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\\Введение

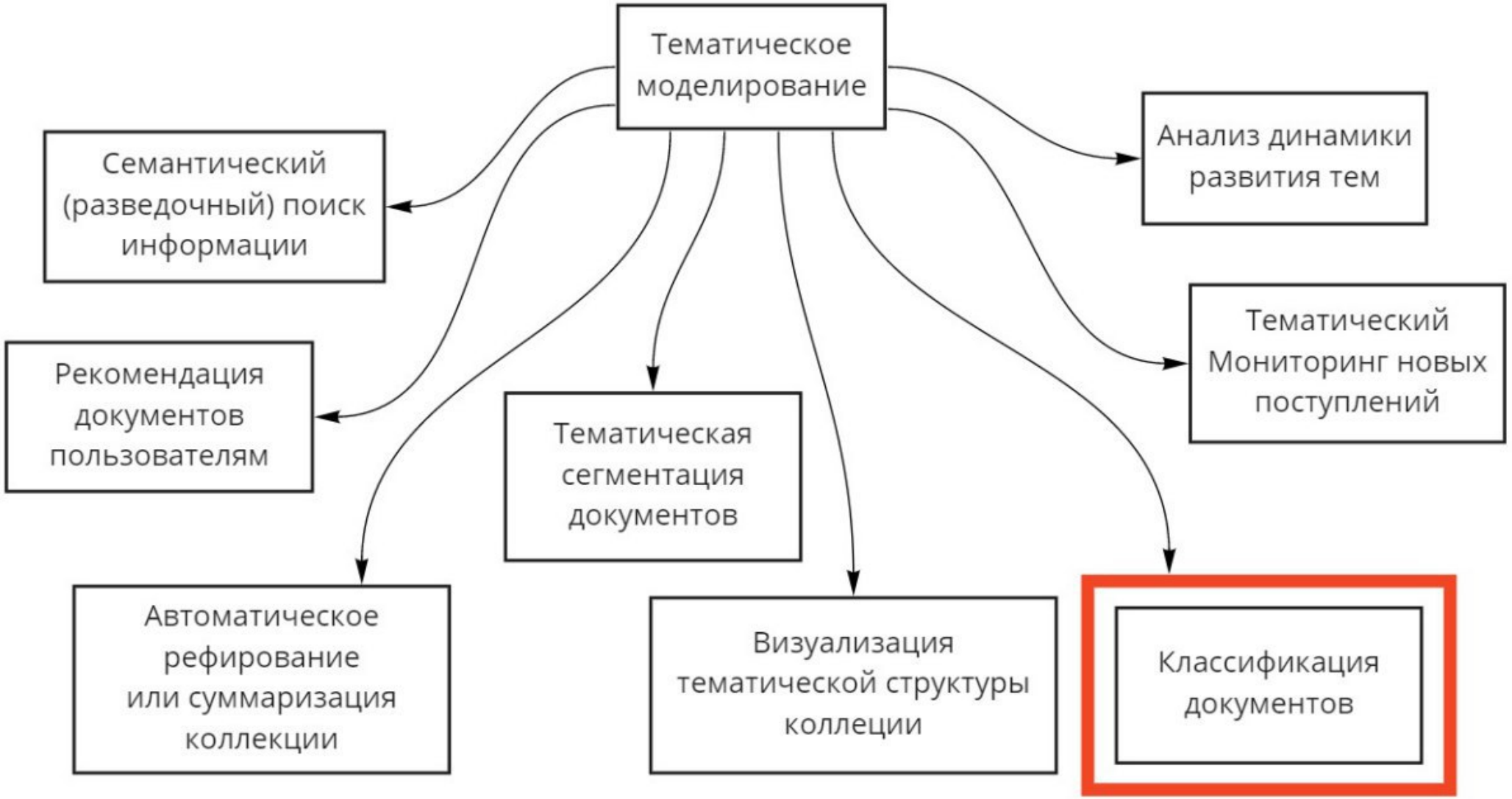
# Цели и задачи

