

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ИМ Н.Э. БАУМАНА

На правах рукописи

УДК //xxx.xxx

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:

// дописать степень, звание
Волкова Лилия Леонидовна

Москва — 2019

Оглавление

1	Реферат	4
1.1	Объект исследования и разработки	4
1.2	Цель и задачи работы	4
1.3	Метод и методология проведения работы	4
1.4	Результаты работы	4
1.5	Основные конструктивные, технологические и технико-эксплуатационные характеристики объекта исследования	4
1.6	Степень внедрения	4
1.7	Рекомендации по внедрению	4
1.8	Область применения	4
1.9	Экономическая эффективность или значимость работы	4
1.10	Прогнозы и предположения о возможных направлениях развития объекта исследования	4
2	Перечень условных обозначений	5
2.1	//	5
3	Введение	6
3.1	//	6
4	Аналитический раздел	7
4.1	//	7
5	Конструкторский раздел	8
5.1	// Алгоритм	8
5.2	// Тесты	8
6	Технологический раздел	9
6.1	//	9
7	Экспериментальный раздел	10
7.1	//	10

8	Заключение	11
8.1	//	11
9	Список источников	12
9.1	//	12
10	Приложения	13
10.1	//	13

1 Реферат

1.1 Объект исследования и разработки

//

1.2 Цель и задачи работы

//

1.3 Метод и методология проведения работы

//

1.4 Результаты работы

//

1.5 Основные конструктивные, технологические и технико-эксплуатационные характеристики объекта исследования

//

1.6 Степень внедрения

//

1.7 Рекомендации по внедрению

//

1.8 Область применения

//

1.9 Экономическая эффективность или значимость работы

//

1.10 Прогнозы и предположения о возможных направлениях развития объекта исследования

//

2 Перечень условных обозначений

2.1 //

3 Введение

Костя пошарил свою работу - глянуть что тут должно быть

3.1 //

4 Аналитический раздел

4.1 //

5 Конструкторский раздел

5.1 // Алгоритм

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

5.2 // Тесты

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

6 Технологический раздел

6.1 //

7 Экспериментальный раздел

7.1 //

8 Заключение

8.1 //

9 Список источников

Ссылка на записи с datafest

Анастасия Янина - работала с Воронцовым - посмотреть ее доклады и статьи

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

* 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

9.1 //

10 Приложения

10.1 //