ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2021 Управление, вычислительная техника и информатика № 56

ИНФОРМАТИКА И ПРОГРАММИРОВАНИЕ

УДК 004.896
DOI: 10.17223/19988605/56/11

Ф.В. Краснов, Е.Н. Баскакова, И.С. Смазневич

ОЦЕНКА ПРИКЛАДНОГО КАЧЕСТВА ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ЗАДАЧ КЛАСТЕРИЗАЦИИ

Исследуются методы оценки качества тематических моделей, способные обеспечить их устойчивое применение для решения практических задач, связанных с анализом набора текстовых документов. На примере задачи мягкой кластеризации показано, что использования метрики средней когерентности тем недостаточно для оценки применимости построенной модели, и целесообразно учитывать показатели связей документов с высококогерентными темами.
Ключевые слова: тематическое моделирование; когерентность темы; мягкая кластеризация; анализ текста; ARTM.

За два десятилетия тематическое моделирование текстовых коллекций зарекомендовало себя как гибкий и надежный инструмент для работы с большими объемами документов. Эффективная реализация этого алгоритма позволяет успешно решать задачи кластеризации и классификации текстов [1], может использоваться для поиска информации [2], аннотирования текстов [3], анализа трендов и новостных потоков [4], обработки мультиязычных данных [5].

Начиная с изобретения метода в 1999 г. [6], принципы и алгоритмы тематического моделирования постоянно развивались. В настоящее время, согласно обзору [7], существует более 50 вариантов построения тематических моделей. Наиболее популярный сегодня метод построения тематической модели – метод латентного распределения Дирихле (Latent Dirichlet Allocation, LDA), созданный авторами D. Blei, A. Ng, M. Jordan и описанный в работе [8], набравшей более 35 тысяч цитирований[1].

Самым перспективным можно назвать метод тематического моделирования с аддитивной регуляризацией (Additive Regulariza Tiontopic Modelling, ARTM) [9], в основе которого лежит алгоритм PLSA (вероятностный латентно-семантический анализ) [6]. В рамках этого подхода был создан теоретический аппарат для построения тематических моделей, оптимизирующих заданный перед началом обучения набор формализованных критериев. Программная реализация метода в библиотеке BigARTM [10] позволяет обучать модели ARTM с высокой скоростью в потоковом (онлайновом) и статическом режимах.

Результатом обучения тематической модели являются две матрицы: Φ-матрица «термы–темы» и Θ-матрица «темы–документы», произведение которых дает исходную матрицу «документы–термы», также известную как модель «мешок слов». Задача тематического моделирования относится к классу некорректных обратных задач, т.е. существует бесконечное множество вариантов Φ и Θ, соответствующих заданной векторной модели текстового корпуса. Поэтому актуальной является задача получения тематической модели с определенным качеством.

Служебным признаком сходимости итерационного процесса обучения модели является ее перплексия. С прикладной точки зрения важно исследовать качество каждой из найденных тем. На прак-

[1] По данным системы «Google Академия».

тике его принято связывать со свойством их интерпретируемости человеком. Для оценки этого показателя могут быть использованы разные стратегии. Обзоры существующих методов были сделаны, например, в работах [11–13].

Наиболее популярные способы вычисления основаны на PMI [14] (или нормализованном варианте – NPMI [15]) и логарифме условной вероятности [16], которые, в свою очередь, базируются на анализе совстречаемости термов. В других подходах статистика совстречаемости используется опосредованно. В работе [17] слова тем помещаются в семантическое пространство, построенное на PMI или NPMI, где методами дистрибутивной семантики рассчитывается попарная схожесть этих слов.

Универсальной метрикой в задачах тематического моделирования считается когерентность тем. Высококогерентная тема содержит семантически связанные между собой слова, что определяется контекстом их использования. Эта метрика подходит для большинства задач текстовой аналитики, решаемых с помощью тематического моделирования. Однако когерентность темы может быть вычислена различными алгоритмами, каждый из которых допускает вариативность набора параметров.

Изучению подходов к оптимизации оценки и методов вычисления тематической когерентности посвящен ряд исследований. Так, в работе [11] отмечается, что значение метрики когерентности оказывается больше у тем с часто употребляемыми термами. В результате получаются понятные, но довольно общие темы, что не всегда приемлемо при работе с отраслевыми документами. Авторы предлагают еще две дополнительные метрики – эксклюзивность и подъем (lift): эксклюзивность темы отражает степень перекрытия между темами, подъем определяет с помощью референсного корпуса степень присутствия специфичных слов в темах модели. В [13] для оценки качества тематической модели предлагается метод автоматического обнаружения лишнего терма (topic intruder). Для этого анализируется не только матрица Φ «темы–термы», но и связи тем с документами в матрице Θ. В [18] исследуется зависимость когерентности темы от выбора числа термов, учитываемых при расчете PMI. В [19] сравниваются различные способы вычисления когерентности и предлагается общий метод оценки совстречаемости термов через исследование близости их векторов.

Большое число исследований говорит о том, что единого «золотого стандарта» по оценке качества тематической модели пока не найдено. При этом академический подход не всегда оправдан в прикладных задачах.

Одна из возможностей прикладного использования тематической модели – мягкая (нечеткая) кластеризация набора текстов. На основе самых значимых слов тем кластеры получают собственные интерпретируемые названия, тогда как при других подходах кластеры формируются анонимными, что понижает объяснимость решения.

Для валидации качества кластеров разработано достаточно много метрик: Partition Coefficient [20], Dunn Index [21], DBI [22] и ее модификации [23–25], Silhouette [26]. Однако все они задействованы в алгоритмах кластеризации и не подходят для уже построенной тематической модели, которая в границах информационной системы может участвовать и в решении других задач.

С прикладной точки зрения в модели важны только темы с высокой когерентностью, их количество (или доля) характеризует потенциал практического использования модели. Интуитивно «хорошей» считается ситуация, когда документ моделируется 2–3 темами с высокими значениями их весов, в то время как документ, где десятки тем присутствуют незначительно, считается нераспознанным моделью (неподдающимся тематическому анализу). При этом в тематической модели всегда будут присутствовать низкокогерентные темы, но они не имеют значения в реальных задачах.

