МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ИМ Н.Э. БАУМАНА

Маркин Кирилл Вадимович

Разработка метода тематического моделирования для новостей на русском языке

Специальность 2301050065 —
«Программное обеспечение вычислительной техники и автоматизированных систем»

Квалификационная работа бакалавра
кандидата в бакалавры

Научный руководитель:
доцент, кандидат технических наук
Клышинский Эдуард Станиславович

Консультант:
старший преподаватель
Волкова Лилия Леонидовна

Москва — 2019

Заменить эту страницу на подписанное ТЗ

Заменить эту страницу на подписанный календарный план

Реферат

Сделать что бы тут был заголовок, но не включался в оглавление

Объект исследования и разработки

Цель и задачи работы

Метод и методология проведения работы

Результаты работы

Основные конструктивные, технологические и технико-эксплуатационные характеристики объекта исследования

Степень внедрения

Рекомендации по внедрению

Область применения

Экономическая эффективность или значимость работы

Прогнозы и предположения о возможных направлениях развития объекта исследования

Перечень условных обозначений

Сделать что бы тут был заголовок, но не включался в оглавление

Добавить условные обозначения (только если встречается более 3 раз)

// Документ -

// Тема -

Оглавление

1 Введение

2 - 3 страницы

Выключить нумерацию введения (Ирина присылала как)

Костя пошарил свою работу - глянуть что тут должно быть

1.1 // актуальность выбранной темы

1.2 // подвести к предметной области и задаче

2 Аналитический раздел

25 – 30 страниц

2.1 Постановка задачи

Целью данной работы является разработка метода тематического моделирования для новостей на русском языке.

Для достижения этой цели необходимо выполнить следующие основные **задачи**:

- Анализ существующих решений и выбор базового алгоритма тематического моделирования для классификация/категоризация новостей на русском языке
- Разработка программного продукта для сбора новостей на русском языке и подготовки данных для последующего анализа
- Подбор методов улучшения алгоритма и значений их параметров
- Обучение модели
- проведение эксперимента

2.2 Задачи тематического моделирования

проводится анализ предметной области

выделяется основной объект исследования

Задачи, для решения которых используется тематическое моделирование разбивают на 2 класса: **Автоматический анализ текста** и **систематизация больших объемов информации**.

В задачах автоматического анализа текста обычно выделяют следующие направления:

- **Классификация и категоризация документов** - необходимо присвоить каждому документу соответствующие классы. Если классы имеют иерархическую структуру - говорят о категоризации.

- **Автоматическое аннотирование документов** - составление краткого обзора на документ, используя наиболее важные фразы.
- **Автоматическая суммаризация коллекций** - решение предыдущей задачи для большой коллекции документов.
- **Тематическая сегментация документов** - разбиение длинного документа части с различными темами.

В задачах систематизации больших объемов информации обычно выделяют следующие направления:

- **Семантический (разведочный) поиск информации** - поиск по коллекции документов на базе тематического моделирования позволяет использовать длинный документ в качестве поискового запроса, а так же находить документы близкие по смыслу даже если ключевые слова, используемые при поиске отсутствуют в результатах поиска.
- **Визуализация тематической структуры коллекции** - все задачи связанные с графическим представлением больших массивов документов.
- **Анализ динамики развития тем** - обычно используется при наличии данных о времени создания документов в коллекции.
- **Тематический мониторинг новых поступлений** - автоматический мониторинг настроенных ресурсов на наличие новых документов, схожих по тематике с настроенным целевым документом.
- **Рекомендация документов пользователям** - создание систем рекомендации на основании данных о просмотренных документов пользователем и его активности.

