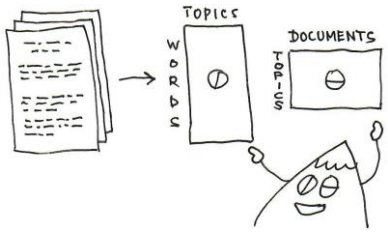
Тематический анализ больших данных

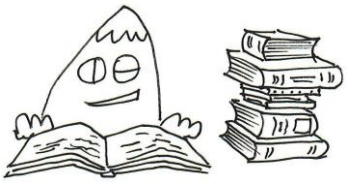
BigARTM — библиотека с открытым кодом для тематического моделирования больших текстовых коллекций и массивов транзакционных данных.

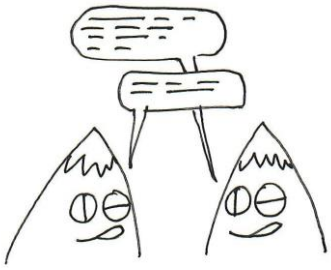
Что такое тематическое моделирование?

Это технология статистического анализа текстов для автоматического выявления тематики в больших коллекциях документов. Тематическая модель определяет, к каким темам относится каждый документ, и какими словами описывается каждая тема. Для этого не требуется никакой ручной разметки текстов, обучение модели происходит без учителя. Похоже на кластеризацию, но тематическая кластеризация является «мягкой» и допускает, чтобы документ относился к нескольким кластерам-темам. Тематическое моделирование не претендует на понимание смысла текста, однако оно способно отвечать на вопросы «о чём этот текст» или «какие общие темы есть у этих текстов».



Для чего используется тематическое моделирование?

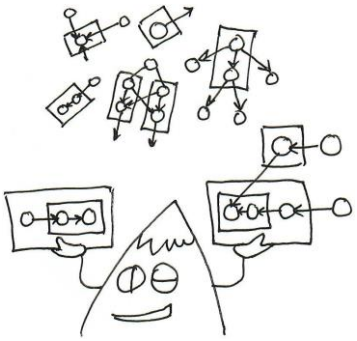




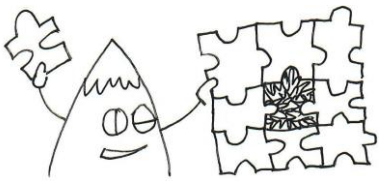
- для разведочного поиска в электронных библиотеках, это поиск по смыслу, а не по ключевым словам
- для обнаружения и отслеживания событий в новостных потоках
- для выявления тематических сообществ в социальных сетях
- для построения профилей интересов пользователей в рекомендательных системах
- для категоризации интентов собеседника и управления диалогом в системах разговорного интеллекта
- для поиска мотивов в нуклеотидных и аминокислотных последовательностях
- для аннотирования изображений
- для поиска изображений по тексту и текстов по изображениям
- для поиска аномального поведения объектов в видеопотоке
- для выявления паттернов поведения клиентов по транзакционным данным.

Слышал про модель LDA. Что-то ещё бывает?

LDA, латентное размещение Дирихле – самая известная и часто используемая тематическая модель. Она была изобретена в 2003 году. Ей предшествовала более простая модель *вероятностного латентного семантического анализа*, PLSA (1999). Позже появились ещё сотни моделей, но разработка каждой из них требовала заново производить математические выкладки и программную реализацию. Ещё одна проблема в том, что задача тематического моделирования имеет бесконечно много решений, и LDA выдаёт лишь одно из них, не предоставляя способа подобрать лучшее решение под конкретную задачу. *Теория аддитивной регуляризации* (ARTM) преодолевает эти затруднения, позволяя собирать модели из готовых модулей в стиле конструктора LEGO. Она появилась в 2014 году, и тогда же стартовал проект BigARTM.



Что такое регуляризация и почему она «аддитивная»?



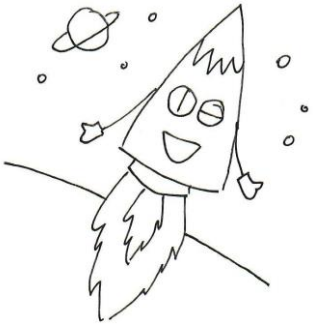
Регуляризация служит для задания желаемых свойств решения с помощью критериев-регуляризаторов. Они способны учитывать нетекстовые данные, улучшать качество классификации текстов, точность и полноту поиска, различность тем, разреженность решения, и т.д. Например, чтобы тематизировать новостной поток, необходимо учитывать время, категории и источники документов, разделять темы на подтемы и создавать новые темы «на лету». *Аддитивная регуляризация (ARTM)* позволяет задавать несколько критериев одновременно, складывать регуляризаторы от разных моделей, создавать комбинации моделей с заданными свойствами под конкретные приложения. В машинном обучении такой подход называется *мультизадачным обучением* (multitask learning).
В тематическом моделировании аддитивная регуляризация приводит к модульной технологии с высокой степенью повторного использования кода, которая и реализована в BigARTM.

