Историческая информатика
Правильная ссылка на статью:
Галушко И.Н. — Применение тематического моделирования для оптимизации процесса поиска релевантных исторических документов (на примере биржевой прессы начала XX в.) // Историческая информатика. – 2023. – № 2. DOI: 10.7256/2585-7797.2023.2.43466 EDN: SKBPNS URL: https://nbpublish.com/library_read_article.php?id=43466

Применение тематического моделирования для оптимизации процесса поиска релевантных исторических документов (на примере биржевой прессы начала XX в.)

Галушко Илья Николаевич

магистр, кафедра исторической информатики, исторический факультет, Московский государственный университет имени М.В. Ломоносова (МГУ)

119234, Россия, г. Москва, ул. Ломоносовский Проспект, 27, корп.4

i.galushko15@gmail.com

Статья из рубрики "Новые методы и технологии обработки исторических источников"

DOI:

10.7256/2585-7797.2023.2.43466

EDN:

SKBPNS

Дата направления статьи в редакцию:

30-06-2023

Аннотация: Ключевой задачей представленной статьи является апробация методики анализа информационного потенциала коллекции исторических источников с помощью тематического моделирования. Некоторые современные коллекции оцифрованных исторических материалов насчитывают десятки тысяч документов, и на уровне отдельного исследователя охват всего доступного наследия представляется затруднительным. Вслед за рядом исследователей мы предполагаем, что тематическое моделирование может стать удобным инструментом предварительной оценки содержания коллекции исторических документов; инструментом отбора только тех документов, в которых присутствует информация, релевантная поставленным исследовательским задачам. В нашем случае в качестве основной коллекции исторических документов была выбрана подборка газеты «Биржевые ведомости». На данном этапе мы можем подтвердить, что в рамках нашего исследования применение тематического моделирования оказалось продуктивным решением для оптимизации процесса поиска исторических документов в объемной коллекции оцифрованных исторических материалов. В то же время необходимо подчеркнуть, что в нашей работе тематическое

моделирование применялось исключительно как прикладной инструмент ускорения поиска и первичной оценки информационного потенциала коллекции документов через анализ выделенных топиков. Наш опыт показал, что по крайней мере для «Биржевых ведомостей» тематическое моделирование с использованием LDA не позволяет делать выводы с позиции применяемой нами методологии содержательного анализа. Данные наших моделей слишком фрагментарны, их можно использовать только для первичной оценки тематик информации, содержащейся в источнике.

Ключевые слова:

тематическое моделирование, латентное размещение Дирихле, Биржевые ведомости, поведенческие финансы, обработка естественного языка, распознавание исторических документов, исторические газеты, поиск исторических документов, машинное обучение, фондовый рынок

Ключевой задачей представленной статьи является апробация методики анализа информационного потенциала коллекции исторических источников с помощью тематического моделирования. Некоторые современные коллекции оцифрованных исторических материалов насчитывают десятки тысяч документов (как, например, «Электронная библиотека исторических документов», созданная Российским историческим обществом (РИО), содержит 294 тысячи распознанных исторических документов [1] – и на уровне отдельного исследователя охват всего доступного наследия представляется затруднительным. Вслед за рядом исследователей [2] мы предполагаем, что тематическое моделирование может стать удобным инструментом предварительной оценки содержания коллекции исторических документов; инструментом отбора только тех документов, в которых присутствует информация, релевантная поставленным исследовательским задачам.

Наше исследование, для которого и была разработана описываемая в статье методика, посвящено изучению доходности ценных бумаг на Санкт-Петербургской фондовой бирже в начале XX в. с позиции поведенческих финансов. Нас интересовали принципы инвестиционной оценки публичных компаний – как определялись приемлемые или недостаточные уровни капитализации; как определялись ценные бумаги, представляющие хороший выбор для помещения капиталов, насколько широко данные методики (если они существовали) применялись в практике биржевой торговли. В качестве одного из основных источников была выбрана газета «Биржевые ведомости», в ежедневных выпусках которой велась биржевая колонка, где печатался комментарий хроникера, в котором описывался настрой участников торгов и нередко приводился подробный анализ текущей ситуации в экономике Российской империи. В колонках «Биржевых ведомостей» часто встречаются аналитические заметки о доходности ценных бумаг: под какой процент размещается очередная эмиссия государственных долговых бумаг; на каком уровне относительно номинала торгуются эти бумаги; соответствует ли предлагаемый процент актуальной статистике денежного рынка и как объяснить курсовую динамику последних дней. Для исследования было решено собрать коллекцию таких заметок, чтобы на ее основе выделить устойчивые аналитические паттерны, характерные для биржевой прессы в вопросах, касающихся доходности ценных бумаг. Мы воспользовались материалами оцифрованного комплекта «Биржевых ведомостей» с сайта Российской национальной библиотеки (447 номеров за 1905 и 1913 года, утренние и вечерние выпуски). В рамках общего исследования доходности ценных бумаг на

фондовом рынке Российской империи нас интересовал ограниченный набор проблем, связанных с поведенческими и институциональными аспектами функционирования фондового рынка Российской империи. Содержание «Биржевых ведомостей», напротив, разнообразно. И, если не считать колонку биржевого хроникера, то встречаются номера, полностью лишенные нужной нам информации. Содержание таких номеров заполнено военными новостями, театральными и литературными обзорами, экономическими рассуждениями небиржевого характера и другими подобными статьями широкого профиля (см. рис. 1 и 2). И в этом контексте логично обратиться к возможностям тематического моделирования в качестве прикладного инструмента для автоматического поиска тех номеров (страниц) из нашей коллекции оцифрованных газетных материалов, которые содержат информацию, касающуюся особенностей функционирования рынка ценных бумаг.

Рисунок 1. Пример организации страницы номера «Биржевых ведомостей».

Рисунок 2. Пример искомого текста. Колонка «Счастливица». Рассказ о том, как некая женщина нашла в магазине сверток с акциями, которые потерял банкир Борисов (Выпуск № 13621 от 28 июня (11 июля)).

* * *

Тематическое моделирование — это метод машинного обучения без учителя, применяемый для определения основных тем коллекции документов (или тем предложений одного документа, который рассматривается в таком случае как совокупность предложений) на основе выделения топиков (про разницу понятий «тема» и «топик» см. Приложение 1). Как правило, топик представляет собой взвешенный по вероятности список слов, которые вместе выражают общее содержание предполагаемой темы [2]. Чем выше коэффициент слова, тем большее значение модель придает этому слову при формировании топика. Так, в Таблице 2 представлен пример двух топиков, определенных нашей моделью для третьей страницы выпуска «Биржевых ведомостей» от 29 апреля 1913 г. (№13521):

Таблица 1. Примеры двух определенных моделью топиков, представленных набором ключевых слов и их вероятностями (см. Приложение 2.1)

Топики
(**1,** '0.023*"склад" + 0.021*"акция" + 0.016*"предприятие" + 0.012*"общество" + 0.012*"пароходных" + 0.012*"товарищество" + 0.012*"транспорт" + 0.012*"страх" + 0.012*"железный" + 0.007*"баланс"')
(**2**, '0.015*"городской" + 0.013*"бумага" + 0.012*"заем" + 0.011*"дорога" + 0.007*"капитал" + 0.007*"облигация" + 0.007*"общество" + 0.007*"акция" + 0.006*"специальный" + 0.006*"иметь"')

