DOI: 10.36535/0203-6460-2021-01-2

Новый подход к тематическому моделированию: от пространства документов к пространству терминов*

Магнус САЛЬГРЕН
(Magnus SAHLGREN)

Исследовательские институты Швеции,
г. Стокгольм, Швеция

В статье рассматривается проблема опоры на документы как базовое понятие для определения взаимодействий термина в стандартных тематических моделях. В качестве альтернативы этой практики мы переформулируем распределения тем в латентные факторы в пространстве сходства терминов. Поясняется идея использования ряда стандартных векторных представлений слов путем построения очень широких окон контекстов. Пространства векторных представлений трансформируются в редкие пространства сходства, а темы извлекаются стандартным способом, перенося факторизацию на пространство заметно меньшего размера. Используется ряд разных способов факторизации и оцениваются различные модели с применением широкого спектра оценочных показателей, включая ранее опубликованные измерения когерентности, а также новые измерения, которые, предположительно, лучше отвечают применениям тематических моделей в реальном мире. Результаты однозначно отражают, что в большинстве случаев модели на основе терминов превосходят стандартные модели на основе документа.

ВВЕДЕНИЕ

Тематические модели часто используются в сценариях реального текстового анализа как способы эффективного изучения данных. Типичным для таких сценариев является обращение к тематической модели со стандартными параметрами данных, а также извлечение некоторого фиксированного числа *n* тем и некоторого фиксированного числа *m* слов по теме, и затем следует интерпретация, составление выводов и окончательного списка терминов. Общим выбором как для *n*, так и для *m* служит примерно 10 единиц. Это подразумевает, что аналитику нужно просмотреть только примерно 100 терминов вместо чтения собрания текстов, содержащего вероятно сотни тысяч или даже миллионов встречающихся слов. С точки зрения эффективности это является ценным средством контент-анализа.

Тематические модели извлекают темы путем раскрытия (латентных) взаимодействий между терминами в пространстве документа. Эта методология, очевидно, предполагает, что данные существуют в четких и постоянных границах документов, в лучшем случае даже справедливо распределение числа слов на документ. К сожалению, это предположение редко встречается в реальных сценариях, где данные существуют в потоках, в группах с нечеткими границами документов или с очень большим разнообразием длины документов. Чтобы осуществить такой сценарий желательно использовать модель, нечувствительную к форматированию вводимых данных. В этой статье описывается и оценивается подобный подход, который представляет в виде векторов процесс тематического моделирования полностью в пространстве терминов. Это делает модель менее чувствительной к форматированию документа и в результате даже более точной.

* Перевод Sahlgren M. Rethinking topic modelling: From document-space to term-document//Findings of the Association for Computational Linguistics: EMNLP 2020, November 16-20. — 2020. —P.2250 -2259. —https://www.aclweb.org/anthology/2020.findings-emnlp.204.pdf

Первоначально работа была мотивирована практическим использованием тематических моделей в анализе сценариев реального мира. В подобных применениях – общих, в частности, для общественных наук, а также сферы безопасности и защиты – аналитик беспокоится только о терминах верхних рангов в окончательном списке терминов. Поэтому мы дополнительно вводим ряд показателей оценки для тематических моделей, которые могут лучше отвечать практическому рассмотрению, чем обычно используемые известные, присущие цели (и в своем большинстве теоретические) измерения оценки. Также предоставляем оценку, считающую тематическое моделирование сценарием аннотации документа и использующую подготовленные вручную аннотации в качестве золотого стандарта. Наши результаты во всех показателях оценки явно демонстрируют, что подходы на основе терминов намного превосходят тематические модели на основе стандартного документа.

ТЕМАТИЧЕСКИЕ МОДЕЛИ НА ОСНОВЕ ДОКУМЕНТА

Тематические модели – семейство методов на основе латентной переменной, предназначенных определять интересные модели во встречаемостях терминов по всему документу. Большинство тематических моделей берут за отправную точку стандартную модель векторного пространства (Vector space model, VSM, т.е. матрицу термин-документ, взвешенную некоторой подходящей схемой взвешивания терминов, такой как TF-IDF). Затем это пространство термин-документ факторизируется на низкоразмерное представление, в котором размеры интерпретируются как темы. Это позволяет и терминам, и документам быть описанными как распределения относительно тем, и соответственно темам – как распределениям относительно терминов и документов. Выбор метода факторизации является основным выбором разработки, когда он касается тематического моделирования. Общие подходы включают разложение по сингулярным числам (Singular Value Decomposition – SVD) [1], факторизацию неотрицательных матриц (Non-negative Matrix Factorization, NMF) [2], латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) [3] и более современные сложные многопараметричекие программы [4, 5].