Значит, BigARTM – это большой ARTM?

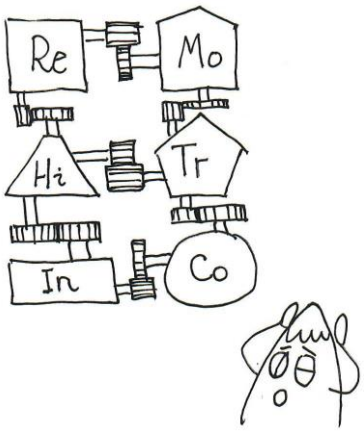
Не совсем. Приставка «big» в названии означает, что реализация модульной технологии ARTM позволяет эффективно обрабатывать большие данные. Что для этого сделано в BigARTM:

- распараллеливание на ядрах центрального процессора,
- пакетная обработка данных, не требующая единовременной загрузки больших данных в оперативную память,
- эффективный алгоритм с линейной вычислительной сложностью по объёму коллекции и по числу тем,
- хранение самых часто обновляемых данных – распределений слов в темах – целиком в оперативной памяти,
- реализация ядра библиотеки на языке C++ с соблюдением современных стандартов промышленного программирования.

BigARTM работает в разы быстрее алгоритмов, реализованных в свободно доступных библиотеках Gensim и Vowpal Wabbit.



Что ещё есть в BigARTM?



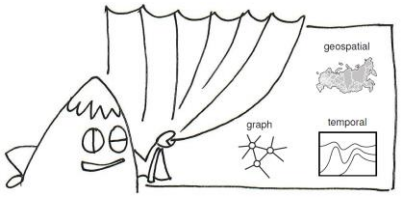
BigARTM реализует несколько механизмов, которые снимают многие ограничения простых моделей типа PLSA или LDA и расширяют спектр приложений тематического моделирования.

- *Regularization.* Регуляризаторы, которые можно комбинировать в любых сочетаниях.
- *Modalities.* Модальности, которыми можно описывать нетекстовые объекты внутри документов.
- *Hierarchy.* Тематические иерархии, в которых темы разделяются на подтемы.
- *Co-occurrence.* Использование данных о совместной встречаемости слов.
- *Intratext.* Внутритекстовые регуляризаторы, обрабатывающие текст как последовательность тематических векторов слов.
- *Transaction.* Тематизация транзакционных данных.

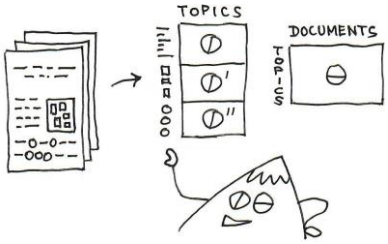
Какие регуляризаторы уже встроены в BigARTM?

- *Сглаживание* заставляет распределение слов в теме (или распределение тем в документе) быть похожим на заданное распределение. Это аналог модели LDA.
- *Разреживание* обнуляет малые вероятности в распределении слов в теме (или в распределении тем в документе).
- *Декоррелирование* делает темы более различными.
- *Отбор тем* позволяет модели избавляться от мелких, неинформативных, дублирующих и зависимых тем.
- *Когерентность* группирует часто совместно встречающиеся слова в одних и тех же темах, улучшая интерпретируемость тем.

Полный список регуляризаторов можно найти в документации.



Про регуляризацию понятно. Что такое «модальность»?

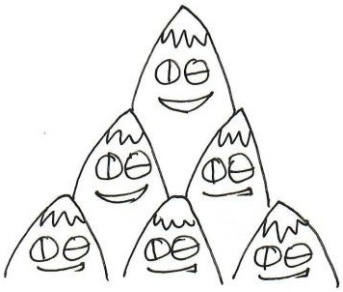


Мультимодальные тематические модели обрабатывают документы, содержащие не только слова, но и токены других модальностей. Это могут быть метаданные документа – авторы, время, источник, классы или рубрики, и т.д. Это могут быть также токены, находящиеся внутри текста – ссылки, теги, словосочетания, именованные сущности, объекты на изображениях, записи о действиях пользователей, и т.д. Модальности помогают строить темы с учётом дополнительной информации. С другой стороны, темы помогают выявлять семантику нетекстовых модальностей, предсказывать или рекомендовать значения пропущенных токенов.

Можно ли языки считать модальностями?