Одним из наиболее популярных методов тематического моделирования, используемых в настоящее время, является латентное распределение Дирихле (LDA), которое представляет собой «генеративную вероятностную модель для коллекций дискретных данных, таких как текстовые корпуса» [3]. Этот метод используется в рамках Digital Humanities для извлечения тем из набора текстов [4]. В этой модели документ (в нашем случае – отдельная страница газетного номера «Биржевых ведомостей») представляет собой смесь топиков, а топик — распределение вероятностей по словарю. Под словарем понимается список всех слов изучаемой коллекции документов – именно словарь задает модели пространство слов, в котором нужно распределить документы таким образом, чтобы сформировать заданное исследователем количество топиков. Важной особенностью LDA является заранее определенное количество топиков, которое устанавливается до начала обучения. Удобной функцией моделей LDA является установление весов, равных 0.001 для слов, определяющих избыточные топики. То есть, если исследователь изначально установил для модели поиск 5 топиков, а по результатам тематического моделирования для трех тем модель установила разные веса слов, изменяющиеся в границах, например, от 0.005 до 0.0021, а для двух других тем – вес 0.001 для всех слов – значит, LDA при заданных настройках не может распределить текст более чем на 3 топика. В ходе наших экспериментов мы установили, что для решения наших исследовательских задач достаточным является порог в 5 топиков.

Создание словаря может дополняться применением различных методов предобработки текстов. В нашей работе использовался классический для компьютерной лингвистики метод лемматизации – приведения всех слов к основной форме. Другой широко применяемый метод предобработки подразумевает поиск и добавление в словарь модели биграмм – слов с двойной основой, выделяемых в тексте по частоте совместной встречаемости (например, «акционерный_капитал», «государственный_рента»). Данный подход мы тестировали изначально, когда планировали создавать LDA модель всей

имеющейся коллекции. При таком подходе в качестве одного документа рассматривается целый номер биржевой газеты. Однако результаты оказались крайне непрактичными – ввиду слишком широкого разброса тем на один выпуск газеты мы не смогли создать интерпретируемые модели. Добавление фильтров (TF-IDF (см. Приложение 1), изменение параметров обучения (например, размера батча – порции данных для итерации обучения), увеличение количества тем (разные варианты в диапазоне от 5 до 30 топиков) – не смогли улучшить ситуацию. Тогда мы решили создавать LDA модели для отдельных номеров газеты, рассматривая их в качестве коллекции предложений. Результаты улучшились, но все еще остались недостаточными для использования тематического моделирования в качестве стабильного поискового алгоритма. Последним и наиболее продуктивным вариантом оказалось построение LDA для каждой отдельной страницы в нашей коллекции «Биржевых ведомостей». Подобный подход позволил нам создать алгоритм отбора релевантных для исследования материалов на основе тематического моделирования, но ввиду ограниченного словаря нам пришлось отказаться от создания биграмм и фильтрации по TF-IDF, что в конечном итоге никак не повлияло на качество итогового моделирования, поскольку мы не ставили задачу исчерпывающего семантического анализа; выявление тем на основе одинарных слов оказалось достаточным для отбора всех номеров газеты, обладающих информацией, релевантной нашим исследовательским задачам. В качестве основной библиотеки для реализации программного кода был выбран Python-модуль Gensim, являющийся одним из наиболее востребованных пакетов NLP-инструментов в мире современной науки о данных[5].

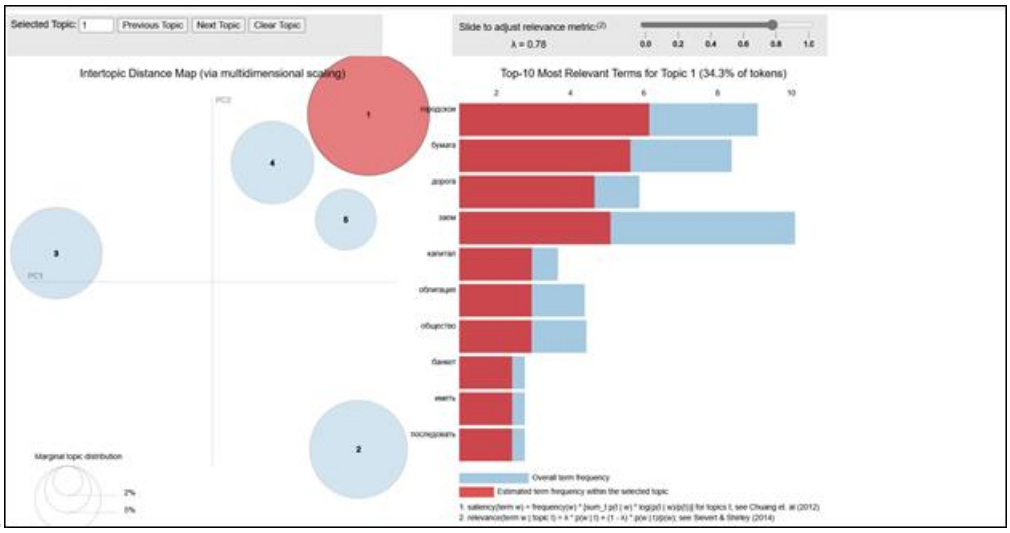


Рисунок 3. Визуализация LDA модели, созданной для 3-ей страницы выпуска №13521 «Биржевых ведомостей» от 29 апреля 1913 г.

В рамках тематического моделирования эксперт должен сам определить название для топика таким образом, чтобы оно отражало семантику отдельных слов, формирующих тему. Для примера по рисунку 8 мы можем однозначно утверждать, что слова из топика №1 явно тяготеют к теме городских займов, на что указывают такие слова, как «городской», «заем», «облигация», «капитал». Архитектура существующих LDA моделей позволяет быстро анализировать содержание выделенных тем на основе поиска по ключевым словам с учетом их «веса». В качестве алгоритма поиска номеров газеты, содержащих нужную нам информацию, мы решили использовать поиск по словам, формирующим топики, с пороговым значением весов слова > 0.01. В качестве смысловых единиц, маркирующих присутствие биржевых сведений в моделируемом документе, мы использовали слова, перечисленные в Таблице 1. Перечень паттернов Fuzzy-match и соответствующих им форм слов [6].

Для всей отобранной коллекции текстов «Биржевых ведомостей» было создано 2411 LDA моделей. Из них наш алгоритм поиска определил в группу «содержащих биржевую информацию» – 457. Разумеется, в значительной степени к таковым относятся страницы с колонкой биржевого хроникера. В качестве определенного доказательства применимости LDA-моделей в подобных источниковедческих задачах отметим, что анализ случайной выборки из 100 номеров «Биржевых ведомостей» показал, что ни одна колонка хроникера не была пропущена поисковым алгоритмом – каждая из них была помещена в сводную таблицу. В качестве иллюстрации мы приведем малую выборку из полей итоговой таблицы, включающую только те страницы и номера, в которых удалось обнаружить биржевые сведения вне колонки биржевого хроникера. Всего таких — 29 моделей. В Таблице 2 представлены примеры определенных моделью наборов топиков. Таблица 3 содержит краткое описание материалов, найденных на страницах, перечисленных в Таблице 2. В приложении мы поместили фотографии страниц, представленных в обеих таблицах.

