процесс работы, ведь в рамках одной компьютерной программы собраны инструменты для анализа и визуализации текстов.

Таким образом, визуализация является мощным инструментом, который обогащает традиционные методы анализа письменных исторических источников. Она позволяет исследователям не только экономить время при работе с большими массивами данных, но и открывает новые перспективы анализа и интерпретации исторических источников.

1. Гарскова, И. М. Новые тенденции в компьютеризованном анализе текстов: концепции, методы, технологии / И. М. Гарскова // История : электронный научно-образовательный журнал. – 2015. – Т. 6. – Выпуск 8(41). – URL: http://history.jes.su/s207987840001255-9-1 (дата обращения: 18.02.2025).

2. Латур, Б. Визуализация и познание: изображая вещи вместе / Б. Латур // Логос: философско-литературный журнал. – 2017. – № 27(117). – С. 95–156.

3. Knorr, K. be Manufacture of Knowledge / K. Knorr. – Oxford : Pergamon Press, 1981. – 189 p.

Косачев Д.П.

АВТОМАТИЗАЦИЯ СБОРА И АНАЛИЗА МЕТАДАННЫХ СТАТЕЙ ИЗ ЦИФРОВОГО АРХИВА «НЬЮ-ЙОРК ТАЙМС» С ИСПОЛЬЗОВАНИЕМ PYTHON

Ключевые слова: историческая информатика, Нью-Йорк Таймс, контент-анализ, Великая Отечественная война, образ СССР.

Исследование роли медиа в формировании общественного мнения и их влияния на ключевые политические процессы в период Второй мировой войны представляет собой важный аспект исторического анализа. В условиях глобального конфликта средства массовой информации, такие как газеты, становились не только источниками информации, но и инструментами пропаганды, консолидации общества. Американские ученые отмечают, что «исследования неизменно выявляют более существенные связи между воздействием медиа и общественным мнением в вопросах внешней политики, чем в вопросах внутренней политики. Эта тенденция частично объясняется тем, что граждане имеют прямой личный опыт во многих сферах внутренней политики, но вынуждены в большей степени полагаться на информацию из массмедиа для оценки и анализа вопросов внешней политики» [6, p. 101].

Цель данной работы заключается в сборе и анализе метаданных статей «Нью-Йорк Таймс» за период 1941–1945 гг., упоминающих СССР, с использованием современных методов автоматизированного сбора и обработки данных. Основное внимание уделяется методике получения базы данных, которая в дальнейшем может быть использована как основа для более детализированного контент-анализа содержания публикаций.

Для достижения поставленной цели был разработан скрипт на языке программирования Python, позволяющий автоматизировать сбор метаданных электронного архива «Нью-Йорк Таймс» через API (Application programming interface). Это позволило создать структурированную базу данных, включающую такие поля, как заголовки статьи, даты публикаций, авторы и ключевые теги, присвоенные сотрудниками газеты.

«Нью-Йорк Таймс» является одним из ведущих мировых печатных изданий на протяжении большей части XX в. В период Второй мировой войны «Нью-Йорк Таймс» обладала значительным влиянием, признаваемым членами правительства за её роль в формировании общественного мнения, особенно в вопросах внешней политики. Судья Верховного суда США в то время Феликс Франкфуртер даже рассматривал «Таймс»

как «квазисудебный орган», подразумевая, что её функции были ближе к функциям судов, чем к другим ветвям власти, подчеркивая огромную ответственность издания перед обществом [7, p. 11].

«Нью-Йорк Таймс» являлось одним из самых массовых печатных изданий в рассматриваемый период. В книге Гэя Тализа приводятся сведения для 1937 г. Он заявляет, что в тот год тираж ежедневных выпусков газеты превысил 500 000 экземпляров, а по воскресеньям достигал 700 000 [10, p. 65]. Одна из немногих общенациональных газет США распространялась в 11 464 городах страны и по всему миру [10, p. 88]. С конца XIX в. она остается в собственности семьи Окс-Сульцбергеров, семья продолжает управлять изданием, подчеркивая его преемственность и уникальную роль в формировании общественного мнения и фиксации исторических событий. В рассматриваемый нами период владельцем был Артур Хейз Сульцбергер, издатель с 1935 по 1961 гг.

В целом исследователи медиа указывают на продемократический характер редакционной политики «Нью-Йорк Таймс» [8]. Однако, в отдельные выборные кампании газета могла поддержать кандидатов от республиканской партии, как было в случае с Эйзенхауэром. В период Второй Мировой войны партийная ориентация и политические предпочтения американского издания менялись. Вот, что пишет Гаррисон Солсбери, бывший журналист «Нью-Йорк Таймс», по поводу поддержки кандидатов в президенты изданием: «Артур Хейз Сульцбергер одобрил поддержку газетой Франклина Д. Рузвельта в 1936 году, но с большой неохотой. И с радостью переключился на Уэнделла Уилки в 1940 году. Сульцбергер поддержал Рузвельта в 1944 году, но только в качестве меры военного времени» [9, p. 90].