Данная статья посвящена изучению поведения метрики когерентности, основанной на нормализованной поточечной взаимной информации, для решения задачи мягкой тематической кластеризации текстов. Цель настоящего исследования – выработка нового подхода к оценке тематической модели, обеспечивающего возможность устойчивого применения методов тематического моделирования в прикладных интеллектуальных информационных системах за счет повышения точности кластеризации, а также уменьшения вычислительной сложности. Гипотеза авторов состоит в том, что среднее число высококогерентных тем на один документ является более эффективной метрикой для оценки качества тематической модели, чем средняя когерентность.

1. Методика

Идея оценки качества тем с помощью когерентности была сформулирована Д. Ньюманом и соавт. в работе [14]. В рамках исследования свойств взаимной информации между термами авторы ввели метрику PMI-Score (1), которую в дальнейшем назвали когерентностью темы:

$$PMI-Score(w)=mean\left\{PMI\left(w_i,w_j\right),ij=1...10(k),i\neq j\right\}, \text{ где } PMI(w_i,w_j)=\log\frac{p\left(w_i,w_j\right)}{p\left(w_i\right)\cdot p\left(w_j\right)}. \quad (1)$$

Когерентность темы PMI-Score вычисляется как средняя поточечная взаимная информация для всех пар различных термов w_i и w_j из массива $w = (w_1, \ldots, w_{10})$, содержащего 10 наиболее весомых термов темы; $p(w_i, w_j)$ – вероятность совместного появления термов w_i и w_j в текстовом окне заданной ширины, $p(w_i)$ – вероятность появления терма w_i в текстовом окне заданной ширины.

Также в [14] были собраны мнения экспертов о тематической модели и констатирована высокая корреляция между когерентностью и человеческими оценками тем. Таким образом, метрика когерентности темы стала основной для оценки ее качества.

Свободный параметр в определении когерентности (1) – число 10, в более поздних работах замененное на k. Для экспертной оценки необходимо было выбирать из каждой темы небольшое количество (k) взвешенных термов, характеризующих ее, тогда как в полной тематической модели количество термов одинаково для всех тем и равно количеству термов в словаре коллекции.

Применительно к тематическим моделям Ньюман использует когерентность в исследовании [27], где она сравнивается с 15 другими метриками и показывается, что сильнее всего когерентность коррелирует с человеческой оценкой у моделей с автоматически выделяемыми темами в количестве $T = 200$ и $T = 400$. Для вычисления совстречаемости термов в [27] введено понятие «скользящего окна» – отрывка исходного текста длиной в 10 термов (позднее длина окна стала задаваться величиной sw – от sliding window). Свободный параметр $k = 10$ стал интерпретироваться как число наиболее весомых термов в темах модели, учитываемое при расчете когерентности.

Д. Мимно и соавт. [16] ввели новую формулу для когерентности темы на основе величины document frequency (DF), которая может быть переписана в терминах данной статьи:

$$C_{df}=\sum_{j=2}^{k}\sum_{i=1}^{j-1}\log\frac{D\left(w_j,w_i\right)+1}{D\left(w_i\right)}, \quad (2)$$

где C_{df} – когерентность темы для документов, $D(w_i)$ – количество документов, где встречается терм w_i, $D(w_i,w_j)$ – количество документов, где встречаются термы w_i и w_j одновременно, w_i – i-й термин в порядке убывания веса в матрице Ф для темы t, k (= 10) – количество термов в теме t (постоянная 1 включена для исключения возможности вычисления логарифма 0).

В отличие от (1) в формуле (2) рассматривается когерентность для документов, а не для термов. Будем различать далее метрики когерентности C_{tf} (1) и C_{df} (2). В формуле (2) еще не используется окно: совстречающимися считаются термы, содержащиеся одновременно в одном или более документах. При этом в обеих формулах (1) и (2) значение когерентности темы зависит от выбора величины k, определяющей набор термов с наибольшими весами. В [16] это число равно 10 при общем количестве тем модели $T = 200$.

В статье [28] предложена нормализованная версия PMI – NPMI, в терминах данной статьи представимая как

$$NPMI\left(t\right)=\sum_{j=2}^{k}\sum_{i=1}^{j-1}\frac{\log\frac{P\left(w_j,w_i\right)}{P\left(w_i\right)P\left(w_j\right)}}{-\log P\left(w_i,w_j\right)}. \quad (3)$$

Когерентность темы t на основе NPMI показала высокую корреляцию с экспертными оценками за счет уменьшения влияния низкочастотных термов.

В работе [19] произведено сравнение вышеперечисленных формул когерентности темы в тематической модели по методике LDA. Представляет интерес, что авторы [19] рассмотрели поведение корреляции когерентности темы с ее человеческой оценкой в зависимости от значения свободного параметра *sw*, определяющего размер скользящего окна, в интервале от 10 до 300 термов. Для всех вариантов расчетов когерентности выявлена общая тенденция: уменьшение корреляции при увеличении $sw > 50$. Также в [19] предложена новая формула для когерентности, в терминах данной статьи представимая следующим образом:

$$C_{tf} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} PMI(w_i, w_j), \text{ где } PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \varepsilon}{P(w_i) P(w_j)}. \quad (4)$$

Нормировочный коэффициент в (4) отражает количество возможных пар среди k термов и позволяет сравнивать когерентности тем при различных значениях k, а PMI может быть заменено на NPMI.

Рассмотрим когерентность для двух вариантов совстречаемости PMI с окном $sw = 10$: по документам (*DF*) и по термам (*TF*), используя только положительные значения (PPMI). Вероятности в формуле Ньюмана (1) можно оценить по частотам:

$p(w_i, w_j) = \frac{n(w_i, w_j)}{n}$ и $p(w_i) = \frac{n(w_i)}{n}$, где $n(w_i) = \sum_{w \in W} n(w_i, w)$, $n = \sum_{w \in W} n(w)$ – размер словаря.