Таблица 2. Примеры топиков и ключевых слов LDA моделей для выборки номеров из коллекции «Биржевых ведомостей» за 1913 год.

	Топик 1	**Топик 2**	**Топик 3**
Биржевые ведомости. 1913, №13521 (29 апр. (12 мая))3.	'0.021*"место" + 0.012*"завод" + 0.012*"съезд" + 0.009*"час" + 0.009*"станция" + 0.009*"октябрист" + 0.009*"академик" + 0.009*"совещание" + 0.009*"вновь" + 0.009*"вода"'	'0.023*"склад" + 0.021*"акция" + 0.016*"предприятие" + 0.012*"общество" + 0.012*"пароходных" + 0.012*"товарищество" + 0.012*"транспорт" + 0.012*"страх" + 0.012*"железный" + 0.007*"баланс"'	'0.015*"городской" 0.013*"**бумага**" 0.012*"**заем**" 0.011*"дорога" 0.007*"**капитал**" 0.007*"**облигация**" 0.007*"**общество**" 0.007*"**акция**" 0.006*"специальный" 0.006*"иметь"'
Биржевые ведомости. 1913, №13529 (3 (16) мая) 2.	'0.020*"**акция**" + 0.014*"**заем**" + 0.011*"рабочий" + 0.009*"май" + 0.007*"русский" + 0.007*"банк" + 0.007*"русско" + 0.007*"**рынок**" + 0.007*"предприятие" + 0.007*"новый"'	'0.015*"остров" + 0.013*"венгерский" + 0.008*"новый" + 0.008*"представитель" + 0.008*"округ" + 0.006*"акция" + 0.006*"май" + 0.006*"арестовать" + 0.006*"вопрос" + 0.005*"запрос"'	'0.010*"время" 0.007*"май" 0.007*"ленский" 0.007*"товарищество" 0.007*"рабочий" 0.007*"договор" 0.007*"держава" 0.006*"последний" 0.006*"подписать" 0.006*"принять"'
Биржевые ведомости. 1913, №13543 (11 (24) мая) 1.	'0.027*"коп" + 0.019*"опер" + 0.019*"лето" + 0.010*"мир" + 0.010*"новый" + 0.010*"торговый" + 0.010*"николаевский" + 0.010*"эстрада" + 0.010*"шт" +	'0.022*"год" + 0.017*"**вклад**" + 0.017*"вопрос" + 0.017*"**прирост**" + 0.012*"**министр**" + 0.012*"проект" + 0.012*"**бумага**" + 0.012*"правительство" + 0.012*"сотрудничать" + 0.012*"ответить"'	'0.022*"театр" 0.015*"касса" 0.015*"невский" 0.015*"май" + 0.015*"вход" + 0.015*"сад" 0.013*"проспект" 0.008*"выпуск" 0.008*"день" 0.008*"адрес"'

	0.010*"дебюты"'			
Биржевые ведомости. 1913, №13553 (17 (30) мая) 1.	'0.011*"день" + 0.011*"новый" + 0.011*"друг" + 0.011*"специально" + 0.011*"представление" + 0.011*"выпуск" + 0.011*"театр" + 0.011*"славянский" + 0.011*"вход" + 0.011*"сербский"'		'0.026*"**акция**" + 0.016*"собор" + 0.014*"час" + 0.010*"день" + 0.010*"новый" + 0.010*"город" + 0.008*"императорский + 0.008*"августейший" + 0.008*"местный" + 0.008*"лицо"'	'0.024*"коп" 0.019*"сегодня" 0.016*"май" 0.013*"акция" 0.013*"маг" 0.013*"невский" 0.013*"проспект" 0.013*"театр" 0.013*"начать" 0.013*"сад"'
Биржевые ведомости. 1913, №13565 (25 мая (7 июня)) 7.	'0.014*"личный" + 0.014*"мыло" + 0.014*"поэтому" + 0.008*"иметь" + 0.008*"убийство" + 0.008*"время" + 0.008*"преступление" + 0.008*"деньга" + 0.007*"станция" + 0.007*"хороший"'		'0.013*"год" + 0.012*"**акция**" + 0.012*"часть" + 0.009*"иметь" + 0.009*"**собрание**" + 0.009*"общий" + 0.009*"**акционер**" + 0.009*"склад" + 0.009*"час" + 0.009*"**правление**"'	'0.023*"книга" 0.016*"корова" 0.012*"местность" 0.012*"иметь" 0.012*"продажа" 0.012*"право" 0.008*"лес" 0.008*"невский" 0.008*"мочь" 0.008*"назначить"'
Биржевые ведомости. 1913, №13587 (8 (21) июня) 2.	'0.020*"пират" + 0.014*"заем" + 0.012*"акция" + 0.010*"дорога" + 0.010*"лодка" + 0.008*"нефтяной" + 0.008*"казна" + 0.008*"железный" + 0.006*"синдикат" + 0.006*"особенно"'		'0.011*"комитет" + 0.011*"неделя" + 0.011*"коп" + 0.006*"иметь" + 0.006*"боль" + 0.006*"принять" + 0.006*"заседание" + 0.006*"денежный" + 0.006*"поклон" + 0.006*"разрешать"'	'0.013*"торговля" 0.011*"министерство" 0.011*"**капитал**" 0.007*"**предприятие**" 0.007*"промышленнос 0.007*"правительство" 0.004*"страна" 0.001*"пират" 0.001*"**акционерный**" 0.001*"политика"'
Биржевые ведомости. 1913, №13597 (14 (27) июня) 2.	'0.016*"президент" + 0.013*"присутствовать" + 0.011*"июнь" + 0.011*"духовенство" + 0.011*"право" + 0.008*"согласие" + 0.008*"запрос" + 0.008*"мир" + 0.008*"выбор" + 0.006*"дело"'		'0.022*"**акция**" + 0.016*"русский" + 0.013*"**предприятие**" + 0.013*"правление" + 0.013*"банк" + 0.010*"**общество**" + 0.010*"**дивиденд**" + 0.010*"доклад" + 0.010*"зав" + 0.010*"русско"')	'0.018*"депутат" 0.014*"дорога" 0.007*"июнь" 0.007*"кредитор" 0.007*"роспись" 0.006*"дело" 0.006*"принять" 0.006*"также" 0.006*"член" 0.006*"вчера"')
Биржевые ведомости. 1913, №13621 (28 июня (11 июля)) 3.	'0.010*"пристав" + 0.009*"дело" + 0.009*"московский" + 0.008*"справка" + 0.007*"кислота" + 0.006*"фальсификация" + 0.006*"проц" +		'0.016*"**акция**" + 0.011*"рубль" + 0.007*"день" + 0.007*"дело" + 0.007*"чиновник" + 0.007*"артели" + 0.005*"пристав" + 0.005*"оказаться" +	'0.012*"июнь" 0.008*"год" 0.008*"общественный" 0.005*"городской" 0.005*"закон" + 0.005*"выпуск" 0.005*"пройти"

	0.006*"салициловый" + 0.005*"пластинка" + 0.005*"чиновник"')	0.005*"помощник" + 0.005*"служба"'	0.005*"управление" 0.005*"жизнь" 0.005*"издать"'
Биржевые ведомости. 1913, №13696 (12 (25) авг.) 1.	('0.012*"**акционер**" + 0.012*"день" + 0.012*"**помещение**" + 0.012*"**акция**" + 0.012*"**кризис**" + 0.012*"класс" + 0.007*"сведение" + 0.007*"третий" + 0.007*"**выпуск**" + 0.007*"подлинный"')	0.009*"дипломатический" 0.008*"август" + 0.008*"театр" + 0.008*"дело" + 0.008*"великий" + 0.008*"июль" + 0.008*"иметь" + 0.008*"переговоры" + 0.008*"вопрос" + 0.008*"близкий"')	'0.002*"дипломатическ 0.002*"переговоры" 0.002*"близкий" 0.002*"вопрос" 0.002*"посол" 0.002*"союз" 0.002*"круг" 0.002*"великий" 0.002*"балканский" 0.002*"сад"'

Таблица 3. Содержание отобранных примеров с указателем на приложение 2 (соответствующие сканы страниц «Биржевых ведомостей»).

