

Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций

Василий Алексеев

Предзащита бакалаврской работы

13 июня 2018



Тема, Интерпретируемость, Когерентность

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

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

Хорошо интерпретируемая тема (самые частые слова)

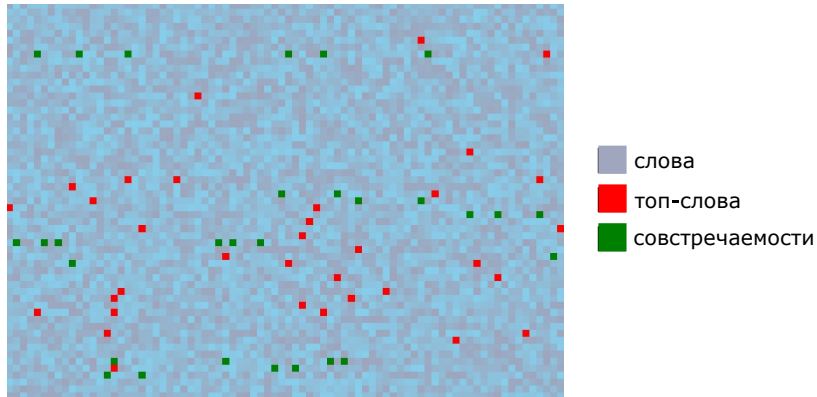
актёр, пьеса, музыкальный, премьера, партер, зритель, продюсер, аудитория, занавес, оркестр

Плохо интерпретируемая тема (самые частые слова)

экспресс, эпитафия, 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 темы



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.

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 (Semantic Closeness): l_2

Значение

Близость векторов 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 $\begin{pmatrix} \text{red} \\ \text{green} \\ \text{blue} \end{pmatrix}$ и $\begin{pmatrix} \text{red} \\ \text{green} \\ \text{blue} \end{pmatrix}$

SemantiC (Semantic Closeness): Var

Значение

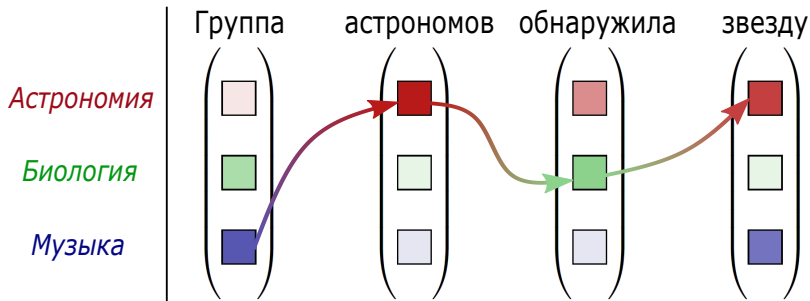
Разброс темы t по близко расположенным словам

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

Значение

Как сильно изменяется тема t 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.
 $l_3=4$

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

Гипотеза

Все тексты сегментированы. Но позиции сегментов не известны

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



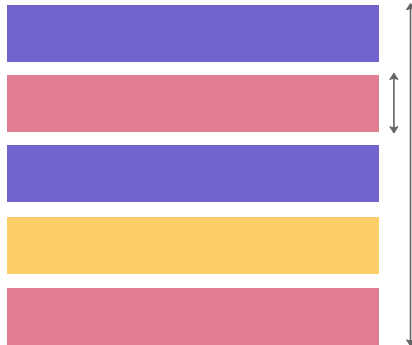
Документ из двух сегментов: про «социологию» и «медицину»

¹<https://postnauka.ru>













Полусинтетический датасет

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

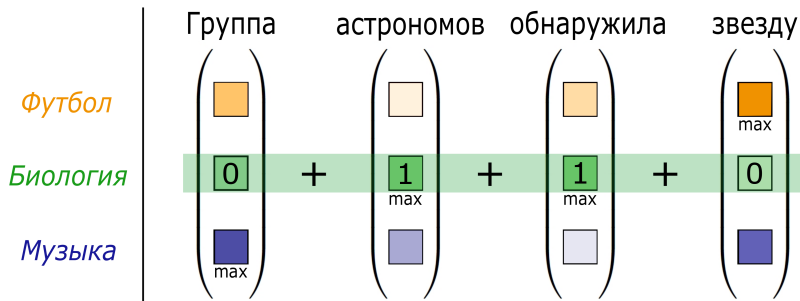
Для каждого слова в полусинтетическом датасете известно, к какой теме оно относится



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

	Группа	астрономов	обнаружила	звезду
Футбол				
Биология				
Музыка				

Количество совпадений темы, предсказываемой моделью $\arg \max_{\tau} p(\tau \mid d, w)$, и темы t на словах сегментов темы t



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

Недостаток подхода с помощью топ-слов

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

Min	0.016
Median	0.048
Mean	0.062
Max	0.28
Total	1.2

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

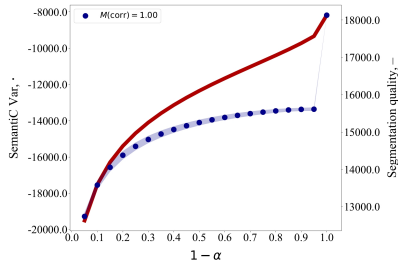
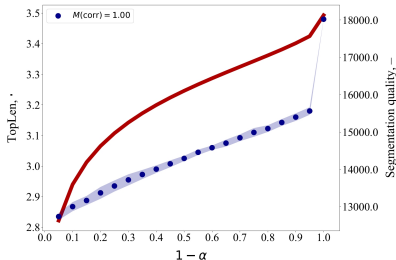
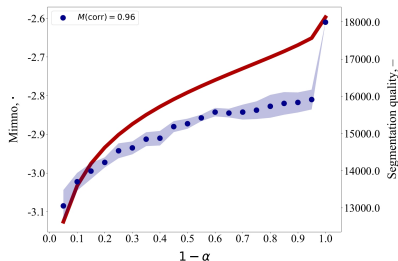
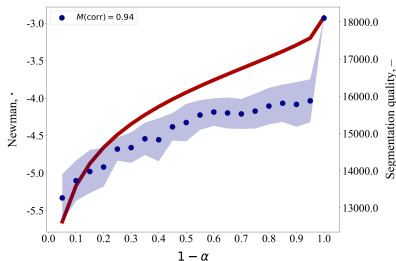
Спирмановские корреляции между когерентностями и качествами сегментации

Ряд моделей: $\Phi(\alpha) = \alpha \cdot \Phi_{bad} + (1 - \alpha) \cdot \Phi_{good} \mid \alpha \in [0, 1)$
от модели коллекции «ПостНаука» Φ_{good} до случайной Φ_{bad}

Coh	Corr
Newman	0.80
Mimno	0.94
SemantiC l_2	0.70
SemantiC Var	1.00
TopLen	1.00
FoCon	1.00

Корреляции при размере сегмента 200 слов
и при 5 темах 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 качества некоторые внутритекстовые методы лучше, чем когерентности по топ-словам.