

# Determination of the Number of Topics Intrinsically: Is It Possible?

Victor Bulatov<sup>1</sup>, **Vasiliy Alekseev**<sup>1</sup>, and Konstantin Vorontsov<sup>2</sup>

<sup>1</sup> Moscow Institute of Physics and Technology, <sup>2</sup> Lomonosov Moscow State University

The 66th MIPT All-Russian Scientific Conference

6 April, 2024

He tethered his horse, which had begun to shiver; fed it; and threw a light blanket over its hindquarters against the chill. He kindled a small fire and prepared a meal, then sat down to wait out the mist, taking up the eastern gourd and composing to its eery metallic tones a chanted lament. The mist coiled around him, sent cold, probing fingers into his meagre shelter. His words fell into the silence like stones into the absolute abyss: 'Strong visions: I have strong visions of this place in the empty times... Far below there are wavering pines... I left the rowan elphin woods to fulminate on ancient headlands, dipping slowly into the glazen seas of evening...'

*(The Pastel City by M. John Harrison)*

He tethered his horse, which had begun to shiver; fed it; and threw a light blanket over its hindquarters against the chill. He kindled a small fire and prepared a meal, then sat down to wait out the mist, taking up the eastern gourd and composing to its eery metallic tones a chanted lament. The mist coiled around him, sent cold, probing fingers into his meagre shelter. His words fell into the silence like stones into the absolute abyss: 'Strong visions: I have strong visions of this place in the empty times... Far below there are wavering pines... I left the rowan elphin woods to fulminate on ancient headlands, dipping slowly into the glasen seas of evening...'

*(The Pastel City by M. John Harrison)*

### Animals

cat  
dog  
horse  
wolf  
hay  
fish-hawk  
hindquarters

### Autumn

cold  
mist  
chill  
evening  
shiver  
metallic  
glasen

### Camp

fire  
warm  
shelter  
safe  
meal  
kindle  
gourd

### Music

song  
guitar  
tone  
string  
compose  
chant  
ballad

### Horror

eery  
lament  
abyss  
silence  
empty  
meagre  
rustle

### Mystery

fate  
ancient  
artifact  
cards  
forget  
vision  
elphin

### Nature

wood  
sky  
sea  
pine  
stone  
rowan  
headland



# Topic Modeling

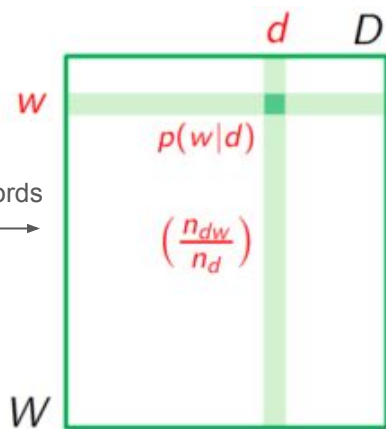
Topic modelling assumes that there are a number of *latent topics* which explain the text collection.

Take some  $T$  (num topics)

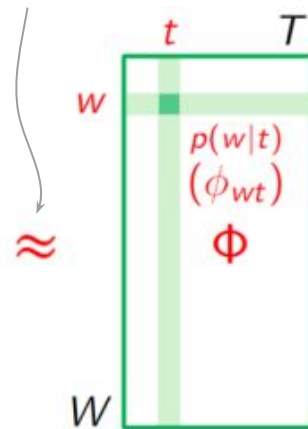


Text collection  
 $D$  (num docs),  $W$  (vocab size)

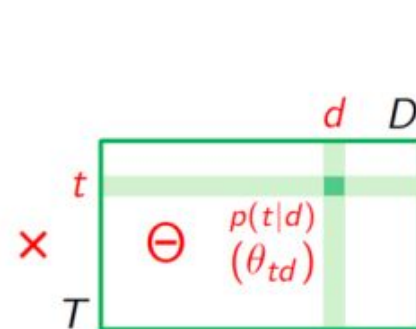
Bag of words



Matrix of word-in-document  
relative frequencies



Matrix of word-in-topic  
probabilities



Matrix of topic-in-document  
probabilities

$\approx$

$\times$

Input

Output

# Topic Modeling

Topic modelling assumes that there are a number of *latent topics* which explain the text collection.

