Секция «Историческая информатика»

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

Научный руководитель – Бородкин Леонид Иосифович

Галушко Илья Николаевич
Аспирант
Московский государственный университет имени М.В.Ломоносова, Исторический факультет, Москва, Россия
E-mail: galushko.iolk@yandex.ru

Ключевой задачей представленного исследования 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 [3] 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. Основная проблема заключается в том, что содержание «Биржевых ведомостей» довольно пространно. 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. И в этом контексте мы решили попробовать применить тематическое моделирование в качестве прикладного инструмента для автоматического поиска тех номеров (страниц) из нашей коллекции, которые содержат информацию, касающуюся особенностей функционирования рынка ценных бумаг.

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. Как правило, топик представляет собой взвешенный по вероятности список слов, которые вместе выражают общее содержание предполагаемой темы [3]. Чем выше коэффициент слова, тем большее значение модель придает этому слову при формировании топика. 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» [2]. 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.

Для всей отобранной коллекции текстов «Биржевых ведомостей» было создано 2411 LDA моделей (мы использовали Python-библиотеку Gensim). 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.

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.

Источники и литература

1) URL: http://docs.historyrussia.org/ru/nodes/1-glavnaya
2) Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020). Topic Modelling Discourse Dynamics in Historical Newspapers. DHN Post-Proceedings.
3) Tze-I Yang, A.J.Torget, R.Mihalcea. Topic modeling in historical newspapers. 2011
4) Thomas Koentges. Measuring Philosophy in the First Thousand Years of Greek Literature. 2020