Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 27-36 recibido 30-04-2021 revisado 08-06-2021 aceptado 09-06-2021

Impact of Text Length for Information Retrieval Tasks based on Probabilistic Topics

Influencia de la Longitud del Texto en Tareas de Recuperación de Información mediante Tópicos Probabilísticos

Carlos Badenes-Olmedo, Borja Lozano-Álvarez, Oscar Corcho
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{cbadenes, ocorcho}@fi.upm.es
{borja.lozano.alvarez}@alumnos.upm.es

Abstract: Information retrieval has traditionally been approached using vector models to describe texts. In large document collections, these models need to reduce the dimensions of the vectors to make the operations manageable without compromising their performance. Probabilistic topic models (PTM) propose smaller vector spaces. Words are organized into topics and documents are related to each other from their topic distributions. As in many other AI techniques, the texts used to train the models have an impact on their performance. Particularly, we are interested on the impact that length of texts may have to create PTM. We have studied how it influences to semantically relate multilingual documents and to capture the knowledge derived from their relationships. The results suggest that the most adequate texts to train PTM should be of equal or greater length than those used to make inferences later and documents should be related by hierarchy-based similarity metrics at large-scale.
Keywords: probabilistic topics, text similarity, hierarchical topics, document retrieval.

Resumen: La recuperación de información ha utilizado tradicionalmente modelos vectoriales para describir los textos. A gran escala, estos modelos necesitan reducir las dimensiones de los vectores para que las operaciones sean manejables sin comprometer su rendimiento. Los modelos probabilísticos de tópicos (MPT) proponen espacios vectoriales más pequeños. Las palabras se organizan en tópicos y los documentos se relacionan entre sí a partir de sus distribuciones de tópicos. Como en muchas otras técnicas de IA, los textos utilizados para entrenar los modelos influyen en su rendimiento. En particular, nos interesa el impacto de la longitud de los textos al crear MPT. Hemos estudiado cómo influye al relacionar semánticamente documentos multilingües y al capturar el conocimiento derivado de sus relaciones. Los resultados sugieren que los textos más adecuados deben ser de igual o mayor longitud que los utilizados para hacer inferencias posteriormente y las relaciones deben basarse en métricas de similitud jerárquicas.
Palabras clave: topicos probabilísticos, semejanza de textos, jerarquía de tópicos, recuperación de documentos.

1 Introduction

Probabilistic Topic Models (PTM) (Hofmann, 2001) (Blei, Ng, and Jordan, 2003) 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 topic distributions, per-document topic distributions or per-document per-word topic assignments. In turn, a topic is a distribution over terms that is biased around those words associated to a single theme. Figure 1 shows some topics that have emerged when creating a topic model with the collection of Wikipedia articles to better understand what

ISSN 1135-5948. DOI 10.26342/2021-67-2

Topic 3	Topic 14	Topic 31
business	game	church
bank	software	bishop
million	games	catholic
development	video	priest
investment	computer	apostolic
management	data	diocese

Figure 1: Topics discovered from the English edition of Wikipedia.

topic means. Each topic is described, this example, by its six most representative words, i.e., those words most present in the documents that mainly contain each topic.

Bag-of-words approach avoids the restriction of word sequences to relate documents based on the use of the same words. PTMs represent texts by probability distributions over a vocabulary created from the whole corpus. Each topic establishes different levels of relevance for each word, and documents are described based on the presence of each topic in their texts. *Latent Semantic Indexing* (Deerwester et al., 1990) (LSI) initially reduced the TF-IDF model using *singular value decomposition* to find the linear subspace that capture most of the information in a collection. LSI has shown to yield high correlation with human perception of similarity (Jung, Ruthruff, and Goldsmith, 2017) and some authors argued that features derived from LSI are able to capture some basic linguistic notions such as synonymy and polysemy. *Probabilistic LSI* (pLSI) (Hofmann, 1999) improved LSI by introducing the concept of topic as a multinomial distribution over the vocabulary of a collection. In pLSI each document is described with a vector of topic proportions, capturing the idea that there is a fixed number of common themes exhibited in a different proportion by the documents in a collection. But it was not able to provide a *generative* process that infers topic proportions for documents not used in the training collection. *Latent Dirichlet Allocation* (LDA) (Blei, Ng, and Jordan, 2003) solved the inferring problem of pLSI by placing a Dirichlet distribution over the topic proportions for the documents and allowing for the discovery of the themes running through the documents. It is considered the simplest *generative* Probabilistic Topic Model. LDA is one of the most widely-used methods when processing texts using NLP techniques and its functionality has been extended to multiple domains (Jelodar et al., 2017).

