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 data are produced or captured daily in any imaginable domain offering the possibility of extracting knowledge from them. Much of them are presented in the form of textual documents such as news articles in digital media or scientific papers in electronic journals. Experts who want to find relevant information in this plethora of documents need to browse through them by reviewing their content and filtering out those related to their interests. Understanding the data is essential to distinguish the valuable texts from others.

Griffiths et al. (2007) argued the knowledge in a text arose from the relations established between their words through the concepts they evoked. Similarly Kenter and de Rijke (2015) considered the knowledge in a collection of documents derives from the relations between them. The knowledge based on associations at different levels, i.e. words and documents, can then help to explore textual data guided by a particular interest.

Texts usually contain noisy, non-relevant information. 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 process them following different annotation techniques (entities, keywords, etc) to leverage their content. Multiple algorithms have been proposed to analyze texts for automatically producing such annotations, from minimal units such as terms and entities, to more general descriptors such as summaries or topics. This is a widely known problem in Artificial Intelligence, particularly in the Information Retrieval (IR) and Natural Language Processing (NLP) fields. However, existing solu-

tions fall under strong technological constraints due to their atomic design. In order to reuse their implementations or the models created, it is usually necessary to use the technology on which they are based. *Processing interoperability* has not been a priority issue in the design of these annotation algorithms. The reuse of the raw data created for a specific task is one of the main objectives imposed, without putting too much interest in how the algorithm (i.e its implementation) could be integrated into a specific software environment without the need to reimplement it. In this thesis we propose *reusable text annotation models and scalable document processing pipelines to integrate them.*

Once documents have been processed, texts are usually converted into numeric vectors to operate with them. The definition and number of dimensions for each vector are key aspects when vector space models (VSM) create their representations. In a common and simple approach, term frequencies guide the creation of a space where each word in the vocabulary is represented by a separate and orthogonal dimension. Term Frequency-Inversed Document Frequency (TFIDF) relativizes the relevance of each term with respect to the entire corpus. The loss of semantic information and the high-number of dimensions are the main drawbacks of this approach that lead to the emergence of other techniques. 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 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. 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 are an example of this approach where each document is expressed by the times that it contains each word, and common distributions of words in the collection emerge as topics. Document vectors are learned from the corpus as a probability distribution of how relevant that document is to a given number of topics (and thus a lower-dimensional space). 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). This thesis proposes a *thematic and low-dimensional feature space suitable for document similarity tasks, especially on big real-world data sets, where documents are described by their most relevant topics.*

However, 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 can be used to reduce this complexity. For instance, computation can be approximated by nearest neighbors (ANN) search problem. 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. The simplex spaces where topics are represented have also been approximated by hash recently (Mao et al., 2017), but sacrificing the exploratory capabilities of topics to support document similarity. The notion of topics is lost and therefore the ability to make thematic explorations of documents. Moreover, metrics in the simplex space are difficult to interpret and the ability to explain the similarity score on the basis of the topics involved in the exploration can be helpful. 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.*

When the information extraction 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. In this thesis we propose *to automatically learn cross-lingual topics to browse multilingual document collections*.

In short, in this work we facilitate the exploration of large document collections with texts written in different languages. We address the problem of programmatically generating annotations for each of the items inside big collections of textual documents, in a way that is computationally affordable and enables a semantic-aware exploration of the knowledge inside it. Our proposal automatically discovers thematic associations between texts and organizes the collection so that it can be browsed through related content.

1.1 Contributions

The work presented in this thesis makes the following contributions:

- **Large-scale Topic-based Text-processing Pipeline**: We define a scalable text processing pipeline following web standards and software best practices for the creation and exploitation of probabilistic topic models.
- **Topic Model-as-a-Service**: We propose a format to distribute and reuse probabilistic topic models.
- **Hierarchical Thematic Annotations**: We present a method to annotate texts by topic hierarchies automatically inferred from their content.
- **Massive Document Comparisons**: We leverage multi-level topic annotations to efficiently index and retrieve related documents while allowing the exploration of the collection by the themes inferred from its texts.

- **Cross-lingual Document Relations**: We introduce a technique to transform probabilistic topics from different languages into a single and shared representation space where texts can be thematically related regardless of the language used.

