Моделирование и анализ данных
2022. Том 12. № 2. С. 20–33
DOI: https://doi.org/10.17759/mda.2022120202
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 2, pp. 20–33
DOI: https://doi.org/10.17759/mda.2022120202
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.855

Разработка метода извлечения ключевых слов на основе вероятностной тематической модели

Романадзе Е.Л.*
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-0351-7235
e-mail: katia_rom.97@mail.ru

Судаков В.А.**
Московский авиационный институт (МАИ), г. Москва, Российская Федерация
ИПМ им. М.В.Келдыша РАН, г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

Кислинский В.Г.****
Московский физико-технический институт (МФТИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-2000-583X

В работе рассмотрена задача тематического моделирования. Для анализа коллекции документов, описывающих товары онлайн-магазина, разработан новый метод извлечения ключевых слов на основе тематического моделирования. Проведен сравнительны анализ базового метода извлечения ключевых слов и предложенного метода. Приведены наглядные результаты, описывающие

Для цитаты:
Романадзе Е.Л., Судаков В.А., Кислинский В.Г. Разработка метода извлечения ключевых слов на основе вероятностной тематической модели // Моделирование и анализ данных. 2022. Том 12. № 2. С. 20–33. DOI: https://doi.org/10.17759/mda.2022120202

***Романадзе Екатерина Левановна,** студент магистратуры, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-0351-7235, e-mail: katia_rom.97@mail.ru

****Судаков Владимир Анатольевич,** доктор технических наук, профессор, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), ведущий научный сотрудник, Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В.Келдыша РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

*****Кислинский Вадим Геннадиевич,** научный сотрудник, Московский физико-технический институт (национальный исследовательский университет) (МФТИ (НИУ)), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-2000-583X

преимущества данного подхода. Полученное решение может быть использовано для упрощения навигации по сайту и поиска релевантных товаров.

Ключевые слова: извлечение ключевых фраз, тематическое моделирование, NLP, LDA, машинное обучение

1. ВВЕДЕНИЕ

Каждый день создается огромное количество текстов, и люди сталкиваются с большим объемом информации, которую нужно анализировать и потреблять. С учетом увеличения окружающих нас данных, этот процесс становится все более сложным, однако не менее необходимым. Актуальном инструментарием, облегчающим решение данной проблемы, является автоматическое извлечение ключевых слов из текста. Извлечение ключевых слов – это определение слов или фраз, которые наилучшим образом выражают основные концепции текстов. Таким образом, опираясь на термины, которые характеризуют документ, человек сможет принять решение о детальном прочтении представленной информации. На практике извлечение ключевых слов решает и многие другие проблемы, такие как:

- упрощение поисковой системы – прокачка индексов/семантический поиск/ расширение поисковых запросов;
- реферирование текстов;
- классификация/кластеризация документов;
- выделение признаков (интерпретируемых топиков) для рекомендательных систем и поведенческого анализа;
- рекламные системы – контекстная реклама и выделение коммерческих интересов [1].

В представленной работе объектом исследования является набор текстов (документов), описывающих товары, представленные в онлайн каталоге интернет-магазина OZON. Предмет исследования – методы извлечения ключевых слов.

Цель работы – извлечение ключевых слов из описаний товаров, что приведет к упрощению их поиска. Для этого были проработаны сопутствующие задачи:

1. Проанализированы полученные данные.
2. Разработан метод для извлечения ключевых слов из документов.
3. Проведен вычислительный эксперимент.
4. Оценена эффективность работы модели.

2. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ

Перед нами стоит задача извлечения ключевых слов, для ее решения рассмотрим исходные данные. Предоставляется коллекция документов, связанная общей тематикой (новости, рассказы, статьи и т.д.) – $\boldsymbol{D}$. Каждый документ $d \in D$ состоит из набора слов $w_1, \ldots, w_{n_d}$, где n_d – это количество слов в документе d. Цель – выделить ключевые слова/фразы K_{l_d} (l_d – это количество ключевых слов в документе d), со-

стоящие из W_{d}, которые будут описывать смысловую часть заданного документа *d*. При этом искомые ключевые слова/фразы K_{l_d} могут быть заданы изначально или вовсе неизвестны.

