Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и математическое моделирование в экономике

МЕТОДЫ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ КОРОТКИХ ТЕКСТОВЫХ ОБЪЯВЛЕНИЙ

(магистерская диссертация)

Студент:
Козлинский Евгений Михайлович

_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)***:**
_____
_____
(подпись консультанта)

Москва 2021

Аннотация

Работа посвящена разработке методов работы с многоуровневом каталогом объявлений. Ставятся задачи по восстановлению меток классов. Строится многоклассовый классификатор на основе мультимодальной тематической модели ARTM. Предложена модификация классического алгоритма TopMine, позволяющая работать с шумными коллекциями.
Ключевые слова: тематическое моделирование; ARTM; рубрикатор; выделение терминов; TopMine; тематический линейный классификатор

Содержание

1 Введение

В современном мире для все большего количества областей деятельности человека требуются агрегаторы информации. Один из признаков удобного агрегатора – интуитивно понятная структура хранения информации.

Существует сайт объявлений об услугах. На этом сайте есть два типа пользователей: исполнители – специалисты, котрорые размещают объявления об оказываемых услугах и заказчики – люди, которым нужно оказание какой-то услуги. Главная задача сайта сделать так, чтобы заказчики и исполнители нашли друг друга. Для удобства обоих сторон на сайте создан каталог услуг, который имеет несколько уровней: профессия, специализация, услуга. Исполнитель, размещая объявление об услуге, может выбрать в какой категории размещать объявление.

Не во всех объявлениях проставлена правильная категория. Исполнители не размечают категории объявлений об услугах по многим причинам, среди них такие: в каталоге нет подходящей категории; исполнитель не нашел подходящую категорию; исполнитель целенаправленно не отметил категорию или указал неправильную категорию. Такие объявления не находятся заказчиками или вводят их в заблуждение. Чтобы улучшить пользовательский опыт заказчиков требуется устранить объявления с не проставленной категорией или объявления с неправильной категорией.

В работе ставится задача об определении категории объявления по его тексту. Задача осложняется большим, более 4 тысяч, числом категорий. Кроме того, нет четкой структуры описания объявления, каждый исполнитель пишет то, что считает нужным, иногда одну фразу. Это затрудняет работу текстовых моделей, предназначенных для работы с полноценными документами или предложениями. Многие исполнители прибегают к оптимизации текстов объявлений с целью улучшить возможность поиска их объявлений поисковыми системами или делают множество похожих объявлений. Это затрудняет использование статистических методов таких как tf-idf, а так же методов, основанных на сравнении близости векторных представлений фраз. В различных категориях может находиться различное по порядкам число объявлений, например: много объявлений в категории «генеральная уборка квартиры» и мало объявлений в категории «ремонт рояля». Это влечет высокую несбалансированностьклассов.

Задачу определения категорий по тексту можно разделить на три последовательные части: предобработка документов, включающая в себя формирование словаря и представления документов с помощью вве-

денного словаря; формирование векторных представлений документов; классификация. Для формирования словаря будут использованы различные подходы выделения фраз и предложена модификация алгоритма TopMine [1]. Для формирования векторных представлений будет использована тематиеская мультимодальная модель ARTM [2]. Для классификации будут использованы линейные методы классификации логистическая регрессия [3] и линейный SVC [4]. Необходимость модификации классического алгоритма TopMine вызвана проблемой трудности применения статистических методов к исходной коллекции по причине многократных повторений одинаковых текстов и фраз некоторыми авторами объявлений.

Существует ряд работ по бесконтрольному нахождению фраз [5]. В них используются подходы языкового моделирования, ранжирования на основе графов и кластеризации.

Есть ряд работ посвященных нахождению тематических терминов во время построения тематической модели или во время постобработки. В TNG [6] используется вероятностная модель, в которой для каждого слова выбирается тема, затем выбирается статус униграммы или биграммы, затем выбирается слово из темоспецифичного распределения униграмм и биграмм. То есть, бинарные переменные и матрица вероятностей переходов образуют дополнительную размерность по темам. В KERT [7] использует бесконтрольные нахождение фраз с помощью частотных методов для каждой темы с ранжированием полученных фраз так, чтобы наверху были наиболее репрезентативные фразы для темы.

В работе [1] был предложен метод нахождения многословных фраз, основанный на статистическом анализе с последующим применением тематической модели. Метод основан на итеративном слиянии фраз, рассчитывая для каждого слияния его значимость. В работе [8] были скомбинированы три подхода к нахождению терминов: синтаксический отбор, статистический анализ и тематический отбор для дополнения признакового описания для задачи ранжирования. Одним из рассмотренных методов статистического анализа был выбран алгоритм TopMine.

2 Постановка Задачи

Разделим задачу на три логические части: формирование словаря n-грамм для описания документов, построение векторных представлений документов и построение модели классификации. Поставим задачу для каждой из частей.

2.1 Векторные представления

Для решения задачи об определении категории объявления по его тексту воспользуемся их векторными представлениями. Векторные представления получим с помощью мультимодальной ARTM. Опишем основные определения и понятия для этого.

2.1.1 Тематическая модель: основные понятия и определения

Тематическое моделирование – подход к анализу текстовых коллекций. Данный подход позволяет провести мягкую кластеризацию объектов коллекции, а так же получить векторные представления термов словаря коллекции.

Обозначим D - множество документов (коллекция). Пусть W - множество термов (словарь). Пусть C - множество категорий (рубрики) Документ d включает в себя набор термов $w \in W$ и в некоторых случаях одну категорию $c \in C$.

Предполагаем, что существует множество латентных (скрытых) переменных T. Такое, что каждое вхождение терма w в документ d связано с проявлением темы t из T.

Предполагаем, что порядок термов в документе не зависит от темы, и что потеря информации о месторасположении терма внутри документа не повлияет на выявление тем (гипотеза мешка слов).

Тогда D можно представить, как выборку троек (d, ω, t), независимо полученных из распределения $p(d, \omega, t)$ на множестве $D \times W \times T$.

