Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 274 ГРУППЫ

«Суммаризация тем в вероятностных тематических моделях»

Выполнил:
студент 4 курса 274 группы
Смирнов Евгений Александрович

Научный руководитель:
профессор РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва, 2016

Содержание

Аннотация

Одной из главных проблем вероятностых тематических моделей является их понимание. Все существующие методы оценки интерпретируемости тем основываются на методе описания мешком терминов. В данной работе предлагается подход для оценки интерпретируемости, основанный на анализе сжатого представления коллекции документов. Для этого объявляется набор требований к тематической модели, для того чтобы считать её интерпретируемой. Для интерпретируемой модели формируется суммаризация тем — список предложений для каждой темы, наиболее точно и полно её описывающий, отранжированный по ценности предложений. В вычислительных экспериментах строится интерпретируемая тематическая модель для коллекции документов конференции ММРО и суммаризация её тем.

Ключевые слова: *интерпретируемость, суммаризация ,когерентность, ARTM, affinity propagation.*

1 Введение

Интерпретируемость тем является одной из острых проблем вероятностного тематического моделирования. Тема считается интерпретируемой, если по предоставленному описанию эксперт может понять, о чём эта тема, и дать ей адекватное название. Мешок терминов, состоящих либо из слов, либо из словосочетаний, является общепринятым методом описания темы.

Большинство методов основаны на привлечении экспертов-асессоров. В работе [1] экспертам предлагалось непосредственно оценивать полезность тем по трёхбалльной шкале. В методе интрузий [2] для каждой найденной темы составляется список из 10 наиболее частотных слов, в который внедряется одно случайное слово. Тема считается интерпретируемой, если подавляющее большинство экспертов правильно указывают лишнее слово. Экспертные подходы необходимы на стадии исследований, но они затрудняют автоматическое построение тематических моделей. В серии работ [1,3,4] был предложен ряд величин, вычисляемых по коллекции автоматически, которые оценивали качество интерпретируемости. Было показано, что среди них величина когерентности лучше остальных коррелирует с экспертными оценками интерпретируемости. Эта величина оценивает частоту наиболее вероятных слова темы встретиться рядом в документах данной коллекции или во внешней политематической коллекции, такой, как Википедия. Когерентность на сегодняшний день остается основной мерой интерпретируемости, вычисляемой автоматически.

Существенным недостатком такого метода оценки интерпретируемости является тот факт, что список топовых терминов темы, вырванных из контекста, не всегда позволяет судить об интерпретируемости темы. В данной работе для оценки интерпретируемости тем предлагается анализировать тексты коллекции документов. Большие объёмы текстов являются значительной проблемой с которой сталкивается такой подход. Естественным решением является анализ сжатого представления исходных документов. Процесс формирования краткого описания отдельного документа называется задачей аннотирования, формирования сжатого представления коллекции — задачей суммаризации [9]. Каждая из тем модели не обязательно относится к какому-то одному конкретному документу коллекции, поэтому для понимания темы нужно располагать кратким содержанием всей коллекции в целом. Для оценки интерпретируемости темы предлагается сформировать наиболее представительный набор предложений, максимально точно и полно описывающий тему, и решить задачу суммаризации для этого набора.

К тематической модели выдвигается ряд требований, для того чтобы считать её подходящей для построения суммаризации, такая модель называется интерпретируемыми. Из предложений, составляющих документы коллекции, для каждой темы выбирается набор наиболее релевантных предложений. С целью устранения дублирования смысла осуществляется их кластеризация. Список суммаризации темы составляется из наиболее ценных предложений кластеров, отранжированных по их ценности. Для измерения качества суммаризации темы вводятся аналогичные задаче классификации оценки точности и полноты, а также оценка представительности суммаризации темы.

Цель данной работы — разработать алгоритм суммаризации тем и методы количественной оценки его качества. Входными данными алгоритма являются тематическая модель и коллекция документов. Алгоритм должен возвращать список ре-

левантных предложений темы, отсортированный так, чтобы первые k его элементов представляли её наиболее полное и точное описание.

В главе 2 ставится задача вероятностного тематического моделирования и описывается метод аддитивной регуляризации тематических моделей ARTM [6]. В главе 3 предлагается последовательность шагов построения алгоритма суммаризации тем. В 4 главе описывается вычислительный эксперимент, в котором строится интерпретируемая тематическая модель и суммаризация её тем.

2 Вероятностная тематическая модель

Пусть D — коллекция текстовых документов, W — множество всех употребляемых в документах терминов. Будем рассматривать каждый документ $d \in D$ как последовательность из n_d терминов $(w_1, w_2, \ldots, w_{n_d})$ из словаря W. Обозначим число вхождений термина w в документ d через n_{dw}.

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как случайная и независимая выборка троек $\{(w_i, d_i, t_i)\}_{i=1}^n$ из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, темы $t \in T$ — *латентными*(скрытыми).

Полагая, что появление термина w в документе d зависит только от темы t и воспользовавшись формулой полной вероятности получим вероятностную модель порождения данных

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d).$$

Обычно число тем $|T|$ много меньше размера коллекции $|D|$ и словаря $|W|$, поэтому задача сводится к поиску приближённого представления заданной матрицы частот

$$F = (\hat{p}_{wd})_{W \times D}\,,\ \hat{p}_{wd} = \hat{p}(w|d) = \frac{n_{dw}}{n_d},$$

в виде произведения $F \approx \Phi\Theta$ двух матриц меньшего размера — *матрицы терминов тем* Φ и *матрицы частот документов* Θ:

$$\Phi = (\phi_{wt})_{W \times D}\,,\ \phi_{wt} = p(w|t),\ \phi_t = (\phi_{wt})_{w \in W}\,,$$

$$\Theta = (\theta_{td})_{T \times D}\,,\ \theta_{td} = p(t|d),\ \theta_d = (\theta_{td})_{t \in T}\,.$$

Матрицы Φ и Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы, представляющие дискретные распределения. Для нахождения параметров Φ и Θ максимизируется логарифм правдоподобия выборки при ограничениях нормировки и неотрицательности:

$$L(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} p\,(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta}; \tag{1}$$

$$\sum_{w \in W} \phi_{wt} = 1,\ \phi_{wt} \geq 0,\ \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geq 0. \tag{2}$$

Серьёзной проблемой при решении этой задачи является неединственность и неустойчивость решения. Правдоподобие (1) зависит только от произведения $\Phi\Theta$, которое определено с точностью до линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические. Требуется метод, позволяющий контролировать преобразование S.

