УДК 004

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ПРИ РАЗРАБОТКЕ ПРОЕКТНЫХ РЕШЕНИЙ

А. А. Новиков, *младший научный сотрудник НИОЦТ ЦПИ ОАО «Гипросвязь», магистрант*
М. В. Стержанов, *доцент кафедры информатики, к. т. н.*
Белорусский государственный университет информатики и радиоэлектроники

Практический опыт применения многочисленных алгоритмов тематического моделирования убеждает, что тематические модели обладают огромным запасом решений, в том числе при разработке проектных решений в различных областях науки. При исследовании методов тематического моделирования для разработки проектных решений также исследуется возможность переноса алгоритмов некоторых методов на язык программирования Python. В рамках исследования предложена примерная последовательность шагов, необходимых и достаточных для предварительной обработки данных, на примере построения тематического профиля национальных интернет-СМИ.

Ключевые слова: интеллектуальный анализ данных, тематическое моделирование, тематическая модель, вероятностные тематические модели, эмбеддинг, токенизация.

ВВЕДЕНИЕ

В настоящее время, благодаря глобальному распространению интернета, большинство пользователей получили возможность в невиданных ранее объемах получать и генерировать информацию. Это стало не только благом, но и породило ряд проблем, связанных со сбором, анализом и систематизацией информации. Аналитики, осуществляющие свою деятельность в различных областях, вынуждены использовать соответствующие этим условиям новые подходы к сбору, анализу и систематизации интересующих данных.

Для описания нового свойства таких данных, которые отличаются большим объемом, высокой скоростью роста и огромным многообразием своих форм, был введен термин «большие данные» (big data). Феномен больших данных сформировал потребность в новых методах обработки и анализа, способных извлекать из этих, как правило, неструктурированных данных полезное знание. Совокупность таких методов обозначается термином «интеллектуальный анализ данных» (data mining).

Из этой совокупности выделяется подмножество, специализирующееся на анализе текстовых данных – интеллектуальный анализ текстовых данных (text mining). Кроме того, выделяется группа методов, которые выполняют задачу тематического моделирования – построения статистических моделей, определяющих тематическую принадлежность каждого документа из корпуса.

Тематическое моделирование – это одна из современных технологий обработки естественного языка (англ. Natural language processing, NLP), активно развивающаяся с конца 90-х годов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Тематическое моделирование не претендует на полноценное понимание естественного языка (англ. natural language understanding, NLU), однако выявление тематики можно считать определенным шагом в этом направлении [1].

Тематическое моделирование принято относить к машинному обучению без учителя, поскольку темы строятся автоматически, но текстовым данным. Для этого не требуется ни разметки, ни словарей, ни баз экспертных знаний. Существуют продвинутые тематические модели, способные учитывать такого рода данные дня улучшения тем и решения трудных задач текстовой аналитики. Такие модели тоже рассматриваются в данной статье.

Тематическая модель, как и нейросетевая, преобразует текст в векторное представление, или эмбеддинг (embedding). Под эмбеддингом понимается возможность уменьшения размерности таких признаков ради повышения производительности модели. Прежде чем говорить о структурированных наборах данных, полезно будет разобраться с тем, как обычно используются эмбеддинги.

Нейросетевые эмбеддинги не интерпретируемы, мы не понимаем смысла их координат. Тематический эмбеддинг – это вектор вероятностей тем. В каждой теме есть наиболее частотные слова, и если модель построена хорошо, то они оказываются связанными по смыслу. Глядя на них, можно сказать, о чем эта тема, составить ее текстовое описание, дать ей название. Наиболее ценное свойство тематических моделей в том, что коллекция сама собой кластеризуется на интерпретируемые темы.

Тематическое моделирование похоже на кластеризацию документов. Отличие в том, что при обычной, «жесткой» кластеризации (hard clustering) документ целиком относится к одному кластеру, тогда как тематическая модель осуществляет мягкую кластеризацию (soft clustering), распределяя содержимое документа по нескольким кластерам-темам. Тематическое моделирование называют также мягкой би-кластеризацией, поскольку каждое слово также распределяется по темам.

Вероятностная тематическая модель (англ. Probabilistic topic model, PTM) определяет вероятности тем в каждом документе и вероятности слов в каждой теме. Такие модели предсказывают вероятности появления слов в документах, но делают это не настолько хорошо, как глубокие нейронные сети типа BERT или GPT-3, но и намного проще и обладают свойством интерпретируемости.