1.2 Thesis Structure

The thesis is structured as follows:

Chapter 2 analyses the state of the art and describes the main concepts handled throughout the thesis. *Chapter 3* presents the research problems and hypotheses that guide our work, 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 efficiently store and search documents 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. This approach is evaluated in *Chapter 8*, where the results are explained in detail. *Chapter 9* provides information on real-world projects where contributions from this thesis have been used. Finally, *Chapter 10* describes 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.
  - José Manuel Gómez-Pérez, Ronald Denaux, Daniel Vila and **Carlos Badenes-Olmedo**. Hybrid Techniques for Knowledge-based NLP - Knowledge Graphs meet Machine Learning and all their Friends. Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP), 69–70. CEUR-WS, 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. Pre-press, no. Pre-press, pp. 1-16. 2020

- Borja Lozano, **Carlos Badenes-Olmedo** and Oscar Corcho. Hierarchical representations of topics to uncover the underlying knowledge of semantically related texts. Proceedings of the 22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW), (under revision). 2020
- **Carlos Badenes-Olmedo**, David Chaves-Fraga, Maria Poveda-Villalon, Ana Iglesias-Molina, Pablo Calleja, Socorro Bernardos, Patricia Martín-Chozas, Alba Fernández-Izquierdo, Elvira Amador-Dominguez, Paola Espinoza-Arias, Luis Pozo, Edna Ruckhaus, Esteban Gonzalez-Guardia, Raquel Cedazo, Beatriz Lopez-Centeno, and Oscar Corcho. Drugs4Covid: Making drug information available from scientific publications. Proceedings of the 19th International Semantic Web Conference (ISWC), (under revision). 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), (under revision). 2020

Chapter 2

Related Work

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

2.1 Text Annotations

The annotation of human-readable documents is a well-known problem in the Artificial Intelligence domain in general and Information Retrieval and Natural Language Processing fields in particular. There already exist a broad set of tools and frameworks able to analyze text for automatically producing such 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 topics or summaries. For example, StanfordNLP Manning et al. (2014) framework allows to perform different operations such as Part-of-Speech (PoS) tagging or Named Entity Recognition in various languages. Others like Mallet[1] or SparkLDA[2] perform topic modeling and clustering. We are focused on the transversal problem of making those standalone tools coexisting under the same solution. Being

[1] `http://mallet.cs.umass.edu`
[2] `https://spark.apache.org/mllib/`

able to effectively integrating them under a common ecosystem helps to seamlessly obtain different kind of annotations and boost the way those solutions can make sense of document collections.

Certain systems among the research and industrial communities have already integrated some of the annotation tools introduced above. For example, Cunningham et al. (2013) works with records from the biomedical domain, where robustness and high precision are prioritized. Therefore they rely on techniques supported by GATE[3] framework, which widely supports hand-crafted, domain specific techniques such as rules or finite state transducers. On the other side of the spectrum we find Li et al. (2012a), where the authors try to annotate text from a much noisier, sparser and error-prone medium: a tweet stream. Therefore they do not rely on any linguistic feature, due to the unpredictable way short social media post are written. We observe how each of those examples has very specific needs and leverages on certain annotation tools in order to accomplish the tasks it was originally created for. In both systems the involved components are highly coupled so they can not be easily extended to contemplate complementary annotation tools or alternative modules.

One crucial problem regarding the re-usability and expansion possibilities of those systems and the tools they leverage on is the language they have been developed in. For example, Mallet uses Java, but others like spaCy [4] are python-based. To the best of our knowledge, there has not been any significant efforts on reconciling into a single architecture such heterogeneous set of tools, therefore minimizing the engineering effort and maximizing scalability of the system so it can be applied to very different domains and textual annotation tasks.

In addition, available annotation systems rely on certain storage solutions that are suited for some tasks but are less adequate others. For example Furlong et al. (2008) uses a relational database (MySQL[5]) to ensure reliability and speed in managing the indexed information. In Rizzo et al. (2015), the authors leverage on Virtuoso triple-store to provide native graph operations over the data. But new requirements may be considered for those systems so different storage needs can come into play.

