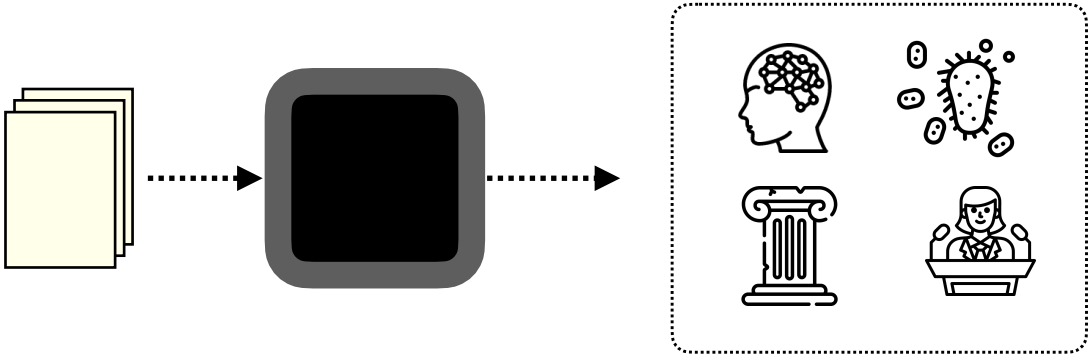
Hierarchical representations of topics to uncover the underlying knowledge of semantically related texts

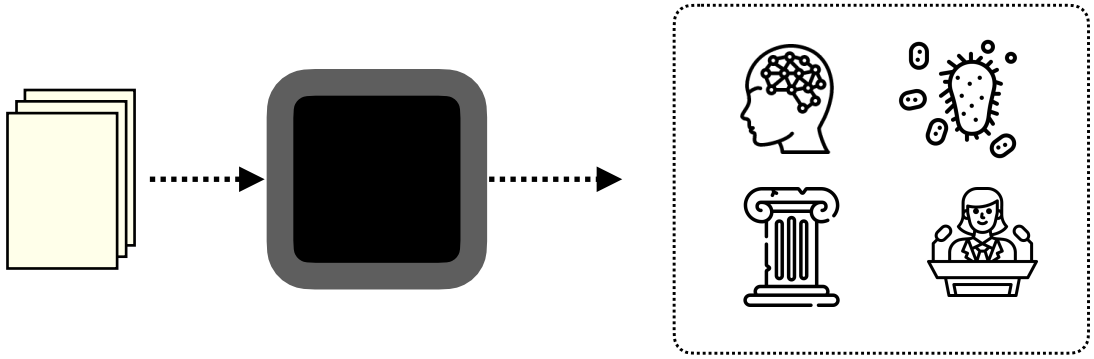
Borja Lozano Álvarez, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Carlos Badenes

