



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

Актуальность

При росте объема новостных потоков актуальна задача автоматизации выделения тем новостей для последующей группировки и анализа.

Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

Задача тематического моделирования - описание такой модели и применение ее к данным.

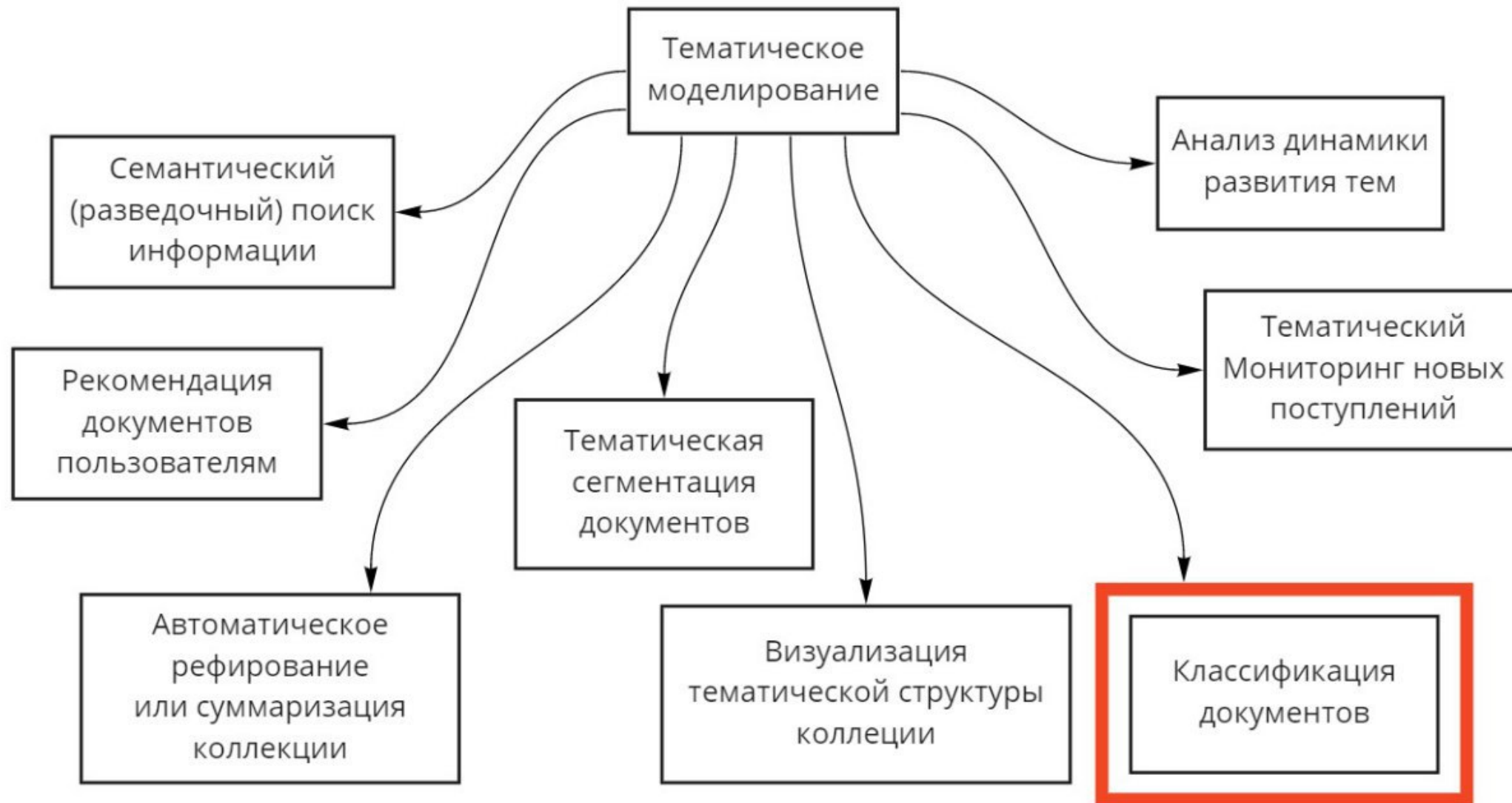
Цели и задачи

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

Задачи:

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

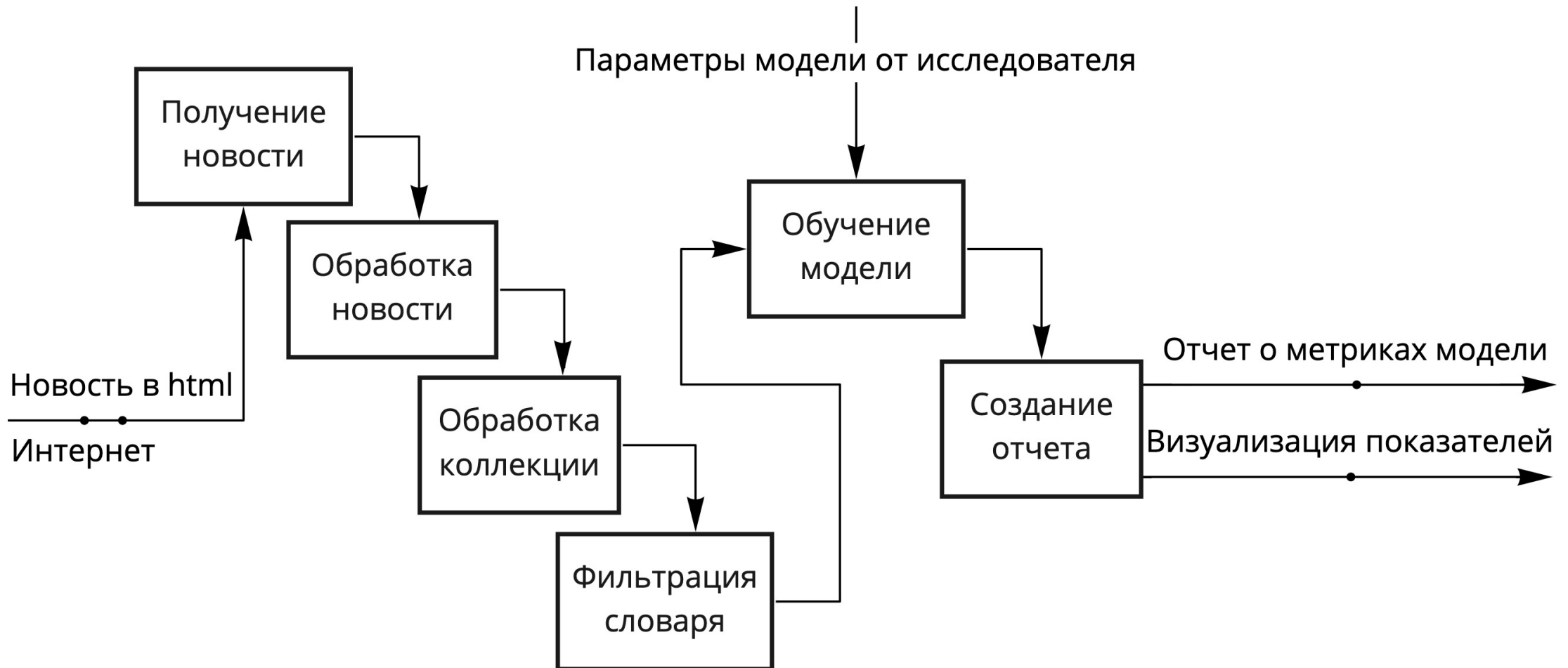
Задачи тематического моделирования



Концептуальная схема

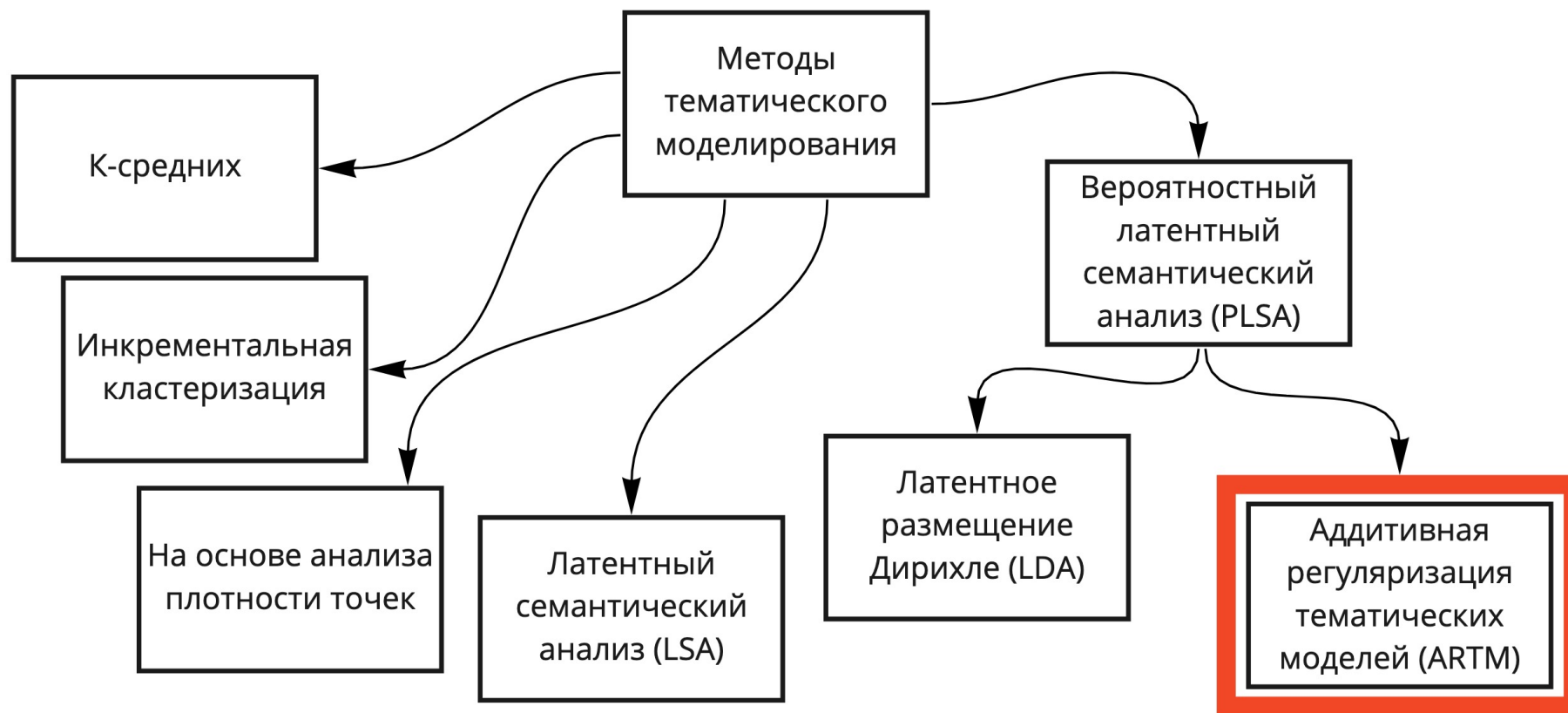


Концептуальная схема



IDEF0 уровня 1

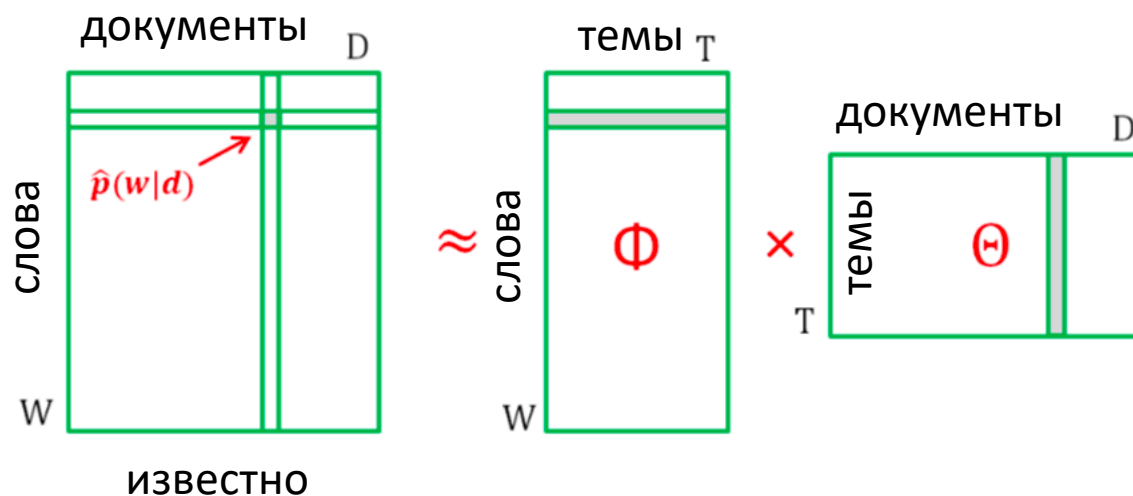
Методы тематического моделирования



В качестве базового метода выбран ARTM.

Следует выбрать регуляризаторы, их порядок применения и веса.

Модель порождения текста



Необходимо найти такие матрицы Φ и Θ , что бы:

- максимизировать сумму вероятностей $p(w|d)$;
- была высока вероятность интерпретируемости тем экспертом.

Регуляризаторы

Регуляризатор – ограничение зависящее от параметров модели.

В данной работе были рассмотрены три регуляризатора.

Разреживающий или сглаживающий регуляризатор матрицы слово-тема

$$R(\Phi) = \tau \sum_{t \in T} \sum_{w \in W} \ln \phi_{wt} \rightarrow \max.$$

Влияет на количество нулевых значений в матрице слово-тема.

Разреживающий или сглаживающий регуляризатор матрицы тема-документ

$$R(\Theta) = \tau \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \rightarrow \max.$$

Влияет на количество нулевых значений в матрице Θ тема-документ.