[3] `https://gate.ac.uk/`
[4] `https://spacy.io`
[5] `https://www.mysql.com/`

For example, column oriented databases (Cassandra[6]) can help to better handle high-volume queries on specific data fields. Same goes with text oriented indexes such as ElasticSearch [7], which can provide customized text-based search operations over the available information.

2.2 Topic-based Relations

Traditional retrieval tasks over large collections of textual documents Hearst and Hall (1999) highly rely on individual features like term frequencies (e.g. TF-IDF). However, new ways of characterizing documents based on the automatic generation of models surfacing the main subjects covered in the corpus have been developed during recent years. Probabilistic Topic Modeling Blei et al. (2010) algorithms are statistical methods 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.

Probabilistic topic models 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. In this way, topic modeling provides 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). This and other topic models such as *Probabilistic Latent Semantic Analysis* (PLSA) Hofmann (2001) are part of the field known as probabilistic 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. While LDA has roots in *Latent Semantic Analysis* (LSA) Deerwester et al. (1990) and PLSA (it was proposed as a generalization of PLSA), it was also influenced by

[6] `http://cassandra.apache.org`
[7] `https://www.elastic.co`

the generative Bayesian framework to avoid some of the over-fitting issues that were observed with PLSA.

This statistical model tries to capture the intuition that documents can exhibit multiple topics. 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 the collection share the same set of topics, but each document exhibits these topics in a different proportion. Documents are represented 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. 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 and they are estimated following some heuristic.

A Dirichlet distribution 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 also it 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.

Unlike a restrictive clustering model, where each document is assigned to one cluster, LDA allows documents to exhibit multiple topics. Moreover, since LDA is unsupervised, the topics covered in a set of documents are discovered from the own corpus; the mixed-membership assumptions lead to sharper estimates of word co-occurrence patterns.

2.2.1 Distance Measures

In a *Topic Model* the feature vector is a topic distribution expressed as vector of probabilities. Taking into account this premise, the similarity between two topic-based resources will be 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 Rao (1982)Lin (1991) 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) Hall et al. (2008) Boyd-Graber and Resnik (2010):

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

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). It 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.

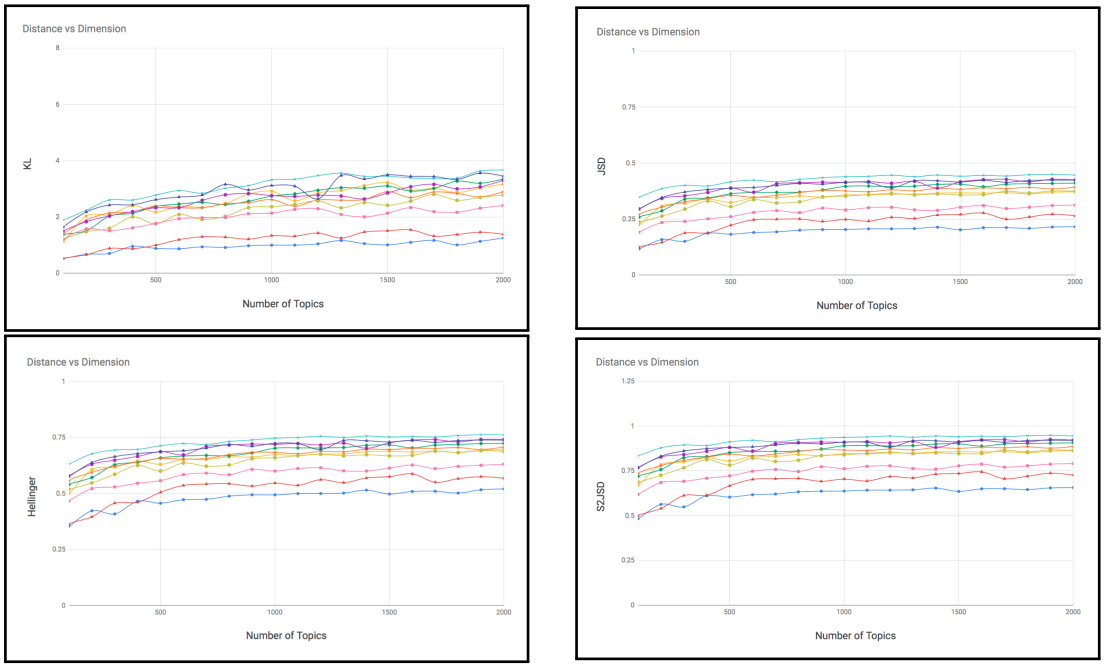

Figure 2.1: Distance values between 10 pair of documents from topic models with 100-to-2000 dimensions.

