

Intra-Text Coherence as a Measure of Topic Models' Interpretability

Vasiliy Alekseev, Victor Bulatov, Konstantin Vorontsov

24rd International Conference on Computational Linguistics and
Intellectual Technologies

1 June 2018



DIALOGUE



Topic, Its Interpretability & Coherence

Topic is a set of words that often occur together in text.

Interpretability of the topic means that a human is able to explain the meaning behind its set of words. However, such human assessment is expensive.

Well Interpreted Topic (Most Frequent Terms)

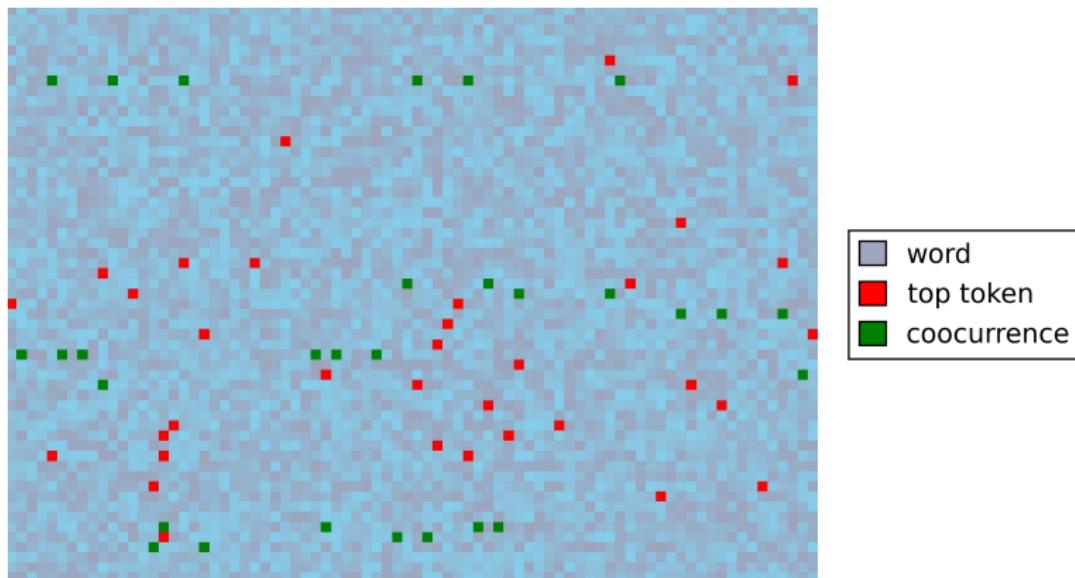
actor, play, musical, premiere, parterre, spectator, producer, audience, backstage, orchestra

Badly Interpreted Topic (Most Frequent Terms)

express, epigraph, foggy, result, image, right, loan, debt, bankrupt, interest

Coherence is an automated method for estimating interpretability, which measures how often 10 most probable terms of the topic occur in close proximity within text.

Purpose of the Study



Problem

Ten most frequent words cover only a small proportion of text.
Co-occurrences are even fewer. Can one rely on the analysis of
these co-occurrences?

Problem

Top-tokens based coherences rely on the list of top ten words. These words reflect only a small part of the whole topic model.

Solution

Evaluate coherence as an average thematic proximity of words closely located in text.

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

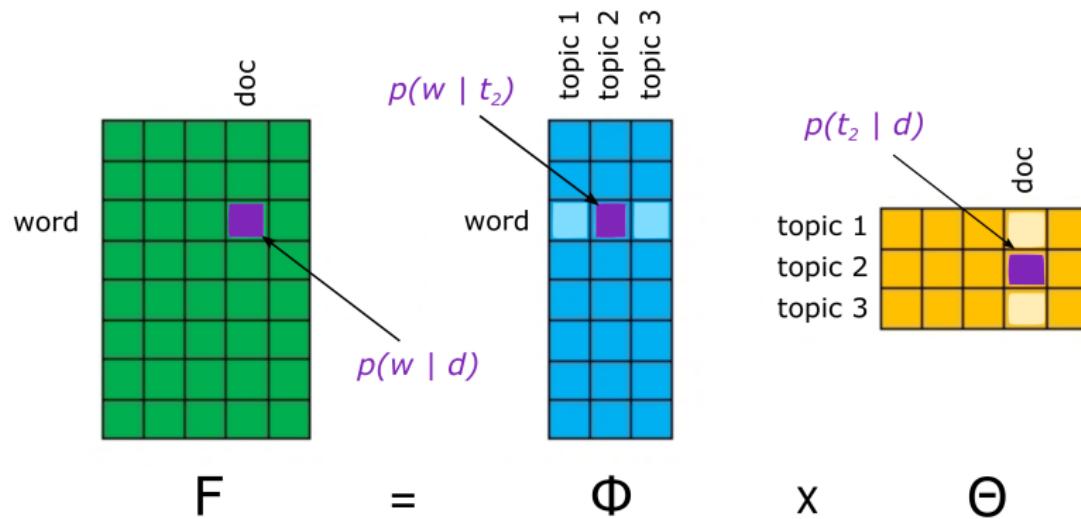
- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Topic Modeling and Matrix Factorization