Предполагаем, что наличие терма w в документе d по теме t зависит от t и не зависит от d (гипотеза условной независимости [9]). Распределение слов по теме описывается как $p(\omega \mid t)$. Причем, выполняется

$$p(\omega \mid d, t) = p(\omega \mid t). \tag{2.1}$$

Обозначим $p(\omega \mid t) = \phi_{\omega t}$ и $p(t \mid d) = \theta_{td}$. Тогда можем записать вероятность появления терма в документе:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t, d) p(t \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \phi_{wt} \theta_{td}. \quad (2.2)$$

Будем наблюдать количество n_{dw} вхождений терма w в документ d.

Тогда, можем заключить информацию о коллекции в матрицу частот вхождений термов в документы размера $|W| \times |D|$ состоящую из элементов n_{wd}.

Примером вероятностной тематической модели является модель PLSA [9]. Для построения модели (2.2) максимизируется логарифм правдоподобия (2.3) при ограничениях нормировки и неотрицательности (2.4) компонент Φ и Θ:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta}; \quad (2.3)$$

$$\sum_{w \in W} \phi_{wt} = 1; \quad \phi_{wt} \geq 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0. \quad (2.4)$$

Для решения этой задачи принято использовать EM-алгоритм. Алгоритм состоит из двух шагов: E-шаг (expectation) и M-шаг (maximization) – которые итерационно повторяются заданное гипперпараметром число раз. На E-шаге по текущим $\phi_{\omega t}$ и θ_{td} строится приближение $p(t \mid d, \omega)$. На M-шаге по текущему $p(t \mid d, \omega)$ вычисляются приближения для $\phi_{\omega t}$ и θ_{td}.

ARTM. Считаем, что $|T|$ много меньше $|D|$ и $|W|$, поэтому задача тематического моделирования сводится к поиску приближённого матричного разложения $F \approx \Phi\Theta$, ранг которого не превышает $|T|$.

Задача стохастического матричного разложения является некорректно поставленной, так как множество её решений в общем случае бесконечно. Если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных матриц S, при условии, что матрицы ΦS и $S^{-1}\Theta$ — стохастические. Добавление дополнительной регуляризации помогает доопределить задачу, а так же наделить матрицы Φ и Θ нужными нам свойствами.

Тематическая модель с адитивной регуляризацией (ARTM) [10] является обобщением модели PLSA. Для ее построения максимизируется линейная комбинация логарифма правдоподобия и нескольких регуля-

ризаторов $R_i(\Phi, \Theta)$, $i = 1, \ldots, k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta); \tag{2.5}$$

при ограничениях (2.4), где τ_i — неотрицательные коэффициенты регуляризации.

Пример регуляризатора сглаживания и разреживания матриц Φ, Θ:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \tag{2.6}$$

Положительные значения β_{wt} и α_{td} соответствуют сглаживанию распределений φ_{wt} и θ_{td}, а отрицательные - разреживанию.

Мультимодальная ARTM. В некоторых задачах документы содержат различные типы информации (например: слова, n-граммы, автор, дата, метка класса). Дополнительная информация может помочь улучшить качество тематической модели. Такие типы информации принято называть модальностями и строить для каждой модальности свою матрицу Φ^m, $\Phi = \bigcup_{m \in M} \Phi^m$, где $m \in M$ - множество модальностей. Построенная с использованием нескольких модальностей тематическая модель называется мультимодальной [11]. Максимизируемый функционал принимает вид:

$$\sum_{m \in M} \alpha_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta); \tag{2.7}$$

при ограничениях (2.4), где α_m —неотрицательные веса модальностей.

2.2 Тематическая модель классификации

Рассматривается задача мультиклассовой классификации с непересекающимися классами. Пусть есть множество возможных документов $\mathbb{D}$. Дано множество документов $D = \{d_1, \ldots, d_{|D|}\} \subset \mathbb{D}$. Существует $y : \mathbb{D} \to C$ - неизвестная зависимость, отображающая множество возможных документов в множество классов. Данная зависимость определяет множество $\{y_1, \ldots, y_{|D|}\}$ - метки классов соответствующих документов такие, что $y_i = y(d_i), i = 1, \ldots, |D|$. Требуется найти алгоритм $a : \mathbb{D} \to C$, приближающий y на всем множестве $\mathbb{D}$. Стандартный подход

к решению этой задачи – поиск алгоритма a в классе линейных классификаторов (2.8):

$$a(X, \omega) = \arg\max_{c \in C} < \omega_c, X >= \tag{2.8}$$

$$= \arg\max_{c \in C} < \phi_c, \theta_d >= \arg\max_{c \in C} \sum_{t \in T} \phi_{ct}\theta_{td} = \arg\max_{c \in C} p(c \mid d) \tag{2.9}$$

где X - матрица векторных представлений документов D, а ω - матрица весов линейного классификатора.

Тематическая модель классификации. В работе [12] строится тематическая модель классификации (2.10) на основе модели LDA. Было показано, что тематическая модель классификации превосходит обычные методы классификации на больших текстовых коллекциях, с большим числом несбалансированных, пересекающихся, взаимозависимых классов.

$$p(c \mid d) = \sum_{t \in T} \phi_{ct}\theta_{td} \tag{2.10}$$

Ее принцып работы таков. В дополнение к основной модальности W в тематическую модель вводится дополнительная модальность для меток класса C. Данная модальность запоняется только у объектов тренировочной выборки. Строится тематическая модель. На основе её матриц Φ и Θ делается предсказание модальности меток класса на объектах тестовой выборки. Таким (2.9) образом матрица Θ служит матрицей признакового описания объектов, а матрица Φ матрицей весов линейной модели.

В работе [13] была построена тематическая модель классификации (2.10) на основе модели ARTM и подтверждены выводы о превосходстве обычных методов классификации.