Несмотря на выбор метода факторизации, все тематические на основе документа модели опираются на предположение, что латентные взаимосвязи между терминами полагаются на тематическое разнообразие в пространстве документов. Это предположение ясно предстает в виде порождаемой истории, переданной с помощью моделей, таких как pLSA [6] или LDA, которая словесно описывает субъективный выбор (ряда) тем для беседы, и для каждой темы выбирается ряд репрезентативных терминов. Эта история несет в себе интуитивный смысл, но отметим, что понятие документа является полностью случайным по отношению к истории; оно входит в историю только как единица текста, являющегося результатом субъекта. Мы утверждаем, что понятие документа является необязательным ограничением для тематических моделей, лимитирующих применение такого рода моделей к данным с собственным форматированием, и что тематические взаимосвязи терминов можно гораздо лучше смоделировать непосредственно в пространстве терминов.

ОТ ПРОСТРАНСТВА ДОКУМЕНТОВ К ПРОСТРАНСТВУ ТЕРМИНОВ

Таким образом мы предлагаем полностью сфокусироваться на пространстве терминов и совершенно отказаться от зависимости относительно понятия документов. Вместо того, чтобы строить векторы терминов для каждого *документа* в данных (т.е. стандартную VSM), мы построим векторные представления слов для всех *терминов* в данных из больших контекстных окон протяженностью примерно в 50 токенов*. Использование таких широких контекстных окон гарантирует, что векторные представления имеют возможность кодирования более широкой, и таким образом, более тематической, контекстуальной информации.

Существует много способов построения векторных представлений слов. В эту статью включены четыре различных подхода:

- Матрица совместной встречаемости (Co-occurrence matrix, COOC), стандартная матрица термин-термин, которая взвешивает подсчеты встречаемости, собранные внутри скользящего контекстного окна, с помощью положительной точечной взаимной информации [7];
- Случайное индексирование (Random Indexing, RI), возрастающая случайная перспективная технология, аккумулирующая векторные представления слов путем сложения векторов случайного индексирования для всех слов в их контекстах [8];
- Word2Vec (W2V), упрощенная многопараметрическая программа, распознающая векторные представления с использованием языка моделирования реальности [9];
- Doc2Vec (D2V), упрощенная многопараметрическая программа, использующая ту же самую архитектуру, что и Word2Vec, но которая умеет предсказывать идентификаторы документов, а не слова [10].

Эти методы имеют свои относительные достоинства и недостатки. Подход COOC прост и направлен, но размерность векторных представлений эквивалентна размеру словаря, который может стать неиспользуемым на больших множествах. RI решает проблему размерности, так как оно использует векторы фиксированного размера, но с дополнительным шумом. Word2Vec широко признана эффективной и точной, но требует дополнительного числа проверочных данных. С другой стороны, Doc2Vec первоначально разрабатывалась для применения к обработке документов, что делает ее интересным кандидатом для более тематически ориентированного применения, такого, какое представлено здесь.

Для каждого окончательного векторного представления слов вычисляем матрицу сходства, содержащую парные сходства между всеми векторами терминов в пространстве векторных представлений. Упрощаем матрицу сходства, удалив из нее менее ценные объекты, создающие для нас разреженное пространство сходства для работы с ним. Это выгодно с вычислительной точки зрения, а также устраняет шум от представлений. Чтобы извлечь темы из пространства сходства, нужно применить любой тип алгоритма, определяющий кластеры

* Размер контекстного окна, безусловно, является параметром, который можно смоделировать и оптимизировать под определенные данные и сценарий анализа. Берем по умолчанию 50 токенов (реализаций) в этих экспериментах и признаем, что применение иного параметра может привести к другим результатам.

или латентные переменные*. В нашем случае выбран ряд методов простой факторизации, включающих:

- Разложение по сингулярным числам (SVD) [11],
- Факторизация неотрицательных матриц (NFM) [2],
- Изучение словаря (Dictionary Learning. DL) [12].

Для каждого из этих методов факторизации извлекаем *n* компоненты (где *n* по умолчанию составляет 10) тем же самым способом, что и для стандартной тематической модели. В экспериментах, приведенных в разделах «Документы против терминов», «Методы факторизации» и «Измерение темы», все результаты представляют собой среднее относительно 10 просмотров разных методов факторизации.

СВЯЗАННЫЕ РАБОТЫ

Есть ряд предыдущих исследований, изучающих использование представлений на основе термина в тематическом моделировании. Одним из примеров является работа авторов [13], которые также основывают свое решение на матрице термин-термин, но их матрица термин-термин является не матрицей встречаемости, а матрицей *корреляции*, созданной из стандартной матрицы термин-документ. Поэтому их модель все еще полагается на данные, форматированные вручную в когерентную документальную структуру. Наоборот, рассматриваемые нами модели не накладывают каких-либо ограничений на форматирование данных, в то же время принимая более строгое определение тематических взаимосвязей в форме расширения контекстных окон, (в нашем случае) – 50 терминов, которое, как правило, значительно меньше и таким образом более точно, чем весь документ.