Рассмотрим теперь основные этапы сбора метаданных статей электронного архива «Нью-Йорк Таймс». После регистрации аккаунта на сайте «Нью-Йорк Таймс» пользователь получает уникальный ключ для подключения к API. Подробная документация, включая описание доступных методов и параметров запросов, размещена на портале для разработчиков [3]. Авторские скрипты, использованные для сбора метаданных, можно найти в открытом доступе на github [4]. Там же размещена итоговая таблица, всего 41 676 статей. Ниже рассмотрим лишь ключевые моменты и приемы, примененные в скрипте.

В запросах для поиска использовались ключевые слова, относящиеся к СССР: 'Russians', 'Russian', 'Russia', 'Soviet', 'USSR', 'Stalin', 'Moscow'. Также можно использовать логические операторы OR и AND, что позволяет формировать более гибкие условия для поиска. Синтаксис составления «ключевого слова» выглядит так: 'Russians+AND+religion'. Подобный запрос вернет метаданные всех статей, где одновременно встречаются слова ‘Russians’ и ‘religion’. Поиск не чувствителен к регистру. Если же мы хотим найти статьи, где встречается хотя бы одно из упомянутых слов, то используем оператор OR: ‘Russians+OR+religion’.

Чтобы охватить весь интересующий период, с июня 1941 г. по сентябрь 1945 г., ввиду суточных лимитов по количеству запросов через API (4000 запросов в день), скрипт по сбору метаданных запускался ежедневно со смещением временного окна периода поиска (29 дней за цикл). Полученные данные в формате JSON сохранялись в структурированную таблицу Excel с помощью библиотеки pandas для дальнейшего анализа.

Денормализация итоговой таблицы проводилась путем добавления туда полей из поля keywords типа dictionary [5]. В результате получена была «длинная» таблица, дополненная полями key_name – категория тегов (тема статьи, персоны, локации и т.д.) и key_value – значение указанного тега для статьи. Для одной статьи может быть

множество значений тегов одной категории, которые будут отличаться значением поля key_rank (Таблица 1).

head-line.main	**web_url**	**pub_date**	**byline.original**	**key_name**	**key_value**	**key_rank**
<заголовок статьи>	<ссылка на статью>	<дата публ-ции>	<автор статьи>	<категория тега>	<значение тега>	<ранг тега>

Таблица 1. Структура данных итоговой таблицы со значениями тегов keywords

Остановимся подробнее на поле keywords. Это поле содержит информацию о темах (subject), персонах (person) и локаций (locations) упомянутых в статье. Значения этих полей заполнялись, индексировались сотрудниками «Нью-Йорк Таймс» с 1913 г. для каждой публикации, попавшей на страницы издания. «Индекс, публикуемый с 1913 г., резюмирует день за днем мировую историю, как она представлена в Нью-Йорк Таймс» [11].

Тема публикации является одним из самых распространенных признаков в исследованиях медиа, использующих метод контент-анализа медиа. Тем самым, мы получаем поле, заполненное квалифицированными сотрудниками авторитетного издания и верифицированное редактором отдела индексирования и таксономии, которое мы можем использовать в исследованиях.

Структура данных, возвращаемых API, включает множество полей, но остановимся только на значимых для дальнейшего анализа. Некоторые из потенциально значимых полей не заполнены для данного периода. Поле key_value, в котором находятся значения «индекса», заполнено более чем в 70% публикаций в период с июня 1941 по сентябрь 1945 г. Также стабильно высокие показатели заполнения в динамике по годам (Рисунок 1). Этого достаточно, чтобы оценивать общие тенденции и предпочтения редакции.

Рисунок 1. Результаты анализа качества данных,
процент заполненности основных полей полученных метаданных

Рассмотрим пример использования значений «индекса» для исследовательского анализа американского издания. Анализ динамики публикаций в американском издании, посвященных СССР, на примере темы религии, показывает, что наибольший интерес к этой теме пришелся на осень 1941 г. Это связано с заявлением С.А. Лозовского от 4 октября 1941 г., в котором подчеркивалась гарантия свободы религиозных обрядов в СССР и активное сопротивление представителей всех конфессий нацизму. Данное заявление способствовало снижению напряженности в американском обществе и укреплению позиций президента Рузвельта в Конгрессе, что в итоге помогло успешному принятию закона о ленд-лизе в отношении СССР [1, с. 146]. Даже восстановление патриаршества в сентябре 1943 г. получило значительно меньше внимания в прессе (Рисунок 2).

