Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Федеральном исследовательском центре
«Информатика и управление» РАН

Потапова Полина Сергеевна

Тематическое моделирование образовательных целей пользователей в системе дистанционного образования

03.04.01 — Прикладные математика и физика

Магистерская диссертация

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

Москва
2020 г.

Содержание

Аннотация

На сегодняшний день все более популярными становятся системы дистанционного образования и системы массовых открытых онлайн курсов, такие как Coursera, edX и другие. Пользователи таких систем уже получают персональные рекомендации образовательного контента на основе своих прошлых действий. Однако для построения индивидуальных образовательных траекторий необходимо учитывать образательную цель пользователя. Предполагается, что пользователь заполняет анкету, в которой формулирует свою цель на естественном языке, а также выбирает варианты ответов на несколько формальных вопросов. Часть анкет размечается экспертами, которые оценивают, насколько конкретно пользователи формулировали свои цели.

В работе предлагается модель классификации, которая определяет степень конкретности образовательных целей пользователя в системе дистанционного образования. Алгоритм основан на тематической модели с аддитивной регуляризацией, которая в явном виде учитывает предположение о разлелении текстов, слов и тем на конкретные и неконкретные.

Проведенные вычислительные эксперименты показали, что тематическая модель специфичности документов справляется с задачей лучше, чем стандартные модели бинарной классификации — наивный байесовский классификатор, логистическая регрессия, а также упрощённая тематическая модель классификации, построенная без моделирования конкретности.

1 Введение

В связи с ускорением темпа жизни и постоянно растущими требованиями рынка труда, люди всё больше заинтересованы в саморазвитии. Поэтому набирают популярность системы массовых открытых онлайн курсов (MOOC-системы), такие как Coursera, edX и другие. Пользователи таких систем хотят получать персональные рекомендации образовательного контента. Они хотят эффектривно проводить своё время, тратить минимальное количество времени на достижение максимального резальтата. Поэтому они ожидают, что рекомендованный контент будет релевантен их образовательным целям.

Существуют различные решения рекомендаций образовательного контента. Рекомендации осуществляются с использованием ключевых слов, есть решения, когда интересы пользователя аккумулируются по текстам описаний всех образовательных курсов, на которые заходил пользователь, а также много других решений, которые будут рассмотрены в разделе 2. Но на сегдняшний день не существует модели рекомендательной системы, которая бы учитывала непосредственно образовательную цель пользователя в явном виде, написанную на естественном языке.

К сожалению, люди не всегда четко и явно знают чего они хотят, не всегда их цель является достаточно конкретной, чтобы по ней вообще имело смысл делать рекомендации. Поэтому актуальной является задача определения конкретности образовательной цели, написанной на естественном языке. Эта задача бинарной классификации. Постановка задачи описана в разделе 5. В разделе 3 описаные существующие подходы к оцениванию специфичности текста. Целью данной работы является построение модели бинарной классификации для определения специфичности образовательных целей, написанных на естественном языке.

В данной работе предлагается решение поставленной задачи при помощи тематического моделирования с аддитивной регуляризацией. В разделе 4 приведена основная теория о тематических моделях с аддитивной регуляризацией. Раздел 6 посвещен описанию идей, лежащих в основе решения. Основной идеей является разделение слов, документов и тем на конкретные и неконкретные. Также модель решения позволяет не только осуществлять бинарную классификацию целей на конкретные и неконкретные, но и подчитывать степень конкретности для каждой цели.

В разделе 7 описаны исходные данные, описаны этапы предобработки текстов, предложены базовые модели решения. Подраздел 7.4 посвещен реализации основного решения - тематической модели с аддитивной регуляризацией.

В разделе 8 анализируется интерпретируемость полученных тем и предлагается способ графической визуализации результатов моделирования в виде круговой диаграммы «темы–специфичность».

В заключении перечисляются результаты, выносимые на защиту.

Проведенные вычислительные эксперименты показали, что тематическая модель специфичности документов справляется с задачей лучше, чем стандартные модели бинарной классификации — логистическая регрессия, наивный байесовский классификатор, а также тематическая модель классификации, построенная без учета конкретности и неконкретности слов, документов и тем.

2 Задачи и методы рекомендаций индивидуальных образовательных траекторий

В связи с увеличением темпа жизни, динамически развивающимся рынком труда и техническим прогрессом, растут и требования к качеству и доступности образования. В последние годы увеличилась популярность систем онлайн образования и массовых открытых онлайн-курсов (Massive Open Online Courses), таких как Coursera, edX и т. п. Это позволило накопить большие объёмы образовательных данных. Пользователи таких систем хотят получать качественные рекомендации образовательного контента, релевантного их образовательным целям.

Потребности времени и накопленные данные привели к появлению задачи индивидуализации обучения.

Исследователь в области наук об образовании Тимошина Т. А. так определяет понятие *индивидуальной образовательной траектории (ИОТ)*: «индивидуальная образовательная траектория студента – это индивидуальный путь в образовании, определяемый студентом совместно с преподавателем, организуемый с учетом мотивации, способностей, психических, психологических и физиологических особенностей обучающегося, а также социально-экономических и временных возможностей субъекта образовательного процесса» [1].

Махныткина О. В. описывает это понятие так: «Под индивидуальной образовательной траекторией будем понимать частично упорядоченный по последовательности изучения набор дисциплин, а также тематику научно-исследовательской работы студента, на которых основывается процесс обучения конкретного учащегося. Индивидуальная образовательная траектория определяется образовательными потребностями, индивидуальными способностями и возможностями учащегося (уровень готовности к освоению программы), а также существующими стандартами содержания образования.» [2]

Таким образом, ИОТ должна учитывать личные мотивации и потребности обучающегося, а также различные его характеристики и возможности.

Долгое время индивидуальное образование один на один с преподавателем считалось самым эффективным. Но предоставить каждому студенту персонального преподавателя для совместных занятий и определения пути в образовании едва ли экономически осуществимо.

Это приводит к появлению задачи автоматического построения индивидуальных образовательных траекторий и рекомендаций образовательных ресурсов в MOOC-системах по личным образовательным целям обучающихся, написанным на естественном языке.

Рассмотрим как решается задача рекомендации образовательного контента в различных областях.

2.1 В российских государственных ВУЗах

В российских государственных ВУЗах результаты, которых должен достичь студент по освоении образовательной программы, описаны в федеральном государственном образовательном стандарте высшего образования (ФГОС ВО). Эти результаты представляют из себя набор компетенций, которыми должен овладеть студент. В работе [3] приведена иерархическая модель оценивания степени овладения компетент-

ностями для студентов-выпускников. А в [4] рассматривается задача нахождения оптимальной индивидуальной образовательной траектории студента с использованием этой модели. Построение ИОТ реализуется на основе методов динамического программирования с учетом иерархической структуры компетентности.

Едва ли можно сказать, что при таком подходе учитываются образовательные цели конерктного студента. Этот метод может помочь ВУЗам выпускать более профессиональных специалистов в соотвутствующих областях, но не поможет студентам достичь желаемого индивидуального образовательного результата.

Авторы работы [5] предлогают моделировать ИОТ с учетом пожеланий студентов. Они ставят задачу нахождения оптимальной ИОТ в терминах теории оптимизации и предлогают оптимизировать степень удовлетворенности студента по всем возможным траекториям.

Пожелания же студентов они учитывают таким образом, что оценивают курсы по выбору и темы НИР по степени привлекательности для студента. Но такая постановка задачи отлична от той, когда мы хотим учитывать конкретную образовательную цель.

2.2 В областях, связанных с рынком труда

В работе [6] говорится, что компетентности, которыми должен овладеть выпускник ВУЗа, должны соответсоовать текущим требованиям рынка труда. Авторы рассматривают метод семантического поиска для рекомендации образовательного контента под заданные требования рынка труда. Релевантность курсов оценивается по семантической близости эмбеддингов требований рынка труда и эмбеддингов текстовых описаний курсов. Курсы ранжируются по релевантности, затем пользователю выдается ранжированный список.

Задача рекомендации образовательного контента актуальна не только в области образования. Она также имеет место в направлениях, связанных с помощью в продвижении по карьерному пути. Например авторы работы [7] разработали продукт под названием Next Job, по текущей должности соискателя позволяющий понять на какую следующую должность он может продвинуться и какими навыками для этого должен обладать. Также Next Job находит релевантные видео с YouTube для развития каждого из этих навыков. Авторы этой работы разработали модель карьерного пути, описывающую какими навыками должен обладать соискатель на новую должность при учете имеющихся навыков на текущей должности. С помощью этой модели для заданного названия текущей должности можно получить множество пар должность-должность и должность-навык. Для пары должность-навык осуществляется поиск релевантных видео. Для предсказания релевантности видео строится бинарный классификатор на основе случайного леса.

Стоит отметить, что авторы этой работы стоят только один шаг в траектории, а это предполагает, что человек четко знает, чего хочет. Это вполне актуально для поиска работы, но не для образования. Преимущество образовательных траекторий в том, что они состоят из последовательности таких шагов. Это дает человеку возможность на неуверенность, траекторию можно корректировать на каждом шагу.

В обеих рассмотренных работах вместо образовательных потребностей обучающегося рассматриваются интересы работодателей.

2.3 В системах дистанционного образования

Основной сферой применения рекомендаций образовательных материалов, конечно, являются массовые открытые онлайн-курсы.

Авторы работы [8] помогают студентам подобрать курсы, закрывающие бреши в их знаниях. По оценкам из колледжа они оценивают, по каким предметам студент наименее успешен, и рекомендуют онлайн-курсы, аналогичные по содержанию этим предметам. Авторы используют вероятностное тематическое моделирование, а именно метод латентного размещения Дирихле с сэмплированием Гиббса (LDA-GS) для обучения двух тематических моделей: модель курсов ВУЗа и модель курсов MOOC-систем. Эти модели позволяют выделить темы курсов колледжа и темы курсов из MOOC. Для рекомендации курсов используется content-based рекомендательная система. В ней для признакового описания каждого курса (и из колледжа, и MOOC) используются их тематические векторы и последовательность векторов, описывающих студентов. Вектор, описывающий студента, состоит из оценок его предпочтений в области искусства, математики, биологии и социальных наук в целом (они одинаковы для всех курсов). Для ВУЗовских курсов известна отметка студента за курс. Задача заключается в предсказании оценки студента за MOOC-курс. Предсказание осуществляется посредством многомерной линейной регрессии. Затем курсы ранжируют по предсказанным оценкам, и рекомендуют 10 курсов с самыми низкими оценками.

Как и в ранее рассмотренных работах, образовательные цели обучающегося не учитываются напрямую. Студенту рекомендуют изучить материалы по предметам с наименее хорошими оценками, хотя, возможно, он хотел бы освоить абсолютно другие направления.

Авторы только что рассмотренной статьи рекомендовали студентам курсы по предметам, в которых они наименее хороши. В работе [9] решается подобная задача. Авторы разработали вероятностную модель, которую можно использовать для рекомендации персонализированных последовательностей уроков с целью помочь учащимся подготовиться к сдаче конкретных заданий.

Но и здесь не учитывается образовательная цель обучающегося, вместо этого его подготавливают к решению конкретных заданий. Такая система наверняка хорошо показала бы себя для подготовки школьников к ЕГЭ. Но нашим целям она не отвечает.

Авторы следующей работы [10] используют для создания рекомендаций действительно много данных о пользователях. Для пары пользователь-курс строится целевая функция в виде взвешенной суммы по трем характеристикам. Одна из этих характеристик описывает латентные интересы пользователя. Строится она следующим образом. С помощью LDA по всем текстам курсов стоится тематическая модель. Тогда в качестве вектора скрытых интересов пользователя используется усредненный тематический вектор всех курсов, на которые заходил пользователь. Другая характеристика описывает социально-демографические особенности пользователя. Всех пользователей кластеризуют по социально-демографическим признакам. В качестве этой характеристики пользователя используют долю таких пользователей из кластера текущего пользователя, которые выбрали этот же курс. Последняя характеристика описывает информацию о курсах. Для упорядоченной пары курсов считается процент пользователей, записавшихся на текущий курс после предыдущего.