Topic models are trained with large corpora of texts, which are generally from the same domain for which we want to make inferences. Documents can be related based on their topics, instead of sequences of words. Topic-based representations bring a lot of potential when applied over different information retrieval (IR) tasks, as evidenced by works in different domains such as health (Nzali et al., 2017), legal (O'Neill et al., 2017), news (He, Li, and Wu, 2017), and hybrid proposals combining topic models and word embedding (Dieng, Ruiz, and Blei, 2020). However, the ability of topics to express the inherent knowledge on which the relationships between documents are built has not been yet analyzed from the texts used to train the models. As far as we know, there are no studies that evaluate how the text length influences on the probabilistic model created to represent and relate semantically documents from their topic distributions by means of similarity functions. In this work we have studied the impact that the text length has, since it determines the space where words can co-occur, to semantically relate documents described in a probabilistic topic space.

This paper is structured as follows: Section 2 presents the state-of-the-art metrics used to compare documents represented by probabilistic topics. The methodology that we have used for the experimentation and how the evaluation was performed is described in section 3. Finally, the results are presented and discussed in section 4, along with the final remarks and future work in section 5.

2 Text Similarity based on Probabilistic Topics

Some works have evaluated LDA models to semantically relate documents. In (Syed and Spruit, 2017), the quality of the topics was measured based on the abstract or the full text of scientific articles. It concluded that full-text was less prone to noisy topics in small datasets. Regarding the ability to relate similar papers, (Badenes-Olmedo, Redondo-García, and Corcho, 2017b) ana-

lyzed similarity relations based on topic distributions when using only sections of scientific papers to describe them (i.e. abstract, method, background, etc). It concluded that the background section allows relating them in a more accurate way than using the abstract.

Distance measures typically used in probabilistic topic models are not based on Euclidean spaces, e.g. cosine-similarity, but consider the simplex space created by the Dirichlet distribution to support the comparisons.

2.1 Density-based Similarity metrics

Documents are represented as vectors of topic distributions in the simplex space created by probabilistic topic models, and distance functions must take into account two considerations, namely *none-negativity* and *sum-equal-one* (Mao et al., 2017) to ensure that the document representation is used as a probability distribution. Metrics such as Jensen-Shannon Divergence (JSD) (Eq.1) (also known as symmetric relative entropy) and Hellinger (He) distance (Eq.2) are commonly used in these spaces (Rus, Niraula, and Banjade, 2013):

$$JSD(Q,D) = \sum_{i=1} q_i \log \frac{2q_i}{q_i + d_i} + \sum_{i=1} d_i \log \frac{2d_i}{q_i + d_i} \quad (1)$$

$$He(Q,D) = \sum_{i=1} \left(\sqrt{q(x_i)} - \sqrt{d(x_i)} \right)^2 \quad (2)$$

However these metrics lack the interpretative capacity offered by topics when comparing documents. Furthermore, in real-world environments where computational cost has to be considered those metrics do not scale well as they require complex operations between all pairs of documents. To address both issues, a set of metrics based on hierarchical representation of topics were proposed, as described next.

2.2 Hierarchy-based Similarity metrics

Similarity metrics based on density functions present four major problems when comparing documents (Badenes-Olmedo, Redondo-García, and Corcho, 2019a):

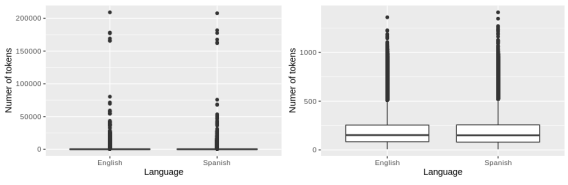
- Pairwise computation of document similarity is costly and grows linearly with the size of the corpus.
- Simplex metrics do not offer a semantic explanation for the similarity obtained.
- Documents that do not share any activated topics (i.e. the bigger components of the topic proportion vector) can still have high similarity due to the sum of distances between the less representative topics (i.e. the smaller components of the topic proportion vector).
- These metrics cannot be extended to support semantic restrictions to enrich queries in the corpus.

