Использование методов тематического моделирования для оценки степени влияния СМИ на общественное настроение

А.В. Чижик

Санкт-Петербургский государственный университет

a.chizhik@spbu.ru

Аннотация

В статье рассматриваются две техники тематического моделирования – LDA и LSA. Они применяются к набору новостных анонсов, которые были опубликованы в период 2020 и 2021 года.
Целью является обнаружение наиболее освещаемых в СМИ тем и последующий анализ динамики существования этих тем в обсуждениях пользователей социальных сетей. Таким образом, тематическое моделирование используется как прикладной метод, при этом на первый план выступает задача качественного выделения тематических кластеров, так как в дальнейшем это повлияет на репрезентативность выводов об общественном настроении.

Ключевые слова: тематическое моделирование, векторная модель, латентный семантический анализ, латентное размещение Дирихле, LDA, LSA

Библиографическая ссылка: Чижик А.В. Использование методов тематического моделирования для оценки степени влияния СМИ на общественное настроение // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной научной конференции «Интернет и современное общество», IMS-2021, Санкт-Петербург, 24 – 26 июня 2021 г. Сборник научных статей). — СПб.: Университет ИТМО, 2021. С. 70-78. DOI: 10.17586/2541-9781-2021-5-70-78

1. Постановка проблемы

Классической задачей в области обработки естественного языка является тематическое моделирование, цель которого – поиск скрытой структуры данных (создание модели коллекции текстовых документов). Такая модель определяет набор тем, которые содержатся в серии документов, что позволяет рассортировать эти документы по различным тематическим категориям. Поскольку количество тем неизвестно, то эта задача относится к пулу задач обучения без учителя (unsupervised learning): на входе присутствует немаркированный набор данных, алгоритм должен самостоятельно провести их логическую классификацию. Таким образом, тематическое моделирование очень похоже на проблему кластеризации данных. С помощью моделирования тем, по сути, происходит группировка текстов, при этом кластеры приобретают интерпретацию как тематические категории. Основное отличие состоит в том, что, оказываясь в категориях тематического моделирования, приходится перейти от более традиционного евклидова векторного пространства к некоторому абстрактному пространству слов. Методы тематического моделирования можно разделить на две основных группы – алгебраические и вероятностные. Латентно-семантический анализ (LSA) относится к алгебраическим методам, а среди вероятностных наиболее популярным является латентное размещение

Дирихле (LDA). Так как LSA и LDA основаны на очень разных математических процедурах, то, очевидно, что в зависимости от типа входных текстовых данных они будут иметь разную степень успеха. При этом алгоритм использования их в рамках прикладной задачи может быть достаточно схож.

2. Набор данных

Для эксперимента была собрана коллекция новостных анонсов, опубликованных на официальной странице интернет-издания «Лента.Ру» в социальной сети «ВКонтакте» за период с 1 января 2020 г. по 20 марта 2021 г. Общее количество слов в исследуемой коллекции – 1766152 слов, среднее количество слов в одном анонсе – 11. В необработанном виде новостные анонсы являются серией текстовых строк, сопровождаемых датой публикации (также была собрана техническая информация об опубликованных постах, однако в этом исследовании она не является значимой). Ниже приведен фрагмент сформированного датасета (рис. 1).

	id	date	likes	reposts	views	comments	text
11	4847468	16-03-2021	69	20	10378	49	Администрация Байдена попыталась связаться с С...
12	4847421	16-03-2021	38	17	8613	20	Роскомнадзор обещает заблокировать Twitter в...
13	4847406	16-03-2021	8	5	6194	2	NaN
14	4847399	16-03-2021	20	16	7398	2	Глава Минвостокразвития предложил построить Ti...
15	4847374	16-03-2021	52	15	8319	2	В конце января пользователи Reddit обвели вокр...
16	4847349	16-03-2021	148	92	12015	9	Немного омолаживающих процедур с Таймыра. Свеж...
17	4847320	16-03-2021	33	18	8346	8	Протесты в Мьянме, вид от первого лица. Там во...