После этих характеристик считается целевая функция. Курсы ранжируются по значению целевой функции, пользователю рекомендуют курсы с наибольшим значением целевой функции.

В этой рекомендательной системе используется действительно много информации о пользователях. Потребности пользователя здесь представляются его скрытыми интересами, подсчитанными по всем курсам, на страницы которых он заходил.

Все рассмотренные работы учитывали интересы обучающихся лишь косвенно или не учитывали вовсе. Ни в одной из рассмотренных статей не был предложен метод рекомендаций, учитывающий образовательные цели обучающегося, написанные на естественном языке. Нам не удалось найти другие работы, предлогающие такой метод. Поэтому мы считаем актуальной задачу построения рекомендательной системы, учитывающей образовательные цели обучающегося.

2.4 Методология постановки целей SMART

Чтобы по цели делать рекомендации, цель должна быть хорошо сформулирована. Размытая формулировка не позвонит понять, достигнута цель или нет. В то же время четко поставленная цель даёт понимание о том, какой конечный результат ожидается, какие шаги нужно предпринять для его достижения, а также понятен признак, по которому можно судить о достижении цели. Существуют разные методологии постановки целей. Самой известной и отвечающей нашим потребностям является методология SMART [11]. Эта методология используется в менеджменте и проектном управлении, но ее можно применить и к другим областям.

SMART — это аббривеатура, она расшифорвывается так: Specific, Measurable, Achievable, Relevant, Time bound. Каждая буква аббривеатуры описывает критерий эффективности поставленных целей, опишем их подробнее.

Specific (Конкретность) Цель должна быть конкретной и ясной, конечный результат описан четко и однозначно. Например «Увеличить словарный запас и навыки чтения художественной литературы в английском языке», а не «Улучшить английский язык».

Measurable (Измеримость) Цель должна быть измеримой. Нужно условиться в чем будет измеряться результат. Необходимо установить признак, по которому можно будет судить о достижении цели. Например «Прочитать “Гарри Поттер и философский камень” в оригинале, понимая 80% текста без словаря» или «Повысить уровень английского языка до B2».

Achievable (Достижимость) Цель должна быть достижимой, ведь от реалистичности достижения цели зависит мотивация того, кто будет ее достигать. Достижимость цели зависит от всех имеющихся ресурсов и ограничений. Ограничениями могут быть финансовое положение, временные ресурсы, доступ к информации, возможность самостоятельно принимать решения, и т.д. Говоря об образовательных целях стоит сказать, что люди часто указывают возраст как преграду для достижения цели.

Relevant (Значимость) Цель должна быть значимой. Её достижение должно быть важно для осуществления более глобальных целей. Значимость - субъективная характеристика. Достижение одного и того же результата может быть важным для одного человека и бесполезным для другого. Цель «Запомнить ТОП-100 английских терминов в области математики и программирования» может быть полезна для человека, которых хочет построить карьеру в области информационных технологий. Но для человека, который просто хочет расширить словарный запас, пользы от достижения подобной цели будет намного меньше, а значит и цель для него менее значима.

Time bound (Ограниченность по времени) Наконец, цель должна иметь временные границы. Должен быть определен финальный срок, превышение которого говорит о невыполнении цели. Например «Выучить 100 слов, относящихся к информационным технологиям, до конца месяца».

Итак, плохо поставленными целями будут, например, такие: «Выучить английский язык», «Разобраться в базах данных». А хорошо сформулированными по SMART будут такие: «Повысить уровень английского языка с B2 до C1 до конца 2020 года, чтобы свободно говорить со знакомыми, которые не знают русского языка», «Пройти курс по SQL на Stepic до 31 августа 2020 года, чтобы стать более востебованным специалистом на рынке труда».

Не все характеристики могут быть определены лишь по тексту цели. К примеру значимость является субъективным понятием. Как уже было сказано, одна и та же цель может быть значимой для одного человека и не важной для другого. Также субъективной характеристикой является достижимость. Нет смысла пытаться предсказать субъективные характеристики.

Для определения наличия объективных характеристик по тексту могут быть использованы разные методы. К примеру, с помощью логистической регрессии можно предсказать вероятность наличия характеристики. Для этих же целей могут быть использованы методы тематического моделирования, которые будут описаны далее. Основной интерес представляет определение степени конкретности образовательной цели. Рассмотрим, какие методы оценки конкретности целей существуют на сегодняшний день.

3 Задачи и методы построения оценок степени специфичности текста

Оценка степени конкретности образовательной цели - слишком узкая задача. Поэтому вместо нее рассмотрим более общую задачу оценки степени специфичности текста. И проанализируем, можно ли применить рассмотренные методы для анализа образовательных целей. Термины «специфичность» и «конкретность» будем использовать как синонимы.

Кажется логичным, что семантическая сложность слова и специфичность - смежные понятия. Вопрос семантической сложности изучен довольно хорошо.

Распространен подход, когда сложность связывается с многозначностью слова. Например в работе [12] обсуждается вопрос нетривиальности зависимости сложности слова от роста полисемии. А в работе [13] сложность определяется по тому, сколько вариантов различных переводов есть в параллельных текстах. Наличие параллельного корпуса текстов позволяет перевести некоторый фрагмент таким образом, что каждое слово фрагмента соответствует ровно одному слову в параллельном тексте. Слова с одним переводом, по мнению авторов, обладают высокой сложностью, а слова с несколькими вариантами переводов - более низкой. Достаточно распространен подход оценивания семантической сложности слова по его длине [14, 15]. Например в работе [14] описывается метрика, оценивающая семантическую сложность текста по длине входящих в него слов. Авторы дали школьникам тексты по биологии и предложили ответить на вопросы по текстам. Сложность текста определялась процентом правильных ответов школьников. В ходе эксперимента авторы выяснили, что более сложными для восприятия оказались тексты, содержащие больший процент длинных слов. Поэтому авторы сделали вывод, что сложность слова пропорциональна его длине.

Вероятно, меру сложности слова можно использовать для оценки специфичности. И все же рассмотрим, какие существуют подходы к оценке непосредственно специфичности слов и текстов.

В этой области весьма распространены статистические методы. К примеру в работе [16] предложены статистические метрики, позволяющие для упорядоченной пары существительных описать специфичность первого слова относительно второго. Авторы предполагают, что специфичные слова встречаются с одними и теми же словами, а не специфичные – с разными. Нужно отметить, что такое предположение очень напоминает гипотезу дистрибутивности из дистрибутивной семантики. Авторы рассматривают несколько метрик, построенных относительно разных модификаторов – частей речи, с которыми встречаются существительные. Для каждого существительного из словаря они подсчитывают нормированное количество различных прилагательных, глаголов и т.д., с которыми встретилось это существительное в корпусе текстов. Для разных модификаторов существуют свои метрики. Также они рассматривают метрики, основанные на энтропии. Лучшей показала себя мерика, описывающая энтропию самого правого модификатора существительного. Авторы использовали большой текстовый корпус, чтобы упорядочить существительные по уровню их специфичности. Они предполагают, что такая семантическая информация может использоваться для автоматического создания или дополнения лексической базы данных, такой как WordNet.

Еще одним интересными примером статистической меры специфичности слова является коэффициент Жуйана. Вот как он описывается во «Введении к частотному словарю современного русского языка» [17]: «Коэффициент Жуйана является лучшим из известных в настоящее время способов измерить, насколько общеупотребительным является слово, или, напротив, насколько оно специфично для отдельных предметных областей.» ... «Коэффициент D отражает равномерность распределения частот в разных сегментах корпуса и вычисляется по следующей формуле:

$$D = 100 \times \left(1 - \frac{\sigma}{\mu\sqrt{n-1}}\right),$$

где n — количество сегментов, на которые разбит корпус, μ — средняя частота слова по всему корпусу (т. е. сумма частот в каждом сегменте, поделенная на n), σ — среднее квадратичное отклонение частоты μ на отдельных сегментах. Для подсчета коэффициента Жуйана корпус разбивается на n равных сегментов (в нашем случае, на 100 частей, размером приблизительно в 90 тыс слов каждый). Тексты в корпусе специально упорядочиваются по функциональным стилям, поэтому тексты одного жанра (например, научные статьи) аккумулируются в пределах небольшого числа сегментов.» ... «Коэффициент вариации $\frac{\sigma}{\mu}$ может принимать значения от 0 (в каждом сегменте частоты одинаковы) до 1 (все словоупотребления встречаются только в одном сегменте). Следовательно, значение D у слов, встречающихся в большинстве документов, близко к 100, а у слов, часто встречающихся лишь в небольшом числе документов, близко к 0.» Логично предположить, что коэффициент Жуйана можно использовать для выделения терминов.

В этом же частотном словаре есть еще одна статистическая метрика — показатель R (range), который «отображает количество сегментов корпуса, в которых встретилось слово».

4 Вероятностное тематическое моделирование

В данном разделе рассматривается применение тематического моделирования для определения наличия SMART характеристик у образовательной цели.

4.1 Задача вероятностного тематического моделирования

Тематическая модель коллекции текстовых документов позволяет определить к каким темам относятся документы и какими словами описываются эти темы.

Пусть D – коллекциятекстовых документов, W – словарь термов коллекции. Термами могут быть слова, словосочетания, термины и т.п. Документ $d \in D$ является последовательностью n_d термов $w_1, \ldots, w_{n_d}$ из словаря W.

Предполагается *гипотеза существования тем*, согласно которой каждое вхождение терма w в документ d связано с некоторой темой t из заданного конечного множества T. Это предположение позволяет рассматривать коллекцию документов как случайную независимую последовательность троек $\{(w_i, d_i, t_i)\}_{i=1}^{n}$ из дискретного распределения $p(w, d, t)$ на конечном вероятностном пространстве $W \times D \times T$.

Также предполагается *гипотеза мешка слов*, согласно которой порядок слов в документе не важен.

Есть *гипотеза условной независимости*, говорящая о том, что появляние терма w в документа d связано с темой t, но не связано с документом d и описывается рапределением $p(w|t)$:

$$p(w, d|t) = p(w|t)$$

Термы w_i и документы d_i являются наблюдаемыми переменными, а темы t_i - латентными.

Запишем вероятностную модель порождения текстовой коллекции документов:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \ d \in D, w \in W, \tag{1}$$

где $p(w|t) = \varphi_{wt}$ – распределение терма $w \in W$ в теме $t \in T$, $p(t|d) = \theta_{td}$ – распределение темы $t \in T$ в документе $d \in D$.

Матрицы $\Phi = (\varphi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$ являются стохастическими. Эта модель описывает процесс порождения текстовой коллекции при известных распределениях $p(w|t)$ и $p(t|d)$.

Тематическое моделирование решает обратную задачу: по известной текстовой коллекции D требуется найти параметры φ_{wt} и θ_{td}, при которых модель порождения данных 1 лучше всего приближает частотные оценки условных вероятностей $p(w|d) = \frac{n_{wd}}{n_d}$.

Таким образом построение тематической модели является задачей стохастического матричного разложения, в которой требуется представить матрицу частот слов в документах $F = (\frac{n_{wd}}{n_d})_{W \times D}$ в виде произведения матриц меньшего размера $\Phi \times \Theta$. Матрица Φ - это матрица распределений слов в темах, она позволяет понять о чем темы, какие в них входят слова. Θ - матрица распределений тем в документах, она позволяет понять о чем документы, какие темы входят в них.

4.2 Аддитивная регуляризация тематических моделей

Найти матрицы Φ и Θ по наблюдаемой коллекции документов D можно с помощью максимизации логарифма правдоподобия текстовой коллекции:

$$\mathcal{L}(\Phi,\Theta) = p(X;\Phi,\Theta) = \prod_{i=1}^{n} p(w_i,d_i) = \prod_{d\in D}\prod_{w\in d} p(w|d)^n_{dw} p(d)^n_{wd} \to \max_{\Phi,\Theta} \tag{2}$$

Прологарифмировав праводподобие, учитывая выражения для $p(w|d)$ и стохастичность матриц, получаем задачу условной оптимизации:

$$\sum_{w\in W}\sum_{d\in D} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{3}$$

$$\sum_{w\in W} \varphi_{wt} = 1, \varphi_{wt} \geq 0, \tag{4}$$

$$\sum_{t\in T} \theta_{td} = 1, \theta_{td} \geq 0 \tag{5}$$

Задача, представленная в таком виде, является некорректно поставленной по Адаму, так как имеет бесконечное количество решений. Если $\Phi\Theta$ - решение задачи стохастического матричного разложения, то и $(\Phi S^{-1})(S\Theta)$ - решение. Это утверждение является справедливым для всех невырожденных матриц S при условии, что ΦS^{-1}, $S\Theta$ - стохастические матрицы.

