Министерство науки и высшего образования
Российской Федерации
«Московский физико-технический институт
(государственный университет)»
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Остроухов Петр Алексеевич

Предобученные по Википедии тематические векторные представления слов

03.04.01 — Прикладные математика и физика

Выпускная квалификационная работа
(магистерская диссертация)

Научный руководитель:
д.ф-м.н. Воронцов К.В.

Москва

2019 г.

Содержание

Аннотация

Построение векторных представлений слов является популярной задачей в области обработки естественного языка. В последнее время большую популярность приобрели обучаемые нейросетевые методы, показывающие хорошее качество на многих задачах. Однако данные векторные представления неинтерпретируемы. То есть, невозможно понять, какой смысл несет та или иная компонента вектора. В данной работе исследуется альтернативный подход на основе тематического моделирования, где векторное представление слова представляется в виде вероятностного распределения по темам коллекции. При этом, все темы несложно проинтерпретировать. Полученные таким образом векторные представления тестируются на задачах классификации документов и близости слов и сравниваются с предобученными векторными представлениями GloVe размерности 300.

1 Введение

В задачах анализа данных перед непосредственным обучением алгоритма для решения той или иной прикладной задачи производится предобработка данных, а также извлечение некоторого количества численных признаков. В отличие от изображений, представляющих собой n-мерные тензор, временных рядов, представляющих собой массивы чисел, или табличных данных, подобное преобразование текстовой информации является далеко не тривиальной задачей. При этом, очевидно, это является первым необходимым необходимым шагом на пути решения более комплексных задач обработки естественного языка, таких как: классификация документов, машинный перевод, поиск синонимов, информационный поиск, категоризация и т.д.

Рассматривая в качестве элементарного объекта текста слово или иной фрагмент текста (последовательность символов, словосочетание, предложение, документ), часто возникает вопрос, как представить этот объект так, чтобы сохранить его семантику. При решении данной задачи исследователи исходят из дистрибутивной гипотезы, согласно которой смысл слова определяется словами, стоящими поблизости (контекстом) [9, 10]. В последние десятилетия наибольшую популярность приобрели векторные модели семантики. Основная идея данного подхода заключается в отображении слова в некоторое векторное пространство, при этом точки, являющиеся образами семантически-близких к нему слов, также окажутся близкими в смысле введенной функции расстояния. Помимо векторных моделей существуют также подход, основанный на использовании тезаурусов [12, 16, 18]. Однако главным недостатком данного метода является отсутствие подобных баз знаний для большинства языков, в то время как для построения векторных моделей необходим только текстовый корпус достаточного размера на рассматриваемом языке.

Используемые в дальнейшем термины:

- *центральное слово* — слово, для которого в данный момент строится векторное отображение;
- *контекст слова* — множество слов с возможными повторениями, находящихся в окне определенного размера по обе стороны относительно центрального слова (к примеру, окно ширины 5 — это 5 слов слева и 5 слов справа от центрального слова), если не оговорено иное;
- *глобальный контекст слова* — объединение контекстов слова по всей коллек-

ции;

- *векторное представление слова* — образ слова в векторном пространстве.

Среди способов построения векторных представлений слов выделяют два основных подхода: частотный и нейросетевой. Частотные методы изучаются с 90-х годов, и в настоящее время являются менее популярными. Суть их заключается в построении матриц частотных распределений совместной встречаемости слов в пределах окна некоторого размера или слов в документах. В построенной матрице каждому слову соответствует определенная строка, которая и выступает в качестве векторного представления слова. Для уменьшения размерности векторного пространства используются методы матричного разложения [15, 22, 28].

Обучаемые нейросетевые векторные модели стали широко применяться с публикаций [6, 7]. Авторы этих статей предложили использовать в качестве векторных представлений слов строки матрицы скрытого слоя однослойной полносвязной нейросети, обученной на задаче предсказания центрального слова по его контексту (CBOW) или на задаче предсказания контекста по центральному слову (Skip-Gram, SGNS). В качестве главного преимущества авторы выдвигали хорошее качество на задаче аналогий. Однако позже высказывались мнения, что как используемый датасет, так и сама постановка задачи являются некорректыми [27]. В дальнейшем появилось множество модификаций и вариаций нейросетевых моделей векторной семантики [2, 5, 8, 24].

Несмотря на то что, на первый взгляд, два описанных выше подхода не имеют ничего общего, в статье [20] приводится доказательство, что модель SGNS неявно факторизует матрицу PMI совстречаемостей слов со сдвигом. В дальнейшем, те же авторы показали, что при правильном подборе гиперпараметров оба метода показывают схожее качество на задачах схожести слов (*word-similarity*) [21], несмотря на столь популярное в последнее время мнение о превосходстве нейросетевых векторных представлений над частотными на большинстве задач.

Основным недостатком всех вышеописанных методов является отсутствие интерпретируемости компонент вектора. В [26] был предложен подход, основанный на построении аддитивной регуляризованной тематической модели (APTM [29]) по совстречаемостям слов, где образы слов представляют собой вероятностные распределения по темам, которые не составляет труда проинтерпретировать. Несмотря на то что этот метод неплохо себя показал на ряде задач, в каждом эксперименте модель

обучалась на разных корпусах с разными параметрами, в отличие от используемого подхода в статьях по нейросетевым моделям, где предобученные на стороннем корпусе векторные представления способны эффективно решать сразу несколько задач.

Целью данной работы является развитие идеи построения тематических векторных представлений слов, описанной в [26]. Предлагается обучить модель APTM на большом корпусе Википедии, извлечь из неё векторные представления слов и протестировать их на прикладных задачах. В рамках работы необходимо решить следующие подзадачи:

1. Предобработать корпус Википедии, преобразовать его в корпус совстречаемостей слов.

2. Построить тематическую модель на полученном корпусе, подобрав гиперпараметры.

3. Извлечь векторные представления из полученной модели и проверить их качество на внешних задачах.

В главе 2 более подробно описываются существующие подходы к решению задачи построения векторных представлений слов: в 2.1 приводится формальная постановка задачи, в частях 2.2 и 2.3 описываются основные частотные и нейросетевые подходы соответственно. В главе 3 приводится описание задачи тематического моделирования: в 3.1 приводится постановка задачи, в 3.2 приводится базовое описание итерационного EM-алгоритма, используемого для построения тематической модели, в 3.3 описывается подход к построению тематической модели APTM, основанный на введении регуляризаторов, в 3.4 приводится метод построения тематических векторных представлений. В главе 4 описываются проведенные эксперименты: в 4.1 приводятся данные о текстовом корпусе а также его предобработке, в 4.2 приведены описания задач, на которых осуществляется тестирование векторных представлений, в 4.3, 4.4, 4.5 приводятся качественные результаты экспериментов, в 4.6 — количественные результаты на наилучших моделях, интерпретируемость тематических векторных представлений приводится в 4.7.

Для построения тематической модели использовалась библиотека с открытым исходным кодом BigARTM.

2 Векторные представления слов

В этой главе и в дальнейшем под термином «частота объекта» будем подразумевать ненормированное число повторений рассматриваемого объекта.

Введем некоторые обозначения. За u и v будем обозначать центральное слово и слово из контекста соответственно, n_{uv} — их частота совместной встречаемости. В случае, если речь идет о задачах матричного разложения, левая матрица будет обозначаться Φ, правая — Θ. Строку, соответствующую слову u в левой матрице обозначим через φ_u, столбец, соответствующий слову v (документу d) в правой матрице — через θ_v (θ_d).

Перед построением модели по текстовому корпусу производится его предобработка. Одним из этапов предобработки является *токенизация* — разбиение документов на *токены*.

Определение 2.1. *Токен — это последовательность символов определенного документа, представляющая собой единую семантическую единицу, используемую для дальнейшей обработки [23].*

То есть, в качестве токенов могут выступать: слово, n-грамма (последовательность из n слов), буквенная n-грамма (последовательность символов, являющаяся частью слова).

2.1 Постановка задачи

Поставим формально задачу построения векторных представлений слов. Обозначим за W множество всех слов какого-либо языка (здесь мы не берем во внимание кросс-язычные векторные представления). Тогда нашей задачей является найти отображение множества слов в векторное веществозначное пространство $f : W \to \mathbb{R}^d$, чтобы при этом семантически близкие друг к другу слова имели близкие друг к другу точки в конечном векторном пространстве. Обозначим через «$w \sim_s w'$» семантическую близость слов w и w'. В таком случае:

$$f : W \to \mathbb{R}^d : \forall w, w_s, w_n \in W : w \sim_s w_s,\ w \not\sim_s w_n \to$$
$$\to \rho\big(f(w); f(w_s)\big) < \rho\big(f(w); f(w_n)\big)$$

в смысле метрики ρ, заданной в векторном пространстве $\mathbb{R}^n$.

