Учет неизвестных слов в вероятностной тематической модели*

С. Н. Карпович[I], А. В. Смирнов[II], Н. Н. Тесля[II]

[I] Акционерное общество «Олимп», г. Москва, Россия
[II] Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», г. Санкт-Петербург, Россия

Аннотация. В работе рассмотрены подходы к учету неизвестных слов в языковых моделях алгоритмов обработки естественного языка. Предложен метод учета неизвестных слов в вероятностном тематическом моделировании, который позволяет определить вероятность новизны документа без обновления параметров модели. Тематические модели рассчитывают вероятностную оценку отнесения слова к темам. Матрица вероятностных отношений слово тема, заполнена апостериорными значениями вероятностей слов, введя в модель понятие штрафа за неизвестность или априорную оценку вероятности для неизвестных слов, можем рассчитать вероятностную оценку новизны документа. Разработан программный прототип метода позволяющий рассчитывать вероятность новизны документа. Проведены эксперименты на корпусе текстов SCTM-ru, демонстрирующие возможности метода для классификации коллекций и потоков текстовых документов, содержащих неизвестные слова, отражающие влияние неизвестных слов на тему документов, сравнивающие результаты классификации тематической модели и модели классификатора.

Ключевые слова: вероятностное тематическое моделирование, обработка текста на естественном языке, учет неизвестных слов, новизна текстовых документов.

DOI 10.14357/20718632200410

Введение

Язык представляет собой постоянно развивающееся, динамическое явление. Под действием внешних и внутренних законов развития происходит изменение языка, появляются новые слова, уходят из обихода устаревающие [1].

В задачах анализа текстовых документов существует вероятность встретить ранее неизвестное слово. Алгоритмы обработки текстов на естественном языке обычно не учитывают неизвестные для моделей слова. Тем не менее, вклад неизвестных слов может быть значителен. Известное для модели слово позволяет определить тему документа. Неизвестное слово может иметь отношение к наиболее вероятной теме документа, или может относиться к новой теме. Если неизвестное слово относится к существующей теме, происходит изменение словаря темы. Если неизвестное слово относится к новой теме, происходит появление новой темы.

Для коротких текстовых документов, к которым относятся сообщения в Twitter и заго-

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 20-07-00904) в части исследования алгоритмов обработки текстовых данных и бюджетной темы № 0073-2019-0005 (весь остальной материал).

ловки новостей, вклад каждого слова в определение темы особенно весом. Если текст состоит из 10 слов, в которых 5 слов неизвестны модели, то определить тему документа получится только с 50% надежностью. В таких случаях важно принимать динамические свойства языка, учитывая вклад неизвестных слов.

В рамках статьи решается задача определения новизны документа, по количеству неизвестных слов, встретившихся в новых документах, без обновления параметров модели, без увеличения размера словаря. Существующие подходы онлайн обучения вероятностных тематических моделей [2–6] в ходе своей работы требуют дорогостоящего обновления параметров модели. Для небольших текстовых корпусов периодическое обновление параметров оправдано, так как позволяет повысить точность модели за разумное время. Построение моделей на больших текстовых корпусах, составленных как выборка документов из веб-ресурсов, требует огромных вычислительных и временных ресурсов, поэтому частое обновление параметров таких моделей неоправданно. Для их практического использования в задачах обработки текста на естественном языке требуются алгоритмы, позволяющие с использованием минимального количества вычислительных ресурсов и без повторного обучения определять тематическую принадлежность текстового документа с оценкой точности предсказания и новизны документа.

Для анализа потоковых данных необходимо принять решение, игнорировать появление неизвестных слов, или с каждым неизвестным словом пересчитывать параметры модели. Игнорирование неизвестных слов приводит к систематическим ошибкам. Учет неизвестных слов позволит своевременно получить сигнал о необходимости обновления параметров модели. Сигналом может стать следующее условие: вероятность новизны документа выше вероятностей каких-либо тем. Предложенный метод основан на алгоритме Positive Example Based Learning-TM (PEBL-TM) [7], в котором определяется вероятностная оценка отнесения документа к известному классу и используется понятие штрафа за неизвестность.

Обзор существующих подходов к учету неизвестных слов в задачах анализа текстовых документов

Для решения задач обработки текста на естественном языке необходимо представить текст в понятном для компьютера векторном виде. Существует несколько способов векторизации текста. Унитарный код (One-Hot) – двоичный код фиксированной длины. Длина такого кода определяется количеством кодируемых объектов и каждому объекту соответствует отдельный разряд кода. При кодировании текста вектору слова соответствует код, длина которого равна размеру словаря модели, составленного из уникальных слов текста, а соответствие вектор-слово определяется значением «1» в соответствующем разряде кода. Такие модели называются линейными моделями с разряженным входом [8]. Весь текстовый документ в унитарном коде будет представлен вектором, длина которого равна размеру словаря модели, а в позициях, найденных в документе слов, записано значение «1». Унитарному кодированию текста свойственно наличие ряда существенных недостатков. Во-первых, не учитывается количество повторений отдельно взятых слов в каждом текстовом документе. Во-вторых, размер словаря может быть очень большим для вычислений. В-третьих, унитарное кодирование не учитывает последовательность расположений слов в документе. Улучшением данного метода кодирования является учет количества повторений слов в тексте. Размер вектора слова будет таким же, как при унитарном кодировании, а вектор документа будет содержать счетчик повторений слов в тексте документа.

Альтернативным подходом является алгоритм tf-idf [9] который учитывает не только количество повторений конкретного слова в документе, но и количество повторений слова в корпусе текстовых документов. Чем чаще слово встречается в конкретном документе и реже во всем корпусе, тем оно более значимо для модели. Различные модификации алгоритма tf-idf используются информационными поисковыми системами для ранжирования ответов по текстовой релевантности. Алгоритмы tf-idf

избавлены от недостатка учета значимости слов, но последовательность расположения слов в документе не учитывается.

Вышеперечисленные методы также как вероятностные тематические модели основаны на предположении, что порядок слов в документе не имеет значения, и документ концептуально представляется в виде «мешка слов» (Bag of Words) [10].

