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)

Оглавление

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 Анализ предметной области

проводится анализ предметной области

выделяется основной объект исследования

6.2.1 Задачи тематического моделирования

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

В задачах автоматического анализ а 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.

6.2.2 Классификация и категоризация документов

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

6.2.3 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

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)

6.4 // 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

IDEF0 метода

7.4 // ? 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

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 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

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

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 //