2.3 Существующие методы

обзор существующих путей/методов/решений и алгоритмов решения

? Графические модели

? pLDA

?JPM - Join Probabilistic Model, AHMM - Aspect Hidden Markov Model, ATM - Autor-Topic Model, CTM - Correlated Topic Model

? dwl.kiev.ua - Дмитрия Владимировича Ландэ

обосновывается необходимость разработки нового или адаптации существующего метода или алгоритма

выводы из обзора (лучше сравнительную таблицу) отсюда актуальность (никто не делал так/улучшаем то-то и то-то)

рассмотреть математику используемых регуляризаторов

добавить математику мультимодальности

2.3.1 Основы кластеризации и классификации документов

В первый раз задача определения и отслеживания тем (TDT, Topic Detection and Tracking) встречается в работе "Topic Detection and Tracking Pilot Study. Final Report."[]. Темой в этой работе называют событие или действие вместе со всеми непосредственно связанными событиями или действиями. Задачей является извлечение событий.

Документы представляются векторной моделью (VSM, Vector Space Model). В такой модели каждому слову сопоставляется определенный вес, вычисляемый по весовой функции.

Базовый вариант весовых функций в таком представлении данных:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D),$$

где

$$TF(t,d) = \frac{freq(t,d)}{max_{W \in D} freq(w,d)}$$

$$IDF(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

пояснить что такое freq

Еще вариант из работы []:

$$w(t,D) = (1 + \log_2 TF(t,D)) \times \frac{IDF(t)}{||\vec{d}||},$$

где $||\vec{d}||$ - номер вектора представляющего документ D. Еще варианты модификаций TF-IDF из работ []:

$$TF' = \frac{TF}{TF + 0.5 + 1.5\frac{l_d}{l_{avg}}},$$

где l_d - длинна документа d, а l_{avg} - средняя длинна документа.

$$IDF' = \frac{\log(IDF)}{\log(N+1)}$$

Для определения расстояния в таком представлении данных использовались различные метрики: дивергенция Кульбака-Лейблера, косинус и другие. В первых работах для решения таких задач использовались алгоритмы кластеризации: метод K-средних, инкрементальная кластеризация и т. д. Каждый кластер описывал то или иное событие.

Главным недостатком такого подхода является однозначность отношения документ-тема. То есть один документ относится к одной теме (событию). В рассматриваемом выше примере про новость финансирования спорта мы увидели, что в одном документе затрагиваются сразу две темы и футбол и финансы. При таком подходе эти данные теряются.

2.3.2 Латентный семантический анализ (LSA)

Dumais et al [] в 1988 году предложил метод LSA. Суть метода в том, что бы спроецировать документы и термины в пространство более низкой размерности. Для этого анализируется совместная встречаемость слов (терминов) в документах. Таким образом задача состоит в том, что бы часто встречающиеся вместе термины были спроецированы в одно и то же измерение семантического пространства.

Дописать что надо по минимуму, что бы был понятен PLSA

2.3.3 Вероятностный латентный семантический анализ (PLSA)

В 1999 году Томасом Хофманом был предложен метод вероятностного латентного семантического анализа (PLSA) []. В вероятностных тематических моделях в отличие от рассмотренных выше методов сначала задается модель, а после с помощью матрицы слов в документах оцениваются ее скрытые параметры. В связи с чем появляется возможность дообучения моделей и упрощается подбор параметров.

Для лучшего понимания алгоритма рассмотрим детальнее процесс написания новости журналистом. Для начала работы он выбирает тему своей новостной статьи. Это, в свою очередь, влияет на то, какие слова он будет использовать. Очевидно, что если журналист решил написать новость про футбол, то слово «мяч» в таком документе появится с большей вероятностью, чем слово «антиматерия». При этом если статья затрагивает финансовую сторону вопроса, то вероятности возникновения слов «мяч» и слово «бюджет» могут сравняться. В таком случае мы можем сказать что такая новость имеет минимум две темы - «спорт» и «финансы», которые в свою очередь и породили слова «мяч» и «бюджет».

Продолжая эту аналогию можно представить себе любую новость как смесь разных тем. А каждое слово, встречающееся в новости как результат срабатывания события упоминания этого слова журналистом из тем, на которые он опирался создавая документ.