2.1 Аддитивная регуляризация ARTM

Для решения проблемы неединственности и неустойчивости используется подход, основанный на многокритериальной регуляризации ARTM. Он позволяет строить

модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора — оптимизационного критерия $R_i(\Phi, \Theta)$, зависящего от параметров модели. Задача сводится к максимизации линейной комбинации критериев $L(\Phi, \Theta)$ и $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \; R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta). \tag{3}$$

2.2 Регуляризаторы

Будем использовать дивергенцию Кульбака–Лейблера как меру различия двух дискретных распределений $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$:

$$KL(p||q) \equiv KL_i(p||q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

Минимизация KL-дивергенции эквивалентна максимизации правдоподобия модели распределения q по эмпирическому распределению p.

Сглаживающий регуляризатор минимизирует различие между распределениями ϕ_t, θ_d и заданными $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_t)_{t \in T}$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Разреживающий регуляризатор основывается на предположении, что каждый документ и каждый термин связан с небольшим числом тем. Он максимизирует KL-дивергенцию между модельными распределениями ϕ_t,θ_d и заданными распределениями $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_t)_{t \in T}$ с большой энтропией, например, равномерным:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Декоррелирующий регуляризатор повышает различность тем, что улучшает интерпретируемость модели. Он минимизирует ковариации между вектор-столбцами ϕ_t, ϕ_s:

$$R(\Theta, \Phi) = -\gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

3 Суммаризация тем

Суммаризация темы — список релевантных предложений темы, отсортированный так, чтобы первые его k элементов представляли ее наиболее полное и точное описание. Неформально к группе релевантных предложений темы следует относить предложения, относящиеся только к этой теме и в совокупности максимально полно представляющие данную тему. Для лучшей интерпретации от предложений будем требовать грамматическую связность и содержание законченной мысли (самодостаточность).

Для формализации требований к алгоритму суммаризации темы t предлагается алгоритм, основанный на следующей последовательности шагов:

1. Выделить наиболее тематичные предложения, точно характеризующие тему.
2. Кластеризовать предложения с целью устранения дублирования смысла.
3. Отобрать наиболее ценные предложения в каждом из кластеров.
4. Отранжировать предложения по максимально быстрому возрастанию суммарной ценности первых k предложений.

Низкое качество интерпретируемости суммаризации темы может быть следствием либо метода построения тематической модели, либо суммаризации темы. В случае низкого качества построенной модели, суммаризация темы является мощным инструментом анализа её интерпретируемости. Предъявим ряд требований к тематической модели, для которой будем в дальнейшем строить суммаризацию тем.

3.1 Построение интерпретируемых моделей

Интерпретируемость тематической модели плохо формализуемое требование. Будем предполагать, что тема хорошо интерпретируема, если практически все термины, относящиеся к ней, не содержатся в других темах. Следовательно из бесконечного множества стохастических матричных разложений $F \approx \Phi\Theta$ требуется рассматривать только те, в которых матрицы Φ и Θ разреженные.

Разобьём множество тем на два непересекающихся подмножества — предметных S и фоновых B тем. *Предметные темы* S состоят из терминов предметных областей, а *фоновые темы* B из слов общей лексики Распределения $p(w|t)$ предметных тем разрежены и декоррелированы, а фоновых сглажены, так как слова из этих тем содержаться во многих документах коллекции.

Для получения модели состоящей из фоновых и предметных тем предлагается комбинация из пяти регуляризаторов: сглаживание фоновых тем в матрицах Φ и Θ, разреживание предметных тем в матрицах Φ и Θ, и декоррелирование предметных тем в матрице Φ:

$$\begin{aligned} R(\Phi,\Theta) \;\; &= -\beta_0 \sum_{t\in S}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d\in D}\sum_{t\in S} \alpha_t \ln \theta_{td} \\ &+\beta_1 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_1 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \\ &-\gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max. \end{aligned}$$

Под тематическим профилем термина w будем понимать вероятность того, что термин w относится именно к теме t:

$$p(t|w) = \frac{\phi_{wt}}{\sum\limits_{t \in S} \phi_{wt}}$$

Назовём множеств *якорных* терминов темы $t \in S$ множество терминов, которые встречаются только в ней:

$$A_t = \{w | p(t|w) = 1\},$$

множеством *не якорных* терминов — множество терминов w, встречающихся не только в теме t:

$$\bar{A}_t = \{w | 0 < p(t|w) < 1\},$$

Лексическим ядром темы t будем называть объединения множеств A_t и $\bar{A}_t$.

Покрытием вероятности темы t множеством терминов $W_n = \{w_i\}_{i=1}^n$ будем называть:

$$C_t(W_n) = \sum_{i=1}^{n} p(w_i|t)$$

Покрытием термина w будем считать мощность множество тем $t \in S$, к которым он относится:

$$C_w = |\{t \in S | p(w|t) \neq 0\}|$$

Формально будем считать предметную тему интерпретируемой, если покрытие вероятности якорными словами близко к единице и вероятностная мера пересечений с другими темами близка к нулю. Тематическую модель будем называть интерпретируемой, если она удовлетворяет следующим условиям:

1. Все предметные темы являются интерпретируемыми.
2. Значение покрытий всех терминов тем незначительно отличаются от единицы.
3. Доля терминов, у которых значение покрытий больше единицы крайне мала.

3.2 Фильтрация тематичных предложений

Будем говорить, что термин w входит в *ядро темы* t, если частотная оценка его встречаемости в теме не равна нулю:

$$\mathcal{K}(t) = \{w \in W \mid p(w|t) \neq 0\}.$$

Пусть дано множество предложений $\{s_1, s_2, \ldots, s_{n_s}\}$, где n_s число предложений во всех документах коллекции. Каждое предложение $s = \{w_i\}_{i=1}^{|s|}$ описывается мешком слов из словаря W. Введём ряд определений для описания свойств тематичных предложений.

