabilistic topic model from a given *domain*. It uses the texts of its *documents* to train a topic model.
  - *binding-key*: *domain.(created;updated)*

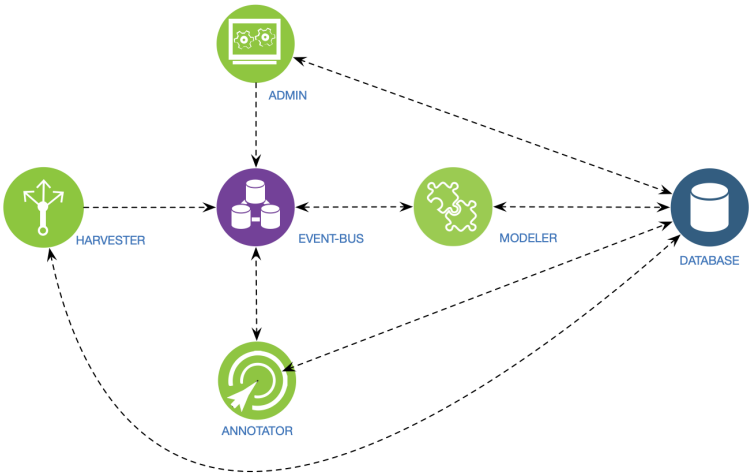


Figure 4.6: Modules (in green), messenger service (in purple) and data storage system (in blue).

  - *routing-key*: *annotation.(created;deleted)*

- **Administrator**: performs user-driven tasks such as reading/writing resources or database queries.

  - *binding-key*: nothing
  - *routing-key*: *domain.(created;updated;deleted)*, *document.(created;updated;deleted)*, *snippet.(created;updated;deleted)*, and *annotation.(created;updated;deleted)*

Each module is wrapped with an Application Program Interface (API) over HTTP, that follows the web standards for the RESTful API development, and a Avro[41]-based interface over TCP for efficiency reasons. Figure 4 shows a sequence diagram that illustrates how modules work to create a topic model when new documents are added to the framework.

[41] `https://avro.apache.org`

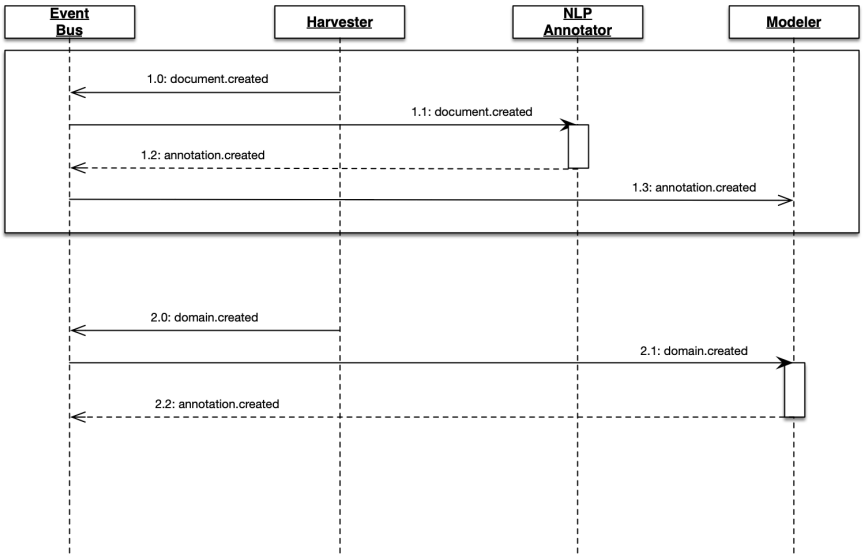


Figure 4.7: Sequence of messages exchanged between modules to create a topic model from the documents added to a domain.

4.2 Reusable Topic Modeling

In order to reuse 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 needed to use it to make inferences and thus avoids unexpected incompatibilities. In this way, we simultaneously facilitate the reuse of topic models and also their scalable execution.

4.2.1 Topic Model Publication

We propose distribute topic models as virtual services hosted in online repositories. Models are then packaged in standard units of software that run reliably applications to explore them from one computing environment to another. A model becomes a standalone, executable package of software that includes everything needed to make inferences: code, data, runtime, system tools, system libraries and settings.