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

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

2.3 Thematic Document Retrieval

Making sense out of the similarity score 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.

Distances 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. Thus, similarity between pairs of documents is dependent on the model

used to represent them when considering this type of metrics. We know that absolute distances between documents vary when we tune hyperparameters differently, but we also see that "relative distances" also change: e.g. for model M1, A is closer to B than C, but according to a M2 trained in the same corpora with different parameters , A is closer to C than 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.

These challenges have been tackled in this thesis 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. A first approach transformed the He divergence into an Euclidean distance so that existing ANN techniques, such as LSH and k-d tree, could be applied Krstovski et al. (2013). But this solution does not consider the special attributions of probability distributions, such as Non-negative and Sum-equal-one. Recently, a hashing schema Mao et al. (2017) taking into account the symmetry has been proposed, non-negativity and triangle inequality features of the S2JSD metric for probability distributions. For set-type data,

Jaccard Coefficient is the main metric used. Some examples are K-min Sketch Li et al. (2012b), 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.

2.4 Multilingual Topic Alignment

Multilingual probabilistic topic models (MuPTM) Vulić et al. (2015) have recently emerged as a group of language-independent generative machine learning models that can be used on large-volume theme-aligned multilingual text. 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 (2009), document classification De Smet et al. (2011) Ni et al. (2011), semantic similarity of words Mimno et al. (2009) Vulić and Moens (2012), information retrieval Vulić and Moens (2013) Ganguly et al. (2012), document matching Platt et al. (2010) Zhu et al. (2013), and others.

Once a PTM or MuPTM has been generated, documents can be represented by data points in a feature space based on topics to detect similarities among them exploiting inference results and using distance calculation metrics on it. Since exact similarity computations are unaffordable for neighbours detection tasks ($O(n^2)$), some algorithms based on approximate nearest neighbor (ANN) techniques have been proposed to efficiently perform document similarity search based on the low-dimensional latent space created by probabilistic topic modelsZhen et al. (2016) Mao et al. (2017). They transform data points from the original feature space into a hash-code space, so that similar data points have larger probability of collision (i.e. having the same hash code). However, the smaller space created by existing hashing methods lose the exploratory capabilities that topic models offer and the explanatory power that topics

have to support the document similarity. The notion of topics is discarded and therefore the ability to make thematic explorations of documents.

In this thesis we take PTM a step further, to make them cross-lingual through hierarchical representations created from a lexical data base. Documents from multi-language corpora are described by hierarchical expressions of multi-lingual concepts and can then be efficiently browsed and related without the need for translation. Hash codes are created from those concept hierarchies to perform document classification and information retrieval tasks on large document collections.

Chapter 3

Research Objectives

The work presented in this thesis aims to facilitate the exploration of large 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 infer 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 studied in this work.

In the third dimension, data structures that efficiently **sorting** texts from their representations based on probabilistic topics were studied. Divisions of space into semantically-related regions are necessary to allow browsing large document collections.

The *representativeness* covered in the previous dimension enables the interpretation of the relations and regions obtained.

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

This chapter introduces our main hypothesis (3.1), its associated research challenges (3.2) and presents the research methodology (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 big 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 down the line (**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 objectives 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 calculate 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[8]) or an operating system (e.g., linux kernel vs on-cloud environments[9]) to load them or to launch the service that deploys them (e.g., ONNX[10]). 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 it.

[8] `https://www.python.org`
[9] `https://vespa.ai`
[10] `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 large 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 Automatic Cross-lingual Topic Alignment

With the ongoing growth in the number of digital articles in different languages, we need annotation methods that enable browsing multi-lingual corpora. 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 table3.2

3.3 Research Methodology

The research presented in this thesis is based on four dimensions or research areas. 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 us 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. 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

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

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.