Другой пример – работа автора [14], который кластеризует векторные представления слов (созданные с помощью Word2Vec), используя Гауссову смесь распределений (Gaussian Mixture Model, GMM). Это предшествующая работа, которая описанными в ней подходами близка к рассматриваемым нами, но есть ряд важных различий. Изучаем диапазон способов векторных представлений слов, используем более широкое контекстуальное окно (50 терминов, а не 11-17) и применяем ряд методов стандартной факторизации вместо GMM, чтобы извлечь кластеры терминов. Несмотря на эти различия, мы считаем работу автора [14] важным стимулом нашего исследования.

Еще одним стимулом для нашей работы является исследование авторов [15], которое объединяет матрицу сходства слов со стандартной моделью NFM на основе документа. Матрица сходства слов строится при помощи модели Skipgram из Word2Vec и используется как дополнительный термин в алгоритме блочного покоординатного спуска, применяемом в решении NFM. Данный подход, удачно названный Факторизацией неотрициальных матриц с поддержкой семантики (Semantics Assisted NFM, SeaNMF), первоначально разработан для данных из коротких документов, в их случае размер контекстных окон, используемых в векторных представлениях Skipgram, равняется длине документов в данных. Авторы [15] утверждают, что разреженность матрицы сходства слов очень выгодна для эффективности этой модели, следовательно, то же самое преимущество применяется и к нашему случаю. Наиболее важное различие между моделью SeaNMF и рассматриваемыми нами подходами состоит в том, что последние полагаются *только* на матрицу сходства слов и таким образом совсем не используют матрицу термин-документ.

В отличие от предыдущих исследований сфокусируемся на общей идее использования векторных представлений слов, а не на SVM как основе для тематического моделирования и сравним ряд разных методов факторизации. Также используем более широкий ряд методов оценки и введем новый ряд мер, лучше соответствующих практическому применению тематических моделей.

МЕТОДЫ ОЦЕНКИ

Поскольку (наибольшее) практическое использование тематических моделей фокусируется только на окончательных списках терминов, это должно стать центром внимания нашей оценки. Проблема здесь состоит в том, что определение качества списков терминов может быть известной субъективной задачей, сравнимой (в шутку) с гаданием на кофейной гуще [16]. Были попытки прийти к более объективным измерениям оценки в тематических моделях, которые, как правило, принимают форму использования разных форм измерения внутренней информации, таких как энтропия, дилемма или когерентность [17-20]. Однако такие теоретические измерения информации не всегда коррелируют с семантической интерпретабельностью, как отмечают авторы [16], и если это так, то неясно, почему *семантическая* интерпретабельность должна коррелировать с *тематической* когерентностью.

В свете этих трудностей несколько заметным представляется то, что аннотации золотого стандарта темы, как правило, не используются в качестве стандартного показателя оценки тематических моделей. Безусловно, вероятнее всего мы не сможем найти такие аннотации для отдельных терминов или списков терминов, но можем найти их на уровне текста. Даже если тематическое моделирование существенно отличается от категоризации или кластеризации текста, то все еще можно использовать категории текста как цели в оценке тематических моделей с учетом того, что категории являются тематическими по сути. То есть, если мы можем найти массив текстов, в котором текст вручную маркируется по одной теме или более, то можно просто сравнить эту золотую стандартную тематическую задачу с тем, что создано тематической моделью.

Один из простых способов сделать это (сравнение), который также дает картину того, как человек-аналитик может использовать результат тематической модели в практическом анализе сценария, состоит в том, чтобы собрать все документы, охваченные каждой темой (т.е. в которой встречается один или более терминов в списке тем), и затем подсчитать перекрытие между этим множеством документов и множеством документов, названных по тематическим категориям в золотом стандарте. Это приведет к доле перекрытия между тематической моделью и золотым стандартом. Утверждаем, что это простой и перспективный способ оценить тематические модели, которые непосредственно отображают применимость на практике. В табл. 2 - 5 назовем этот показатель «Истиной».

* Наши первоначальные эксперименты включали стандартные методы кластеризации, такие как k-means, агломеративную кластеризацию и методы на основе плотности, но при использовании кластеризации не было обнаружено никакого значительного улучшения по сравнению с факторизацией.

Человек-аналитик может также интересоваться и другими приведенными ниже факторами:

• Перекрытие: сколько тем перекрывается? Определим это количество как долю идентичных терминов в темах; чем их меньше, тем, предположительно, это лучше с точки зрения аналитика.

• Охват: сколько данных охватывают темы? Аналитик может отдать предпочтение такому решению – определим это количество как долю текстов, содержащих термины в темах; подходит ли это к большим или малым охватам зависит от сценария анализа.

• Уникальность: как часто термины из разных тем совместно встречаются в одном и том же документе? Если мы хотим низкий охват данных (т.е. небольшие и фокусные темы), мы вероятно будем страдать от более высокой уникальности тем, тогда как преследуя цель большого охвата данных, нам надо ожидать более низкой уникальности тем.

• Отделение: насколько векторные представления отличаются между темами? Это измеряется как среднее различие между сходством по косинусу терминов *внутри* темы и сходством по косинусу терминов *между* темами.