Рис. 1. Фрагмент исследуемого набора данных

Так как прикладной целью исследования было выявление корреляции между новостной повесткой в СМИ и формирующимся общественным мнением, то был собран еще один набор данных за тот же временной период, который раскрывал присутствующие в социальных сетях обсуждения обычных людей на предложенные темы. Источником послужил публичный Telegram-чат новостного канала Mash – «MACX», в котором участники на фоне официальных новостей обсуждают текущие события в мире, высказывая свою позицию (выводы), а также дополнительно освещая тему (факты). Этот датасет включил в себя 56171 слов, среднее количество слов в одном сообщении - 13. Формат коллекции – текст и дата его публикации (то есть, аналогичен описанному выше).

Как правило, тематическое моделирование предполагает достаточно длинные текстовые объекты в качестве исследуемой единицы (например, полный текст статьи), это связано с тем, что большее количество слов в документе помогает четче очертить потенциальную тему, а также составить объемный тематический словарь. Однако специфика новостных анонсов позволяет ожидать надежное ядро семантическое содержания за счет лаконичности дискурса, свойственного этому типу журналистского контента. Также необходимо отметить, что общий объем собранных записей обеспечил достаточную глубину анализа данных.

3. Построение векторного представления текста

Для того чтобы текст было возможно использовать в качестве входных данных в любом алгоритме, необходимо преобразовать языковую сущность (слова, предложения, параграфы или текст в его полном объеме) в набор чисел (числовой вектор). Оперируя такими векторами, в частности, становится легко представить в геометрическом пространстве близость слов друг к другу. Такой вектор называется word embedding. В случае данного

исследования процессу векторизации подвергались по отдельности каждый новостной анонс (и далее – каждое сообщение пользователя из второго набора данных).

Самый простой подход к векторизации модель «мешка слов» (Bag-of-words model, BoW), в рамках которой пренебрегается порядок слов, составляется словарь присутствующих слов, таким образом, каждое слово становится возможным превратить в вектор по длине такого словаря. Такой вектор показывает, сколько раз каждое слово из словаря встречается в конкретном документе.

Этот способ векторизации называется one-hot-encoding, и, в целом, дает необходимое количество операций над закодированным текстом для того, чтобы его можно было успешно проанализировать. Так, например, можно сложить векторы всех слов в предложении и получить вектор суммы. Также такой набор векторов дает информацию о том, насколько часто в предложении встречаются разные слова. К тому же векторы предложений можно сравнивать между собой. Распространенной альтернативой этого метода является использование статистической меры TF-IDF, которая вычисляет относительную частоту слов в документе по сравнению со всем корпусом, помогая таким образом оценить важность каждого слова внутри конкретного документа. Данный метод полезен при анализе коллекции неравномерных по длине текстов в качестве противодействия большим значениям, которые более длинные документы имели бы по сравнению с короткими, если бы использовались необработанные подсчеты. Однако исследуемый набор данных обладает двумя важными характеристиками, во-первых, в нем присутствуют тексты примерно одинаковой длинны, во-вторых, новостные анонсы являются короткими текстами, поэтому использование метода TF-IDF, скорее всего, приведет не к улучшению, а к снижению качества векторизации в рамках дальнейшей задачи тематического моделирования. Поэтому для этого исследования был выбран наивный BoW подход, который в конечном итоге принес терм-документную матрицу (document-term matrix), где каждая строка соответствует новостному анонсу, а каждый столбец – отдельному слову. Отметим, что при кодировании текстовой информации из набора данных были отсечены стоп-слова (например, предлоги и союзы) с целью обеспечения большей репрезентативности.