Тематический профиль предложения s — среднее значение профилей его терминов:

$$p(t|s) = \frac{1}{|s|} \sum_{w \in s} p(t|w)$$

Информативность предложения s относительно темы t — число слов из ядра $\mathcal{K}(t)$:

$$\mathrm{I}(s,t) = \sum_{w \in W} [w \in \mathcal{K}(t)].$$

Тематичность предложения s относительно темы t — близость его тематического профиля по расстоянию косинус к вырожденному распределению этой темы:

$$\mathrm{T}(s,t) = 1 - \frac{p(t|s)}{\|p(t|s)\|_2}.$$

Пусть $\forall s,t\ \mathrm{T}(s,t) < \mathrm{T}_0$ тематические профиль предложения s относительно темы t достаточно близок к вырожденному распределению темы t, тогда T_0 будем называть *порогом тематичности*.

Пусть $\forall s,t\ \mathrm{I}(s,t) > \mathrm{I}(s,t)$ предложение s содержит достаточное число терминов из ядра $\mathcal{K}(t)$ для точного описания темы t, тогда I_0 будем называть *порогом информативности*.

Пусть заданы T_0 и I_0, тогда предложения s будем называть *тематичным* для темы t, если $\mathrm{T}(s,t) < \mathrm{T}_0$ и $\mathrm{I}(s,t) > \mathrm{I}_0$. Таким образом фильтрация тематичных предложении сводиться к выбору параметров T_0 и I_0.

3.3 Кластеризация предложений

Среди тематичных предложений могут оказаться достаточно схожие по смыслу, что может существенно усложнить выбор наиболее ценных из них. Предлагается кластеризовать тематичные предложения для сокращения их числа без потери качества интерпретируемости.

Расстояние между предложениями s_i и s_j относительно темы t можно оценить по их мешкам слов из ядра темы $K(t)$ с помощью коэффициента Жаккара :

$$\mathcal{J}(s_i, s_j|t) = \frac{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cap s_j]}{\sum\limits_{w \in \mathcal{K}(t)} [w \ \in s_i \cup s_j]}$$

Будем считать, что предложения тем более похожи, чем больше содержат одинаковых высоковероятных слов из ядра темы. Для того чтобы формализовать это предположение будем оценивать расстояния между предложениями посредством модифицированного коэффициента Жаккара:

$$\mathcal{J}`(s_i, s_j|t) = \frac{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cap s_j] p(t|w)}{\sum\limits_{w \in \mathcal{K}(t)} [w \in s_i \cup s_j] p(t|w)}$$

Центры кластеров, являющиеся одновременно объектами выборки будем называть экземплярами. Многие популярные методы кластеризации, основанные на выборе k центров задают на первом шаге начальные положения экземпляров случайным

образом, затем итеративно переопределют их таким способом, чтобы новый набор давал меньшую ошибку. Такой подход чувствителен к начальному приближению, поэтому принято перезапускать алгоритм несколько раз, назначая другие объекты выборки экземплярами для последующего выбора лучшего решения. Тем не менее, такой метод работает лишь в том случае, когда число кластеров мало и как следствие шансы одной случайной инициализации сойтись к хорошему решению велики. Заведомо можно считать, что большая коллекции документов образует значительное число кластеров, что означает неприменимость такого подхода для решения нашей задачи.

Предлагается использовать алгоритм affinity propagation [5], основанный на подходе, полагающем на каждой итерации все точки выборки возможными экземплярами. Рассматривая каждую точку как узел сети, метод рекурсивно посылает весовые сообщения по рёбрам, пока набор экземпляров и кластеров достаточно хорош. В каждый момент времени вес сообщения отражает текущее правдоподобие того, что одна вершина является экземпляром другой.

Входными данными для алгоритма являются оценки схожести точек $s(i,k)$. Они отражают насколько хорошо точку k следует выбирать экземпляром для точки i. Алгоритм не требует на входе выбирать число кластеров, однако, присутствует возможность задавать предпочтения в выборе экземпляров. Точка k будет выбрана в качестве экземпляра тем более вероятнее, чем больше значение $s(k,k)$.

Точки обмениваются между собой двумя типами сообщений, каждое из которых отражает один из видов конкуренции между точками. После каждой итерации алгоритма возможно определить экземпляры и относящиеся к ним точки. Вес сообщения $r(i,k)$ отправленного точкой i кандидату k быть её экземпляром отражает насколько подходящим является выбор k в роли экземпляра i по сравнению с другими кандидатами. Вес сообщения $a(i,k)$ от кандидата k в экземпляры точки i отражает насколько свойственно для i выбирать k экземпляром, рассматривая другие точки выбравшие k экземпляром.

Алгоритм affinity propagation:

1. $\forall i,k\ a(i,k) \leftarrow 0$

2. Повторять до сходимости:

   $\forall i,k\ r(i,k) \leftarrow s(i,k) - \max\limits_{k' \neq k}\{a(i,k') + s(i,k')\}$

   $\forall i \neq k, k\ a(i,k) \leftarrow \min\left(0, r(k,k) + \sum\limits_{i' \notin \{i,k\}} \max(0, r(i',k))\right)$

   $\forall k\ a(k,k) \leftarrow \sum\limits_{i' \neq k} \max(0, r(i',k)$

3. Центроиды кластеров: $c = \{c_1, \ldots, c_N\}$, где

   $c_i = \arg\max\limits_k [a(i,k) + r(i,k)]$

Для применения алгоритма к решению задачи кластеризации тематичных предложений в качестве функции схожести будем использовать ранее определённый коэффициент Жаккара следующим образом $s(i,j|t) = 1 - \mathcal{J}'(i,j|t)$.

3.4 Ранжирование предложений

Суммаризация темы, должна быть сформирована из наиболее ценных предложений кластеров, отранжированных таким образом, чтобы первые k из них максимально полно описывали тему. Ранжировать список выдачи суммаризации будем исходя из оценок *ценности предложений* и *ценности кластеров*, их содержащих.