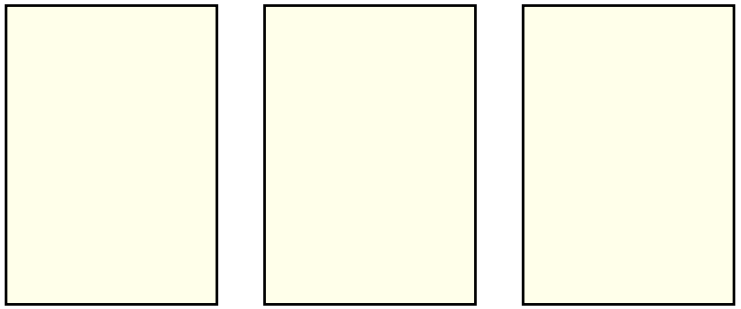
borja.lozano.alvarez@alumnos.upm.es

Topic Models

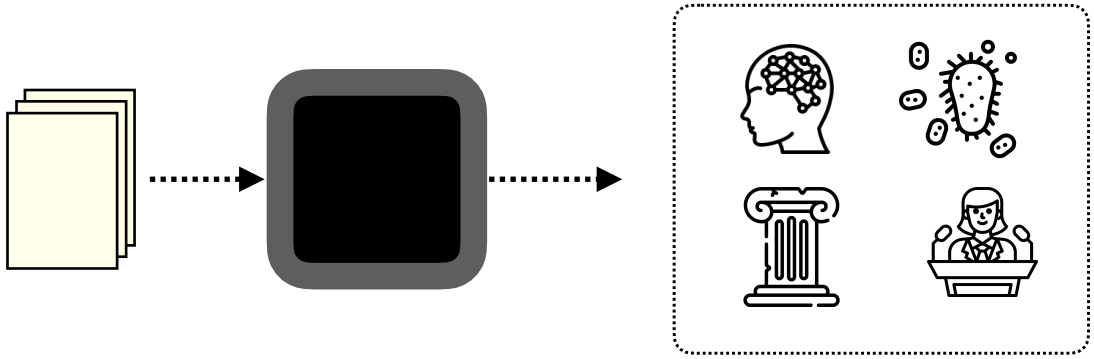


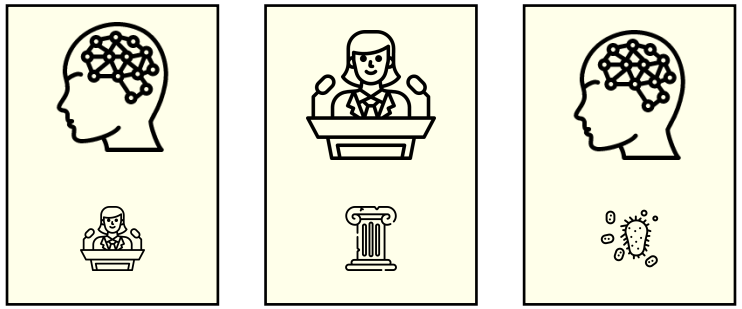
Topic Models



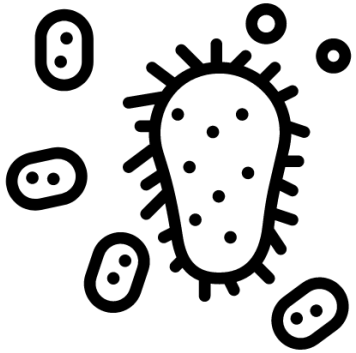


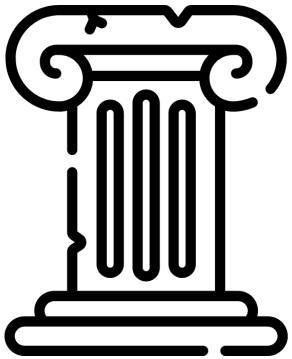
Topic Models

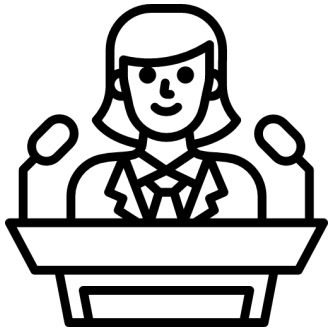




Topics







Topics

In Probabilistic Topic Models a **topic** is a multinomial distribution over the vocabulary

Topics

In Probabilistic Topic Models a **topic** is a multinomial distribution over the vocabulary

	People	Cognitive	President	Neuron	Ballot
	0.08	0.4	0.01	0.5	0.01
	0.25	0.03	0.35	0.07	0.2
	0.2	0.01	0.48	0.01	0.3

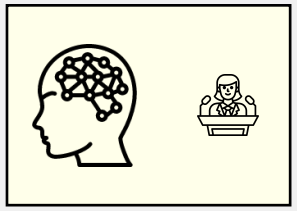
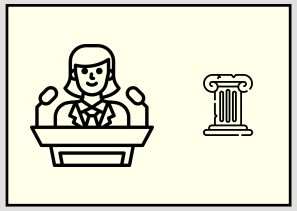
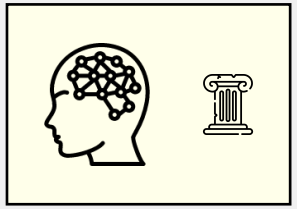
Topics