Многие современные методы векторного представления слов используют нейросетевые алгоритмы, применяющие для представления слова так называемые плотные вектора признаков, где каждый признак представляется вектором в d-мерном пространстве. Эти векторные представления основаны на моделях Continuous bag of words (CBOW) и skip-gram (SG) [8, 11, 12], учитывающих контекст слова. Главное достоинство плотных представлений – способность к обобщению [8]. К таким моделям относят Word2Vec [11], GloVe [13], BERT [14], FastText [15], Generative Pre-trained Transformer (GPT-3) [16]. Их также называют семантическими моделями векторного пространства языка. Каждое слово представлено в виде вещественного вектора. Вектора близких по смыслу слов в данном представлении похожи. Модели позволяют находить интересные закономерности, выполняя простые алгебраические операции над векторами слов. Так, например, в работе [11] было показано, что вектор слова *«король»* минус вектор слова *«мужчина»* плюс вектор слова *«женщина»* приводит к вектору, наиболее близкому к слову *«королева»*. Модель BERT, обученная на больших текстовых корпусах и настроенная для решения конкретной задачи, предлагает решение 11 задач обработки естественного языка, в частности, предсказание пропущенного слова в предложении [14]. Модель GPT-3 содержит 175 миллиардов параметров, что в 10 раз больше, чем у любой предыдущей не разреженной языковой модели [16]. Важное отличие GPT-3 в том, что для всех задач она применяется без тонкой настройки параметров, при этом показывает отличные результаты при решении многих задач обработки естественного языка, включая задачи по переводу, ответы на вопросы, расшифровки слов, даже генерации новостных статей которые оценщикам-людям трудно отличить от статей, написанных людьми. Модель GPT-3 умеет использовать новые слова в предложении после того, как они были определены только один раз.

Имея большие успехи в определении семантической близости отдельных слов, эти модели сложно использовать в изначальном виде для представления целых документов, поскольку результирующий вектор документа, полученный через сумму векторов встретившихся слов, сложно интерпретировать. Для векторизации предложений и целых документов используют модификации описанных моделей, такие как doc2vec [17], sent2vec [18], Sentense-BERT [19].

Алгоритмы классификации текстовых документов зачастую определяют сходство между документами сравнивая векторные представления этих документов. Классификатор обучается на обучающем множестве документов, затем выполняется проверка точности классификации на тестовом множестве. Если по каким-либо причинам обучающее множество документов не содержит все слова, которые используются в тестовом множестве, то эти слова не повлияют на решение классификатора.

Учитывая сложности векторизации больших корпусов документов, исследуются задачи выбора оптимального размера словаря языковой модели и сокращения размера словаря без существенной потери качества. Традиционный подход к сокращению размера словаря модели опирается на частоту слов, при котором выбирается пороговая частота, отсекающая редко используемые слова. В работе [20] предложен альтернативный метод сокращения размера словаря языковой модели, на основе вариационного отсева. В работе [21] проведены эксперименты по сокращению размера словаря языковой модели, без значительной потери в качестве работы алгоритмов классификации.

Алгоритм учета неизвестных слов в вероятностных тематических моделях

Вероятностная тематическая модель [22, 23] определяет с какой вероятностью каждый документ относится к той или иной теме и из каких слов состоит каждая тема. Словарь тематических

ческой модели состоит из слов, на которых строилась тематическая модель. Учитывая динамические свойства языка, всегда присутствует вероятность встретить новые слова. Традиционные тематические модели, игнорируют появление новых слов. В работах [2-5] предложены алгоритмы онлайн тематического моделирования, которые позволяют обновлять параметры тематической модели по мере поступления новых документов с новыми словами. В работе [6] предложен метод определения тем для новых слов с помощью произведения Адамара. Для эффективного определения темы нового, неизвестного слова необходимо наличие нескольких документов, в которых встретилось это слово. Если неизвестное слово встретилось лишь в одном документе, то точно определить его тему невозможно.

При разработке алгоритма учета неизвестных слов в вероятностных тематических моделях использовался разработанный авторами ранее алгоритм классификации текстовых документов Positive Example Based Learning-TM (PEBL-TM) [7]. Данный алгоритм использует вероятностную тематическую модель, построенную на множестве документов, представленных экземплярами одного класса. Подход позволяет использовать тематическую модель как одно-классовый классификатор. Модель обучается только на известных примерах документов, и затем она используется для отбора похожих из коллекции или даже потока текстовых документов.

Для учета неизвестных слов используем такой же подход, будем рассчитывать вероятности известных слов и назначать штраф за неизвестные слова. Размер штрафа — это априорная оценка вероятности отнесения неизвестного слова к новой теме. Зависит от степени уверенности в данных для обучения модели. Если точность и полнота данных для обучения высока, то высоким должен быть и штраф за неизвестность. Если существуют опасения на счет точности или полноты данных для обучения, то следует экспериментальным путем подбирать подходящее значение штрафа. Значение по умолчанию 1, неизвестные слова будут оказывать максимальное влияние на новизну документов.

Определим математическое описание вероятностной тематической модели. Пусть D – коллекция текстовых документов, W - словарь терминов. Каждый документ $d \in D$ представляет собой последовательность терминов $(w_1, \ldots, w_{n_d})$ из словаря W.

Исходя из вероятностного уточнения гипотезы «мешка слов», множество $\Omega = D \times W \times T$ является конечным вероятностным пространством с неизвестной функцией вероятности $p(d,w,t)$. Появление термов w в документе d по теме t зависит от темы, но не зависит от документа, и описывается общим для всех документов распределением $p(w|d,t) = p(w|t)$, которое называется гипотезой условной независимости. С учетом гипотезы условной независимости по формуле полной вероятности распределение термов в документах $p(w|d)$ описывается вероятностным распределением (EM) слов в темах $\varphi_{wt} = p(w|t)$ с учетом весов $\theta_{td} = p(t|d)$:

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

Для вычисления φ_{wt} и θ_{td} используется [24], являющийся итерационным алгоритмом, в котором на E шаге вычисляется ожидаемое значение функции правдоподобия, а на M шаге вычисляется оценка максимального правдоподобия. Модификация EM-алгоритма для вероятностного тематического моделирования рассмотрена в работе [25].

Вероятностная модель появления пары «документ-слово» может быть записана тремя эквивалентными способами:

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

где: p(t) – неизвестное априорное распределение тем в коллекции;

p(d) – априорное распределение на коллекции документов, эмпирическая оценка $p(d) = n_d / n$, где $n = \sum_d n_d$ – суммарная длина всех документов, а n_d – длина документа в словах;

p(w) – априорное распределение на множестве слов, эмпирическая оценка $p(w) = n_w / n$, где n_w – число вхождений слова w во все документы.