Наиболее важными признаками для определения ценности предложений являются тематичность и информативность, которые показывают какую долю предложения составляют слова из ядра темы и вероятность того, что предложение является представителем именно данной темы. Ценность предложения зависит от его положения в тексте, находящиеся ближе к началу документа\раздела довольно часто содержат основную мысль. Качество суммаризации темы существенно зависит от длины её предложений. Очень длинные предложения, наполненные большим числом лингвистических приёмов нежелательны, так как они увеличивают время работы асессора. Поэтому число слов в предложениях и число знаков препинания являются важными признаками в формировании списка суммаризации темы. Хорошими лингвистическими признаками ценности предложения являются синтаксическая связность и отсутствие анафоры.

Число предложений, составляющих кластер, а также число документов, к которым принадлежат эти предложения будем считать важнымы признаками, характеризующими ценность кластера. Чем плотнее кластер, тем выше его ценность. Плотность кластера будем оценивать долей предложений содержащих k общих слов.

Пусть для списков суммаризаций тем, имеются экспертные оценки релевантности предложений, тогда задача ранжирования предложений суммаризации сводится к подбору весовых коэффициентов при признаках. Такая задача решается любым из алгоритмов классификации.

3.5 Оценка качества суммаризации

Будем обозначать $S(k,t)$ суммаризацию темы t, состоящую из k предложений. Для оценки качества суммаризаций $S(k,t)$ предлагается ряд характеристик, по которым будем определять достаточно ли хороша модель для отправки на оценку экспертами-асессорами. Введём понятия точности и полноты суммаризации аналогично тому, как они вводятся в задаче классификации.

Будем считать, что чем больше суммарная вероятность слов из предложений суммаризации $S(k,t)$, входящих в ядро темы, тем сильнее суммаризация покрывает эту тему. Формально полноту суммаризации $S(k,t)$ будем вычислять следующим образом:

$$\sum_{w \in W} [w \in S(k,t)] p(w|t)$$

Будем считать, что чем больше доля вероятности слов, составляющих суммаризацию $S(k,t)$, входящих в ядро темы, тем точнее она описывает эту тему. Вероятность слов не входящих в ядро темы, будем оценивать их математическим ожиданием попасть в ядро других тем. Формально точность суммаризации $S(k,t)$ определим следующим образом:

$$\frac{\sum\limits_{i=1}^{k}\sum\limits_{w\in s_i}[w\in \mathrm{K}(t)]p(w|t)}{\sum\limits_{i=1}^{k}\sum\limits_{w\in s_i}[w\in \mathrm{K}(t)]p(w|t)+\mathbb{E}_{t`\neq t}p(w|t`)}$$

Будем говорить, что суммаризация темы согласуется с лексическим ядром темы, если распределение вероятности слов суммаризации темы сходится к распределению слов лексического ядра темы. Меру согласованности суммаризации темы $S(k,t)$ с ядром темы $(K)(t)$ будем называть *представительностью* и формально определим следующим образом:

$$\cos(p(w|S(k,t)),p(w|t))$$

Сравнивать интерпретируемости тем по топовым словам и списку суммаризации темы предлагается на оснований экспертных оценок прилагаемой анкеты.

Анкета для экспертного оценивания интерпретируемости тем и качества суммаризации:

1. Число интерпретаций темы по списку топовых слов
2. Число интерпретаций темы по списку топовых предложений
3. Число общих интерпретаций по обоим спискам
4. Число неинтерпретируемых предложений в суммаризации

Для проверки корректности введённых количественных оценок качества и полноты следует сравнить их с точностью и полнотой, полученными на основании оценок экспертов.

4 Вычислительный эксперимент

Вычислительные эксперименты производились на данных конференции ММРО. Была построена интерпретируемая тематическая модель коллекции документов. Для полученной модели была построена суммаризации тем. Из суммаризаций тем для экспертного оценивания была выбрана суммаризация на основании количественных оценок качества точности и полноты.

Описание данных. Параметры коллекции документов конференции "Математические методы распознавания образов"(ММРО):

- Размер коллекции $|D| = 1061$
- Число предложений ≈ 70999
- Размер словаря $|W| = 7805$

Среда разработки. Для реализации алгоритма использовался язык Python. Построение тематической модели осуществлялось методами библиотеки BigARTM [7]. Для кластеризации предложений использовался алгоритм affinity propagation из библиотеки scikit-learn [8].

4.1 Построение интерпретируемой модели

В эксперименте была построена тематическая модель, состоящая из 50 предметных и 15 фоновых тем. На первых 50 итерациях использовались регуляризаторы сглаживания фоновых тем и декоррелирования предметных тем. На следующих 50 итерациях производилось разреживание предметных тем. Матрицы Φ и Θ, построенной модели, получилось достаточно разреженными, что видно из Рис. 1.