Исходя из дистрибутивной гипотезы, представив глобальный контекст слова $u \in W$ в виде множества W_u, можно переписать постановку в виде:

$$f : W \to \mathbb{R}^d : \forall u, u_s, u_n \in W : u \sim_s u_s,\ u \not\sim_s u_n \to$$
$$\to \rho\big(f(u|W_u); f(u_s|W_{u_s})\big) < \rho\big(f(u|W_u); f(u_n|W_{u_n})\big).$$

2.2 Частотные векторные представления

2.2.1 Pointwise Mutual Information (PMI)

Стоит отметить, что помимо обычной частоты совместной встречаемости слов в некотором окне, довольно популярной её модификацией является Pointwise Mutual Information (PMI) [4]:

$$PMI(u, v) = \log \frac{P(u, v)}{P(u)P(v)},$$

пришедшая из теории информации. PMI является мерой асоциативности слов и показывает, как видно из формулы, отношение вероятности совместного появления термов u и v к произведению вероятностей их появления по отдельности. Заменив вероятности на частотные распределения, получим:

$$PMI(u, v) = \log \frac{n_{uv}|W|}{n_u n_v}.$$

Так как существуют такие термы u и v, которые никогда не встречаются вместе, для них $PMI(u, v) = \log 0 = -\infty$. Чтобы избавиться от данного недостатка, используется модификация этого метода — Positive PMI:

$$PPMI(u, v) = \max\{0, PMI(u, v)\}.$$

2.2.2 Hyperspace Analogue to Language (HAL)

Суть метода HAL заключается в построении матрицы взвешенных совстречаемостей слов $N = \{n_{uv}\}_{u,v \in W}$ внутри скользящего окна определенного размера (где W — множество слов рассматриваемого корпуса), при этом n_{uv} приобретает вес, обратно-пропорциональный удаленности контекстного слова v от центрального.

Основным недостатком данной модели является слишком сильный сдвиг в сторону высокочастотных термов. Какие-то общеупотребимые слова могут вносить необоснованно большое количество информации в представления слов, вследствие чего образы слов, не являющихся семантически близкими, окажутся таковыми в векторном

пространстве. Нивелировать данный недостаток можно, используя вместо обычных частот статистику PPMI.

2.2.3 Latent Semantic Analysis (LSA)

В методе LSA строится матрица частот слов в документах $N = \{n_{wd}\}_{w \in W, d \in D}$ (где D — множество документов корпуса). В отличие от HAL, данный подход представляет каждый документ в виде *мешка слов* (*bag of words*), то есть не учитывает порядок следования слов в документе. Для построения низкоразмерного латентного семантического пространства используется сингулярное разложение (SVD): $N = U\Sigma V^T$, и выбираются T главных компонент. Далее в качестве матрицы векторных представлений выбирается $\Phi = U_T\sqrt{\Sigma_T}$ или $\Phi = U_T\Sigma_T$. Таким образом, LSA решает задачу низкорангового матричного разложения:

$$N = \Phi\Theta, \ |\Phi| = |W| \times T, \ |\Theta| = T \times |D|.$$

2.3 Нейросетевые векторные представления

Существуют нейросетевые векторные представления, явно факторизующие матрицу совстречаемостей слов с обучаемыми параметрами [24].

Однако многие нейросетевые методы основаны на решении задач, не связанных напрямую с построением векторных представлений, где векторные представления получаются в качестве побочного продукта. К таким задачам можно отнести задачу генерации текста (BERT [2], ELMo [5]), предсказания контекста по центральному слову (Skip-Gram [7], SGNS [6] или слова по контексту (CBOW [7]).

Далее будут рассмотрена модель GloVe, которая используется в качестве бейзлайна в данной работе, а также модели SGNS и CBOW, давшие серьезный толчок в развитии нейросетевых векторных представлений.

2.3.1 SGNS, CBOW

В [7] Миколов предложил эффективный метод построения векторных представлений на основе однослойной нейросети без нелинейностей, где в качестве матрицы векторных представлений слов выступает матрица весов скрытого слоя $|W| \times T$, где W — словарь токенов коллекции, T — размерность векторного пространства. Для начала, рассмотрим модель Skip-Gram. Пусть размер скользящего окна равен l.

Обозначим за C_u контекст слова u размера $2l$. Функционал качества можно записать в виде:

$$\mathcal{L}_i = \log p(v \in C_u | u) = \sum_{v \in C_u} \log p(v|u) =$$
$$= \sum_{v \in C_u} \log \frac{\exp \sum_t \varphi_{vt}\theta_{tu}}{\sum_{w \in W} \exp \sum_t \varphi_{wt}\theta_{tu}}.$$

Для модели Continious Bag Of Words (CBOW) этот функционал выглядит несколько иначе:

$$\mathcal{L}_i = \log p(u|v \in C_u) = \log \frac{\exp \sum_t \left(\varphi_{ut} \sum_{v \in C_u} \theta_{tv}\right)}{\sum_{w \in W} \exp \sum_t \left(\varphi_{wt} \sum_{v \in C_u} \theta_{tv}\right)} =$$
$$= \sum_{v \in C_u} \log \frac{\exp \sum_t \varphi_{ut}\theta_{tv}}{\sum_{w \in W} \exp \sum_t \left(\varphi_{wt} \sum_{v \in C_u} \theta_{tv}\right)}.$$

Можно увидеть, что единственное отличие между этими двумя моделями заключается в способе нормировки. При этом, в обеих моделях используются две матрицы векторных представлений: для центральных слов и для слов из контекста. Матрица контекстов не берется во внимание, и векторные представления слов извлекаются из другой матрицы. Оптимизация происходит согласно принципу максимума правдоподобия: необходимо просуммировать все $\mathcal{L}_i$ и промаксимизировать полученное выражение с помощью градиентных методов. К примеру, для Skip-Gram конечный функционал выглядит, как:

$$\sum_{u \in V_u} \sum_{v \in V_v} n_{uv} \log p(v|u) \to \max_{\Phi,\Theta}. \tag{2.1}$$

В [6] было предложено несколько модификаций метода Skip-Gram. Так как функция softmax является довольно вычислительно-затратной (при нормировке необходимо выполнять суммирование по всему словарю), в статье были предложены её более эффективные аппроксимации. *Иерархический softmax* (*hierarchical softmax*) представляет выходной слой нейросети в виде дерева Хаффмана, листья которого соответствуют словам из словаря. Таким образом, сложность вычисления снижается с $O(|W|)$ до $O(\log_2 |W|)$. Второй альтернативой является *негативное семплирование* (*negative sampling*), где оптимизируемый функционал заменяется на

$$\sum_{u \in W} \sum_{v \in W} n_{uv} \big(\log \sigma\langle\varphi_v; \theta_u\rangle + k\mathsf{E}_{v' \sim P_n(u)} \log \sigma(-\langle\varphi_{v'}; \theta_u\rangle)\big) \to \max_{\Phi,\Theta}.$$

Матожидание берется по шумовым словам, не принадлежащим контексту слова («негативным» объектам), и оценивается по методу Монте-Карло семплированием

из всего текстового корпуса (отсюда и название «негативное семплирование»):

$$\mathsf{E}_{v'\sim P_n(u)} \log \sigma(-\langle \varphi_{v'}; \theta_u \rangle) \approx \frac{1}{k} \sum_{s=1}^{k} \log \sigma(-\langle \varphi_{v_s}; \theta_u \rangle), \;\; v_s \sim P_n(u).$$

Количество семплов k является гиперпараметром, подбираемым пользователем. Таким образом, задача переформулируется в виде задачи бинарной классификации, где посредством логистической регресии моделируется вероятность принадлежности слова v контексту слова u: $P(v \in C_u|u)$. Причем, как видно из уравнения, роль негативного семплирования заключается в максимизации вероятности непринадлежности шумовых слов контексту центрального слова: $P(v' \notin C_u|u)$. Распределение, из которого осуществляется семплирование, также является гиперпараметром. Авторы на основе своих эмпирических наблюдений предлагают использовать униграммное распределение в степени 3/4: $v' \sim \left(\frac{n_{v'}}{n}\right)^{3/4}$, где n — длина коллекции в термах. Данная модификация модели Skip-Gram получила название Skip-Gram Negative Sampling (SGNS).