In Probabilistic Topic Models a **topic** is a multinomial distribution over the vocabulary

	People	Cognitive	President	Neuron	Ballot
topic_0	0.08	0.4	0.01	0.5	0.01
topic_1	0.25	0.03	0.35	0.07	0.2
topic_2	0.2	0.01	0.48	0.01	0.3

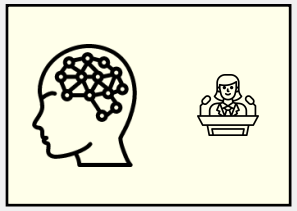
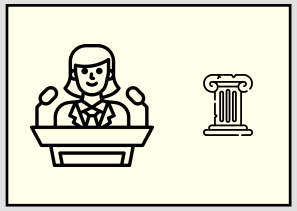
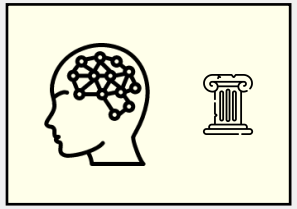
Documents

In Probabilistic Topic Models a **<u>document</u>** is a multinomial distribution over the topics

			
	0.9	0.01	0.09
	0.01	0.2	0.79
	0.8	0.05	0.15

Documents

In Probabilistic Topic Models a **<u>document</u>** is a multinomial distribution over the topics

	topic_0	topic_1	topic_2
	0.9	0.01	0.09
	0.01	0.2	0.79
	0.8	0.05	0.15

Comparing documents in PTM

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

Comparing documents in PTM

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

Problems

Comparing documents in PTM

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

Problems

Pairwise computation of document similarity is costly and grows linearly with the size of the corpus.

Comparing documents in PTM

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

Problems

Probability metrics do not offer a semantic explanation for the similarity obtained.

Comparing documents in PTM

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

Problems

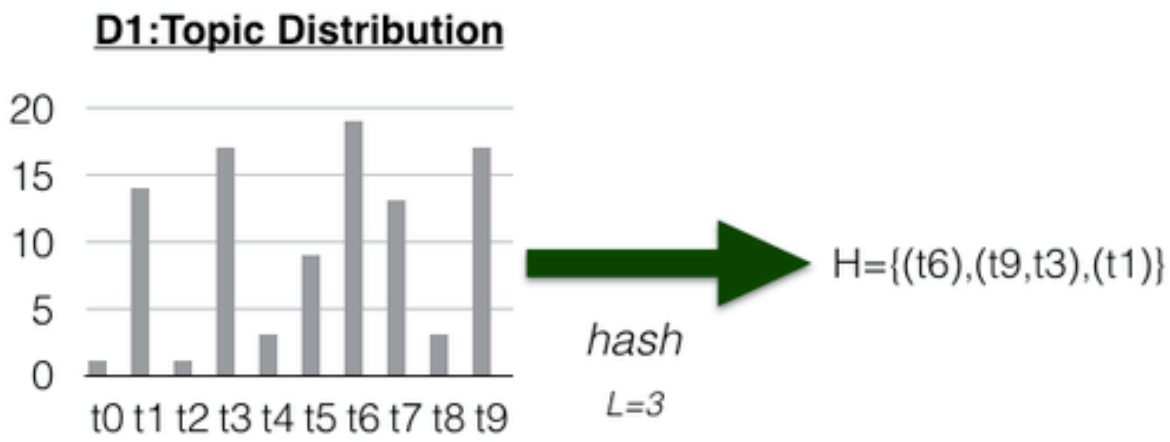
These metrics cannot be extended to support semantic restrictions to enrich queries in the corpus.