Вероятностные тематические модели находят множество применений. Это выявление трендов в новостных потоках, патентных базах, научных публикациях, многоязычный информационный поиск, классификация и категоризация документов, тематическая сегментация текстов, суммаризация текстов, поиск тематических сообществ в социальных сетях, тегирование веб-страниц, обнаружение текстового спама. Тематическое моделирование может быть успешно использовано в том числе для решения социологических задач. С их помощью, например, можно определить разницу в освещении одних и тех же событий и фактов национальными и иностранными средствами массовой информации с использованием глобальной компьютерной сети интернет (далее – интернет-СМИ) [1] или увидеть, как менялось отношение интернет-СМИ к конкретному событию, объекту, субъекту и т. д. в определенный временной интервал.

Темой данного исследования является изучение разнообразия методов тематического моделирования при разработке проектных решений, перенос алгоритмов некоторых методов на мультипарадигменный язык программирования Python. В качестве примера показывается последовательность шагов, необходимых для предварительной обработки данных, для построения тематического профиля национальных интернет-СМИ. Обосновывается выбор количества тем для построения модели тематического моделирования.

ОСНОВНЫЕ ЭТАПЫ, НЕОБХОДИМЫЕ ДЛЯ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОГО ПРОФИЛЯ НАЦИОНАЛЬНЫХ ИНТЕРНЕТ-СМИ

Сбор данных

На данном этапе предполагается определить совокупность в построении тематического профиля национальных интернет-СМИ. Основу исследуемого пула составят новостные статьи интернет-СМИ, опубликованные в исследуемый период (истекший 2022 год). Интернет-СМИ, в целях построения указанного профиля, считается веб-сайт, ставящий своей задачей выполнение функции средства массовой информации в интернете и ориентированный на людей, живущих в Республике Беларусь.

Новостная статья – не просто текст. Это документ в электронном виде, позволяющий установить его целостность и подлинность, который подтвержден путем идентификации его принадлежности к официальному, зарегистрированному в установленном законом Республики Беларусь порядке интернет-ресурсу, выступающему в роли интернет-СМИ, имеющий свою структуру. В его структуре интерес представляют такие элементы, как собственно текст (тело статьи), название, дата публикации, комментарии и принадлежность к национальному интернет-СМИ.

Для сбора статей будет использован парсер на мультипарадигменном языке программирования Python. Предполагаемое количество собранных таким образом статей составит более 10 тыс. единиц.

Предварительная обработка данных

Обработка данных – чрезвычайно важный этап интеллектуального анализа текста для построения тематического профиля. На этом этапе происходит удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду.

Также на данном этапе осуществляется удаление из каждой статьи уникальных (специфических) признаков, свидетельствующих о ее принадлежности к какому-либо источнику. Внимательное изучение получаемых текстов статей показал, что редакции каждого СМИ устанавливают собственные правила оформления статей. Эти правила касаются способа оформления источников данных, ссылок на упоминания имен авторов статей, специфических для данного СМИ условных обозначений и т. д.

После исключения уникальной (специфической) информации данные из различных источников объединяются в единый корпус и подвергаются дальнейшей обработке.

Токенизация

Это своего рода предварительная обработка; идентификация базовых единиц, подлежащих обработке. Без этих четко разделенных базовых единиц невозможно провести какой-либо анализ или генерацию. Идентификация единиц, которые не нуждаются в дальнейшей декомпозиции для последующей обработки, является чрезвычайно важной. Ошибки, сделанные на данном этапе, вызовут больше ошибок на более поздних этапах обработки текста. Именно токены являются теми первичными элементами, которые непосредственно участвуют в процессе анализа.

Два основных признака токена: лингвистическая значимость и методологическая полезность.

Для токенизации могут использовать различные алгоритмы, например Pattern.

Лемматизация

После токенизации и удаления ненужных токенов, являющихся знаками препинания, документы будут представлены от набора неких символов к документам, представленным в виде списка слов. Дальнейшие наши шаги будут направлены на уменьшение длины этого списка, т. е. на снижение общего количества токенов. Цель таких шагов – снижение вычислительной сложности анализа данных.

Для компьютера различные формы одного и того же слова являются совершенно разными словами. Существует два способа приведения различных словоформ к одной лексеме: стемминг и лемматизация.