Алгоритм учета неизвестных слов в модели описан в листинге 1. На вход алгоритма подается набор документов, на которых строится вероятностная тематическая модель. В разработанном методе вероятностная тематическая модель строится на основании тематической привязки документов к темам, заданной в обучающих данных [26]. Такой подход удобнее для проведения экспериментов с оценкой качества работы алгоритма. Но предлагаемый алгоритм может быть адаптирован для работы с неразмеченными данными, то есть вероятностная тематическая модель может быть построена без информации о принадлежности документов к темам [22, 23, 27], в этом случае модель выполняет разделение множества документов для обучения на заданное количество кластеров, и определяет вероятности отнесения новых документов к одному из уже имеющихся в модели кластеру или задает высокую оценку вероятности новизны документа.

Листинг 1: Алгоритм учета неизвестных слов в вероятностной тематической модели

1. Вход: коллекция документов D, штраф за неизвестность *P(new)=1*, документы, класс которых необходимо определить d_{new}.
2. Выход: вероятностная оценка отнесения документа к теме и оценка новизны.
3. Построить вероятностную тематическую модель на коллекции документов.
4. Для всех $w \in d_{new}$.
5. Добавляем новую тему к модели *T= T +1*.
6. Если слова нет в модели, $p(w|t) = P(new) = 1, P(topic) = 0$, где *P(topic)* – вероятность отнесения слова *w* к известной теме *t*.
7. Если слово есть в модели, считаем $p(w|t) = P(topic) = n_{dwt(topic)} / n_{dw}$, $P(new) = n_{dwt(new)} / n_{dw}$, где $n_{dwt(topic)}$ – число вхождений слова *w* связанного с известной темой *t* в документе *d*, $n_{dwt(new)}$ – число вхождений нового слова w связанного с неизвестной темой *t* в документе *d*.
8. Считаем вероятность отнесения документа к классу $p(d|t) = \sum_{w \in d} P(topic)$.

Вероятностная оценка отнесения нового документа к темам будет содержать оценку новизны документа, которую можно условно воспринимать как новую тему. По оценке вероятности новизны документа, можно судить об уверенности в отнесении документа к известным темам и можно принимать решение о необходимости полного переобучения вероятностной тематической модели с использованием всех новых документов.

Эксперимент с корпусом текстов SCTM-ru

На языке разработки Python с использованием библиотеки машинного обучения scikit-learn [28] создан прототип программы, реализующий предложенный метод. В качестве экспериментальных данных использовался корпус SCTM-ru [29], созданный специально для тестирования задач тематического моделирования. Источником документов для данного корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 12 тыс. документов, 320 авторов, почти 12000 уникальных категорий. События, описанные в документах, распределены с ноября 2005 года по январь 2017 года. В корпусе SCTM-ru насчитывается 2,5 млн словоупотреблений, состоящих только из букв русского алфавита. Словарный состав корпуса составляет 262 тыс. уникальных словоформ. Каждая новость содержит указанные автором темы. Обычно перед автором новости не стоит задача перечислить все возможные темы, к которым новость может иметь отношение. Тем не менее указанные темы дают весомые основания полагать, что новость сильно связана с ними. Для проведения экспериментов построим несколько вероятностных тематических моделей на различных наборах данных. Используем классические метрики алгоритмов классификации (точность, полнота, F-мера) для оценки результатов экспериментов, а также будем считать количество документов с высокой оценкой новизны. Оценка новизны документа считается высокой в том случае, когда она выше максимальной вероятностной оценки принадлежности документа к какой-либо из известных тем. В экспериментах, результаты которых представлены

далее, будет продемонстрирована работа алгоритма на искусственных данных, проведено сравнение предлагаемого метода с моделью классификации и определена зависимость штрафа модели от показателей данных для обучения. Поскольку в работе не ставилась задача онлайн-переобучения модели, выбор аналогов предложенного метода является сложным. Для части экспериментов в качестве аналога, с которым возможно сравнение, был выбран метод логистической регрессии, как наиболее близкий по смыслу и получаемым результатам.

Работа алгоритма на искусственных данных. Корпус текстовых документов представлен тремя выдуманными документами: «почему спорт футбол бег», «почему бег полезен здоровье», «почему автомобили дороги». Слово «почему» встречается в каждом документе. Слово «бег» встречается в двух из трех документов. Остальные слова встречаются по одному разу в документе. Каждый документ относится к одной из трех тем: «спорт», «здоровье», «автомобили». Используя документы из искусственного корпуса и знание о их тематической принадлежности построим вероятностную тематическую модель, как описано в работе [26], и создадим модель классификатора используя логистическую регрессию из библиотеки scikit-learn. Сравнение результатов классификации для известных и неизвестных слов представлено на Рис. 1.

