Московский государственный технический университет имени Н.Э. Баумана

Разработка метода тематического моделирования для новостей на русском языке

Автор:
студент группы ИУ7-81
Маркин Кирилл Вадимович

Научный руководитель:
доцент, кандидат технических наук
Клышинский Эдуард Станиславович

Консультант:
старший преподаватель
Волкова Лилия Леонидовна

Москва 2019

Актуальность

Из-за огромного количества различных новостных потоков стало сложно выделять действительно актуальную для себя информацию.

Разрабатываемый мною метод тематического моделирования новостей будет применяться для распределения новостного потока на различные нужные для пользователя темы.

Это решение будет особенно актуально пользователям, интересующимся узкими темами, которые не распределяются журналистами на категории.

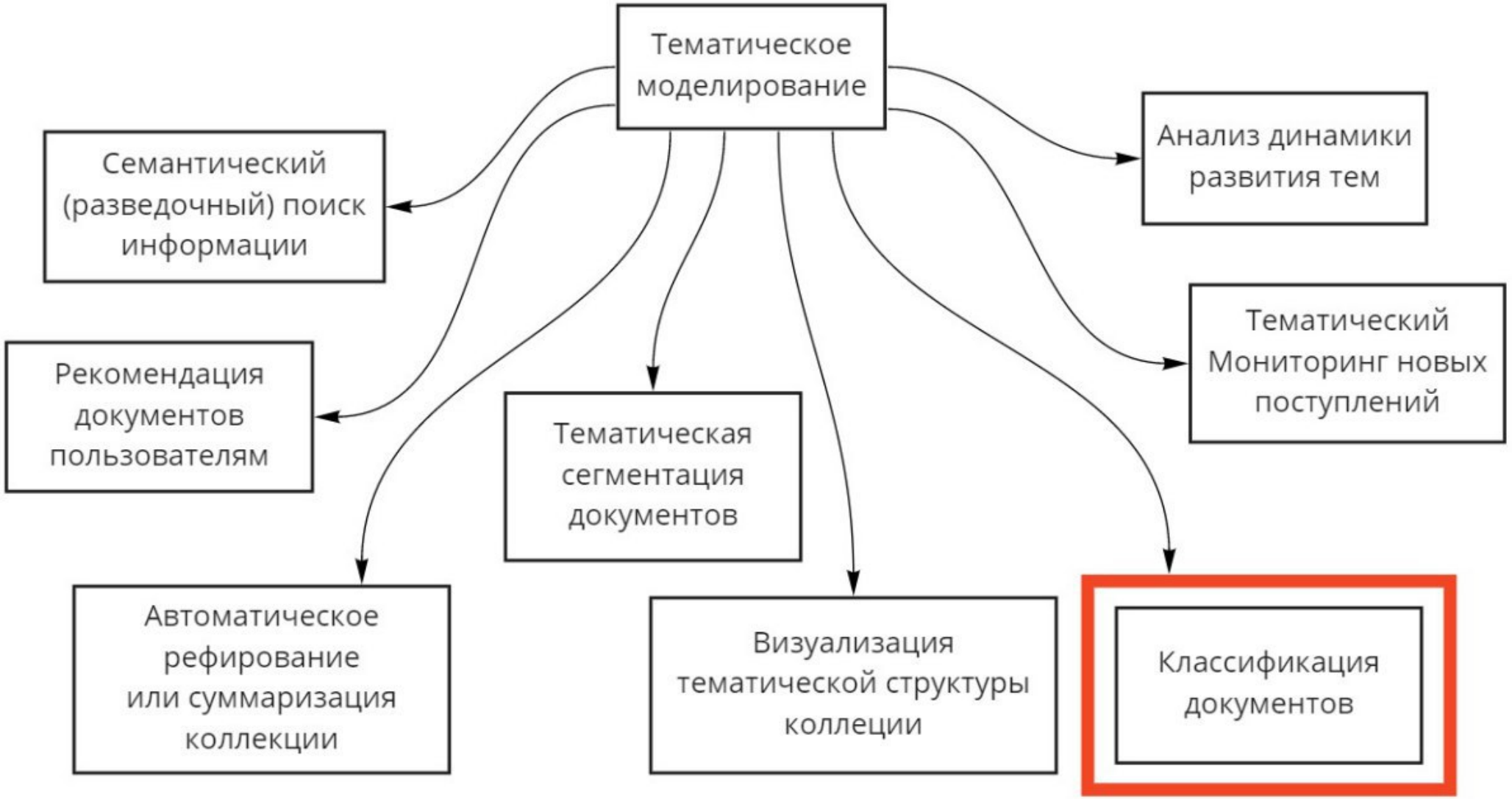
Цели и задачи

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

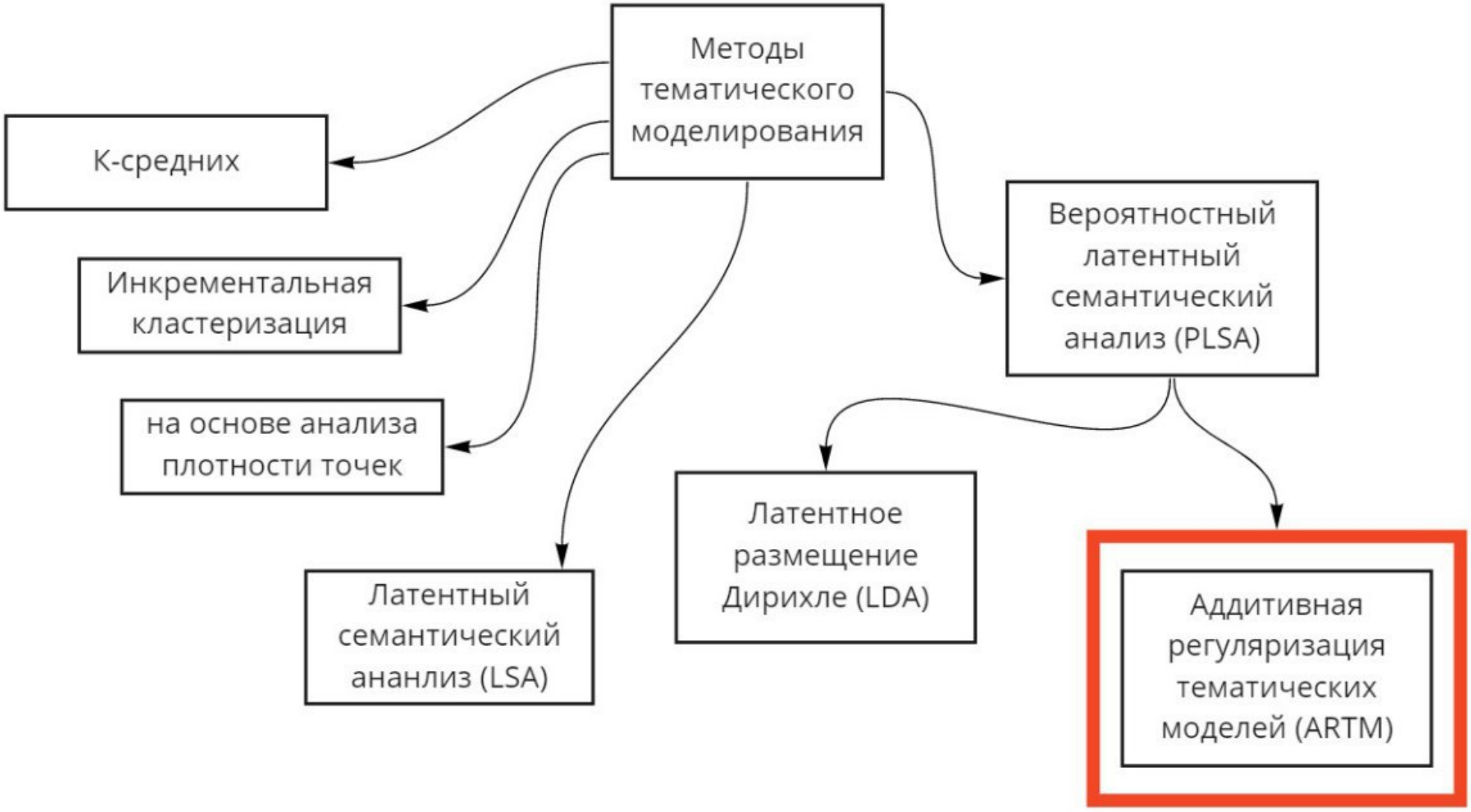
Задачи:

- анализ существующих решений и выбор базового алгоритма тематического моделирования для классификации новостей на русском языке;
- разработка программного продукта для сбора новостей на русском языке;
- разработка программного продукта для подготовки данных для последующего анализа;
- подбор методов улучшения алгоритма и значений его параметров;
- обучение модели;
- проведение параметризации метода;
- проведение апробации метода;
- составление рекомендаций о применимости предложенного метода.