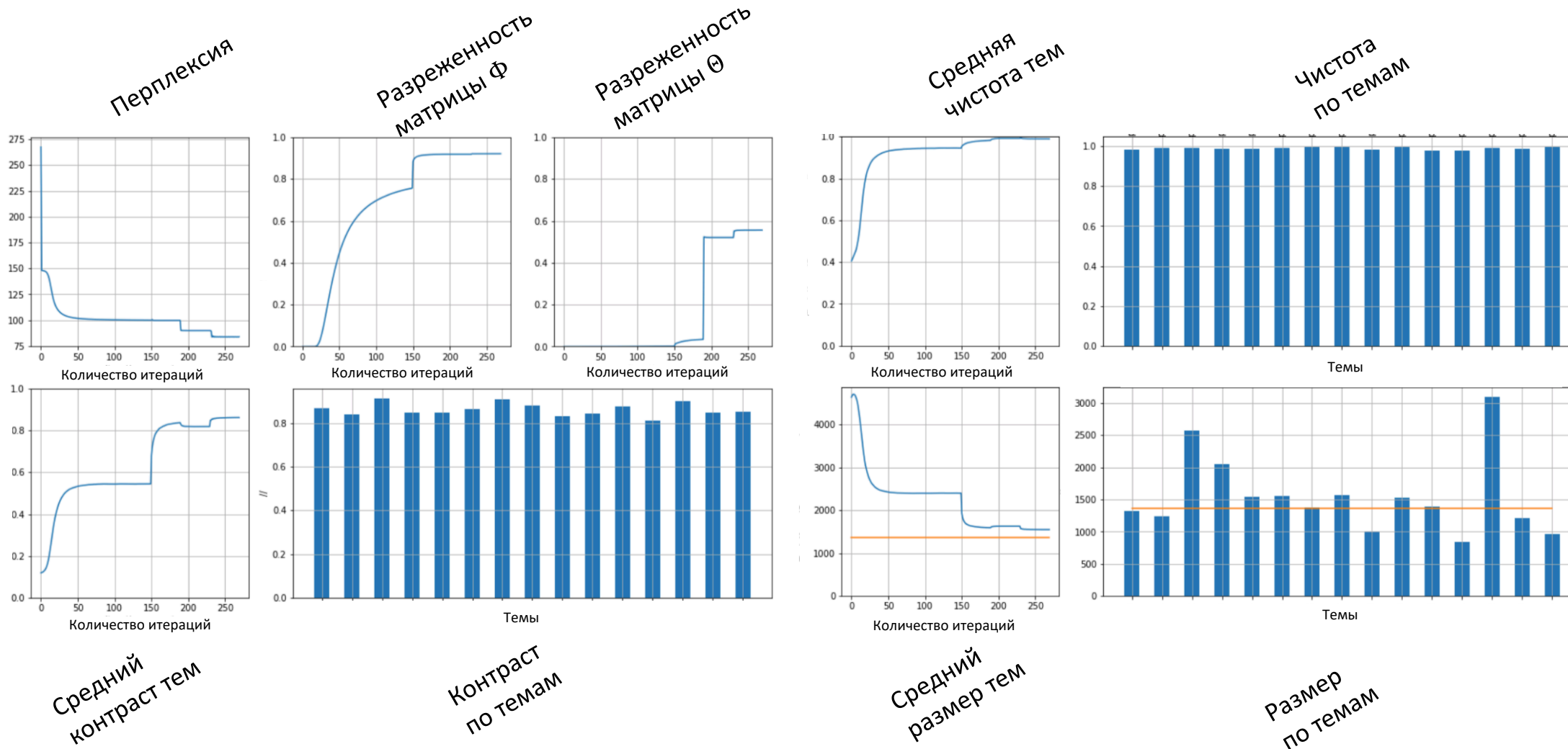
Регуляризатор декоррелирования тем

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max$$

Влияет на попарную корреляцию тем как столбцов матрицы Φ слово-тема.

Оценки

Для оценки модели была предложена визуализация статистики метрик.