To alleviate these issues, a new approach to compare topic distributions was proposed (Badenes-Olmedo, Redondo-García, and Corcho, 2019a) that reduces the topic distributions vector to a hierarchical set-type vector. Documents are described by sets of topics grouped into three relevance levels. They are compared using the Jaccard index, a metric that compares how similar two sets are by how many objects they share. In our experiments, a linear distribution of weights (i.e $w_i = 3 - i$) has been used to add up the hierarchy levels (Eq.3):

$$WJL(H^A, H^B) = \sum_{i=0}^{L} \sum_{j=0}^{L} w_i w_j * \frac{|H_i^A \cap H_j^B|}{|H_i^A \cup H_j^B|} \quad (3)$$

3 Experiments

Our study is aimed at evaluating how text length influences the probabilistic topics that are created from a document corpus, and how this influences the calculations and results of state-of-the-art similarity metrics to semantically relate documents. Several document retrieval tasks were designed from annotated document collections. The study considers both multilingual and monolingual scenarios. From a collection of documents manually tagged with categories, we train topic



(a) Before text processing. (b) After text processing.

Figure 2: bag-of-words size.

models to create representation spaces where texts are projected and compared to identify similar documents. We divide the original corpus into several datasets by grouping documents with similar length to measure how text length influences the relations obtained. A probabilistic topic model is trained for each dataset, and is used to make inferences across all datasets. In this way we evaluate the performance of topic models to relate similar documents when the length of the texts used in training and inferences vary (Fig.3).

3.1 Corpora

A multilingual corpora was created from the English and the Spanish editions of the JRC-Acquis (Steinberger et al., 2006) and DGT-Acquis (Steinberger et al., 2014) datasets. It contains 135.836 legislative texts of the European Union (EU) from the 1950s to 2011. The length of the texts is calculated using white spaces and punctuation marks to distinguish terms. More advanced techniques based on phrases or entities could have been used, but we want to avoid the noise they might introduce in their inferences. The median length of the texts, since Acquis is a parallel corpus, is 152 terms for English texts and 150 terms for Spanish texts with a high variance from less than 7 terms in the shortest texts to more than 1.300 terms in the longest texts (Table 1). The distribution of documents according to their number of tokens is shown in Figure 3.

Documents are annotated with the EuroVoc taxonomy, which follows the International Standards for processing the documentary information of the EU institutions (ISO 2788-1986 and ISO 5964-1985). It is a multilingual thesaurus with 7,193 concepts/labels from 21 domain areas such as politics, international relations, law, economics, etc. In our study we used the 452 root concepts identified in (Badenes-Olmedo, Redondo-García, and Corcho, 2019b) to categorize documents. In this way we ensure independence between probabilistic topics when creating the models from these categories. This is a restriction imposed by topic models as they are described by density functions.

		English	**Spanish**
#Documents		67781	68055
#Terms	*Median*	152	150
	Mean	204.13	203.54
	Variance	36080.66	37074.97
	Min	7	6
	Max	1360	1411

Table 1: Multilingual corpora created from the JRC and DGT Acquis datasets.

3.2 Text Pre-Processing and Topic Model Training

Texts were pre-processed to remove common stopwords and domain-specific ones based on topic distributions. Rare terms with extremely low total document frequency were also removed. Words were lemmatized and transformed to lower-case. A lower and an upper limit on the number of words were defined to homogenize the size of bag-of-words. These bounds are based on the interquartile range (Fig. 2) and are commonly applied in the state-of-the-art (Schofield, Magnusson, and Mimno, 2017).

Topic models were created using the Gibbs sampling implementation from *librAIry*(Badenes-Olmedo, Redondo-García, and Corcho, 2017a) system. By default it only uses *verbs*, *nouns*, *proper nouns* and *adjectives* to create the models. The Dirichlet priors $\alpha = 0.1$ and $\beta = 0.01$ were set following the conclusions from (Hu et al., 2014). Models with 50, 100, 300 and 500 topics were considered to analyze their ability to capture the knowledge needed to accurately relate similar documents.