• Время: сколько времени занимает процесс факторизации (т.е. заключить тему в) пространства сходства? Для воспроизводства и сравнения используем функции факторизации из scikit-learn (https://scikit-learn.org/) по возможности с установкой по умолчанию.

Также включаем измерения когерентности UCI [18] и UMASS [19] в качестве сравнения. Эти меры суммируют значения PMI (точечной взаимной информации, Pointwise mutual information – PMI) всех пар слов в темах; мера UMASS подсчитывает совместные встречаемости внутри целого документа, тогда как мера UCI определяет границы совместной встречаемости слов внутри скользящего окна:

$$UMASS\left(w_i, w_j, \in\right) = \log \frac{D\left(w_i, w_j\right) + \in}{D\left(w_j\right)} \qquad (1)$$

$$UCI\left(w_i, w_j, \in\right) = \log \frac{p\left(w_i, w_j\right) + \in}{p\left(w_i\right) p\left(w_j\right)} \qquad (2)$$

Придерживаясь мнения авторов [21], устанавливаем $\in < 1$, в нашем случае $\in = 0{,}001$.

ЭКСПЕРИМЕНТЫ

Следующие эксперименты используют ряд разных массивов данных, взятых из двух разных источников. Первым источником данных является Swedish news, собранный вручную и аннотированый по темам людьми-экспертами. Этот массив данных хорошо соответствует сценарию анализа реального мира. Однако поскольку источник Swedish news – относительно небольшой и публично недоступен*, также создаем ряд искусственно аннотированных массивов данных на английском языке на основе англоязычной Википедии (English Wikipedia). Различные массивы данных подробно представлены в табл. 1 и в последующих разделах статьи.

Все эксперименты в статье обрабатываются на машине Intel Xeon E5-2620 2,40 Гц центрального процессора и 192 Гб оперативной памяти. Все техники факторизации осуществляются на стандартных установках и применениях в версии 0.20.0 метода scikit-learn. Используем применения метода Gensim (https://radimrehurek.com/gensim/) из Word2Vec и Doc2Vec со стандартными параметрами установки и внутренние применения Python из COOC и RI. Применение RI доступно на сайте: https://ghetto.sics.se/mange/ri.

Данные на основе Swedish News

Шведский массив данных состоит из новостных статей, собранных из ведущих шведских газет (Svenska dagbladet, Dagens nyheter, Aftonbladet, Expressen) авторами [22]. Каждая новостная статья аннотировалась вручную по нескольким разным категориям экспертами отделения журналистики, медиа и коммуникации Университета г. Гетеборга. Была использована категория HuvudÄmne (по-английски *main topic*, основная тема) как золотой стандарт названия, так как она явно представляет основную тему новостной статьи. На практике полезная тематическая модель должна быть способной минимально идентифицировать эти 34 различные основные темы из данных. Данные содержат 895 новостных статей с общим числом токенов – 366 456. Средняя длина документов составляет около 400 терминов, с очень высокой вариативностью. Для данных шведского массива игнорируются термины с частотой меньше 5.

Таблица 1

Массивы данных, используемые в экспериментах

Данные	#Тексты	#Токены	#Типы	#Темы	Минимальная частота
Swedish News	895	366 456	33 358	34	5
English Wikipedia	100 000	14 784 214	269 741	40 109	10
English Wikipedia (небольшие темы)	213 656	30 873 801	273 056	125 397	20
English Wikipedia (средние темы)	112 653	16 316 965	173 509	11 194	10
English Wikipedia (большие темы)	1 273	196 378	13 398	20	5

*Данные можно получить, контактируя с авторами исследования Swedish news, [22].

Таблица 2

Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
VSM	NMF	-7,72	58,73	0,17	1,00	0,11	0,21	0,18	123,20
	LDA	**-7,24**	53,37	0,36	1,00	0,11	0,17	0,19	202,44
COOC	NMF	-9,92	95,50	0,08	1,00	0,28	0,35	0,28	356,54
RI	NMF	-8,46	145,74	0,03	0,74	0,91	0,31	**0,34**	361,56
W2V	NMF	-10,41	**159,60**	0,00	0,26	0,92	0,37	0,31	468,50
D2V	NMF	-15,54	146,11	0,00	0,55	0,79	**0,45**	0,23	477,92

Примечание: Результаты для массива данных Swedish News по различным векторным представлениям (VSM, COOC, RI, Word2Vec, Doc2Vec) относительно 7 разных показателей оценки, включая оценки когерентности UMASS и UCI, тематическое перекрытие, тематический охват, уникальность, отделение и перекрытие с истиной. Также приводится время (сек.) обработки для каждой факторизации. Все оценки представляют собой среднее относительно 10 реализаций.