Биржевые ведомости. 1913, № 13521 (29 апр. (12 мая)). 3 страница Приложение 2.1	Колонка «Может ли город обойтись без ссуды у банков» в разделе «Государственная Дума». На странице мы находим подробный доклад гласного петербургской городской думы по вопросу об организации очередного займа для покрытия муниципальных расходов Санкт-Петербурга. Автор высказывает сомнение в необходимости нового выпуска и считает, что продажа ценных бумаг, имеющихся на балансе городской думы, сможет обеспечить финансовые потребности столицы. К докладу прилагаются подробные статистические выкладки со ссылкой на «отчет с.-петербургского городского общественного управления за 1911 г.»
Биржевые ведомости. 1913, № 13529 (3 (16) мая). 2 страница Приложение 2.2	Колонка «Разоблачение Керенского», посвященная оправданию забастовки Ленских рабочих. В тексте члены правления ленского товарищества (а также неназванные акционеры) обвиняются в использовании административных связей для сокрытия тяжелых условий труда рабочих.
Биржевые ведомости. 1913, № 13543 (11 (24) мая). 1 страница Приложение 2.3	Колонка «Бюджетные прения» в разделе «Государственная Дума». Приводятся основные тезисы речи А.И. Шингарева о медлительности железнодорожного строительства. Далее следует оценка недавних слов министра финансов В.Н. Коковцова о текущем состоянии фондового рынка, удержавшегося на своих уровнях в контексте общего падения цен на мировых рынках.
Биржевые ведомости. 1913, № 13553 (17 (30) мая). 1 страница	Объявление правления Волжского акционерного общества маслобойных и химических заводов «Салолин» об открытии подписки на акции дополнительного выпуска с приведением условий участия.

Приложение 2.4	
Биржевые ведомости. 1913, № 13565 (25 мая (7 июня)). 7 страница Приложение 2.5	Объявление правления Акционерного общества С.-Петербургских товарных складов о проведении чрезвычайного общего собрания акционеров с повесткой мероприятия.
Биржевые ведомости. 1913, № 13587 (8 (21) июня). 2 страница Приложение 2.6	Колонка «Речь Коновалова» в разделе «Государственная Дума». А.И. Коновалов поднял проблему несовершенства акционерного законодательства. Депутат говорил о длительности процедуры разрешения акционерных предприятий и высказывался против ограничений, устанавливаемых при разрешении акционерных обществ для поляков и евреев.
Биржевые ведомости. 1913, № 13597 (14 (27) июня). 2 страница Приложение 2.7	Обращение к владельцам облигаций Сестрорецкой железной дороги председателя конкурсного управления, юрисконсультанта министерства финансов К.К. Дыновского: сообщение о ликвидации текущего предприятия и организации продажи дороги в собственность другому обществу, готовому взять на себя обязательства по реорганизации дороги. Приводится сообщение представителя группы кредиторов дороги Л.Л. Балинского о том, что купоны не оплачивались дорогой с конца 1906 г.
Биржевые ведомости. 1913, № 13621 (28 июня (11 июля)). 3 страница Приложение 2.8	Колонка «Счастливица». Рассказ о том, как некая женщина нашла в магазине сверток с акциями, которые потерял банкир Борисов. Этот банкир находился в доме предварительного заключения по подозрению в растрате этих акций. Редакция высказывает подозрение, поскольку «утерянные» в июне акции были внезапно найдены в магазине в сентябре.
Биржевые ведомости. 1913, № 13696 (12 (25) авг.) 1 страница Приложение 2.9	Объявление от правления акционерного общества «Северный ломбард» об обмене временных свидетельств на подлинные акции.

***

Анализ материалов, найденных с помощью LDA моделей, был включен в основное содержание нашего исследования, посвященного изучению поведенческих практик, связанных с анализом доходности ценных бумаг. Таким образом, был реализован подход, при котором содержательные задачи исследования решались традиционными историческими методами, предполагающими детальное изучение текста источника и встраивание отдельного документа в единую систему с другим имеющимся материалом – архивными источниками и сочинениями биржевых практиков [7].

На данном этапе мы можем подтвердить, что в рамках нашего исследования применение тематического моделирования оказалось продуктивным решением для оптимизации процесса поиска исторических документов в объемной коллекции оцифрованных исторических материалов. В то же время необходимо подчеркнуть, что в нашей работе тематическое моделирование применялось исключительно как прикладной инструмент ускорения поиска и первичной оценки информационного потенциала коллекции документов через анализ выделенных топиков. Наш опыт показал, что по крайней мере для «Биржевых ведомостей» тематическое моделирование с использованием LDA из библиотеки Gensim не позволяет делать выводы с позиции применяемой нами методологии содержательного анализа, предполагающей работу с внутренним содержанием источника на уровне аналитических практик. Данные наших моделей слишком фрагментарны, их можно использовать только для первичной оценки тематик информации, содержащейся в источнике. Безусловно, мы должны учитывать возможность дальнейшего усложнения моделей через применение аддитивной регуляризации [8] и совершенствование OCR-распознавания в сторону разбиения страницы на документы по каждой отдельной колонке, – что может существенно повысить способность модели улавливать более глубокую семантику газетного текста. Подводя итоги, мы бы хотели отметить прикладную значимость продолжения исследований практической применимости тематического моделирования в решении задач источниковедения.

Для интересующихся читателей мы оставляем ссылку на доступ к предложенному автором программному коду и набору исходных данных [9].

Приложение 1: Словарь определений

Тематическое моделирование – это технология статистического анализа текстов для автоматического выявления тематики в больших коллекциях документов. Тематическая модель определяет, к каким темам относится каждый документ, и какими словами описывается каждая тема. Для этого не требуется ручная разметка текстов, обучение модели происходит без учителя. Этот процесс можно сравнить с кластеризацией, но тематическая кластеризация является «мягкой» и допускает, чтобы документ относился к нескольким кластерам-темам. Тематическое моделирование не претендует на понимание смысла текста, однако оно способно отвечать на вопросы «о чём этот текст» или «какие общие темы есть у этих текстов» [10].