3.3 Experimental Scenarios

Our task consists in searching for related documents to a given text, using representations based on different trained probabilistic topics, and comparing this result with the set of related documents, based on their overlap in terms of Eurovoc categories (Figure 3). The ability of probabilistic topics to capture the inherent knowledge of the corpus and allow documents to be related to each other from their vector representations is evaluated by comparing the relationships obtained by this process with those obtained from the manual labels they share.

Each document in the original corpus is manually annotated with EuroVoc categories. The original set of categories was reduced to 452 independently identified areas. Documents that share the same categories are considered to be semantically related and serve as a ground-truth to validate

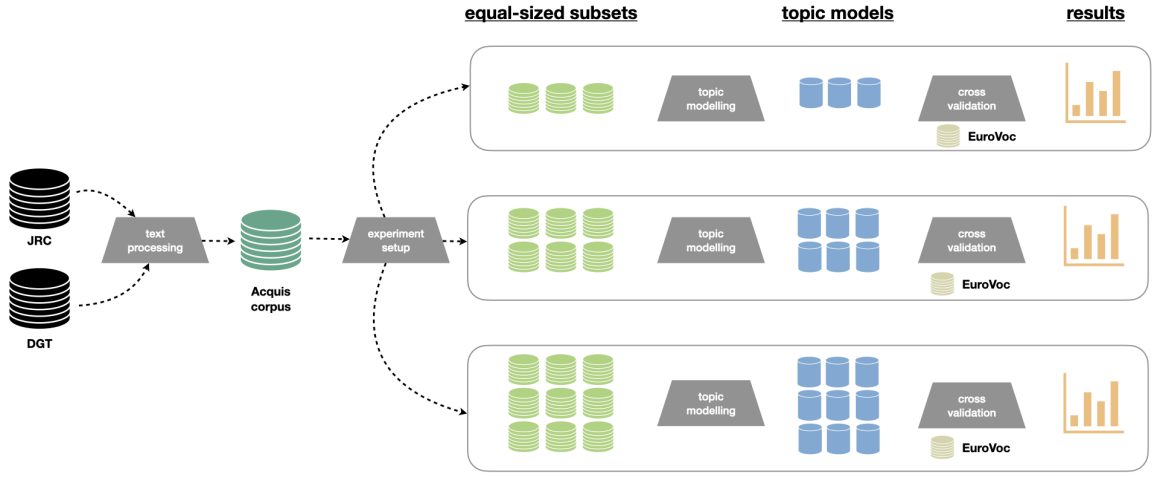


Figure 3: Preparation of experiments by creating topic models for each subset of the original corpus and cross-validated with EuroVoc thesaurus.

the unsupervised approach based on probabilistic topics.

Three evaluation scenarios were created, each of them dividing the initial corpora into subsets of the same size with texts of similar length. We have considered 3, 6 and 9 divisions of the original corpus in order to have enough detail when analyzing the results. The higher the number of divisions, the greater the detail but the lower the number of documents in each subset and this may affect the quality of the trained topic model. With these three scenarios we have an adequate balance between detail and quality of topic models.

Documents were pre-processed to filter out verbs, proper names, nouns and adjectives (i.e tokens) and to create bag-of-words with them. The inter-quartile index (± 1.5) was taken into account to discard too short or long texts (see Table 2).

Data was divided into a sample subset (5%) for testing and the rest (95%) was used to train a topic model. The test set was described by topic distributions based on the trained model. State-of-the-art distance metrics were used to compare them and to obtain the most similar ones. The top10 most similar documents are evaluated in terms of Mean Average Precision (MAP) with the top10 obtained when comparing them from the EuroVoc labels. MAP allows evaluating on average how good the results of a query are by taking the mean of all average precisions for the first 10 results when comparing a list of retrieved documents and the ground truth.

#Divisions	Partition	#Docs	Median	Mean
3	1	22,594	43	49.75
	2	22,593	152	152.76
	3	22,594	337	409.88
6	1	11,297	20	20.74
	2	11,297	84	78.75
	3	11,297	129	128.42
	4	11,296	175	177.11
	5	11,297	255	261.76
	6	11,297	517	558.03
9	1	7,532	13	14.77
	2	7,531	43	45.49
	3	7,531	89	88.98
	4	7,531	121	120.56
	5	7,531	152	151.96
	6	7,531	185	185.76
	7	7,531	236	238.73
	8	7,531	337	346.14
	9	7,532	619	644.73

Table 2: tokens per partition and division.