Методом устронения недоопределенности является регуляризация. Этот метод предполагает прибавление к функции правдоподобия слогаемого, называемого регуляризатором. Регуляризатор описывает предположения о том как должны выглядеть матрицы Φ и Θ, учитывая специфику решаемой задачи. Например обычно ожидается, что темы будут интерпретируемыми, различными, разреженными. Конечная практическая задача также может накладывать дополнительные предположения.

Аддитивная регуляризация тематических моделей заключается в максимизации логарифма правдоподобия взвешанной суммы регуляризаторов:

$$\sum_{w\in W}\sum_{d\in D} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta}, \tag{6}$$

где τ_i – неотрицательный коэффициент регуляризации.

При $R(\Phi,\Theta) \equiv 0$ мы имеем дело с тематической моделью PLSA.

Существует много различных регуляризаторов, рассмотрим основные из них.

Регуляризатор сглаживания / разреживания тем основывается на *гипотезе разреженности*. Она заключается в предположении, что любая тема описывается небольшим числом термов, то есть значительная часть вероятностей φ_{wt} и θ_{td} равна нулю. Разреженность является необходимым условием для интерпретируемости тем.

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \tag{7}$$

При коэффициентах $\alpha_0, \beta_0 > 0$ имеем дело с регуляризатором сглаживания. Он минимизирует кросс-энтропию между столбцами $\vec{\varphi}_t$ и фиксированным распределением $\vec{\beta} = (\beta_w : w \in W)$, а также между столбцами $\vec{\theta}_d$ и распределением $\vec{\alpha} = (\alpha_t : t \in T)$. Применение регуляризатора сглаживания приводит к тому, что распределения $\vec{\varphi}_t$ и $\vec{\theta}_d$ становятся похожи на фиксированные распределения $\vec{\beta}$ и $\vec{\alpha}$ соответственно. Если в качестве фиксированных распределений $\vec{\beta}$ и $\vec{\alpha}$ выбрать равномерные распределения, применение регуляризатора приводит к сглаживанию тем. То есть к тому, что чем более общеупотребимым является слово, тем с большей вероятностью оно войдет в тему. В этом случае модель эквивалентна модели Латентного размещения Дирихле (LDA).

При коэффициентах $\alpha_0, \beta_0 < 0$ получаем регуляризатор разреживания. Он, напротив, максимизирует кросс-энтропию, заставляя распределения $\vec{\varphi}_t$ и $\vec{\theta}_d$ становится непохожими на распределения $\vec{\beta}$ и $\vec{\alpha}$ соответственно. Примерение этого регуляризатора приводит к тому, что чем более общеупотребительным является слово, то есть чем большую частоту оно имеет, тем менее вероятно оно будет входить в отдельные темы. То есть в темы, полученные с применением регуляризатора разреживания, будут вхожить редкие слова.

Предметные и фоновые темы. Чтобы модель была интерпретируемой, каждая темы должна состоять из термов, характерных для одной предметной области и редко встречающихся в других. Для этого матрицы Φ и Θ должны быть разраженными. А темы разделяться на 2 вида - предметные и фоновые. Фоновые темы состоят из общеупотребительных слов. Предметные темы состоят из специфических слов, терминов, характерных для конкретной предметной области. Предметные темы представляют наибольший интерес. Их распределения $p(w|t)$ должны быть разрежены и различны. Распределения $p(t|d)$ тоже должны быть различны, так как ожидается, что предметная тема присутствует в относительно небольшом числе документов.

Регуляризатор декоррелирования тем используется для увелияения различности тем между собой. Чем более различны темы, тем более интерпретируемой и информативной является модель. Для этого минимизируется сумма попарных ковариаций между всеми парами тем:

$$R(\Phi) = -\tau \sum_{t,s \in T} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \tag{8}$$

Применение этого регуляризатора приводит к выделению фоновых тем - слова общей лексики группируются в отдельные темы.

Сглаживание фоновых тем, разреживание предметных и декоррелирование столбцов матрицы Φ позволяет существенно повысить интерпретируемость тем [18–20].

Обучение Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Известно [19, 21], что в таком случае локальный максимум задачи (6) с ограничениями (4), (5) удовлетворяет следующей системе уравнений со вспомогательными переменными

$p_{tdw} = p(t|d,w)$:

$$p_{tdw} = \underset{t\in T}{norm}(\varphi_{wt}\theta_{td}); \tag{9}$$

$$\varphi_{wt} = \underset{w\in W}{norm}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{10}$$

$$\theta_{td} = \underset{t\in T}{norm}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \tag{11}$$

где оператор $norm_{t\in T}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$.

Решение данной системы уравнений с помощью метода простой итерации эквива-лентно EM-алгоритму, где на E-шаге мы пересчитываем значения p_{tdw}, а на M-шаге вычисляем φ_{wt} и θ_{td}.

4.3 Мультимодальные тематические модели

Мультимодальные тематичесие модели описывают данные, содержание метаданные помимо основного текста. Такие метаданные называются модальностями. В качестве токенов модальностей могут выступать слова и словосочетания естественного языка, хэштеги, жанры или категории, ключевые слова, пользователи рекомендательных систем, моменты времени и т.д. Каждая модальность описывается своими токенами и имеет словарь, свой для каждой модальности.

Пусть M - множество модальностей. Модальность $m \in M$ описывается своим словарем W_m. Объединение словарей по всем модальностям $m \in M$ обозначим W. Тематическая модель модальности m:

$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T}\varphi_{wt}\theta_{td}, \tag{12}$$

где $w \in W_m, d \in D$.

Каждой модальности m соответствует стохастическая матрица $\Phi_m = (\varphi_{wt})_{W_m\times T}$. Матрицы Φ_m всех модальностей, записанные в столбец, образуют матрицу Φ. Распределение тем в каждом документе общее для всех модальностей.

Построение мультимодальной тематической модали заключается в максимизации взвешанной суммы логарифма правдоподобия модальностей и регуляризаторов:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W_m} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta}, \tag{13}$$

$$\sum_{w\in W_m}\varphi_{wt} = 1, \varphi_{wt} \geq 0, \sum_{t\in T}\theta_{td} = 1, \theta_{td} \geq 0. \tag{14}$$

Веса τ_m позволяют сбалансировать модальности по их важности.

4.4 Критерии качества тематических моделей

Существуют внешние и внутренние критерии качества тематических моделей.

Внешние критерии появляются от требований конечной решаемой задачи. Ведь часто построение тематитческой модели не является конечной целью, а строится для решения какой-то другой задачи. К примеру, тематическая модель может строиться для классификации документов. Тогда внешними критериями качества модели будут критерии качества классификации. Например площадь под ROC-кривой, precision, recall, accuracy.

Перплексия Наиболее распространённым внутренним критерием качества тематической модели является *перплексия*. Это мера несоответствия модели $p(w|d)$ термам w, которые встречаются в документах d. Она определяется через log-правдоподобие (1) или (13) каждой модальности m:

$$P_m(D;p) = \exp\left(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W_m} n_{dw}\ln p(w|d)\right), \tag{15}$$

где $n_m = \sum\limits_{d\in D}\sum\limits_{w\in W_m} n_{dw}$ - длина коллекции по m-ой модальности.

Чем меньше перплексия, тем лучше модель предсказывает появление термов w в документах d.

Обобщающую способность тематической модели принято оценивать по перплексии, посчитанной по отложенной выборке (hold-out perplexity). Так как перплексия обучающей выборки является заниженной из-за переобучения.

5 Постановка задачи предсказания специфичности документа

.

Дано Выборка $X = (x_i, y_i)_{i=1}^{l}$, x_i - признаковое описание образовательной цели (документа), $y_i \in \{0, 1\}$ - метка класса. Более подробное описание исходных данных может быть найдено в подразделе 7.1.

Найти Алгоритм $a(x) : X \to Y = \{0, 1\}$, который по образовательной цели предсказывает, является ли она конкретной.

Таким образом требуется решить задачу бинарной классификации.

Варианты алгоритма $a(x)$:

baseline:	наивный байесовский классификатор, логистическая регрессия, простая тематическая модель классификации
основное решение:	тематическая модель классификации, учитывающая, что специфичные слова должны быть только в специфичных темах и не должны попадать в не специфичные

Критерий В качестве метрик качества будем использовать площадь под ROC кривой (roc_auc), accuracy, precision и recall.

6 Разработка модели специфичности документов

Основная идея построения модели специфичности документов и слов заключается в следующем. Есть конкретные D_1 и неконкретные D_0 документы (цели). О конкретности документа мы знает из разметки. Слова, входящие в эти докуметы бывают двух типов: W_1 - специфичные, конкретные, узкоспециализированные слова, которые входят в специфичные документы, а также W_0 - неконкретные, общие, абстрактные слова. Разделение словаря W на W_0 и W_1 относится ко всем модальностям, а не только к модальности слов и n-грамм. Наше предположение заключается в том, что неконкретные слова W_0 входят во все темы, в то время как узкоспециализированные, конкретные слова W_1 входят только в специфичные, конкретные темы, описывающие конкретные цели.

Это соображение моделируется разделением тем $T = T_0 \cup T_1$ на конкретные T_1 и неконкретные T_0.

Конкретные слова W_1, входящие в конкретные документы D_1 должны обладающие высокой степенью тематичности. Например, это могут быть слова, которые попадают в топ термов, описывающих тему. *Тематичное слово* w имеет высокую вероятность $p(t|w)$ для какой-то темы t. Мы ожидаем, что специфичный документ $d \in D_1$ состоит из слов, некоторые из которых имеют высокую тематичность, то есть высокую веростность $p(t|w)$ для какой-то темы t. В тот момент как у не специфичного документа $d \in D_0$ все слова имеют низкую вероятность $p(t|w)$ для всех тем $t \in T$, так как описываются общеупотребительными словами. Распределения $p(t|w)$ у слов такого документа близко к равномерному.

Будем обозначать степень специфичности документа $d \in D$ через $s_d \in [0, 1]$, степень специфичности слова $w \in W$ через $s_w \in [0, 1]$.

К примеру, цель «Хочу быть счастливым.» обладает низкой специфичностью s_d, так как состоит из общеупотребительных слов. «Хочу разбираться в нейронных сетях и тематическом моделировании.» - пример конкретного документа, в котором есть тематичные слова.

Таким образом мы хотим построить модель, которая бы учитывала разделение слов, документов и тем на конкретные и неконкретные. Вид такой модели представлен на рис. 1. Обсудим модель более побробно.