«процесс порождения текстового документа вероятностной тематической моделью.png»

Вставить картинку

Допущения

- Порядок слов в документе не важен (bag of words).
- Слова в документах генерируются темой, а не самим документом.
- Порядок документов в коллекции не важен.
- Каждое отношение документ-слово (d,w) связано с некоторой темой $t \in T$.

- Коллекция представляет собой последовательность троек документ-слово-тема (d,w,t).
- В теме не большое число образующих слов.
- В документе используется не большое число тем.

Пусть:

- D - коллекция документов размера n_d с документами d.
- W - словарь терминов размера n_w со словами w.
- T - список тем размера размера n_t с темами t.
- n_{dw} - количество использований слова w в документе d.
- Каждый документ состоит из слов: $d \subset W$
- $p(w|d)$ - вероятность появления слова w в документе d
- $p(w|t)$ - вероятность появления слова w в теме t
- $p(t|d)$ - вероятность появления темы t в документе d
- $\hat{p}(w|d) = \frac{n_d w}{n_d}$ - наблюдаемая частота слова w в документе d

Требуется найти параметры вероятностной порождающей тематической модели. То есть представить вероятность появления слов в документе $p(w|d)$ в виде:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Запишем вероятности $p(w|t)$ в матрицу $\Phi = (\phi_{wt})$, а вероятности $p(t|d)$ в матрицу $\Theta = (\theta_{td})$. Тогда вероятность появления слов в документе можно представить в виде матричного разложения:

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

«матричное разложение.png»

Вставить картинку

То есть решается задача обратная к генерации текста (работе журналиста). Необходимо по имеющийся коллекции документов понять какими распределениями матриц ϕ_{wt} и θ_{td} она могла быть получена.

Понятие стохастической матрицы

Теперь, воспользовавшись принципом максимума правдоподобия с ограничениями на элементы стохастических матриц, если максимизировать логарифм правдоподобия получается:

$$\begin{cases} \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \to max_{\Phi,\Theta}; \\ \sum_{w\in W} \phi_{wt} = 1; & \phi_{wt} \geq 0; \\ \sum_{t\in T} \theta_{td} = 1; & \theta_{td} \geq 0. \end{cases}$$

2.3.4 Латентное размещение Дирихле (LDA)

Задача в таком виде поставлена не корректно так как существует больше одного решения этой системы:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'.$$

То есть результаты будут зависеть от стартовых значений параметров модели и при кадом обучении будут различаться. Но так же это означает, что есть возможность модифицировать алгоритм, сужая пространство решений. Введем для этого критерий регуляризации $R(\Phi,\Theta)$ - некоторый функционал, соответствующий прикладной задаче, для которой обучается модель. Рассмотрим задачу максимизации регуляризованного правдоподобия:

$$\begin{cases} \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to max_{\Phi,\Theta}; \\ \sum_{w\in W} \phi_{wt} = 1; & \phi_{wt} \geq 0; \\ \sum_{t\in T} \theta_{td} = 1; & \theta_{td} \geq 0. \end{cases}$$

В 2003 году Дэвидом Блеем, Эндрю Энджи и Маклом Джорданом был предложен метод латентного размещения Дирихле (LDA) [. На дан-

ный момент это одна из самых цитируемых статей по тематическому моделированию. Они предложили решать задачу со следующим регуляризатором:

$$R(\Phi,\Theta) = \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td},$$

$$\beta_w > 0,$$

$$\alpha_t > 0,$$

где β_w и α_t - параметры регуляризатора.

Для понимания метода введем понятие дивергенции Кульбака-Лейблера для дискретных распределений.
Пусть даны два дискретных распределения $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$, тогда дивергенция Кульбака-Лейблера

$$KL(P||Q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

Дивергенция Кульбака-Лейблера обладает следующими свойствами:

- Неотрицательность:
  Добавить формулу

- Несимитричность:
  Добавить формулу

Дивергенция Кульбака-Лейблера связана с максимумом правдоподобия:
Добавить формулу

Пусть P - эмпирическое распределение. Q - параметрическая модель распределения с параметром α. При минимизации дивергенции Кульбака-Лейблера (максимизации правдоподобия) определяется такое значение α, при котором P как можно лучше соответствует модели.