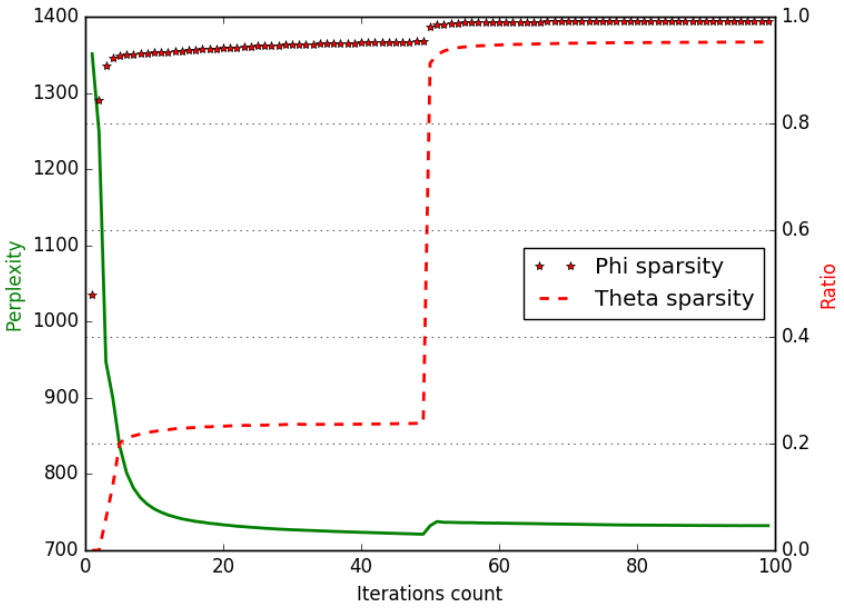
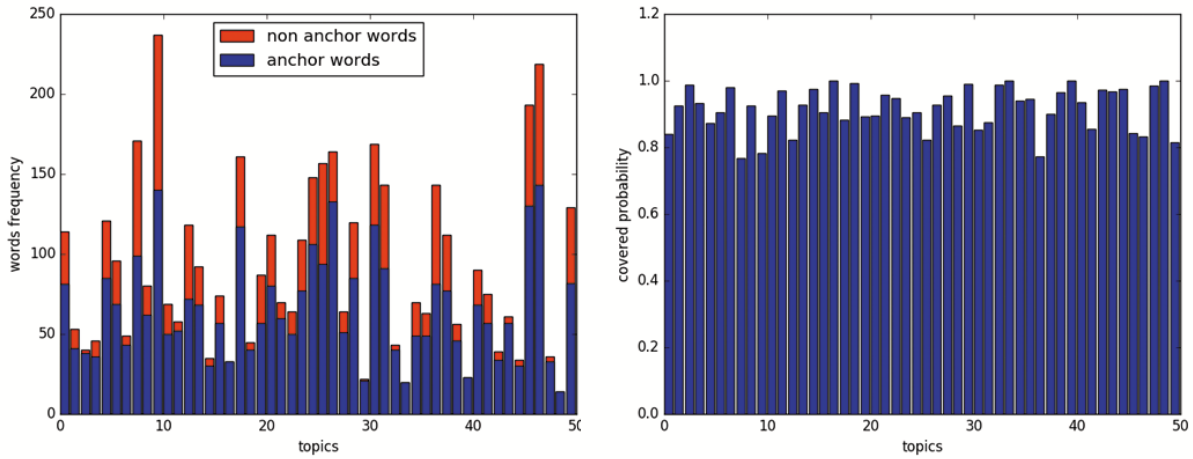


Рис. 1: Зависимость перплексии и разреженности матриц Φ и Θ от числа итераций

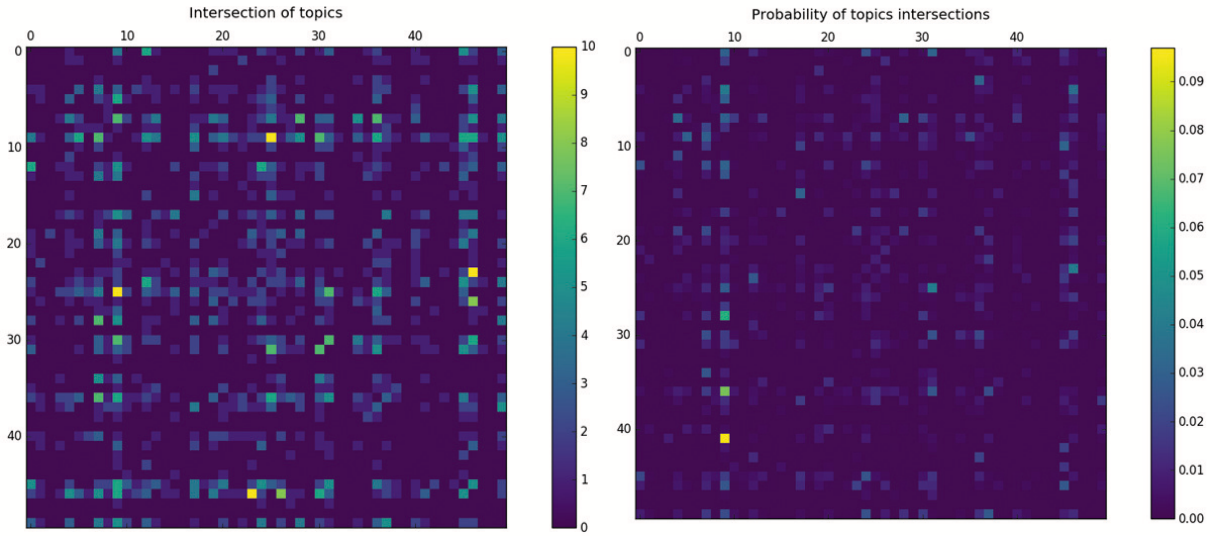
Из Рис. 2 видно, что практически во всех темах доля якорных слов составляет более 65%, причём покрытие вероятности каждой темы более 75%. Число не якорных слов существенно зависит от мощности лексического ядра темы, при крайне больших размерах лексического ядра число не якорных слов становится значительным.



(a) (b)

Рис. 2: Число якорных слов(a) и покрытия вероятностей(b) тем

Темы пересекаются не более чем по 10 терминам, причём вероятностная мера их пересечения составляет не более 0.1, согласно Рис. 3. Доля терминов, у которых покрытие больше единицы мало, как видно из Рис. 4. Таким образом построенная модель удовлетворяет предъявленным требованиям интерпретируемости.



(a) (b)