Topic model describes documents using *latent* topics

- $\varphi_{wt} \equiv p(w | t)$ – how often word w appears in topic t
- $\theta_{td} \equiv p(t | d)$ – probability of topic t within document d



Original Dataset: Three Top-Words of Each of 19 Topics

2000 *monotopic* PostNauka¹ articles. Topics found with BigARTM²

Topic	First Top-Word	Second Top-Word	Third Top-Word
1: математика	математика (0.016)	задача (0.008)	декарт (0.008)
2: технологии	технология (0.015)	робот (0.012)	сеть (0.010)
3: физика	частица (0.027)	электрон (0.015)	кварк (0.015)
4: химия	химия (0.021)	молекула (0.019)	материал (0.016)
5: земля	земля (0.029)	планета (0.028)	атмосфера (0.012)
6: астрономия	звезда (0.039)	галактика (0.031)	вселенная (0.019)
7: биология	клетка (0.027)	организм (0.011)	мозг (0.010)
8: медицина	пациент (0.016)	препарат (0.012)	заболевание (0.012)
9: психология	психология (0.009)	мозг (0.009)	психолог (0.008)
10: экономика	экономика (0.016)	страна (0.010)	цена (0.008)
11: история	история (0.010)	историк (0.007)	власть (0.006)
12: политика	государство (0.014)	политика (0.012)	политический (0.011)
13: социология	социология (0.013)	социолог (0.009)	социальный (0.008)
14: культура	культура (0.015)	фильм (0.007)	искусство (0.006)
15: образование	университет (0.021)	образование (0.014)	школа (0.013)
16: язык	язык (0.077)	слово (0.037)	словарь (0.011)
17: философия	философия (0.018)	философ (0.013)	философский (0.008)
18: религия	святынище (0.010)	религия (0.007)	царь (0.006)
19: россия	россия (0.028)	страна (0.009)	русский (0.009)

¹<https://postnauka.ru>

²Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections, 2015

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

- Coherence $\Big|_t = \text{Average}_{w_i, w_j \text{ from } k \text{ top-words}} \text{PMI}(w_i, w_j)$

- Newman³: $\text{PMI}(w_i, w_j) = \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$

- Mimno⁴ : $\text{PMI}(w_i, w_j) = \ln \frac{D(w_i, w_j) + 1}{D(w_i)}$

- $p(w_i), p(w_i, w_j)$ – probability to find word w_i and two words w_i, w_j in a context window of given size
- $D(w_i), D(w_i, w_j)$ – number of documents containing word w_i and two words w_i, w_j in a context window of given size

³Newman et al. Automatic Evaluation of Topic Coherence, 2010

⁴Mimno et al. Optimizing Semantic Coherence in Topic Models, 2011

Drawbacks of Top-Tokens Based Approach

Top token-based coherences may ignore more than 98% of words of the documents' collection.

	PostNauka, %	Wikipedia, %
Minimum	0.016	0.0065
Median	0.048	0.029
Mean	0.062	0.036
Maximum	0.28	0.11
Total	1.2	1.7

The proportion of corpus contributing to the co-occurrence counts of top 10 most frequent words for each topic

Drawbacks of Top-Tokens Based Approach

A single top token "частиц" out of the first 10 ones is seen.
The wide range of less strong topical words is ignored by the top-tokens based coherences.

Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of topic 3: физика with top 10 in bold: частица, электрон, кварк, атом, энергия, вселенная, фотон, физика, физик, эксперимент, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Meaning

Semantic proximity of closely located words in text

	Группа	астрономов	обнаружила	звезды
Astronomy	$\begin{pmatrix} \text{pink} \\ \text{green} \end{pmatrix}$	$\begin{pmatrix} \text{red} \\ \text{light green} \end{pmatrix}$	$\begin{pmatrix} \text{red} \\ \text{green} \end{pmatrix}$	$\begin{pmatrix} \text{red} \\ \text{light green} \end{pmatrix}$
Biology	$\begin{pmatrix} \text{green} \\ \text{blue} \end{pmatrix}$	$\begin{pmatrix} \text{light green} \\ \text{light blue} \end{pmatrix}$	$\begin{pmatrix} \text{light blue} \\ \text{blue} \end{pmatrix}$	$\begin{pmatrix} \text{light green} \\ \text{blue} \end{pmatrix}$
Music	$\begin{pmatrix} \text{blue} \\ \text{purple} \end{pmatrix}$			

Pairs of close vectors are compared: e.g. $\begin{pmatrix} \text{pink} \\ \text{green} \\ \text{blue} \end{pmatrix}$ vs $\begin{pmatrix} \text{red} \\ \text{light green} \\ \text{blue} \end{pmatrix}$

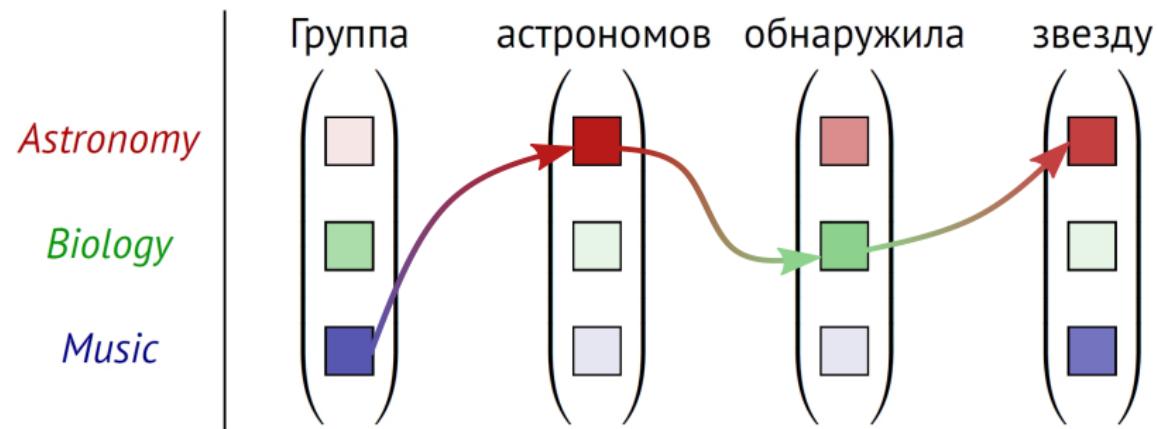
Meaning

How much the meanings of adjacent words differ, according to the topic model

	Low variance			High variance		
Literature	русский	поэт	Пушкин	Толстой	Рассел	Эйлер
Philosophy						
Mathematics						

Meaning

Estimation of how much the focus of a conversation drifts



Meaning

Average length of the topic in text

Count words of topic t , penalizing when word of another topic is encountered.

Пример. Тема $t = \text{"Чёрные дыры"}$

Группе $\underbrace{\text{астрономов удалось}}_{l_1=2}$ обнаружить $\underbrace{\text{звезды, обращающуюся}}_{l_2=2}$
вокруг $\underbrace{\text{чёрной дыры на рекордно близком}}_{l_3=4}$ расстоянии.

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

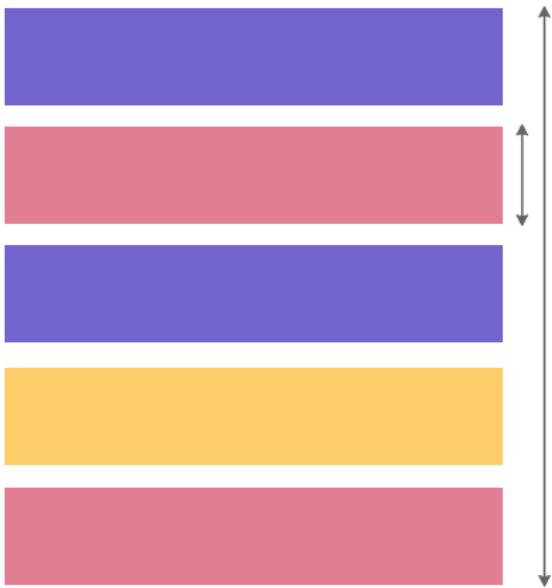
The better the coherence is, the better it should describe the ability of a topic model to figure out the segmentation structure!