Для слов, которые присутствовали в обучающих данных, модели предсказывают наи-

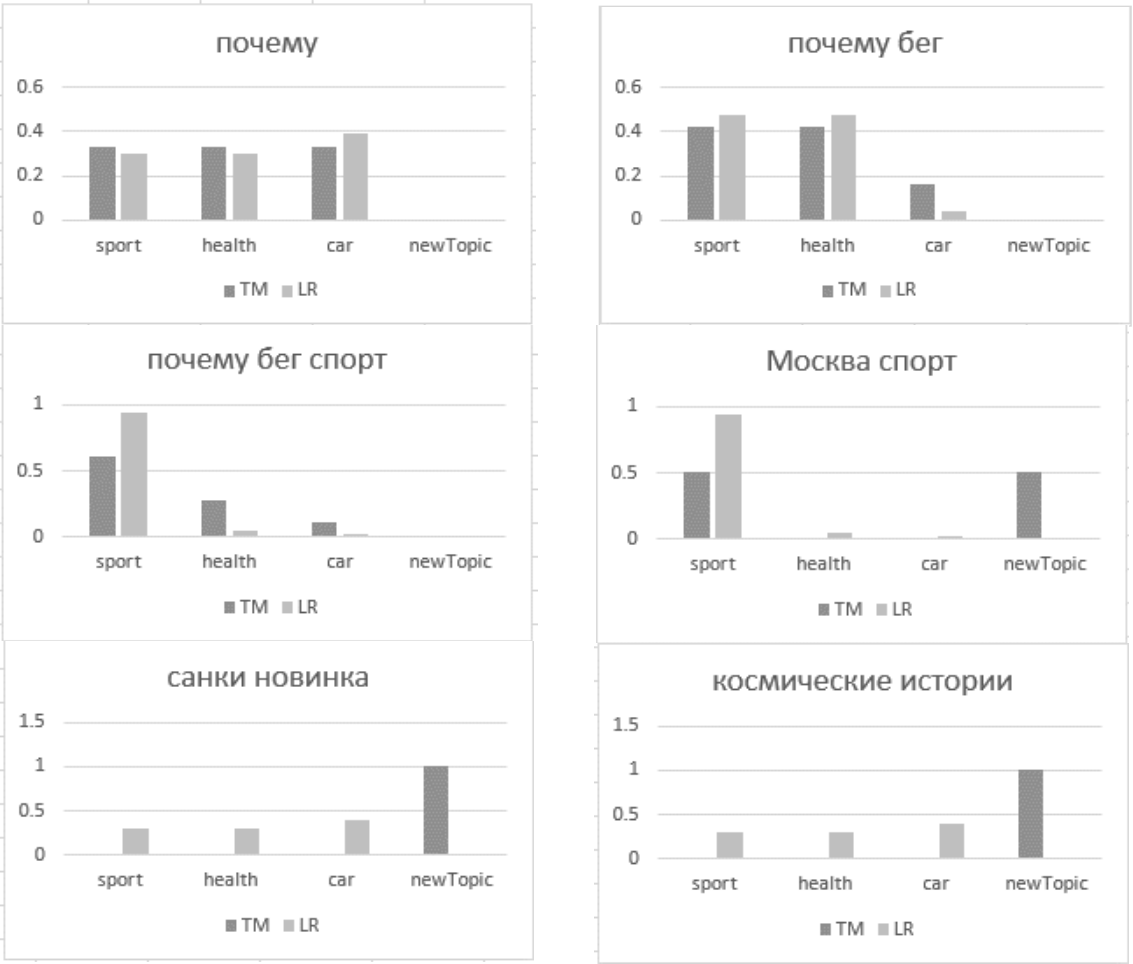
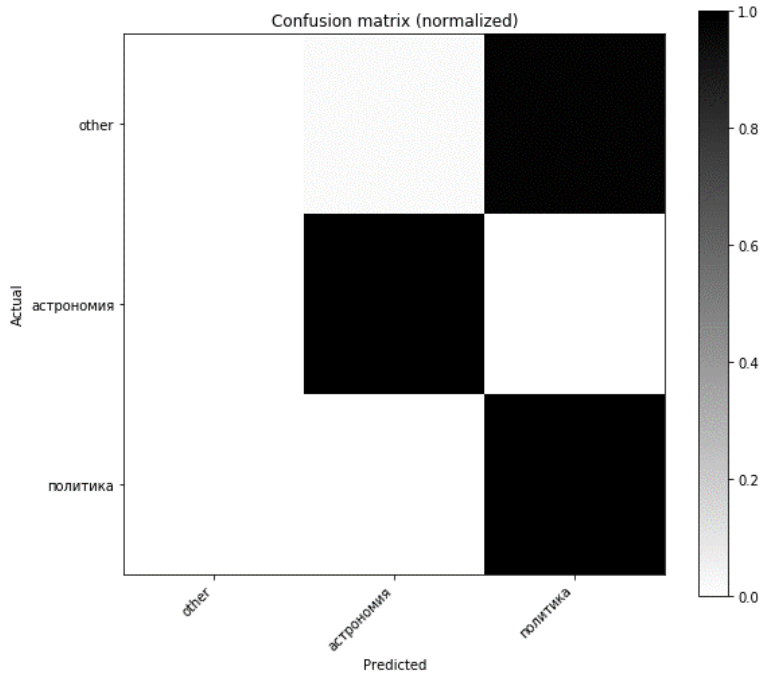


Рис. 1. Сравнение результатов классификации для искусственных данных между тематической моделью и классификатором

более вероятную тему. С появлением в тестовых данных неизвестных слов точность модели классификатора снижается. Фраза «Москва спорт» состоит из одного известного и одного неизвестного слова, вследствие чего вероятность новизны документа по оценке тематической модели составляет 0.5, и с такой же вероятностью тематическая модель относит эту фразу к теме «спорт». В реальных условиях подобное поведение может приводить к ошибкам. Если тестовые данные состоят только из неизвестных слов, то тематическая модель дает высокую оценку новизны документа, а модель классификатора ошибочно относит документ к известным темам. Для более полной демонстрации подобной ошибки, на Рис. 1 также представлены результаты классификации для других слов с использованием той же модели, на основе тестовых фраз «санки новинка», «космические истории».

Проведем сравнительный эксперимент между вероятностной тематической моделью и моделью логистической регрессии, в качестве классификатора, на реальных данных корпуса текстов SCTM-ru. В данном эксперименте производится оценка модели по метрикам качества алгоритмов классификации. Обучаем модели текстами документов, относящимися к темам «политика» и «астрономия». Темы для обучения были выбраны случайным образом. Для подготовки данных из корпуса были отобраны все документы, относящиеся к выбранным темам, и разделены на два множества для тестирования и обучения. Итоговое обучающее множество содержит 105 документов. Для шума, к тестовым данным были добавлены документы, относящиеся к темам «спорт», «Россия», «Европа», «футбол», эти документы в тестовых данных отнесены к общей теме «other». Тестовое множество содержит 979 документов, из которых 384 документа относятся к темам политика или астрономия. Задача моделей найти в тестовом множестве документов те, что относятся к темам политика и астрономия. Матрица ошибок для модели классификатора представлена на Рис. 2. На нем видно, что модель классификатора ошибочно отнесла документы с высокой новизной к теме «политика».