Классификация по тематическим векторам. Допускается, что вероятностная тематическая модель могла не очень хорошо обучиться, из-за особенностей коллекции, таких как размеры коллекции и малом количестве документов в некоторых классах. Но при этом ею могли быть быть получены хорошие векторные представления документов. Из этого можем сделать предположение, что доучивание матрицы весов линейной модели с прежней матрицей Θ улучшит качество модели. Доучим матрицу весов модели ω по признаковому описанию из Θ документов из D с различными функциями потерь $\mathcal{L}$:

$$\sum_{i=1}^{|D|} \mathcal{L}(a, c_i) \to \min_{\omega} \tag{2.11}$$

Функции потерь. Пусть $g(x, \omega) = 0$ - уравнение разделяющей поверхности, где $g(x, \omega)$ - разделяющая функция такая что $a(x, \omega) = \arg\max_{c \in C} g(x, \omega)$. Отступом (margin) объекта x_i называется величина $M_i(\omega) = g(x_i, \omega) y_i$. Различные функции потерь можно выразить через отступ M. Среди функций потерь рассмотрим кусочно линейную (SVM) и логарифмическую (LR):

$$SVM: \ \mathcal{L}(M) = (1 - M)_+ \tag{2.12}$$

$$LR: \ \mathcal{L}(M) = \log_2(1 + e^{-m}) \tag{2.13}$$

2.3 Формирование словаря n-грамм

2.3.1 Выбор N-грамм

N-грамма – последовательность из n слов, или иных единиц языка.
Коллокация – n-грамма из слов, встречающаяся сильно чаще, чем ожидается, если бы это были отдельные независимые слова.
Словосочетание – коллокация из связанных по смыслу, а также грамматически, обычно служит для обозначения конкретного понятия (термина).

Для повышения качества тематической модели будем добавлять в словарь различные коллокации. Для отбора коллокаций будем использовать модернизацию алгоритма TopMine [1].

Терминами обычно называют n-граммы со следующими свойствами:

- frequency – высокая частотность в коллекции
- collocation – коллокация
- completeness – максимальная по включению цепочка слов (полнота)
- syntactic connectedness – грамматическая корректность
- topicality – тематичность

2.3.2 Классический TopMine

Данный метод основан на идее слияния рядом стоящих n-грамм. А именно, алгоритм начиная с $n = 1$ подсчитывает насколько часто стоят подряд две n-граммы:
$C(a_1, \ldots, a_k)$ – хэш-таблица частот k-грамм.

$A_{d,k}$ – множество позиций i в документе d, с которых начинаются все частые k-граммы: $C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k$.
Основной шаг алгоритма выглядит так: если $(i \in A_{d,k})$ и $(i+1 \in A_{d,k})$ то увеличиваем на единицу $C(w_{d,i}, \ldots, w_{d,i+k})$, где $i = 1, \ldots, n_d$.

Algorithm 2.1 TopMine N-grams

Input: Коллекция D, пороги ε_k
$C(w) := n_w$ для всех $w \in W$
$A_{d,0} := \{1, \ldots, n_d\}$
for $k := 1, \ldots, k_{max}$ **do**
  for $d \in D$ **do**
    $A_{d,k} := \{i \in A_{d,k-1} \mid C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k\}$
    for $i \in A_{d,k}$ **do**
      if $i+1 \in A_{d,k}$ **then**
        ++$C(w_{d,i}, \ldots, w_{d,i+k})$
      end
    end
  end
  Оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$
end
Result: Хэш-таблица частот $C(a_1, \ldots, a_k),\ k = 1, \ldots, k_{max}$

2.3.3 Модификация TopMine: SourceTopMine

Оригинальный алгоритм из статьи [1] обладает несколькими недостатками.

Первый недостаток – алгоритм навязывает необходимое свойство термина completeness, то есть выделяет только максимальную по включению цепочку слов, отбрасывая все составляющие. Это не всегда полезно, например, 3-грамма «заказать шкаф купе» будет создана благодаря 2-граммам «заказать шкаф» и «шкаф купе», которые сами по себе тоже выглядят полезными. Но алгоритм TopMine удаляет n-граммы, которые послужили материалом для создания (n+1)-грамм. Способ борьбы с этой проблемой был предложен в работе [14], он заключается в игнорировании удаления.

Второй недостаток – при подсчете хэш-таблицы частот $C(a_1, \ldots, a_k)$ учитываются все вхождения n-граммы в одном документе. Это заставляет алгоритм выдавать за коллокации n-граммы, встречающиеся только в одном документе, но много раз. Такие n-граммы как правило бесполезны для решения практических задач. Чтобы устранить этот недостаток

в работе предлагается вместо частоты вхождений n-граммы в коллекцию подсчитывать число документов, в коллекции, в которых n-грамма была найдена. Ниже предложен алгоритм с учетом этой модификации SourceTopMine.
$C(a_1, \ldots, a_k)$ – хэш-таблица частот k-грамм.
$C_{d,k}(a_1, \ldots, a_k)$ – множество частот k-грамм в документе d.
$A_{d,k}$ – множество позиций i в документе d, с которых начинаются все частые k-граммы: $C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k$.

Input: Коллекция D, пороги ε_k

$C(w) := n_w$ для всех $w \in W$
$A_{d,0} := \{1, \ldots, n_d\}$
for $k := 1, \ldots, k_{max}$ **do**
  for $d \in D$ **do**
    $C_{d,k} := \{\}$;
    $A_{d,k} := \{i \in A_{d,k-1} \mid C(w_{d,i}, \ldots, w_{d,i+k-1}) \geqslant \varepsilon_k\}$
    for $i \in A_{d,k}$ **do**
      if $i + 1 \in A_{d,k}$ **then**
        $C_{d,k}$ *add* $(w_{d,i}, \ldots, w_{d,i+k})$
      end
    end
    for $(w_{d,i}, \ldots, w_{d,i+k}) \in C_{d,k}$ **do**
      $++C((w_{d,i}, \ldots, w_{d,i+k})$
    end
  end
  оставить только частые k-граммы: $C(a_1, \ldots, a_k) \geqslant \varepsilon_k$
end
Result: хэш-таблица частот $C(a_1, \ldots, a_k),\ k = 1, \ldots, k_{max}$

Благодаря этой модификации мы можем низко опускать пороги ε_k для $k > 2$.
Полагаем, что если фразу из трех и более слов употребили хотя бы в двух различных документах, то велика вероятность того, что она не случайна. Для этого предлагается выставлять порог $\varepsilon_k = 2$ для $k > 2$.