Схема эксперимента для решения задачи представлена на рисунке 1. В основу метода входит модель тематического моделирования – LDA. Для получения наиболее быстрых и качественных результатов данные обрабатываются стандартными методами, а также посредством статистической меры IDF. Непосредственное извлечение ключевых слов происходит на последнем этапе.

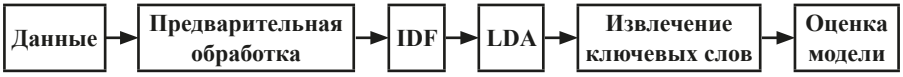

Рис. 1. Схема эксперимента

Далее рассмотрим каждый пункт по отдельности, предварительно ознакомившись с теоретической основой и базовыми понятиями, применяемыми в работе с текстом, а также непосредственно в тематическом моделировании.

3. ПРЕДОБРАБОТКА

Для решения задач NLP данные первым делом подвергаются предварительной обработке. Ее цель заключается в очищении текста от незначимых и неинформативных данных, что упрощает и облегчает работу. Правильная предварительная обработка в первую очередь сказывается на качестве оценки результатов. Далее опишем базовые способы, применяемые к текстам в качестве предварительной обработки.

В первую очередь из текстов убирается вся пунктуация, так как анализу подлежат непосредственно слова. Далее слова приводятся к нижнему регистру поскольку одно и то же слово, написанное в разных регистрах, моделью будет восприниматься как два разных слова. Получившиеся тексты подвергаются токенизации. Токенизация – это задача разделения текста на части, называемые токенами, таким образом каждый документ можно представить в виде списка слов или словосочетаний, из которых он состоит.

Далее данные подвергаются лемматизации или стеммингу. Лемматизация – это приведение слов к нормальному виду. Например, слово «красивое» будет преобразовано в «красивый», слово «убежал» в «убежать». Данные преобразования считаются наиболее точными, но как правило занимают больше времени. Стемминг – это процесс отбрасывания окончаний или других изменяемых частей слов. Например, слово «красивое» будет преобразовано в «красив», а слово «убежал» в «беж». Подобные преобразования занимают меньше времени, но могут приводить к спорным результатам, так как урезав большую часть слова может быть утерян его смысл и урезанное слово будет в дальнейшем трактоваться некорректно. Следует отметить, что стемминг в большей степени применяется к английским словам, так как лексически выдает более точные результаты, чем при применении к словам русского языка. В конечном счете после лемматизации/стемминга алгоритму проще воспринимать связь между словами и анализировать их, так как одно и то же слово, представленное

в разных падежах, регистрах или времени, должно восприниматься одинаково, так как имеет как правило идентичную смысловую нагрузку.

Следующим шагом осуществляется удаление стоп-слов. Под стоп-словами подразумеваются слова, которые часто встречаются во всех документах или в документах представленной тематики. Они считаются общими и теряют свою ценность за счет частоты, так как явно не являются ключевыми для документов и потому не несут смысловой нагрузки. Также в список стоп-слов могут быть включены слова, которые, наоборот, крайне редко встречаются. Если слово было встречено всего один раз во всей коллекции документов, то оно также имеет малую эффективность в анализе, поэтому подобные слова следует исключить.

Перечисленная обработка является фундаментальной для работы с текстами, но в отдельных случаях могут потребоваться дополнительные преобразования. К примеру, помимо отдельно стоящих слов можно выделять устойчивые фразы, словосочетания автоматическими методами или же рассматривать для анализа только определенные части речи, также с помощью регулярным выражений можно избавиться от лишних данных.

TF-IDF (сокращение от term frequency – inverse document frequency) – это статистическая мера для оценки важности слова в документе, который является частью коллекции.

Для ее расчета вычисляются меры TF (term frequency – частота слова), и IDF (inverse document frequency – обратная частота документа). При этом TF оценка слов меняется от документа к документу, а IDF оценка слова одинаковая для слов внутри каждого документа.

TF оценивает частоту некоторого слова внутри документа по следующей формуле:

$$tf\left(t,d\right)=\frac{n_t}{\sum_k n_k}\quad , \tag{1}$$

где n_t – число вхождений слова t в документ, а в знаменателе – общее число слов в данном документе.