Таблица 3

Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
	NMF	-9,07	56,44	0,16	1,00	0,10	0,13	0,00	14,462,20
VSM	LDA	-2,31	58,55	0,34	1,00	0,12	0,15	0,01	12,399,78
COOC	NMF	1,62	152,08	0,00	1,00	0,95	0,27	0,03	8,895,03
RI	NMF	**11,55**	178,86	0,00	0,74	0,95	0,11	**0,06**	10,169,69
W2V	NMF	-5,67	141,64	0,00	0,57	0,86	**0,50**	0,01	12,098,17
D2V	NMF	9,63	**185,88**	0,00	0,14	0,97	0,44	0,02	9,571,91

Примечание: Результаты для массива данных English Wikipedia по различным векторным представлениям (VSM, COOC, RI, Word2Vec, Doc2Vec) относительно 7 разных показателей оценки, включая оценки когерентности UMASS и UCI, тематическое перекрытие, тематический охват, уникальность, отделение и перекрытие с истиной. Также приводится время (сек.) обработки для каждой факторизации. Все оценки представляют собой среднее относительно 10 реализаций.

Данные на основе Wikipedia

Так как массив данных Swedish News сравнительно небольшой и публично недоступен, также включаем ряд более крупных массивов данных на основе случайных выборок статей из English Wikipedia. Выборки созданы случайным выбором текста параграфов из Wikipedia и использованием названий векторных представлений Wikipedia как названия темы для текста. Два примера таких тем – «Изменение климата в Финляндии» и «Майк Тайсон против Мишеля Спинкса». Также используем вероятностную выборочную стратегию, создающую в среднем 20 текстовых выборок по теме, со стандартным отклонением, равным примерно 10, и минимальным числом выборок – около 5.

Как видно в табл. 1, создано 4 разных массива данных на основе этой стратегии[*]. Первый содержит 100 000 текстов с общим числом 14 784 119 униграмм терминов. Средняя длина документа составляет примерно 150 терминов со стандартным отклонением около 50 (самый длинный документ содержит примерно 1 тыс. терминов, а самый короткий – 50). Чтобы иметь возможность изучать влияние размера темы на тематические модели, создаем три разных массива данных с разнообразным числом текстов на тему. Создаем данные для небольших, среднего размера и больших тем, в которых маленькими темами являются такие, которые имеют 5 и меньше текстов по теме, большие темы – 50 и более текстов на тему, а попадающие между ними считаются среднего размера. Это приводит к 125 397 небольшим темам, содержащим 30 873 748 токенов, 11 194 темам среднего размера с наличием 16 316 954 токенов и 20 большим темам, содержащим 196 378 токенов. Используем минимальную частоту порога (10 встречаемостей) для английских данных, исключение составляют данные небольших тем, где мы вместо этого устанавливаем минимальную частоту порога – 20 встречаемостей, и данные больших тем, в которых используем в качестве порога 5 встречаемостей.

Документы против терминов

В первой группе экспериментов сравниваем тематические модели на основе документа с моделями на основе термина. Включаем две разных модели на основе документа – NMF и LDA,[*] обе применяются к стандартной VSM с помощью взвешивания TF-IDF. Сравниваем эти основные модели с четырьмя разными моделями на основе термина, которые используют NMF как метод факторизации[**]; стандартная матрица совместной встречаемости взвешивается с помощью PPMI (COOC), Random Indexing (RI), Word2Vec (W2V) и Doc2Vec (D2V).

Табл. 2 отражает результаты данных массива Swedish News. Базовые модели на основе документа получают более высокие оценки измерения UMASS на основе документа, но значительно более низкие оценки изме-

[*] Массивы данныхWikipedia можно скачать через ссылку: https://bit.ly/33hhyiQ

[*] Используем NMF и LDA, поскольку они являются наиболее общими методами факторизации, применяемыми в стандартных тематических моделях.

[**]Используем здесь NMF, поскольку она относительно устойчива. Сравнение разных методов факторизации для моделей на основе термина представлено в табл. 4.

рения UCI на основе слова. Модели на основе документа имеют более высокое перекрытие между темами и они также охватывают больше данных, но за счет меньшего числа уникальных тематических задач. Модели на основе термина имеют более высокое среднее отделение между терминами внутри, а не вдоль тем, и они больше соответствуют тематическому распределению вручную; лучшей моделью относительно перекрытия с истинными названиями является RI, которое охватывает 34 % золотого стандарта.

Табл. 3 представляет результаты данных английского массива. В этом случае отмечаем, что модели на основе терминов значительно превосходят модели на основе документа не только по измерению UCI, но и по измерению UMASS, за исключением W2V, которая имеет более низкую оценку, чем основная модель VSM+LDA. Снова отметим, что модели на основе документов имеют более высокий охват тем там, где модели на основе терминов совсем не имеют ни одного перекрытия для данных английского массива. Отметим также, что модели на основе документов стремятся охватить больше данных, чем модели на основе терминов, и что модели на основе терминов имеют более уникальные тематические распределения. Модели на основе терминов также имеют более высокое среднее отделение между терминами внутри, а не вдоль тем, и они также стремятся лучше соответствовать аннотациям золотого стандарта – н отмечается очень низкое перекрытие для всех моделей по данным английского массива; наилучшей моделью в этом случае также является случайное индексирование, которое имеет перекрытие с составленными людьми аннотациями только на уровне 6 %.