Document consisting of sociology and medicine segments

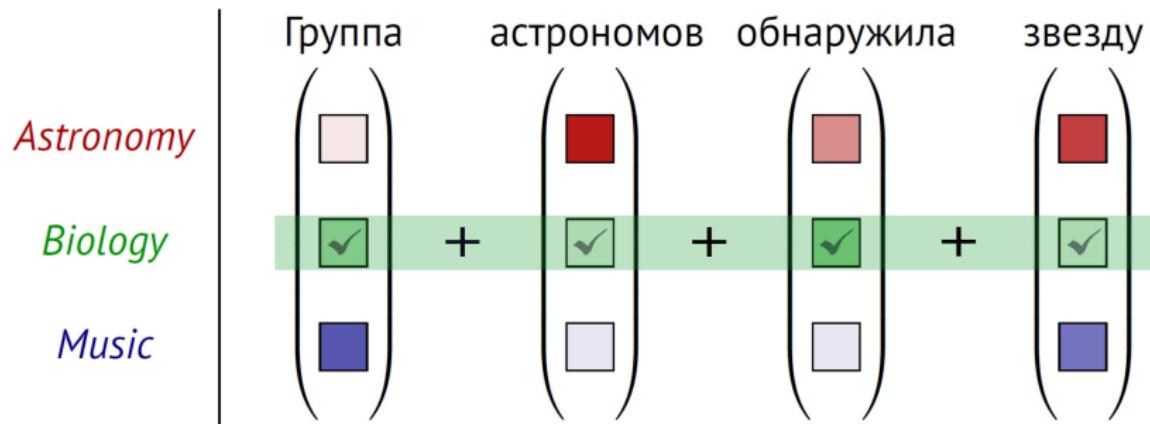
Dataset Generation

- We “cut” the monotopical documents into smaller monotopical segments.
- We “sew” them together in random order to produce a new document.
- We know topic labels for each word and can use them as ground truth.



Segmentation Quality. Soft

Sum of components $p(t | d, w)$ for topic t



Segmentation Quality. Strict

Number of coincidences of predicted topic and actual topic t

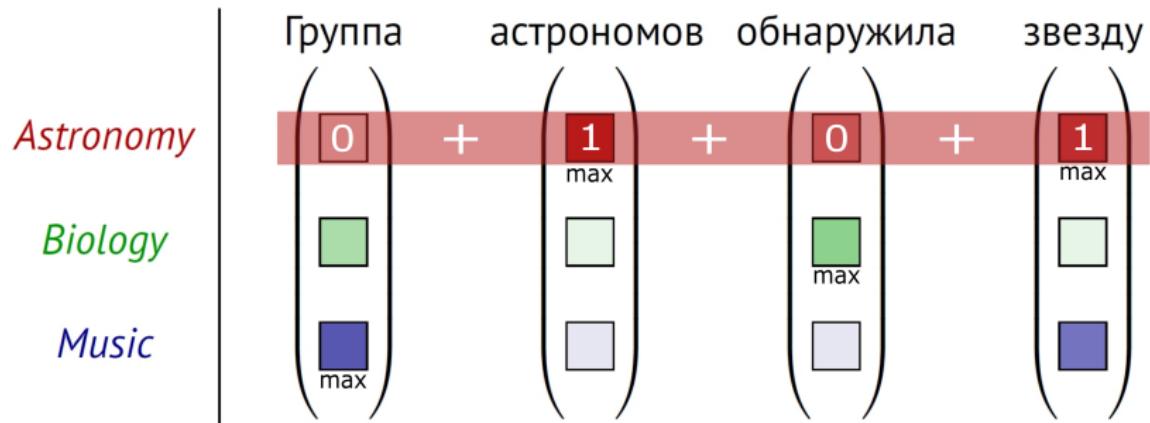


Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Illustration of a Bad Model Segmenting Text

topic 16 : язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся **предположения, желания**. Нормальный **африканский** грамматический **приём** — не говорить "я это сделаю" или "это будет", а сказать "это возможно" или "я хочу это сделать". Они говорят о будущем, но "попадают" в будущее непрямым путём.

topic 12 : политика

И я посылаю деньги **борцам** за независимость Курдистана, участвуя в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных гражданств. В литературе последних **десяти** лет бытуют такие выражения, как гендерное гражданство и **экономическое** гражданство. Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5500	11000	-4.8	-3.1	-13	0.95	-37000	2.9	-140000
16000	38000	-3.7	-2.7	-3.7	0.70	-8100	3.5	-54000

- SQ (S), SQ (H) – Soft and Strict segmentation qualities
- N, M – Newman, Mimno
- SC, TL, FC – SemantiC, TopLen, FoCon