IDF оценивает обратную частоту документов, включающих в себя некоторое слово и выражается в представленной ниже формуле.

$$idf\left(t,D\right)=ln\frac{\left|D\right|}{\left|\left\{d_i\in D|t\in d_i\right\}\right|} \tag{2}$$

где $\left|D\right|$ – число документов в коллекции, $\left|\left\{d_i\in D|t\in d_i\right\}\right|$ – число документов из коллекции D, в которых встречается t.

Таким образом для оценки слов с помощью TF-IDF перемножаются две рассмотренные меры, и по итогу вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

$$tf\text{-}idf\left(t,d,D\right)=tf\left(t,d\right)\times idf\left(t,D\right) \tag{3}$$

Больший вес по TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. Мера TF-IDF часто используется в задачах анализа текстов и информационного поиска, например, как один из критериев релевантности документа поисковому запросу, при расчёте меры близости документов при кластеризации.

4. ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Тематическое моделирование – одно из современных направлений обработки естественного языка. Тематическая модель коллекции текстовых документов определяет к каким темам относится каждый документ и какие слова образуют каждую тему. Алгоритм описывает темы дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем. Такой подход напоминает кластеризацию, однако, отличие в том, что при кластеризации документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет мягкую кластеризацию, разделяя документ между несколькими кластерами.

Исходные данные – коллекция текстовых документов D, при этом каждый документ d из D представляется как последовательность термов $W_d = \left(w_1, \ldots, w_{n_d}\right)$, где n_d – количество термов документа d. Термами считаются слова, словосочетания, цифры или иные сущности, которые входят в документ, в зависимости от того какой предварительной обработке подверглись документы. Предполагается, что каждый документ описан одной или несколькими темами, а темы различаются частотой употребления отдельных термов. Таким образом коллекцию документов можно представить в виде последовательности троек . . Так как термы и документы известны, на их основе предполагается выявить темы. В связи с этим требуется найти:

1) число тем;
2) слова, характерные для каждой темы, и их распределения;
3) принадлежность документов к темам.

Ниже представлены задачи, которые можно решить с помощью данной модели:

- ранжирование документов по заданной тематике;
- ранжирование документов по степени сходства с заданными документом;
- определение тематики различных сущностей (конференций, журналов и т.д.);
- определение тематики авторов.

Для решения задачи необходимо определить основные предположения вероятностных тематических моделей. Так, предполагается, что:

- порядок документов в коллекции не влияет на результат;
- порядок термов внутри документа также не влияет на результат – используется «мешок слов»;
- термы, которые часто встречаются во всех документах, не имеют смысловой нагрузки и поэтому удаляются из списка (стоп-слова);
- слова, написанные в разных формах, считаются одинаковыми и приводятся к одному виду;

- каждая тема $t \in T$ описывается неизвестным распределением $p(w|t)$ на множестве термов $w \in W$;
- каждый документ $d \in D$ описывается неизвестным распределением $p(t|d)$ на множестве тем $t \in T$;
- гипотеза условной независимости $p(w|t,d) = p(w|t)$. Она предполагает, что появление термов в документе d по теме t зависит непосредственно от темы и описывается общим распределением $p(w|t)$ [5].

Распределение термов в документе $p(w|d)$ описывается вероятностной смесью распределений термов в темах $\phi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$.

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d) = \sum_{t \in T} \phi_{wt}\, \theta_{td} \tag{4}$$

Для поиска приближенных значений матриц ϕ_{wt} и θ_{td} максимизируется логарифм правдоподобия:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\, \theta_{td} \to \max_{\Phi, \Theta}. \tag{5}$$

Для максимизации (5) вычисляется EM-алгоритм, в котором итерационно чередуются E и M шаги. Рациональный алгоритм представлен на рисунке 2, где ϕ_{wt} – вероятности термов w в каждой теме t; θ_{td} – вероятности тем t в каждом документе d; n_{wt} – число троек, в которых терм w связан с темой t; n_{td} – число троек, в которых терм документа d связан с темой t; n_t – число троек, связанных с темой t; n_{tdw} – число троек, в которых терм w документа d связан с темой t; n_d – длина документа d в термах; n_{dw} – число вхождений терма w в документ d [2].