МЕТОДЫ ФАКТОРИЗАЦИИ

Обращаемся к эффектам использования различных методов факторизации для разных представлений. Табл. 2 и 3 показывают, что разница между NMF и LDA для модели на основе документа заметно ощутима для более крупного английского массива данных, в котором LDA действует несколько лучше, чем NMF. Для небольшого шведского массива данных существенной разницы нет.

Табл. 4 демонстрирует эффекты использования различных методов факторизации с применением моделей на основе терминов. Включаем три разных метода факторизации для двух разных векторных представлений (COOC и W2V) в этих результатах. Заметим, что NMF ведет к лучшим результатам для обоих векторных представлений с участием данных шведского массива, но эти результаты являются более разнородными для данных английского массива. Что касается как векторных представлений COOC, так и W2V, то изучение словаря приводит к наилучшим измерениям UMASS, UCI. SVD ведет к наилучшему разделению внутри и вдоль тем для векторных представлений COOC, а NMF – для векторных представлений W2V. Изучение словаря (DL) приводит к лучшему перекрытию аннотаций тем с участием людей для COOC, тогда как нет различия в перекрытии между разными методами факторизации для векторных представлений W2V. SVD является самым быстрым методом с участием применения технологии scikit-learn.

Измерение темы

Поскольку темы, как правило, поступают в разных размерах, то представляется уместным задать вопрос, как различные модели обрабатывают разные по размеру темы. Как указывается в разделе «Данные на основе Wikipedia», мы используем три массива данных с темами разного размера; небольшие темы, охватывающие по меньшей мере 5 текстов каждая, большие темы – 50 текстов каждая, и темы среднего размера, охватывающие от 5 до 50 текстов каждая. Табл. 5 отражает результаты модели LDA на основе документа, векторные представления W2V на основе NMF, а также векторные представления RI на основе SVD. Включаем RI в этот пример, так как оно очень хорошо выполняется на небольших и среднего размера темах.

Таблица 4

Шведский массив									
Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
COOC	NMF	**-9,08**	**121,04**	0,00	1,00	0,72	**0,31**	**0,34**	204,13
	SVD	-11,24	101,69	0,03	1,00	0,30	0,28	0,30	**100,34**
	DL	-13,60	117,97	0,05	0,95	0,57	0,26	0,24	222,05
W2V	NMF	**-8,80**	**150,04**	0,00	0,28	0,97	**0,46**	**0,47**	244,76
	SVD	-9,03	139,70	0,00	1,00	0,88	0,38	0,29	**98,82**
	DL	-12,83	149,31	0,00	0,40	0,88	0,37	0,37	221,82
Английский массив									
Векторное представление	Model	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина	Время
COOC	NMF	2,16	152,68	0,00	1,00	0,93	0,26	0,02	8,167,09
	SVD	-4,89	117,62	0,02	1,00	0,31	**0,30**	0,01	**3,529,35**
	DL	**11,41**	**162,32**	0,07	0,55	0,95	0,19	**0,05**	3,818,89
W2V	NMF	-7,51	137,2	0,00	0,60	0,85	**0,49**	0,01	8,892,46
	SVD	-2,35	150,28	0,00	0,82	0,76	0,39	0,01	**4,965,02**
	DL	**1,13**	**162,53**	0,00	0,44	0,88	0,36	0,01	5,973,62

Примечание: Результаты использования различных методов факторизации (NMF, SVD и Dictionary Learning) для векторных представлений COOC и Word2Vec в массивах данных Swedish News (вверху) и English Wikipedia (внизу). Время обработки приводится в сек. (с участием применений в методе scikit-learn), и все оценки представляют собой среднее относительно 10 реализаций.

Таблица 5

Измерение темы	Векторное представление	Модель	UMASS	UCI	Перекрытие	Охват	Уникальность	Отделение	Истина
	VSM	LDA	4,93	102,12	0,13	1,00	0,18	0,27	0,00
Небольшое	W2V	NMF	0,63	161,66	0,00	0,17	0,92	**0,53**	0,00
	RI	SVD	**16,24**	**188,14**	0,00	0,38	0,93	0,09	**0,01**
	VSM	LDA	4,76	107,27	0,10	1,00	0,20	0,27	0,04
Среднее	W2V	NMF	2,84	168,03	0,00	0,25	0,89	**0,49**	0,01
	RI	SVD	**15,29**	**186,31**	0,00	0,31	0,85	0,11	**0,06**
	VSM	LDA	-10,77	90,33	0,06	1,00	0,23	0,28	0,31
Большое	W2V	NMF	**-9,93**	**136,60**	0,00	0,30	0,97	**0,54**	**0,88**
	RI	SVD	-10,15	104,26	0,01	1,00	0,56	0,28	0,41

Примечание: Эффективность модели LDA на основе документа, W2V на основе NMF и RI на основе SVD для данных с темами разного размера. Как и в табл. 2 - 4, все оценки представляют собой среднее относительно 10 реализаций.