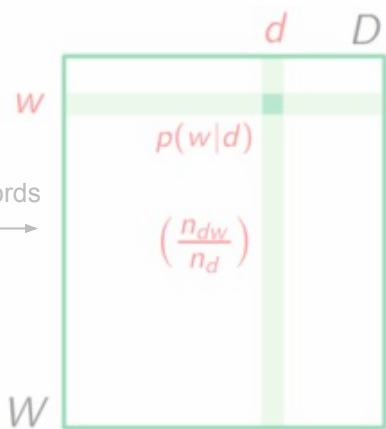
Take *what* T (num topics)?

Er zwingt sich über dann, diesen Geruch für sich zu ertragen, plötzlich, irgendwo. Es gelingt ihm fast immer. Wie einem Meister plötzlich bei und wieder ein Meistertisch gelingt. Seine ganze Kindheit sei aus Geruchlichkeit zusammengesetzt, zusammengesetzt haben sie sich zu seiner Kindheit. Nicht vor sei es, ständig in Bewegung. Und aus Worten und Halbspielen, aus der Angst vor Ungewissen, solchen Tieren, bunten Gesetzen, tiefen Flüssen, Hunger, Zukunft. Er hat in seiner Kindheit Ungewissen, Hunger, wilde Tiere und stoffende Flüsse kennengelernt. Auch Zukunft, Abscheu. Der Krieg hat ihm ermöglicht zu wissen, was Leute, die den Krieg nicht kennen, niemals sehen. Die Geduld, wasche in seinem Leben sei ein dem Land ab, dass von Großvater war unruhig, genauso unruhig wie er selbst. Die Großmutter geruchlich, natürlich, unangenehm für gewisse Menschen. Der Großvater sah ein dem Fabel mit in Land, schafte, in Gespräche, in Finsternisse hinein. »Herrn« mündlich waren die Gefährten, sagt er, der Vorleser war sein allergrößter Verlust. Die Eltern kümmerten sich wenig um ihn, mehr um den ein Jahr älteren Bruder, von dem sie alles erwarteten, was sie von ihm nicht erwarteten: eine glückliche Zukunft, übertrage Zukunft. Matrie Liebe und mehr Taschengeld hatte von Bruder immer bekommen. Wie sie erwarteten, erwartete sein Bruder sie nie. Mit einem Schwester verband ihn ein viel zu schwachen Band, als dass es halten könnte. Später klang es so wie über dem Ozean hinweg, schweben sich Botschaften, von Europa nach Mexiko, von Mexiko nach Europa, versuchten aus ihrer Verbundenheit eine Liebe zu machen, eine Abhängigkeit, was ihnen vielleicht auch gelang. »Sie schwelte mit immer, diesmal im Jahr, so oft wie ich hier, sagte er. Aus dem Moment und tief in den Gedanken die seine Gedanken, die immer dünner wurden. Mit dem Tod der Großeltern ging es in Finsternis, die nicht mehr aufhören wird...

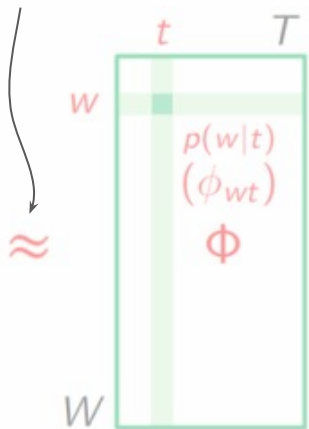
Dass stach auch der Vater, die Mutter folgte ihm ein Jahr

Text collection  
D (num docs), W (vocab size)