Матрица Φ приобретает блочный вид и состоит из 4 блоков. Φ_{00} описывает распределения неконкретных слов из W_0 в некокретных темах T_0. Распределения неконкретных слов должно быть близко к равномерному. Матрица $\Phi_1 0$ полностью состоит из нулей. Она описывает распределение конкретных слов W_1 в неконкретных темах. Согласно нашему предположению, конкретные слова должны входить только в конкретные тем T_1 и не входить в T_0. Именно поэтому эта подматрица нулевая. Подматрица Φ_{01} описывает распределение общих, неконкретных слов в конкретных темах. Подматрица Φ_{11} представляет наибольший интерес. Она описывает распределения конкретных, узкоспециализированных слов в конкретных темах. Эта матрица должна быть разреженной и декоррелированной - мы ожидаем, что конкретная тема описывается малым числом слов, и слова, описывающие одну тему не похожи на слова другой темы. Такой вид матрицы Φ относится ко всем модальностям. Блок матрицы Φ, описывающий модальность меток классов, принимает тривиальный вид: она состоит из 0 и 1. Напомним, что матрица Φ стохастическая по столбцам для

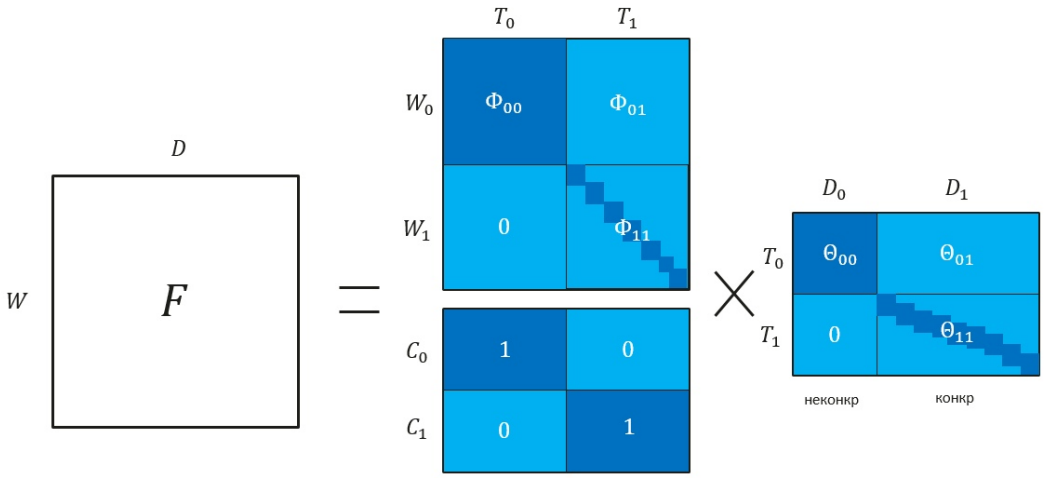

Рис. 1. Вид тематической модели специфичности документов и слов
F - матрица частот слов в документах, Φ - матрица распределений слов по темам, Θ - матрица распределений тем по документам.

каждой модальности. В словаре модальности меток класса всего два слова: токен нулевого и первого классов. Поэтому подматрица Φ_{00} в данном случае - это вектор, состоящий из единиц, длина его равна количеству неконкретных тем, Φ - вектор, состоящий из нулей, аналогичной длины. Φ_{01} - нулевой вектор, длина которого равна количеству конкретных тем, а Φ_{11} - вектор, состоящий из единиц, его длина равна количеству специфичных тем.

Матрица Θ также имеет блочный вид. Соображения, лежащие в основе её построения, аналогичны соображениям для построения матрицы Φ. Наибольший интерес для нас представляет подматрица Θ_{11}, так как именно она описывает распределение конкретных тем в конкретных документах. Необходимо чтобы $\Theta_1 1$ была разреженной и декоррелированной, так как ожидается, что документ относится к небольшому числу тем, и темы не похожи между собой.

Специфичность s_d документа d будем вычислять как вероятность относиться к конкретной теме $T_1 \subset T$:

$$s_d = \sum_{t \in T_1} p(t|d) = \sum_{t \in T_1} \theta_{td} \tag{16}$$

Специфичность s_w слова w также будем вычислять как вероятность относиться к конкретной теме $T_1 \subset T$:

$$s_w = \sum_{t \in T_1} p(t|w) = \sum_{t \in T_1} \varphi_{wt} \frac{n_t}{n_w} \tag{17}$$

О том как построить модель с матрицами описанного вида написано в разделе 7.4, описывающем реализацию этой модели.

7 Реализация модели специфичности документов

7.1 Описание исходных данных

Для начала разберемся с тем, что из себя представляют исходные данные. Для сбора данных была составлена анкета, в которой людям предлогалось описать свои цели и ответить на разные вопросы о себе и целях.

В анкетировании приняли участие около 11 тысяч человек. Каждый человек написал по 3 цели. То есть имелось около 33 тысяч различных целей. В таблице 1 кратко описана информация, собираемая анкетированием.

Вопросы анкеты	**Варианты ответов**
Формулировка цели	
Моя первая цель	*Описание цели*
Ответы на вопросы про цель	
Представляете ли Вы себе результат по этой цели?	Да, четко
	Да, нечетко
	Нет
К какой тематической области относится Ваша цель?	Математика и IT
	Естественные науки
	Творчество и создание нового
	...
Каким может быть первый шаг для достижения цели?	Не знаю, с чего начать
	Описание первого шага
Сколько времени может занять достижение цели?	Не знаю
	Нет жестких сроков
	Описание срока
...	...
Социально-демографическая информация	
Возраст	*Возраст*
В каком населенном пункте Вы проживаете?	*Населенный пункт*
Какое у Вас образование?	Среднее общее образование
	Высшее образование
	Ученая степень
	...
Какое Ваше основное занятие на данный момент?	Учусь в школе, гимназии
	Учусь в вузе
	Работаю в организации
	Предприниматель
...	...

Таблица 1. Информация из анкет
Таблица кратко описывает данные, собираемые анкетой. Варианты ответов, *выделенные курсивом*, говорят о том, что этот ответ пользователи писани на естественном языке.

Далее было необходимо разметить данные. Размечать все 33 тысячи дорого и долго, поэтому обработано было меньшее количество. Было выявлено, что наиболее подробно люди описывали первую цель из трёх, а потом, видимо уставали, и давали менее подробные и качественные ответы. Поэтому было решено размечать только первую цель для каждого человека. Для разметки были привлечены 3 эксперта-разметчика. Они разметили 6 тысяч первых целей. Таким образом имелось 18 тысяч уникальных разметок. Таблица 2 кратко описывает некоторую информацию, полученную в ходе разметки данных.

Признак	**Краткое описание**	**Область значений**
Specific	Конкретность	{0, 1}
Achievable	Достижимость	{0, 1}
Time_bound	Ограниченность по времени	{0, 1}
Education	Цель относится к образованию	{0, 1}
Unambiguity	Цель не разделяется на несколько целей	{0, 1}
...	...	...

Таблица 2. Информация из разметки
Таблица описывает некоторую информацию, полученную благодаря разметке данных. Также были размечены аттракторы целей и другая техническая информация.

Итак, было собрано 6000 уникальных разметок. После удаления шумовых и дублирующихся данных, а также после того, как были оставлены только цели, для которых хотя бы один разметчик указал, что цель относится к образованию, из 6000 целей осталось 5153 цели.

Из 5153 человек, заполнивших анкету, 1178 женщин, 887 мужчин, предпочли не указывать пол 3088 человек. Высшее образование имеют 771 человек, два и более высших 305 человек, среднее профессиональное - 60, неоконченное высшее - 271, среднее общее образование - 170, ученая степень имеется у 369 человек, основное общее образование - 42 человека, категорию «другое» выборали 92 человека, предпочли не указывать информацию 3073 человека. В сфере образования занято 792 человека, с IT и программированием связано 314 человек, следующей по популярности сферой занятости являются органы управления - 159 человка, затем сфера науки и культуры - 112 человек, в каждой из других различных сфер занято менее 100 человек в каждой, не стали указывать информацию 3105 человек.

7.2 Анализ и предобработка исходных данных

Для предобработки текстов целей были применены токенизация, приведение слов к нижнему регистру, приведение к начальной форме (лемматизация), удаление стоп слов. Для каждой цели был выделен набор биграмм.

На графиках рисунка 2 представлены распределения длин целей до и после предобработки. Также длину целей описывает таблица 3. Проанализировав графики и таблицу можно сделать вывод, что описывая свои цели саморазвития, люди не дают воли своей фантазии - они сипользуют очень мало слов для описания целей. Большинство людей описывает свои цели менее чем 10 словами. 50% людей использует

всего лишь 3 слова для описания цели! 75% используют 6 слов, а после предобработки такие цели описываются 4 словами. Предобработка привела к тому, что некоторые цели стали иметь нулевую длину. Изначально имелось 3957 уникальных целей из 5153, после предобработки уникальных целей осталось 3521.

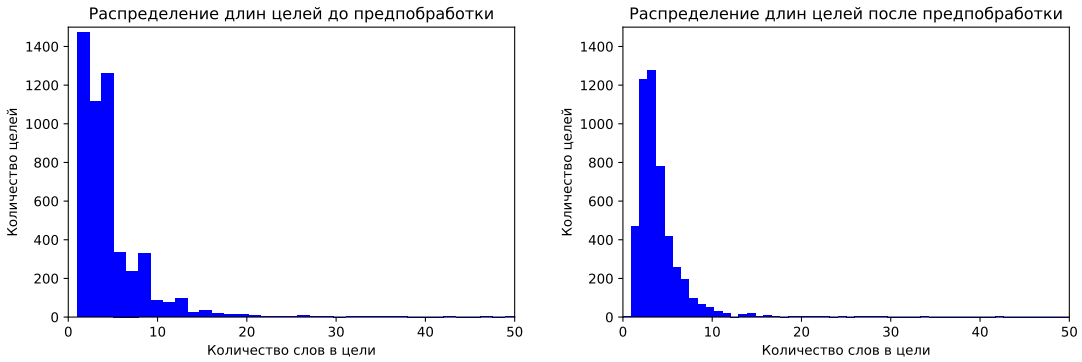

Рис. 2. Гистограммы длин целей до и после предобработки
Графики распределения длин слов в целях до предобработки и после. Можно увидеть, что для описания целей люди использовали преимущественно менее 10 слов.

Без применения методов машинного обучения ясно, что цель, описанная 3 словами едва ли является конкретной в терминах SMART. А таких целей большинство. В трети случаев разметчики оценивали конкретность ориентируясь только на текстовое описание цели. В оставшихся двух третях случаев разметчики также видели ответы респондентов на вопросы анкеты. Учитывая все это, логично предположить, что размеченная выборка целей окажется несбалансированной по классам, и класс неконкретных целей окажется больше, чем класс конкретных целей. Однако на практике это оказалось не совсем так. Один разметчик отнес 3158 целей к неконкретным, 1955 к конкретным. Другой 2637 к неконкретным, 2516 к конкретным. Третий 2813 к неконкретным, 2340 к конкретным.

	до	после
количество, шт	5153	4955
min	1	0
25%	2	2
50%	3	3
75%	6	4
max	220	139

Таблица 3. Распределение длин целей до и после предобработки

В таблице 4 приведены примеры некоторых целей, которые были единогласно отнесены разметчиками к конкретными или неконкретным. При этом они оценивали цели только по тектовому описанию, то есть без информации из анкет.

Можно увидеть, что неконкретные цели являются более абстрактными и общими, часто видно, что люди сами точно не знают чего хотят. Характерным признаком конкретных целей является наличие конечного результата, достижение которого ожидается по достижении цели. В то же время можно заметить и некоторую нелогичность. 5 неконкретная цель и 4 конкретная крайне похожи по смыслу, однако одну с уверенностью отнесли к одному классу, а другую - к другому. То же самое можно сказаь про 1 неконкретную и 9 конкретную цели. 10 конкретная цель ка-

жется действительно конкретной с одной стороны - указана точная научная область, компьютерные науки. С другой стороны это огромная область, и человек хочет разобраться в ней в целом. А если сравнить эту цель со 2 конкретной и 5 конкретной, встаёт вопрос о том, почему эти три цели отнесены к одному классу.

	Неконкретные цели
1	Получить нормальную работу
2	Самоопределение
3	Достижение желаемых целей
4	Найти свое место (любимую профессию, чувство гармонии)
5	Более профессионально изучение компьютера
6	Общий уровень культуры и развитие
7	Внутренняя удовлетворенность, уверенность в себе, развитие
8	Идти в ногу с развитием общества
9	Получить пользу от самообразования
10	Расширить рамки мировоззрения
	Конкретные цели
1	Повышение результативности труда
2	Развитие профессиональных качеств и навыков
3	Получить образование
4	Повысить уровень знаний возможности компьютера
5	Углубление профессиональных знаний
6	Стать проф фотографом
7	Получить степень магистра
8	Закончить вуз с красным дипломом
9	Быть конкурентноспособным на рынке труда
10	Хочу стать хорошим специалистом в компьютерных науках в целом

Таблица 4. Примеры конкретных и неконкретных целей
В таблице приведены примеры целей, которые разметчики единогласно посчитали конкретными и неконкретными. Пунктуация авторов сохранена.