Так же, предполагаем, что термы встреченные только в одном документе не влияют на его тематическое представление, а значит не влияют на качество решения задачи о восстановлении меток класса. Заметим, что множество k-грамм $k > 2$, полученных с помощью SourceTopMine при $\varepsilon_k = 2$ для $k > 2$ включает в себя все термы, полученные алгорит-

мом TopMine, исключая n-граммы, встречающиеся ровно в одном документе.

Таким образом предложен алгоритм SourceTopMine, в котором решена проблема многократных повторов в текстах. Так же для данного алгоритма существует обоснованное, интерпретируемое, не зависящее от размеров коллекции значение гипперпараметра $\varepsilon_k = 2$, что может избавить нас от его подбора.

2.4 Формальная постановка задачи восстановления категорий

Дано:
D - база объявлений.
$d \in D,\ d = w_{d,1}, \ldots, w_{d,N_d}$ – текст описания объявления об услуге, где N_d – число слов в описании.
$C = \{c_1, c_2, \ldots, c_{Nc}\}$ – Множество классов услуг.
Задачи:
Востановить класс услуги c_i объявления d_j там, где он пропущен.
Критерии качества:
accuracy по асессорской разметке объявлений услуг без класса.

3 Вычислительные эксперименты

3.1 Получение и обработка данных

Дана коллекция текстовых документов – описаний объявлений о предоставлении услуг. Текст каждого объявления не пустой и его длинна $\leqslant 2000$ символов. Существует трехуровневая строгая иерархия: профессия, специализация и услуга. Каждое объявление имеет метки класса профессии и специализации. Некоторые объявления имеют метки класса услуги. В эксперименте участвуют объявления только из одной профессии – «Ремонт и строительство». В ней содержится 700 классов уровня услуг. На которые приходится 850к объявлений, из них 260к не имеют меток класса услуги. Такие объявления полезны для моделей составления словаря и для тематических моделей, но бесполезны для моделей задачи классификации.

Для оценки качества модели классификации используем разметку 800 объявлений.

Дана коллекция терминов, которая получена из набора синонимов к некоторым услугам и специализациям, полученной из экспертной разметки. Она включает в себя 3813 термин. Из них 2006 содержатся в текстах описаний объявлений, то есть их можно будет в них найти. Будем использовать эту коллекцию для оценки качества моделей выделения n-грамм.

Все тексты и термины были лематизированны, приведены в нижний регистр, избавлены от знаков препинания, специальных символов и лишних отступов, были удалены стоп слова. Все документы, переставшие содержать слова после данной процедуры были удалены из коллекции.

Примеры объявлений:

Класс: Ремонт пластиковых окон
Текст: Ремонт пластиковых окон Ремонт пластиковых окон под ключ

Класс: Ремонт и замена оконных ручек
Текст: Модели, предлагаемые на профильном рынке, отличаются лишь типом рабочего элемента. Повсеместно распространены следующие их категории: двухсторонние и односторонние; с замком, подразумевающим использование ключа; с фиксатором; в виде лепестка либо ракушки.

3.2 Сравнение терминов классического TopMine и модифицированного

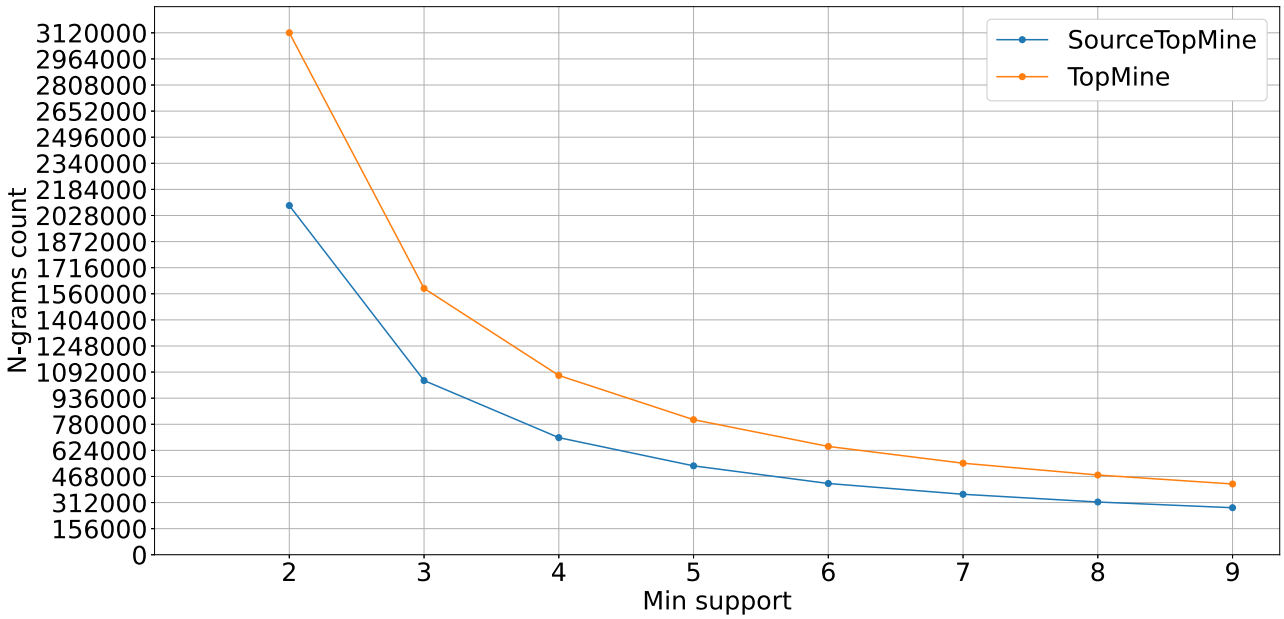

Рис. 1: Зависимость количества найденных терминов от порога минимального вхождения в текстовую коллекцию. От 2 до 10 вхождений

В данном эксперименте сравниваются результаты работы оригинального алгоритма TopMine с предложенной в данной работе модификацией SourceTopMine на коллекции описанной выше, а так же производится подбор гиперпараметра min support ε_k для обоих моделей на основе количества и качества полученных терминов.