Comparing documents in PTM

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

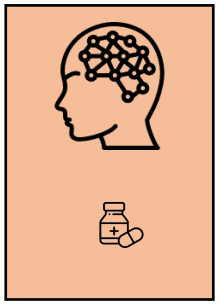
Problems

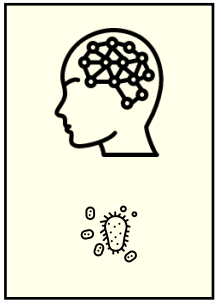
Hierarchical representations of topics

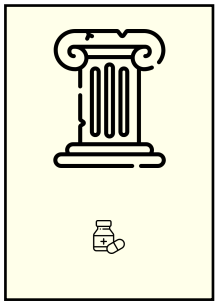


Source: Badenes-Olmedo, C., Redondo-García, J., Corcho, O.: **Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms.**

Hierarchical representations of topics

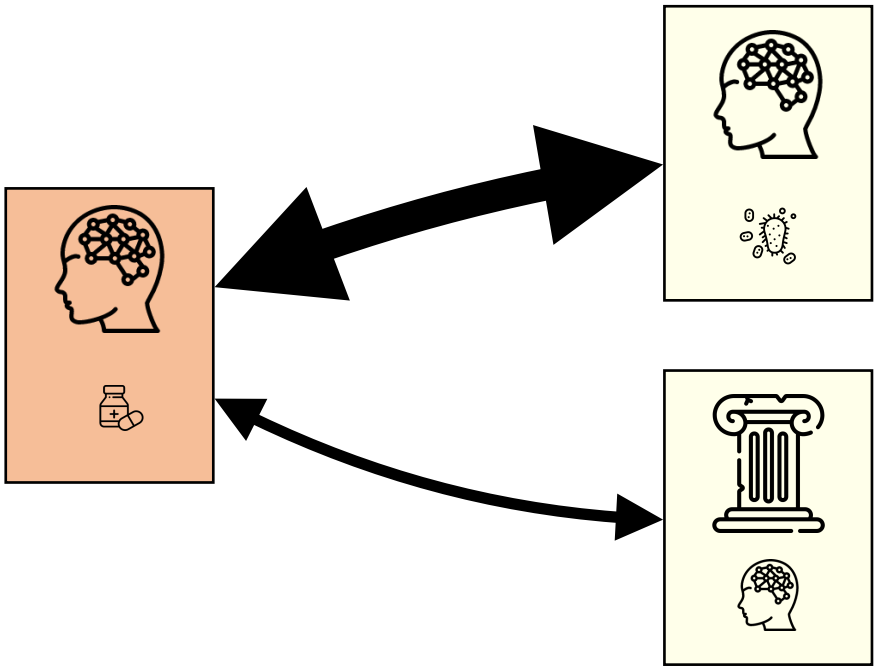






Hierarchical representations of topics

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



Metric Comparison & Vocabulary Size

- How well does the hierarchical metrics perform in comparison to state-of-the-art metrics, both in performance and accuracy
- How does the size (in number of tokens) of the documents affect the document similarity task when using PTM

DATA

EuroVoc Thesaurus

- Multilingual thesaurus with a taxonomy of 7,193 concepts/labels from 21 domains
- The concepts from which all categories derive leaved us with 452 root concepts *
- Two documents are relevant if they share categories

*Source: Badenes-Olmedo, C., Redondo-García, J.L., Corcho, O.: **Scalable Cross-Lingual Document Similarity through Language-Specific Concept Hierarchies**

DATA

Acquis Corpus

- Corpus of the Official Journey of the EU constructed by merging the JRC Acquis (manually annotated) and the DGT Acquis

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

Table 1: Number of documents and tokens by dataset

EVALUATION

Ground truth

- To evaluate a PTM with a test collection:
  - Create ground truth by creating a **relevant** list for each document.
  - Create a **retrieved** list by pairwise comparison with the PTM representation
  - Compare both list for all documents in the test-collection to obtain MAP@10