Следовательно, формулу PPMI можно записать как

$$PPMI(w_{di}, w_{dj}) = \left[\log \frac{n(w_{di}, w_{dj}) n}{n(w_{di}) n(w_{dj})} \right]_+, \quad (6)$$

где

$$n(w_{di}, w_{dj}) = \sum_{d=1}^{|D|} \sum_{i=1}^{N_d} \sum_{j=1}^{N_d} \left[0 < |i - j| \le k \right] \text{ – величина } CoocTF, \quad (7)$$

оценка, обозначающая, сколько раз пара термов w_i и w_j встретилась в коллекции внутри окна заданной ширины, либо

$$n(w_{di}, w_{dj}) = \sum_{d=1}^{|D|} \left[\exists i, j : 0 < |i - j| \le k \right] \text{ – величина } CoocDF, \quad (8)$$

оценка, обозначающая, в скольких документах коллекции встретилась пара термов w_i и w_j хотя бы один раз в окне заданной ширины.

Согласно [19, 29], формулы (6)–(8) дают высокую корреляцию когерентности тем с человеческой оценкой их качества.

В рассмотренных исследованиях анализируется когерентность отдельных тем либо усредняется когерентность всех тем модели, чтобы оценивать общее качество тематической модели и сравнивать методики тематического моделирования [30]. Авторы [19, 29, 31] отмечают, что, упорядочив темы по этой метрике, можно выбрать основные темы коллекции, однако характер распределения когерентности тем в этих работах не изучен, что подсказало авторам направление исследований.

Для эксперимента в рамках данной работы была выбрана метрика среднего количества высококогерентных тем в одном документе T_D: при уменьшении этой величины улучшается качество кластеризации текстовой коллекции с помощью тематической модели.

2. Описание эксперимента

Цель эксперимента – изучить поведение метрики когерентности темы в зависимости от свободных параметров тематической модели. Для достижения этой цели были сформулированы следующие задачи:

1. Выбрать три различные методики построения тематической модели.

2. Выбрать два набора данных с различными характеристиками (длина одного документа, размер словаря, количество документов коллекции).

3. Выбрать формулу для расчета когерентности темы.

4. Разработать форму визуального представления распределения когерентности темы в зависимости от свободных параметров тематической модели; определить характер зависимости распределения когерентности.

5. Разработать методику определения прикладного качества тематической модели.

Для эксперимента были выбраны три методики построения тематической модели: генеративная – LDA (*alpha* = *T*/50, *beta* = 0,01) [8], вероятностная – PLSA [6], вероятностная с аддитивной регуляризацией – ARTM [9].

Для реализации этих алгоритмов использовалась библиотека BigARTM как наиболее производительная. Токенизация проводилась методом NLTK, приведение к нормальной форме – PyMorphy. Предобработка и анализ документов указанными методами и алгоритмами осуществлялись в рамках платформы анализа текста SemanticTech[1].

В качестве данных были выбраны два набора русскоязычных документов:

– «Тайга»: 7 695 текстов средней длины (среднее количество термов в одном документе – 145) со словарем объемом 40 тысяч термов [32];

– ГОСТ: 1 066 длинных текстов (среднее количество термов в документе – 1 391) со словарем объемом 23 тысяч термов (сформирован авторами[2]).

Для расчета когерентности темы была использована формула

$$C_{PPMI} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} PPMI\left(w_i, w_j\right), \tag{9}$$

где PPMI рассчитывается по формуле (6). Для расчета когерентности по словарной частоте C_{PPMI}^{TF} совстречаемость термов определяется по формуле (7), для расчета когерентности по документарной частоте C_{PPMI}^{DF} совстречаемость термов определяется по формуле (8).

3. Результаты

На рис. 1 для выбранных наборов данных показана зависимость средней когерентности от количества тем (*T*) для трех вариантов построения тематических моделей по формулам (6)–(8).

Видно, что для корпуса «Тайга» наибольшим значением средней когерентности обладает методика LDA в варианте DF в районе значения $T = 600$ и в районе значения $T = 300$ в случае TF. Для ГОСТ наибольшим значением средней когерентности обладает ARTM в районе $T = 500$ (DF) и $T = 400$ (TF). Для «Тайги» при TF все методики имеют ярко выраженные максимумы, в случае DF методики ARTM и PLSA выходят на асимптотический рост при $T > 500$. Наименьшими средними значениями когерентности обладает тематическая модель набора «Тайга» по методике PLSA при DF. В ходе эксперимента остановка обучения моделей для получения зависимостей, отраженных на рис. 1, была сделана при достижении схожих значений перплексии около $P = 300$, что позволяет считать сравнение методик оправданным.

Рассмотрим детально распределение когерентности тем для разных значений *T* и различных методик построения тематической модели. На рис. 2 приведены кривые распределения когерентности тем в разных случаях для набора «Тайга». Из него видно, что распределения принципиально не отличаются: с ростом количества тем модели растет максимальное значение когерентности тем. Эти детали распределений теряются при усреднении когерентности, сделанном на рис. 1.

Далее в качестве метрики когерентности была использована метрика C_{PPMI}^{TF}, поскольку, как видно на рис. 1, 2, она более чувствительная для данной коллекции и имеет ярко выраженные экстремумы, что позволяет выбрать оптимальное число тематических компонент для построения модели.