Применяя тематическое моделирование, важно проводить разграничение между топиками и, собственно, темами. Топики являются результатом статистической обработки коллекции документов и состоят из слов, которым в зависимости от выбранной статистической модели (LDA, LSA, BertTopic и т.д.) была присвоена определенная значимость, на основе которой моделью было сделано предположение, что эти топики (кластер значимых слов) формируют тему документа на семантическом уровне. Однако важно подчеркнуть, что процедура признания, что определенный топик действительно отражает присутствующую в тексте документа тему, является сугубо экспертной. Исследователь определяет, насколько семантически релевантным текстовой коллекции оказался сформированный моделью набор топиков.

TF-IDF (*от англ.* TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно

пропорционален частоте употребления слова во всех документах коллекции.

ПРИЛОЖЕНИЕ 2

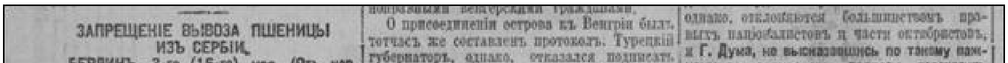
Страницы газеты «Биржевые ведомости», выявленные методом автоматизированного поиска по содержанию комплекта LDA моделей


№ 13521. 29-го апрѣля 1913 г. БИРЖЕВЫЯ ВѢДОМОСТИ.

Можетъ-ли городъ обойтись безъ ссуды у банковъ?

О запрещеніи славянскаго банкета.

Рисунок 2. 1. Биржевые ведомости. 1913, № 13521 (29 апр. (12 мая)). Третья страница


ЗАПРЕЩЕНІЕ ВЫВОЗА ПШЕНИЦЫ ИЗЪ СЕРБІИ.

ОТЪѢЗДЪ СЕРБСКИХЪ ДЕЛЕГАТОВЪ ВЪ ЛОНДОНЪ.

ОГОВОРКИ БОЛГАРІИ ОТНОСИТЕЛЬНО ПОДПИСАНІЯ ПРЕДВАРИТЕЛЬНАГО МИРА.

ПОЧЕМУ БОЛГАРІЯ ПОСПѢШИЛА СЪ ПРИНЯТІЕМЪ ПРЕДВАРИТЕЛЬНЫХЪ УСЛОВІЙ МИРА?

Русско-французскія желѣзнодорожныя концессіи въ Турціи.

Шпіономанія въ Швеціи.

РАЗОБЛАЧЕНІЯ КЕРЕНСКАГО.

Рисунок 2. 2. Биржевые ведомости. 1913, № 13529 (3 (16) мая). Вторая страница

ГОСУДАРСТВЕННАЯ ДУМА.

Первая сессія. — 44-е засѣданіе

Суббота, 11-го мая.

Бюджетныя пренія.

Продолжаются бюджетныя пренія. Вчера, кромѣ предсѣдателя бюджетной комиссіи М. М. Алексѣенко и министра финансовъ В. Н. Коковцова, выступилъ ШИНГАРЕВЪ, который лишь успѣлъ начать свою рѣчь.

РѢЧЬ ШИНГАРЕВА.

ШИНГАРЕВЪ продолжаетъ говорить о дефектахъ новой русской землеустроительной комиссіи, утверждаетъ, что культурныя мѣропріятія проводятся у насъ съ помощью кнута, высылокъ и арестовъ и доказываетъ, что въ результатѣ нашего землеустроительства оказывается громадная масса бездомнаго пролетаріата, того ходячаго бобыля, который никуда не можетъ пристроиться.

Касаясь попутно и дѣятельности крестьянскаго банка, ораторъ переходитъ къ нашей внѣшней торговлѣ:

— Россія, — говоритъ онъ, — единственная страна, въ которой бюджетъ превышаетъ торговый оборотъ, что влечетъ за собою непропорціональность нашего бюджета по его экономическому базису.

Англійская торговля осталась на одномъ уровнѣ. Германія увеличила свой ввозъ въ Персію въ четыре раза, Бельгія въ 8 разъ, а нашъ русскій вывозъ за послѣднее десятилѣтіе ничтоженъ, и мы еще строимъ добрососѣдской Германіи дорогу для того, чтобы ввозить нѣмецкіе товары въ сѣверную Персію.

Мы могли бы удвоить нашъ активный торговый балансъ, если бы могли, какъ говорилъ министръ финансовъ, производить все сами, а не привозить изъ-за границы. Но почему мы не производимъ все то, что привозимъ изъ-за границы?

Россія производитъ ленъ на весь міръ, чуть-ли не въ 8 разъ больше, чѣмъ всѣ остальныя державы міра. А что же, мы обрабатываемъ нашъ ленъ? Ни Боже мой. Мы его вывозимъ въ сырьѣ.

Мы вывозимъ больше, чѣмъ ввозимъ шелка, шерсти, сала, кожъ и другихъ продуктовъ.

ШИНГАРЕВЪ говоритъ о медленности нашего желѣзнодорожнаго строительства, о плохомъ состояніи дорогъ и переходитъ къ той части рѣчи министра, которая касалась недавняго паденія бумагъ на міровыхъ рынкахъ:

— Министръ съ нѣкоторой гордостью заявилъ, что наши бумаги не очень упали. Позволю себѣ напомнить, что до того, что было до войны, они не поднялись.

Министръ отмѣчаетъ развитіе вкладовъ и акціонерныхъ капиталовъ. Но мы знаете, напримѣръ, что въ англійскихъ банкахъ — два процента казенныхъ денегъ, а въ русскихъ — двадцать процентовъ. У насъ — огромное количество казенныхъ денегъ фигурируетъ въ банкахъ. Въ то время, какъ у насъ количество вкладовъ въ частныхъ коммерческихъ банкахъ выросло за послѣдніе 3 года на 800 милл. руб., въ Америкѣ прирость вкладовъ въ годъ равняется 560 милл.

Министръ гордится приростомъ вкладовъ, но въ то время, какъ въ Америкѣ прирость вкладовъ доходитъ до 336 милл. въ годъ, у насъ — до 85 милл. руб.

Министръ говорилъ о прогрессѣ. Да, мы движемся впередъ, но съ необычайной медленностью, съ поразительнымъ убожествомъ, съ остановками, съ болѣзненными процессами. Чѣмъ обидѣла насъ исторія? Чѣмъ не надѣлила насъ судьба, что насъ обгоняютъ державы съ меньшимъ составомъ населенія, съ убогими, природными богатствами?

Отвѣтъ ясенъ — міръ внѣшній, о немъ проявили много заботы, даже, какъ будто, слишкомъ много. А міръ внутренній — забота о хозяйствѣ, наукѣ и техникѣ, о внутреннемъ мірѣ, о которомъ мечтаетъ вся громадная страна, — что сдѣлало наше правительство?

Экономика и политика неотдѣлимы, и на-

Рисунок 2. 3. Биржевые ведомости. 1913, № 13543 (11 (24) мая). Первая страница

НА СЕМЕНОВСКОМЪ ПЛАЦУ

БѢГА

Завтра 18 Мая Начало въ 2 часа дня.

С.-ПЕТЕРБУРГСКІЙ ТОРГОВЫЙ БАНКЪ

Невскій пр., № 7—9, домъ Вавельбергъ

ОТДАЮТСЯ ВЪ НАЕМЪ

БЕЗОПАСНЫЕ ЯЩИКИ (САРЕS)

отъ 3 рублей

НЕСГОРАЕМАЯ КЛАДОВАЯ

НА ХРАНЕНІЕ СЪ УПРАВЛЕНІЕМЪ.