Позже в статье [20] было показано, что оптимизация функционала качества SGNS соответствует неявной матричной факторизации матрицы shifted PMI (sPMI):

$$sPMI(u, v, k) = PMI(u, v) - \log k = \log \frac{n_{uv}}{n_u n_v} - \log k \approx \Phi\Theta, \tag{2.2}$$

где k — параметр негативного семплирования, Φ и Θ — матрицы размера $|W| \times T$ и $T \times |W|$ соответственно.

2.3.2 GloVe

В отличие от предыдущего подхода, авторы Global Vectors (GloVe [24]) изначально представляют свою модель, как факторизацию матрицы логарифмов частот встречаемостей одного слова в контексте другого по всему текстовому корпусу, проводя аналогию с методами HAL и LSA. Помимо этого, для каждого слова вводятся обучаемые параметры сдвига b_u и $\tilde{b}_v$. Чтобы нивелировать вклад маленьких значений частот совстречаемостей, и уменьшить вклад больших значений, вводится весовая функция $f(n_{uv})$, которая является неубывающей, штрафует за большие значения, и $f(0) = 0$. В качестве функционала качества используется взвешенный метод наименьших квадратов:

$$\sum_{u\in W} \sum_{v\in W} f(n_{uv}) \big(\langle \varphi_u, \theta_v \rangle + b_u + \tilde{b}_v - \log n_{uv}\big)^2 \to \min_{\Phi,\Theta,b,\tilde{b}}. \tag{2.3}$$

При этом сами авторы в качестве весовой функции выбрали:

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha}, & x < x_{\max}, \\ 1, & \text{иначе}, \end{cases}$$

с гиперпараметрами $x_{\max} = 100$, $\alpha = 3/4$.

Векторное представление слова w строится, как сумма значений соответствующих ему строки матрицы Φ и столбца матрицы Θ: $\vec{w} = \varphi_w + \theta_w$.

3 Вероятностное тематическое моделирование

Основной задачей тематического моделирования является определение того, к каким темам относится тот или иной документ. Это возможно благодаря тому, что каждый документ из коллекции представляется в виде вероятностного распределения по темам, а тема, в свою очередь, представляется в виде вероятностного распределения по словам.

3.1 Постановка задачи

Обозначим множество документов коллекции за D, а множество всех слов — за W. При этом, каждый документ $d \in D$ представляется в виде мешка слов: $d = \{w_1, ..., w_{n_d}\}$. Задача тематического моделирования заключается в представлении распределения слов в документах в виде взвешенной смеси распределений тем по словам, где веса равны распределениям документов по темам. Используя формулу полной вероятности и гипотезу условной независимости (распределение тем по словам не зависит от документов), можно записать задачу следующим образом:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \tag{3.1}$$

где φ_{wt} — распределение тем по словам, а θ_{td} — распределение документов по темам, T — множество тем.

Положив $F = \big(p(w|d)\big)^{|W|\times|D|}$, $\Phi = \big(\varphi_{wt}\big)^{|W|\times|T|}$, $\Theta = \big(\theta_{td}\big)^{|T|\times|D|}$, уравнение (3.1) можно переписать в матричном виде:

$$F = \Phi\Theta. \tag{3.2}$$

Чаще всего $|T| << |D|$, $|T| << |W|$, поэтому (3.2) — задача низкорангового матричного разложения. Так как мы хотим на выходе получить стохастические матрицы, столбцы которых являются вероятностными распределениями, необходимо наложить дополнительные ограничения неотрицательности элементов каждой из матриц, а также нормировки столбцов в единицу.

Задача решается через максимизацию правдоподобия по всем парам слов и документов $\{w_i, d_i\}_{i=1}^n$:

$$p(\{w_i, d_i\}_{i=1}^n; \Phi, \Theta) = \prod_{i=1}^n p(w_i|d_i)p(d_i) = \prod_{w \in W}\prod_{d \in D} p(w|d)^{n_{dw}} \underbrace{p(d)^{n_{dw}}}_{\text{const}} \to \max_{\Phi,\Theta},$$

где n_{dw} — частота слова w в документе d. Прологарифмируем данное выражение, отбросим константное распределение документа, добавим введенные ранее ограничения на задачу, и, с учетом (3.1), получим:

$$\log \prod_{w \in W} \prod_{d \in D} p(w|d)^{n_{dw}} = \sum_{w \in W} \sum_{d \in D} n_{dw} \log \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta}, \tag{3.3}$$

$$\varphi_{wt} \geq 0, \quad \sum_{w \in W} \varphi_{wt} = 1, \quad \theta_{td} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1. \tag{3.4}$$

Описанная выше модель называется Probabilistic Latent Semantic Analysis (PLSA [14]).

3.2 EM-алгоритм

Введем некоторые обозначения:

- n_{tdw} — число троек, в которых слово w документа d связано с темой t
- $n_{wt} = \sum_{d \in D} n_{tdw}$ — число троек, в которых слово w связано с темой t,
- $n_{td} = \sum_{w \in d} n_{tdw}$ — число троек, в которых любое слово из документа d связано с темой t,
- $n_t = \sum_{d \in D} n_{wt}$ — число троек, связанных с темой t,
- $n_d = \sum_{w \in d} n_{wd}$ — длина документа в словах.

Представив вероятностные распределения из (3.3) в виде их частотных оценок, получим:

$$p(w|d) = \frac{n_{wd}}{n_d}, \quad \varphi_{wt} = p(w|t) = \frac{n_{wt}}{n_t}, \quad \theta_{td} = p(t|d) = \frac{n_{td}}{n_d}. \tag{3.5}$$

Видно, что все эти вероятности выражаются через n_{tdw}. При этом, введя дополнительную переменную $p(t|d,w) = n_{tdw}/n_{dw}$, можно выразить n_{tdw} по формуле Байеса:

$$n_{tdw} = n_{dw} p(t|d,w) = n_{dw} \frac{\varphi_{wt} \theta_{td}}{\sum_{t' \in T} \varphi_{wt'} \theta_{t'd}}. \tag{3.6}$$

Задача (3.3), (3.4) решается с помощью *итерационного EM-алгоритма*, где сначала случайным образом инициализируются матрицы Φ и Θ, далее на E-шаге вычисляется n_{tdw}, и на M-шаге обновляются матрицы Φ и Θ, до тех пор, пока процесс не сойдется.

Для больших коллекций может использоваться модификация обычного EM-алгоритма — *онлайн EM-алгоритм* [1, 13]. Основная идея его заключается в том,

что матрица Φ может сойтись до окончания первой итерации, вследствие чего может хватить одной итерации по всей коллекции. По факту, построенная таким образом тематическая модель не обладает высоким качеством, так что может потребоваться несколько проходов онлайн-алгоритма.

3.3 Аддитивная регуляризация тематической модели

Задача (3.2) является некорректно поставленной по Адамару: у неё не существует единственного решения. Действительно, для любых матриц Φ и Θ мы можем ввести некоторую обратимую матрицу $S^{|T|\times|X|}$ такую, что $F = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$ тоже будет решением, при условии что Φ', Θ' — стохастические с неотрицательными элементами. Некорректно поставленные задачи решаются путем введения в функционал качества дополнительных аддитивных взвешенных членов, называемых *регуляризаторами*, которые доопределяют задачу, учитывая её особенности. Таким образом, задачу (3.3), (3.4) можно переписать следующим образом:

$$\sum_{w\in W}\sum_{d\in D} n_{dw}\log\sum_{t\in T}\varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}, \tag{3.7}$$

$$\varphi_{wt}\geq 0, \quad \sum_{w\in W}\varphi_{wt} = \{0,1\}, \quad \theta_{td}\geq 0, \quad \sum_{t\in T}\theta_{td} = \{0,1\}, \tag{3.8}$$

где $R(\Phi,\Theta) = \sum_{i=1}^{n}\tau_i R_i(\Phi,\Theta)$ — взвешенная сумма регуляризаторов, τ_i — неотрицательные коэффициенты регуляризации, являющиеся гиперпараметрами. Стоит отметить, что здесь мы также добавили возможность появления нулевых столбцов в матрицах Φ и Θ, что может быть последствием действия регуляризаторов.

Приведем теорему, описывающую алгоритм вычисления параметров регуляризованной тематической модели.

Теорема 3.1. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда точка (Φ,Θ) локального экстремума задачи (3.7), (3.8) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:*

$$p_{tdw} = norm_{t\in T}(\varphi_{wt}\theta_{td}); \tag{3.9}$$

$$\varphi_{wt} = norm_{w\in W}\left(n_{wt} + \varphi_{wt}\frac{\partial R(\Phi,\Theta)}{\partial\varphi_{wt}}\right); \quad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{3.10}$$

$$\theta_{td} = norm_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R(\Phi,\Theta)}{\partial\theta_{td}}\right); \quad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}. \tag{3.11}$$

Доказательство приводится в [29].