Наиболее примечательным в этих результатах табл. 5 является то, что ни одна из моделей не работает хорошо относительно перекрытия с названиями золотого стандарта на небольших и среднего размера темах. Векторные представления RI с факторизацией SVD получают удивительно высокие оценки UMASS и UCI, и это единственная модель с каким-либо видимым перекрытием с истинными названиями для небольших тем (едва заметное 1% перекрытие), а также имеет наибольшее перекрытие для тем среднего размера (6%). Что касается больших тем, то все модели работают значительно лучше относительно перекрытия с истинными названиями; модель на основе документа имеет перекрытие в 31%, RI – в 41%, а W2V – очень высокое перекрытие в 88%.

ОБСУЖДЕНИЕ

Как очевидно из описанных в этой статье экспериментов, различные тематические модели имеет разные свойства, и правильный выбор тематической модели зависит от определенной информационной потребности сценария конкретного анализа. Даже если модели на основе терминов в целом превосходят стандартные модели на основе документа по всем данным и показателям, используемым в этой статье, все еще могут иметь место ситуации, в которых модель на основе документа будет более подходящей в использовании. Такой сценарий может быть в случае, если аналитику требуется решение с большим охватом данных; модели на основе документа стремятся привести к более высокому охвату данных, но здесь возможно перекрытие между темами, и распределение темы (подсчитывается как встречаемость тематических терминов в документах) является менее уникальным в сравнении с моделями на основе терминов.

С другой стороны, модели на основе терминов дают больше уникальных тем с меньшим перекрытием и лучшим отделением между темами. Модели на основе терминов также достигают более высоких значений во всех показателях оценки (UMASS и UCI когерентность, представление отделения и перекрытие с истинными названиями) с исключением в данных массива Swedish News, в котором модели на основе документа ведут к более высокой когерентности UMASS. В целом различие между моделями на основе документа и на основе терминов ниже при рассмотрении измерения UMASS, чем при изучении измерения UCI, которое может быть объяснено тем фактом, что первые используют документы в качестве единиц подсчета совместной встречаемости, а вторые – слова.

Заметим, имеется большое противоречие между реализациями, что затрудняет предоставление какого-либо определенного вывода относительно выбора оптимальной разработки тематической модели на основе терминов. Определенные методы факторизации кажутся более подходящими к определенным представлениям и определенным данным. В целом NMF, кажется, работает лучше в этих экспериментах для большинства векторных представлений слов в массиве данных Swedish News, а изучение словаря (DL) лучше работает в этих экспериментах для данных массива English Wikipedia. С другой стороны, если темы небольшие, то SVD, вероятно, работает лучше, в частности, для векторных представлений RI.

Что касается различных типов векторных представлений слов, то отмечаем, что модель COOC, как правило, приводит к самому высокому охвату данных, за ней следует RI, которое также стремится иметь наилучшее перекрытие с аннотациями золотого стандарта, сделанными людьми, за исключением случая больших тем, когда Word2Vec значительно лучше. Заметим, что и Word2Vec, и Doc2Vec имеют высокое среднее отделение терминов в отличие от тем, но добавление документальной информации в Doc2Vec не кажется полезным для вывода о теме.

Подчеркнем, что данные, используемые в этих экспериментах, содержат только одну тему в документе, тогда как многие другие сценарии тематического моделирования оперируют многими темами в документе. Мы не считаем, что это ограничение должно иметь какое-то влияние на общий характер наших результатов, так как модели на основе терминов в высшей степени применимы к сценариям со многими темами. Предложенное сравнение золотого стандарта также непосредственно применимо к данным со многими темами.

ЗАКЛЮЧЕНИЕ

Статья демонстрирует большую полезность просмотра вывода о теме в тематических моделях в целях выявления (скрытых) латентных факторов в пространстве терминов, чем в пространстве документов. Предлагается простая модель на основе терминов, использующая

стандартные векторные представления слов с участием методов стандартной факторизации. Несмотря на их простоту, такие модели на основе терминов превосходят все тестируемые модели на основе документа по всем показателям оценки, используемым в статье. Также предлагается задача тематической категоризации, применяющая тематические аннотации золотого стандарта, а также ряд других показателей, которые могут лучше отвечать анализу сценариев реального мира, чем тип внутренних измерений, широко используемый в литературе по тематическим моделям. Использование этих дополнительных измерений стимулирует нас характеризовать различные свойства тематических моделей, а также сделать сознательный выбор разработки тематической модели для определенных информационных потребностей.

Наши эксперименты демонстрируют, что оптимальная модель вероятнее всего должна быть ориентированной на данные и конкретную задачу, и что оптимальный выбор определенных представлений и методов факторизации очевидно будет различаться от случая к случаю. Тем не менее, в качестве надежной основы предлагаем использовать Word2Vec представления с участием факторизации NMF.