4. Обзор моделей LSA и LDA

Скрытый семантический анализ (LSA) был предложен для задач тематического моделирования в 2004 году [1]. В основе метода лежит идея о том, что слова будут встречаться в похожих частях текста, если они имеют одинаковое значение. На вход в LSA модель поступает матрица, состоящая из m документов и n слов (созданная с применением ранее описанных методов, или любых других способов векторизации текстовых данных). Затем происходит процедура факторизации матрицы: алгоритм раскладывает матрицу по сингулярным значениям, благодаря чему получаются три новые матрицы (на рис.2 отображена суть процесса разложения), линейная комбинация которых является достаточно точным приближением к исходной матрице.

Основная идея заключается в том, что матрица (topic matrix), получившаяся при перемножении новых ортогональных, которая содержит только k первых линейно независимых компонент исходной матрицы, отражает структуру зависимостей, которые латентно присутствовали в исходной матрице. Каждая из n строк этой матрицы представляет собой документ, а каждый из первых k столбцов соответствует теме. Тогда (i, j)-тая запись может считаться мерой присутствия темы j в документе i.

Чтобы отсортировать документ по тематической категории, достаточно узнать наибольшее значение каждой строчки (argmax), которое будет соответствовать наиболее широко представленной теме. Отметим, что количество тематических категорий (параметр k), на которое алгоритм будет делить тексты, является задаваемым параметром.

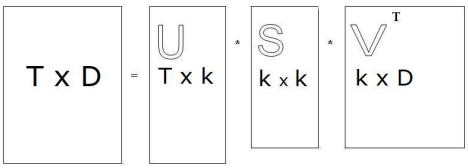

Рис. 2. Разложение матрицы размерности (T X D) на матрицу термов U размерности (T X k), матрицу документов V размерности (k X D) и диагональную матрицу S размерности (k X k), где k – количество сингулярных значении диагональной матрицы S.

Латентное размещение Дирихле (LDA) было представлено в 2003 году [2]. как генеративная вероятностная модель для коллекций дискретных данных. Важный идейный момент LDA заключается в том, что вероятностные модели удобно понимать и представлять в виде порождающих процессов (generative processes), то есть последовательно описывать, как порождается единица данных, а имено каждое слово в документе (указывая вероятностные распределения). Основа метода – предположение о том, что в каждом документе смешаны разные темы, а в каждой теме – присутствует определенное распределение слов. Интуитивно прочитывается два уровня агрегирования: 1) распределение по категориям (к примеру, новости об экономике, политические новости и т.п.), 2) распределение слов внутри категории (например, «деньги» и «акции» актуальны текстах об экономике и финансах). Поэтому документы рассматриваются как распределения вероятностей по скрытым темам, а эти темы - как распределения вероятностей по словам. При этом существует большое количество слов, которые появляются в текстах любой тематики с одинаковой вероятностью. Поэтому удаление стоп-слов и для этого метода – важный шаг реализации алгоритма.

Теоретическое обоснование LDA полагается на использование понятия взаимозаменяемости (теорема де Финетти [3]), используя которую можно получить внутридокументную статистическую структуру через смешанное распределение. Итак, метод предполагает, что процесс порождения каждого слова состоит в том, чтобы сначала выбрать тему по распределению, соответствующему документу, а затем выбрать слово из распределения, соответствующего этой теме. То есть, чтобы отсортировать новостные анонсы по тематическим кластерам, LDA обращается к априорным значениям распределения Дирихле, использует вариационные байесовские методы для вывода скрытых параметров распределения, которые затем характеризуют различные темы.

Как и в случае с LSA, количество тем является гиперпараметраметром, который задается модели на входе. Результат работы алгоритма представляется в форме матрицы, но каждая из строк теперь представляет собой распределение вероятностей, определенное по темам для каждого документа. Поэтому (i,j)-тое значение этой тематической матрицы может интерпретироваться как вероятность того, что заголовок i принадлежит теме j (точнее, как доля слов в заголовке, относящихся к теме j). Для получения оценочной категории темы каждого новостного анонса, необходимо вычислить наибольшее значение каждой строчки.