Оператор norm преобразует вектор значений в вероятностное распределение путем нормировки и положительной срезки (обнуления отрицательных элементов):

$$\operatorname{norm}(x) = \left\{ \frac{\max(x_i, 0)}{\sum_{j=1}^{n} \max(x_j, 0)} \right\}_{i=1}^{n}.$$

Данная теорема модифицирует способ обновления матриц Φ и Θ на M-шаге EM-алгоритма. При этом, E-шаг остается тем же самым. Также можно заметить, что, обнулив функционал суммы регуляризаторов $R(\Phi, \Theta) = 0$, мы снова придем к формулам (3.5), (3.6).

Далее будут рассмотрены наиболее популярные регуляризаторы декоррелирования и разреживания, которые планируется в будущем применить в тематической модели, используемой для построения векторных представлений слов.

3.3.1 Регуляризатор декоррелирования

Так как мы хотим, чтобы каждый столбец матрицы Φ описывал одну конкретную тематику, необходимо, чтобы распределения каждой из тем были максимально различными. Для этого вводится регуляризатор декоррелирования, в котором минимизируются попарные скалярные произведения столбцов матрицы Φ:

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t' \in T} \sum_{t \in T \setminus \{t'\}} \langle \varphi_t, \varphi_{t'} \rangle.$$

Тогда формула M-шага, согласно (3.15), выглядит следующим образом:

$$\varphi_{wt} = \operatorname{norm}_{w \in W} \Big(n_{wt} - \tau \varphi_{wt} \sum_{t' \in T \setminus t'} \varphi_{wt'} \Big).$$

3.3.2 Регуляризатор сглаживания/разреживания

Темы можно разделить на *фоновые* и *предметные*. Согласно *гипотезе разреженности*, у каждой предметной темы существует некоторое *семантическое ядро* — термины, описывающие данную тему, и с помощью которых тему можно проинтерпретировать. Вследствие этого подразумевается, что предметные темы имеют разреженные распределения. Фоновые темы же содержат в себе слова общей лексики, и их распределение по словам гораздо более сглаженно. Для учета вышеописанных свойств задачи вводится регуляризатор сглаживания или разреживания матрицы Φ.

Помимо этого, подразумевается, что каждый документ может принадлежать небольшому числу тем, следовательно такие документы должны иметь разреженные

распределения по темам, и можно ввести регуляризатор сглаживания или разреживания матрицы Θ.

Объединив два описанных регуляризатора, получим:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \log \varphi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \log \theta_{td},$$

где β_{wt}, α_{td} — коэффициенты регуляризации. Данный функционал можно проинтерпретировать, как линейную комбинацию кросс-энтропий между равномерными распределениями слов в темах, и тем в документах и параметрами модели φ_{wt}, θ_{td} соответственно (распределение вероятности для равномерного распределения входит в β_{wt} и α_{td}). При отрицательных коэффициентах регуляризации происходит сглаживание распределений, при положительных — разреживание распределений.

Формулы М-шага, согласно (3.15), (3.16), выглядят так:

$$\varphi_{wt} = \operatorname{norm}_{w \in W}(n_{wt} + \beta_{wt});$$
$$\theta_{td} = \operatorname{norm}_{t \in T}(n_{td} + \alpha_{td}).$$

3.3.3 Мультимодальная тематическая модель

Помимо текста документа в коллекциях часто присутствуют различные метаданные: авторы, комментарии, категории и т.п.

Предполагается, что использование различных видов токенов и метаданных может благоприятно сказаться как на качестве самой тематической модели и интерпретируемости тем, так и на качестве решения задачи, для которой используется тематическое моделирование. Для этого по всей коллекции документов для каждого типа данных (различные типы токенов и метаданных), называемых *модальностями*, строится собственная матрица частот слов в документах. При этом, словари каждой модальности попарно не пересекаются. Тематическая модель, построенная на таких данных, называется *мультимодальной*.

Обозначим за M множество модальностей, W_m — словарь, соответствующий m-й модальности, объединение всех словарей — за W. Тогда тематическую модель, построенную по рассматриваемой модальности можно представить, как:

$$p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}, \;\; w \in W_m, \;\; d \in D.$$

При этом, каждой модальности соответствует стохастическая матрица Φ_m. При построении модели все эти матрицы конкатенируются по первой координате, в результате чего получается матрица размера $|W| \times |T|$, где все распределения тем по словам

общие, вследствие чего улучшается интерпретация тем (к примеру, в топе наиболее вероятных слов у темы могут оказаться категории документов).

Мультимодальная модель обучается методом максимизации линейной комбинации правдоподобий тематических моделей, соответствующих каждой из модальностей:

$$\sum_{m\in M}\tau_m\sum_{w\in W^m}\sum_{d\in D}n_{wd}\log\sum_{t\in T}\varphi_{wt}\theta_{td}+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \tag{3.12}$$

$$\varphi_{wt}\geq 0,\quad \sum_{w\in W^m}\varphi_{wt}=\{0,1\},\quad \theta_{td}\geq 0,\quad \sum_{t\in T}\theta_{td}=\{0,1\}, \tag{3.13}$$

где веса τ_m являются гиперпараметрами, позволяющими сбалансировать вклад каждой из модальностей в мультимодальную тематическую модель.

Теорема 3.2. *Пусть функция $R(\Phi,\Theta)$ непрерывно дифференцируема. Точка (Φ,Θ) локального экстремума задачи (3.12), (3.13) удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw}=p(t|d,w)$:*

$$p_{tdw}=norm_{t\in T}(\varphi_{wt}\theta_{td}); \tag{3.14}$$

$$\varphi_{wt}=norm_{w\in W}\left(n_{wt}+\varphi_{wt}\frac{\partial R(\Phi,\Theta)}{\partial\varphi_{wt}}\right);\quad n_{wt}=\sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw}; \tag{3.15}$$

$$\theta_{td}=norm_{t\in T}\left(n_{td}+\theta_{td}\frac{\partial R(\Phi,\Theta)}{\partial\theta_{td}}\right);\quad n_{td}=\sum_{m\in M}\sum_{w\in d\cap W^m}\tau_m n_{dw}p_{tdw}.. \tag{3.16}$$

Доказательство приводится в [29].

3.3.4 Иерархическая тематическая модель

Иерархическая тематическая модель позволяет моделировать отношения гипонимии и гиперонимии между темами. Такая модель представляется в виде древовидной структуры, где каждый узел означает тему, и на каждом последующем уровне узлов больше, чем на предыдущем. Для моделирования межуровневых связей в модель вводятся распределения тем по подтемам: $\psi(s|t)$.

Построение иерархической модели осуществляется итеративно: уровень за уровнем. Пусть уже построено l уровней тематической модели. Чтобы построить $(l+1)$-й уровень, необходимо представить распределения тем из l-го уровня в виде взвешенной смеси распределений тем по подтемам из дочернего уровня:

$$p(w|t)=\sum_{s\in S}p(w|s)p(s|t)=\sum_{s\in S}\varphi_{ws}\psi_{st}, \tag{3.17}$$

где S — множество тем $(l+1)$-го уровня. Представив эту задачу в матричном виде, получим:

$$\Phi^l \approx \Phi\Psi,$$

где Φ — стохастическая матрица со столбцами, соответствующими распределениям слов по подтемам. Стоит отметить, что, так как $|S| > |T|$, матрицы Φ и Ψ обладают более высоким рангом, чем Φ^l, поэтому это уже не является задачей низкоранговой матричной факторизации.

Представим (3.17) в виде регуляризатора:

$$R(\Phi, \Psi) = \tau \sum_{w \in W} \sum_{t \in T} n_{wt} \log \sum_{s \in S} \varphi_{ws}\psi_{st}.$$

Как видно, данное выражение с точностью до обозначений совпадает с (3.3). Таким образом, вместо добавления регуляризатора в построенную l-уровневую модель можно добавить $|T|$ псевдодокументов, где каждый псевдодокумент соответствует теме родительского уровня. Частоты слов в каждом псевдодокументе будут равны: $\tau n_{wt} = \tau n_t \varphi_{wt}$. При этом, в получившемся разложении на месте матрицы Θ будет стоять матрица Ψ.

