Заключение

члена диссертационного совета Лукашевич Натальи Валентиновны
по содержанию диссертации Булатова Виктора Геннадьевича
«Методы оценивания качества и многокритериальной оптимизации
в библиотеке TopicNet»,
представленной на соискание ученой степени
кандидата технических наук по специальности 05.13.18
Дата защиты 28.12.2020

Оценка соответствия диссертации требованиям Положения о присуждении ученых степеней кандидата наук, доктора наук в МФТИ (далее - Положение):

1. **Актуальность тематики диссертации:**

Настоящее время характеризуется объемными потоками неструктурированной текстовой информации, имеется большое количество задач анализа поступающих текстовых данных. Известным инструментом анализа текстовых коллекций являются подходы на основе вероятностного тематического моделирования. Для экспертного анализа особенно важно получение качественных интерпретируемых тематик, обсуждаемых в анализируемой коллекции. Однако для этого может потребоваться применение достаточно сложных настроек параметров вероятностных тематических моделей.

Диссертация Булатова В.Г. посвящена вопросам оценки интерпретируемости порождаемых тем, автоматическим оценкам интерпретируемости, а также проблеме автоматического подбора параметров тематического моделирования на основе автоматических оценок.

2. **Научная новизна выносимых на защиту результатов:**

Следующие результаты работы являются новыми:

1. Предложена новая методология многокритериального выбора моделей на основе концепций «дерева экспериментов», «кубов гиперпараметров» и «рецептов моделирования» в рамках теории аддитивной регуляризации тематических моделей (ARTM).
2. Предложен новый способ построения иерархических тематических моделей с разными весами модальностей на разных уровнях иерархии.

3. **Теоретическая и практическая значимость диссертационной работы:**

Теоретическая значимость работы состоит в развитии теории аддитивной регуляризации тематических моделей (ARTM). Вводятся понятия внутритекстовой когерентности, относительных и абсолютных коэффициентов регуляризации, фактора балансировки, дерева экспериментов, куба гиперпараметров, рецепта моделирования.

Практическая значимость работы состоит в том, что предложенные подходы и методы реализованы в библиотеке тематического моделирования с открытым кодом TopicNet, которая может быть использована для решения различных прикладных задач анализа текстовых и транзакционных данных.

4. **Полнота опубликования основных результатов диссертации в рецензируемых научных изданиях**:

Основные результаты диссертации опубликованы в рецензируемых научных изданиях.

5. **Вопросы и замечания** :

По работе имеются следующие замечания:

1. Предложенный в работе подход к анализу качества тематических моделей на основе анализа внутритекстовых фрагментов очень похож на подход, описанный в работе Lund, J., Armstrong, P., Fearn, W., Cowley, S., Byun, C., Boyd-Graber, J., & Seppi, K. (2019, July). Automatic Evaluation of Local Topic Quality. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 788-796), которая, олнако, не цитируется.
2. На стр. 109. описывается процедура обработки именованных сущностей в задаче классификации интентов. Не очень понятно, как можно извлекать названия компаний (организаций), названий улиц (городов) на основе размеченной коллекции PERSONS-1000, в которой размечены только имена людей (тег PERSON). Почему модуль морфологического анализа pymorphy2 упоминается как инструмент извлечения именованных сущностей на основе правил. На стр. 116 упоминается извлечение именованных сущностей по словарю, ранее никаких словарей для именованных сущностей не встречалось.
3. Стр. 117 указывается, что «прием ведёт к значительному улучшению качества на всех трёх проверочных выборках». Вместе с тем на двух коллекциях показан рост на 1 процентный пункт. В работе не обосновывается, что это улучшение значительно.

6. **Общая характеристика диссертации (не включает резолютивную часть):**

Материал диссертации изложен последовательно и логично. Структурные составляющие диссертационной работы (введение, главы, заключение, библиографический список, приложения) позволяют получить полное представление о проделанных исследованиях и полученных результатах. Работа содержит обширный обзор подходов к анализу качества вероятностных тематических моделей.

Считаю, что диссертация Булатова В.Г. представляет собой существенный вклад в развитие методов автоматического анализа потоков неструктурированной текстовой информации на основе инструментов вероятностного тематического моделирования.

Ведущий научный сотрудник
Лаборатории анализа информационных ресурсов
НИВЦ МГУ имени М.В. Ломоносова

д.т.н. Лукашевич Наталья Валентиновна
Телефон: 8-9261446163
E-mail: louk_nat@mail.ru

08.12.2020

Подпись Лукашевич Н.В. заверяю

Ведущий специалист отдела кадров