Вход: коллекция D, число тем |T|, начальные приближения матриц ϕ_{wt} и θ_{td};
Выход: параметры ϕ_{wt} и θ_{td}

1 **Повторять**

2 Обнулить n_{wt}, n_{td}, n_t для всех $d \in D, w \in W, t \in T$;

3 **Для всех** $d \in D, w \in W$

4 $$n_{tdw} \leftarrow \frac{n_{dw} \phi_{wt} \theta_{td}}{\sum_t \phi_{wt} \theta_{td}} \text{ для всех } t \in T$$

5 Увеличить n_{wt}, n_{td}, n_t на n_{tdw} для всех $t \in T$;

6 $$\phi_{wt} \leftarrow \frac{n_{wt}}{n_t}$$ для всех $w \in W$, $t \in T$;

7 $$\theta_{td} \leftarrow \frac{n_{td}}{n_d}$$ для всех $d \in D$, $t \in T$;

8 **Пока** ϕ_{wt} и θ_{td} не сойдутся;

Рис. 2. Рациональный EM-алгоритма для тематической модели

Данный алгоритм называют EM-алгоритмом, где на E-шаге (expectation) происходит оценка условного распределения латентных тем n_{tdw} по формуле Байеса для всех терминов в документах, а на M-шаге (maximization) по этим вероятностям вычисляются частотные оценки матриц ϕ_{wt} и θ_{td}.

5. АДДИТИВНАЯ РЕГУЛЯРИЗАЦИЯ И LDA

Для решения проблем неустойчивости и неединственности используются регуляризаторы. На искомое решение накладываются дополнительные ограничения. Аддитивная регуляризации или подход ARTM основан на идее многокритериальной регуляризации, при котором вводятся еще *n* критериев $R_i(\Phi, \Theta)\ i = 1,2,\ldots,n$. Взвешенная сумма всех таких критериев:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta), \tag{6}$$

где τ_i – неотрицательный коэффициент регуляризации, максимизируется совместно с основным критерием правдоподобия:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \tag{7}$$

Задача решается также EM-алгоритмом, модифицируя M шаг следующим образом:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), \tag{8}$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right). \tag{9}$$

LDA (latent Dirichlet allocation – латентное размещение Дирихле) является наиболее цитируемой моделью тематического моделирования. Основная идея заключается в предположении, что матрицы Θ и Φ являются случайными векторами и порождаются распределением Дирихле с гиперпараметрами $\alpha \in \mathbb{R}^T$ и $\beta \in \mathbb{R}^W$ соответственно:

$$Dir(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1},\ \alpha_t > 0,\ \alpha_0 = \sum_t \alpha_t,\ \theta_{td} > 0,\ \sum_t \theta_{td} = 1; \tag{10}$$

$$Dir(\phi_t; \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1},\ \beta_w > 0,\ \beta_0 = \sum_w \beta_w,\ \phi_{wt} > 0,\ \sum_w \phi_{wt} = 1; \tag{11}$$

где $\Gamma(z)$ – гамма-функция, β_w и α_t столбцы матриц β и α, а α_0 и β_0 – коэффициенты регуляризации. Параметры распределения Дирихле связаны с математическим ожиданием порождения случайных векторов: $\mathrm{E}\theta_{td} = \frac{\alpha_t}{\alpha_0}, \mathrm{E}\phi_{td} = \frac{\beta_t}{\beta_0}$ [4].

В терминах ARTM модель LDA выражается через сглаживающие регуляризаторы следующим образом:

$$R(\Phi, \Theta) = \beta_0 \sum_{t,\, w} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d,t} \alpha_t \ln \theta_{td} \ . \quad (12)$$

Подставив данный критерий формулу М-шага, получим новое его представление:

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \beta_0 \beta_w \right), \quad (13)$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \alpha_0 \alpha_w \right). \quad (14)$$

6. РЕАЛИЗАЦИЯ ЭКСПЕРИМЕНТА

Цель работы заключается в упрощении навигации по сайту путем извлечения ключевых слов из описаний товаров. В качестве данных имеется набор annotations из 400 тысяч описаний, характеризующих товары онлайн магазина OZON. Так как у анализируемых данных отсутствуют искомые, заранее известные ключевые слова, то точность эксперимента будет оцениваться как эмпирически, так и за счет заранее отобранных данных, для которых ключевые слова уже прописаны. Таким образом мы имеем три набора данных: annotations – обучающая выборка и 500N-KPCrowd-v1.1, SemEval2017 – тестовые выборки.