Минимизация дивергенции Кульбака-Лейблера эквивалентна максимизации правдоподобия.

Пусть

Добавить формулу

- некоторый вектор над словарем W. Со словами w.

При

Добавить формулу

вероятность

Добавить формулу

этого слова по темам будет сглаживаться приближаясь к

Добавить формулу

:

Добавить формулу

При

Добавить формулу

значение

Добавить формулу

наоборот будут разреживаться, удаляясь от

Добавить формулу

к нулю :

Добавить формулу

то есть в матрице Ф будет больше нулевых элементов или близких к нулю.

2.3.5 Аддитивная регуляризация тематических моделей (ARTM)

Неединственность решения максимизации регуляризованного правдоподобия позволяет накладывать сразу несколько ограничений на модель, этот метод называется аддитивной регуляризацией тематических моделей (ARTM).

То есть :

Добавить формулу

Где :

Добавить формулу

- коэффициенты регуляризации, а

Добавить формулу

- регуляризаторы.

При таком подходе возникает проблема поиска коэффициентов, которая обычно решается добавлением регуляризаторов в модель по одному и оптимизации соответствующих коэффициентов в ходе пробных запусков моделей.

2.3.6 Решение задачи максимизации регуляризованного правдоподобия

Решение задачи в общем виде аналитическими методами слишком сложно. Однако, если выбирать гладкие регуляризаторы, то можно воспользоваться условием Крауша-Куна-Таккера. Получится система уравнений:

Добавить формулу

Где

Добавить формулу

Такую систему можно решить численным методом простых итераций. В данном случае его называют EM-алгоритм.

Для получения результата необходимо итерационно выполнять E-шаг и M-шаг до достижения требуемой точности.

E-шаг :

Добавить формулу

Где

Добавить формулу

M-шаг :

Добавить формулу

Где

Добавить формулу

Этот процесс можно организовать параллельно, если обновлять мат-

рицу Φ по порциям, после анализа очередного пакета документов. Обычно уже после просмотра нескольких первых десятков тысяч документов матрица Φ получается уже устоявшиеся и остается только тематизировать остальные документы

2.3.7 Выбор алгоритма

Добавить выбор алгоритма

В данной работе рассматривается задача классификации и категоризации документов. В качестве документов выступают новости на русском языке. Необходимо с помощью выбранного метода и способов его усовершенствования разбить коллекцию новостей на темы, интерпретируемые человеком и получить возможность оценивать новый документ (новость) на принадлежность этим темам.

Особенностью тематического моделирования является возможность не использовать в процессе построения модели размеченные данные. То есть темы, на которые разбивается коллекция так же создаются по ходу формирования модели.

2.3.8 Формализованное описание проблемы

Откуда брать данные и какие они бывают

описание критериев сравнения нескольких реализаций метода или алгоритма

Входные данные:

- Коллекция новостей на русском языке на разные темы в сети интернет.

Выходные данные:

- Обученная тематическая модель с настроенными регуляризаторами.
- Список тем с образующими их словами
- Названия тем

Получение данных:

- Парсинг новостных агрегаторов
- Парсинг крупных новостных сайтов

Подготовка данных:

- Удаление форматирования текста
- Исправление опечаток
- Слияние слишком коротких текстов
- Выделение терминов
- Приведение слов к нормальной форме (лемматизация)
- Удаление слишком частых слов
- Удаление слишком редких слов

2.4 // Функциональные требования к

Что мы хотим получить (это и будет "мостиком"к конструкторской)

Для решения задачи классификации и категоризации новостей на русском языке необходимо, чтобы программная реализация собирала новости из ресурсов сети Интернет, обрабатывала их в формат, необходимый для работы модели, создавала и обучала модель. При обучении необходимо подобрать наилучший комплект регуляризаторов, их параметров и коэффициентов. Также должная быть возможность последующего повторного использования и дообучения модели.