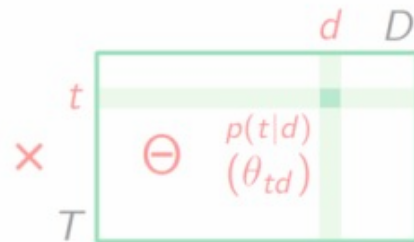
Bag of words



Matrix of word-in-document relative frequencies



Matrix of word-in-topic probabilities



Matrix of topic-in-document probabilities

Input

Output

# Number of Topics



Art



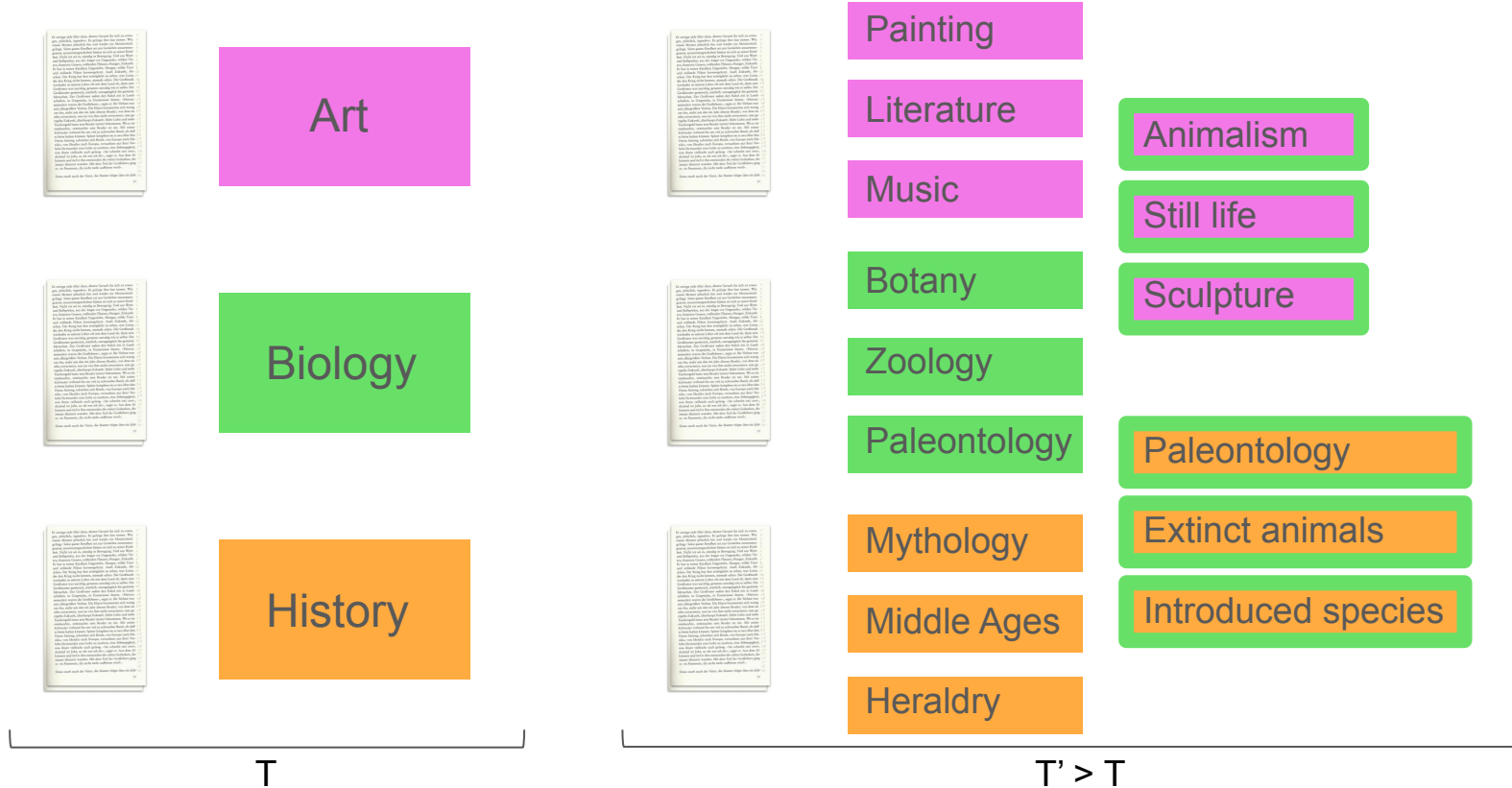
Biology



History

T

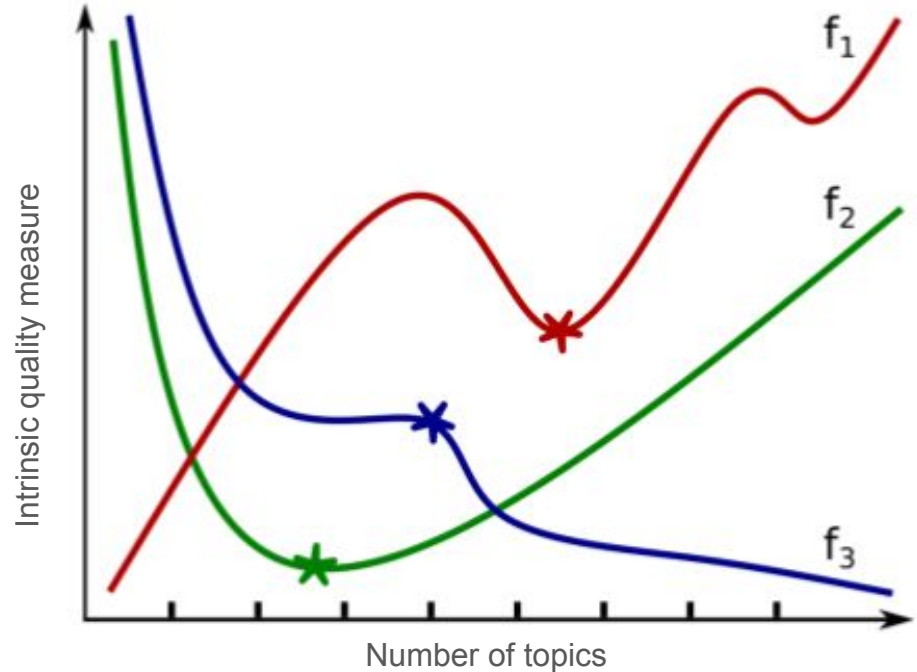
# Number of Topics?



# Determination of the Number of Topics Intrinsically

**Purpose:** find out if intrinsic model quality criteria help in determining the number of topics.

**Solution:** train models with different number of topics and select the optimal number as corresponding to the best quality.



Expected possible dependencies of the intrinsic quality criterion on the number of topics.



# Related Work

- ldatuning

*Perplexity, topic diversity for LDA.*

- TOM

*Topic diversity, topic model stability.*

- OCTIS

*Topic diversity, coherence but without determining the number of topics.*

Nikita M., Chaney N. [Ldatuning: Tuning of the latent dirichlet allocation models parameters](#). – 2016. ([github](#))

Guille A., Soriano-Morales E. P. [TOM: A library for topic modeling and browsing](#). – 2016. ([github](#))

Terragni S. et al. [OCTIS: Comparing and optimizing topic models is simple!](#) – 2021. ([github](#))

# Intrinsic Quality Measures

- **Perplexity** (↓)

Measure of model's “surprise” when it sees text.

- **Diversity and sufficiency** (*D-avg-COS*, *D-Spectral*; ↑)

If the number of topics is too large, the model produces a lot of small similar topics.