Предыдущие рассуждения приводят к выводу о том, что собранные данные не очень хорошие.

Описывая данные, стоит также сказать, что в словаре, составленном по всей коллекции документов, 3533 слова и словосочетания, из них 113 ключевых слов используются для описания данных из анкет.

7.3 Базовые модели

В качестве базового решения можно рассмотреть три модели: наивный байесовский классификатор, логистическая регрессия и тематическая модель классификации.

Выборка была разделена на данные для обучения и теста в соотношении 0.7 к 0.3 со стратификацией. Стратификация позволяет получить в тестовой выборке такое же соотношение классов как в данных для обучения.

Наивный байесовский классификатор В качестве признакового описания цели использовался вектор счетчиков слов, входящих в документ. Таким образом были получены дискретные разреженные признаки. Логично предположить, что распределение признаков похоже на мультиномиальное распределение и использовать его для восстановления распределений признаков.

Логистическая регрессия В логистической регрессии в качестве векторов признаков целей также использовались вектора счестчиков вхождений слов в документ. Логистическая регрессия была построена с применением L1-регуляризации.

Тематическая модель классификации Мультимодальная тематическая модель позволяет более эффективно использовать имеющиеся данные. Цели описываются тремя модальностями: @ngram - модальность n-грамм, включающая уни- и биграммы, модальность @info, описывающая информацию, полученную из анкет и модальность @specificity меток класса.

Были подобраны веса модальностей, дающие лучшее качество. Довольно интересным является факт того, что оптимальный вес модальности информации @info оказался меньше, чем модальности n-грамм. Дело в том, что в половине случаев цель описывается все лишь 3 n-граммами, в то время как токенов модальности @info существенно больше, так как в анкете было много вопросов. Было проведено разреживание матрицы Φ по всем модальностям, разреживание матрицы Θ, а также декорреляция матрицы Φ. Для каждого регуляризатора были подобраны оптимальные коэффициенты.

Построение тематических моделей осуществлялось с использованием библиотеки тематического моделирования BigARTM [21] и TopicNet.

Результаты, которые показали все три базовых модели приведены в таблице 5, на графике 3 изображены ROC-кривые, соответствующие этим решениям.

	MultinomialNB	LogRegression	ARTM_classification
roc_auc	0.843594	0.843779	0.787452
precision	0.788851	0.792321	0.705634
recall	0.702256	0.682707	0.753383
accuracy	0.782784	0.778077	0.749159

Таблица 5. Результаты базовых решений

Можно увидеть, что наивный байесовский классификатор и логистическая регрессия с L1-регуляризацией дают примерно одинаковое качество классификации. Тематическая модель классификации дает более плохой результат. Более тонкая настройка параметров модели скорее всего позволила бы повысить качество. Благодаря полученным результатам можно установить порог, с которым будет сравниваться качество основного решения. Основное решение должно давать качество, лучшее чем 0.85 по метрике roc_auc площади под ROC-кривой.

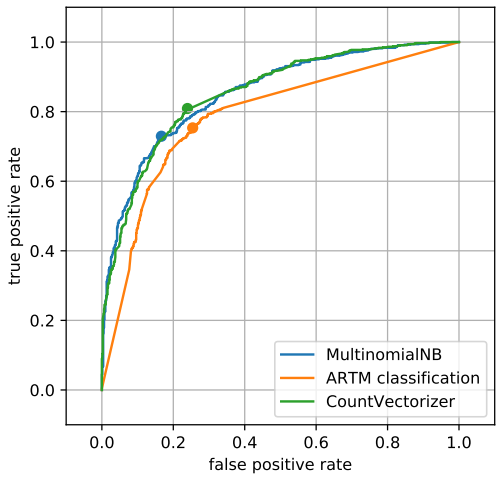

Рис. 3. ROC-кривые для базовых решений

7.4 Реализация модели специфичности документов

Основным решением является тематическая модель специфичности документов, учитывающая разделение слов, документов и тем на конкретные и неконкретные. В разделе 6 описаны соображения, лежащие в основе модели и ожидаемые свойства модели. Φ и Θ должны иметь блочный вид. Рисунок 1 демонстрирует как именно должны выглядеть эти матрицы. В этом разделе будет описано как построить такую модель.

Для начала стоит сказать, что модель будет строиться с использованием трёх модальностей: @ngram - модальность слов и словосочетаний, @info - модальность информации, полученной из анкет и @specificity - модальность меток классов. Модальность @ngram состоит из 3418 токенов, @info из 113 токентов, @specificity из 2 токенов: 0 означает неконкретность, 1 - конкретность. Всего в матрице Φ 3533 токена по всем модальностям.

Рассмотрим как построить подматрицу Φ_{00}, описывающую распределения неконкретных токенов W_0 в неконкретных темах T_0. Построим вспомогательную тематическую модель. А именно тематическую модель неконкретных целей, обученную на неконкретных документах D_0. После того как модель построена, инициализируем подматрицу Φ_{00} нашей итоговой основной модели матрицей Φ из модели неконкретных целей. Поступим таким образом со всеми модальностями, для модальности @specificity подматрица Φ_{00} принимает тривиальный вид и состоит из единиц.

Подматрицу Φ_{10} инициализируем нулями. Φ_{01} и Φ_{11} инициализируем случайными числами из равномерного распределения на $[0, 1]$ таким образом, чтобы вероятности конкретных слов W_1 были в 100 раз выше вероятностей неконкретных. Это необходимо, так как мы хотим, чтобы в конкретных темах была высокая вероятность у узкоспециализированных, предметных слов и значительно более низкая у общеупотребительных. Нормируем Φ_{01} и Φ_{11} таким образом, чтобы сумма элементов в стобцах матрицы Φ равнялась единице, то есть чтобы матрица продолжала быть стохастической. Матрицу Θ инициализируем случайным образом - числами из равномерного распределения на $[0, 1]$, соблюдая условие стохастичности.

Для того, чтобы матрицы преобрели желаемый вид как на рис. 1, необходимо добавить соответствующие регуляризаторы и провести тонкую насройку коэффициентов влияния регуляризаторов на модель. Но добавлять их имеет смысл только после того как модель сошлась без регуляризации. За исключением регуляризатора сглаживания, который можно использовать сразу, он не импортит модель. Сначала подберем параметры для модели с единственным регуляризатором сглаживания, затем перейдем к настройке основных регуляризаторов.

Вот какие параметры можно настраивать для текущей модели: веса модальностей @ngram, @info и @specificity, количество специфичных и общих тем, а также коэффициент сглаживания подматрицы Φ_{00}, о нём поговорим подробнее.

Матрица Φ_{00} взята из модели неспецифичных целей. Она описывает распределения неконкретных слов в неконкретных целях. Она уже выглядит так как надо, нам не зачем изменять её. Для того, чтобы зафиксировать состояние матрицы Φ_{00}, воспользуемся регуляризатором сглаживания.

Была получена модель с оптимальными параметрами, дающая наилучший результат по качеству классификации. Лучшая модель выбиралась по метрике roc_auc площади под ROC-кривой. Здесь с далее демонстрация настройки оптимальный параметров осуществляется следующим образом: все параметры модели оптимальны и зафиксированы и изменяется только один параметр, демонстрация которого производится. Демонстрация графиков и таблиц для подбора всех параметров кажется излишней, поэтому будут продемонстрированы только некоторые параметры.

Существенный вклад в качество модели внесла настройка весов модальностей. ROC-кривые для моделей с разными весами и таблицы, более детально описывающие получаемые метрики качества, могут быть найдены на рис. 4 и таб. 6 для @ngram, рис. 5 и таб. 7 для @info, рис. 6 и таб. 8 для @specificity.

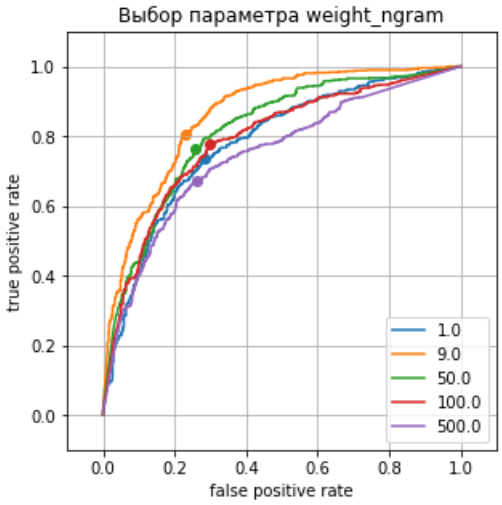

Рис. 4. Влияние веса модальности @ngram на качество модели

Таблица 6. Влияние веса модальности @ngram накачество модели

w_ngram	roc_auc	precision	recall	accuracy
1.0	0.7792	0.6772	0.7353	0.7249
9.0	**0.8621**	**0.7399**	**0.8045**	**0.7861**
50.0	0.8134	0.7055	0.7639	0.7518
100.0	0.7883	0.6789	0.7759	0.7357
500.0	0.7450	0.6762	0.6721	0.7094

Весьма интересными параметрами являются количество специфичных и неспецифичных тем. Влияние числа тем на качество модели отражают таблица 9 и рисунок 7 для неспецифичных тем и рисунов 8 для специфичных тем. Видно, что наибольшего

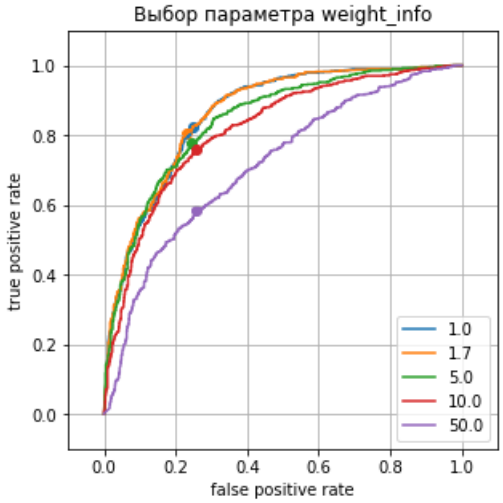

Рис. 5. Влияние веса модальности @info на качество модели

Таблица 7. Влияние веса модальности @info накачество модели

w_info	roc_auc	precision	recall	accuracy
1.0	0.8601	0.7296	0.8240	0.7848
1.7	**0.8621**	**0.7399**	**0.8045**	**0.7861**
5.0	0.8424	0.7190	0.7774	0.7646
10.0	0.8154	0.7047	0.7609	0.7505
50.0	0.7139	0.6467	0.5864	0.6718

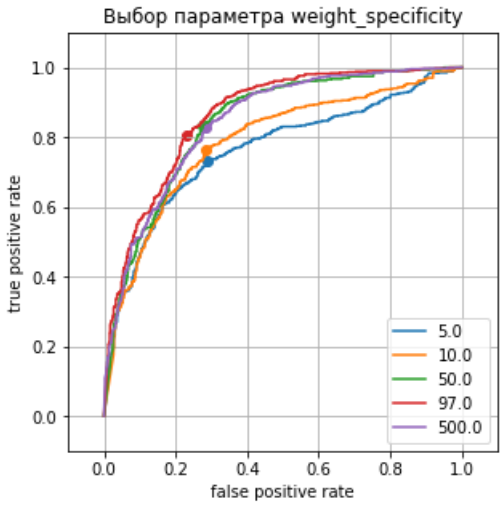

Рис. 6. Влияние веса модальности @specificity на качество модели

Таблица 8. Влияние веса модальности @specificity накачество модели

w_sp	roc_auc	precision	recall	accuracy
5.0	0.7609	0.6721	0.7308	0.7202
10.0	0.7862	0.6860	0.7624	0.7377
50.0	0.8413	0.7080	0.8315	0.7713
97.0	**0.8621**	**0.7399**	**0.8045**	**0.7861**
500.0	0.8435	0.7028	0.8285	0.7666

качества модель достигает при 3 неконкретрых темах. С количеством специфичных тем ситуация интереснее. Видно, что чем больше тем, тем хуже становится качество. А между 4 и 6 специфичными темам разница не большая, качество изменяется не сильно. 6 тем было выбрано потому что при таком значении параметра модель действительно даёт лучшее качество, но не только по этой причине. При 4 темах качество примерно такое же, однако стоит помнить, что мы имеем дело с тематической моделью, и можем ориентироваться не только на качество классификации, но и на перплексию. При 4 темах перплексия получается больше, чем при 6, значит 6 тем - лучший вариант. Если внимательно посмотреть на документы, которые модель относит к конкретным темам, становится понятно, что в коллекции точно больше 4 тем.

