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

Оглавление

1 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

2 Перечень условных обозначений

2.1 //

3 Введение

2 - 3 страницы

Костя пошарил свою работу - глянуть что тут должно быть

3.1 // актуальность выбранной темы

3.2 // перечисляются задачи, которые необходимо решить для достижения этой цели

4 Аналитический раздел

25 – 30 страниц

4.1 // проводится анализ предметной области и выделяется основной объект исследования

4.2 // обзор существующих методов и алгоритмов решения

4.3 // обосновывается необходимость разработки нового или адаптации существующего метода или алгоритма

4.4 // обзор существующих методов и алгоритмов решения

4.5 // формализованное описание проблемы предметной области

4.5.1 описание входных и выходных данных

4.5.2 описание критериев сравнения нескольких реализаций метода или алгоритма

4.5.3 описание способов тестирования разработанного, адаптированного или реализованного метода или алгоритма

4.5.4 описание функциональных требований к разрабатываемому программному обеспечению

5 Конструкторский раздел

25 – 30 страниц

5.1 // обосновать последовательность этапов выполнения

5.2 // Алгоритм сбора данных

Мой написанный код для парсинга

Уже предварительно собранные открытые данные

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

5.3 // Алгоритм анализа

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

5.4 // 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

5.5 // Тесты

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

6 Технологический раздел

20 - 25 страниц

6.1 обоснованный выбор средств программной реализации

6.2 описание основных (нетривиальных) моментов разработки

6.3 методики тестирования созданного программного обеспечения

6.4 информация, необходимая для сборки и запуска разработанного программного обеспечения

7 Экспериментальный раздел

10 - 15 страниц

7.1 экспериментов 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

7.2 качественное и количественное сравнение с аналогами

8 Заключение

Посмотреть по правилу презентации - какие здачи ставили и какие выполнили

8.1 //

9 Список источников

9.1 // Разобрать

Ссылка на записи с datafest

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

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

9.2 // 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

dwl.kiev.ua - Дмитрия Владимировича Ландэ

Можно сделать сервис на РИА новости

Можно сделать сервис на агрегаторы новостей

10 Приложения

схемы, листинги программного кода, наборы тестов и др

10.1 //