Делаем вывод, что модели на основе терминов являются конкурентными, если не превосходящими, в сравнении с традиционными моделями на основе документа, с рядом дополнительных выгод, включая независимость документального форматирования и относительную устойчивость к размеру темы. Хотя изучаемые в этой статье модели превосходят модели на основе документов по всем показателям, считаем наш подход на основе терминов простой основной моделью с большими возможностями в улучшении.

Благодарность. Данная работа была выполнена при частичной поддержке Шведского научного агентства FOI и Шведского научного совета (грант 2017-02429, Лингвистические изучения обществ). Автор выражает признательность Магнусу Роселлю (Шведское научное агентство FOI), Йоханнесу Йохансонну (Отделение журналистики, медиа и коммуникаций, Университет г. Гетеборга), а также Стефану Дальбергу (Отделение гуманитарных и общественных наук, Шведский университет, расположенный в центре Швеции) за участие в дискуссиях по статье. Автор особо благодарит Бенгта Йохансонна (Отделение журналистики, медиа и коммуникаций, Университет г. Гетеборга) за предоставление доступа к аннотированному массиву данных Swedish News.

ЛИТЕРАТУРА

1. *Deerwester S., Dumais S. T., Furnas G. T., Landauer T. K., Harshman R.* Indexing by latent semantic analysis// Journal of the American Society for Information Science. — 1990. — Vol. 41, No. 6. — P. 391–407.

2. *Lee D.D., Seung H.S.* Algorithms for non-negative matrix factorization/ T. K. Leen, T. G. Dietterich, and V. Tresp (eds.), Advances in Neural Information Processing Systems 13, pages 556–562. — MIT Press, 2001.

3. *Blei D. M., Ng A. Y., Jordan M.I.* Latent dirichlet allocation// Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.

4. *Cao Z., Li S., Liu Y., Li W., Ji H.* A novel neural topic model and its supervised extension// AAAI Conference on Artificial Intelligence. — 2015.

5. *Miao Y., Grefenstette E., Blunsom P.* Discovering discrete latent topics with neural variational inference//Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 2410–2419. — JMLR.org., 2017.

6. *Hofmann T.* Probabilistic latent semantic analysis// Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, page 289–296, San Francisco, CA, USA. — Morgan Kaufmann Publishers Inc., 1999.

7. *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings// Transactions of the Association for Computational Linguistics. — 2015. — Vol. 3. — P. 211–225.

8. *Sahlgren M.* An introduction to random indexing// Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE), Copenhagen, Denmark. — 2005.

9. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space// Proceedings of International Conference on Learning Representations (ICLR). —2013.

10. *Le Q., Mikolov T.* Distributed representations of sentences and documents//Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32, ICML'14, page II–1188–II–1196. — JMLR.org., 2014.

11. *Golub G. H., Van Loan C. F.* Matrix Computations, third edition. — The Johns Hopkins University Press, 1996.

12. *Mairal J., Bach F., Ponce J., Sapiro G.* Online dictionary learning for sparse coding// Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, page 689–696, New York, NY, USA. — Association for Computing Machinery, 2009.

13. *Arora S., Ge R., Halpern Y., Mimno D., Moitra A., Sontag D., Wu Y., Zhu M.* A practical algorithm for topic modeling with provable guarantees//Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13), pages II–280–II–288. — JMLR.org., 2013.

14. *Sridhar V.K. R.* Unsupervised topic modeling for short texts using distributed representations of words// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 192–200. — Association for Computational Linguistics, 2015.

15. *Shi T., Kang K., Choo J., Reddy C. K.* Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations //Proceedings of the 2018 World Wide Web Conference (WWW'18), pages 1105–1114, Republic and Canton of Geneva, Switzerland. — International World Wide Web Conferences Steering Committee, 2018.

16. *Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. M.* Reading tea leaves: How humans interpret topic models // Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09, pages 288–296, USA. — Curran Associates Inc., 2015.

17. *Wallach H. M., Murray I., Salakhutdinov R., Mimno D.* Evaluation methods for topic models//Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), pages 1105–1112, New York, NY, USA. — ACM, 2009.

18. *Newman D., Noh Y., Talley E., Karimi S., Baldwin T.* Evaluating topic models for digital libraries//Proceedings of

the 10th Annual Joint Conference on Digital Libraries, JCDL '10, page 215–224, New York, NY, USA. — Association for Computing Machinery, 2010.

19. *Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models//Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, page 262–272, USA. — Association for Computational Linguistics, 2011.

20. *Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D.* Exploring topic coherence over many models and many topics// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL '12, pages 952–961, Stroudsburg, PA, USA. — Association for Computational Linguistics, 2012.

21. *Stevens K., Kegelmeyer P., Andrzejewski D., Buttler D.* Exploring topic coherence over many models and many topics//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952–961, Jeju Island, Korea. — Association for Computational Linguistics, 2012.

22. *Johansson B., Strömbäck J.* Kampen om mediebilden: nyhetsjournalistik i valrörelsen 2018.— Institutet för Mediestudier, 2019.