Набор данных annotations состоит в общей сумме из 29,5 млн слов, из них 260 тысяч уникальных. Описания представлены на русском языке. На рисунке 3 продемонстрирована частота встречаемых слов в документах. Как видно большая часть встречается только один раз.

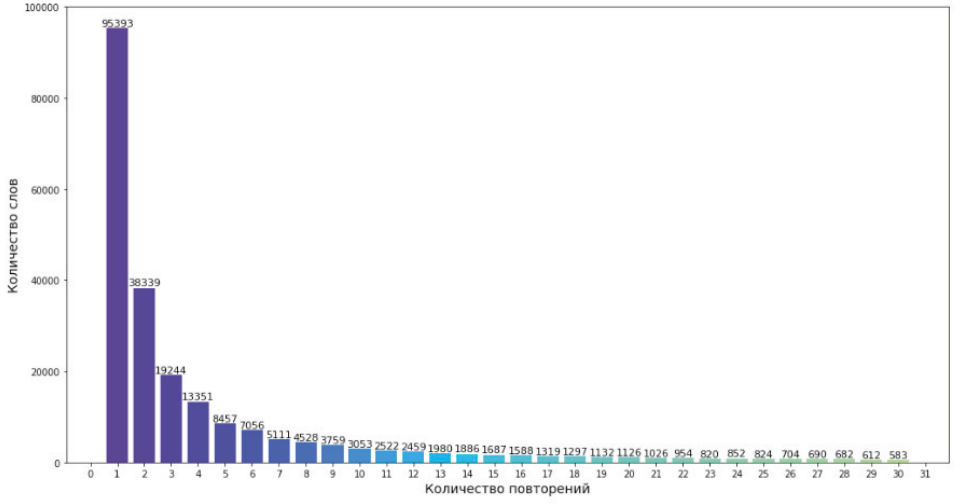

Рис. 3. Частота встречаемых слов в документах

Рисунок 4 показывает количества слов в документах. В среднем описание товара состоит из 20–25 слов.

Набор данных 500N-KPCrowd-v1.1 представляет собой сводку новостей вместе с заголовками. Датасет описан на английском языке. Для каждой новости есть определенный набор заданных ключевых слов. Всего представлено 500 новостных статей.

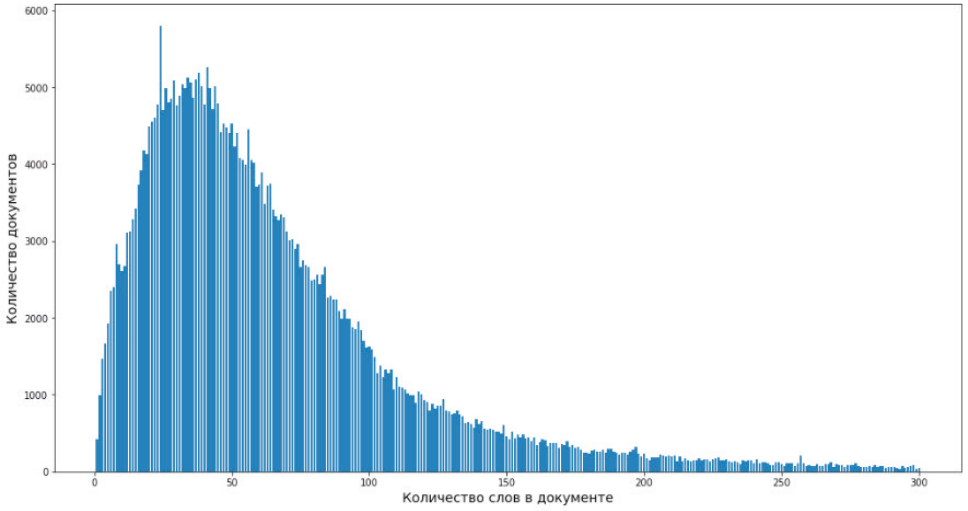

Рис. 4. Частота слов в документе

Набор SemEval2017 состоит из 500 документов, выбранных из журнальных статей ScienceDirect, равномерно распределенных по областям информатики, материаловедения и физики. Статьи представлены на английском языке

В таблице 1 приведены основные характеристики рассматриваемых датасетов.