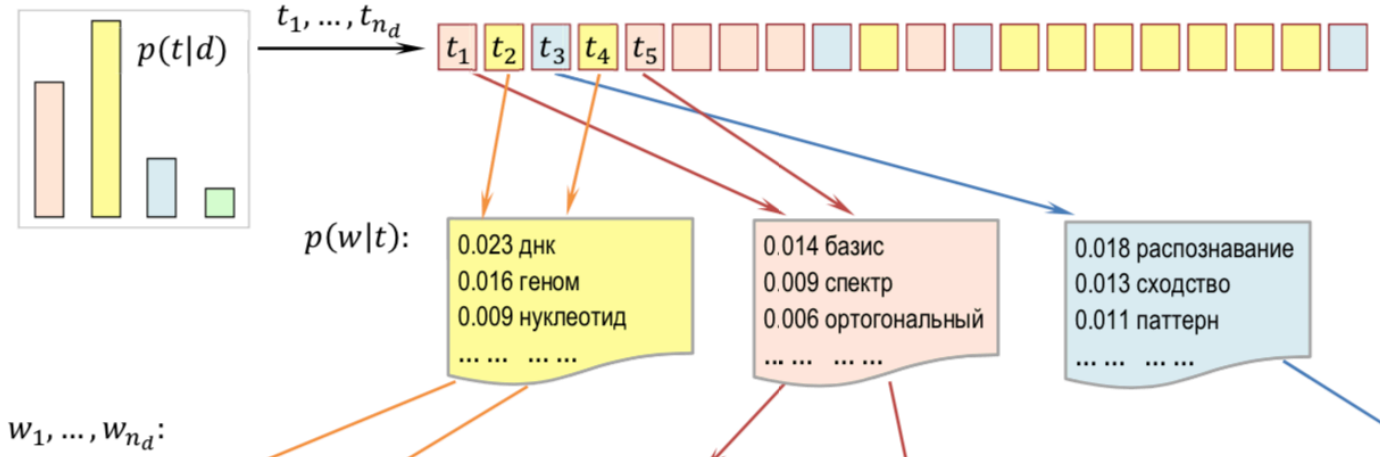
Задачи тематического моделирования



Методы тематического моделирования



Описание задачи

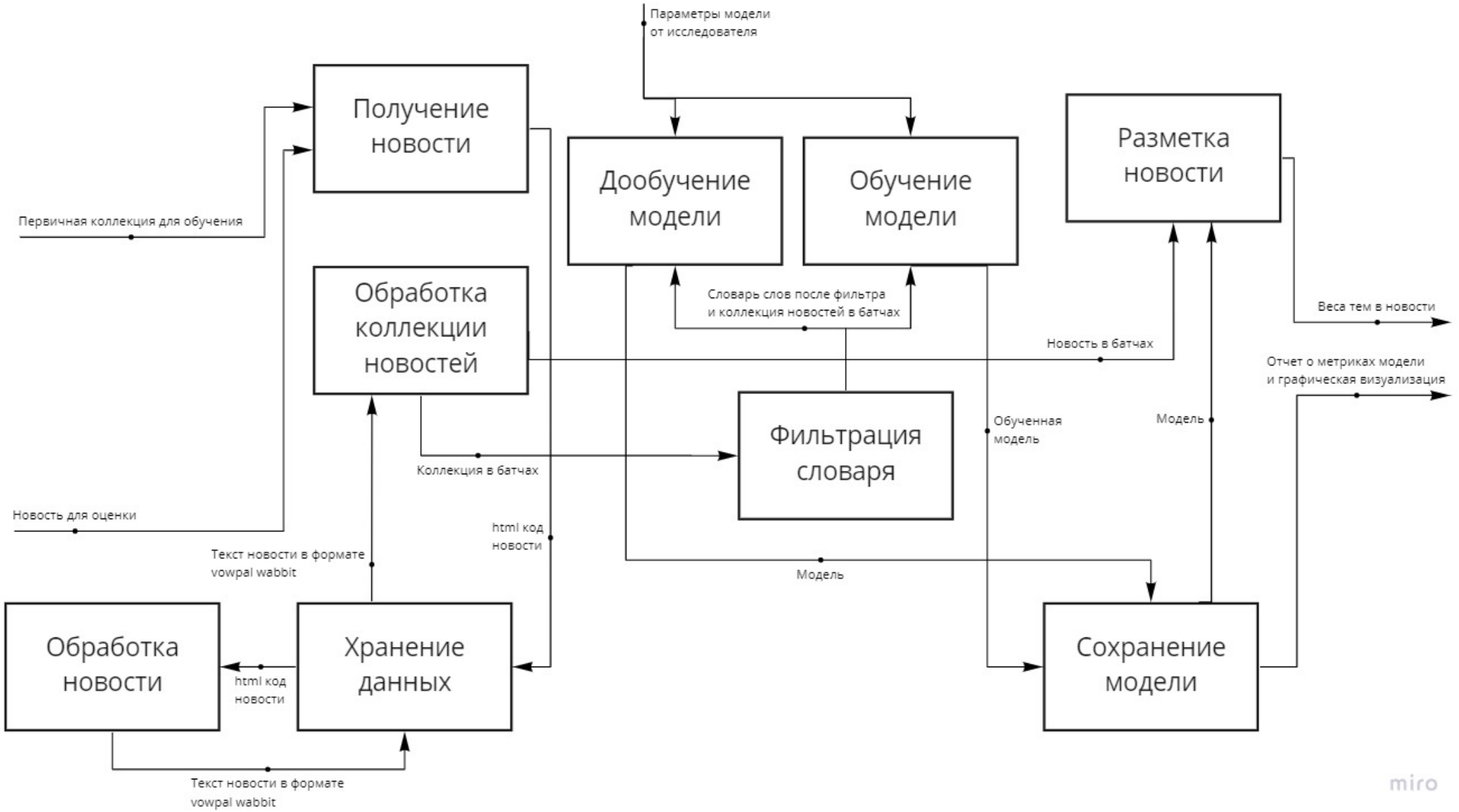


Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Изображение из работы //