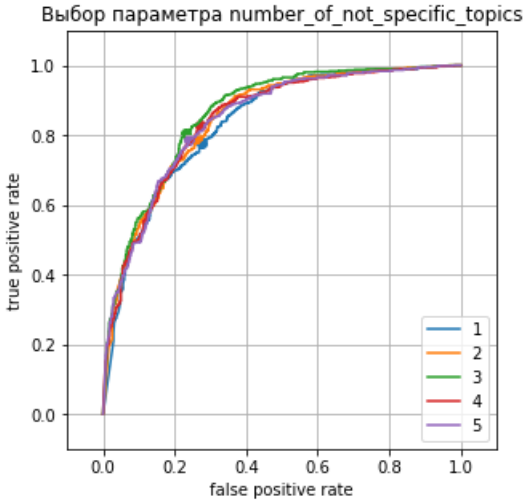

Рис. 7. Зависимость качества модели от количества неспецифичных тем

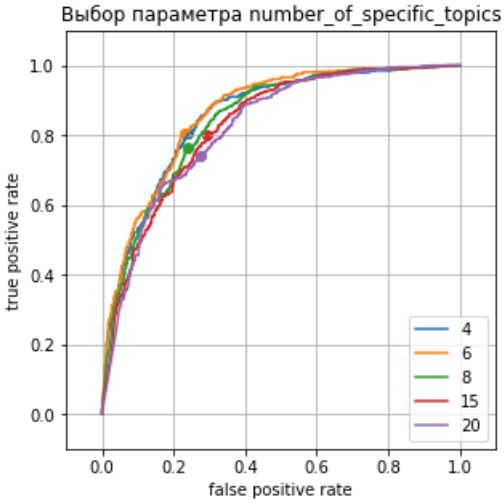

Рис. 8. Зависимость качества модели от количества специфичных тем

numer_not_sp_topics	roc_auc	precision	recall	accuracy
1	0.8354	0.6948	0.7774	0.7478
2	0.8474	0.7075	0.7894	0.7599
3	**0.8621**	**0.7399**	**0.8045**	**0.7861**
4	0.8458	0.7118	0.8285	0.7733
5	0.8460	0.7253	0.7864	0.7713

Таблица 9. Зависимость качества модели от количества неспецифичных тем

Более подробное о выборе числа конкретных тем при учете документов, относимых к получаемым темам, можно прочитать в разделе 8.

После установления оптимальных параметров для модели можно добавлять регуляризаторы и искать коэффициенты для них.

Необходимо, чтобы среди элементов подматрицы Φ_{11} преобладали нули. Это нужно для того, чтобы темы описывались небольшим количеством слов. Также требуется, чтобы темы были попарно непохожи. Для достижения этих свойств введём регуляризаторы разреживания и декоррелирования подматрицы Φ_{11} по всем модальностям. Зависимость качества модели от коэффициента разреживания Φ_{11} по модальности @ngram отражена на рис. 9, изменения метрик качества можно увидеть в таблице 10. Демонстрация влияния именно этого коэффициента на качество модели выбрана как наиболее информативная и наглядная. Разреживание Φ_{11} по модальности @ngram наиболее сильно отразилось на качестве итоговой модели.

Также необходимо сделать столбцы Θ_{11} разреженными и декоррелированными, так как предполагается, что каждый документ относится к небольшому числу тем, а темы попарно различны. Для этого воспользуемся регуляризаторами разреживания и декорреляции матрицы Θ_{11}.

Значения метрик качества классификации, которого удалось достичь при реали-

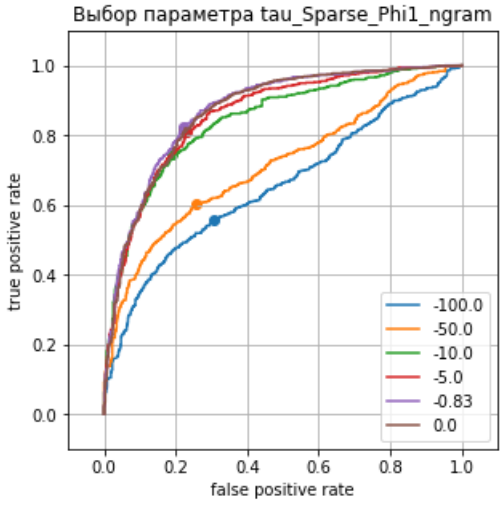

Рис. 9. Зависимость качества модели от коэффициента разреживания Φ_{11} по модальности @ngram

Таблица 10. Зависимость качества модели от коэффициента разреживания Φ_{11} по модальности @ngram

tau	roc_auc	precision	recall	accuracy
-100.00	0.6516	0.5942	0.5548	0.6314
-50.00	0.7106	0.6535	0.6015	0.6792
-10.00	0.8410	0.7592	0.7443	0.7800
-5.00	0.8576	0.7390	0.8090	0.7868
-0.83	**0.8716**	**0.7520**	**0.8255**	**0.8002**
0.00	0.8664	0.7458	0.8075	0.7908

зации основной модели, представлены в таблице 11. Для сравнения также приведены значения метрик качества базовых решений.

	ARTM_main_model	MultinomialNB	LogRegression	ARTM_classification
roc_auc	0.871657	0.843594	0.843779	0.787452
precision	0.752055	0.788851	0.792321	0.705634
recall	0.825564	0.702256	0.682707	0.753383
accuracy	0.800269	0.782784	0.778077	0.749159

Таблица 11. Сравнение результатов качества разных решений

8 Описание и анализ результатов

8.1 Интерпретация тем

Итак, была построена тематическая модель, позволяющая осуществлять бинарную классификацию - предсказывать, является образовательная цель конкретной или нет. Для модели было установлено количество специфичных и абстрактных тем. Абстрактные темы не представляют большого интереса, но все таки приведем некоторую информацию, которую построенная модель позволяет о них получить. В таблице 12 приведены наиболее характерные слова для каждой темы, а в таблице - наиболее характерные абстрактные цели для каждой темы.

topic_0	topic_1	topic_2
знание	новое	повышение
навык	стать	уровень
получение	свой	новый
развитие	хороший	приобретение
новый	расширение	умение
получить	узнать	рост
новое	кругозор	большой
компетенция	человек	работа
новый_знание	узнать_новое	знание_умение
получение_новый	познание	повышение_уровень

Таблица 12. Наиболее характерные слова для неконкретных тем

topic_0	Получить дополнительные знания и навыки, которые пригодятся в дальнейшем.
	Получение необходимых знаний для решения задач, стоящих перед мной..
	Получение навыков и знаний для достижения материального благополучия.
	получить новые знания, которые пригодятся мне в дальнейшем.
	Получить новые знания и навыки в области преподаваемых дисциплин.
topic_1	Хочу лучше понимать людей, уметь контактировать не только с айтишниками.
	Стать более эрудированной личностью, интересным собеседником.
	Развивать мышление, стать умнее и быть более гибкой сознанием..
	Понимание процессов изменений во внешней среде..
	Понять, что я делаю лучше всего и совершенствоваться в этом.
topic_2	Чем больше знаю я, тем больше знают мои ученики.
	Повышение личной эффективности и карьерный рост.
	Дойти до уровня при котором я смогу устроиться на работу, которую хочу.
	повышение уровня своей эрудиции, правовой и общей культуры;.
	Повышение уровни сложности проектов, в которых я могу участвовать.

Таблица 13. Наиболее характерные цели для неконкретных тем

Действительно, цели, которые были отнесены к неконкретным, довольно размыто описывают конечный результат, который люди ожидают достичь.

Рассмотрим, как выглядят конкретные темы. В таблице 14 представлены наиболее характерные слова для каждой из специфичных тем, таблица 15 описывает наиболее характерные цели, соответствующие каждой конкретной теме.

topic_3	закончить	**topic_6**	язык
	университет		английский
	учёба		английский_язык
	окончить		выучить
	закончить_университет		программирование
	научиться		изучить
	вуз		уровень
	владение		иностранный
	кандидатский		иностранный_язык
	магистратура		выучить_английский
topic_4	стать	**topic_7**	диссертация
	свой		работа
	квалификация		рынок
	бизнес		защитить
	повышение		специалист
	карьерный		докторский
	цифровой		труд
	рост		рынок_труд
	повышение_квалификация		стать
	карьерный_рост		докторский_диссертация
topic_5	профессиональный	**topic_8**	получить
	уровень		образование
	повышение		высокий
	компетенция		диплом
	повысить		высокий_образование
	навык		стать
	развитие		получить_высокий
	свой		получить_диплом
	повышение_профессиональный		область
	профессиональный_уровень		степень

Таблица 14. Наиболее характерные слова для каждой из конкретных тем.

Видно, что конкретные темы действительно отличаются друг от неконкретных. Здесь наиболее вероятными являются цели для которых понятен конечный результат, это действительно конкретные темы. Также видно, что документы и слова, относящиеся к одной теме - действительно об одном и том же. Таблица 16 содержит названия конкретных тем. Они были сформулированы исходя из того какие документы входят в эти темы.

В разделе 7.4 говорилось о выборе количества конкретных и неконкретных тем. Остановимся на этом моменте ещё раз. Конечной задачей было построить модель с высоким качеством бинарной классификации. Поэтому было выбрано такое количество тем, которое дает высокое качество классификации и адекватную перплексию. При этом если смотреть на полученную модель только как на тематическую, забыв про задачу классификации, то можно увидеть, что эта модель не идеальна. Если бы конечная задача была - построить хорошую тематическую модель, стоило бы или сделать больше конкретных тем, или построить иерархическую тематическую модель.

topic_3	Получить высшее экономическое образование.
	Закончить медицинский вуз.
	Закончить вуз с красным дипломом.
	Получить два высших образования.
	С отличием окончить вуз.
topic_4	Стать грамотным специалистом в своей сфере.
	Актуальность. Сейчас изучаю методологию преподавания финансовой грамотности.
	Наращивать экспертность в технологизации исследовательского мышления.
	На данный момент сдать международный экзамен.
	Личная эффективность на работе.
topic_5	Развитие навыков управления и проектной деятельности.
	Научиться формировать и управлять эффективной командой.
	Развить в себе компетенции управляющего, чтобы возглавить отдел в Ростехе.
	Развить компетенции проектного менеджмента.
	развитие профессиональных управленческих лидерских качеств.
topic_6	Изучение английского языка до уровня Effective operational proficiency or advanced.
	Изучить английский язык до уровня не ниже upper intermediate.
	владеть иностранным языком на уровне, позволяющем учиться в США.
	говорить на английском языке на уровне носителей англоязычных стран.
	выучить английский и китайский языки.
topic_7	Подготовка и защита докторской диссертации.
	Проведение исследования, подготовка и защита докторской диссертации.
	Защитить докторскую диссертацию не позднее 2022 года.
	Написать докторскую диссертацию по политическим наукам.
	Защита докторской диссертации до 2035 года.
topic_8	Получить ученую степень.
	Получить степень магистра.
	Получить уч.степень доктора педагогических наук.
	Получение степени MBA.
	Получить высшую степень преподавателя (доктор информационных наук).

Таблица 15. Наиболее характерные цели для каждой из конкретных тем.

Такую, у которой есть родительские темы, их не много, и каждая родительская тема разделяется на несколько дочерних.