- **Clustering** (*SilhC*, *CHI*; ↑)

How similar an object (word) is to its own cluster (topic) compared to other clusters.

- **Stability** (↓)

Models with the “incorrect” number of topics are unstable (differ from each other).

# Intrinsic Quality Measures

- **Information-theoretic** (*AIC*, *BIC*, *MDL*; ↓)

Balance between model complexity and the goodness of fit (“model complexity minus model likelihood”).

- **Entropy** (*Rényi*; ↓)

“Correct” number of particle states (word topics) should correspond to the equilibrium state, which is characterised by the minimum of entropy.

- **Top-tokens**

Nonrandomness (*Coherence*; ↑) and specificity (*Lift*; ↑) of topic top words.

# Methodology

FOR EACH dataset:

FOR EACH topic\_model:

FOR EACH random\_seed:

FOR EACH t FROM t\_min(dataset) TO t\_max(dataset):

init(topic\_model, random\_seed)

train(topic\_model, t)

quality = eval(topic\_model)

draw\_on\_plot(t, quality)

t\_opt = analyze\_plot() # search for pronounced min/max

# Models

- **PLSA**: a simple topic model without any hyperparameters aside from  $T$ .
- **LDA**: a well-known topic model, having priors for  $\Phi$  and  $\Theta$  distributions.
- **Decorrelated** (ARTM): attempts to reduce pairwise topic correlations.
- **Sparse** (ARTM): divides its topics into background and specific (sparse).
- **Sparse decorrelated** (ARTM): sparse and decorrelated simultaneously.

Hofmann, T. [Probabilistic latent semantic analysis](#). – 1999.