Для решения задач построения тематического профиля целесообразнее использовать лемматизацию, поскольку получаемые в результате этого процесса леммы интерпретировать легче, чем усеченные основы слов (стемминг), значение которых не всегда легко восстановить.

Лемматизация – это процесс приведения словоформы к лемме – ее нормальной (словарной) форме. В русском языке нормальная форма имени существительного имеет именительный падеж и единственное число, для прилагательных добавляется требование мужского рода, а глаголы, деепричастия и причастия в нормальной форме должны стоять в инфинитиве.

Удаление стоп-слов

Дальнейшие усилия по уменьшению количества токенов связаны с удалением так называемых стоп-слов. Эти слова, сами по себе почти не неся полезного смысла, тем не менее необходимы для нормального восприятия текста. Чаще всего к разряду стоп-слов относятся служебные части речи – предлоги, союзы, частицы. Будучи широко распространенными в любых текстах, они мало могут сказать о его специфике, а следовательно, и о теме.

В качестве базы для списка стоп-слов предполагается использовать русские стоп-слова из программы NLTK (платформа для создания программ Python для работы с данными естественного языка). Однако такой список нельзя считать достаточно полным, он включает (по состоянию на 01.01.2023) в себя 301 слово, покрывая лишь самые основные случаи (это, который, такой, некоторый, другой, тот и др.) [3].

Тематическое моделирование

Построение тематической модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. В терминах кластерного анализа тема – это результат бикластеризации, то есть одновременно кластеризации и слов, и документов по их семантической близости. Обычно выполняется нечеткая кластеризация, то есть документ может принадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием.

Что касается конкретных методов тематического моделирования, то в данном исследовании будет использован метод латентного размещения Дирихле (latent Dirichlet allocation, LDA). Созданный в 2003 г. [4], сейчас LDA, безусловно, лидирует среди других вероятностных тематических моделей.

На выходе после обучения модели LDA получаются векторы, показывающие, как распределены темы в каждом документе, и распределения, показывающие, какие слова более вероятны в тех или иных темах. Таким образом, из результатов LDA легко получить для каждого документа список встречающихся в нем тем, а для каждой темы – список характерных для нее слов, т. е. фактически описание темы.

Для тематического моделирования предполагается использовать программы Gensim и Mallet.

Однако прежде, чем приступать к тематическому моделированию, необходимо произвести предварительную обработку данных, специфичную для данного этапа, а именно удаление редко встречающихся токенов. Удаленные токены представляют собой слова с ошибками, цифры, гиперссылки, английские слова (в том числе написанные транслитом), имена собственные и просто редкие слова.

Определение оптимального количества тем

Определение оптимального числа тем – важная подзадача в тематическом моделировании, поскольку ее решение существенно влияет на осмысленность получаемого набора тем. Занижение числа тем приводит к чрезмерно общим результатам. Завыше-

ние приводит к невозможности разумной интерпретации. Оптимальное число тем зависит от числа документов в анализируемом корпусе: в малых корпусах оптимальным является, как правило, меньшее число тем. Однако не существует однозначного метода определения оптимального количества тем, и часто это количество определяется «на глазок», исходя из личного мнения исследователя.

Тем не менее можно говорить о том, что самым распространенным способом является расчет перплексии. Данная мера основана на значении правдоподобия, именно она использовалась в оригинальном исследовании [4], где впервые был представлен метод LDA.

Популярность тем у читателей

Зная количество комментариев к новостным статьям, мы можем определить самые резонансные темы, подсчитав, статьи какой тематики комментируют чаще всего, а какой реже.

Мы можем получить показатель комментируемости темы путем сложения рассчитанных для каждого документа произведений вероятности присутствия темы в документе на количество комментариев в данном документе. Полученное таким образом значение само по себе ничего не значит, важно лишь то, как оно различается от темы к теме. Поэтому для наглядности мы можем без последствий принять наибольшее значение комментируемости за 100 %, а для остальных тем рассчитать долю, которую они составляют от этих 100 %.

Однако построение тематической модели является некорректно поставленной оптимизационной задачей, имеющей бесконечно много решений. Согласно теории регуляризации А. Н. Тихонова, решение такой задачи возможно доопределить и сделать устойчивым. Для этого к оптимизационному критерию добавляется регуляризатор – дополнительный критерий, учитывающий специфические особенности данных или предметной области. Можно добавить регуляризатор, сильно улучшив один критерий качества, затем добавить второй и улучшить модель по другому критерию, и так несколько раз.