3 Конструкторский раздел

25 – 30 страниц

3.1 // обосновать последовательность этапов выполнения

3.2 // Алгоритм сбора данных

как будем извлекать данные (без кода пока)

Мой написанный код для парсинга

Уже предварительно собранные открытые данные

https://newspaper.readthedocs.io/en/latest/ - возможный инструмент для парсинга

25 500 новостей (там суммарно 9 000 000 слов - я посчитал) за все время существования media.zone (я сам написал парсер, могу его же натравить на любой другой новостной ресурс) - уже скачены и лежат на моем компьютере

statmt.org - это не совсем подходит нам, тут новости короткие совсем. Но тоже скачал на всякий случай поиграться - тут суммарно 8,4 гигабайта чистого текста - уже скачены и лежат на моем компьютере

webhose.io - 290 000 новостей - уже скачены и лежат на моем компьютере

Можно сделать сервис на РИА новости

Можно сделать сервис на агрегаторы новостей

3.3 // Алгоритм анализа

разработка метода

Базовый алгоритм: ARTM (bigartm.readthedocs.io)

Предобработка текста: лемматизация, удаление стоп-слов, ngrams

Используем модальности (дата публикации, ссылки на другие документы, авторы)

Используем производные от статьи данные по различным алгоритмам (записываем в модальности) - алгоритмы еще не выбраны

IDEF0 метода

3.4 // ? Что делаем

Можно попробовать обучаться на месяце/неделе/дне (и это в теории можно вынести в экперимент) и выдавать как меняются темы

решить иерархически ли хотим строить темы или многое ко многим

3.5 // Оценка

как будем оценивать (без кода)

Разбиение на 2 части и замеры разницы оценки - устойчивость - Через предложение разбивать статью можно попробовать

Толока - описание теста - выбрать лишнее слово, подумать что еще можно

3.6 // Требования к программе

4 Технологический раздел

20 - 25 страниц

4.1 // обоснованный выбор средств программной реализации

4.2 // описание основных (нетривиальных) моментов разработки

4.3 // методики тестирования созданного программного обеспечения

4.4 // информация, необходимая для сборки и запуска разработанного программного обеспечения

5 Экспериментальный раздел

10 - 15 страниц

5.1 // эксперименты и их результаты

Можно поиграть с периодом обучение и сравнения данных (месяц/неделя/день) и смотреть где лучше (?что лучше)

Можно поиграть с размером новости и посмотреть как от этого зависят результаты

5.1.1 // проводим апробацию

5.1.2 // анализируем результаты

5.2 // качественное и количественное сравнение с аналогами

оцениваем адекватность и качество

5.3 // даём рекомендации о применимости метода/софта

6 Заключение

6.1 // отчитаться по каждому пункту тз/по каждой задаче и цели

6.2 // сказать про перспективы (мы все уже не умрём)

7 Список источников

7.1 // Разобрать

- Ссылка на записи с datafest
- Воронцов - книги и лекции
- Ученики Воронцова - доклады и статьи
- Анастасия Янина - работала с Воронцовым - посмотреть ее доклады и статьи
- Потапенко Анна - работала с Воронцовым - посмотреть ее доклады и статьи
- "Диалог NLP Конференция
- курсы на курсере
- dwl.kiev.ua - Дмитрия Владимировича Ландэ
- Обзор
- Topic Detection and Tracking Pilot Study. Final Report.

7.2 // Датасеты

- 25 500 новостей (там суммарно 9 000 000 слов - я посчитал) за все время существования media.zone (я сам написал парсер, могу его же натравить на любой другой новостной ресурс) - уже скачены и лежат на моем компьютере
- statmt.org - это не совсем подходит нам, тут новости короткие совсем. Но тоже скачал на всякий случай поиграться - тут суммарно 8,4 гигабайта чистого текста - уже скачены и лежат на моем компьютере
- webhose.io - 290 000 новостей - уже скачены и лежат на моем компьютере
- Можно сделать сервис на РИА новости
- Можно сделать сервис на агрегаторы новостей

8 Приложения

добавить схемы, листинги программного кода, наборы тестов и др

8.1 //