[1] Система мониторинга и семантического анализа юридических документов SemanticTech – собственная разработка NAUMEN R&D; свидетельство о регистрации программы для ЭВМ № 2021610340 от 13.01.2021. URL: https://new.fips.ru/iiss/document.xhtml?faces-redirect=true&id=3aee969154040e2049c98bc60a571dcd

[2] Краснов Ф.В., Баскакова Е.Н., Смазневич И.С. Принцип построения корпуса нормативно-технических документов. PREPRINTS.RU. 2021. DOI: 10.24108/preprints-3112181

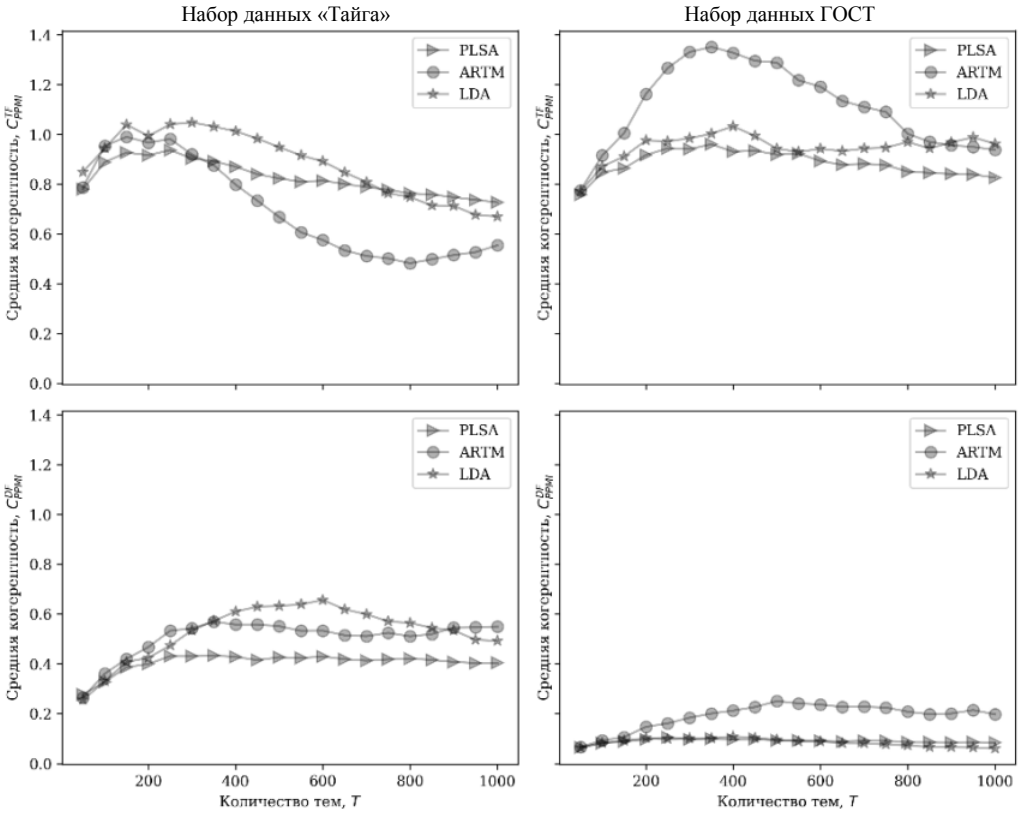

Рис. 1. Средняя когерентность для различных наборов данных и методик построения тематических моделей
Fig. 1. Average coherence for different datasets and methodologies for constructing thematic models

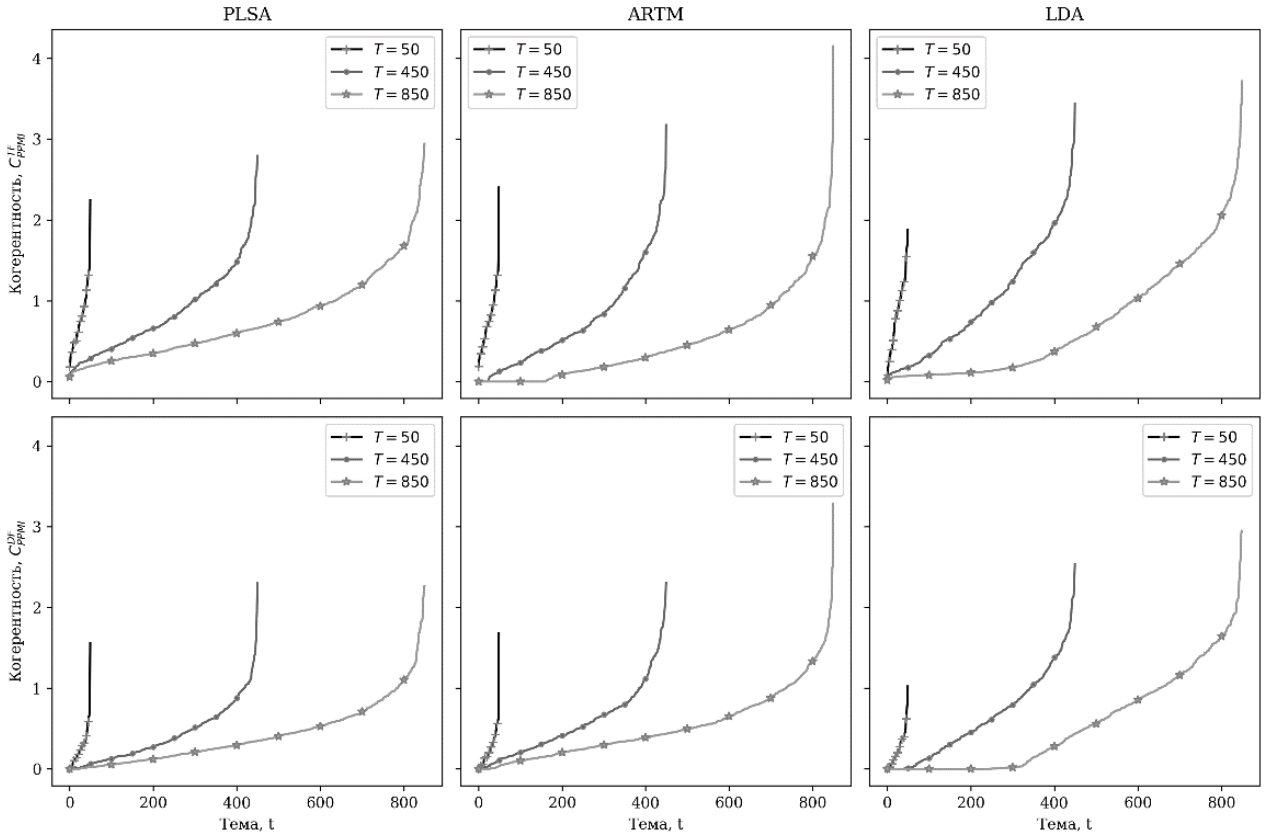

Рис. 2. Когерентность тем для методик ARTM, LDA, PLSA и набора данных «Тайга» в вариантах DF и TF
Fig. 2. Coherence of topics for ARTM, LDA, PLSA methods and the "Taiga" dataset in DF and TF variants