5. Описание эксперимента

Текстовые данные были предобработаны в следующей последовательности: разбиение текстов на токены; удаление спецсимволов, ссылок и пунктуации; удаление стоп-слов и лемматизация токенов. Далее текст был векторизован при помощи CountVectorizer (библиотека scikit-learn), который возвращает закодированные вектора с длиной всего словаря (поэтому векторы разреженные) и информацией, сколько раз каждое слово появилось в документе. Так возникает терм-матрица, которая будет подаваться на вход в оба алгоритма тематического моделирования.

Первым этапом анализа стало выявление наиболее частотных слов в наборе данных (без учета стоп-слов), что дало возможность оценить словарь исходных данных (рис. 3).

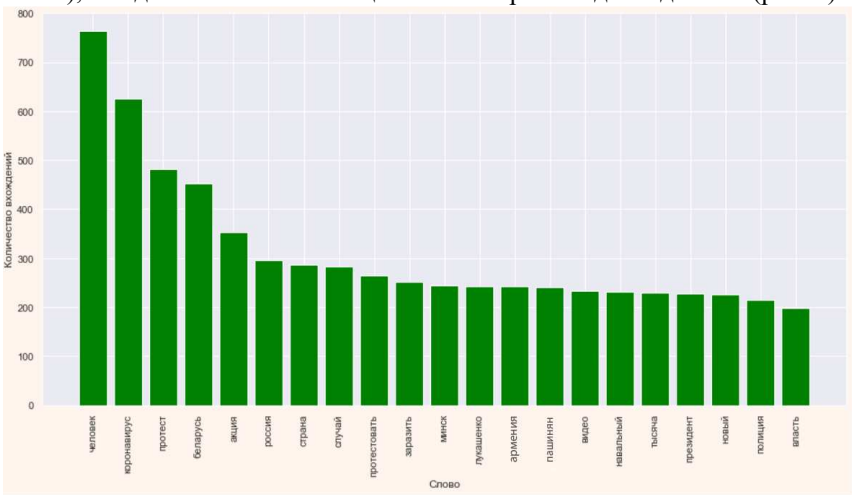

Рис. 3. 20 наиболее встречаемых в наборе данных слов

Получившаяся диаграмма показывает, что, во-первых, проведенной предобработки текстовых данных оказалось достаточно, так как наиболее часто встречающиеся слова выглядят интерпретируемо, во-вторых, интуитивно прочитывается несколько тем.

5.1. Латентно-семантический анализ

Модель LSA реализовывается с помощью TruncatedSVD (библиотека scikit-learn). Количество искомых тем было выбрано эмпирически – 20 (это же число будет использоваться и для модели LDA). Взяв argmax для каждого новостного анонса в получившейся матрице, были получены и отсортированы прогнозы тем для всех объектов в выборке. В каждом выделившемся тематическом топике были найдены наиболее часто встречающиеся слова (для более легкой дальнейшей интерпретации) – рис.4.