3.4 Тематические векторные представления слов

Как уже упоминалось в разделе 2.3, SGNS неявно факторизует матрицу sPMI совстречаемостей (2.2). Создатели GloVe же изначально ставили задачу, как явную факторизацию матрицы логарифмов совстречаемостей слов с дополнительными обучаемыми сдвигами и весами (2.3). Проводя аналогию с этими моделями, рассмотрим несколько модифицированную постановку задачи таматического моделирования. А именно, вместо матрицы частот слов в документах будем рассматривать матрицу совстречаемостей слов, элементами которой будут частоты слов внутри псевдодокументов — глобальных контекстов других слов. Тогда задача тематического моделирования (3.1) примет следующий вид:

$$p(v|u) = \sum_{t \in T} p(v|t)p(t|u) = \sum_{t \in T} \varphi_{vt}\theta_{tu}. \tag{3.18}$$

или

$$F' = \Phi\Theta,$$

где Φ и Θ — стохастические матрицы со столбцами, соответствующими распределениям тем по словам и слов по темам соответственно. Так как в (3.18) нет четкого

разделения на центральные слова и слова контекста (используется только информация о совстречаемости), то в качестве векторных представлений слов можно брать как столбцы матрицы Θ, так и столбцы матрицы Φ, пересчитанные по формуле Байеса:

$$p(t|v) = \frac{p(v|t)p(t)}{p(v)} = \frac{\varphi_{vt}p(t)}{p(v)} = \frac{n_{vt}}{n_v}. \tag{3.19}$$

При этом, так как θ_{tu} нельзя просто получить из φ_{vt} по формуле Байеса, очевидно, векторные представления получатся разными.

Запишем метод максимизации правдоподобия для (3.18):

$$\sum_{u\in W}\sum_{v\in W} n_{uv}\log p(v|u) = \sum_{u\in W}\sum_{v\in W} n_{uv}\log\sum_{t\in T}\varphi_{vt}\theta_{tu} \to \max_{\Phi,\Theta}, \tag{3.20}$$

$$\varphi_{vt} \geq 0, \quad \sum_{v\in W}\varphi_{vt} = 1, \quad \theta_{tu} \geq 0, \quad \sum_{t\in T}\theta_{tu} = 1. \tag{3.21}$$

Как можно заметить, данный функционал очень похож на оптимизируемый функционал из Skip-Gram (2.1) с ограничениями неотрицательности и нормировки.

4 Эксперименты

В данном разделе будут описаны проведенные эксперименты. Для начала стоит отметить, что в наших экспериментах тематические векторные представления вычисляются по формуле Байеса из матрицы Φ согласно (3.19).

Так как целью работы является не столько построение качественной тематической модели с хорошо интерпретируемыми матрицами, сколько построение интерпретируемых векторных представлений слов, которые могут решать некоторые прикладные задачи не хуже существующих аналогов, рассматриваются только значения внешних критериев качества, в то время как внутренние критерии, такие как перплексия и когерентность, не берутся во внимание. В качестве задач, на которых будет осуществляться тестирование, были выбраны классификация документов и задача близости слов. Далее будет более подробно сказано об этих задачах, а также о коллекциях, на которых они решались.

В качестве бейзлайна были выбраны векторные представления GloVe [24], с размерности 300, предобученные по объединенным корпусам Википедии 2014 года и English Gigaword Fifth Edition.

4.1 Описание корпуса и предобработка

В качестве текстового корпуса был взят дамп Википедии за 2018 год на английском языке. Всего в нём 4352108 документов, каждый из которых принадлежит некоторому множеству категорий. В наших экспериментах использовались модальности слова и категорий. Подбор весов модальности категорий не оказал существенного влияния на внешние критерии качества, при этом значительно улучшив внутренние критерии. Так как в данной работе акцент ставится именно на внешние критерии, модальности категорий, как и модальности слов, был присвоен единичный вес.

При извлечении данных из дампа использовались средства библиотеки с открытым исходным кодом gensim, с небольшими модификациями. После извлечения корпус предобрабатывался и приводился к формату, воспринимаемому библиотекой BigARTM. При предобработке документов все слова приводились к нижнему регистру. Также удалялись стоп-слова (*and, the, i, you, ...*) и пунктуация. Токенизацию по словам gensim производит автоматически.

Для построения тематической модели по совстречаемостям, необходимо преобразовать коллекцию документов в коллекцию псевдодокументов, где в качестве псев-

додокумента выступает глобальный контекст слова. Необходимый инструментарий для построения такой «псевдоколлекции» уже реализован и встроен в библиотеку BigARTM, где псевдодокумент представляется в виде множества слов, входящих в глобальный контекст, совместно с их частотами. Функционал данного инструментария позволяет выбирать параметры ширины контекстового окна (`cooc_window`) и нижней границы частоты слова внутри глобального контекста (`cooc_min_tf`), ниже которой слова отбрасываются. В наших экспериментах были установлены фиксированные значения `cooc_window`=10, `cooc_min_tf`=20. Также в BigARTM реализована возможность построения псевдоколлекции, где вместо частот слов внутри глобальных контекстов используются PPMI. Корпус, построенный таким образом с теми же параметрами, что и обычная псевдоколлекция, будет использоваться для обучения тематической модели в некоторых экспериментах.

Ввиду большого размера полученной коллекции для построения тематических векторных представлений берутся 100000 слов, обладащих самой высокой частотой по всей коллекции, остальные отбрасываются.

4.2 Рассматриваемые прикладные задачи

4.2.1 Классификация документов

Опишем постановку задачи классификации документов. Пусть D — некоторая выборка документов, Y — множество всех классов, которому каждый документ может принадлежать. Обозначим за $f : D \to \mathbb{R}^T$ процедуру построения векторного представления документа (в данной работе векторное представление документа рассчитывается, как среднее арифметическое векторных представлений его слов), за $g : \mathbb{R}^T \to Y$ — модель классификации. Улучшить качество решения данной задачи в смысле некоторого критерия можно как путём улучшения отображения f, так и g. Стоит отметить, что в данном эксперименте целью является построение векторного отображения, лучшего, чем бейзлайн на каком-то стандартном алгоритме классификации. Поэтому для классификации используется логистическая регрессия с перебором параметра L_2-регуляризации по сетке, и в дальнейшем алгоритм классификации не оптимизируется. Поставим задачу более формально:

- **Дано**: множество пар документ-метка класса $\{(d, y)\}_{d \in D,\ y \in Y}$, притом каждый документ может принадлежать только одному классу.

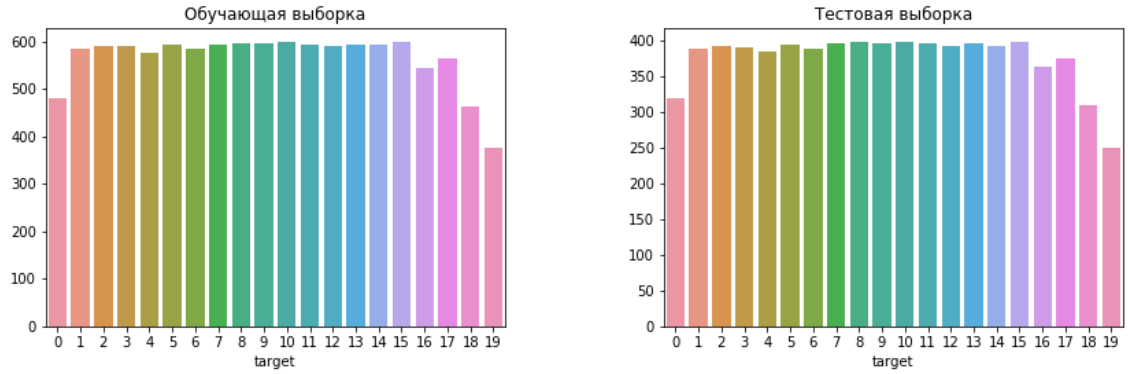

Рис. 1: Распределения объектов по классам в обучающей и тестовой выборках в датасете 20 Newsgroups

- **Найти**: такое векторное представление f, что:

$$\mathcal{Q}\Big(\Big\{\big(g(f(d)),y\big)\big|\ d\in D,\ y\in Y\Big\}\Big)\to\max,$$

  где $\mathcal{Q}$ — критерий качества.