ЛЛОЙДЪ

АВТОМОБИЛИ и ГРУЗОВИКИ

Невскій, 97. В. ЖУКОВСКІЙ, Телеф. 21-55.

Согласно положенію Совѣта Министровъ ВЫСОЧАЙШЕ утвержденному 15-го марта 1913 года,

„Правленіе Волжскаго Акціонернаго Общества маслобойныхъ и химическихъ заводовъ „САЛОЛИНЪ“

Т-во Ресторанъ „МЕТРОПОЛЬ“

Состоящая подъ ВЫСОЧАЙШИМЪ ЕГО ИМПЕРАТОРСКАГО ВЕЛИЧЕСТВА покровительствомъ

IV МЕЖДУНАРОДНАЯ АВТОМОБИЛЬНАЯ ВЫСТАВКА

въ Михайловскомъ манежѣ

Закрытіе 19-го Мая.

Плата за входъ 1 руб., для учащихся въ формѣ 50 коп., для учащихся группами при руководителѣ отъ 11 час. утра до 1 часу дня — 25 коп.

Выставка открыта ежедневно отъ 11 ч. у. до 11 ч. в.

Отличившіеся на всѣхъ международныхъ и военныхъ испытаніяхъ

ГРУЗОВОЗЫ БЮССИНГЪ

На осеннемъ пробѣгѣ Военнаго Вѣдомства удостоены отзыва „ОТЛИЧНЫЕ“.

Русско-Америк. Т-во тепловыхъ двигателей.

Ст. Петербургъ, Думская, 7, телефонъ 561-27.

НА ВЕРСТОВОЙ ГОНКѢ 14-го мая

Рисунок 2. 4. Биржевые ведомости. 1913, № 13553 (17 (30) мая). Вторая страница

Наилучшее, единственно благоустроенное здоровое мѣсто для жилья

ЗИМОЙ и ЛѢТОМЪ

Лѣсъ, озеро, абиссинскіе колодцы.

Разсрочка

ежемѣсячный платежъ на 3 года

МУРТЫШКИНО

допродаются остальные

ДАЧНЫЕ УЧАСТКИ

ГИМНАЗІЯ и РЕАЛЬНОЕ УЧИЛИЩЕ

д-ра Н. П. ШЕПОВАЛЬНИКОВА,

Каменноостровскій пр., 24 (прот. Лицея). Тел. 119-99.

Со всѣми правами (для учителей и учениковъ) гимназій и реальныхъ училищъ Мин. Нар. Пр.

Директоръ д-ръ мед. Шеповальниковъ.

ПРАВЛЕНІЕ

Акціонернаго Общества С.-Петербургскихъ товарныхъ складовъ

имѣетъ честь пригласить г.г. акціонеровъ Общества на чрезвычайное общее собраніе, имѣющее состояться въ С.-Петербургѣ, въ помѣщеніи Калашниковскаго Отдѣленія С.-Петербургскаго Международнаго Коммерческаго Банка (Полтавская, 12), 18-го Іюня сего года, въ 8 час. вечера.

Предметы занятій:

1) Обсужденіе и рѣшеніе вопроса объ увеличеніи основного капитала Общества, путемъ дополнительнаго выпуска акцій.

2) Обсужденіе и рѣшеніе вопроса о пріобрѣтеніи новаго недвижимаго имущества и способѣ его оплаты.

Г.г. акціонеры, желающіе участвовать въ Общемъ Собраніи, благоволятъ предъявить не позже 4 час. дня, 11-го Іюня 1913 года, въ Правленіе Общества, (Полтавская, 12), принадлежащія имъ акціи или временныя свидѣтельства, или же удостовѣренія въ принятіи оныхъ на храненіе или залогъ какъ государственныхъ, такъ и дѣйствующихъ на основаніи Правительствомъ утвержденныхъ уставовъ частныхъ кредитныхъ учрежденій, съ обозначеніемъ въ удостовѣреніяхъ нумеровъ акцій и временныхъ свидѣтельствъ.

ТОРГИ

Измайловскій пр. 16.

СТАНЦІЯ „ГОРСКАЯ“

ПРОДАЖА ДАЧНЫХЪ УЧАСТКОВЪ.

Безпроцентная разсрочка на пять лѣтъ.

Планы высылаются по первому требованію.

КОМНАТЫ отъ 70 рубл. НА СЕЗОНЪ

сдаются въ ЕРМОЛОВСКОМЪ ПАНСІОНАТѢ при ст. Ермоловской.

Прекрасн. сухая мѣстность, больш. сосновый садъ.

Справки: Ст. Ермоловская. Телефонъ № 93.

Рисунок 2. 5. Биржевые ведомости. 1913, № 13565 (25 мая (7 июня)). Седьмая страница

2 БИРЖЕВЫЯ ВѢДОМОСТИ. № 13587. 8-го іюня 1913 г.

ДЛЯ ДАЧИ и ДЕРЕВНИ ПИДЖАКИ

Пахнетъ Сентябремъ.

Борьба съ рѣчными пиратами.

Рисунок 2. 6. Биржевые ведомости. 1913, № 13587 (8 (21) іюня). Вторая страница

душіе взглядовъ по всѣмъ пунктамъ существующаго между обѣими державами соглашенія.

Лондонъ, 13-го (26-го) іюня. (СПА). По свѣдѣніямъ агентства Рейтера, **Пишонъ и Камбонъ имѣли сегодня продолжительное совѣщаніе съ Греемъ и Никольсономъ** въ министерствѣ иностранныхъ дѣлъ.

Позже Грей и Никольсонъ совѣщались свыше часа съ президентомъ Пуанкаре въ Сентъ-Джемскомъ дворцѣ; во время совѣщанія присутствовали также Пишонъ и Камбонъ.

Лондонъ, 13-го (26-го) іюня. (СПА). Въ бесѣдѣ съ представителемъ агентства Рейтера Пишонъ сказалъ, что президентъ **Пуанкаре вынесъ прекрасное впечатлѣніе отъ поѣздки и былъ глубоко тронутъ оказаннымъ ему пріемомъ**, такъ какъ пріѣздъ президента оказалъ сильное вліяніе на укрѣпленіе согласія между Франціей и Англіей. Выяснилось, что услуги, оказываемыя всему міру этимъ согласіемъ, заключаются въ поддержаніи международнаго мира въ интересахъ всей Европы.

Обмѣнъ мнѣній съ сэромъ Эдуардомъ Греемъ не только подтвердилъ прежнее впечатлѣніе, но и выяснилъ также полное согласіе обоихъ кабинетовъ по обще-политическимъ вопросамъ, а равно и по всѣмъ вопросамъ, касающимся сохраненія мира. Пріѣздъ президента явился, такимъ образомъ, элементомъ, способствующимъ общему умиротворенію.

ВЫБОРНЫЕ ЗАПРОСЫ.

Покончивъ съ бюджетомъ, Г. Дума спѣшитъ прежде всего ликвидировать выборные запросы. Уже сессія на исходѣ и интересъ къ выборнымъ запросамъ значительно выдохся.