Матрица ошибок для тематической модели представлена на Рис. 3. Тематическая модель корректно определила новизну для большей части документов. При этом тематическая модель допустила ошибку, неверно оценив часть документов про «астрономию» как документы



Рису. 2. Матрица ошибок для модели классификатора

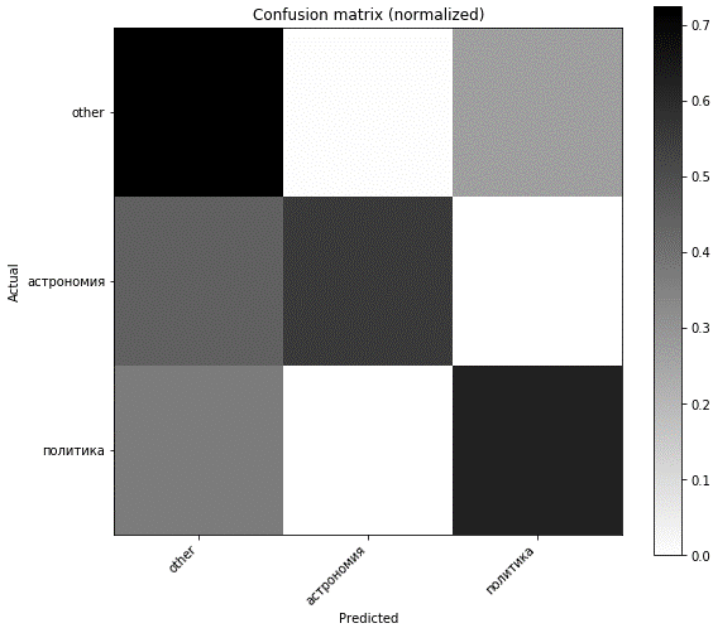


Рис. 3. Матрица ошибок тематической модели

с высокой новизной. Такая ошибка могла произойти по причине малого количества обучающих примеров по теме «астрономия». Часть неизвестных слов, встретившихся модели, на самом деле имеет отношение к теме «астрономия». Если обучить модель с использованием всех слов, представляющих интересующую тему, то такой ошибки можно избежать.

Сравнение метрик оценки качества моделей представлено в Табл. 1. Итоговые оценки метрик точности, полноты и средневзвешенной F-меры для результатов предсказания у модели классификатора низкие. Это связано с тем, что модель классификации не учитывает влияние неизвестных слов на определение темы, не определяет новизну документов. Оценки тематической модели, позволяют утверждать, что предложенный метод эффективен для определения новизны документов и может быть использован в практических задачах, когда необходимо отобрать документы определенных тем из коллекции документов, относящихся к различным темам.

Далее был проведен эксперимент на данных корпуса текстов SCTM-ru, демонстрирующий влияние данных, используемых для обучения модели на оптимальный размер штрафа. Вероятностная тематическая модель была построена на основе 10 тыс. документов из данного корпуса. Словарь тематической модели содержит 86 182 слова. Модель состоит из 4 057 тем. Для тестирования используются 100 документов корпуса, которые не участвовали в обучении. В тестовых документах встретилось 881 неизвестное слово, максимально в одном документе встретилось 107 неизвестных слов, среднее количество неизвестных слов на документ 8.81, медиана 6. При штрафе за неизвестность равным 1 (значение по умолчанию), 98 документов получили высокую оценку новизны. Высокая оценка означает, что вероятность новизны была выше вероятности отнесения к какой-либо теме.

Табл. 1. Сравнение метрик оценки качества моделей

Модель	Точность	Полнота	Средневзвешенная F-мера
Тематическая модель	0.69	0.69	0.69
Модель Классификатора	0.15	0.39	0.22

Такой результат может казаться странным, в документе встретили одно неизвестное слово и оценили документ с максимальной новизной. Как было отмечено выше, размер штрафа следует выбирать, опираясь на информацию о корпусе для обучения. В данном случае количество тем модели очень высоко, даже с учетом большого размера словаря, велик шанс получить низкие оценки тематической вероятности для отдельных слов. Построенная модель позволяет выяснить максимальные значения вероятностей для слов наиболее характерных для каждой темы. Среднее значение вероятности от полученных максимальных значений для построенной модели составляет 0.31, медиана 0.2. Разница между заданным значением штрафа и средним значением максимальных вероятностей слов велика, что приводит к смещению вероятностной оценки в сторону новизны документа. При значении штрафа равному среднему от максимальных значений вероятностей слов (0.31), 90 документов получили высокую оценку новизны. При значении штрафа равному медианному от максимальных значений вероятностей слов (0.2), 71 документ из 100 получил высокую оценку новизны. Результаты эксперимента представлены на Рис. 4. Таким образом, оптимальный размер штрафа за неизвестное слово следует выбирать, опираясь на понимание данных задачи.

Проведем эксперимент на реальных данных корпуса текстов SCTM-ru, демонстрирующий применимость метода для отбора документов на интересующие темы из набора текстовых данных, включающих документы различных тем. Целью эксперимента является выяснение того, какое количество документов будет классифицировано верно, при условии обучения вероятностной тематической модели на экземплярах документов, относящихся к нескольким заданным темам, при том, что тестовые данные будут включать документы не имеющие отношения к заданным темам. Текстовые данные для обучения состоят из двух случайно выбранных тем, 595 документов, 6 823 слов. Данные для тестирования состоят из 7 510 документов, из которых 87 определены авторами к одной из отобранных тем. Следует учитывать, что перед авторами документов не стояла задача указывать тему, поэтому в тестовом наборе может быть больше документов относящихся к интересующим нас темам, в которых авторы не проставили тему. Среднее значение вероятности от полученных максимальных значений для построенной модели составляет 0.99, медиана 0.99. Средние значения высоки, поэтому размер штрафа модели может быть высоким, используем для штрафа значение по умолчанию. Вероятность новизны документа указывается темой «other». Результат эксперимента представлен в матрице ошибок на Рис. 5. По матрице ошибок видно, что документы, относящиеся к теме «политика», были корректно классифицированы. При этом часть документов из темы «Россия» были ошибочно оценены как документы с высокой новизной, такое могло произойти в случае недостатка обучающих примеров для темы. Для обеспечения высокой

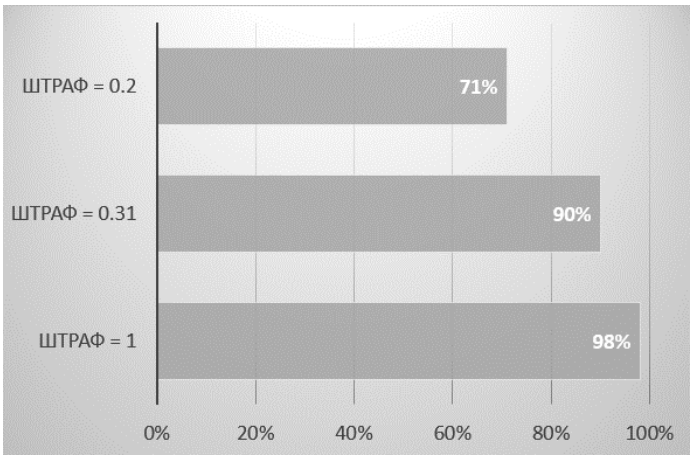


Рис. 4. Доля документов с высокой новизной, в зависимости от штрафа за неизвестное слово