Диаграмма метода

Диаграмма метода

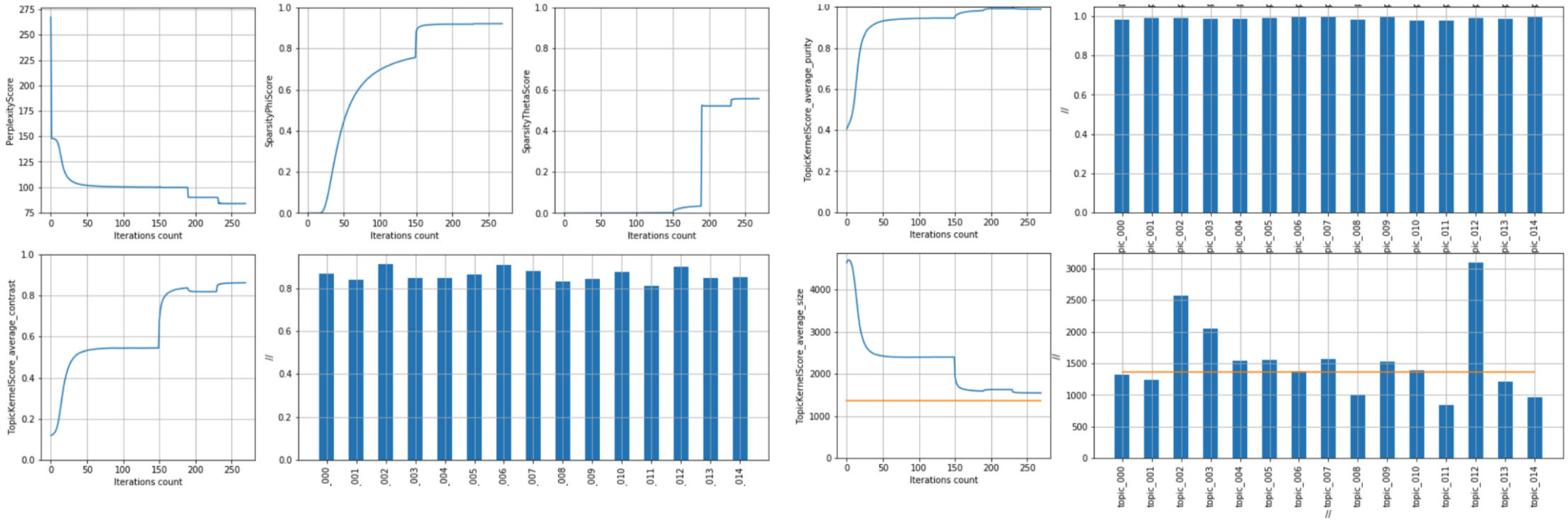


Список технологий

Python3 – основной язык программирования
SQLite – база данных
Beautifulsoup – для работы с html файлами
Pymystem3 – для лемматизации текстов