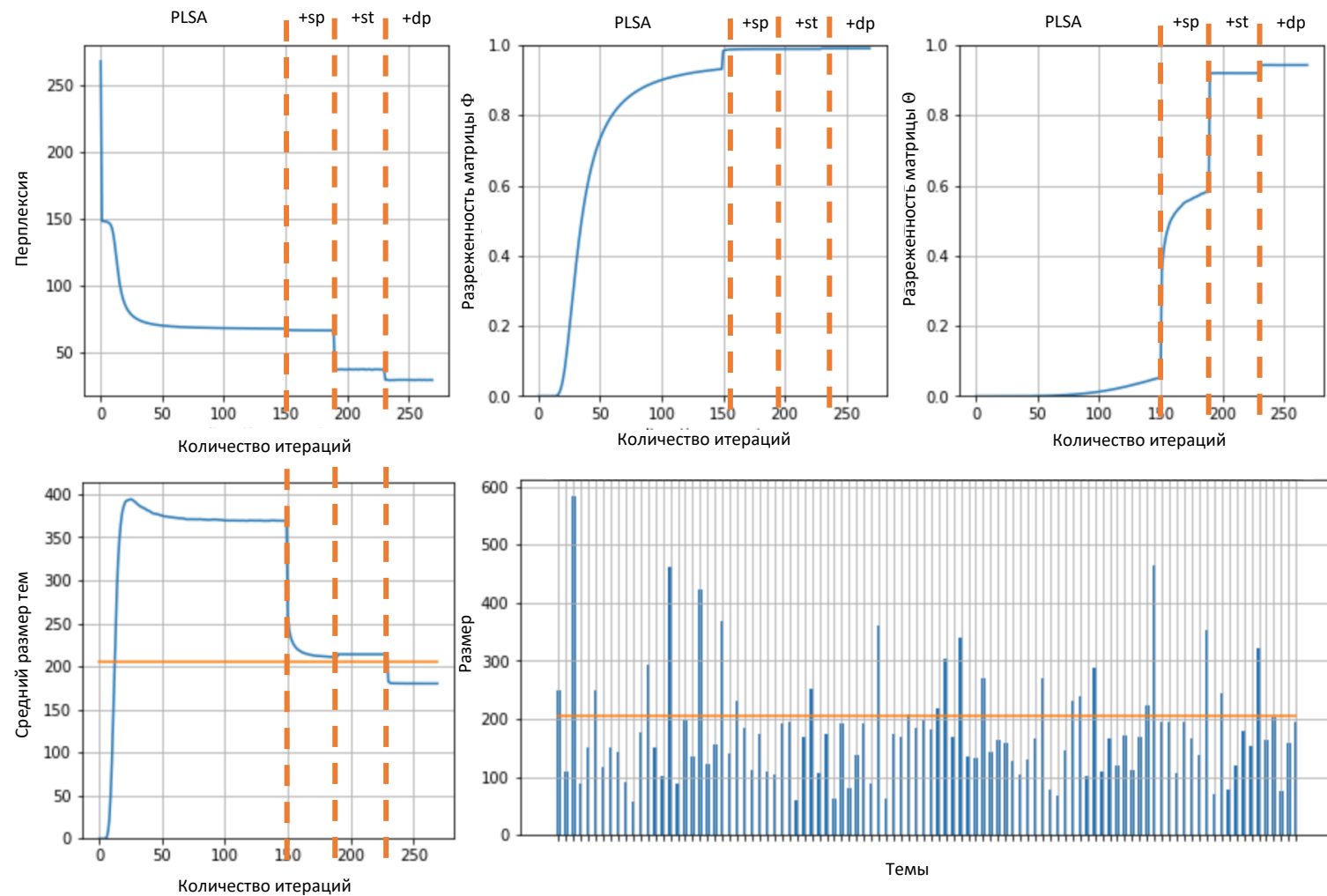
Исследование

Название модели коллекция + регуляризаторы	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
...

Рекомендуется
применять
регуляризаторы
последовательно.

Рекомендуемый
порядок
регуляризаторов:
sp, st, dp.

Метрики для модели



4_2_ria_24000_100t_plsa+sp+st+dp

Результаты: 10 наиболее релевантных теме слов

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 плохой темы:

- подробно памятник письмо наследие охрана спецслужба справка реставрация сноудена запрос

Рекомендации

- При скачивании данных из сети интернет рекомендуется сохранять их в базу данных, если есть такая возможность.
- Процесс обработки данных крайне желательно выстраивать в несколько потоков при этом необходимо решить проблему с заблокированной базой при одновременном обращении.
- Все процедуры по получению и обработке данных следует организовать в виде независимых сервисов, которые можно останавливать и снова запускать в произвольном порядке без ущерба для данных.
- Массивы данных, взятые из интернета не так сложно получить самостоятельно. Часто это лучший путь. Данные будут чище и, возможно, будет возможность использовать что-то, чего нет у других.
- Стоит отделить процедуру подготовки коллекции для обучения и процедуру предварительной подготовки каждого документа по отдельности.
- Следует хранить дату обработки текста, что бы при изменении в процедуре обработки можно было перезапустить процесс на старых записях, а не на всей базе данных.
- Необходимо максимально аккуратно обрезать словарь. Вместе с часто используемыми словами легко отфильтровать ценные для модели данные.
- При обучении модели ARTM следует сначала пройти по коллекции до того как сойдется (перестанет изменяться) перплексия, и только после этого начать добавлять регуляризаторы.
- Регуляризаторы стоит добавлять по одному. После чего продолжать обучение пока модель не сойдется.
- Коэффициенты при регуляризаторах стоит сначала выбирать небольшими по модулю, постепенно увеличивая их значение.
- Количество тем в рассмотренном примере задавалось исследователем, но стоит попробовать указывать избыточное количество тем и уменьшать их еще одним не рассмотренным в данной работе регуляризатором.
- Если у темы слишком большой размер по сравнению с другими темами – необходимо проверить ее на фоновые слова, характерные для любой темы.

Заключение

Разработан метод тематического моделирования новостей на русском языке.

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