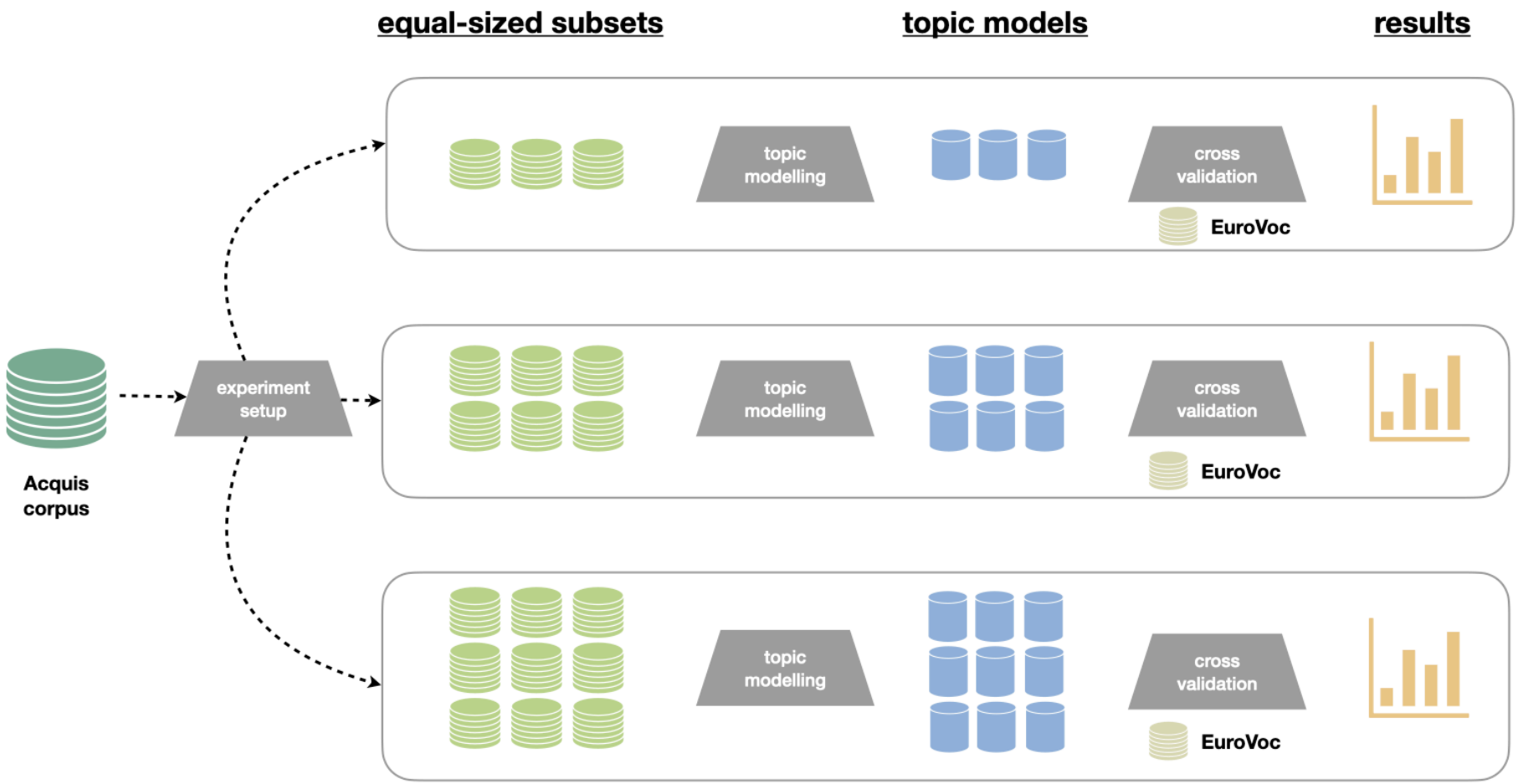
Results

Materialisation of Knowledge in Topics

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**

Methodology



Results

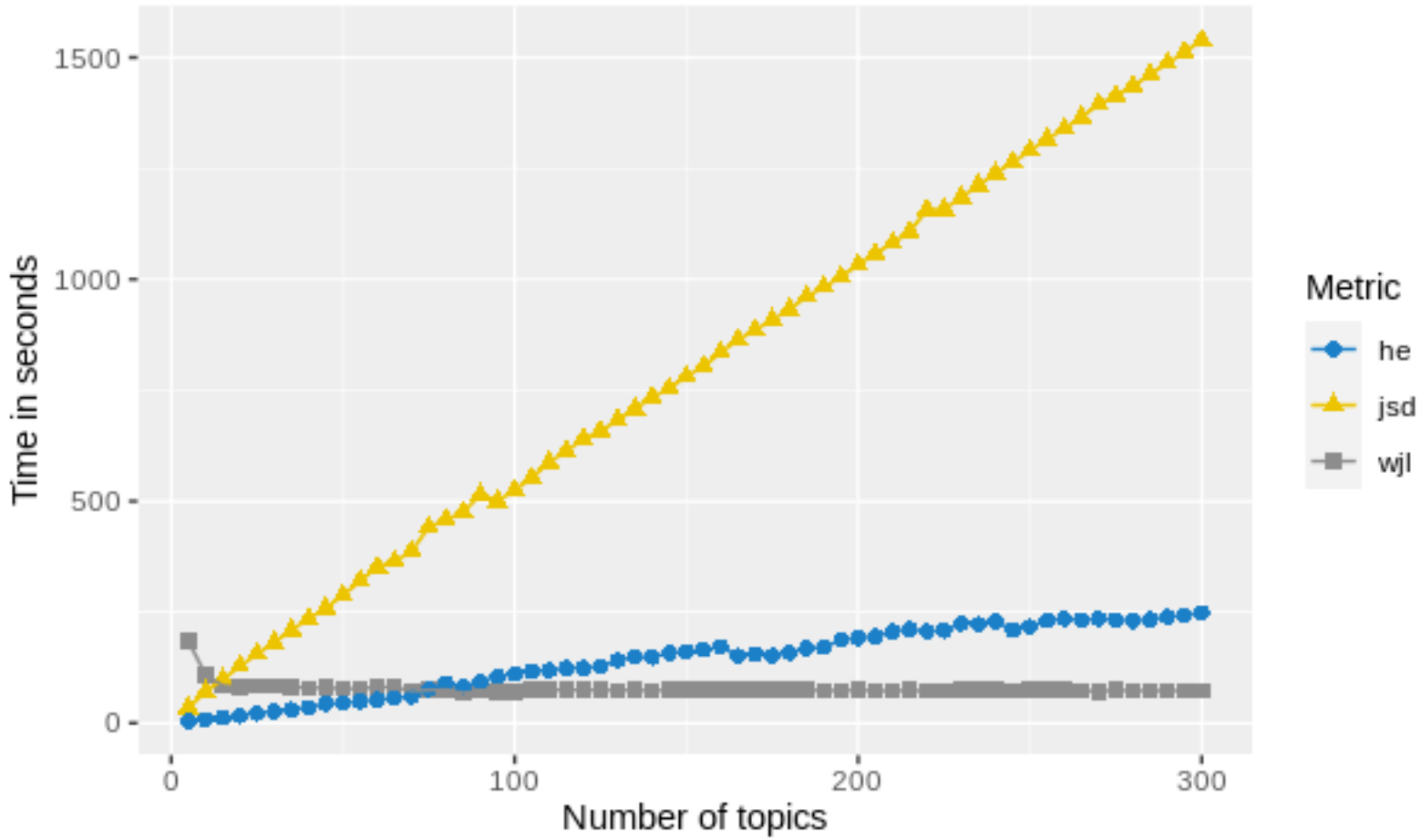
Influence of Text Length

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**

Results

Computational time



Conclusion

- If we assuming 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
- 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

Hierarchical representations of topics to uncover the underlying knowledge of semantically related texts

Borja Lozano Álvarez, Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
Carlos Badenes

borja.lozano.alvarez@alumnos.upm.es