BigARTM – реализация базового алгоритма
Matplotlib – для визуализации метрик модели

Оценки

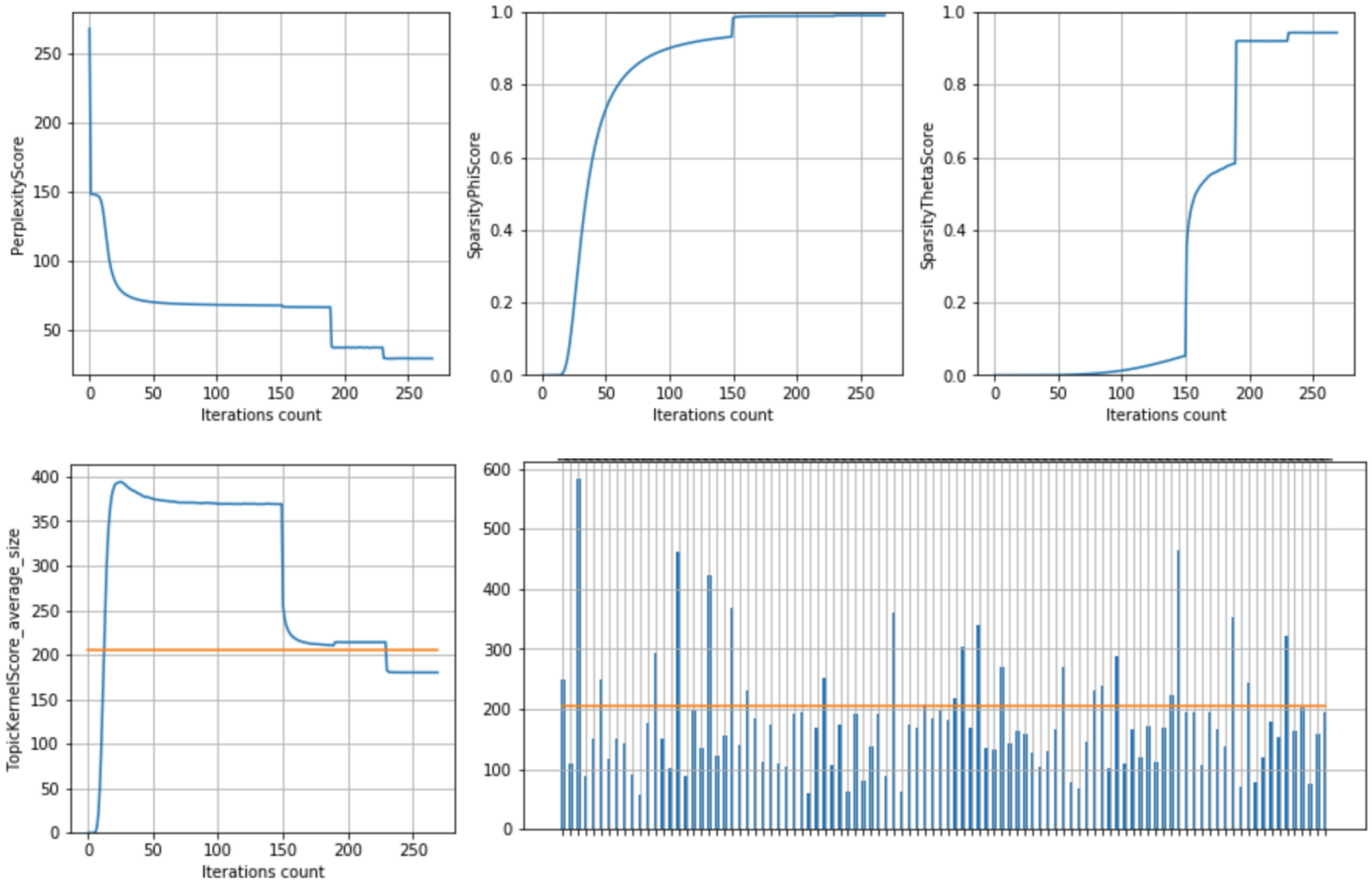
Для оценки модели была реализован функционал, выводящий всю необходимую статистику в удобном графическом представлении.