При сравнении будем обращать внимание на метрику количества N-грамм в полученном словаре (*N-grams count*). Считаем, что чем меньше *N-grams count*, тем лучше, так как увеличение словаря влечет увеличение числа обучаемых параметров модели, что усложняет модель. Терминов в разметке мало и они есть только в некоторых классах, поэтому мы не рассматриваем напрямую полноту и точность выделения полезных n-грамм, а вводим метрику *terms detected* - число найденных n-грамм принадлежащих коллекции терминов синонимов, про которые мы точно знаем, что они полезны и осмыслены. Считаем, что чем больше *terms detected*, тем лучше полнота, чем больше отношение *terms detected* к *N-grams count*, тем лучше точность.

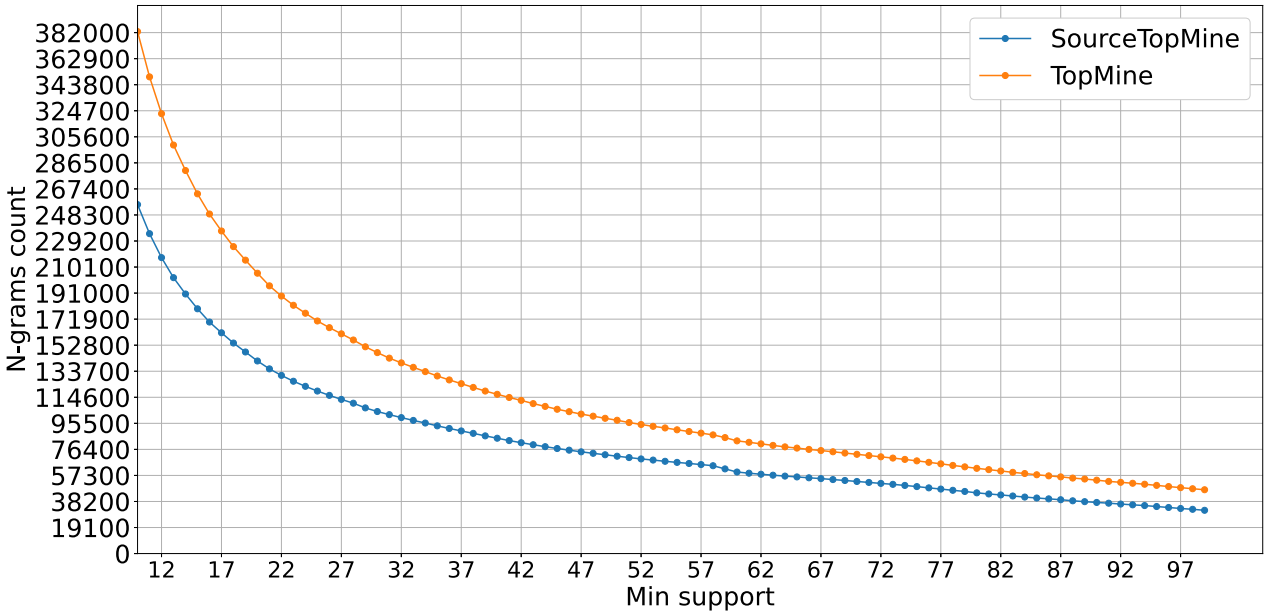

Рис. 2: Зависимость количества найденных терминов от порога минимального вхождения в текстовую коллекцию. Начиная с 10 вхождений

В качестве документов для SourceTopMine служат объединения всех объявлений от одного автора. Таким образом, в SourceTopMine вместо количества вхождений терма в коллекцию подсчитывается количество авторов, которые употребили терм.

На графиках (2) и (1) представлена зависимость количества найденных термов от порога *Min support*, который в случае TopMine является вхождением терма в коллекцию, а в случае SourceTopMine является числом авторов, употребивших терм.

На графике (3) по оси Y количество терминов из экспертной разметки, которое попало в список найденных термов при различных ограничениях *Min support*. По оси X количество N-грамм в словаре. Значение *Min support* есть порядок точки на графике справа налево, начиная с 2. Самая правая точка соответствует $Min\ support = 2$, следующая соответствует 3.

Из графиков (2) и (1) следует, что модель SourceTopMine выделяет меньше термов. В результатах обеих моделей большая часть термов встречается всего несколько раз. По графику (3) видно, что с $Min\ support = 2$ TopMine находит больше терминов из разметки, чем SourceTopMine, тк некоторые из терминов разметки встретились только у одного автора. Так же видно, что график SourceTopMine ле-