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 that they can be assigned to new documents. Such a high volume of data made difficult to process it manually, so we needed to automatize 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 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[11], 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 Section 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*

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

*was computationally affordable and enabled a semantic-aware exploration of the knowledge inside it (**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 Open Research Corpus[12] in domains such as Computer Science, Neuroscience and Biomedicine (Badenes-Olmedo et al., 2017c) (Badenes-Olmedo et al., 2017a) (Badenes-Olmedo et al., 2019a).

The main contributions under this dimension are described in Section 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;

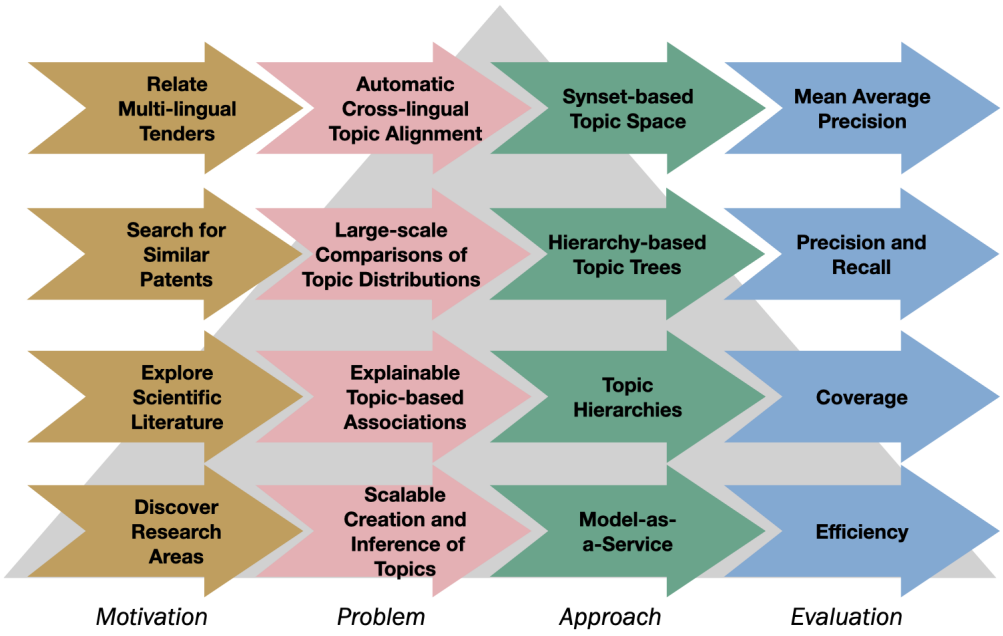

Figure 3.1: Research dimensions of the thesis. The first ones must be overcome before reaching higher dimensions.

[12] `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 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[13]. The relations between patents derived from their manual categorization were compared with those automatically obtained from their topic distributions (Badenes-Olmedo et al., 2020)(Badenes-Olmedo et al., 2019a).

The main contributions under this dimension are described in Section 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

[13] `https://www.uspto.gov/ip-policy/economic-research/research-datasets`

Portuguese editions of JCR-Acquis[14] corpora and revealed promising results on classifying and sorting documents by similar content across languages (Badenes-Olmedo et al., 2019b)(Badenes-Olmedo et al., 2019a).

The main contributions under this dimension are described in Section 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.

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

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

Scalable Creation and Inference of Topics

This chapter presents *librAIry*[15], our text management platform that combines natural language processing techniques with automatic learning algorithms to analyze large collections of documents. It serves as a technological framework where we can implement the advances of our research and measure its performance.

4.1 Document Workflow

Given the huge amount of textual data about any domain that is daily being produced or captured in any imaginable domain, it becomes crucial to provide mechanisms for programmatically processing this raw data so we can make sense out of it: discarding all the noisy, non-relevant information and keeping only the data that can bring value for the involved agents (general consumers, experts, companies, investors...).