Так как тема **topic_5** про управленческие навыки на самом деле является более общей. Это тема про разные гибкие навыки, а цели про развитие управленческих качеств просто попали в топ-5 наиболее характерных для этой темы. При этом в эту же тему попали цели про тайм менеджмент, публичные выступления и личную эффективность, то есть про развитие soft skills. Аналогично дело обстоит с темой **topic_4** про развитие профессиональных компетенций. В нее вошли цели про профессиональное развитие в большом количестве разных сфер, профессий. Темы, характеризующие развитие в конкретных профессиях могли бы быть дочерними темами в иерархической модели. А вот с темой **topic_7** про защиту докторской диссертации все уже хорошо, а также с темой про получением высшего образования **topic_3**. Их нет смысла делить на более мелкие темы. Если только не считать, что под высшим образованием люди считают и бакалавриат, и магистратуру, и специалитет. Тему

topic_3	Получить высшее образование
topic_4	Стать грамотным специалистом в своей сфере
topic_5	Развить профессиональные управленческие качества
topic_6	Изучить иностранный язык
topic_7	Защитить докторскую диссертацию
topic_8	Получить ученую степень

Таблица 16. Названия тем

topic_6 про изучение иностранных языков иностранным языков тоже тоже едва ли стоит дробить. Правда, при желании и ее можно разделить на конкретные языки. Интересный факт: в основном люди хотят выучить английский и китайский.

8.2 Графическая визуалиазция «темы–специфичность»

Для всех документов была подсчитана степень специфичности. Вычисления проводились по формуле 16, степень специфичности для документа определеяется как вероятность относится к любой из конкретных тем. На рисунке 10 представлена специфичность всех документов. Видно, что документы, относящиеся к специфичным темам действительно имеют высокую конкретность, а к неспецифичным - низкую.

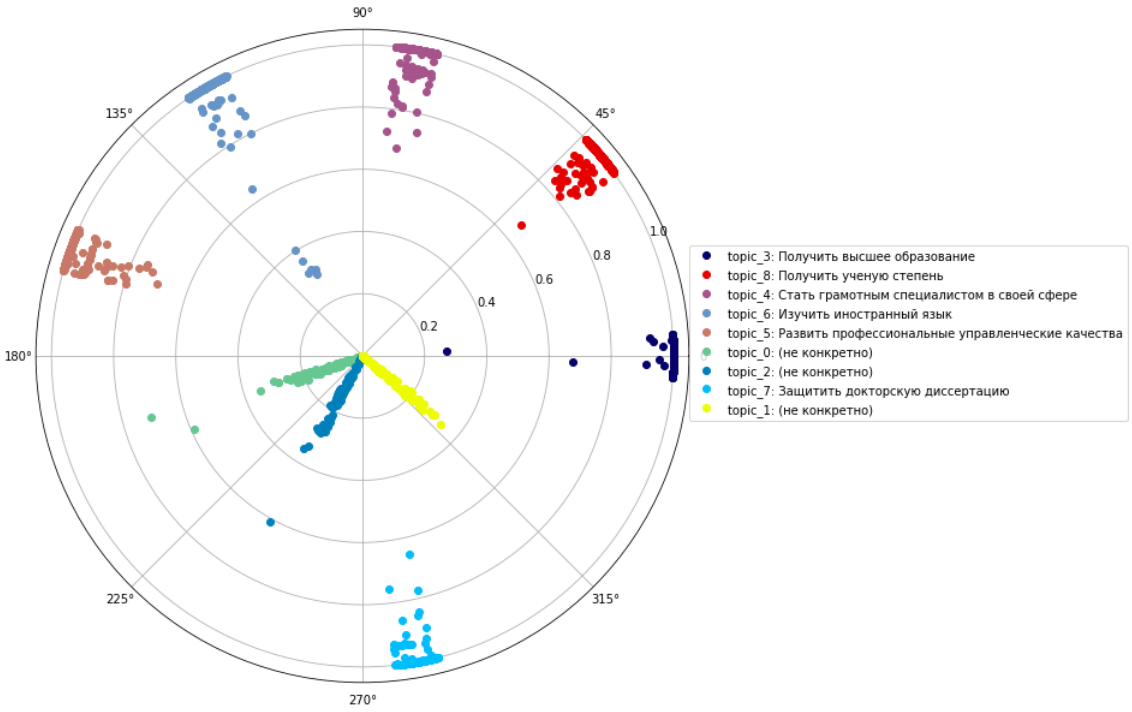

Рис. 10. Специфичность всех документов

На следующих графиках отдельно представлены специфичность неконкретных (рис. 11) и конкретных (рис. 12) документов.

Видно, что четко выделаются специфичные и неспецифичные документы. Сначала была построена тематическая модель неспецифичных целей по d0, она помогла выделить неспецифичные темы **topic_0, topic_1, topic_2**. Темы **topic_3, . . . ,**

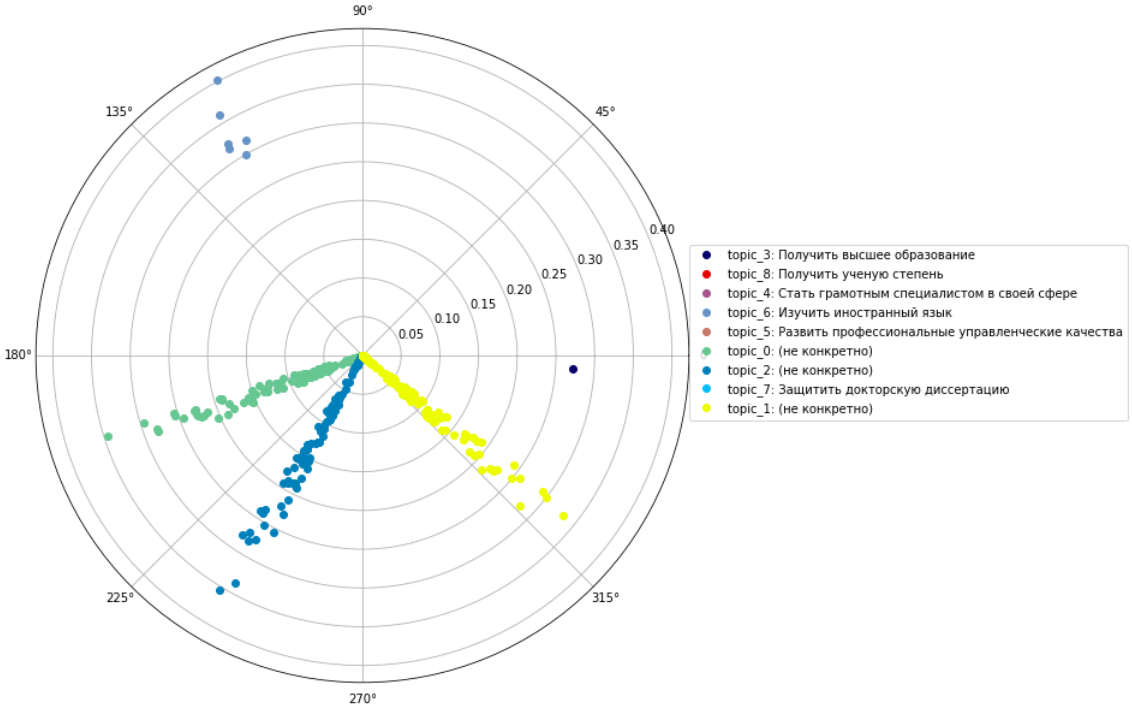

Рис. 11. Специфичность документов из d0

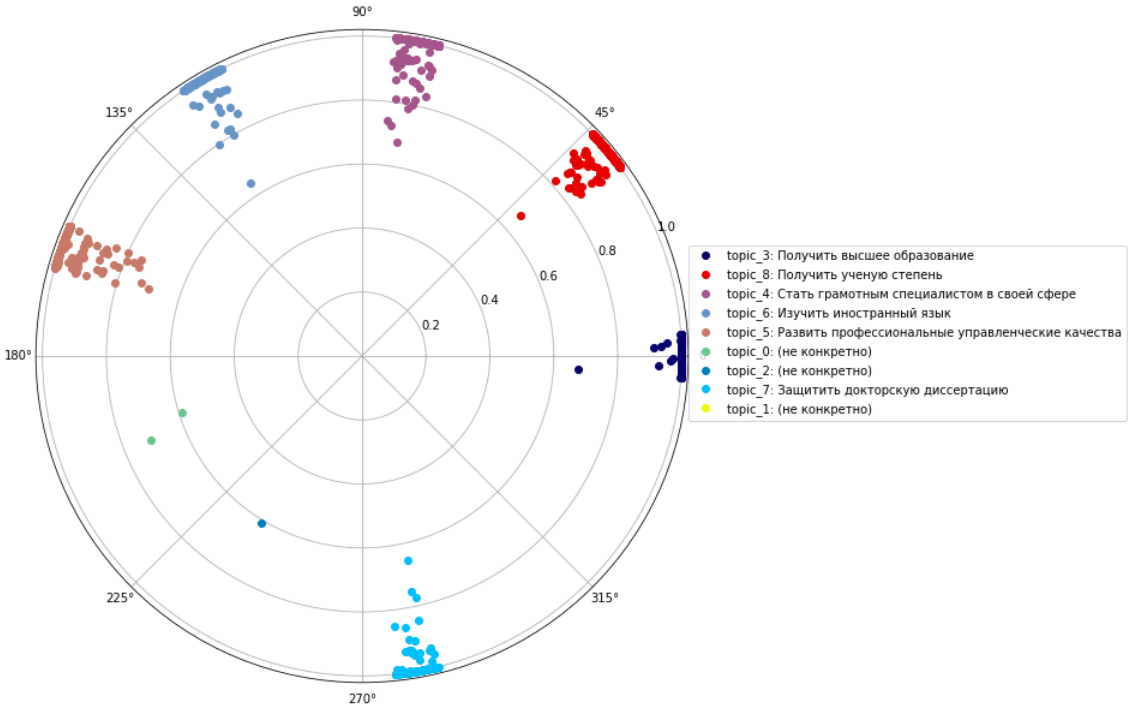

Рис. 12. Специфичность документов из d1

topic_8 являются конкретными. Видно, что у не специфичных документов мера конкретности менее 0.4, а у специфичных - более 0.8. Видно, что между неспецифичными и специфичными документами существует разрыв: почти нет документов, у которых мера специфичности была бы от 0.4 до 0.6.

На рисунке 13 представлена визуализация специфичности всех слов коллекции.