Рис. 3: Мощность(a) и вероятностная мера(b) пересечения тем

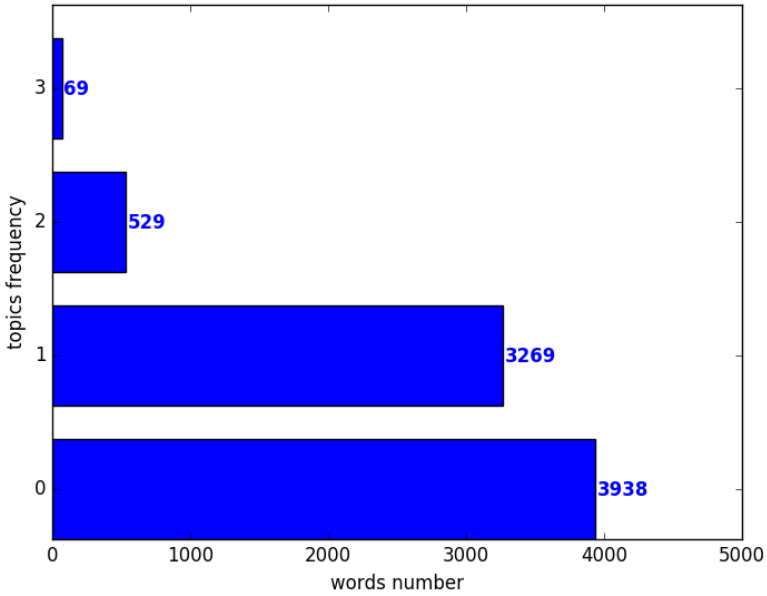
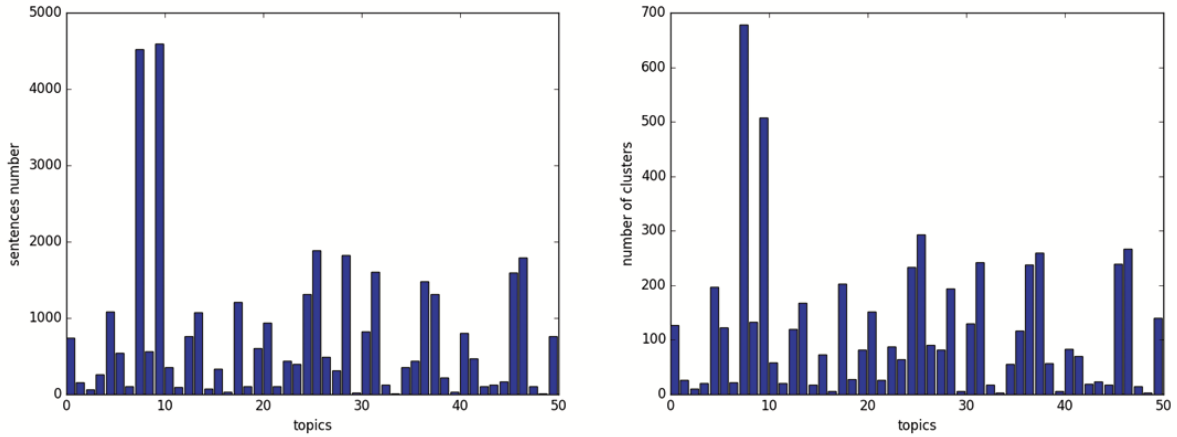


Рис. 4: Покрытия терминов

4.2 Суммаризация тем

Предобработка данных. Из множества предложений были удалены предложения, состоящие только из слов фоновых тем, а также предложения, у которых мешок слов содержал менее двух слов. После фильтрации множество содержало 54808 предложений.

Фильтрация предложений. Для каждой темы было сформировано множество тематичных предложений. В множество были отобраны предложения, содержащие не менее двух слов из ядра темы и тематичностью не более 0.5. Число тематичных предложений, как и число образовавшихся кластеров, существенно зависит от числа якорных слов в теме, что видно из Рис. 5 и Рис. 2.



(a) (b)

Рис. 5: Число тематичных предложений(a) и кластеров(b)

Качество кластеризации. Для каждой точки i из кластера c, положим $a(i)$ равным расстоянию до ближайшие точки из кластера c, за $b(i)$ обозначим минимальное расстояние до точки, не принадлежащей кластеру c. Коэффициент Silhouette показывает насколько похожие точки лежат в одном кластере вместе с точкой i:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Silhouette score является средним значением коэффициента Silhouette по всей выборке и показывает насколько точно построились границы между кластерами.

Предложения, объединённые в один кластер, должны быть крайне похожи между собой. Для оценки меры близости предложений вычисляется вероятностная мера попарных мер схожести предложений, образующих один кластер. Описанную оценку качества будем называть *плотностью кластеризации* и вычислять следующим образом:

$$1 - \mathbb{E}_c \mathcal{J}`(i, j|t)$$

Согласно Рис. 6 вероятностная мера пересечения в кластерах выше 0.3, значения Silhouette score превышают 0.5. Это означает, что алгоритм объединяет в кластеры довольно похожие предложения.

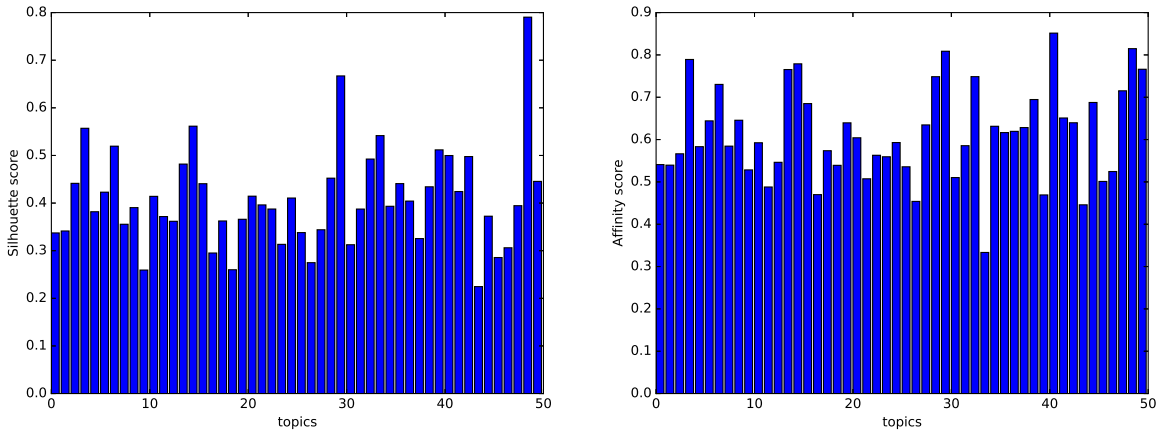


Рис. 6: Меры качества кластеризации

Ранжирование предложений. Для формирования списка суммаризации тем в качестве наиболее ценных предложений были выбраны экземпляры, отранжированные по ценности. В качестве функции ценности было выбрано произведение тематичности экземпляра на мощность кластера по числу предложений.