While some specific tools already allow for advanced sense-making operations, others opt for composing a solution where different analysis techniques are integrated under a uniform data schema. However, this integration involves significant efforts on reconciling data sources, coordinating processing operations, and efficiently exploiting results from the execution of those techniques. There is the need for a more flexible paradigm where tools and algorithms for textual document analysis, from different programming languages and technologies, can operate independently and in a collaborative manner

[15] http://librairy.linkeddata.es

creating a common document oriented workflow through their actions. In the context of the scientific publications, the personalized recommendation of research papers based on their content is a key novel feature for performing a smart selection of relevant resources over very big collections of scientific content. From the set of values and different attributes extracted from the papers and by generating advanced knowledge models about the information they contain we can bridge across the different relevant pieces of information and allow users to navigate them in a more efficient and powerful way. This knowledge about a specific document is frequently acquired by different techniques focused on revealing certain aspects of it, that are later combined to achieve one particular task.

The architecture presented in this thesis aims to ease the way different software modules work together and lays the foundation for efficiently process big volumes of textual documents in a distributed, decoupled manner.

4.2 librAIry

librAIry is a framework where different text mining tools, available in various languages and technologies, can operate in a distributed, high-performance and isolated manner creating a common workflow through their actions. Instead to work towards a predefined sequence of actions, synchronization across modules is achieved through the aggregation of the operations executed by them in response to an emergent chain of events. This raises both technical and functional challenges to coordinate multiple executions. From the technical point of view, isolated environments and communication mechanisms are provided so initially dissimilar tools can be executed with maximum guarantees. From the functional point of view, all executions are coordinated to reach a final result as aggregation of partial results derived from each execution.

4.2.1 Functional Features

The architecture is articulated around three main concepts: (1) the **resource** such as *document*, a *part-of-a-document*, or a *domain*. (2) the **actions** performed over them: *create*, *update* or *delete* a resource. And (3) the new **state** that is reached by the resource after an action is performed, such as *created*, *updated* or *deleted*. An **event** is a message containing details about those three aspects, published on a shared

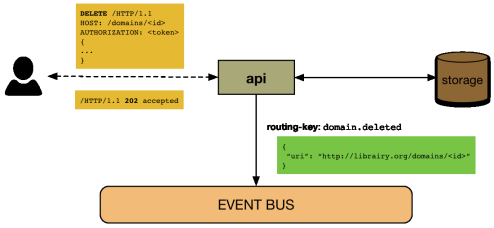

Figure 4.1: Domain deleted flow.

event-bus available for all the modules deployed in the framework. This will, in turn, allow that any module can perform actions on one or more resources in response to a new state reached by a given resource. Actions executed in parallel from distributed environments.

4.2.1.1 Resources

Two main kinds of resources are considered: those derived from external sources such as (1) *documents* from textual files (e.g. a research paper), (2) *parts* from logical divisions of a *document* (e.g. rhetorical classes or sections), and (3) *domains* from sets of *documents* (e.g. a conference or journal), and those derived from processing the previous ones such as *annotations*.

To better illustrate this model, consider to explore the research papers published at the SIGGRAPH conference in 2016[16]. First, every paper will be materialized as a new *document* containing the full-text. Immediately after, the *document* will be automatically associated to several *parts*, each of them grouping sentences by rhetorical class (e.g. approach, background, challenge, future work and outcome) and by section (e.g abstract, introduction). Finally, a new *domain* will be created grouping all these *documents*. Different analysis will be performed extending the initial set of resources with more annotations at several representational levels: at *document level*, full-text based annotations are provided such as named-entities, compounds and descriptive tags. At *relational level*, connection between resources are found (e.g. semantic similarity-based relationships). And finally, at *domain level* annotations such as tags and summaries are composed describing the corpus of *documents*.

[16]http://s2016.siggraph.org

4.2.1.2 Event-based Paradigm

An event illustrates a performed action, i.e. a resource and its new state. It follows the Representational State Transfer (REST)Fielding and Taylor (2002) paradigm, but taking into account the state reached after an action, i.e *created*, *deleted* or *updated*. Thus, an event contains the resource type and the new state reached by a specific resource.

4.2.1.3 Linked Data Principles

Data in *librAIry* is individually addressable and linkable Turchi et al. (2012) following the Linked Data principles defined by T. Berners-Lee Bizer et al. (2009). Thus, resources (i.e. a *domain*, a *document*, a *part* or an *annotations*) have: (1) a URI as name, (2) a retrievable (or dereferenceable) HTTP URI so that it can be looked up, (3) a useful 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.