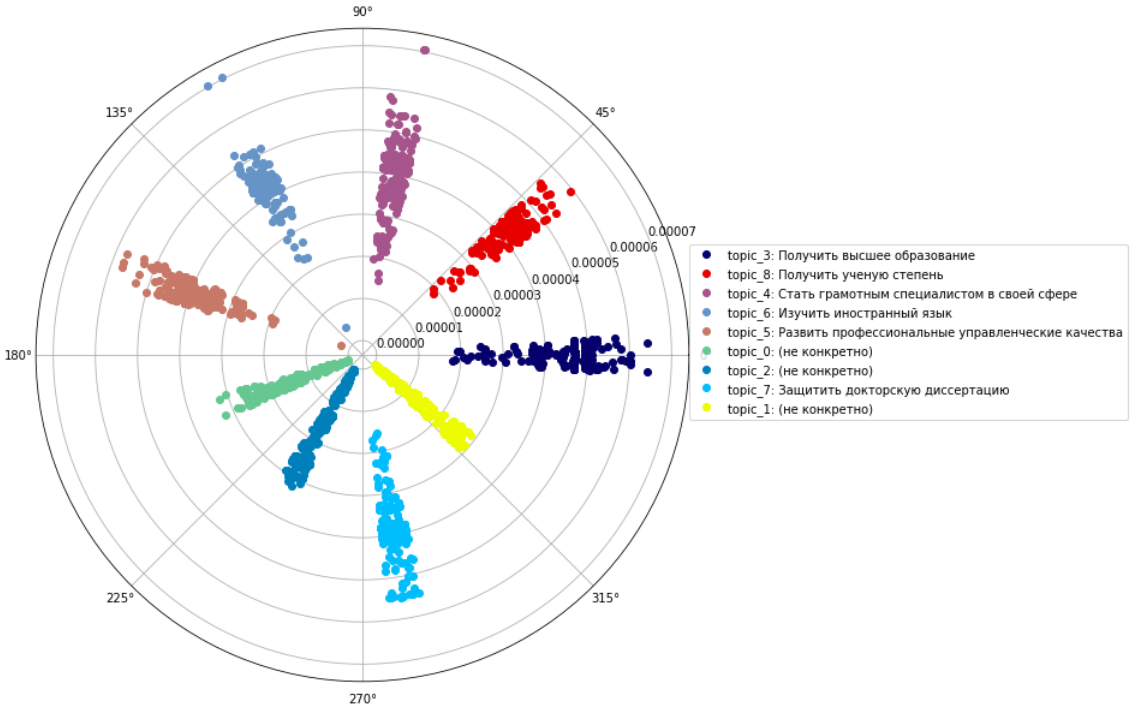

Рис. 13. Специфичность всех слов

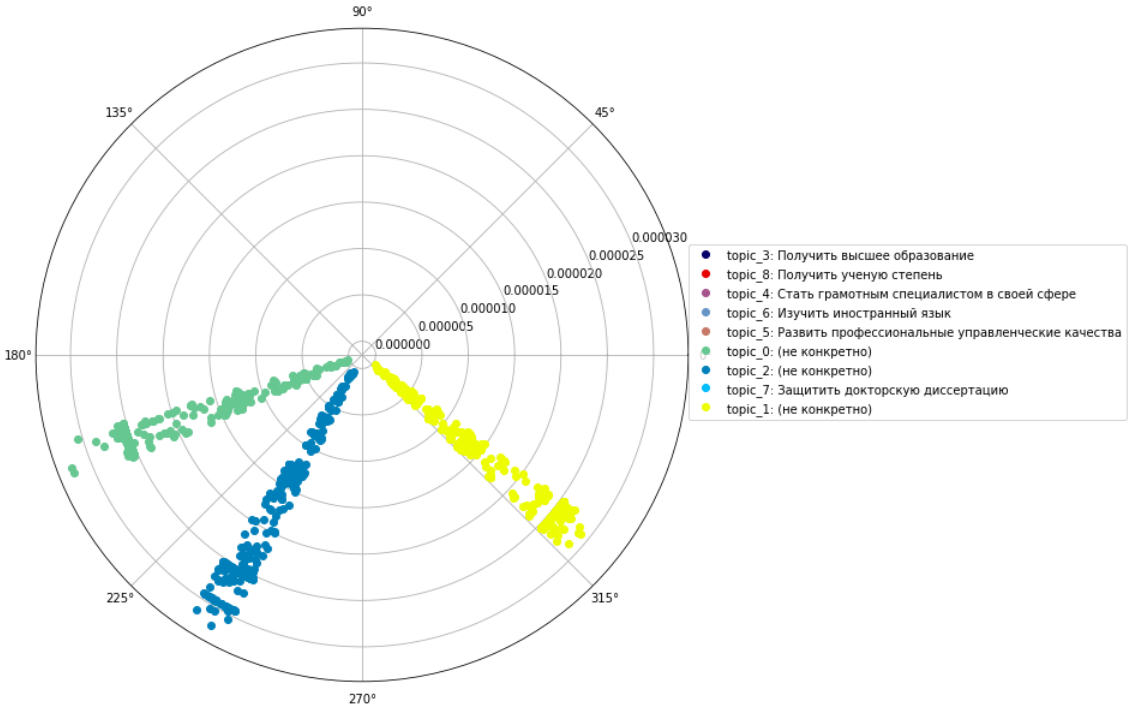

Рис. 14. Специфичность слов неконкретных тем

Видно что слова, как и документы, явно разделяются на специфичные и абстрактные, общие. Видимо, общими словами стоит называть те, у которых степень

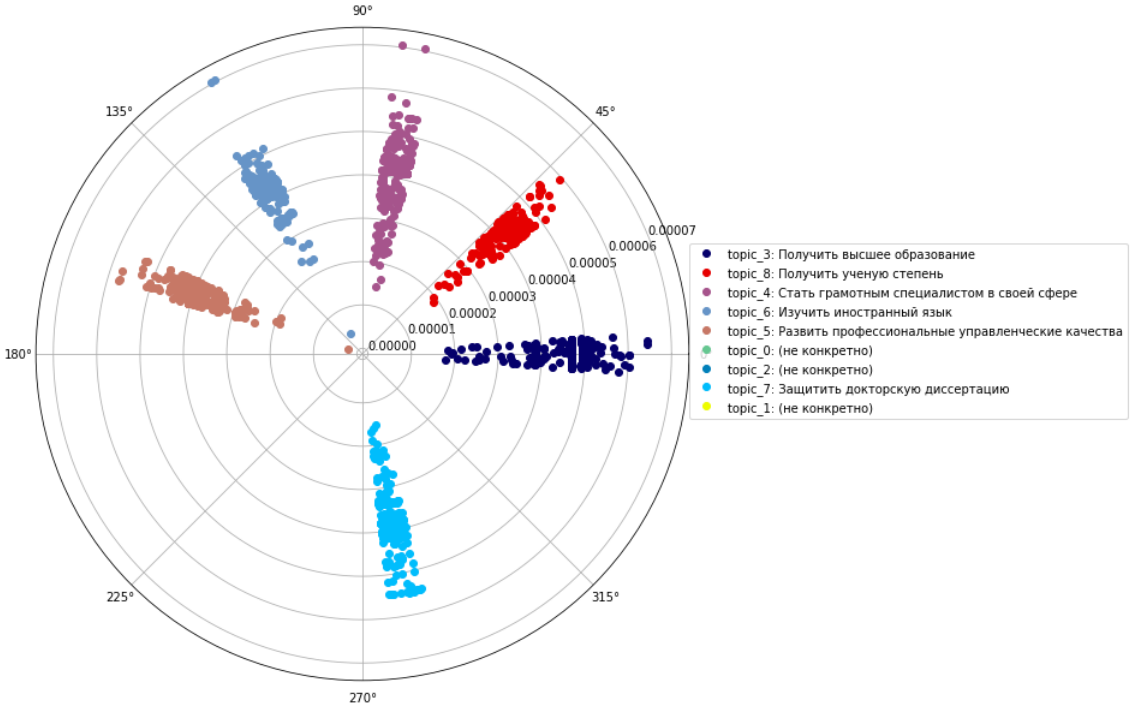

Рис. 15. Специфичность слов конкретных тем

конкретности менее 0.000035. Общие слова входят во все темы коллекции, в тот момент как предметные, специфичные слова входят только в конкретные темы. Это именно тот результат, которого мы хотели добиться.

Следует понимать, что это число, 0.000035, зависит от конкретной коллекции документов, от ее размера. При построении модели на другой коллекции, это число оказалось бы другим. В качестве такого порогового числа, отделяющего предметные слова от общих, стоит взять меру специфичности самого конкретного слова из не конкретных. Так сказать, меру специфичности лучшего из худших.

9 Заключение

В работе предложен подход к решению задачи определения конкретности образовательных целей, описанных на естественном языке и виде ответов на вопросы формальной анкеты. Предложен алгоритм бинарной классификации образовательных целей, основанный на вероятностном тематическом моделировании с аддитивной регуляризацией, позволяющий оценивать степень конкретности целей. Модель основана на предположении о явном разделении целей, тем и слов на конкретные и неконкретные. Предполагается, что неконкретные (фоновые) слова могут встречаться в любых документах, тогда как конкретные (специфичные) слова встречаются только в конкретных документах.

Результаты, выносимые на защиту

- Вероятностная тематическая модель мультимодальных анкетных данных с частичным обучением для оценивания тематики и конкретности образовательных целей пользователей в системе дистанционного образования.
- Способ обучения тематической модели, предполагающий явное разбиение коллекции документов, множества тем и словаря слов на специфичные (конкретные) и фоновые (неконкретные).
- Способ графической визуализации результатов тематического моделирования в виде круговой диаграммы «темы–специфичность».
- Результаты вычислительных экспериментов, показывающие, что предложенная тематическая модель позволяет точнее классифицировать анкетные данные, чем стандартные модели бинарной классификации (логистическая регрессия, наивный байесовский классификатор, обычная тематическая модель классификации).

Направления дальнейших исследований. Решавшаяся в данной работе задача является первым шагом к созданию диалоговой системы, которая за небольшое число уточняющих вопросов ведёт пользователя от исходных неконкретных целей к рекомендации вполне определённой образовательной траектории. По мере прохождения образовательной траектории система-советчик периодически уточняет и корректирует цели пользователя с учётом пройденных им курсов, его достижений и информации, полученной от него в диалоге.

Некоторые задачи, которые необходимо решить для создания системы-советчика:

- Построение иерархической тематической модели для детализации конкретных тем путём разбиения их на подтемы.
- Использование текстовых описаний и таксономии образовательного контента для связывания целей и образовательной траектории.
- Использование информации о пройденных курсах и достигнутых результатах пользователя для уточнения целей и образовательной траектории.

Список литературы

[1] Тимошина Т. А. Концепция выстраивания индивидуальной образовательной траектории студента // Педагогика и психология как ресурс развития современного общества : сб. ст. 2-й Междунар. науч.- прак. кон. (Рязань, 7–9 окт. 2010 г.). Рязань, 2010. С. 315–320

[2] Махныткина О. В. Моделирование индивидуальной образовательной траектории с учетом пожеланий студентов //Сборник научных трудов SWorld. – 2013. – Т. 7. – №. 2. – С. 31-33.

[3] Алгазин Г. И., Чудова О. В. Информационные технологии комплексной оценки компетентности выпускника вуза //Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2009. – Т. 7. – №. 3.

[4] Махныткина О. В. Моделирование и оптимизация индивидуальной траектории обучения студента //Рукопись. Текст.: автореф. дис. канд. техн. наук. – 2013. – Т. 5.

[5] Махныткина О. В. Моделирование индивидуальной образовательной траектории с учетом пожеланий студентов //Сборник научных трудов SWorld. – 2013. – Т. 7. – №. 2. – С. 31-33.

[6] Ботов Д. С., Дмитрин Ю. В., Кленин Ю. Д. Семантический поиск учебных дисциплин под требования рынка труда на основе нейросетевых моделей языка //Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2019. – Т. 19. – №. 2.

[7] Chern A. et al. Automated Discovery and Classification of Training Videos for Career Progression //arXiv preprint arXiv:1907.11086. – 2019.

[8] Apaza R. G. et al. Online Courses Recommendation based on LDA //SIMBig. – 2014. – С. 42-48.

[9] Reddy S., Labutov I., Joachims T. Latent skill embedding for personalized lesson sequence recommendation //arXiv preprint arXiv:1602.07029. – 2016.

[10] Jing X., Tang J. Guess you like: course recommendation in MOOCs //Proceedings of the International Conference on Web Intelligence. – 2017. – С. 783-789.

[11] Doran G. T. There'sa SMART way to write management's goals and objectives //Management review. – 1981. – Т. 70. – №. 11. – С. 35-36.

[12] Raukko J. Polysemy as complexity //A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday. SKY Journal. – 2006. – Т. 19. – С. 357-361.

[13] Melamed I. D. Measuring semantic entropy //Tagging Text with Lexical Semantics: Why, What, and How?. – 1997.

[14] Mikk J., Uibo H., Elts J. Word length as an indicator of semantic complexity //Text as a linguistic paradigm: levels, constituents, constructs. – 2001. – С. 187-195.

[15] Lewis M. L., Frank M. C. The length of words reflects their conceptual complexity //Cognition. – 2016. – Т. 153. – С. 182-195.

[16] Caraballo S. A., Charniak E. Determining the specificity of nouns from text //1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. – 1999.

[17] Ляшевская О. Н., Шаров С. А. Введение к частотному словарю современного русского языка // Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. С. 8. URL: http://dict.ruslang.ru/freq.pdf

[18] Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»(Бекасово, 4–8 июня 2014 г.). – 2014. – №. 13. – С. 20.

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

[20] Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге //Машинное обучение и анализ данных. – 2016. – Т. 2. – №. 2. – С. 173-186.

[21] Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 370-381.

Список иллюстраций

Список таблиц