

Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций

Василий Алексеев

Предзащита бакалаврской работы

13 июня 2018



Тема, Интерпретируемость и Когерентность

Тема характеризуется набором слов, которые часто совместно встречаются в тексте. *Топ-слова* темы — её самые частые слова.

Интерпретируемость темы означает, может ли человек по словам темы объяснить, о чём она, дать ей подходящее название.

Хорошо интерпретируемая тема (самые частые слова)

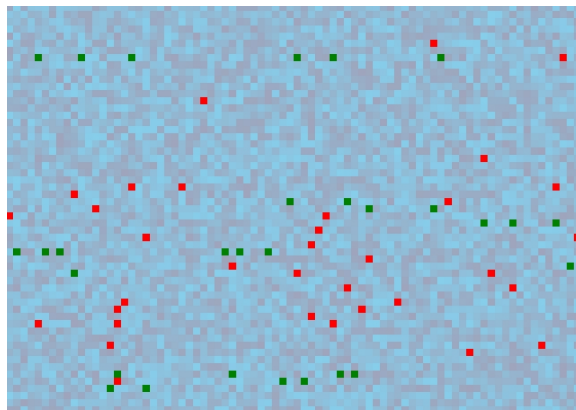
актёр, пьеса, музыкальный, премьера, партер, зритель, продюсер, аудитория, занавес, оркестр

Плохо интерпретируемая тема (самые частые слова)

экспресс, эпиграф, туманный, результат, образ, право, заём, иероглиф, лак, футбол

Когерентность — это автоматический способ оценки интерпретируемости, когда оценивается, как часто топ-слова темы встречаются недалеко друг от друга в тексте.

Проблема подхода к оценке интерпретируемости через совстречаемости топ-слов тем



- слова
- топ-слова
- совстречаемости

Десять топовых слов покрывают малую часть всего текста.
Совстречаемостей этих слов (то есть позиций топ-слов, когда рядом с ними есть другие топ-слова) ещё меньше.

Проблема

Когерентности по топ-словам опираются на заданное количество самых частых слов темы. Этот список слов несёт информацию лишь о части тематической модели. Помимо этого, оценивать интерпретируемость с помощью экспертов дорого и затратно.

Решение













Смотреть, как тема распределена по *всем* словам текста. Считать когерентность темы как среднюю схожесть слов, близко расположенных в тексте.

- 1 Внутритекстовые когерентности
- 2 Полуавтоматическая оценка качества функций когерентности
 - Полусинтетический датасет
 - Качество сегментации
- 3 Эксперименты

SemantiC (L2, Cos): Semantic Closeness

Значение



















Близость близко расположенных в тексте слов темы t

	Группа	астрономов	обнаружила	звезду
<i>Астрономия</i>				
<i>Биология</i>				
<i>Музыка</i>				

Сравниваются пары векторов слов: например $\begin{pmatrix} \text{dark red} \\ \text{light green} \\ \text{light blue} \end{pmatrix}$ и $\begin{pmatrix} \text{dark red} \\ \text{light green} \\ \text{dark blue} \end{pmatrix}$

Значение

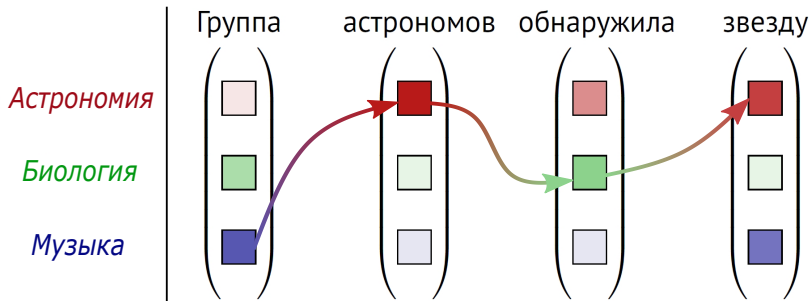
Разброс темы t по близко расположенным словам

	Малый разброс			Большой разброс		
	русский	поэт	Пушкин	Толстой	Рассел	Эйлер
<i>Литература</i>						
<i>Философия</i>						
<i>Математика</i>						

Значение

Как сильно изменяется тема среди смежных слов

Метод не привязан к теме, он сразу даёт значение когерентности для *тематической модели* как целого.



Значение

Средняя длина темы внутри текста

Считает слова темы t , штрафую, когда встречается слово другой темы.