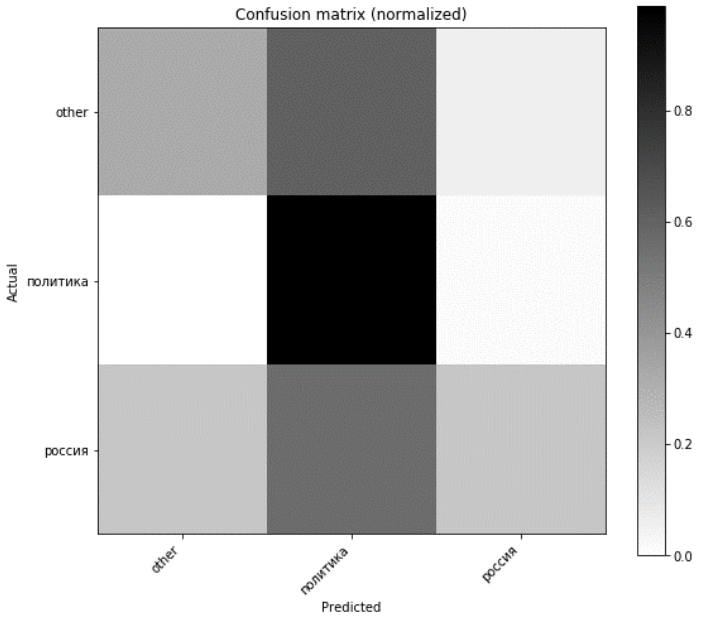


Рис. 5. Матрица ошибок для тематической модели

точности предсказания следует использовать репрезентативный набор данных. Проведенный эксперимент подтверждает возможность использования метода для отбора документов интересующих тем из набора или потока текстовых документов.

Проведем эксперимент, демонстрирующий влияние размера штрафа на определение новизны документа. Будем считать количество документов с высокой оценкой новизны, снижая размер штрафа, на одном и том же наборе тестовых документов. В эксперименте используются имена существительные в нормальной словоформе из корпуса SCTM-ru, используемого в качестве источника данных. Данные для обучения включают в себя 732 документа, относящиеся к темам, «Россия» и «политика», словарный состав модели насчитывает 8 392 слова, среднее значение максимальных вероятностей наиболее характерных для каждой темы слов составляет 0.99. Для тестирования используются 338 документов, которые не вошли в обучающие данные. В ходе эксперимента итеративно строятся вероятностные тематические модели, снижается размер штрафа, и подсчитывается количество документов с максимальной новизной. Результат представлен на Рис. 6. При размере штрафа более 0.6 все документы имеют высокую оценку новизны. Значительное снижение количества документов с высокой новизной наблюдается при штрафе менее 0.2. Таким образом, можно предположить, что размер штрафа за неизвестное слово в вероятностной тематической модели следует задавать с учетом не только данных для обучения, но и стоящей перед исследователем задачи. Если требуется в неизвестном наборе текстовых документов найти те, что имеют отдаленное отношение к интересующим темам, когда важна полнота, то размер штрафа должен быть установлен ниже. Если же требуется найти, только те документы, которые имеют высокую вероятность отнесения к интересующим темам, когда важна точность, то размер штрафа должен быть выше. На графике приведена средневзвешенная F-мера, качество классификатора снижается по мере снижения размера штрафа.

Заключение

В результате проделанной работы был разработан метод для определения новизны документов, позволяющий учитывать неизвестные слова в вероятностной тематической модели. Предложенный метод может быть успешно использован для определения классов как статического корпуса документов, так и для потока текстовых документов.

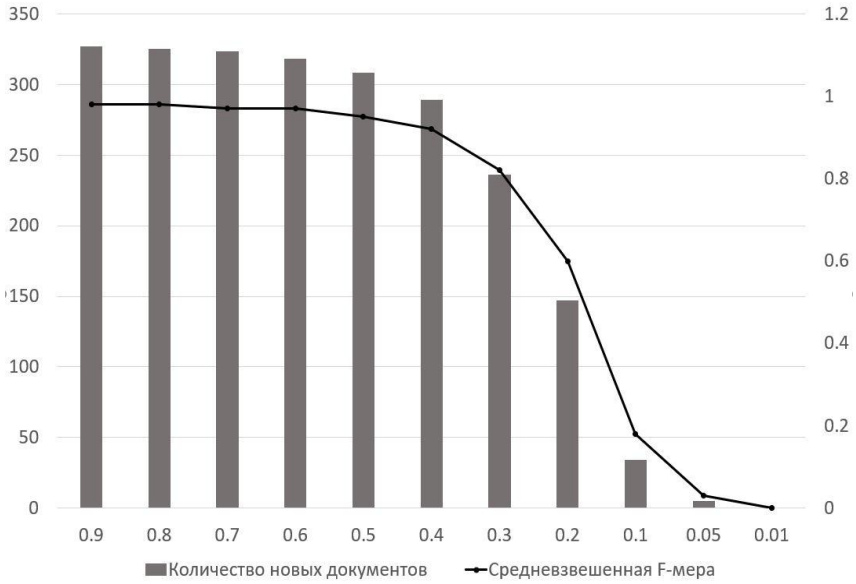


Рис. 6. Количество новых документов и средневзвешенная F-мера в зависимости от размера штрафа за неизвестность

Вероятностная тематическая модель, применяемая в методе, используется для вычисления вероятности новизны документа и вероятностных оценок отнесения документа к какой-либо известной теме. Информация о новизне документа позволит повысить точность существующих методов классификации. В практических задачах велика вероятность встретить новые слова при обработке данных, их учет и получение вероятностной оценки новизны документа, позволяет создавать системы с обратной связью, своевременно реагирующие на изменение словарного состава обрабатываемых данных. Метод также позволяет отбирать документы, относящиеся к интересующим темам, при условии обучения вероятностной тематической модели только на экземплярах, имеющих отношение к интересующим темам, а это позволяет использовать более компактные вероятностные тематические модели, содержащие только словарь из интересующих тем.