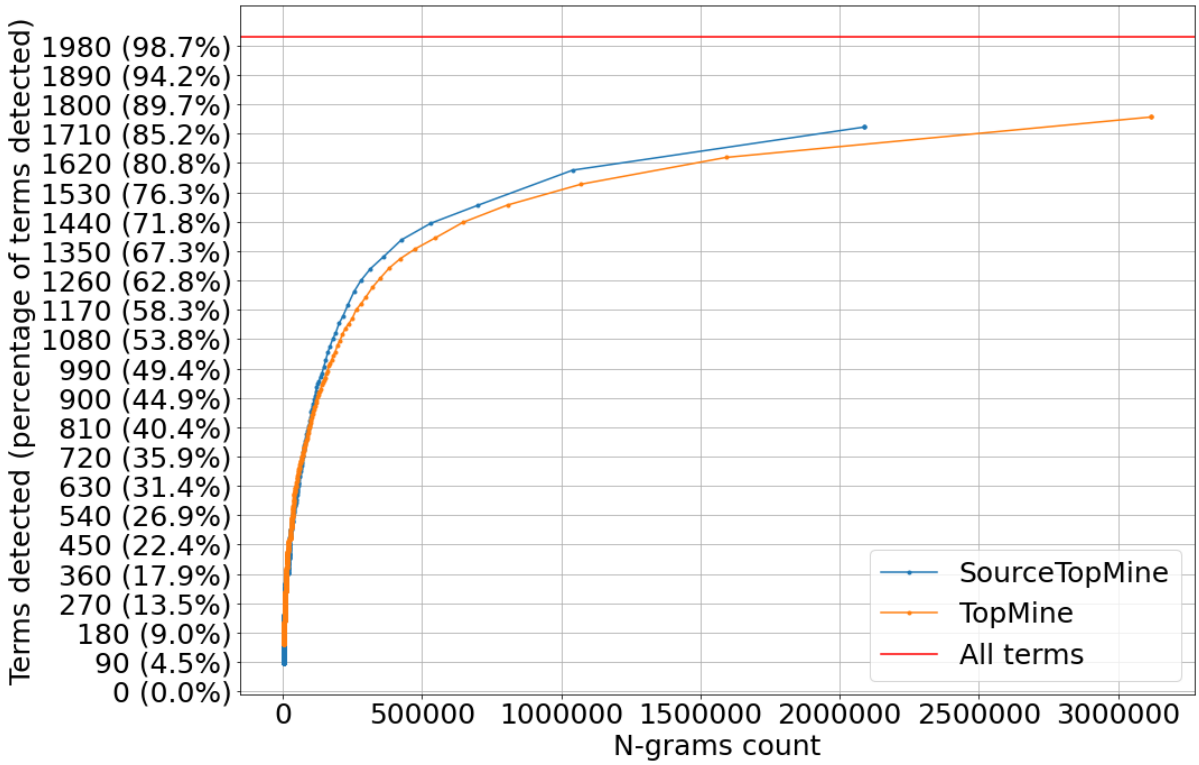

Рис. 3: Количество найденных терминов разметки (и доля найденных) от размера словаря N-грамм

жит строго выше графика TopMine, из чего можно сделать вывод, что SourceTopMine показывает лучшую точность, чем TopMine. То есть, при необходимости ограничить размер словаря термов сверху SourceTopMine покажет лучшую полноту. В дальнейших экспериментах будет установлено ограничение на размер словаря в 1000000 (1 миллион) термов.

3.3 Сравнение методов классификации

Для задачи восстановления категорий использовалось несколько алгоритмов.

Описанные алгоритмы будут отличаться по методу формирования словаря, построению тематической модели и методу классификации.

Рассмотрено три метода формирования словаря: слова вместо термов; термы из TopMine; термы из SourceTopMine. Для построения мультимодальной тематической модели PLSA использована библиотека BigARTM [15]. В качестве дополнительной модальности использовалась категория уровня услуг у тех документов, у которых она известна. Для классификации использованы методы опорных векто-

ров и логистической регрессии, реализованных в svm.LinearSVC и linear_model.LogisticRegression соответственно из пакета sklearn, над векторными представлениями документов, полученными из матрицы Θ тематичческой модели.

Так же, для классификации использованы два алгоритма на основе эвристик.

В первом алгоритме "PredModal"предсказывается модальность категории уровня услуг (2.9). То есть, берем аргумент максимальной компоненты вектора вероятностей наличия модальности в документе, полученного из обученной тематической модели:

$$r_{pred}(d) = \arg\max_{r \in R} p(r|d).$$

Во втором алгоритме "Topic2rubric"предсказываем рубрику таким образом:

$$r_{pred}(d) = \arg\max_{r \in R} n(r|t_d),$$

где

$$t_d = \arg\max_{t \in T} p(t|d),$$

$$n(r|t) = \sum_{d \in D_t} I(r(d) = r),$$

$$D_{t_0} = \bigcup_{d \in D} d \text{ , где } \arg\max_{t \in T} p(t|d) = t_0.$$

$r(d)$ - рубрика документа d. То есть, в данной эвристике каждому документу сопоставляется наиболее вероятная тема, каждой теме сопоставляется ровно одна рубрика.

3.3.1 Подбор гипперпараметров тематической модели

В данном разделе проиллюстрирован подбор гипперпараметров модели тематического моделирования. Подбирался вес модальности меток классов, число итераций EM-алгоритма для обучения тематической модели и количество тем.