В табл. 1 приведены значения метрики качества кластеров в корпусе ГОСТ. Значения когерентности C_{PPMI}^{TF} получены при размере окна $sw = 5$. Под порогом подразумевается вес связи документа и темы в матрице Θ. Среднее количество тем в документе и среднее количество высококогерентных тем в документе обозначены как K_t и K_{hct} соответственно.

Таблица 1

Метрики качества кластеров текстов корпуса ГОСТ

Количество тем в модели	Методика моделирования	Порог < 10^{-4}		Порог < 10^{-3}		Порог < 10^{-2}	
		K_{hct}	K_t	K_{hct}	K_t	K_{hct}	K_t
50	PLSA	10,7	25,3	6	14,5	2,6	6,4
	ARTM	12,5	25,6	8	16,5	3,6	7,5
	LDA	27	50	24,8	46,1	3,8	7,7
100	PLSA	13,7	33,4	6,8	16,6	2,6	6,4
	ARTM	16,4	36	9,7	21,4	3,8	8,5
	LDA	57,8	99,7	36,3	64,8	3,6	6,9
200	PLSA	19,6	39	8,8	17,2	3,3	6,1
	ARTM	29,8	48,3	17,1	27,6	5,7	9,4
	LDA	123,4	196,1	30,1	50,6	3,6	6
500	PLSA	15,3	35,5	6,2	13,9	2,3	4,6
	ARTM	40,7	78,4	25,5	52,4	7,3	14,8
	LDA	211,5	413	8,8	22,9	2,3	4,8

Результат эксперимента показал, что вне зависимости от используемой методики построения модели в среднем можно достичь сокращения числа тем документов в 2 раза, анализируя среди всех тем только высококогерентные темы (ВКТ). В прикладных задачах это обеспечит более качественную кластеризацию документов с точки зрения человеческой оценки. Анализ только ВКТ документа позволит сократить усилия на исключение шумовых тем, принимая во внимание только те семантические связи, которые являются весомыми в рамках обрабатываемой коллекции документов.

Однако необходимо удостовериться, что учет только ВКТ не приведет к снижению качества распределения документов по темам для пользователя.

В рамках данного эксперимента высококогерентными темами считаются темы со значением когерентности больше 1; это означает, что каждый терм встречается с каждым другим термом один раз (среди 10 наиболее значимых термов каждой темы).

При учете только ВКТ из 1 066 документов корпуса ГОСТ 161 документ (15%) не имеет связи ни с одной темой, при этом 54 документа не имеют явно выраженных (с весом более 0,4) связей ни с одной из тем в Θ вне зависимости от значения их когерентности, – такие документы исключаются из рассмотрения. Интерес представляют документы, имеющие весомую (> 0,4) связь с автоматически выделенной темой, которая не является высококогерентной. При подробном анализе были выявлены темы с когерентностью, близкой к пороговой (= 1), связанные большим весом с несколькими документами (пример такой темы приведен в табл. 2), не имеющими весомых связей с другими темами. Такие документы также образуют кластер.

Таблица 2

Пример кластера документов ГОСТ, связанных с низкокогерентной темой, близкой к порогу

Topic_4. Когерентность – 0,82. Термы (веса): покрытие (0,08), нанесение (0,04), состав (0,04), приготовление (0,03), валик (0,02), масло (0,02), приложение (0,02), древесина (0,01), влажность (0,01), конструкция (0,01)	
Вес темы в документе	Название документа
0,79	ГОСТ 25130–82. Покрытие по древесине вспучивающееся огнезащитное ВПД. Технические требования
0,78	ГОСТ 25131–82. Государственный стандарт Союза ССР. Покрытие по стали вспучивающееся огнезащитное ВПМ-2. Технические требования
0,67	ГОСТ 23790–79. Покрытие по древесине фосфатное огнезащитное. Технические требования
0,65	ГОСТ 25665–83. Покрытие по стали фосфатное огнезащитное на основе минеральных волокон. Технические требования
0,63	ГОСТ 23791–79 (1985). Покрытие по стали фосфатное огнезащитное. Технические требования

Далее, со снижением уровня когерентности, обнаруживалось все меньше связей между документами и темами, существенных для решения прикладных задач, таких как автоматическое формирование рубрикатора, тематический поиск, маршрутизация документов и т.д.

Отдельно была рассмотрена группа документов, имеющих весомую связь в матрице Θ (> 0,6) с низкокогерентной темой (НКТ). Если учитывать только ВКТ, то в некоторых документах теряется весомая связь с НКТ и основной темой становится ВКТ, что не снижает качества анализа с точки зрения пользователя. Примеры таких документов приведены в табл. 3.