Таблица 1

Характеристики датасетов

Данные	Язык	Количество документов	Количество слов	Количество уникальных слов
annotations	Русский	400000	29.5 млн	260 т
500N-KPCrowd-v1.1	Английский	500	116 т	16 т
SemEval2017	Английский	500	49 т	8 т

Предобработка данных осуществлялась с помощью языка программирования python. На этапе предобработки с помощью регулярных выражений данные почистили от пунктуации html-тэгов, а также удалили слова, состоящие из одного буквы. Токенизация и удаление стоп-слов было осуществлено с помощью библиотека nltk. Для приведения слов к единой форме была использована лемматизация из nltk.stem. WordNetLemmatizer() и pymorphy2.MorphAnalyzer(). По итогу обработки получили релевантные данные для каждого документа. Во-первых, данных стало гораздо меньше, что способствует ускорению работы алгоритмов, во-вторых, сами слова представлены в удобном виде, подлежащем анализу.

Следующим этапом применялся статистический метод IDF для отбора кандидатов, наиболее информативных слов. Получив результаты, слова с наибольшими и наименьшими значениями IDF были удалены, так как первые можно отнести в группу стоп-слов, которые слишком часто встречаются, а вторая группа наоборот – редкие единичные слова, которые также не подлежат анализу.

Итоговые данные были переведены в формат BOW (Bag-of-words) и разбиты на темы с помощью тематического моделирования. Для реализации была использована модель LDA с аддитивной регуляризация из библиотеки BigARTM. Bag-of-Words или мешок слов – это модель, часто используемая при обработке текстов, представляющая собой неупорядоченный набор слов, входящих в обрабатываемый текст. В этой модели документ представляется в виде мешка его слов с сохранением информации об их количестве [3].

В ходе эксперимента данные были разбиты на 150 тем. На рисунке 6 представлена визуализация тем, для наглядности выведены по два слова из 13 тем. Для визуализации за основу была взята матрица ϕ, в которой хранятся вероятности попадания слова в определенную тему. Размерность матрицы была уменьшена до 2 с помощью t-SNE – алгоритм машинного обучения, базовый принцип которого заключается в сокращении попарных расстояний между точками при сохранении их относительного расположения. Иными словами, алгоритм отображает многомерные данные на пространство более низкой размерности, при этом сохраняя структуру соседства точек.

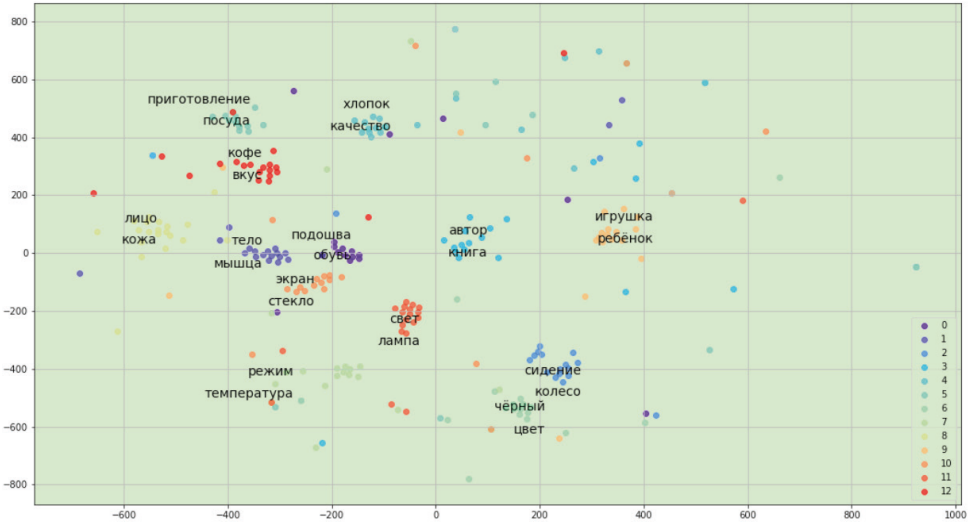

Рис. 6. Распределение тем

Для удобства было выведено по 30 первых слов в каждой теме. График напоминает задачу кластеризации, как и было упомянуто выше. Есть слова, которые строго попадают под свою тематику, а также слова, которые находятся в промежуточном состоянии, так как могут с приблизительно одинаковой вероятностью относится к нескольким темам одновременно.

7. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

По итогам работы модели каждому документу были присвоены топ 3 темы наиболее подходящие и из каждой темы выделены топ 3 слова. Пример отобранных ключевых слов показан на рисунке 7.

Коврик-пазл «Сказочная принцесса» – увлекательная, развивающая игрушка, на которой малыш сможет лежать, сидеть, ползать и играть. Коврик состоит из двадцати четырех текстурированных сегментов, с красочным изображением принцесс, цифр от 0 до 9 и других волшебных предметов. Таким образом, в игровой форме малыш развивает не только визуальное и тактильное восприятие предметов, но и в игровой форме научится счету. Коврик помогает развитию мелкой моторики, воображения, мышления и пространственного восприятия ребенка.

Коврик-пазл создан из экологически чистого полимерного материала – ЭВА (этиленвинилацетат).

Основные достоинства:

– хорошо амортизирует
– гипоаллергенный
– не электризуется
– легкий
– стойкий к воздействию химических веществ
– препятствует образованию и размножению грибка, а также болезнетворных бактерий

Для детей с 10 месяцев

Размер коврика в собранном состоянии: 62 х 93 х 1 см

Размер упаковки: 32 х 32 х 6 см

Рис. 7. Пример отобранных ключевых слов для описания

На рисунке 7 представлено описание товара «коврик-пазл «Сказочная принцесса». Каждая тема показана отдельным цветом. Ключевые слова «коврик», «материал» и «размер» характеризуют сам товар – коврик, слова следующей тема – «игрушка», «малыш» и «ребенок» объединены детской тематикой, что указывает на то, для кого предназначен товар, а следующая тройка слов – «развивает», «восприятие» и «мышление» указывают на пользу товара. Следовательно, по данному примеру можно сказать, что алгоритм отработал корректно.

Оценка качества модели была произведена с помощью расчетов для тестовой выборки. Результаты приведенного эксперимента сравнивались с базовым алгоритмом, в основу которого вошло выделение ключевых слов с помощью метода TF-IDF, то есть отбиралось также по 9 ключевых слов из каждого документа, которые имеют наибольший вес.

Так как у данных 500N-KPCrowd-v1.1 и SemEval2017 есть заранее известные ключевые слова, на их основе была посчитана метрика precision, которая вычисляет соотношение количества правильно отобранных ключевых к общему количеству искомых ключевых слов:

$$precision = \frac{TP}{TP + FP},$$

где TP – количество истинных срабатываний, а FP – количество ложных срабатываний соответственно.

Для сравнения к тестовым данным был применен базовый алгоритм, который оценивал слова по TF-IDF. По результатам базового алгоритма precision для датасета 500N-KPCrowd-v1.1 precision составил 0.38 и 0.43 соответственно. После проведения описанного эксперимента метрика возросла и составила 0.56 для 500N-KPCrowd-v1.1 и 0.62 для SemEval2017. Подробнее результаты описаны в таблице 2.

Таблица 2

Результаты эксперимента

Данные	Базовый алгоритм (TF-IDF)	Представленный метод
500N-KPCrowd-v1.1	0.38	0.56
SemEval2017	0.43	0.62

8. ЗАКЛЮЧЕНИЕ

В результате проделанной работы извлечены ключевые слова из описаний товаров для упрощения поиска. Была разработана модель тематического моделирования LDA, основанная на подходе ARTM. Данные, используемые для проведения эксперимента, – список описаний товаров в размере 400 тысяч документов. Для проведения эксперимента данные были подвержены предварительной обработке, выделении кандидатов в ключевые слова с помощью статистических методов и обработаны с помощью модели LDA для получения результатов. Оптимальное решение было достигнуто выделением 150 тем из всего списка документов. Конечный вывод ключевых слов состоял из выделения наиболее значимых трех слов по каждой из трех приоритетных тем, выделенных для конкретного документа. Качество работы оценивалось на тестовых данных – готовые данные 500N-KPCrowd-v1.1 и SemEval2017 с выделенными ключевыми словами. Метрика precision показала результаты в 0.56 и 0.62 соответственно, что выше результата базового алгоритма.