- **Критерий**: (квадратными скобками здесь обозначается предикат)

$$acc=\frac{\sum_{\{(d,y)\}}[g(f(d))=y]}{|D|}.$$

Для тестирования на данной задаче были выбраны коллекции Large Movie Review Dataset [19] и 20 Newsgroups [17]. Large Movie Review Dataset — коллекция отзывов о фильмах с сайта `imdb.com`. И в обучающую, и в тестовую выборку входят 12500 положительных и 12500 отрицательных отзывов. 20 Newsgroups содержит в себе 18846 документов новостей, почти равномерно распределенных по 20 категориям (см. рис 1). В обучающую выборку входит 11314 документов, а в тестовую — 7532 документа. Так как классы рассматриваемых корпусов являются более-менее сбалансированными, в качестве критерия качества было выбрано отношение количества правильно классифицированных объектов ко всему объемум выборки (*accuracy*). Предобработка обоих корпусов аналогична предобработке документов Википедии. Токенизация производилась по всем пробельным символам.

4.2.2 Задача близости слов

Поставим формально задачу близости слов (*word similarity*):

- **Дано**: выборка пар слов совместно с человеческими оценками их семантической близости $\{(w_1^i, w_2^i), s^i\}_{i=1}^n$.
- **Найти**: такое векторное представление слов, что нормированное скалярное произведение образов слов будет коррелировать с оценками их семантической близости:

$$f : W \to \mathbb{R}^T :$$
$$\rho\Big(\big\{c\big(f(w_1^1); f(w_2^1)\big), ..., c\big(f(w_1^n); f(w_2^n)\big)\big\}, \big\{s^1, ..., s^n\big\}\Big) \to \max,$$

  где c — нормированное скалярное произведение (или косинусная мера).
- **Критерий**: ρ — корреляция Спирмена.

В качестве датасетов были выбраны WordSimilarity-353 (WS353) [25], MEN [3], SimLex-999 [11]. WS353 состоит из двух подвыборок: 153 объекта, где каждый объект — это пара слов с 13 оценками их семантической близости, 200 объектов, где каждый объект — это пара слов с 16 оценками их семантической близости. Все оценки лежат внутри отрезка $[0; 10]$, для каждой пары слов оценки усреднялись. MEN содержит 3000 объектов, где каждый объект — это пара слов с одной оценкой их семантической близости по 50-бальной шкале. SimLex-999 содержит 999 объектов, где каждый объект — это пара слов с одной оценкой их семантической близости в отрезке $[0; 10]$.

Предобработка данной коллекции не производилась.

4.3 Вид EM-алгоритма

Как известно из подраздела 3.2, существует два вида EM-алгоритма. Причём онлайн EM-алгоритм лучше применим к большой текстовой коллекции ввиду более высокой скорости сходимости. Так как мы имеем дело с довольно большим корпусом, первым делом было решено исследовать, как влияет выбор типа EM-алгоритма на внешние критерии.

Так как Википедия является ресурсом, покрывающим совершенно разные и, порой, совершенно не связанные области знаний, для её качественного описания, скорее всего, может понадобиться большое количество тем. Как видно из графиков на рис. 2, онлайн-алгоритм гораздо быстрее сходится к более высокому уровню качества, что нам на руку, так как одну итерацию алгоритм на большом количестве тем делает довольно долго. В итоге, было решено остановиться именно на нём.

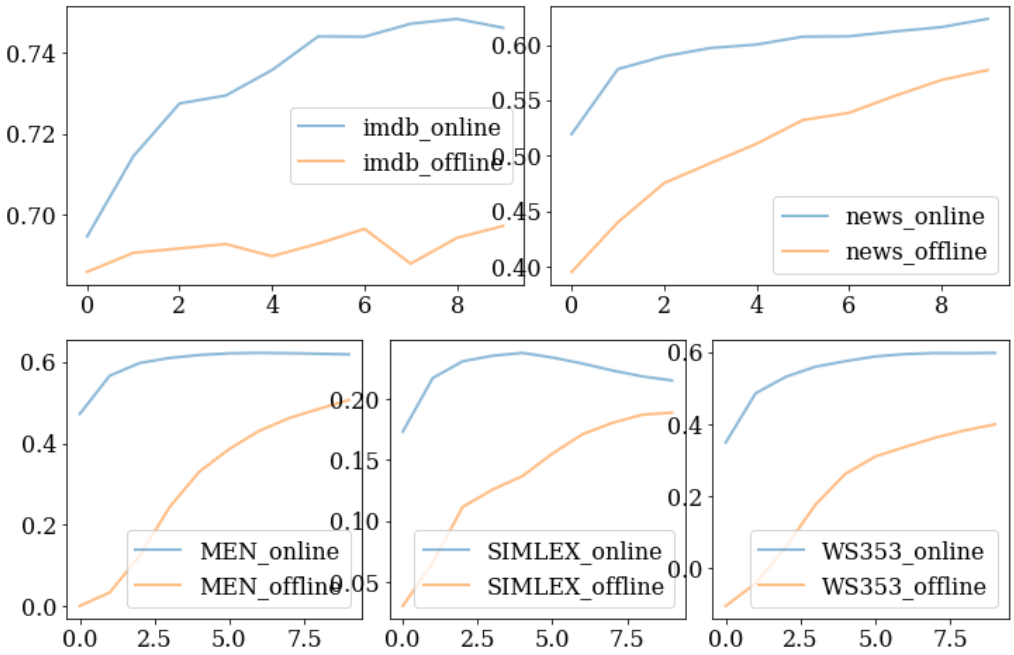

Рис. 2: Вид EM-алгоритма ($|T| = 100$)

4.4 Количество тем

Следующим шагом было исследование влияния количества тем на внешние критерии качества.

На рисунке 3 сравниваются две модели: 100 и 500 тем. Как видно из графиков, увеличение количества тем действительно положительно сказывается на метриках качества во всех датасетах кроме SimLex-999. В наших экспериментах приемлемые результаты были получены на 750 и 2000 темах, что будет показано позже.

4.5 Модификации текстового корпуса

4.5.1 Извлечение «узких» контекстов

Общеупотребимые слова могут встречаться во многих документах и не принадлежать какой-то конкретной теме. Вследствие этого их вероятностное распределение по темам будет получаться почти равномерным, что может ухудшить качество тематической модели. Во избежание этого решено было попробовать оставить в коллекции слова, необладающие подобным свойством, и посмотреть, как вследствие этого меняется качество. Перед этим введем несколько определений.

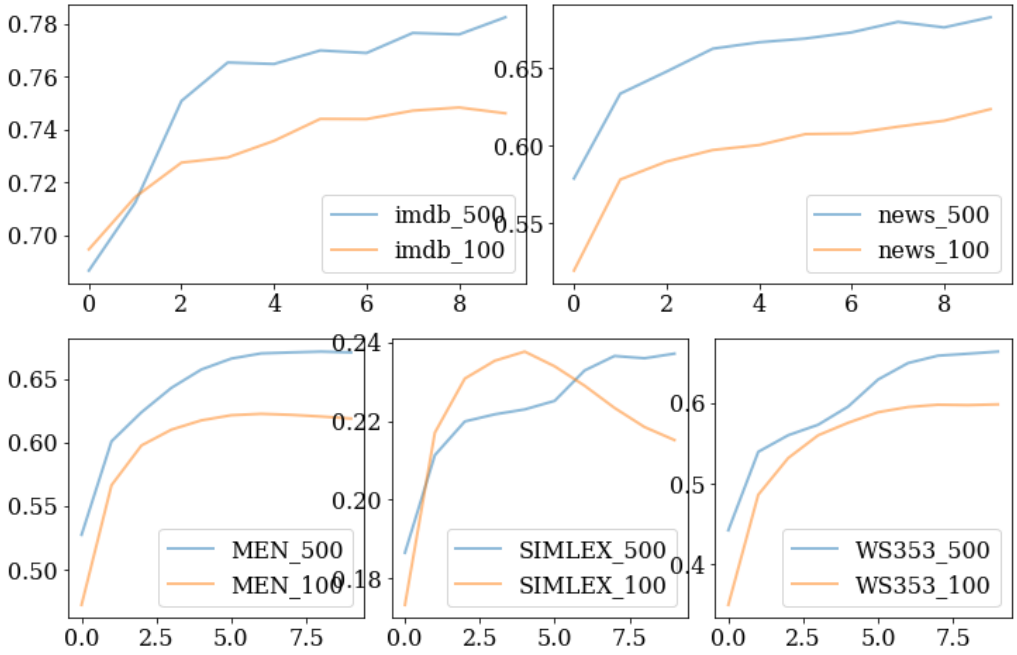

Рис. 3: Количество тем

Определение 4.1. ***Документной частотой*** *называется число документов коллекции, в которых встретилось данное слово:*

$$DF(w) = |\{d \in D|\ n_{dw} > 0\}|.$$

Определение 4.2. ***Энтропией*** *слова u называется энтропия глобального контекста этого слова:*

$$H(u) = -\sum_{v \in C_u} p(v|u) \log p(v|u).$$

Максимум энтропии достигается при равномерном распределении. Следовательно, чем меньше энтропия терма, тем дальше распеределение его контекста от равномерного, и тем меньше различных слов входит в его контекст (тем «уже» его контекст). Это означает, что рассматриваемое слово может оказаться термином в какой-то теме, и, таким образом, улучшать тематическую модель. Известно, что $H(w) \leq C + \beta \log DF(w)$, где C и β — некоторые константы. Разобьем все документные частоты слов на r интервалов, и множество слов на r подмножеств согласно этим интервалам: $W_i = \{w \in W|\ DF^{(i)} \leq DF(w) < DF^{(i+1)}\}$. Получим, что область значений функции энтропии для слов с одинаковой документной частотой ограничено сверху логарифмом документной частоты. Так как при увеличении документной