Таблица 3

Пример документов, по которым произошла замена ключевой темы с НКТ на ВКТ

Тема (когерентность)	Вес темы в документе	Состав темы
ГОСТ 12.1.041–83. Система стандартов безопасности труда. Пожаровзрывобезопасность горючих пылей. Общие требования		
ВКТ (1,80)	0,26	Topic_79: вещество (0,06), температура (0,04), взрыв (0,03), смесь (0,03), испытание (0,02), взрывоопасный (0,02), самовоспламенение(0,02), воспламенение (0,02), горючий (0,01), концентрация (0,01)
НКТ (0,45)	0,63	Topic_123: пыль (0,02), оборудование (0,02), аппарат (0,02), норма (0,02), водоснабжение (0,01), вода (0,01), услуга (0,01), коммунальный (0,01), лампа (0,01), система (0,01)
ГОСТ 20522–96. Грунты. Методы статистической обработки результатов испытаний		
ВКТ (1,35)	0,1	Topic_96: грунт (0,31), глинистый (0,02), песок (0,02), природный (0,01), степень (0,01), грунтовый (0,01), состав (0,01), крупнообломочный (0,01), органический (0,01), плотность (0,01)
НКТ (0,79)	0,76	Topic_11: значение (0,13), коэффициент (0,06), формула (0,05), приложение (0,04), характеристика (0,03), испытание (0,03), величина (0,02), результат (0,02), определение (0,01), средний (0,01)

В остальных случаях документы, имеющие весомую связь только с НКТ, при удалении этой темы из рассмотрения считаются нераспознанными тематической моделью, что не противоречит их первоначальному статусу, поскольку низкая когерентность темы означает ее неинтерпретируемость.

Таким образом, эксперимент подтвердил, что использование в информационных приложениях связей документов только с ВКТ не ухудшает результаты тематического анализа, а позволяет концентрироваться на малом количестве ключевых тем, понятных пользователям. Следовательно, показатель качества тематической модели можно вычислять с помощью доли высококогерентных тем в модели. Сокращение количества тем в модели на 30% за счет использования только ВКТ позволило снизить среднее количество связей «документ–тема», рассматривая только хорошо интерпретируемые темы, и повысить среднюю когерентность модели на 20%.

4. Дискуссия

Для исследования корреляции когерентности темы с человеческим восприятием важны как наибольшие, так и наименьшие значения когерентности, чтобы сравнивать их и с позитивными, и с негативными экспертными оценками. Для прикладных же задач темы с невысокой когерентностью не представляют интереса как неинтерпретируемые в силу низкой совстречаемости образующих их термов. Для оценки качества тематической модели критически важны именно высококогерентные темы.

В зависимости от прикладной задачи для этого могут быть использованы разные метрики. В частности, при кластеризации важным условием качестватематической модели является присутствие каждой ВКТ в ненулевом числе документов коллекции, причем с достаточным весом.

Дополнительным критерием качества модели может выступать требование минимизации числа ВКТ в каждом из документов коллекции либо в большинстве из них. На практике в прикладных информационных системах достаточным является выявление в документе от одной до трех основных тем. Таким образом, на модель может быть наложено ограничение по среднему количеству ВКТ, входящих в число основных тем в одном документе коллекции.

При таком подходе всегда останутся документы, не относящиеся к ВКТ, поэтому некоторые документы коллекции при кластеризации должны быть признаны «непонятыми». Чтобы сократить их количество, можно ввести еще одно требование: число документов, содержащих ВКТ в качестве одной из основных тем, должно быть максимально, а документов без ВКТ – минимально.

В то же время для достижения этой цели недостаточно просто увеличивать долю ВКТ в модели. При довольно большом их количестве некоторые из них будут состоять из одинаковых термов. Наличие таких похожих ВКТ снижает прикладную ценность тематической модели. Поэтому необходимо выделять ВКТ с неповторяющимися термами. В ARTM это позволяет сделать регуляризатор декоррелирования.

Кроме того, если при определении тематики корпуса ориентироваться только на значение когерентности, существует риск потери автоматически выделенного кластера документов, сильно отличающихся по смыслу от остальных. Простое снижение порога когерентности в этом случае нивелирует выгоду от использования в приложениях только ВКТ, понятных пользователю. Таким образом, необходимо устанавливать гибкий порог по определению уровня когерентности.

Заключение

В статье исследуется возможность разработки метрики качества тематической модели, способной обеспечить устойчивое практическое применение тематического моделирования при решении задач обработки неструктурированных текстовых данных в прикладных интеллектуальных информационных системах. Рассматривается задача мягкой кластеризации коллекции документов.

Прикладные задачи по мягкой кластеризации коллекции документов предъявляют повышенные требования к метрике для выбора наилучшей тематической модели. Авторы показали, что использование метрики средней когерентности приводит к излишним затратам ресурсов для нахождения оптимальной тематической модели. В работе продемонстрировано, что на практике для достижения устойчивой работы приложений достаточно использовать в качестве метрики среднее число высококогерентных тем на один документ.

В рамках исследования проведен эксперимент по тематическому моделированию двух корпусов русскоязычных текстов: «Тайга» (7 695 документов) и ГОСТ (1 066 документов) по методам LDA, PLSA и ARTM. Изучено распределение когерентности тем в зависимости от параметров расчета этой метрики для каждого метода и установлено, что с увеличением общего числа тем в модели растет максимальное значение их когерентности, однако при усреднении метрики по всем темам модели эта информация теряется.

По окончании эксперимента проведена проверка на предмет возможных потерь в результатах кластеризации при исключении из рассмотрения низкокогерентных тем. На примерах продемонстрировано, что существенных потерь, критичных с точки зрения человеческой оценки, при сужении тем модели до высококогерентной их части не происходит, что подтвердило сформулированную в исследовании гипотезу.

ЛИТЕРАТУРА

1. Rubin T.N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine learning. 2012. V. 88, № 1-2. P. 157–208.
2. Янина А., Воронцов К. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных. 2016. Т. 2, № 2. С. 173–186.
3. Litvak M., Vanetik N., Liu C., Xiao L., Savas O. Improving summarization quality with topic modeling // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. 2015. P. 39–47.
4. Zhang J., Song Y., Zhang C., Liu S. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora // Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010. P. 1079–1088.
5. Ni X., Sun J.T., Hu J., Chen Z. Mining multilingual topics from wikipedia // Proceedings of the 18th international conference on World wide web. 2009. P. 1155–1156.
6. Hoffmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York : ACM, 1999. P. 50–57.