Но разсмотрѣть ихъ нужно и Г. Дума вяло слушаетъ ораторовъ.

Говоритъ с.-д. СКОБЕЛЕВЪ, который еще разъ доказываетъ, что незакономѣрности были, что с.-д-овъ всячески преслѣдовали и заканчиваетъ:

— **Наша избирательная система, это — дѣтище государственнаго переворота 3-го іюня, является шедевромъ того, до чего додумалась когда-нибудь реакціонная мысль въ Европѣ.** Выборы въ 4-ю Думу съ очевидностью доказали, что установленіе всеобщаго избирательнаго права является завѣтомъ не только демократіи, но и элементарной политической честности.

Націоналистъ свящ. ДРОЗДОВСКІЙ чуть-ли не въ первый разъ за всю сессію поднимается на трибуну, дабы протестовать противъ нападокъ на духовенство.

— Духовенство присутствуетъ здѣсь по волѣ Самодержца Всероссійскаго, — говоритъ онъ, — **а разъ духовенство имѣетъ право участвовать въ выборахъ, то оно должно проявить свое право.**

— Вы предвидите, что документы, на которыхъ построены запросы, неосновательны и вамъ приходится, господа, отъ стыда опустить глаза. **Довольно брани на православное духовенство. Довольно унижать достоинство Г. Думы и свое собственное.** Духовенство знаетъ цѣль вашихъ выпадовъ. Оно не смутится и будетъ парировать ваши удары.

Благодарное своему Государю за дарованное право участія въ выборахъ, оно исполнило свой долгъ честно, добросовѣстно, смѣло и прямо смотритъ въ глаза, не чувствуя за собой никакой вины.

Справа раздаются аплодисменты, а слѣва возгласъ:

— Ни стыда, ни совѣсти.

* * *

Выступаютъ священники КАРПИНСКІЙ и НАСЕЛЕНКО, которые также защищаютъ духовенство.

Говоритъ еще ВОРОНКОВЪ.

...вплетъ:

— Я былъ предусмотрителенъ и, когда меня спросили въ канцеляріи Г. Думы, къ какой партіи я принадлежу, я шутя отвѣтилъ: Буду принадлежать къ фракціи к.-д...

Это говоритъ бывшій кандидатъ въ московскія городскія головы, членъ 3-й Гос. Думы.

Возрожденіе сестрорѣцкой жел. дор.

Вновь избранный предсѣдателемъ, учрежденнаго по дѣламъ приморской с.-петербургско-сестрорѣцкой жел. дор. конкурснаго управленія, юрисконсультъ министерства финансовъ д. с. с. **Н. Н. Дьяковскій** о судьбѣ дороги говоритъ:

— Мы постараемся принять самыя рѣшительныя мѣры, что въ самомъ непродолжительномъ будущемъ окончательно ликвидировать сестрорѣцкую жел. дор. и дать возможность пріобрѣсти ее на наиболѣе выгодныхъ для кредиторовъ условіяхъ одной изъ многочисленныхъ компаній, которыя заявили конкурсу о своемъ желаніи купить дорогу.

При этомъ, новымъ владѣльцамъ дороги, кромѣ финансовыхъ, будутъ поставлены и условія о совершенной реорганизаціи дороги.

Конечно, облигаціонеры должны быть удовлетворены, но дорогу передадимъ мы только такому солидному обществу, которое согласится ввести электрическую тягу, переустроить курортъ на подобіе лучшихъ заграничныхъ лечебныхъ мѣстъ, построить гостиницы и вообще способствовать превращенію всей обслуживаемой сестрорѣцкой дорогой мѣстности въ благоустроенный уголокъ, который бы петербуржцы полюбили также, какъ они теперь любятъ Финляндію.

Изъ числа многочисленныхъ предложеній иностранныхъ компаній есть нѣсколько, предусматривающихъ всѣ эти улучшенія, такъ что выборъ можно будетъ произвести безъ затрудненія.

Будутъ также приняты мѣры, чтобы наши законодательныя учрежденія въ возможно скоромъ времени утвердили правила для продажи дороги.

Чрезвычайное общее собраніе кредиторовъ въ этомъ смыслѣ уже отправило на имя

Рисунок 2. 7. Биржевые ведомости. 1913, № 13597 (14 (27) июня). Вторая страница

чать на коробкахъ время изготовленія консервовъ. Тогда бы публика была болѣе гарантирована, что ей не отпускаются коробки, простоявшія въ магазинѣ три года и больше.

Старые консервы особенно вредны, такъ какъ если они содержатъ кислоту, то она можетъ разъѣдать стѣнки коробокъ и въ консервахъ получится примѣсь металлическихъ солей.

Кондитерскія издѣлія и прохладительные напитки.

Для предупрежденія прокисанія такихъ продуктовъ, какъ пастила, нерѣдко употребляется салициловая кислота, для подкрашиванія конфектъ и прохладительныхъ напитковъ употребляются каменноугольныя краски. Сахаръ замѣняется сахариномъ.

Борьба съ фальсификаціей.

— Нашъ законъ,—говоритъ въ заключеніе проф. Прожбытекъ,—не знаетъ даже термина фальсификація, и торговцы, уличенные въ фальсификаціи, подлежатъ отвѣтственности или за обманъ, или за причиненіе вреда здоровью покупателя. Послѣднее часто доказать трудно.

И въ одномъ, и въ другомъ случаѣ на фальсификаторовъ налагаются гораздо болѣе слабыя наказанія, чѣмъ это предусмотрѣно иностранными кодексами. Проектъ новаго уголовнаго уложенія устанавливаетъ болѣе точно понятіе фальсификаціи и болѣе суровыя наказанія за нее.

Обязанность слѣдить за доброкачественностью продуктовъ, поступающихъ въ продажу, лежитъ на полиціи, санитарныхъ врачахъ и санитарахъ, торговыхъ врачахъ, обязанныхъ слѣдить за мѣстами продажи и приготовленія съѣстныхъ продуктовъ.

II.

Счастливица.

Іоганнъ Казариной, женщинѣ болѣе чѣмъ средняго достатка, посчастливилось: 26-го сентября 1912 г., входя въ какой-то магазинъ, по Невскому просп., она нашла свертокъ, въ которомъ оказалось десять акцій ростово-владикавказской жел. дороги, стоящихъ свыше 33 тыс. рублей.

Въ тотъ же день Казарина доставила находку въ управленіе сыскной полиціи, гдѣ оказалось, что заподозрѣнный въ растратѣ этихъ самыхъ акцій артельщикъ биржевой барона Штиглица артели, нѣкій Борисовъ, которому артель не повѣрила, когда онъ пропажу акцій объяснялъ утерей ихъ, заключенъ въ домъ предварительнаго заключенія.

Страннымъ показалось, что Борисовъ утерялъ акціи въ іюнѣ, а Казарина нашла ихъ въ сентябрѣ. Какъ это произошло—слѣдствіе не выяснило, но зато выяснило, что Казарина съ Борисовымъ никогда не были знакомы.

За отсутствіемъ уликъ дѣло противъ Борисова было прекращено, а акціи постановлено выдать артели по принадлежности.