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

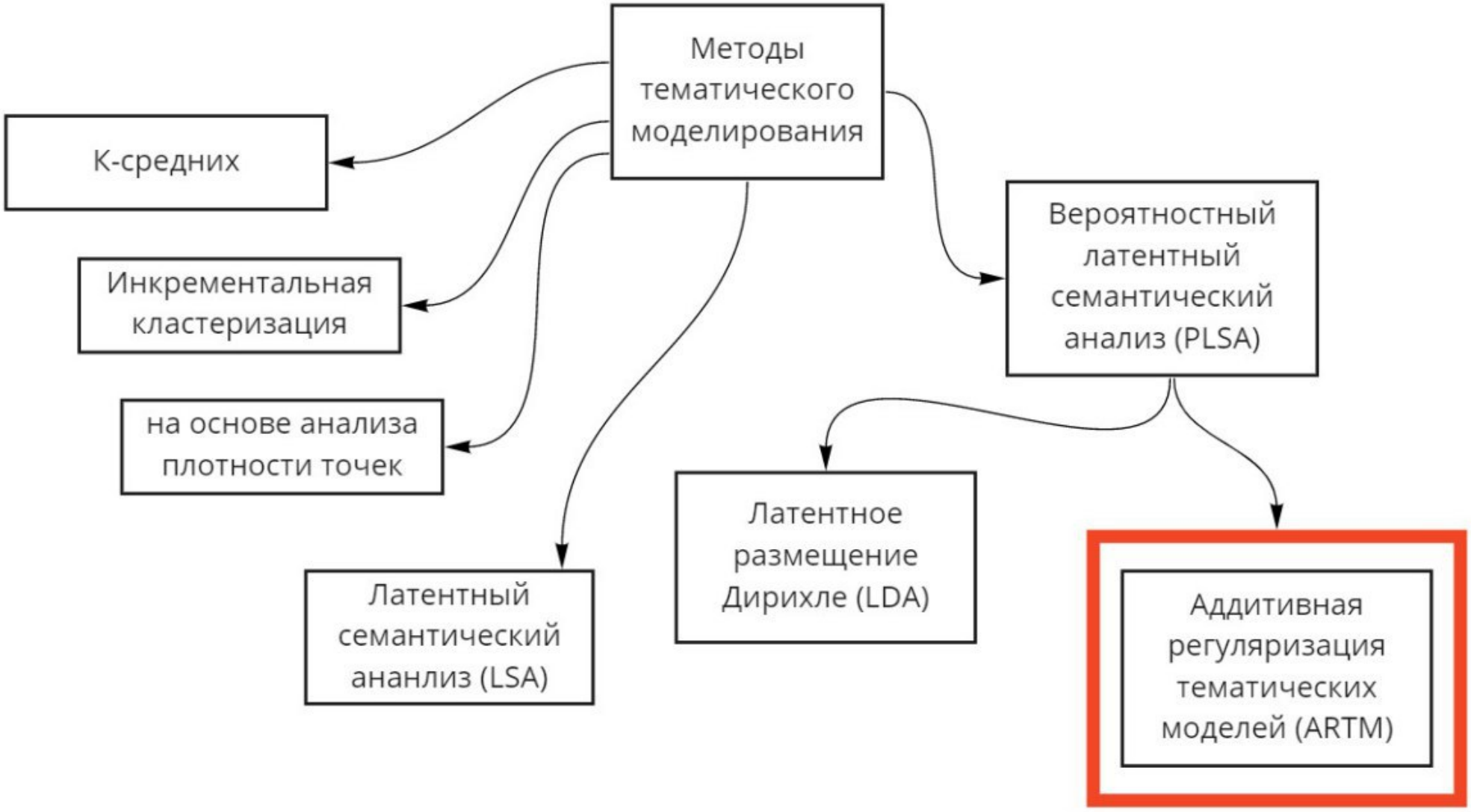
Задачи:

- 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.