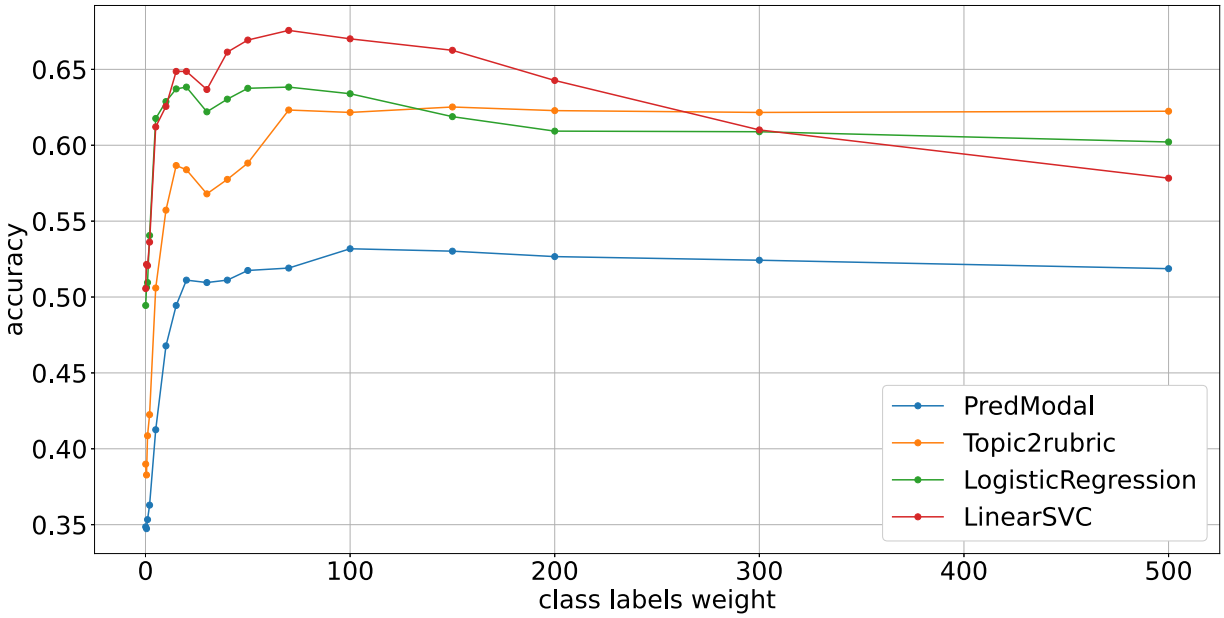

Рис. 4: Качество классификации на различных моделях в зависимости от веса модальности меток класса в тематической модели

Из графика (4) можно сделать вывод, что для всех моделей кроме $PredModal$ оптимум веса модальности меток класса к весу основной модальности находится в окрестности $70:1$, для модели $PredModal$ в окрестности $100:1$. Так же, и графика видно, что наличие модальности метки класса значительно влияет на качество всех моделей, тк при минимальном рассмотренном весе $1:10$ качество моделей значительно хуже, чем в оптимуме.

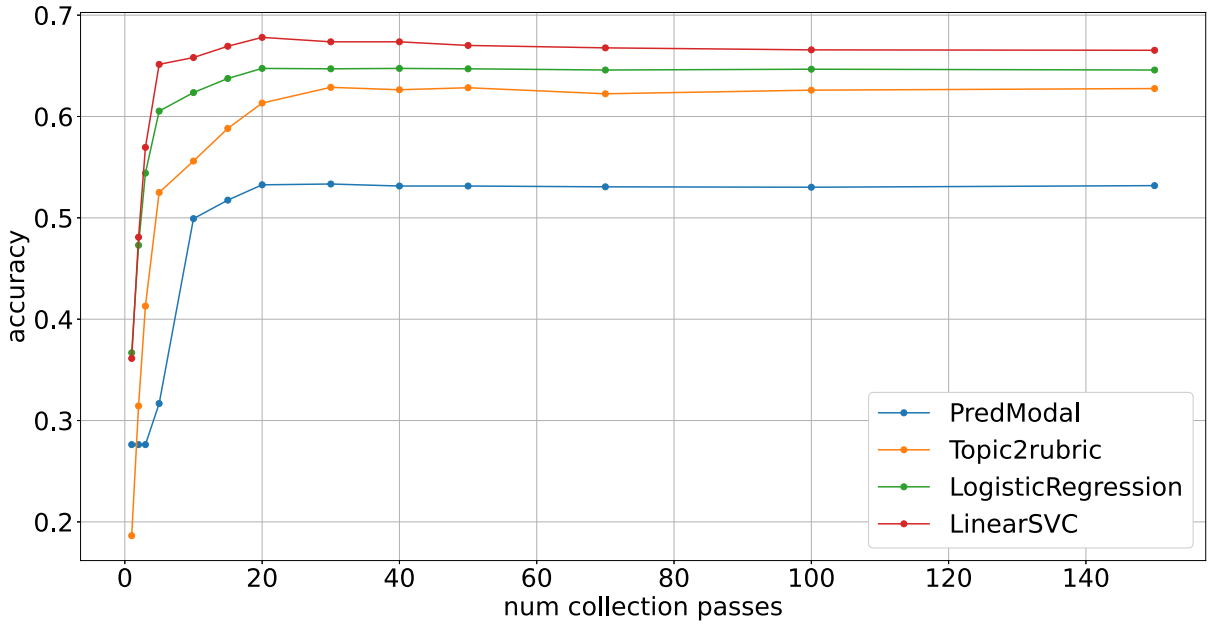

Рис. 5: Качество классификации на различных моделях в зависимости от количества итераций EM-алгоритма при построении тематической модели

По графику (5) видно, что итоговое качество моделей не сильно зависит от количества проходов EM-алгоритма по коллекции после порога в 20 проходов. Однако локальный оптимум расположен на 20 или 30 проходах, в зависимости от модели. Это может объясняться тем, что перплексия модальности меток классов имеет локальный минимум в районе 20 итераций (6).

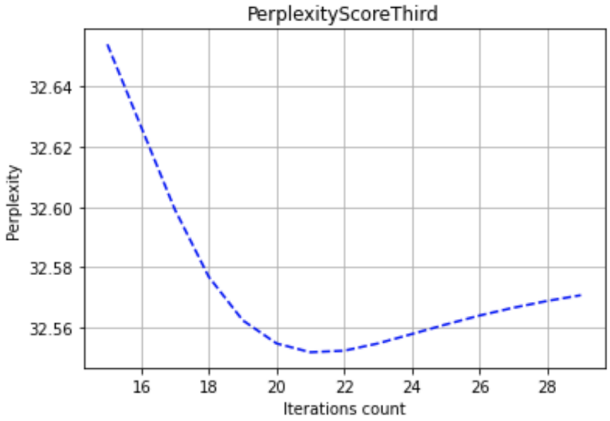

Рис. 6: Перплексия модальности меток класса в зависимости от количества итераций EM-алгоритма при построении тематической модели