4 *Results*

As expected, models trained with small documents perform worse than those with large documents specially for the supervised splits where groups had the same number of documents(see tables A.1 to A.6). The smallest document group still had the worst performance in the unsupervised split, but, due to groups not having the same number of documents, the biggest document group usually had the second worst P@k, specially for larger texts.

4.1 Categorization based on Topics

A topic-based similarity to all documents in corpus is calculated according to density and hierarchy-based metrics (described in Section 2) for each test document.

Since several topic models have been created for each dataset (with 50,100,300 and 500 topics), the precision results for each model were averaged following the mean average precision (MAP) metric. Thus, results reflect the capacity of each topic model to automatically capture the knowledge required to relate documents from their texts. As shown in Table 3, the use of probabilistic topics to automatically relate documents offers a performance with an accuracy above 0.8. This performance is slightly higher for English texts than for Spanish texts. We suspect that this is due to the difference in quality of the text processing tools for each language (i.e. lemmatized, PoS, etc.).

Among the metrics used to relate documents, the JSD metric performs better than the other density-based (e.g. Hellinger) and hierarchy-based (e.g. WJL) measures. However, it seems that density-based metrics perform worse than hierarchy-based metrics when the number of topics is high. This could be due to the fact that topics have different levels of specificity and density-based metrics assume that all topics are equally descriptive, since they all have the same weight when measuring distance. The sum of distances of the less representative topics for JSD is higher as the number of topics diverge from its optimum number of topics (i.e between 100 and 300 topics in Table 3). However hierarchy-based metrics only take into account the most relevant topics, and this behavior not only makes them robust to dimensional changes in the models, but also

Acquis (MAP@10)

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

Table 3: Performance of density-based metrics.

Acquis-3 (MAP@10)

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

Table 4: MAP of density- and hierarchical-based distances from a corpus divided into three subsets.

seems to improve its accuracy for higher dimensions.

We can conclude that automatically generated annotations from topic models offer a knowledge close to that offered by categories manually assigned from the EuroVoc thesaurus in the Acquis legal corpus to relate texts. In the case of large and heterogeneous collections, i.e. with a high number of different topics, it would be more appropriate to annotate documents by topic hierarchies than using densities. In view of these results, the knowledge offered by topics allows automatically discovering what is being treated in a collection of documents, and the knowledge offered by its hierarchical representation allows understanding why documents are related in a similar way as it would be done with manually assigned labels.

4.2 Text Length Impact

To better understand how the length of the texts used for training affects the creation of probabilistic topics, we evaluated three different scenarios where the original corpus is divided into subsets with similar text sizes. In the first scenario we have created three equal sets and compared the performance using density-based metrics (i.e. JSD) and hierarchy-based metrics (i.e. WJL) for a document retrieval task. Table 4 shows the mean average precision when using a training set (columns) and a test set (rows) from among the 3 subsets into which the initial corpus was divided. The same experiment has been repeated in an analogous way for the scenarios with 6 (Table 5) and 9 (Table 6) subsets. Our aim is to analyze if there is any behavior that is common in all of them.

Models created from texts, i.e. training set, with greater or equal length to the texts used in the inferences, i.e. test set,

Acquis-6 (MAP@10)

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

Table 5: MAP of density- and hierarchical-based distances from a corpus divided into six subsets.

offered better performance in document retrieval tasks regardless of the language used. This behavior appears in the tables 4, 5 and 6 in the cells whose column is greater than or equal to its row. This is evidenced by the fact that those models performed better for almost all sets. Although for some evaluations of small documents models trained with large texts didn't yield the best performances they were not significantly different from the best models. For small documents both metrics performed similarly.

On the other hand, the performance of WJL significantly outperformed JSD for longer documents (i.e. higher columns). A remarkable case is the table 6. The results for the evaluation of the 9^{th} set (group with biggest document) with the 9^{th} model (trained with the biggest document set) were 13% better in the English case and 21% better, suggesting that, with enough text data, PTM models produce small variations in topic proportion vectors from which WJL metric benefits.