Постановка оптимизационной задачи максимума апостериорной вероятности (англ. Maximum a posteriori, MAP) позволяет интерпретировать логарифм априорного распределения как вероятностный регуляризатор, отделить его от конкретной модели и использовать в других моделях. Аддитивность регуляризаторов приводит к модульной технологии тематического моделирования, которая реализована в проекте BigARTM [5].

BigARTM – библиотека с открытым кодом, основанная на теории ARTM. Она имеет расширяемый встроенный набор регуляризаторов и метрик качества, реализует онлайновый и офлайновый многопоточный пакетный EM-алгоритм, обеспечивающий высокую эффективность обработки больших коллекций на одном компьютере.

Аддитивная регуляризация тематических моделей (англ. Additive regularization of topic models, ARTM) – многокритериальный подход к построению вероятностных тематических моделей коллекций текстовых документов. Охватывает наиболее известные тематические модели PLSA, LDA и многие байесовские модели. Является альтернативой байесовскому обучению тематических моделей.

Вероятностный латентный семантический анализ (англ. Probabilistic Latent Semantic Analysis, PLSA) – вероятностная тематическая модель представления текста на естественном языке. Модель называется латентной, так как предполагает введение скрытого (латентного) параметра – темы. Применяется в задаче тематического моделирования.

Latent Dirichlet Allocation (LDA) – популярный алгоритм моделирования тем реализованные в том числе в пакете Gensim. Основная задача алгоритмов TM, заключается в том, чтобы полученные темы были хорошего качество, понятными, самозначимыми и разделенными.

При решении прикладных задач комбинирование готовых регуляризаторов позволяет строить модели с заданными свойствами без дополнительных математических выкладок и программирования. Создание такой технологии в рамках байесовского подхода едва ли возможно. В отличие от байесовского обучения, ARTM стандартизирует реализацию тематических моделей в виде модульной расширяемой библиотеки регуляризаторов. Процедуры хранения и передачи данных, распараллеливания EM-алгоритма, оценивания качества моделей являются общими для широкого класса моделей и их композиций [6].

ЗАКЛЮЧЕНИЕ

Таким образом, исследование множества вероятностных тематических моделей показало высокий потенциал разнообразных вариантов их применения в практической (прикладной) плоскости, в том числе при разработке проектных решений в различных областях науки. При этом тематические модели могут выступать универсальным базисом с гибкой структурой, адаптивной к конкретному проекту, либо задаче в различных областях (СМИ, социологические, маркетинговые и иные исследования). В качестве примера по формированию такого рода базисной структуры предложены основные этапы, необходимые и достаточные, для построения тематического профиля национальных интернет-СМИ.

Исследования по данной теме будут продолжены, в том числе в области переноса алгоритмов некоторых методов на мультипарадигменный язык программирования Python.

ЛИТЕРАТУРА

1. Воронцов, К. В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM / К. В. Воронцов. – М, 2020. – 95 с.
2. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis // JASIS (41) 1990. – pp. 391–407.
3. Hofmann, Thomas. Probilistic latent semantic analysis / Thomas Hofmann // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. – 1999. – pp. 50–57.
4. Агеев, М. С., Добров, Б. В., Лукашевич, Н. В. Автоматическая рубрикация текстов: методы и проблемы / М. С. Агеев, Б. В. Добров, Н. В. Лукашевич // Ученые записки Казанского государственного университета. Серия «Физико-математические науки». – 2008. – Т. 150. – № 4. – С. 25–40.
5. Айсина, Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов / Р. М. Айсина // Машинное обучение и анализ данных. – 2015. – Т. 1. – № 11. – С. 1584–1618.
6. Воронцов, К. В., Потапенко, А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем / К. В. Воронцов, А. А. Потапенко // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). – Вып. 13 (20). – М: Изд-во РГГУ. – 2014. – С. 676–687.

Practical experience in the application of numerous topic modeling algorithms convinces us that topic models have a huge reserve of solutions, including the development of design solutions in various fields of science. In the study of topic modeling methods in the development of design solutions, the possibility of transferring the algorithms of some methods to the Python programming language is also investigated. The study proposes an approximate sequence of steps necessary and sufficient for preliminary data processing on the example of building a thematic profile of national online media.

Keywords. Data mining, topic modeling, topic model, probabilistic topic models, embedding, tokenization.

Получено 07.03.2023.