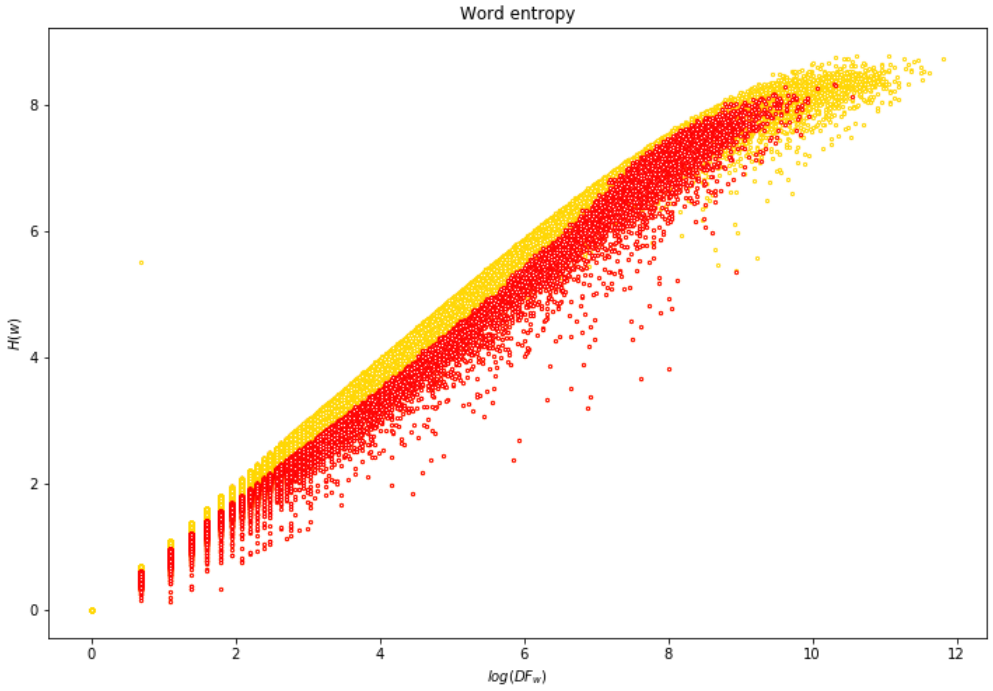

Рис. 4: Зависимость энтропии от документной частоты (желтый — все слова, красный — слова с узкими контекстами

частоты энтропия может возрастать даже для слов-терминов, вместо отбрасывания множеств W_i с большими DF, стоит прибегнуть к квантильной регрессии: для каждой DF оставить лишь те слова, чья энтропия меньше некоторого τ-квантиля. Таким образом, останутся только слова, удовлетворяющие условиям:

$$\begin{cases} DF^{(i)} \leq DF(w) < DF^{(i+1)} \\ H(w) \leq H^{(i)}, \end{cases} \tag{4.1}$$

где $H^{(i)}$ — τ-квантиль вариационного ряда энтропий для i-го интервала.

Определение 4.3. *Центральное слово $u \in W$ обладает **узким контекстом**, если для некоторого $i = 1, ..., r$ выполняется условие* (4.1).

В данной работе разбиение документных частот проводилось поэлементно, и в каждом интервале содержится только одна документная частота. График зависимости $H(w)$ от $\log DF(w)$ при $\tau = 0.05$, можно увидеть на рисунке 4. Можно видеть, что для многих значений DF большая часть плотности распределения энтропии сосредоточена в наибольших значениях вариационного энтропийного ряда.

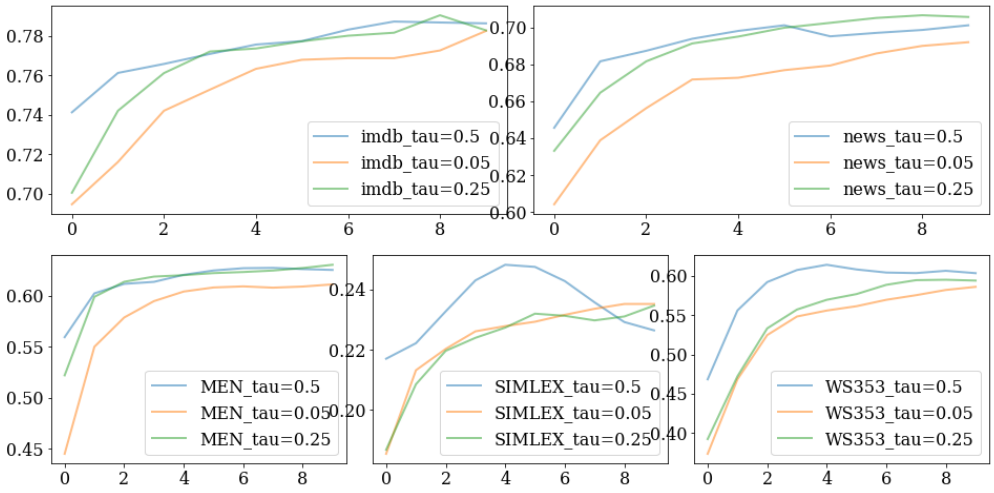

Рис. 5: Различные значения τ

На рис. 5 можно видеть, как меняется качество в зависимости от τ с ростом числа итераций. Можно увидеть, что в среднем модель, построенная на узких контекстах с $\tau = 0.25$ ведет себя почти так же, как и при 0.5, а где-то и лучше. Также эта модель почти везде превосходит модель с $\tau = 0.05$. В связи с этим в дальнейших экспериментах модель, построенная на узких контекстах будет рассматриваться именно при $\tau = 0.25$.

4.5.2 Эксперименты на разных корпусах

Помимо обычного корпуса и корпуса узких контекстов, было решено попробовать строить тематическую модель по матрице PPMI (упоминалось в подразделе 2.2.1) вместо обычных частот совстречаемостей.

Как видно из графиков (см. рис 6), PPMI гораздо быстрее сходится, но при этом наименьшее качество имеет на задаче близости слов. Модель узких контекстов на классификации сходится быстрее и имеет схожее качество или даже лучшее, как и у PPMI. Модель, построенная на обычном корпусе, ведет себя наилучшим образом или примерно на том же уровне на всех задачах, проигрывая только на классификации новостей. Решено было остановиться на этой модели, хотя стоит протестировать все подходы для большего количества тем.

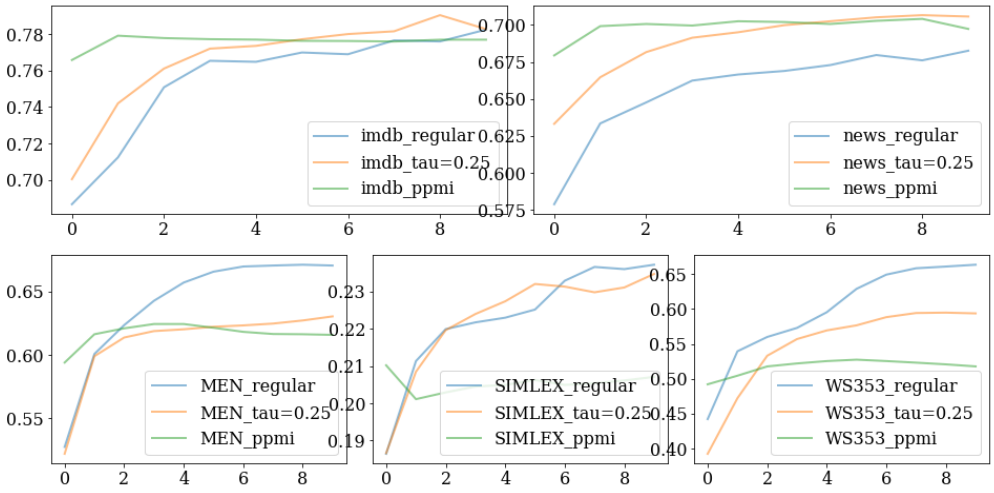

Рис. 6: Сравнение результатов на обычном корпусе, PPMI и узких контекстах при $\tau = 0.25$

type	T	ep	imdb	news	MEN	SIMLEX999	WS353
glove	300	-	**0.835**	**0.730**	**0.737**	**0.371**	0.543
Φ_b	750	4	0.755	0.710	0.647	0.252	**0.594**
Φ_b	750	5	0.763	0.713	0.649	0.259	0.585
Φ_b	750	6	0.768	0.718	0.645	0.263	0.576
Φ_b	750	7	0.769	0.719	0.642	0.266	0.573
Φ_b	750	8	0.769	0.721	0.634	0.270	0.558
Φ_b	750	9	0.766	0.723	0.630	0.268	0.542