4.3 Time Required for Comparisons

The perception of efficiency on hierarchy-based metrics to calculate distances in topic models for large corpora was analyzed by capturing the computational time (in seconds) that each metric used to compare the documents (Fig.4). For almost all number of topics, hierarchical-based metrics are so much faster than probabilistic ones. However, for small number of dimensions (i.e. topics) the topic proportion vector is not sparse enough to identify any relevant topics from the uninformative ones, resulting in hierarchies containing all topics for every document. In other words, all documents share at least one topic. Increasing the number of topics alleviates this problem to the point of archiving an almost constant time for more than 35 topics. Although density-based metrics (e.g. JSD and HE) increased their computational time linearly with the representation size, JSD calculation requires computing two logarithms for each dimension in the document representation, which is way more time consuming than the HE metric square-roots. For the same reason, with a small number of dimensions, pairwise comparison is faster using the probabilistic metrics than the hierarchical metrics.

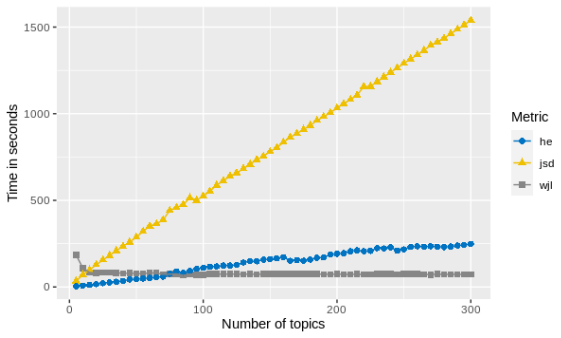


Figure 4: Time required to perform information retrieval tasks on a corpus of 100K documents described with different number of topics.

Acquis-9 (MAP@10)

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

Table 6: MAP of density- and hierarchical-based distances from a corpus divided into nine subsets.

5 Conclusions

In this paper we have studied the impact that the length of texts has, since they determine the space where words can co-occur, to semantically relate documents described in a probabilistic topic space. We have also studied the ability of probabilistic topics to automatically cluster related texts, and the performance of density-based and topic hierarchy-based distance measures. Multiple document retrieval tests were performed on a collection of legal documents, comparing the results obtained by this unsupervised approach, with the results obtained using manual annotations. Representation methods based on probabilistic topics have proven to be reasonably accurate in annotating semantically related documents with the same categories. State-of-the-art metrics based on densities and hierarchical representations of topics were evaluated to measure document similarity. Regardless of the approach used, the knowledge captured by word distributions (i.e topics) to automatically relate legal texts has shown an accuracy close to 0.8 compared to relations based on EuroVoc categories.

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

If we consider that the complexity of a text is directly proportional to its length, probabilistic models are not able to generalize the knowledge they acquire during their training to process more complex texts. In other words, the knowledge captured by probabilistic topics to group semantically related documents can only be applied to texts of equal or lesser length than those used during training.

In addition, the larger the corpus and the more topics it contains (i.e. the more diverse the content of its documents), the more appropriate it is to use similarity metrics based on hierarchical representations of the topics (see Figures 5 and 6). The noise introduced by the less present topics in a text is adverse to density-based metrics. The relationships suggested when manually annotating documents are therefore based on a small group of labels. Under these conditions, PTM can guide the corpus exploration by providing an unsupervised method to thematically annotate documents and potentially giving insight of the relations between documents.

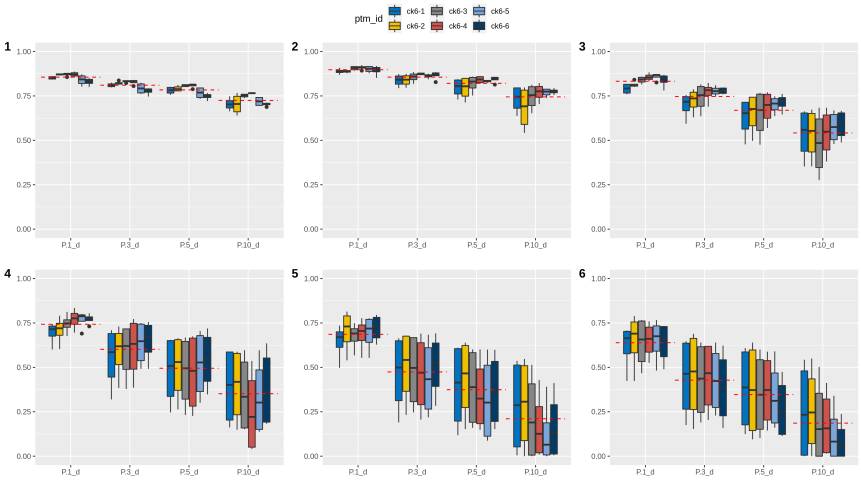


Figure 5: Test comparisons in 6 partitions based on JSD.