Пример для темы $t = \text{«Чёрные дыры»}$

Группе астрономов удалось обнаружить звезду, обращающуюся
 $l_1=2$ $l_2=2$
 вокруг чёрной дыры на рекордно близком расстоянии.
 $l_3=4$

Гипотеза

Все тексты сегментированы. Но позиции сегментов не известны

2000 *монотематических* статей «ПостНауки»¹ разрезаются на сегменты одинаковой длины и сшиваются в новые документы.



Документ из двух сегментов: про «социологию» и «медицину»

¹<https://postnauka.ru>













Полусинтетический датасет

Чем лучше функция когерентности, тем лучше она должна описывать способность тематической модели угадывать сегментную структуру текста

Для каждого слова в полусинтетическом датасете известно, к какой теме оно относится

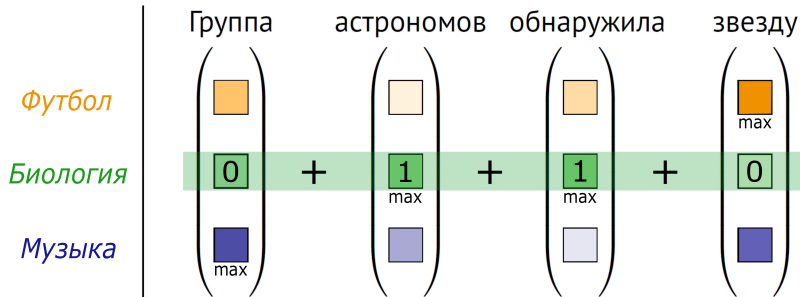


Сумма $p(t \mid d, w)$ по всем словам в сегментах темы t

	Группа	астрономов	обнаружила	звезду
Футбол				
Биология				
Музыка				

Качество сегментации: Hard

Количество совпадений между темой, предсказываемой моделью $\arg \max_{\tau} p(\tau \mid d, w)$, и действительной темой среди слов в сегментах темы t



Когерентности по топ-словам могут игнорировать более 98% слов текста коллекции документов

Min	0.016
Mean	0.062
Max	0.28
<hr/>	
Total	1.2

Часть корпуса (%), которая занята встречаемостями десяти топовых слов для тем «ПостНауки»

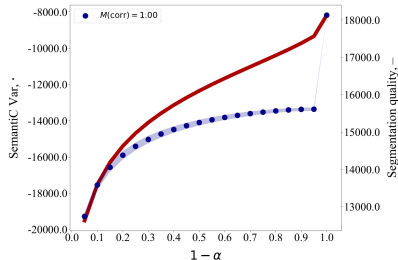
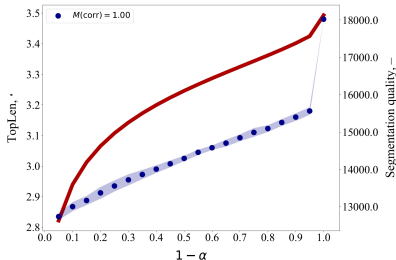
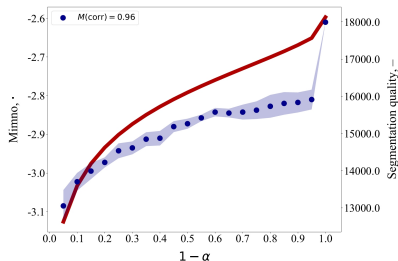
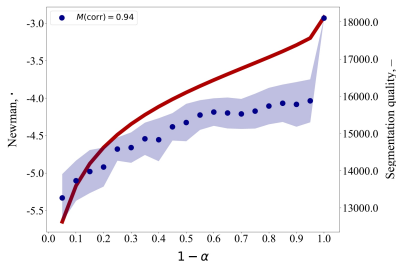
Спирмановские корреляции между когерентностями и качествами сегментации

Ряд тематических моделей: от «хорошей» модели коллекции «ПостНаука» до плохой модели, матрица Φ которой взята из некоторого вероятностного распределения

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

Результаты для сегментов с размерами 50, 200 и 400 слов и 5 темами
в каждом документе

Когерентности и качества сегментации как функции качества тематической модели



- Проиллюстрирован недостаток когерентностей по топ-словам: покрытие лишь малой части текстовой коллекции.
- Предложен полуавтоматический метод оценки качества функций когерентности: по корреляции с качеством сегментации полусинтетического текста тематическими моделями.
- Представлены методы *внутритекстовой* когерентности. По предложенной функции оценки качества некоторые новые методы показывают лучшие результаты, чем когерентности по топ-словам.