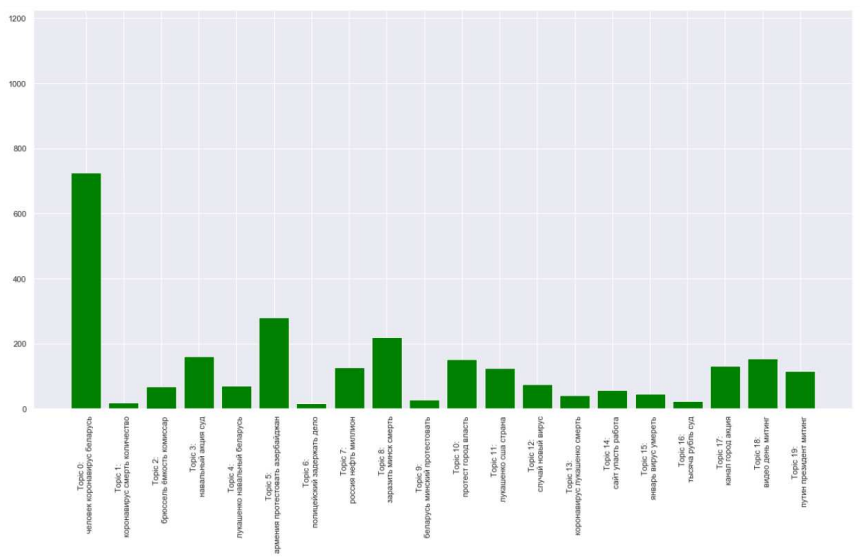

Рис. 4. Результат работы LSA, в каждой найденной теме выделены три наиболее частотных слова для визуализации

На гистограмме видно, что модель LSA в целом определила некоторые темы, которые интуитивно очерчивались и благодаря общему анализу частот слов. При этом распределение тем неравномерно, что свидетельствует о том, что одни темы более распространены, чем другие, в новостных репортажах.

Далее полученные вектора были преобразованы с использованием техники нелинейного снижения размерности t-SNE [4] для их отображения в двухмерное пространство. Таким образом, двадцатимерные тематические векторы были сжаты в двухмерные представления, чтобы выделить кластеры (рис. 5).

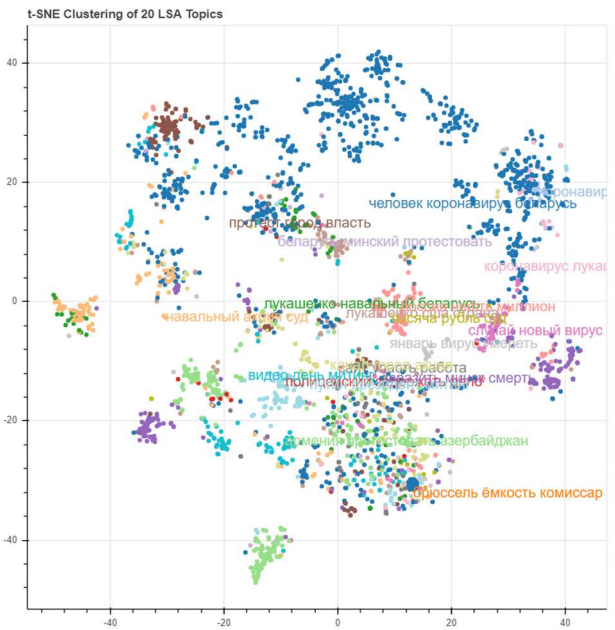

Рис. 5. Визуализация полученных тематических кластеров (модель LSA)

Хотя полученные выше тематические категории казались в целом согласованными, диаграмма рассеяния показывает, что разделение между этими категориями условное, есть участки, где кластера накладываются друг на друга. К такому результату в частности могло привести то, что один и тот же термин мог быть одинаково важен для нескольких тем одновременно.

5.2. Латентно-семантический анализ

LDA также реализован с использованием библиотеки scikit-learn (LatentDirichletAllocation класс). Также, как описано выше, для LSA по итогу работы алгоритма был вычислен argmax каждой записи в матрице, чтобы получить прогнозируемую категорию темы для каждого новостного анонса. Затем эти тематические категории были охарактеризованы по наиболее часто используемым словам, что проиллюстрировано на гистограмме (рис. 6).