В качестве примера для оценки интерпретируемости рассмотрим список топовых слов и суммаризацию темы №26, жирным шрифтом выделены слова из лексического ядра темы. По списку топовых терминов, данная тема интерпретируется, как относящаяся к области здравоохранения. Из суммаризации темы становится ясным, что тема включает себя: «выявление групп психологического риска среди беременных женщин», «тональный и темпоральный анализ» и «интеллектуальную поддержку принятий решений для врачей».

Список топ слов темы №26:
заболевание, больной, врач, речь, звук, лечение, симптом, речевой, медицинский, акустический, диагноз, высота, гласный, беременный, диагностик, обследование, депрессия, эмоциональный, психологический, согласный

Суммаризация темы №26:

1. *Обычно для* ***установления диагноза*** *проводится биопсия.*
2. *В* ***женской консультации*** *роддома 1 г.*
3. *Грудная и сердечно-****сосудистая хирургия****.*
4. *Данные клинического* ***обследования*** *этих* ***больных*** *были систематизированы по параметрам (признакам для распознавания).*
5. *О признаках* ***эмоций*** *в* ***речи***
6. ***Тональный*** *и* ***темпоральный*** *анализ.*
7. *Поскольку необходимо осуществлять широкомасштабную* ***диагностику*** *и* ***профилактику депрессии*** *для вышеупомянутого контингента, то необходимость создания интеллектуальной системы* ***экспресс****-диагностики и профилактики депрессии не вызывает* ***сомнений****.*
8. *На первом этапе несколько* ***врачей*** *независимо указали крайне* ***тяжелых*** *новорожденных и* ***благополучных новорожденных****.*
9. *Данная работа посвящена исследованию* ***применимости*** *существующих* ***артикуляционных*** *моделей для анализа фонетических и* ***паралингвистических*** *особенностей* ***речи****, выявлению корреляций между* ***лингвистическими*** *признаками* ***эмоций*** *и их* ***акустическими*** *параметрами, построению алгоритмов распознавания* ***эмоционального*** *состояния* ***говорящего****.*
10. *Выявление групп* ***психологического*** *риска среди* ***беременных женщин****.*
11. *Траектория* ***высоты звука*** *для высказывания «Сама понимаешь».*
12. *Замечено, что* ***гласные****, на которые* ***приходится акцент*** *уровня фразы, могут иметь два максимума интенсивности, что объясняется особенностями* ***эмоциональной фразовой акцентуации****.*
13. ***Врач*** *в процессе* ***лечения больного*** *часто принимает решения на основе недостаточной информации, когда статистические методы малоэффективны.*
14. *Интеллектуализация поддержки принятия диагностических и* ***профилактических*** *решений по* ***депрессии*** *на основе интеллектуальной системы* ***ДИАПРОД****.*

Оценка качества суммаризации. Рассмотрим две темы, существенно отличающиеся по размеру лексического ядра Таблица 1. Из Рис. 7 и Рис. 8 видно, что количественные оценки точности и полноты зависят от размера лексического ядра темы.

Тема	№1	№9
Размер лексического ядра	237	53
Число якорных слов	140	41
Число тематичных предложений	4597	153
Число кластеров	507	26
Silhouette score	0.34	0.26
Плотность кластеризации	0.46	0.47

Таблица 1: Сравнение характеристик тем №1 и №9.