Таблица 1: $T = 750$

4.6 Лучшие результаты

Одной из наилучшим образом показавших себя моделей является модель, построенная на 750 темах, результаты можно видеть в таблице 1. Данная модель превосходит смогла превзойти бейзлайн на 5% только на задаче WS353, причем качество на разных задачах ведет себя по-разному: к примеру, в то время как на классификации качество монотонно возрастает, на WS353 оно убывает уже после 4 итерации.

Увеличив количество тем до 2000, мы смогли получить лучшее качество также на задаче классификации новостей, хотя, как видно из таблицы 2, хорошее качество на классификации и WS353 достигаются на разных итерациях. Стоит также отметить, что на 750 темах качество на WS353 было выше. Возможно, это связано с тем, что WS353 является не самым качественным датасетом, и его недостатки не раз упо-

type	T	ep	imdb	news	MEN	SIMLEX999	WS353
glove	300	-	**0.835**	0.730	**0.737**	**0.371**	0.543
Φ_b	2000	4	0.811	0.724	0.643	0.272	**0.553**
Φ_b	2000	5	0.821	0.730	0.641	0.285	0.537
Φ_b	2000	6	0.825	0.733	0.631	0.283	0.534
Φ_b	2000	7	0.828	0.737	0.617	0.276	0.517
Φ_b	2000	7	0.827	**0.739**	0.612	0.277	0.497

Таблица 2: $T = 2000$

Тема 1	Тема 2	Тема 3
aikido	asheboro	risk
ueshiba	greensboro	outcom
parkour	ashevill	bias
karat	kannapoli	consequ
dojo	fairborn	factor

Таблица 3: Топы слов для нескольких тем при $T = 500$

минались в литературе, потому как на SimLex-999 качество растет при увеличении количества тем, и авторы этого датасета старались нивелировать недостатки WS353.

4.7 Интерпретируемость векторных представлений

В таблице 3 приводятся примеры слов, обладающие наибольшей вероятностью при разных темах. Как видно, данные темы несложно проинтерпретировать: Тема 1 соответствует боевым искусствам, Тема 2 — городам, Тема 3 — что-то, связанное с риском, последствиями, результатами.

5 Заключение

В данной работе были получены интерпретируемые тематические векторные представления, обученные по корпусу Википедии. Проведен анализ качества модели на различных модификациях исходного корпуса совстречаемостей. Также были получены результаты, превосходящие бейзлайн на некоторых задачах.

В дальнейшем планируется исследовать поведение модели с бóльшим количеством тем на модифицированных корпусах, добавить регуляризаторы декоррелирования и разреживания матрицы Θ, использовать иерархическую тематическую модель, добавить дополнительные модальности. Также следует исследовать другие способы построения векторных представлений документов для задачи классификации и протестировать данную модель на других задачах и других языках.

Список литературы

[1] *Bassiou N. K., Kotropoulos C. L.* Online plsa: Batch updating techniques including out-of-vocabulary words // *IEEE transactions on neural networks and learning systems.* — 2014. — Vol. 25, no. 11. — Pp. 1953–1966.

[2] Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // *arXiv preprint arXiv:1810.04805.* — 2018.

[3] *Bruni E., Tran N.-K., Baroni M.* Multimodal distributional semantics // *Journal of Artificial Intelligence Research.* — 2014. — Vol. 49. — Pp. 1–47.

[4] *Church K. W., Hanks P.* Word association norms, mutual information, and lexicography // *Computational linguistics.* — 1990. — Vol. 16, no. 1. — Pp. 22–29.

[5] Deep contextualized word representations / M. E. Peters, M. Neumann, M. Iyyer et al. // *arXiv preprint arXiv:1802.05365.* — 2018.

[6] Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen et al. // Advances in neural information processing systems. — 2013. — Pp. 3111–3119.

[7] Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // *arXiv preprint arXiv:1301.3781.* — 2013.

[8] Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // *Transactions of the Association for Computational Linguistics.* — 2017. — Vol. 5. — Pp. 135–146.

[9] *Firth J. R.* A synopsis of linguistic theory, 1930-1955 // *Studies in linguistic analysis.* — 1957.

[10] *Harris Z. S.* Distributional structure // *Word.* — 1954. — Vol. 10, no. 2-3. — Pp. 146–162.

[11] *Hill F., Reichart R., Korhonen A.* Simlex-999: Evaluating semantic models with (genuine) similarity estimation // *Computational Linguistics.* — 2015. — Vol. 41, no. 4. — Pp. 665–695.

[12] *Hirst G., St-Onge D. et al.* Lexical chains as representations of context for the detection and correction of malapropisms // *WordNet: An electronic lexical database.* — 1998. — Vol. 305. — Pp. 305–332.

[13] *Hoffman M., Bach F. R., Blei D. M.* Online learning for latent dirichlet allocation // advances in neural information processing systems. — 2010. — Pp. 856–864.

[14] *Hofmann T.* Probabilistic latent semantic analysis // UAI. — 1999.

[15] Indexing by latent semantic analysis / S. Deerwester, S. T. Dumais, G. W. Furnas et al. // *Journal of the American society for information science.* — 1990. — Vol. 41, no. 6. — Pp. 391–407.

[16] *Jarmasz M., Szpakowicz S.* Roget's thesaurus and semantic similarity // *Recent Advances in Natural Language Processing III: Selected Papers from RANLP.* — 2004. — Vol. 2003. — P. 111.

[17] *Lang K.* Newsweeder: Learning to filter netnews // Machine Learning Proceedings 1995. — Elsevier, 1995. — Pp. 331–339.

[18] *Leacock C., Chodorow M.* Combining local context and wordnet similarity for word sense identification // *WordNet: An electronic lexical database.* — 1998. — Vol. 49, no. 2. — Pp. 265–283.

[19] Learning word vectors for sentiment analysis / A. L. Maas, R. E. Daly, P. T. Pham et al. // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Portland, Oregon, USA: Association for Computational Linguistics, 2011. — June. — Pp. 142–150.
`http://http://www.aclweb.org/anthology/P11-1015`

[20] *Levy O., Goldberg Y.* Neural word embedding as implicit matrix factorization // Advances in neural information processing systems. — 2014. — Pp. 2177–2185.

[21] *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings // *Transactions of the Association for Computational Linguistics.* — 2015. — Vol. 3. — Pp. 211–225.

[22] *Lund K., Burgess C.* Producing high-dimensional semantic spaces from lexical co-occurrence // *Behavior research methods, instruments, & computers.* — 1996. — Vol. 28, no. 2. — Pp. 203–208.

[23] *Manning C., Raghavan P., Schütze H.* Introduction to information retrieval // *Natural Language Engineering.* — 2010. — Vol. 16, no. 1. — Pp. 100–103.

[24] *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — Pp. 1532–1543.

[25] Placing search in context: The concept revisited / L. Finkelstein, E. Gabrilovich, Y. Matias et al. // *ACM Transactions on information systems.* — 2002. — Vol. 20, no. 1. — Pp. 116–131.

[26] *Potapenko A., Popov A., Vorontsov K.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks // Conference on Artificial Intelligence and Natural Language / Springer. — 2017. — Pp. 167–180.

[27] *Rogers A., Drozd A., Li B.* The (too many) problems of analogical reasoning with word vectors // Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017). — 2017. — Pp. 135–148.

[28] *Schütze H., Pedersen J.* A vector model for syntagmatic and paradigmatic relatedness // Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research / Oxford. — 1993. — Pp. 104–113.

[29] *Vorontsov K., Potapenko A.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2014. — Pp. 29–46.