Blei D. M., Ng A. Y., Jordan M. I. [Latent dirichlet allocation](#). – 2003.

Vorontsov K. et al. [Bigartm: Open source library for regularized multimodal topic modeling of large collections](#) //AIST 2015.

# Datasets

Dataset	$D$	$W$	$T_{\text{expected}}$	$T_{\text{min}}$	$T_{\text{max}}$
WikiRef220	220	4839	5	2	20
20NG	18846	2174	15–20	3	40
Reuters	10788	5074	90	5	150
Brown	500	7409	10–20	5	25
StackOverflow	895621	3430	40	5	60
PostNauka	3404	8417	15–30	5	50
ruwiki-good	8603	236018	10/90	5	100

$D$  — number of documents,  $W$  — size of vocabulary,  $T$  — number of topics (expected  $T$ , and min/max values to be used in the experiments). Preprocessing: lemmatization, stop-words removal.

# Results

Three features to summarize the behaviour of metrics:

- **Jaccard**: independence of the result from model random initialization (↓)
- **Informativity**: readability of obtained plots (↑)
- **Expected**: precision of the metric providing an expected number of topics (↑)

According to the results, *the number of topics is not a well-defined property of a particular corpus.*

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
Diversity	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
Clustering	D-avg-JH	0.302	0.053	0.022
	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
average coherence	0.780	0.472	0.208	
uni-theta-divergence	0.470	0.197	0.047	

# Results

Three features to summarize the behaviour of metrics:

- **Jaccard**: independence of the result from model random initialization (↓)
- **Informativity**: readability of obtained plots (↑)
- **Expected**: precision of the metric providing an expected number of topics (↑)

According to the results, *the number of topics is not a well-defined property of a particular corpus.*

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
Diversity	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
Clustering	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
	D-avg-JH	0.302	0.053	0.022
Clustering	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
	average coherence	0.780	0.472	0.208
uni-theta-divergence	0.470	0.197	0.047	



# Results

Three features to summarize the behaviour of metrics:

- **Jaccard**: independence of the result from model random initialization (↓)
- **Informativity**: readability of obtained plots (↑)
- **Expected**: precision of the metric providing an expected number of topics (↑)

According to the results, *the number of topics is not a well-defined property of a particular corpus.*

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
Diversity	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
Clustering	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
	D-avg-JH	0.302	0.053	0.022
Clustering	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
	average coherence	0.780	0.472	0.208
uni-theta-divergence	0.470	0.197	0.047	

# Results

Three features to summarize the behaviour of metrics:

- **Jaccard**: independence of the result from model random initialization (↓)
- **Informativity**: readability of obtained plots (↑)
- **Expected**: precision of the metric providing an expected number of topics (↑)

According to the results, *the number of topics is not a well-defined property of a particular corpus.*

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
Diversity	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
Clustering	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
	D-avg-JH	0.302	0.053	0.022
Clustering	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
	average coherence	0.780	0.472	0.208
uni-theta-divergence	0.470	0.197	0.047	

# Results

Three features to summarize the behaviour of metrics:

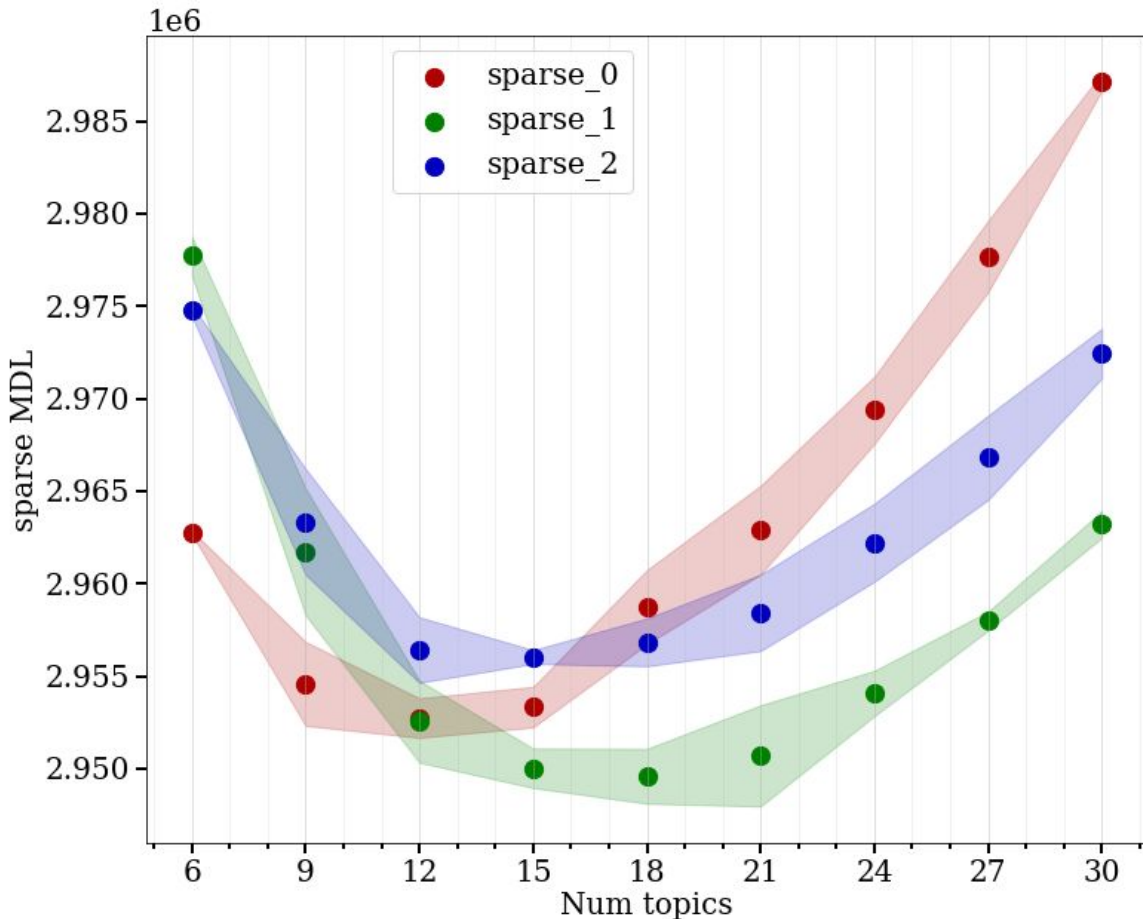
- **Jaccard**: independence of the result from model random initialization (↓)
- **Informativity**: readability of obtained plots (↑)
- **Expected**: precision of the metric providing an expected number of topics (↑)

According to the results, *the number of topics is not a well-defined property of a particular corpus.*

	Score	Jaccard	Informativity	Expected
Information-theoretic	AIC	0.280	0.542	0.578
	AIC sparse	0.219	0.111	0.100
	BIC	0.128	0.444	0.461
	BIC sparse	0.274	0.164	0.128
	MDL	0.096	0.488	0.414
	MDL sparse	0.282	0.428	0.256
Diversity	renyi-0.5	0.470	0.507	0.425
	renyi-1	0.356	0.475	0.394
	renyi-2	0.230	0.299	0.183
	D-Spectral	0.456	0.144	0.083
	D-avg-L2	0.682	0.250	0.119
	D-cls-H	0.595	0.245	0.189
Clustering	D-avg-JH	0.302	0.053	0.022
	lift	0.383	0.123	0.033
	holdout-perplexity	0.228	0.025	0.019
	perplexity	0.218	0.023	0.014
	CHI	0.277	0.157	0.008
	SilhC	0.233	0.079	0.028
average coherence	0.780	0.472	0.208	
uni-theta-divergence	0.470	0.197	0.047	