4.2.2 Framework Architecture

Following a publisher/subscriber approach, all the modules in the framework can publish and read events to notify and to be notified about the state of a resource. Therefore, the system flow is not unique and is not explicitly implemented, instead distributed and emergent flows can appear according to particular actions on resources.

4.2.2.1 Event-Bus

We use the Advanced Message Queuing Protocol (AMQP) as the messaging standard in *librAIry* to avoid any cross-platform problem and any dependency to the selected message broker. This protocol defines: *exchanges*, *queues*, *routing-keys* and *binding-keys* to communicate publishers and consumers.A message sent by a publisher to an exchange is tagged with a routing-key. Consumers matching that routing-key with the binding-key used to link the queue to that exchange will receive the message. In *librAIry* this 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

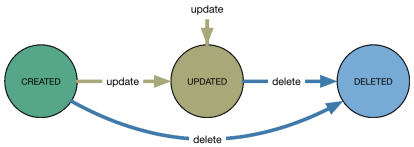

Figure 4.2: Resource states.

(e.g. ́domains.created ́for new domains), or multiple type events (e.g. #.created for all new resources).

4.2.2.2 API

A HTTP-Rest Application Program Interface (API) was designed for interaction with end-users. Any external operation motivated by a user will be handled here. Some of them, usually those related to reading operations, will be completely managed by this module getting all the data from the internal storage. However, those operations implying a modification of the status of some resource (e.g. creation of a *document*), may be also performed by other modules listening for that type of event asynchronously. This module publishes to the following routing-keys: *domain.(created;updated;deleted)*, *document.(created;updated;deleted)*, *part.(created;updated;deleted)*, and *annotation.(created;updated;deleted*

4.2.2.3 Storage

Multiple types of data can be handled in this ecosystem. Inspired in the Data Access Object (DAO) pattern, we have created a Unified Data Manager (UDM) providing access to any type of data used in the system. Three types of databases have been considered:

- **column-oriented database**: Focused on unique identified and/or *structured data*. This storage allow us searching key elements across resources.
- **document-oriented database**: Focused on indexing raw text. This storage allow us to execute advanced search operations over all the information gathered about a textual resource.
- **graph database**: Focused on relations. This storage allow us exploring resources through the relationships between them.

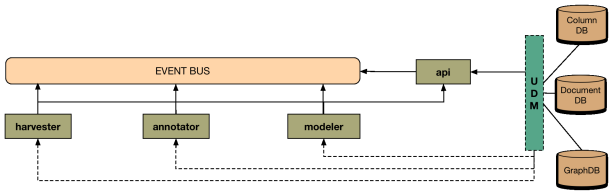

Figure 4.3: Modules.

4.2.2.4 Modules

The modules composing *librAIry* have been designed following the microservices architectural style. A module is a cohesive (i.e. it implements only functionalities strongly related to the concern that it is meant to model Dragoni et al. (2016)) and independent process working on the framework with a specific purpose. This purpose is defined by both the routing-key and the binding-key associated to the events handled by the module.

These are the main types of modules identified in *librAIry*:

- **Harvester**: creates system resources such as *documents*, *parts* and *domains*, from local or remote located textual files.
  - Listening for: nothing
  - Publishing to: *document.(created)*, *part.(created)*, *domain.(created;updated)*
- **Annotator**: retrieves named-entities, compounds, lemmas and other annotations resulting of Natural Language Processing (NLP) task execution from *documents* and *parts*.
  - Listening for: *document.(created;updated)*, *part.(created;updated)*
  - Publishing to: *annotation.(created;deleted)*
- **Modeler**: builds representational models from a given *domain*.
  - Listening for: *domain.(created;updated)*
  - Publishing to: *annotation.(created;deleted)*

4.3 Model-as-a-Service

...

4.4 Summary

In *librAIry*, existing algorithms and tools coming from different technologies can work collaboratively to process and analyze large collections of textual resources which has been successful applied to some real scenarios [17].

A new model definition based on the previously mentioned principle of maximizing information re-usability and minimize irrelevant data is being studied to create a more fine-grained resource design. New domains, in the sense of particular vocabularies or specific textual formats, are also being analyzed to be included into the system via specific harvesters or more precise annotators. Moreover, a template-based mechanism oriented to facilitate the integration of new tools and techniques into the system is being built to make easier to develop new modules as well as increasing the available modules at Docker-Hub.