Законное желаніе Казариной получить съ артели добровольное вознагражденіе встрѣтило возраженіе, и она до фактической выдачи артели акцій предъявила искъ въ размѣрѣ одной трети стоимости находки, и искъ этотъ судомъ уже обезпеченъ путемъ наложенія ареста на подлежащія выдачѣ акціи на сумму 11 тыс. рублей. Интересы Казариной защищаетъ прис. пов. О. С. Трахтеренъ.

Курьезъ.

Курьезную исторію, происшедшую на закончившемся недавно харьковскомъ съѣздѣ по народному образованію, разсказываетъ «Утро Россіи». Эта исторія напоминаетъ гоголевскаго «Ревизора». Московскій коммерческій институтъ командировалъ на съѣздъ одного изъ своихъ слушателей, войдя въ частное соглашеніе съ городской управой относительно провѣзда его по желѣзнымъ дорогамъ. На съѣздѣ студента поэтому приняли за официальнаго представителя московскаго городского самоуправленія, оказывали ему соотвѣтствующій почетъ, и рѣшили избрать предсѣдателемъ съѣзда, чтобы въ его лицѣ почтить представляемую имъ столицу. Въ концѣ-концовъ, недоразумѣніе выяснилось, и слушатель института предсѣдателемъ изобрать не былъ.

Кто переусердствовалъ?

Рисунок 2. 8. Биржевые ведомости. 1913, № 13621 (28 июня (11 июля)). Третья страница

„НОЧЬ ЛЮБВИ".
Нач. въ 8½ час. веч.
Завтра новая оперетта въ 3 д.
„Король Магазина".

Зрительный Залъ и Народная Аудиторія Имени Его Высочества Принца А. П. Ольденбургскаго при НАРОДНОМЪ ДОМѢ ИМПЕРАТОРА НИКОЛАЯ II.
сегодня «Царь-плотникъ», оп. «Лебединое озеро.» «Тяжелые дни», «Преступленіе и наказаніе», «ЖИВОЙ ТРУПЪ».

САДЪ и ТЕАТРЪ „Эденъ" (Газовая, 23, тел. 36-10).
К. К. Баумвальдтъ.

АРКАДІЯ
Новая деревня

Автомобили
Делагэ
ПАРИЖЪ. Спб., Итальянская, 29.

ПРАВЛЕНІЕ АКЦІОНЕРНАГО ОБЩЕСТВА
„СѢВЕРНЫЙ ЛОМБАРДЪ"
(Владимірскій пр., 19).

ПЕРЕѢЗЖАЮЩИМЪ
И ВНОВЬ УСТРАИВАЮЩИМСЯ
предлагаетъ
Городской Установочный Отдѣлъ

ПО РОССІИ: Заба
Выставка
ФЕЛЬЕТ

ДЕШЕВО и ЭЛЕГАНТНО
М. ЗАКГЕЙМЪ
ВЕРНУЛСЯ,
Д-ръ И. ФЕЙЕРТАГЪ,

Рисунок 2. 9. Биржевые ведомости. 1913, № 13696 (12 (25) авг.). 1 страница

Библиография

1. URL: http://docs.historyrussia.org/ru/nodes/1-glavnaya
2. Tze-I Yang, A.J.Torget, R.Mihalcea (2011). Topic modeling in historical newspapers.
3. Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020). Topic Modelling Discourse Dynamics in Historical Newspapers. DHN Post-Proceedings.
4. Koentges, Thomas (2020). Measuring Philosophy in the First Thousand Years of Greek Literature.
5. Egger, Roman (2020). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts.
6. Галушко И.Н. Корректировка результатов OCR-распознавания текста исторического источника с помощью нечетких множеств (на примере газеты начала XX века) // Историческая информатика. – 2023. – № 1. – С. 102-113.
7. Представленная статья является частью моей магистерской диссертации по теме:

«Поведенческие аспекты анализа доходности ценных бумаг на фондовом рынке Российской империи в начале XX века: контент-анализ биржевых нарративов». Найденные LDA-алгоритмом выпуски «Биржевых ведомостей» в данной работе рассматривались в сочетании с материалами фонда №143 ЦГАМ (Московский биржевой комитет) и трудами биржевых практиков начала XX в. (Васильев А.А. Биржевая спекуляция, теория и практика. СПб., 1912.).

8. Воронцов К. В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2020.
9. GitHub. URL: https://github.com/iodinesky/Topic-modeling-in-historical-newspapers
10. Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2023.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Рецензируемая статья посвящена анализу информационного потенциала коллекции исторических источников с помощью тематического моделирования. В данном случае тематическое моделирование, пригодное для различных направлений анализа текстов, используется как инструмент предварительной оценки содержания коллекции исторических документов для отбора текстов, релевантных исследовательским запросам. В качестве массива текстов, на котором проводится апробация предлагаемой методики, используется пресса (газета «Биржевые ведомости» за 1905 и 1913 гг.).
Для исследования были взяты оцифрованные комплекты газет с сайта Российской национальной библиотеки (447 номеров). Автором была поставлена задача использования тематического моделирования для автоматического поиска тех газетных страниц, где есть информация о функционировании рынка ценных бумаг. В процессе поиска использовался хорошо апробированный в компьютерной лингвистике метод лемматизации. Использовались вероятностные LDA-модели, часть из которых была автоматически отобрана для дальнейшего анализа как содержащая биржевую информацию. Далее проводился содержательный анализ отобранного материала.
Актуальность работы не вызывает сомнений, поскольку поиск и отбор необходимой для дальнейшего исследования информации является на сегодняшний день серьезной проблемой, на решение которой уходит огромное количество времени. Любой способ адекватного решения подобных вопросов – это большой вклад в научно-исследовательскую практику.
Научная новизна работа также очевидна. Тематическое моделирование почти не освоено в исторической науке, появляются только первые опыты его использования, а вопрос о степени его полезности в рамках творческой лаборатории профессионального историка до сих пор остается открытым.
Статья построена не вполне традиционным образом, поскольку носит во многом экспериментальный характер. После постановки проблемы практически сразу начинается логический раздел, посвященный созданию методики использования тематического моделирования для поиска информации. Далее анализируются результаты поиска. Показательно, что, с одной стороны, автор констатирует продуктивность созданной методики для оптимизации поиска, с другой, – подчеркивает, что полученные модели слишком фрагментарны и их можно использовать только для первичной оценки тематики информации источника. Статья дополнена примерами

топиков и ключевых слов, фрагментами отобранных текстов, словарем определений и фотографиями газетных страниц. Следует отметить, что безусловно интересная и новаторская статья рассчитана на подготовленного читателя, знакомого с основами анализа текстов. В то же время продуманный стиль статьи облегчает понимание довольно сложных и не всегда привычных для традиционного взгляда историка вещей, о которых идет речь.

Библиография статьи содержит достаточный для подобных исследований список, хотя, думается, что не лишним было бы добавить работ на русском языке.

Статья фактически является приглашением к обмену мнениями и дискуссиям по рассматриваемой проблематике, скорее всего, ей обеспечена хорошая цитируемость в силу актуальности рассмотренных вопросов.

Публикация исследования, связанного с методами машинного обучения, рассчитана на определенный круг читателей, который, безусловно, будет достаточно широким, поскольку любая работа, связанная с интеллектуальной обработкой данных, вызывает сегодня большой интерес научной общественности. Статья рекомендуется к публикации.