

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

1 Техническое задание

Заменить эту страницу на подписанное ТЗ

2 Календарный план

Заменить эту страницу на подписанный календарный план

3 Реферат

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

4 Перечень условных обозначений

Добавить условные обозначения (только если встречается более 3 раз)

// Документ -

// Тема -

Оглавление

1	Техническое задание	2
2	Календарный план	3
3	Реферат	4
4	Перечень условных обозначений	5
5	Введение	8
5.1	// актуальность выбранной темы	8
5.2	// подвести к предметной области и задаче	8
6	Аналитический раздел	9
6.1	Постановка задачи	9
6.2	Задачи тематического моделирования	9
6.3	Существующие методы	10
6.3.1	Основы кластеризации и классификации документов	11
6.3.2	Латентный семантический анализ (LSA)	12
6.3.3	Вероятностный латентный семантический анализ (PLSA)	13
6.3.4	Латентное размещение Дирихле (LDA)	15
6.3.5	Аддитивная регуляризация тематических моделей (ARTM)	16
6.3.6	Решение задачи максимизации регуляризованного правдоподобия	16
6.3.7	Выбор алгоритма	16
6.3.8	Формализованное описание проблемы	17
6.4	// Функциональные требования к	18
7	Конструкторский раздел	19
7.1	// обосновать последовательность этапов выполнения	19
7.2	// Алгоритм сбора данных	19
7.3	// Алгоритм анализа	19

7.4	// ? Что делаем	20
7.5	// Оценка	20
7.6	// Требования к программе	20
8	Технологический раздел	21
8.1	// обоснованный выбор средств программной реализации .	21
8.2	// описание основных (нетривиальных) моментов разработки	21
8.3	// методики тестирования созданного программного обеспечения	21
8.4	// информация, необходимая для сборки и запуска разработанного программного обеспечения	21
9	Экспериментальный раздел	22
9.1	// эксперименты и их результаты	22
9.1.1	// проводим апробацию	22
9.1.2	// анализируем результаты	22
9.2	// качественное и количественное сравнение с аналогами .	22
9.3	// даём рекомендации о применимости метода/софта . . .	22
10	Заключение	23
10.1	// отчитаться по каждому пункту тз/по каждой задаче и цели	23
10.2	// сказать про перспективы (мы все уже не умрём)	23
11	Список источников	24
11.1	// Разобрать	24
11.2	// Датасеты	24
12	Приложения	25
12.1	//	25

5 Введение

2 - 3 страницы

Костя пошарил свою работу - глянуть что тут должно быть

5.1 // актуальность выбранной темы

5.2 // подвести к предметной области и задаче

6 Аналитический раздел

25 – 30 страниц

6.1 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

6.2 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 документах пользователем и его активности.

6.3 Существующие методы

обзор существующих путей/методов/решений и алгоритмов решения

Классификация и кластеризация документов, VSM (Vector Space Model)

LSA - Латентно-семантическое индексирование, SVD - Singular Value Decomposition

? Графические модели

PLSA - Probabilistic latent semantic analysis

LDA - Latent Dirichlet allocation - латентное размещение Дирихле - специальный регуляризатор для Баеса

? pLDA

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

ARTM - Additive Regularization for Topic Modeling

Обзор

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)

6.3.1 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." [1]. 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 [1]:

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

где $\|\vec{d}\|$ - номер вектора представляющего документ D . Еще варианты модификаций TF-IDF из работ [1]:

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

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

6.3.2 Латентный семантический анализ (LSA)

Dumais et al [1] 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

6.3.3 Вероятностный латентный семантический анализ (PLSA)

В 1999 году Томасом Хофманом был предложен метод вероятностного латентного семантического анализа (PLSA) [1]. 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 подробнее 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 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
- $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ - 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 имеющейся 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:

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

6.3.4 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:

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

В 2003 году Дэвидом Блеем, Эндрю Энджи и Маклом Джорданом был предложен метод латентного размещения Дирихле (LDA) [1]. На данный 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.

6.3.5 Аддитивная регуляризация тематических моделей (ARTM)

6.3.6 Решение задачи максимизации регуляризованного правдоподобия

6.3.7 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.

6.3.8 Формализованное описание проблемы

Необходимая существующая математика

описание входных и выходных данных

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

6.4 // Функциональные требования к

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

7 Конструкторский раздел

25 – 30 страниц

7.1 // обосновать последовательность этапов выполнения

7.2 // 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

7.3 // 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

7.4 // ? Что делаем

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

решить иерархически ли хотим строить темы или многое ко многим

7.5 // 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

7.6 // Требования к программе

8 Технологический раздел

20 - 25 страниц

- 8.1 // обоснованный выбор средств программной реализации
- 8.2 // описание основных (нетривиальных) моментов
разработки
- 8.3 // методики тестирования созданного программного
обеспечения
- 8.4 // информация, необходимая для сборки и запуска
разработанного программного обеспечения

9 Экспериментальный раздел

10 - 15 страниц

9.1 // эксперименты и их результаты

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

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

9.1.1 // проводим апробацию

9.1.2 // анализируем результаты

9.2 // качественное и количественное сравнение с аналогами

оцениваем адекватность и качество

9.3 // даём рекомендации о применимости метода/софта

10 Заключение

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

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

11 Список источников

11.1 // 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.

11.2 // 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

12 Приложения

добавить схемы, листинги программного кода, наборы тестов и др

12.1 //