Опираясь на свойство мультимодальности вероятностной тематической модели [30, 31], предложенный подход может быть использован для учета нетекстовых свойств документов, что позволит расширить круг практических задач применения алгоритма.

Программный прототип метода, а также программный код и результаты экспериментов доступны по ссылке https://github.com/cimsweb/PEBL-ML-TM

Литература

1. Крылова М. Н. Язык как динамическая система //Модели, системы, сети в экономике, технике, природе и обществе. – 2014. – №. 1 (9).
2. Wang C., Blei D., Heckerman D. Continuous time dynamic topic models. preprint arXiv:1206.3298. – 2012.
3. Hoffman M., Bach F. R., Blei D. M. Online learning for latent dirichlet allocation. Advances in neural information processing systems. – 2010. – С. 856-864.
4. Zhai K., Boyd-Graber J. L. Online Latent Dirichlet Allocation with Infinite Vocabulary. ICML (1). – 2013. – Т. 28. – С. 561-569.
5. Lau J. H., Collier N., Baldwin T. On-line Trend Analysis with Topic Models: \# twitter Trends Detection Topic Model Online. COLING. – 2012. – С. 1519-1534.
6. Карпович С.Н. Тематическая модель с бесконечным словарем // Информационно-управляющие системы. 2016. No6С. 43-49. doi:10.15217/issn1684-8853.2016.6.43(ВАК)
7. Карпович С. Н., Смирнов А. В., Тесля Н. Н. Одноклассовая классификация текстовых документов с использованием вероятностного тематического моделирования //Искусственный интеллект и принятие решений. – 2018. – №. 3. – С. 69-77.
8. Гольдберг Й. Нейросетевые методы в обработке естественного языка. – ДМК-Пресс, 2019.

9. Berger A., Lafferty J. Information retrieval as statistical translation //ACM SIGIR Forum. – New York, NY, USA: ACM, 2017. – Т. 51. – №. 2. – С. 219-226.
10. Wallach H. M. Topic modeling: beyond bag-of-words //Proceedings of the 23rd international conference on Machine learning. – 2006. – С. 977-984.
11. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.
12. Rong X. Word2vec parameter learning explained //arXiv preprint arXiv:1411.2738. – 2014.
13. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
14. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
15. Joulin A. et al. Fasttext. zip: Compressing text classification models //arXiv preprint arXiv:1612.03651. – 2016.
16. Brown T. B. et al. Language models are few-shot learners //arXiv preprint arXiv:2005.14165. – 2020.
17. Lau J. H., Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation //arXiv preprint arXiv:1607.05368. – 2016.
18. Le Q., Mikolov T. Distributed representations of sentences and documents //International conference on machine learning. – 2014. – С. 1188-1196.
19. Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks //arXiv preprint arXiv:1908.10084. – 2019.
20. Chen W. et al. How large a vocabulary does text classification need? a variational approach to vocabulary selection //arXiv preprint arXiv:1902.10339. – 2019.
21. Chirkova N., Lobacheva E., Vetrov D. Bayesian compression for natural language processing //arXiv preprint arXiv:1810.10927. – 2018.
22. Hoffman T. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. — 1999. – С. 50-57.
23. Blei D.M., Ng A.Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. — 2003. – Т. 3. – №. Jan. – С. 993-1022.
24. Moon T. K. The expectation-maximization algorithm //IEEE Signal processing magazine. – 1996. – Т. 13. – №. 6. – С. 47-60.
25. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования //Машинное обучение и анализ данных. – 2013. – Т. 1. – №. 6. – С. 657-686.
26. Карпович С.Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI//Труды СПИИРАН. – СПб.,2016. –Т. 4. –No. 47. –С. 92-104(ВАК, Scopus)
27. Vorontsov K., Potapenko A. Additive regularization of topic models //Machine Learning. – 2015. – Т. 101. – №. 1-3. – С. 303-323.
28. Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of machine learning research. – 2011. – Т. 12. – №. Oct. – С. 2825-2830.
29. Карпович С.Н. Русскоязычный корпус текстов SCTM-ru для построения тематических моделей // Труды СПИИРАН. –СПб., 2015.–No39. С. 123-142. УДК 004.912(ВАК)
30. Ianina A., Vorontsov K. Regularized multimodal hierarchical topic model for document-by-document exploratory search //2019 25th Conference of Open Innovations Association (FRUCT). – IEEE, 2019. – С. 131-138.
31. Vorontsov K. et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections //Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. – 2015. – С. 29-37.

Карпович Сергей Николаевич. Акционерное общество «Олимп» 121099, г. Москва, ул. Новый Арбат, д.36. Руководитель группы развития поиска портала Правительства Москвы, кандидат технических наук. Количество печатных работ 13. Область научных интересов: тематическое моделирование, обработка текстов на естественном языке, информационный поиск, машинное обучение. E-mail: cims@yandex.ru

Смирнов Александр Викторович. Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», г. Санкт-Петербург, Россия. Главный научный сотрудник, доктор технических наук, профессор. Количество печатных работ: 450. Область научных интересов: интеллектуальное управление конфигурациями виртуальных и сетевых организаций, логистика знаний, поддержка принятия решений. e-mail: smir@iias.spb.su

Тесля Николай Николаевич. Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук», г. Санкт-Петербург, Россия. Старший научный сотрудник, кандидат технических наук. Количество печатных работ: 61. Область научных интересов: онтологии, слияние онтологий, управление знаниями, обработка текстов на естественном языке. e-mail: teslya@iias.spb.su

Penalty for Unknown Words in Topic Model

S. N. Karpovich[I], A. V. Smirnov[II], N. N. Teslya[II]

[I] JSC "Olympus", Moscow, Russia
[II] St. Petersburg Federal Research Center of the Russian Academy of Sciences

Abstract. The paper considers approaches to accounting for unknown words in language models used in natural language processing algorithms. A method is proposed for accounting for unknown words in probabilistic topic modeling, which allows to determine the probability of a document's novelty in relation to existing topics. Topic models calculate the probabilistic assessment of classifying a word to some topic. The word-topic probabilistic relationship matrix in such a model is filled with posterior values of word probabilities. To calculate the probabilistic assessment of a document's novelty, this paper proposes to introduce the concept of a penalty for obscurity or an a priori probability estimate for unknown words into the model. A software prototype has been developed that allows calculating the probability of a document's novelty taking into account various penalty values. Experiments were conducted on the SCTM-ru text corpus, demonstrating the capabilities of the method for classifying collections and flows of text documents containing unknown words that reflect their influence on the topic of documents. During the experiments, the classification results were also compared using a thematic model and a classifier model based on logistic regression.
Keywords: topic modeling, natural language processing, penalty unknown words