There are several technologies that can virtualize services. Among them, Docker[42]

[42] `https://www.docker.com/`

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[43] so they can be easily downloaded and run on any destination machine. Containers not only offers virtualization advantages, but also version and license control, since they handle some information that we use in the following way:

- **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) and assembled in an *annotation* module of *librAIry*, it can be used to perform multiple tasks through several interfaces. These models, like the other modules of the framework, provide a double API over HTTP and TCP, based on Swagger(Fig. 4.8) and Avro[44] technologies respectively. These services based on probabilistic topic models, have to support the typical tasks of use of this type of models. The *conditions under which the model has been created*, both in relation to the corpus (i.e. number of documents, vocabulary size,), to the NLP tasks (i.e stopwords, PoS filtering) and to the model itself (i.e. hyperparameters), must be registered and available for query. The *topics and 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. It is also required to *infer topic*

[43] `https://hub.docker.com/repositories`

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

distributions in new texts. This ability of topic models is the most extended, since it allows clustering documents with similar topic distributions. And finally, another widely used task when using these models is the *assignment of categories* (i.e. topics) to new texts. Unlike the previous task, where the presence of each topic in a text is measured, now only the most relevant topics (e.g. top3) are considered.

A summary of the tasks that a topic model can provide and their scope can be found in Table 4.1.

Tasks	Scope
T1: Reproducibility	create the same topic model
T2: Exploration	browse topics and words
T3: Inference	calculate topic distributions

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

In order to cover these tasks, the model 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. A topic model is *explorable* (T2) when the word distributions of each topic are known. A topic model is *interpretable* (T3, T4) when the presence of each topic can be measured in a text. Table 4.2 summarizes the operations offered by a topic model service and the tasks that can be performed through its web interface.

The rest of the section describes how the model implements these operations through service methods. An online model is publicly available as web service for review[45].

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 HTTP-GET requests in `/settings` and by TCP requests in *getSettings*. An online example is available here[46].

A detailed list of parameters is shown below:

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

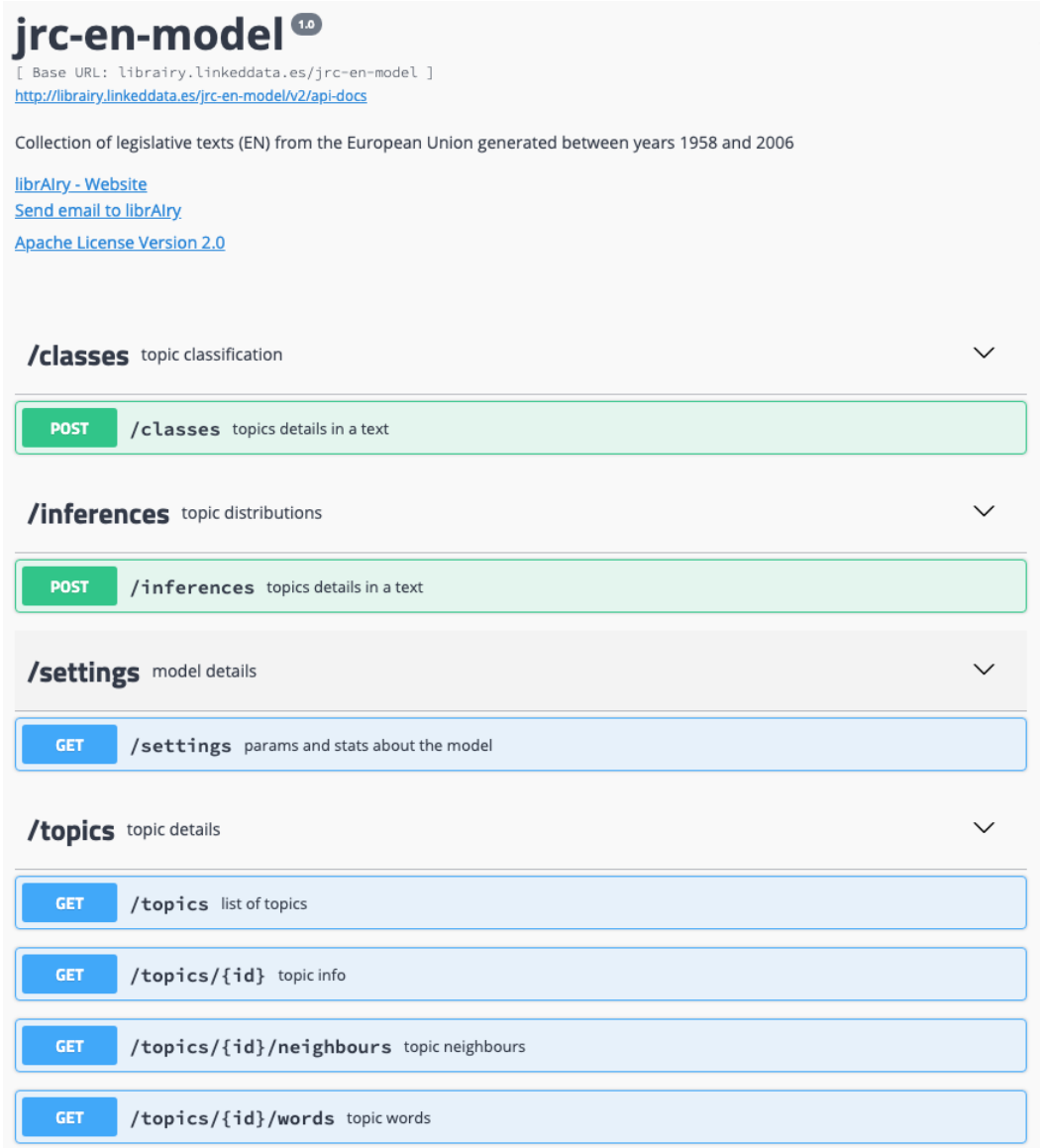


Figure 4.8: Swagger-based web interface of a probabilistic topic model service created with *librAIry*.

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 training parameters (i.e number of iterations, seed, likelihood)	T1: Reproducibility
O4: reading of topics described by word distributions	T2: Exploration
O5: reading of words described by topic distributions	T2: Exploration
O6: calculation of topic distributions in texts	T3: Inference

Table 4.2: Operations offered by a Probabilistic Topic Model.

- **algorithm**: method used to train the model (e.g. LDA, LLDA).
- **date**: date of model creation. It follows the ISO-8601.
- **params**: configuration used during training.
  - **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)

- **params**: some performance metrics.
  - **loglikelihood**: model fit score with respect to the training set.
  - **vocabulary**: number of unique words.
  - **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 Model API offers four different methods for exploring topics and word distributions:

- **Topic List**(O4). By means of the HTTP-GET *topics*[47] method and 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 10 most representative words.

- **Topic Detail**(O4). By means of the HTTP-GET *topics/_id_* [48] method and TCP *getTopic* method, a topic identified by *id* is described by providing *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* [49] method and TCP *getTopicWords* method, a list of word distributions for the topic identified by *id* is provided. Every word has its weight with respect to the topic.

- **Topic Neighbours**(O4). By means of the HTTP-GET *topics/_id_/neighbours* [50] method and TCP *getTopicNeighbours* method, a list of topics related to the topic

[47] `http://librairy.linkeddata.es/jrc-en-model/topics`
[48] `http://librairy.linkeddata.es/jrc-en-model/topics/0`
[49] `http://librairy.linkeddata.es/jrc-en-model/topics/0/words`
[50] `http://librairy.linkeddata.es/jrc-en-model/topics/0/neighbours`

identified by *id* is provided. The distance between topics is measured from their word distributions.

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). The method is available by an HTTP-POST request to `/inferences` or TCP request to *createInference* with a JSON message with the following data:

- **text**: raw text to be analyzed

4.3 Summary

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

We tested and validated *librAIry* by using the framework in some real world scenarios such as DrInventor[51], where thousands of scientific publications were processed, or TheyBuyForYou[52], where hundreds of thousands of public procurement texts were analyzed, or CorpusViewer[53], 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 platform for large scale text processing*).

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 4.1 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 4.6 shows the four modules involved

[51] `https://ec.europa.eu/digital-single-market/en/news/dr-inventor-personal-research-computer-assistan`

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

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

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 4.7, 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 separately or easily integrated into the *librAIry* framework where can be processed in a scalable way. Regardless of the API used, since it works both over HTTP and over TCP, there are four types of tasks that guide the definition of the service: reproducibility, exploration, inference and classification. Tables 4.1 and 4.2 detail the operations that support these tasks. This 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

Chapter 6

Large-scale Comparisons of Topic Distributions

Chapter 7

Cross-lingual Document Similarity

Chapter 8

Evaluation

8.1 Evaluation Metrics

8.2 Text Representativeness

8.3 Large-scale Text Processing

librAIry has been used in some real scenarios such as a research-paper repository for the European project DrInventor [54], a support to decision makers for analyzing patents and public aids for the ICT sector, and also as a book recommender for an online content platform. This has allowed us to identify some weak and strong points of the framework and iterate over the architecture to come with the described solution.