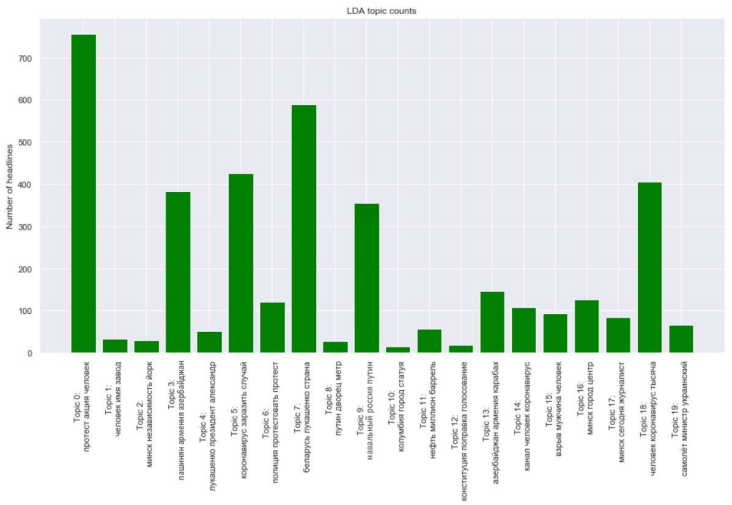

Рис. 6. Результат работы LDA, в каждой найденной теме выделены три наиболее частотных слова для визуализации

Получившиеся результаты отличаются от результатов, полученных с помощью LSA: выделенные темы более последовательны, к тому же распределение тем выглядит более убедительно. Вероятно, это следствие вариационного алгоритма Байеса, который начинается с равных априорных значений для всех категории и только постепенно обновляет их по мере прохождения через набор данных.

Для интерпретируемого сравнения LDA с LSA полученная с использованием этого метода тематическая матрица также была спроецирована в двухмерное пространство (рис. 7).

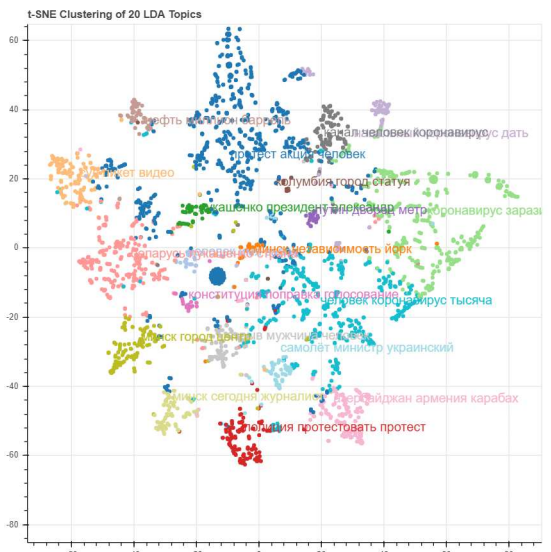

Рис. 7. Визуализация полученных тематических кластеров (модель LDA)

Отображение кластеров в двумерное пространство четко показывает, что LDA сработало для анализируемых данных гораздо лучше: тематические кластеры четко разделены между собой, к тому же каждая тема отсортирована в почти непрерывные области (инверсивная картина наблюдалась на рис. 5).

В качестве завершающего этапа исследования было решено выделить 4 наиболее популярные темы, детектированные алгоритмом LDA, чтобы обратиться ко второму набору данных с целью выявления динамики обсуждения в Telegram-канале тем, которые настолько активно обсуждались в СМИ (рис. 8). Поиск проводился по ключевым словам, выделенным на предыдущем этапе.

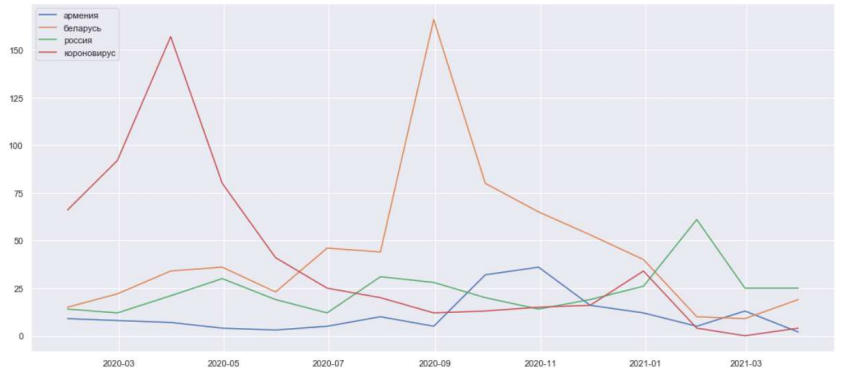

Рис. 8. Активность обсуждения тем протестов в России, Армении и Беларуси в сравнении с динамикой развития темы короновируса (анализ настроений при этом не учитывался)