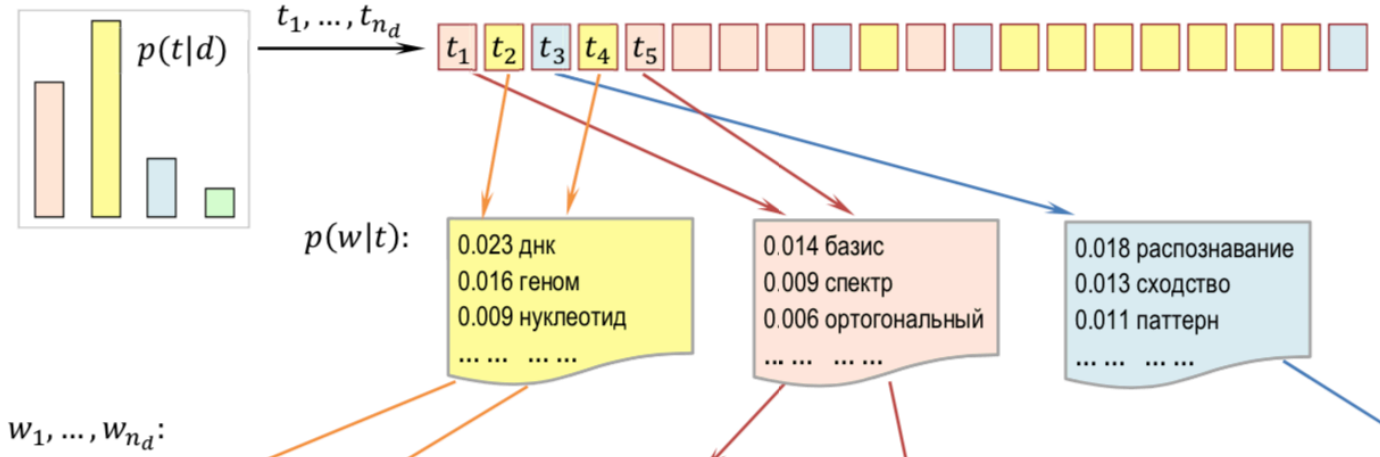
# Задачи тематического моделирования



# Методы тематического моделирования



# Описание задачи

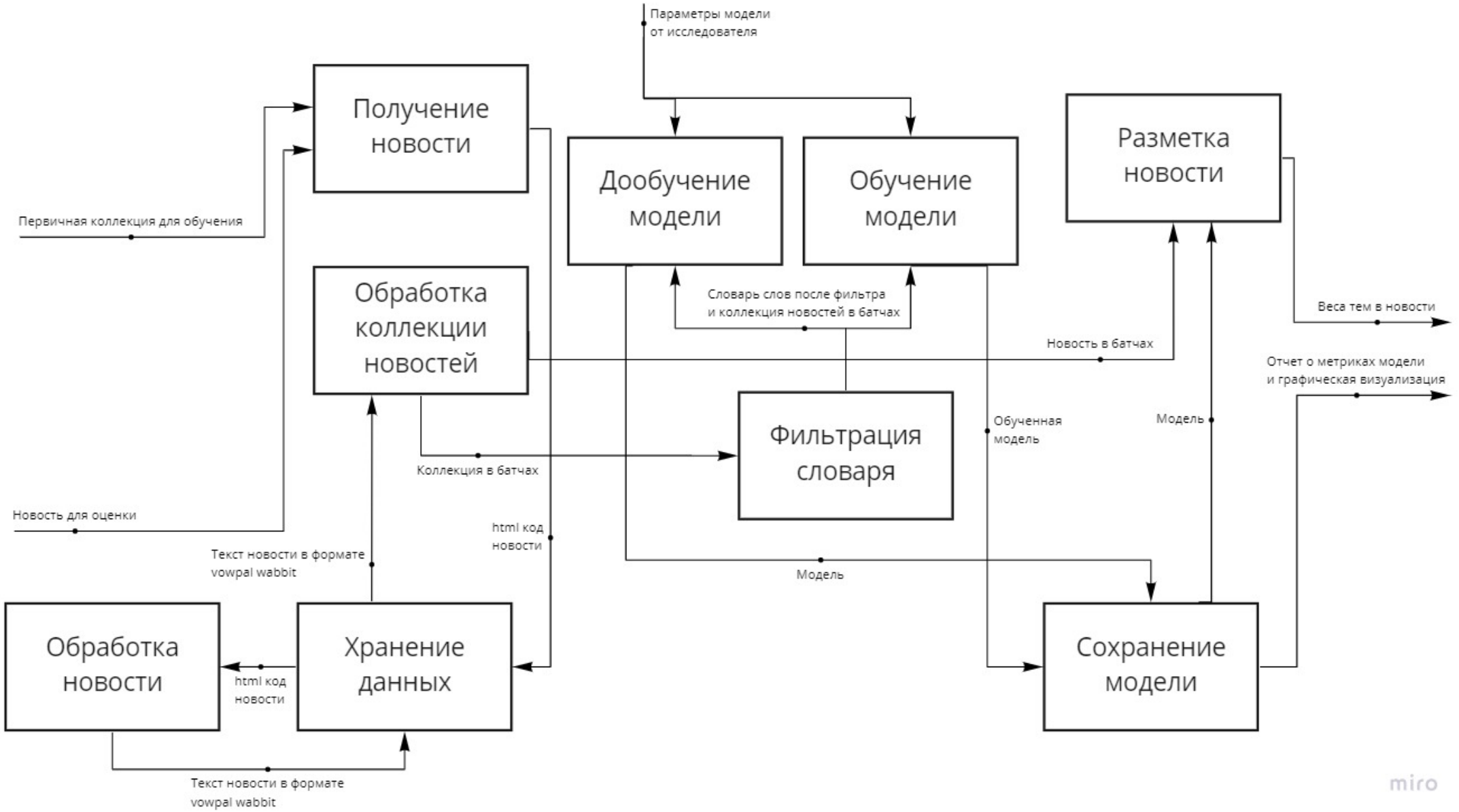



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным 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