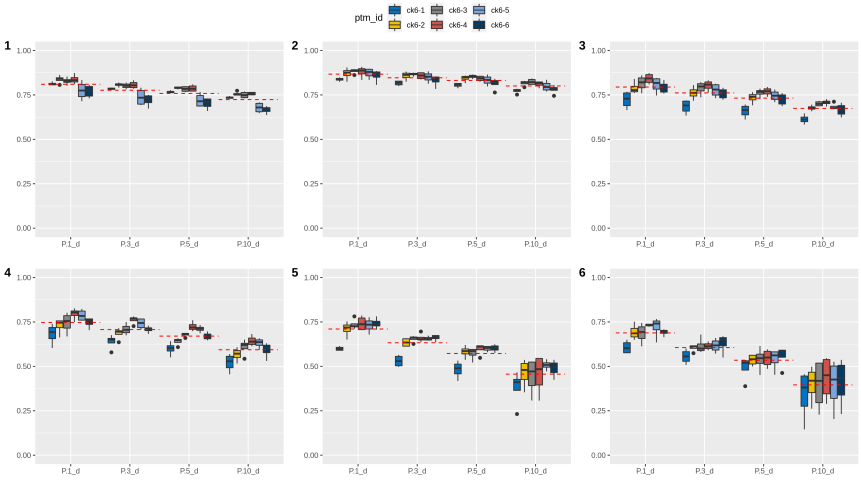


Figure 6: Test comparisons in 6 partitions based on WJL.

There are still challenges and questions that will have to be solved in future work, namely finding the influence of the weights in the hierarchical-based metric; analyzing the complexity of texts beyond their lengths, taking into account the rhetoric of its discourse to represent scientific texts (.e.g. using only the paragraphs describing the approach or the method to create the topic distributions); And even observing their behaviour in different languages.

Acknowledgments

This work is supported by the project *KnowledgeSpaces* with reference PID2020-118274RB-I00, financed by the Spanish Ministry of Science and Innovation.

References

Badenes-Olmedo, C., J. Redondo-García, and O. Corcho. 2019a. Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms. *Semantic Web Journal*.

Badenes-Olmedo, C., J. Redondo-García, and O. Corcho. 2019b. Legal document retrieval across languages: topic hierarchies based on synsets. *Proceedings of the 1st Workshop on Iberlegal co-located with 32nd International Conference on Legal Knowledge and Information Systems organized by the Foundation for Legal Knowledge Based Systems (JURIX)*.

Badenes-Olmedo, C., J. L. Redondo-García, and O. Corcho. 2017a. Distributing text mining tasks with librairy. In *DocEng 2017 - Proceedings of the 2017 ACM Sym-*

posium on Document Engineering, pages 63–66, August.

Badenes-Olmedo, C., J. L. Redondo-García, and O. Corcho. 2017b. An initial analysis of topic-based similarity among scientific documents based on their rhetorical discourse parts. In *Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci)*, pages 15–22.

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

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

Dieng, A. B., F. Ruiz, and D. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

He, J., L. Li, and X. Wu. 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.

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

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Hu, Y., K. Zhai, V. Eidelman, and J. Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176.

Jelodar, H., Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. 2017. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey.

Jung, K. H., E. Ruthruff, and T. Goldsmith. 2017. Document similarity misjudgment by lsa: Misses vs. false positives. *Cognitive Science*.

Mao, X.-L., B.-S. Feng, Y.-J. Hao, L. Nie, H. Huang, and G. Wen. 2017. S2JSD-LSH: a locality-sensitive hashing schema for probability distributions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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

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

Rus, V., N. Niraula, and R. Banjade. 2013. Similarity measures based on latent dirichlet allocation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 459–470. Springer.

Schofield, A., M. Magnusson, and D. Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

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

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *The 5th International Conference on Language Resources and Evaluation - Proceedings p. 2142-2147*, May.

Syed, S. and M. R. Spruit. 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174.