7. Vayansky I., Kumar S.A.P. A review of topic modeling methods // Information Systems. 2020. V. 94. P. 101582.
8. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation // Journal of machine Learning research. 2003. V. 3, № 1. P. 993–1022.
9. Vorontsov K.V. Additive regularization for topic models of text collections // Doklady Mathematics. 2014. V. 89, № 3. P. 301–304.
10. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open source library for regularized multimodal topic modeling of large collections // International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham., 2015. С. 370–381.
11. Krasnashchok K., Jouili S. Improving topic quality by promoting named entities in topic modeling // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018. V. 2: Short Papers. P. 247–253.
12. Omar M. et al. LDA topics: Representation and evaluation // Journal of Information Science. 2015. V. 41, № 5. С. 662–675.
13. Bhatia S., Lau J.H., Baldwin T. An automatic approach for document-level topic model evaluation // arXiv preprint. arXiv:1706.05140.2017. URL: https://arxiv.org/pdf/1706.05140.pdf
14. Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries // Proceedings of the 10th Annual Joint Conference on Digital libraries. 2010. С. 215–224.
15. Lau J.H., Newman D., Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014. С. 530–539.
16. Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. P. 262–272.
17. Aletras N., Stevenson M. Evaluating topic coherence using distributional semantics // Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) : Long Papers. 2013. P. 13–22.
18. Lau J.H., Baldwin T. The sensitivity of topic coherence evaluation to topic cardinality // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. С. 483–487.
19. Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015. С. 399–408.
20. Bezdek J.C. Cluster validity with fuzzy sets // Journal of Cybernetics. 1973. V. 3 (3). P. 58–73.
21. Dunn J.C. Well-separated clusters and optimal fuzzy partitions // Journal of cybernetics. 1974. V. 4. No. 1. С. 95–104.
22. Davies D.L., Bouldin D.W. A cluster separation measure // IEEE transactions on pattern analysis and machine intelligence. 1979. № 2. С. 224–227.
23. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering validity checking methods: part II // ACM Sigmod Record. 2002. V. 31, № 3. С. 19–27.
24. Xie X.L., Beni G. A validity measure for fuzzy clustering // IEEE Transactions on Pattern Analysis & Machine Intelligence. 1991. № 8. С. 841–847.
25. Краснов Ф.В. Оценка оптимального количества тематик в тематической модели: подход на основе качества кластеров // International Journal of Open Information Technologies. 2019. V. 7, № 2. P. 8–15.
26. Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. V. 20. P. 53–65.
27. Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010. P. 100–108.
28. Bouma G. Normalized (pointwise) mutual information in collocation extraction // Proceedings of GSCL. 2009. P. 31–40.
29. O'Callaghan D., Greene D., Carthy J., Cunningham P. An analysis of the coherence of descriptors in topic modeling // Expert Systems with Applications. 2015. V. 42, № 13. P. 5645–5657.
30. Syed S., Spruit M. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation // 2017 IEEE International conference on data science and advanced analytics (DSAA). IEEE, 2017. С. 165–174.
31. Nikolenko S.I., Koltcov S., Koltsova O. Topic modelling for qualitative studies // Journal of Information Science. 2017. V. 43, № 1. P. 88–102.
32. Shavrina T., Shapovalova O. To the Methodology of Corpus Construction for Machine Learning: "Taiga" Syntax Tree Corpus and Parser // Корпусная лингвистика : тр. Междунар. конф. СПб., 2017. С. 78–84.

Поступила в редакцию 22 декабря 2020 г.

Krasnov F.V., Baskakova E.N., Smaznevich I.S. (2020) ASSESSMENT OF THE APPLIED QUALITY OF TOPIC MODELS FOR CLUSTERING PROBLEMS. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 56. pp. 100–111

DOI: 10.17223/19988605/56/11

The average coherence is the key quality assessment parameter for topic models; it reflects most closely the human topic evaluation. Scientific literature describes a variety of methods to calculate topic coherence, each featuring their pros and cons from the point of view of both scientific validity and practical utility.

The current paper studies the potential for practical application of different coherence calculation methods in real informational systems, which were tested on two text corpora, namely «Taiga» (7695 documents) и «GOST» (1066 documents), using several algorithms of topic modeling (LDA, ARTM, PLSA). The problem of soft clustering of a text document collection was considered.

The article describes the conducted experiment aimed to reduce the total number of the surveyed model topics to the set of high-coherent topics (HCT). Various ways of documents connection to the high-, mid-, and low-coherent topics have been examined; it has been demonstrated that the applied quality of the topic model is not diminished when the low-coherent topics are removed from the scoop, while depending on the specific features of the objective and the data being processed the coherence threshold value can be reduced.

The topic coherence used to analyze the results of transition from the complete model to the set of HCT was calculated using the following formula, determined as the most valid one for the task:

$$PPMI\left(w_{di}, w_{dj}\right) = \left[\log \frac{n\left(w_{di}, w_{dj}\right) n}{n\left(w_{di}\right) n\left(w_{dj}\right)}\right]_{+}, \text{ where } n\left(w_{di}, w_{dj}\right) = \sum_{d=1}^{|D|} \sum_{i=1}^{N_d} \sum_{j=1}^{N_d} \left[0 < |i-j| \le k\right] \text{ is the } CoocTF \text{ value, defining the co-}$$

occurrence of the terms w_i and w_i within the sliding window of a preset width in all the document; $|D|$ is the document collection size, N_d is the volume of the document d.

The main outcome of the research is the conclusion that topic model application for practical use in informational systems requires focusing not on the average coherence values, but on the topics featuring high values of coherence. To assess the applied quality of a topic model, a complex procedure was developed based on topic coherence calculating and involving some extra criteria. It demonstrated better effectiveness in soft clustering of a text collection, than evaluating a topic model by the metrics of its average coherence.

Keywords: topic modeling; topic coherence; soft clustering; text analysis; ARTM.