Illustration of the Good Model Segmenting Text

topic 16 : язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю" или "это будет", а сказать "это возможно" или "я хочу это сделать". Они говорят о будущем, но "попадают" в будущее непрямым путём.

topic 12 : политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных гражданств. В литературе последних десяти лет бытуют такие выражения, как гендерное гражданство и экономическое гражданство. Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5500	11000	-4.8	-3.1	-13	0.95	-37000	2.9	-140000
16000	38000	-3.7	-2.7	-3.7	0.70	-8100	3.5	-54000

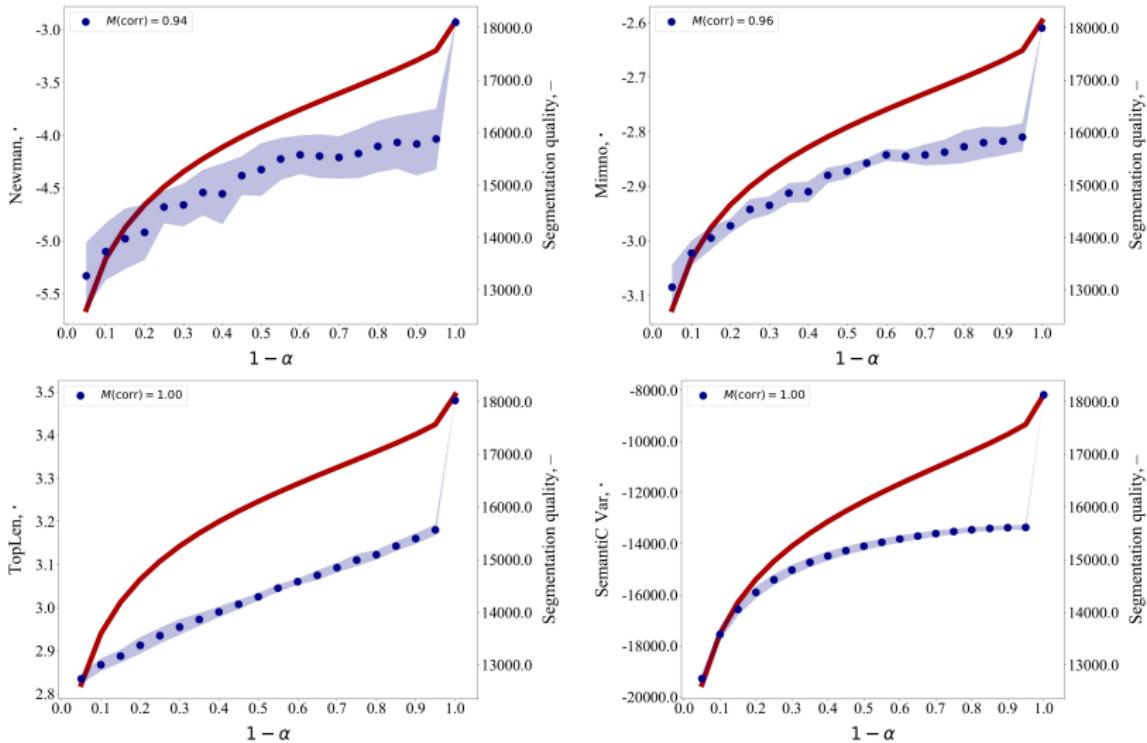
- SQ (S), SQ (H) – Soft and Strict segmentation qualities
- N, M – Newman, Mimno
- SC, TL, FC – SemantiC, TopLen, FoCon

Spearman Correlations Between Coherences & Segmentation Quality

Coh	Corr	Coh	Corr	Coh	Corr
Newman	0.75	Newman	0.80	Newman	0.85
Mimno	0.96	Mimno	0.94	Mimno	0.97
SC L2	0.92	SC L2	0.70	SC L2	0.59
SC Cos	-0.97	SC Cos	-0.97	SC Cos	-0.96
SC Var	1.00	SC Var	1.00	SC Var	1.00
TopLen	1.00	TopLen	1.00	TopLen	1.00
FoCon	1.00	FoCon	1.00	FoCon	1.00

Results for datasets with sizes of segments: 50, 200 and 400 words – and with 5 topics in each document

Coherences & Segmentation Quality as Functions of Topic Model Quality



- New methods of calculating coherence of a topic are proposed.
These methods take into account the whole text and, in the problem under consideration, outperform the top-tokens based ones.
- An automatic method for evaluating coherences functions is introduced.
It is based on the comparison of coherence values with text segmentation qualities for a range of topic models.