Литература

1. *Augenstein, I., Das, M., Riedel, S., Vikraman, L. and McCallum, A.* (2017) Semeval 2017 task 10: Scienceie – extracting keyphrases and relations from scientific publications. In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3–4, 2017, 546–555. URL: https://doi.org/10.18653/v1/S17–2091.
2. *Апишев М.А.* Эффективные реализации алгоритмов тематического моделирования: дис. канд. физ-мат наук: 230401. – М., 2020. – 152 с.
3. *Воронцов К.В.* Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf.
4. *Воронцов К., Потапенко А.А.* Аддитивная регуляризация тематических моделей // Доклады Академии наук. – 2014. – Т. 456, № 3. – С. 268–271.
5. *Коршунов Антон, Гомзин Андрей.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН, 2012. Т. 23. – С. 215–244.

Development of a Keyphrase Extraction Method Based on a Probabilistic Topic Model

Ekaterina L. Romanadze*
Moscow Aviation Institute, Moscow, Russia
ORCID: https://orcid.org/0000-0003-0351-7235
e-mail: katia_rom.97@mail.ru

Vladimir A. Sudakov**
Moscow Aviation Institute Moscow, Russia,
IPM them. M.V. Keldysh RAS, Moscow, Russia
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

Vadim G. Kislinsky****
Moscow Institute of Physics and Technology, Moscow, Russia,
ORCID: https://orcid.org/0000-0003-2000-583X

The article considers the task of topic modeling. A new method for extracting keywords has been developed based on topic modeling to analyze a collection of documents describing the goods of an online store. A comparative analysis of the basic method for extracting keywords and the proposed method was carried out. Illustrative results are presented that describe the advantages of this approach. The resulting solution can be used to simplify site navigation and search for relevant products.

Keywords: keyword extraction, topic modeling, NLP, LDA, machine learning

For citation:
Romanadze E.L., Sudakov V.A., Kislinsky V.G. Development of a Keyphrase Extraction Method Based on a Probabilistic Topic Model. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2022. Vol. 12, no. 2, pp. 20–33. DOI: https://doi.org/10.17759/mda.2022120202 (In Russ.,abstr. in Engl.).

***Ekaterina L. Romanadze,** Graduate Student, Moscow Aviation Institute (National Research University)(MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0003-0351-7235, e-mail: katia_rom.97@mail.ru

****Vladimir A. Sudakov,** Doctor of Technical Sciences, Professor, Moscow Aviation Institute (National Research University), MAI, Moscow, Russia, Leading Researcher, Federal State Institution "Federal Research Center Institute of Applied Mathematics. M.V. Keldysh of the Russian Academy of Sciences (IPM named after M.V. Keldysh RAS), Moscow, Russia, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

*****Vadim G. Kislinsky,** Researcher, Moscow Institute of Physics and Technology (National Research University) (MFTI), Moscow, Russia, ORCID: https://orcid.org/0000-0003-2000-583X

References

1. Augenstein, I., Das, M., Riedel, S., Vikraman, L. and McCallum, A. (2017) Semeval 2017 task 10: Scienceie – extracting keyphrases and relations from scientific publications. In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3–4, 2017, 546–555. URL: https://doi.org/10.18653/v1/S17–2091.
2. Apishev M.A. Effective implementation of topic modeling algorithms: dis. cand. physics and mathematics: 230401. – M., 2020. – 152 p.
3. Vorontsov K.V. Probabilistic topic modeling: theory, models, algorithms and design BigARTM. URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf.
4. Vorontsov K., Potapenko A.A. Additive regularization of thematic models // Reports of the Academy of Sciences. – 2014. – T. 456, № 3. 268–271 p.
5. Korshunov Anton, Gomzin Andrey. Thematic modeling of natural language texts // Proceedings of the Institute for System Programming of the Russian Academy of Sciences, 2012. T. 23. p. 215–244.

Получена 18.04.2022
Принята в печать 23.06.2022

Received 18.04.2022
Accepted 23.06.2022