[17] `http://drinventor.dia.fi.upm.es`

Chapter 5

Explainable Topic-based Associations

5.1 Topic Relevance

5.2 Topic-based Clustering

5.3 Summary

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 [18], 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[19]: (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

[18] http://drinventor.eu
[19] 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 [20] containers and uploaded to Docker-Hub [21] 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

[20] https://www.docker.com
[21] 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

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. 15

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. 15

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. 27

Badenes-Olmedo, C., Redondo-Garcia, J., and Corcho, O. (2017b). Distributing Text Mining tasks with librAIry. In *17th ACM Symposium on Document Engineering (DocEng)*. 26

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. 27, 28, 29

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. 29

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2020). Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms. *Semantic Web*. 28

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. 27

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. 34

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6):55–65. 11

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

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

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

Celikyilmaz, a., Hakkani-Tur, D., and Tur, G. (2010). LDA Based Similarity Modeling for Question Answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 1–9. 13

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. 3

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLOS Computational Biology*. 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. 13

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. 15

De Smet, W. and Moens, M.-F. (2009). 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. 16

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. 16

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*, 41:391–407. 11

Divoli, A., Nakov, P., and Hearst, M. A. (2012). Do peers see more in a paper than its authors? *Advances in Bioinformatics*. 9

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. 36

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860. 14

Fielding, R. T. and Taylor, R. N. (2002). Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology*, 2(2):407–416. 34

Furlong, L. I., Dach, H., Hofmann-Apitius, M., and Sanz, F. (2008). Osirisv1. 2: a named entity recognition system for sequence variants of genes in biomedical literature. *BMC bioinformatics*, 9(1):84. 10

Ganguly, D., Leveling, J., and Jones, G. (2012). Cross-Lingual Topical Relevance Models. In *Proceedings of COLING 2012*, pages 927–942. 16

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. 2

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. 3

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244. 1, 13

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

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. 3

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. 11

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

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. 16

Kenter, T. and de Rijke, M. (2015). Short Text Similarity with Word Embeddings Categories and Subject Descriptors. *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag. (CIKM 2015)*, pages 1411–1420. 1

Krstovski, K. and Smith, D. A. (2011). A Minimally Supervised Approach for Detecting and Ranking Document Translation Pairs. In *Workshop on Statistical MT*. 3

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. 15

Kulis, B. and Grauman, K. (2012). Kernelized Locality-Sensitive Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1092–1104. 15

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., and Lee, B.-S. (2012a). Twiner: Named entity recognition in targeted twitter stream. pages 721–730, New York, NY, USA. ACM. 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. 16

Li, P., Owen, A. B., and Zhang, C.-H. (2012b). One Permutation Hashing. *Advances in Neural Information Processing*. 16

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. 13

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. 2

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*. 9

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*. 3, 15, 16

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). 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. 16

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. 16

O'Neill, J., Robin, C., O'Brien, L., and Buitelaar, P. (2017). An analysis of topic modelling for legislative texts. *CEUR Workshop Proceedings*, 2143. 3

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. 16

Rao, C. R. (1982). Diversity: Its Measurement, Decomposition, Apportionment and Analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 44(1):1–22. 13

Rizzo, G., Troncy, R., Corcho, O., Jameson, A., Plu, J., Hermida, J. C. B., Assaf, A., Barbu, C., Spirescu, A., Kuhn, K.-D., et al. (2015). 3cixty@ expo milano 2015: Enabling visitors to explore a smart city. 10

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. 13

Sciences, E. R. S. f. P. and life (2016). Harnessing the power of content - Extracting value from scientific literature: the power of mining full-text articles for pathway analysis Harnessing the Power of content. 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. 3

Terasawa, K. and Tanaka, Y. (2007). Spherical LSH for Approximate Nearest Neighbor Search on Unit Hypersphere. In *Algorithms and Data Structures*, pages 27–38. 15

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. 34

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. 15

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. 16

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. 16

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. 16

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. 15

Westergaard, D., Stærfeldt, H.-h., Tønsberg, C., Jensen, L. J., and Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*. 9

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. 16

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. 3, 16

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. 16