KRASNOV Fedor Vladimirovich (Candidate of Technical Sciences, Expert, Researcher, Department of Information Management Systems, NAUMEN R&D, Yekaterinburg, Russian Federation).
E-mail: fkrasnov@naumen.ru

BASKAKOVA Elena Nikolaevna (Leading System Analyst, Department of Information Management Systems, NAUMEN R&D, Yekaterinburg, Russian Federation).
E-mail: enbaskakova@naumen.ru

SMAZNEVICH Irina Sergeevna (Business Analyst, Department of Information Management Systems, NAUMEN R&D, Yekaterinburg, Russian Federation).
E-mail: ismaznevich@naumen.ru

REFERENCES

1. Rubin, T.N., Chambers, A., Smyth, P. & Steyvers, M. (2012) Statistical topic models for multi-label document classification. *Machine Learning*. 88(1-2). pp. 157–208. DOI: 10.1007/s10994-011-5272-5
2. Yanina, A. & Vorontsov, K. (2016) Multimodal topic models for exploratory search in a collective blog. *Mashinnoe obuchenie i analiz dannykh – Machine Learning and Data Analysis*. 2(2). pp. 173–186. DOI: 10.21469/22233792.2.2.04
3. Litvak, M., Vanetik, N., Liu, C., Xiao, L. & Savas, O. (2015) Improving summarization quality with topic modeling. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*. pp. 39–47. DOI: 10.1145/2809936.2809944
4. Zhang, J., Song, Y., Zhang, C. & Liu, S. (2010) Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1079–1088. DOI: 10.1145/1835804.1835940
5. Ni, X., Sun, J. T., Hu, J. & Chen, Z. (2009) Mining multilingual topics from Wikipedia. *Proceedings of the 18th International Conference on World Wide Web*. pp. 1155–1156. DOI: 10.1145/1526709.1526904
6. Hoffmann, T. (1999) Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM. pp. 50–57.
7. Vayansky, I. & Kumar, S.A.P. (2020) A review of topic modeling methods. *Information Systems*. 94. pp. 101–582. DOI: 10.1016/j.is.2020.101582
8. Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*. 3(1). pp. 993–1022.
9. Vorontsov, K. (2014) Additive regularization for topic models of text collections. *Doklady Mathematics*. 89(3). pp. 301–304. DOI: 10.1007/s10994-014-5476-6
10. Vorontsov, K., Frei, O., Apishev, M., Romov, P. & Dudarenko, M. (2015) BigARTM: Open source library for regularized multimodal topic modeling of large collections. *International Conference on Analysis of Images, Social Networks and Texts*. Springer, Cham. pp. 370–381.

11. Krasnashchok, K. & Jouili, S. (2018) Improving topic quality by promoting named entities in topic modeling. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers)*. pp. 247–253. DOI: 10.18653/v1/P18-2040
12. Omar, M. et al. (2015) LDA topics: Representation and evaluation. *Journal of Information Science*. 41(5). pp. 662–675. DOI: 10.1177/0165551515587839
13. Bhatia, S., Lau, J.H. & Baldwin, T. (2017) An automatic approach for document-level topic model evaluation. *ArXiv preprint* arXiv:1706.05140. [Online] Available from: https://arxiv.org/pdf/1706.05140.pdf
14. Newman, D., Noh, Y., Talley, E., Karimi, S. & Baldwin, T. (2010) Evaluating topic models for digital libraries. *Proceedings of the 10th annual joint conference on Digital libraries*. pp. 215–224. DOI: 10.1145/1816123.1816156
15. Lau, J.H., Newman, D. & Baldwin, T. (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 530–539. DOI: 10.3115/v1/E14-1056
16. Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011) Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp. 262–272.
17. Aletras, N. & Stevenson, M. (2013) Evaluating topic coherence using distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. pp. 13–22.
18. Lau, J.H. & Baldwin, T. (2016) The sensitivity of topic coherence evaluation to topic cardinality. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 483–487. DOI: 10.18653/v1/N16-1057
19. Röder, M., Both, A. & Hinneburg, A. (2015) Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*. pp. 399–408. DOI: 10.1145/2684822.2685324
20. Bezdek, J.C. (1973) Cluster validity with fuzzy sets. *Journal of Cybernetics*. 3(3). pp. 58–73. DOI: 10.1080/01969727308546047
21. Dunn, J.C. (1974) Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*. 4(1). pp. 95–104. DOI: 10.1080/01969727408546059
22. Davies, D.L. & Bouldin, D.W. (1979) A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*. 2. pp. 224–227.
23. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002) Clustering validity checking methods: part II. *ACM Sigmod Record*. 31(3). pp. 19–27. DOI: 10.1145/601858.601862
24. Xie, X.L. & Beni, G. (1991) A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. 8. pp. 841–847. DOI: 10.1109/34.85677
25. Krasnov, F. (2019) Evaluation of Optimal Number of Topics of Topic Model: An Approach Based on the Quality of Clusters. *International Journal of Open Information Technologies*. 7(2). pp. 8–15.
26. Rousseeuw, P.J. (1987) Silhouettes a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 20. pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7
27. Newman, D., Lau, J.H., Grieser, K. & Baldwin, T. (2010) Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 100–108.
28. Bouma, G. (2009) Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*. pp. 31–40.
29. O'Callaghan, D., Greene, D., Carthy, J. & Cunningham, P. (2015) An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*. 42(13). pp. 5645–5657. DOI: 10.1016/j.eswa.2015.02.055
30. Syed, S. & Spruit, M. (2017) Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. pp. 165–174. DOI: 10.1109/DSAA.2017.61
31. Nikolenko, S., Koltsova, O. & Koltsov, S. (2017) Topic modelling for qualitative studies. *Journal of Information Science*. 43(1). pp. 88–102. DOI: 10.1177/0165551515617393
32. Shavrina, T. & Shapovalova, O. (2017) To the Methodology of Corpus Construction for Machine Learning: "Taiga" Syntax Tree Corpus and Parser. *Korpusnaya lingvistika* [Corpus linguistics]. Proc. of the International Conference. pp. 78–84.