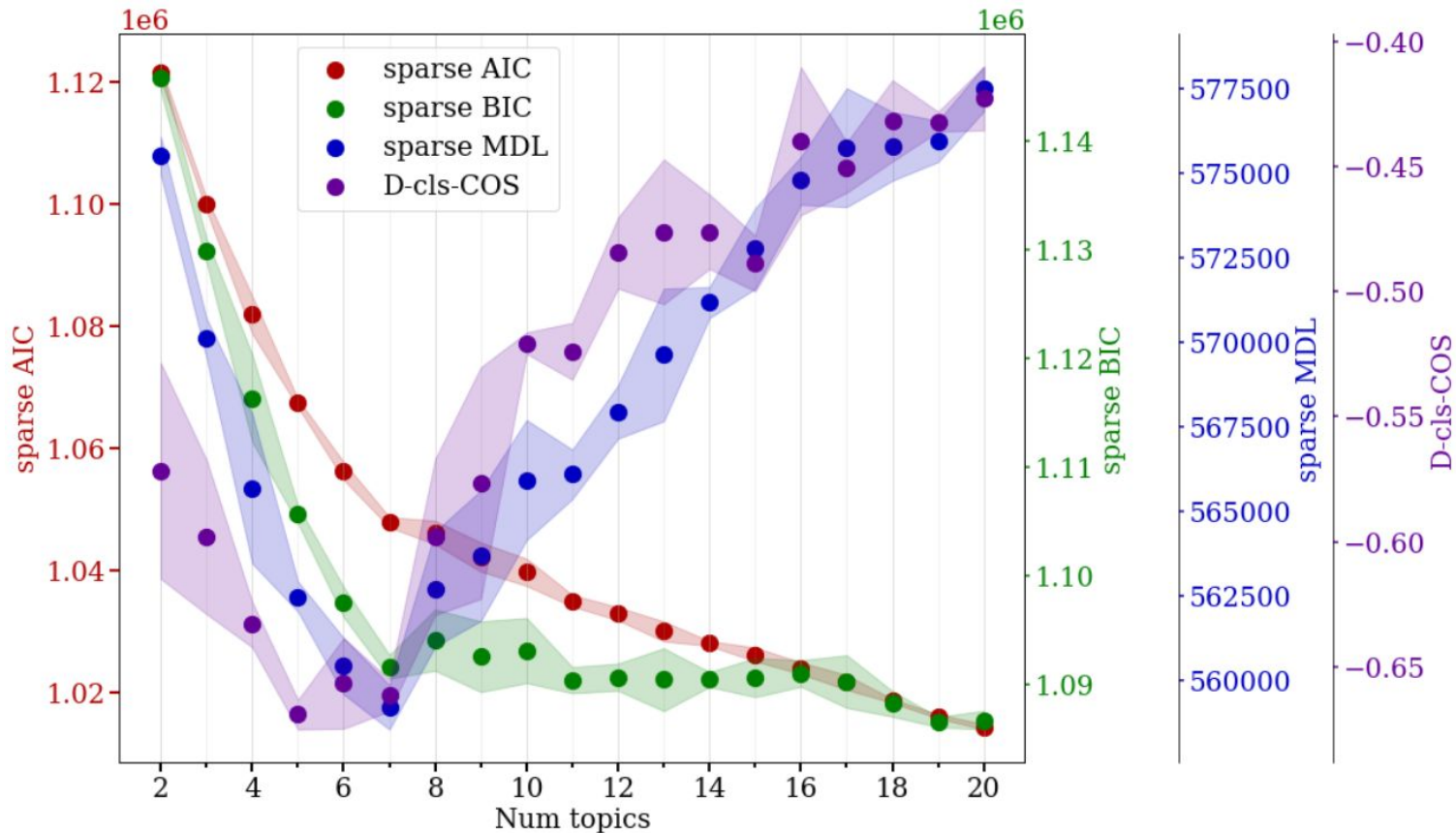
# Results

- Optimal number of topics depends on the model.
- Randomness causes variance.



Sparse MDL criterion for models with different sparsity hyperparameter values (WikiRef220).

# Results



Different criteria often do not agree with each other (but sometimes they do). A set of quality metrics exploring various T for PLSA (WikiRef220). Cosine-based diversity is taken with a negative sign. All metrics agree with 7 being a reasonable value for T.

# Conclusion

- Number of topics is a method- and a model-dependent quantity.
- Number of topics is not an absolute property of a particular corpus.
- Perplexity is not helpful for finding the number of topics.
- Simplest approaches (AIC, BIC, MDL; Rényi) achieve best results.

Recommendations (based on evidence):

- Examine several related measures.
- Information-theoretic methods (AIC, BIC, MDL) are better employed in conjunction.

Recommendations (based on reflections on the topic):

- Select a model according to a secondary task.
- Build a hierarchy of topics and prune it afterwards.
- Utilize the process of human (semi-) supervision.

*The main purpose of topic modeling should be the search for such a method of model training which, given the number of topics, results in a model whose topics in the absence of external criterion are all interpretable.*