The following modules have been developed[55]: (1) a ***general-purpose harvester*** which retrieves text and meta-information from PDF files in local or remote file-system; (2) a ***research paper-oriented harvester*** focused on collecting and processing more specific textual files (e.g. scientific papers) creating both *documents* and *parts* inferred from the rhetorical classes of the paper; (3) a ***Stanford CoreNLP-based Annotator*** which discovers named-entities, compounds and lemmas from *documents* and *parts*; (4) a ***Topic Modeler*** based on Latent Dirichlet Allocation (LDA) which creates probabilistic topic models for each *domain* in the framework. They are annotated with the set of topics (i.e. ranked list of words) discovered from the corpus, and both *documents* and *parts* of that domain are also annotated by the vector of probabilities to belong

[54] http://drinventor.eu
[55] https://github.com/librairy

to these topics. It uses the Spark implementation of the algorithm; and (5) a ***Word Embedding Modeler*** which creates a *word2vec* model from the *documents* contained in a *domain*.

Due to linear scalability and high performance features, Cassandra has been used to support the column-oriented storage functionality, Elasticsearch as document-oriented storage and Neo4j as graph-oriented storage.

All modules in *librAIry* have been packaged as Docker [56] containers and uploaded to Docker-Hub [57] to facilitate the installation of the system.

Maximizing information re-usability and minimize irrelevant data, becomes specially important when the system handles large collections of data (around million of documents). Fine-grained resource definitions have been key to achieve this, so modules execute actions only when really necessary. When a new *domain* is created, for instance, a new Topic Model is trained for that *domain* and is used to calculate the semantic similarity between the *documents* (and the *parts*) in that domain. If a new *document* (or *part*) is added to that *domain*, the model is trained again and the semantic similarities are re-calculated. However, this becomes unfeasible when the domain is frequently updated and it is composed by a large number of documents. One solution has been to define a new type of resource between domains and documents, models, that describes the representational state (e.g. topic model) of a collection of documents. Thus the model is only re-trained when a significant amount of *documents* are added to the sampling data set and not to the entire *domain*. This less transient model is used to calculate semantic similarities between the *document* collection (and *parts*) inside a *domain* in a more efficient way. Following this more precise execution of tasks, the routing-keys should include the URI of the implied resource into the definition, not only in the content of the message. It would allow modules listening to both the type of a resource or to a specific resource (or subsets, via regular expressions).

While the storage modules are always used to save/update/delete a resource, they are not always required from the end-user. The graph storage, for instance, makes sense when a path between two *documents* or *parts* is requested for a given *domain*. However, some *domains* are not intended to be explored by their linked resources. A

[56] https://www.docker.com
[57] https://hub.docker.com/u/librairy/

more fine/grained definition of resources will allow graph-storage being only used when necessary.

On the other hand, distributed execution of NLP tasks (not only in threads, but also in machines) has proved to be especially useful to handle large collection of *documents*. It requires less processing time than a monolithic solution (e.g. CoreNLP application) and it also provides a dynamic load balancing between modules.

8.4 Topic-based Clustering

8.5 Cross-lingual Similarity

8.6 Conclusions

Chapter 9

Experiments

9.1 Polypharmacy and Drug-drug Interactions

...

9.2 Corpus Viewer

...

9.3 ODS Classifier

...

9.4 Drugs4Covid

...

Chapter 10

Conclusions

10.1 Assumptions and Restrictions

...

10.2 Contributions

...

10.3 Impact

...

10.4 Limitations

...

10.5 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

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

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

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

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

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

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

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

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. 12

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

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.

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. 12

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

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

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

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

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

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.

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. 20

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. 20

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

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

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.

Ganguly, D., Leveling, J., and Jones, G. (2012). Cross-Lingual Topical Relevance Models. In *Proceedings of COLING 2012*, pages 927–942. 20

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

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

Hall, D., Jurafsky, D., and Manning, C. D. (2008). *Studying the History of Ideas Using Topic Models*. 12

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.

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

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.

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

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

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

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

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

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 *2nd International Workshop for Natural Language Processing Open Source Software (NLP-OSS)*.

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

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

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.

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. 20

Moritz, M. and Ì^chler, 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

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.

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. 20

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*.

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. 20

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

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

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

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

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

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. 20

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. 20

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. 18

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

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.

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.

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

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. 20