Да, *мультиязычные тематические модели* реализуются как частный случай мультимодальных. Параллельные или сравнимые тексты на нескольких языках образуют один документ, и слова разных языков считаются в нём модальностями. Мультиязычные модели позволяют создавать системы кроссязычного и мультиязычного тематического поиска, в которых запрос даётся на одном языке, а ответ может быть получен на других языках. Например, по тексту патента на русском языке можно искать близкие патенты на английском. Если в своей коллекции нет параллельных текстов, а мультиязычный поиск нужен, то её можно дополнить параллельными текстами из Википедии.

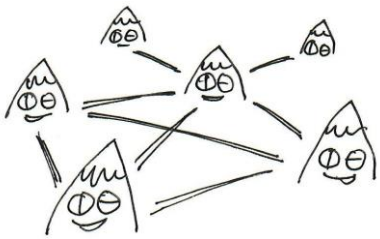
Иерархии тоже реализуются через модальности?



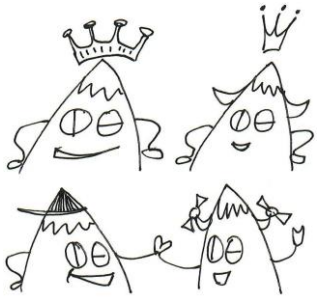
Это возможно, но более эффективным оказался другой подход. Вообще, *иерархические тематические модели* используются для автоматической рубрикации текстов. В BigARTM тематическая иерархия строится сверху вниз по уровням. Каждая дочерняя тема связывается с одной или несколькими родительскими. Каждая родительская тема может разделиться на несколько подтем, либо перейти на следующий уровень целиком. При построении каждого следующего уровня темы родительского уровня обрабатываются наряду с самой коллекцией как большие «псевдо-документы». Оказалось, что это работает лучше, чем использование родительских тем в качестве модальности.

Зачем нужны модели сочетаемости слов?

Тематические модели сочетаемости (word co-occurrence) строятся не по частотам слов в документах, а по частотам совместного употребления пар слов в общих контекстах. Построить такую модель довольно просто: надо сформировать коллекцию *псевдо-документов*, в которой каждый документ соответствует слову и содержит все слова, встречавшиеся с ним в общем контексте (например, в одном предложении) повсюду в коллекции. Такая модель называется *тематической моделью сети слов* (word network topic model, WNTM). В чём отличие от обычной модели, и чем это лучше? В основе данного подхода лежит *дистрибутивная гипотеза*: «смысл слова в языке определяется совокупностью всех слов, встречающихся в его локальных контекстах». Любая тематическая модель строит для каждого слова его векторное представление в виде распределения вероятностей тем. Но в моделях сочетаемости эти векторы точнее отражают смыслы слов в их локальных контекстах и точнее решают задачи семантической близости слов и документов.



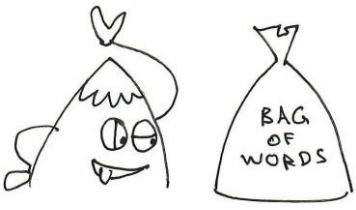
Кажется, это похоже на word2vec. Но ведь тематические модели – это другое?



В обоих подходах слова и документы получают *векторные представления* (embedding) фиксированной размерности, которые помогают решать различные задачи текстовой аналитики. В этом они похожи. Отличие в том, что тематический вектор является вероятностным распределением. Каждая координата в нём равна вероятности соответствующей темы, при этом каждая тема описывается характерными ключевыми словами или фразами. Поэтому тематические векторные представления оказываются *интерпретируемыми*, и это одно из ключевых преимуществ тематических моделей. Векторные представления семейства x2vec (word2vec, doc2vec, node2vec и другие) таким свойством не обладают. Ещё одно отличие в том, что тематическая модель находит общую тематическую структуру коллекции, а не только векторные представления отдельных слов и документов.

Зачем нужны внутритекстовые регуляризаторы?

Внутритекстовые регуляризаторы позволяют учитывать порядок слов, синтаксические связи, деление текста по предложениям и абзацам и другую внутритекстовую информацию. Важным их применением является тематическая *сегментация текстов*. Благодаря механизму регуляризации, не только темы определяют сегментацию, но и сегментация может влиять на темы. Внутритекстовая регуляризация позволяет отойти от гипотезы «мешка слов» – самого критикуемого допущения в тематическом моделировании. Есть и другие способы частичного учёта порядка слов, например, в моделях сочетаемости слов (WNTM) или при выделении словосочетаний в текстах. Однако механизм внутритекстовых регуляризаторов – наиболее общий и гибкий. Он позволяет создавать специальные регуляризаторы для выявления и анализа внутренней тематической структуры текста.

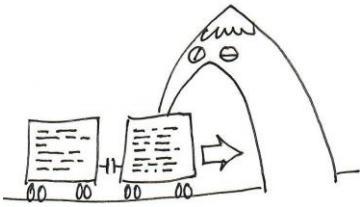


Что такое транзакционные данные?