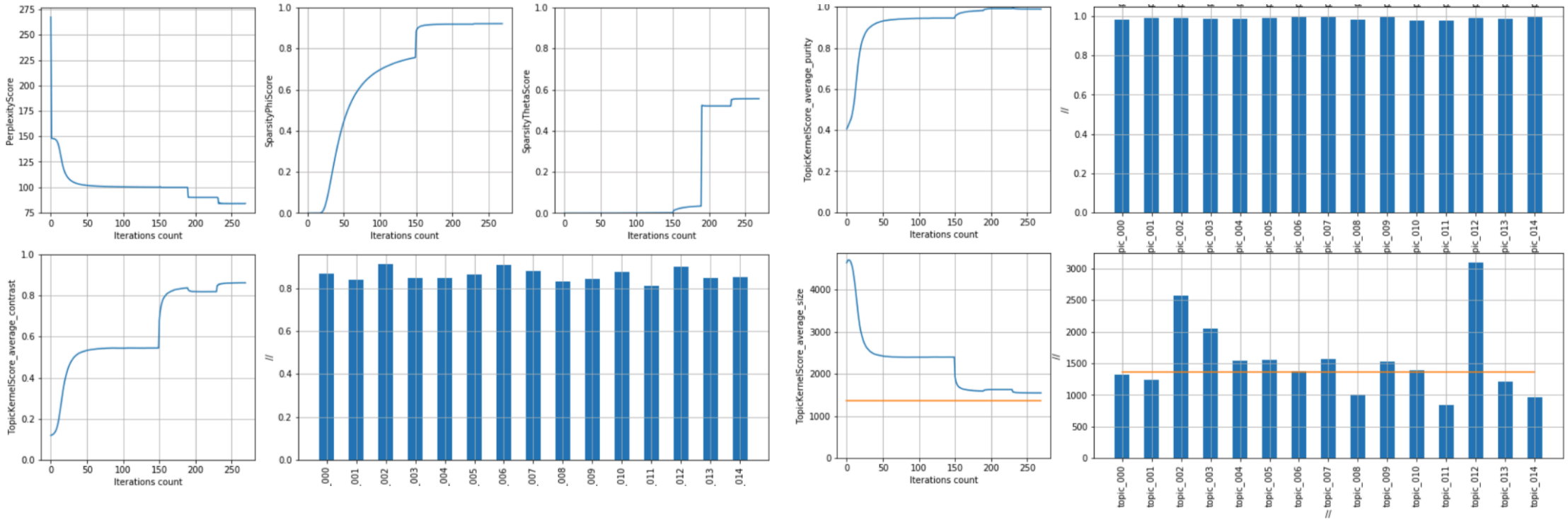


# Список технологий

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 графическом представлении.



# Результаты

# Результаты

# 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.