Исследование

Название	Перплексия	Разреженность матрицы слово-тема	Разреженность матрицы тема-документ	Средний контраст тем	Средняя чистота тем	Средний размер тем
Стадия исследования базовых значений параметров						
0_0_zona_23000_15t_plsa	5,8766	0,5250	0,0013	0,5531	0,9571	210
0_0_zona_23000_15t_plsa+sp	5,8674	0,8797	0,1112	0,6403	0,9842	187
0_0_zona_23000_15t_plsa+sp+st	5,5030	0,8798	0,7139	0,5578	0,9987	217
0_0_zona_23000_15t_plsa+sp+st+dp	5,4354	0,9472	0,7943	0,9830	0,9995	93
0_0_zona_23000_15t_plsa+(sp+st+dp)	5,4455	0,9423	0,7993	0,9523	0,9990	99
...	...	...	...	...	...	...
1_1_zona_23000_10t_plsa	6,0201	0,4458	0,0005	0,5395	0,9670	326
1_1_zona_23000_10t_plsa+sp+st+dp	8,6023	0,9540	0,8853	1,0000	1,0000	83
...	...	...	...	...	...	...
3_1_zona_23000_15t_plsa	1355,7546	0,6934	0,0000	0,5372	0,9306	2339
3_1_zona_23000_15t_plsa+sp+st+dp	1208,1539	0,9220	0,3538	0,8647	0,9861	1473
3_1_zona_23000_15t_plsa+(sp+st+dp)	1250,5999	0,9124	0,3297	0,3297	0,9767	1601
Стадия исследования количества тем						
...	...	...	...	...	...	...
4_0_ria_24000_15t_plsa	100,0766	0,7567	0,0018	0,5444	0,9453	2397
4_0_ria_24000_15t_plsa+sp+st+dp	84,1027	0,9225	0,5561	0,8611	0,9899	1549
4_1_ria_24000_50t_plsa+sp+st+dp	33,3470	0,9817	0,9034	0,9891	0,9999	372
4_2_ria_24000_100t_plsa+sp+st+dp	29,1980	0,9912	0,9436	0,9996	0,9999	180
4_3_ria_24000_150t_plsa+sp+st+dp	25,3347	0,9941	0,9611	1,0000	1,0000	120
...	...	...	...	...	...	...

Метрики



4_2_ria_24000_100t_plsa+sp+st+dp

Результаты

Пример хороших тем:

- Наука: ученый исследование коллега примерно лаборатория эксперимент изучение анализ изучать метод
- Футбол: футбольный футболист зенит спартак динамо цска поле локомотив болельщик забивать
- Литература: книга автор писатель написать название поэт литература библиотека рождаться литературный
- Деньги: продажа кредит капитал сделка актив сбербанк доля кредитный банковский прибыль

Пример плохих тем:

- предложение оценка точка необходимо существовать речь вариант особый зрение принцип
- подробно памятник письмо наследие охрана спецслужба справка реставрация сноудена запрос

Рекомендации

Заключение

В результате данной работы был разработан метод тематического моделирования новостей на русском языке.

Были решены следующие задачи:

- проанализированы существующие решения и выбран базовый алгоритм тематического моделирования для классификации новостей на русском языке;
- разработан программный продукт для сбора новостей на русском языке и подготовки данных для последующего анализа;
- разработан программный продукт для подготовки данных для последующего анализа;
- подобраны методы улучшения алгоритма и значений его параметров;
- обучена модель;
- проведена параметризация метода;
- проведена апробации метода;
- составлены рекомендаций о применимости предложенного метода.