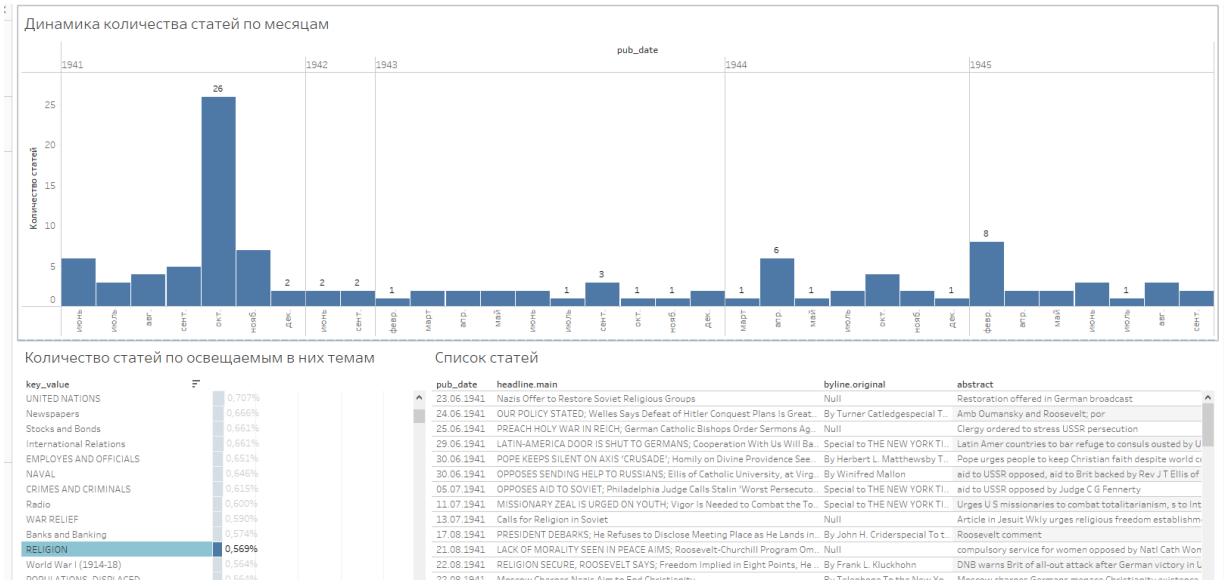

Рисунок 2. Динамика количества публикаций с тегом «Religion» для статей с упоминанием СССР

Сформированная база метаданных «Нью-Йорк Таймс» по статьям, упоминающим СССР за период 1941–1945 гг. представляет собой ценный ресурс для исторических исследований медиа, служа каркасом для более детального контент-анализа. Она позволяет не только изучать общие тенденции в освещении ключевых событий, таких как в приведенном примере по вопросу религиозной свободы в СССР, но и обогащать анализ дополнительными авторскими признаками кодирования статей при проведении более сложных методов контент-анализа. Современные методы NLP (Natural Language Processing) могут быть эффективно применены для анализа заголовков и аннотаций статей, что открывает новые возможности для выявления семантических паттернов и тематических акцентов. Однако важно учитывать, что «индекс», присвоенный сотрудниками «Нью-Йорк Таймс», может быть субъективным и неполным. Тем не менее, даже с учетом этих ограничений, база метаданных остается мощным инструментом для изучения роли медиа в формировании общественного мнения и их влияния на ключевые политические процессы, такие как укрепление антигитлеровской коалиции.

1. Великая Отечественная война 1941–1945 годов : в 12 т. – Т. 8 : Внешняя политика и дипломатия Советского Союза в годы войны. – М. : Кучково поле, 2014. – 864 с.

2. Dallek, R. Franklin D. Roosevelt and American Foreign Policy, 1932–1945 / R. Dallek. – NY : Oxford University Press, 1979. – 680 p.

3. https://developer.nytimes.com/get-started (date of access: 17.03.2025).