Выводы

Скрытый семантический анализ (LSA) и скрытое распределение Дирихле (LDA) использовались для определения присутствующих в новостных анонсах тем. Модели LSA не удалось добиться большого разделения между выделенными кластерами, хотя выделенные темы выглядели достаточно интерпретируемыми, если ориентироваться на наиболее частотные слова. В то же время алгоритм LDA продемонстрировал большой потенциал для подобных исследований, добившись хорошего разделения между темами. Для дальнейшей оптимизации LDA-модели видится целесообразным выбор количества тематических групп производить путем оптимизации показателей качества. Например, использовать показатель согласованности, который измеряет семантическое сходство между наиболее часто встречающимися словами в теме. Максимально увеличив показатель согласованности, удастся добиться еще более хорошего качества реализованной модели.

Полученный график активности пользователей социальных сетей показывает, что СМИ оказывают влияние на динамику обсуждений той или иной темы (на рис. 8 видны пики максимального внимания и наибольшей апатии к каждой из выделенных тем, которые совпадают со временем появления новых релевантных информационных поводов), однако при этом наблюдается способность индивидов к самостоятельной оценке актуальности освещаемой средствами массовой информации повестки дня (это отчетливо видно по общему распределению количества обсуждений каждой из тем).

Литература

[1] Bellegarda J.R. Latent Semantic Language Modeling for Speech Recognition // Mathematical Foundations of Speech and Language Processing, IMA, 2004. Vol. 138. P. 73-103.
[2] Blei D., Ng A., Jordan M. Latent Dirichlet allocation // Journal of Machine Learning Research 3. 2003. P. 993-1022.
[3] Barlow R. E. Introduction to de Finetti (1937) Foresight: Its Logical Laws, Its Subjective Sources // Breakthroughs in Statistics. 1992. P. 127-133.

[4] Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE // Journal of Machine Learning Research. 2008. Vol. 9. P. 2579-2605.

Comparing LDA and LSA Topic Models for Indicating Trends of Public Mood

A.V. Chizhik

Saint Petersburg State University

This study would work on topic modeling focused on the algorithm employing Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). The data collection of news announcements, that were published between 2020 and 202, is used as the main data resours with unstructed text. The stages of preprocessing include cleansing, stemming, and stop words. The advantages of LSA are fast and easy to implement. LSA, on the other hand, doesn't consider the relationship between documents in the corpus, while LDA does. This study shows that LDA gives a better result than LSA.

Keywords: topic modeling, text embeddings, LDA, LSA

Reference for citation: Chizhik A.V. Comparing LDA and LSA Topic Models for Indicating Trends of Public Mood // Computer Linguistics and Computing Ontologies. Vol. 5 (Proceedings of the XXIV International Joint Scientific Conference «Internet and Modern Society», IMS-2021, St. Petersburg, June 24-26, 2021). - St. Petersburg: ITMO University, 2021. . 70 – 78. DOI: 10.17586/2541-9781-2021-5-70-78

Reference

[1] Bellegarda J.R. Latent Semantic Language Modeling for Speech Recognition // Mathematical Foundations of Speech and Language Processing, IMA, 2004. Vol. 138. P. 73-103.
[2] Blei D., Ng A., Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research 3. 2003. P. 993-1022.
[3] Barlow R. E. Introduction to de Finetti (1937) Foresight: Its Logical Laws, Its Subjective Sources. Breakthroughs in Statistics. 1992. P. 127-133.
[4] Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE. Journal of Machine Learning Research. 2008. Vol. 9. P. 2579-2605.