DOI 10.14357/20718632200410

References

1. Krylova M. N. Jazyk kak dinamicheskaja sistema [Language as a dynamic system] // Modeli, sistemy, seti v ehkonomike, tekhnike, prirode i obshhestve [Models, systems, networks in economics, engineering, nature and society]. – 2014. – №. 1 (9). 1.
2. Wang C., Blei D., Heckerman D. Continuous time dynamic topic models. preprint arXiv:1206.3298. – 2012.
3. Hoffman M., Bach F. R., Blei D. M. Online learning for latent dirichlet allocation. Advances in neural information processing systems. – 2010. – С. 856-864.
4. Zhai K., Boyd-Graber J. L. Online Latent Dirichlet Allocation with Infinite Vocabulary. ICML (1). – 2013. – Т. 28. – С. 561-569.
5. Lau J. H., Collier N., Baldwin T. On-line Trend Analysis with Topic Models: \# twitter Trends Detection Topic Model Online. COLING. – 2012. – С. 1519-1534.
6. Karpovich S.N. Tematicheskaja model s beskonechnym slovarem [Topic Model with an Infinite Vocabulary] // Informazionno-Upravlyaushie Sistemy [Information & Control Systems]. 2016. No6S. 43-49. doi:10.15217/issn1684-8853.2016.6.43 (VAK).
7. Karpovich S. N., Smirnov A. V., Teslja N. N. Odnoklassovaja klassifikacija tekstovykh dokumentov s ispol'zovaniem verojatnostnogo tematicheskogo modelirovanija [Positive Example Based Learning-TM] // Iskusstvennyjj intellekt i prinjatie reshenijj [Artificial Intelligence and Decision Making]. – 2018. – №. 3. – S. 69-77.
8. Goldberg Y., Hirst G. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers(2017) //9781627052986 (zitiert auf Seite 69).
9. Berger A., Lafferty J. Information retrieval as statistical translation //ACM SIGIR Forum. – New York, NY, USA: ACM, 2017. – Т. 51. – №. 2. – С. 219-226.
10. Wallach H. M. Topic modeling: beyond bag-of-words //Proceedings of the 23rd international conference on Machine learning. – 2006. – С. 977-984.
11. Mikolov T. et al. Efficient estimation of word representations in vector space //arXiv preprint arXiv:1301.3781. – 2013.
12. Rong X. Word2vec parameter learning explained //arXiv preprint arXiv:1411.2738. – 2014.
13. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
14. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
15. Joulin A. et al. Fasttext. zip: Compressing text classification models //arXiv preprint arXiv:1612.03651. – 2016.
16. Brown T. B. et al. Language models are few-shot learners //arXiv preprint arXiv:2005.14165. – 2020.
17. Lau J. H., Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation //arXiv preprint arXiv:1607.05368. – 2016.
18. Le Q., Mikolov T. Distributed representations of sentences and documents //International conference on machine learning. – 2014. – С. 1188-1196.

19. Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks //arXiv preprint arXiv:1908.10084. – 2019.
20. Chen W. et al. How large a vocabulary does text classification need? a variational approach to vocabulary selection //arXiv preprint arXiv:1902.10339. – 2019.
21. Chirkova N., Lobacheva E., Vetrov D. Bayesian compression for natural language processing //arXiv preprint arXiv:1810.10927. – 2018.
22. Hoffman T. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. — 1999. – С. 50-57.
23. Blei D.M., Ng A.Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. — 2003. – Т. 3. – №. Jan. – С. 993-1022.
24. Moon T. K. The expectation-maximization algorithm //IEEE Signal processing magazine. – 1996. – Т. 13. – №. 6. – С. 47-60.
25. Vorontsov K.V., Potapenko A.A. Modifikacii EM-algoritma dlja verojatnostnogo tematicheskogo modelirovanija [EM-like algorithms for probabilistic topic modeling] // Mashinnoe obuchenie i analiz dannyh [Machine Learning and Data Mining]. – 2013. – Т. 1. – №. 6. – С. 657-686.
26. Karpovich S.N. Mnogoznachnaja klassifikacija tekstovykh dokumentov s ispol'zovaniem verojatnostnogo tematicheskogo modelirovanija ml-PLSI [Multi-Label Classification of Text Documents using Probabilistic Topic Model ml-PLSI.] // Trudy SPIIRAN [SPIIRAS Proceedings]. – SPb.,2016. –Т. 4. –No. 47. –S. 92-104 (VAK, Scopus)
27. Vorontsov K., Potapenko A. Additive regularization of topic models //Machine Learning. – 2015. – Т. 101. – №. 1-3. – С. 303-323.
28. Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of machine learning research. – 2011. – Т. 12. – №. Oct. – С. 2825-2830.
29. Karpovich S.N. Russkojazychnyjj korpus tekstov SCTM-ru dlja postroenija tematicheskikh modelejj [The Russian language text corpus for testing algorithms of topic model] // Trudy SPIIRAN [SPIIRAS Proceedings]. –SPb., 2015.– No39. C. 123-142. UDK 004.912(VAK)
30. Ianina A., Vorontsov K. Regularized multimodal hierarchical topic model for document-by-document exploratory search //2019 25th Conference of Open Innovations Association (FRUCT). – IEEE, 2019. – С. 131-138.
31. Vorontsov K. et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections //Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. – 2015. – С. 29-37.

Karpovich S. N. PhD. JSC "Olympus" st. Novy Arbat, d.36, Moscow, 121099, Russia, e-mail: cims@yandex.ru

Smirnov A. V. Dr.Habil. Professor. St. Petersburg Federal Research Center of the Russian Academy of Sciences, 39, 14 Line, St.Petersburg, 199178, Russia. e-mail: smir@iias.spb.su

Teslya N. N. PhD. St. Petersburg Federal Research Center of the Russian Academy of Sciences, 39, 14 Line, St.Petersburg, 199178, Russia. e-mail: teslya@iias.spb.su