4. https://github.com/dmitry-kosachev/NYT_archive_metadata_parsing_on_USSR?tab=readme-ov-file (date of access: 17.03.2025)
5. https://github.com/dmitry-kosachev/NYT_archive_metadata_parsing_on_USSR/blob/main/flattering_the_keywords_column.ipynb (date of access: 17.03.2025)
6. The News and Public Opinion: Media Effects on Civic Life / M. McCombs, R. Lance Holbert, S. Kiousis, W. Wanta. – Cambridge : Polity Press, 2011. – 210 p.
7. Leff, L. Buried by the Times : the Holocaust and America's Most Important Newspaper / L. Leff. – New York : Cambridge University Press, 2005. – 426 p.
8. Puglisi, R. Being The New York Times: the Political Behaviour of a Newspaper / R. Puglisi // The B.E. Journal of Economic Analysis & Policy – Vol. 11 – 2011. Iss. 1 (Contributions), Art. 20.
9. Salisbury, H. E. Without Fear or Favor: the New York Times and its Times / H. E. Salisbury. – NY Times Books, 1980. – 648 p.
10. Talese, G. The Kingdom and the Power: Behind the Scenes at The New York Times / G. Talese. – NY, Random House Trade Paperback, 2007. – 596 p.
11. Times Index Names Editor; Gephart to Be Consultant – 28.09.1960. – URL: https://timesmachine.nytimes.com/timesmachine/1964/09/28/118537970.html?pageNumber=30&login=smartlock&auth=login-smartlock (date of access: 17.03.2025).

Чэчулін З.В.

“ТЭКІ НАРУШЭВІЧА” ЯК КРЫНІЦЫ ПА ГІСТОРЫІ І РЭКАНСТРУКЦЫІ АРХІВА ВЯЛІКІХ КНЯЗЁЎ ЛІТОЎСКІХ (ДРУГАЯ ПАЛОВА XIV – ПЕРШАЯ ПАЛОВА XV СТ.)

Ключавыя словы: Адам Нарушэвіч, “Тэкі Нарушэвіча”, Мацей Догель, архіў вялікіх князёў літоўскіх, архіў Каралеўства Польскага, акт, дакумент.

“Тэкі (папкі) Нарушэвіча” былі створаныя ў перыяд з 1781 па 1785 гг. [14, s. 19] і складаюцца з 231 рукапіснага тома, якія ўтрымліваюць спісы (копіі) 38270 гістарычных крыніц. Каардынатарам праекта быў вядомы дзеяч навукі, вялікі пісар ВКЛ (з 1781 г.) Адам Нарушэвіч (1733–1796) [1, с. 348]. Падставай для стварэння “тэк” была праца А. Нарушэвіча над “Гісторыяй польскага народа”, для напісання якой патрабавалася выявіць і скапіяваць крыніцы. Храналагічна “тэкі” ахопліваюць перыяд з 1227 па 1729 г. [27] і сістэматызаваныя па храналагічнаму прынцыпу. Аднак з улікам таго, што перапішчыкі ў некаторых выпадках памыляліся з устанаўленнем правільнай даты дакумента, гэта часткова вяло да парушэння гэтага прынцыпу.

На дадзены момант з 231 тома “тэк Нарушэвіча” 217 тамоў знаходзяцца на захоўванні ў Бібліятэцы князёў Чартарыйскіх у Кракаве [14, s. 22]. Цікавячы нас перыяд прадстаўлены тамамі № 7–17. Капіісты de visu працавалі з арыгіналамі дакументаў з аддзела “Lithuania” архіва Каралеўства Польскага, які быў перавезены ў 1765 г. з Кракаўскага замка ў Варшаўскі каралеўскі замак [24, s. 125]. На гэта указваюць адпаведныя запісы, змешчаныя ў правых верхніх кутах аркушаў: “Ex Originali in archivo Regnis”, або “Ex ms arch. Stan. Aug. Regni”. Храналагічна першы дакумент, скапіяваны з арыгінала, датуецца 7 лістапада 1367 г. [13, № 65, k. 261–262], апошні – 6 чэрвеня 1393 г. [14, № 142, k. 565–566]. Гэтая частка дакументаў архіва вялікіх князёў літоўскіх трапіла ў архіў Каралеўства Польскага яшчэ ў 1386 г. пры ажыццяўленні каранацыйных мерапрыемстваў Ягайлам, а архіў князя Скіргайлы трапіў у склад архіва Каралеўства Польскага або ўвосень 1393 г., калі Скіргайла пачаў знаходзіцца пры каралеўскім двары Уладзіслава II Ягайлы [5, с. 32], або пасля смерці Скіргайлы 10 студзеня 1395 г. у Кіеве [4, с. 23].

Што датычна храналагічна пазнейшых дакументаў, то Нарушэвіч і яго група капіістаў не працавалі з іх арыгіналамі de visu. Адносна гэтага ў правых верхніх кутах аркушаў маюцца адпаведныя запісы: “Ex archivo Regnis Collect. Dogieli”. У некаторых выпадках да апошніх запісаў дадавалася: “Ex originali Niesvisciense”, што ўказвае на