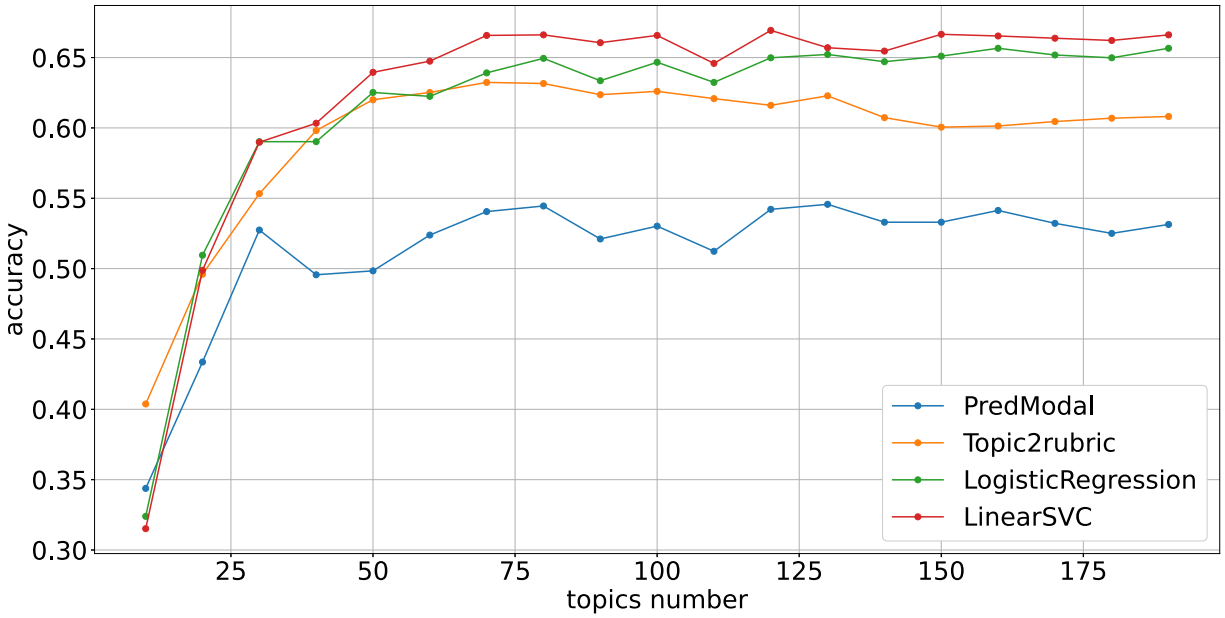

Рис. 7: Качество классификации на различных моделях в зависимости количества тем тематической модели

По графику (7) можно сделать вывод, что с 80 тем графики моделей выходят на плато и в дальнейщем не растут. .

3.3.2 Результаты экспериментов

	PredModal	Topic2rubric	**LinearSVC**	LogisticRegression
Words only	0.359	0.506	0.552	0.444
TopMine	0.523	0.605	0.638	0.637
SourceTopMine	0.530	0.626	**0.666**	0.647

Таблица 1: Точность восстановления меток класса на различных классификаторах, на различных методах формирования словарях термов.

	PredModal	Topic2rubric	**LinearSVC**	LogisticRegression
Words only	0.359	0.506	0.552	0.444
TopMine	0.491	0.597	0.626	0.610
SourceTopMine	0.527	0.614	**0.666**	0.640

Таблица 2: Точность восстановления меток класса на различных классификаторах, на различных методах формирования словарях термов при ограничении размера словаря в 1000000 (1 миллион) термов.

В таблицах (1) и (2) приведены результаты экспериментов с различными методами классификации при оптимальных для них гипперпараметрах. Таблицы различаются условием на ограниченность размера словаря N-грамм для построения тематической модели. В таблице (1) оно отсутствует. В таблице (2) присутствует на уровне 1000000 (1 миллион) N-грамм.

Из обеих таблиц видно, что наилучший результат на всех рассмотренных методах формирования словаря показывает классификатор с помощью метода опорных векторов. Это подтверждает работоспособность построенного алгоритма тематической классификации. И свидетельствует о превосходстве его над методом PredModal на данной коллекции. Среди методов формирования словаря лучший результат на SourceTopMine. Это подтверждает обоснованность предложения о пользе модификации алгоритма TopMine.

Разница между значениями точности для одних и тех же комбинаций методов между таблицами (1) и (2) иллюстрирует влияние ограничения размера словаря на итоговое решение задачи классификации. На результаты полученные с помощью метода TopMine ограничение словаря повлияло значительно сильнее, что согласуется с результатами эксперимента про количество найденных терминов из разметки. .

4 Заключение

В работе было предложено улучшение алгоритма TopMine, направленное на уменьшение объема словаря и улучшение качества на задаче классификации.

Предложен алгоритм восстановления меток классов документов на основе их векторных представлений, полученных с помощью тематической модели. Даны рекомендации по подбору гипперпараметров модели.

Список литературы

[1] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han. Scalable topical phrase mining from text corpora, 2014.

[2] Воронцов К. В. Вероятностное тематическое моделирование: обзор моделей и регуляризационный подход. 2020.

[3] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85:41–75, 10 2011.

[4] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 08 2008.

[5] K. Hasan and Vincent Ng. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In *COLING*, 2010.

[6] Xuerui Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702, 2007.

[7] Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. Automatic construction and ranking of topical keyphrases on collections of short documents. In *Proceeding of 2014 SIAM International Conference on Data Mining*. SIAM – Society for Industrial and Applied Mathematics, April 2014.

[8] Полушин В. В. Тематические модели для ранжирования рекомендаций текстового контента. 2017.

[9] Thomas Hofmann. Hofmann, t.: Unsupervised learning by probabilistic latent semantic analysis. machine learning 42(1-2), 177-196. 42:177–196, 01 2001.

[10] K. V. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, May 2014.

[11] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101:1–21, 12 2014.

[12] Timothy Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Mach Learn*, 88, 07 2011.

[13] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Ianina. Non-bayesian additive regularization for multimodal topic modeling of large collections. 10 2015.

[14] Попов А. С. Выделение множества тематик в неразмеченной коллекции диалогов. 2019.

[15] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: Open source library for regularized multimodal topic modeling of large collections. pages 370–381, 04 2015.