Транзакциями называются взаимодействия между объектами. Например, транзакция {u,b,p} в рекламной сети – «пользователь u кликнул баннер b, расположенный на странице p»; или финансовая транзакция {b,s,g} – «покупатель b купил товар g у продавца s». Примером транзакции в тексте является вхождение слова в документ или появление пары слов в общем контексте. Предложения, фразы, словосочетания или синтагмы также являются примерами текстовых транзакций. *Транзакционные* или *гиперграфовые* тематические модели основаны на предположении, что транзакция объединяет слова или иные объекты, имеющие общие темы. Это наиболее общий вид тематических моделей, которые можно строить с помощью BigARTM. Они применимы не только к текстовым данным, но и к транзакционным протоколам, порождаемым разнообразными информационными системами — торговыми, финансовыми, производственными, транспортными, системами массового обслуживания и т.д.

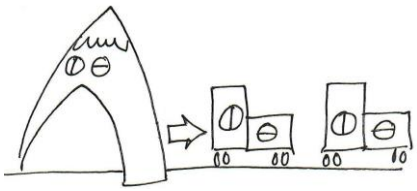
Как готовить данные для BigARTM?

BigARTM не предназначен для решения задач текстовой аналитики «под ключ». Пользователь сам определят, какая необходима предобработка входных данных и постобработка выходных. Перед обращением к BigARTM часто используются следующие методы предварительной обработки текстов:



- удаление слишком редких слов, разметки и «прочей грязи»,
- исправление опечаток,
- лемматизация или стемминг,
- удаление слишком частых слов (стоп-слов),
- автоматическое выделение терминов или коллокаций,
- выделение именованных сущностей,
- синтаксический парсинг (для некоторых Intratext-механизмов),
- вычисление частот парной сочетаемости слов.

Какие данные получаются на выходе, и как их использовать?



Основные результаты тематического моделирования находятся в двух матрицах:

- *матрица Фи «слова-темы»* содержит распределение вероятностей слов для каждой темы; они нужны, чтобы интерпретировать темы и показывать их пользователям;
- *матрица Тета «темы-документы»* содержит распределение вероятностей тем для каждого документа; они используются в качестве векторных представлений документов для поиска, классификации, визуализации документов.

Кроме того, есть побочные результаты моделирования:

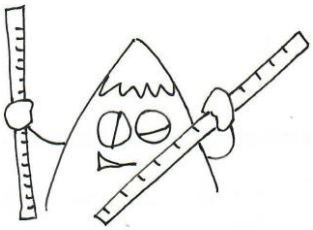
- *распределения вероятностей тем для каждого слова в каждом документе*; они служат для анализа тематической структуры документа и поиска информации внутри документов;
- *метрики качества модели*, вычисляемые на каждой итерации; они используются для контроля качества моделирования и подбора стратегии регуляризации.

Какие метрики качества вычисляются в BigARTM?

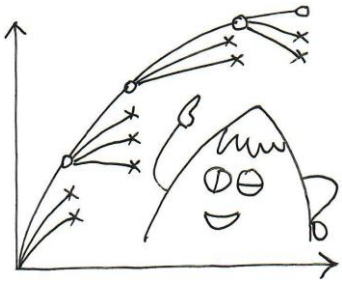
BigARTM располагает встроенными *метриками качества* (scores), и позволяет добавлять свои.
Некоторые метрики, доступные «из коробки»:

- *Перплексия.* Общепринятая мера качества моделей языка.
- *Разреженность.* Доля вероятностей, близких к нулю, в матрице Фи или Тета, соответственно.
- *Чистота и контрастность* оценивают различность тем.
- *Когерентность* наиболее вероятных слов темы. Является общепринятой мерой интерпретируемости темы.
- *Доля фоновых слов.* Если она велика, это может свидетельствовать о вырожденности модели.

Метрики качества пересчитываются на каждой итерации по каждому обработанному пакету данных и позволяют контролировать качество модели в процессе её построения.



Что такое стратегия регуляризации и как её выбирать?



Регуляризаторы во многом подобны лекарствам: в малых дозах они бесполезны, в больших становятся ядом, а некоторые их сочетания приводят к плохо предсказуемым последствиям. Комбинирование регуляризаторов требует проведения экспериментов по подбору коэффициентов, управляющих силой их воздействия на модель. Это требует определённого экспериментального опыта и знакомства с лучшими практиками в аналогичных исследованиях с BigARTM.
Стратегия регуляризации определяется набором регуляризаторов, последовательностью их включения и правилами изменения коэффициентов в ходе итераций. Обычно регуляризаторы включают по очереди, подбирая для каждого коэффициент регуляризации «методом проб и ошибок».

Чтобы разобраться в деталях, что почитать?

Документация по BigARTM есть на сайте bigartm.org.
Теория описана здесь (на русском языке):
www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
и в статье (на английском):
fruct.org/publications/fruct21/files/Koc.pdf.