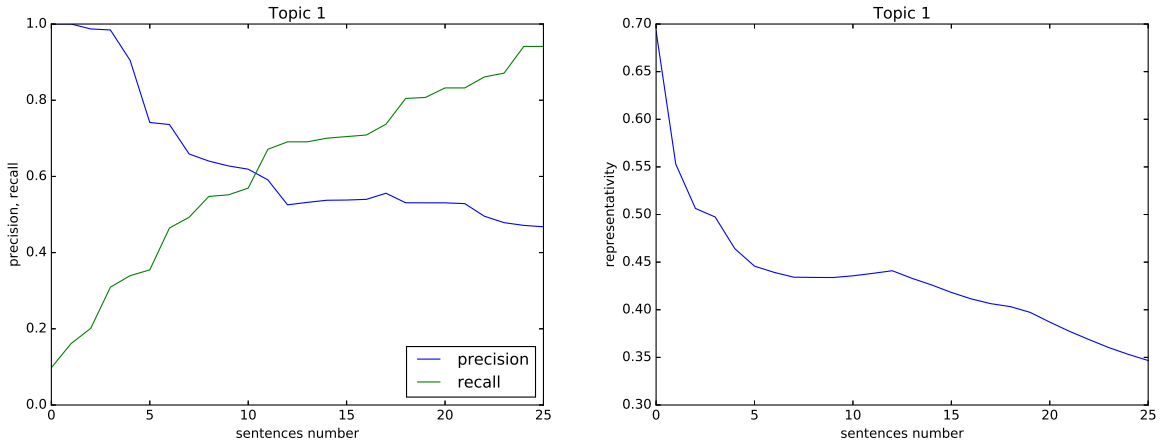


Рис. 7: Количественные оценки качества суммаризации темы №1

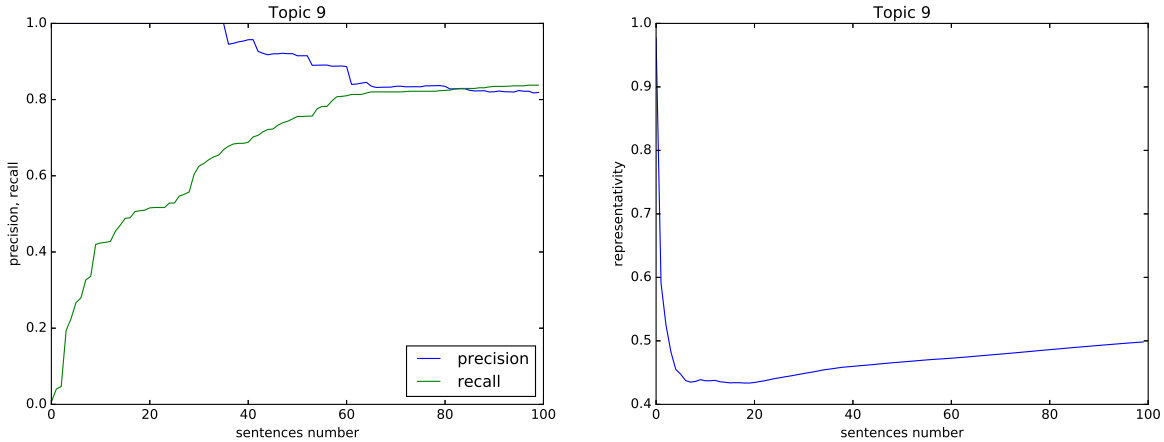
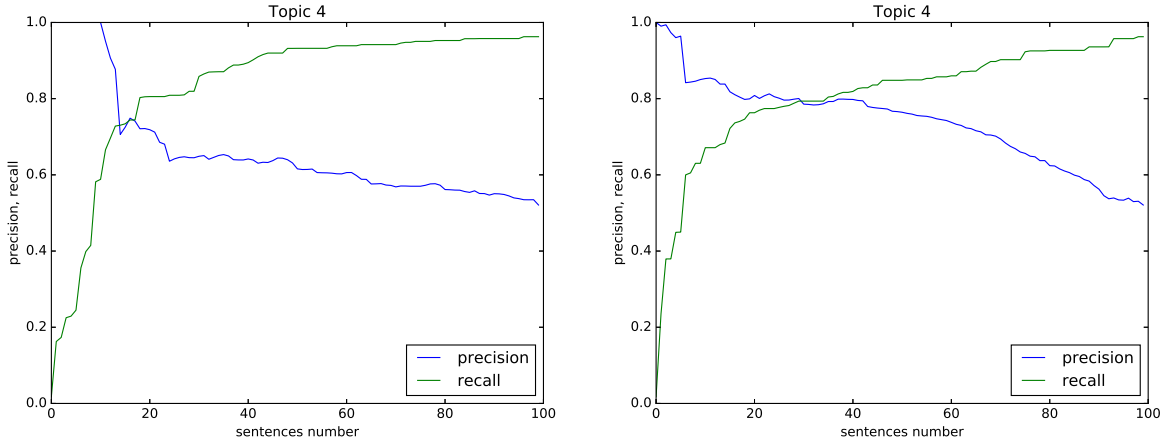


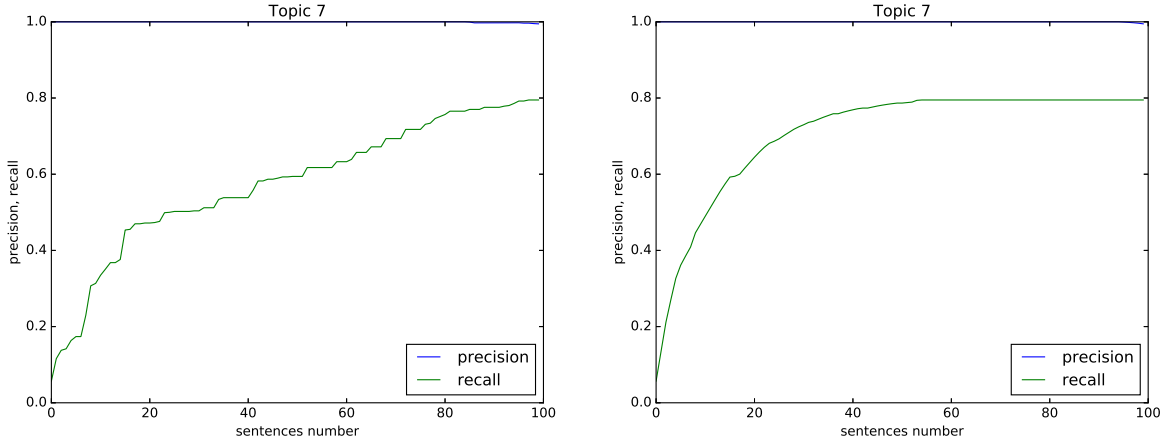
Рис. 8: Количественные оценки качества суммаризации темы №9

Количественные оценки качества суммаризации тем существенно зависят от порядка предложений. Простой перестановкой предложений можно улучшить качество, что видно из Рис. 9 и Рис. 10.



(a) (b)

Рис. 9: Количественные оценки точности и полноты суммаризации темы №4 без перестановки строк(a) и с перестановкой строк(b)



(a) (b)

Рис. 10: Количественные оценки точности и полноты суммаризации темы №7 без перестановки строк(a) и с перестановкой строк(b)

Для оценки качества суммаризации тем и проверки корректности количественных оценок точности и полноты требуются экспертные оценки экспертов-асессоров.

5 Заключение

Разработан алгоритм суммаризации тем в вероятностных тематических моделях и его количественный оценки качества. Проведёны вычислительные эксперименты на коллекции документов конференции ММРО. В экспериментах построена тематическая модель, отвечающая предъявленным требованиям интерпретируемости. Для каждой темы из модели построена суммаризация тем. Показано, что суммаризация тем существенно зависит от размера лексического ядра темы и числа тематичных предложений и от метода ранжирования списка выдачи. Для улучшения качества алгоритма и проверки корректности количественных оценок точности и полноты требуются экспертные оценки на основании предложенной анкеты.

Список литературы

[1] Newman D., Karimi S., Cavedon L. External evaluation of topic models // Australasian Document Computing Symposium. — December 2009. — Pp. 11–18.

[2] Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Blei D. M. Reading tea leaves: How humans interpret topic models // Neural Information Processing Systems (NIPS). — 2009. — Pp. 288-296.

[3] 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. — HLT '10. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. — Pp. 100-108.

[4] Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — EMNLP '11. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. — Pp. 262-272.

[5] Frey B. J., Dueck D. Clustering by passing messages between data points // science. — 2007. — Т. 315. — №. 5814. — Pp. 972-976.

[6] Воронцов К.В. Аддитивная регуляризация тематических моделей коллекции текстовых документов. Доклады РАН, 2014. Т. 455., №3. 268-271

[7] BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Dudarenko Analysis of Images, Social Networks and Texts, 370-381

[8] Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. – 2011. – Т. 12. – №. Oct. – Pp. 2825-2830.

[9] Radev D. R. et al. Centroid-based summarization of multiple documents //Information Processing & Management. – 2004. – Т. 40. – №. 6. – Pp. 919-938.