Историческая информатика
Правильная ссылка на статью:
Галушко И.Н. — Корректировка результатов OCR-распознавания текста исторического источника с помощью нечетких множеств (на примере газеты начала XX века) // Историческая информатика. – 2023. – № 1. DOI: 10.7256/2585-7797.2023.1.40387 EDN: OCFBSP URL: https://nbpublish.com/library_read_article.php?id=40387

Корректировка результатов OCR-распознавания текста исторического источника с помощью нечетких множеств (на примере газеты начала XX века)

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:

10.7256/2585-7797.2023.1.40387

EDN:

OCFBSP

Дата направления статьи в редакцию:

06-04-2023

Аннотация: Наша статья посвящена попытке применения современных методов NLP для оптимизации процесса распознавания текста исторических источников. Любой исследователь, решивший воспользоваться инструментами распознавания отсканированных текстов, столкнется с рядом ограничений точности конвейера (последовательности операций распознавания). Даже наиболее качественно обученные модели могут давать существенную ошибку по причине неудовлетворительного состояния дошедшего до нас источника: порезы, изгибы, кляксы, стертые буквы – всё это мешает качественному распознаванию. Наше предположение состоит в том, что, используя заранее заданный набор слов, маркирующих присутствие интересующей нас темы, с помощью модуля нечетких множеств (Fuzzy sets) из NLP-библиотеки SpaCy, мы сможем восстановить по шаблонам те слова, которые по итогам процедуры распознавания оказались распознаны с ошибками. Для проверки качества процедуры восстановления текста на выборке из 50 номеров газеты «Биржевые ведомости» мы посчитали оценки количества слов, которые бы не вошли в семантический анализ из-за неправильного распознавания. Все метрики были посчитаны также с использованием паттернов нечетких множеств. Оказалось, что в среднем на номер «Биржевых ведомостей» приходится 938.9 слов, маркирующих тему нашего исследования –

торговые и финансовые операции с ценными бумагами. Из них изначально правильно распознаются в среднем 87.2% слов. Примерно 119.6 слов (в среднем на 50 номеров) содержат опечатки, связанные с некорректным распознаванием. Благодаря использованию алгоритмов нечетких множеств нам удалось эти слова восстановить и включить в семантический анализ. Мы считаем, что восполнение 12.8% слов, потенциально относящихся к изучаемой теме – это хороший результат, существенно повышающий качество дальнейшего семантического анализа текста методами компьютерного моделирования.

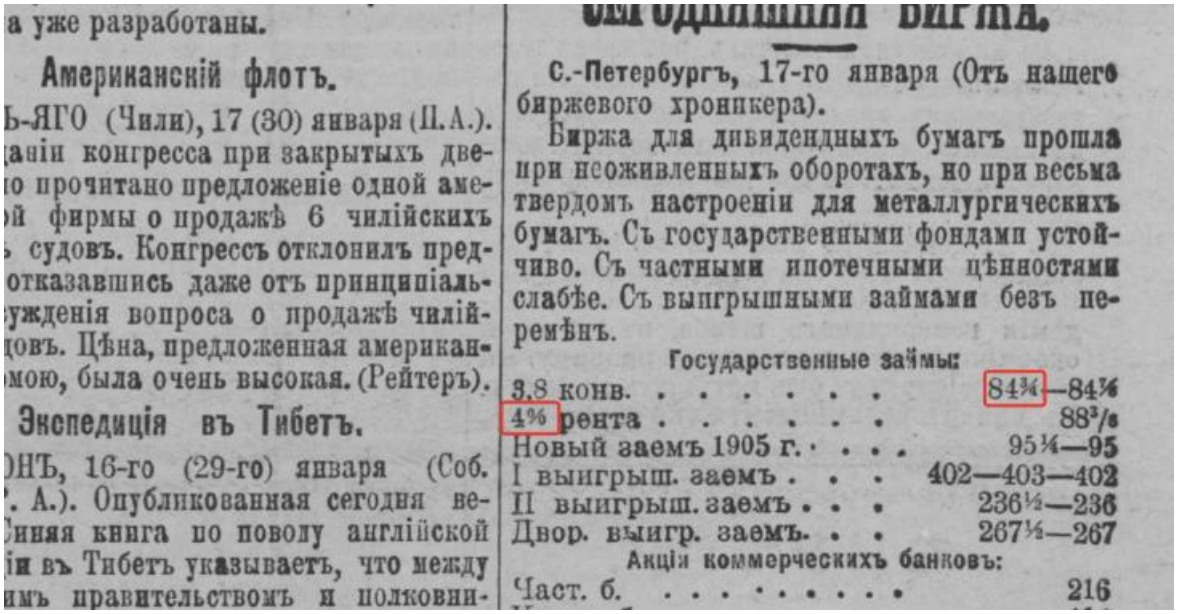
Ключевые слова:

распознавание исторических источников, исправление OCR, нечеткие множества, обработка естественного языка, предобработка текста, Биржевые ведомости, расстояние Левенштейна, контент-анализ, тематическое моделирование, исторические газеты

На современном этапе социально-гуманитарные науки всё чаще обращаются к методам машинной обработки текстов. Для специалистов в области исторической информатики стали классическими задачи сетевого анализа смысловых категорий газетных текстов [1]; определения авторства по устойчивым паттернам стиля, формализованного в грамматических категориях текста [2]; семантического сравнения корпусов текстов методами контент-анализа [3]. Появляются новые направления: разведочный анализ исторических источников посредством тематического моделирования [4], применение искусственного интеллекта в задачах восстановления поврежденных источников [5]. Все эти направления объединяет требование к качеству оцифрованного и распознанного текста. Сегодня уже излишним представляется обсуждение вопроса о степени влияния качества подготовки текста на итоговые результаты моделей любого уровня сложности [6]. Мировое сообщество исследователей настолько ясно осознает необходимость создания наиболее точных и гибких инструментов распознавания текста, что данную область можно назвать чуть ли не единственной вспомогательной дисциплиной исторической науки, для которой развитие методов AI и NLP (natural language processing) привело к знаковому росту количества публикаций. Только за последние три года на сайте *Paperswithcode* можно найти более 40 публикаций, посвященных проблематике перевода исторических источников в машиночитаемый вид [7]. И мы говорим только о публикациях, авторы которых оставили сообществу все свои программные наработки в открытом доступе.

Наша статья посвящена попытке применения современных методов NLP для оптимизации процесса распознавания текста исторических источников. Сегодня существует несколько популярных платформ для проведения этой процедуры. В этом контексте безусловно стоит отметить европейский проект Transkribus, прямо декларирующий своей целью распознавание именно исторических текстов. Ключевым преимуществом своей платформы разработчики считают наличие большого количества предобученных языковых моделей для разных эпох (как для рукописных, так и для печатных текстов). Transkribus – это платформа для исследователей, требующая внесения ежегодных взносов и предоставляющая право долевого участия в проекте. Речь идет о доступе к огромной библиотеке пользователей шаблонов, созданных с помощью встроенного инструментария платформы. Для русского языка доступны пять таких моделей: две для печатного шрифта XVIII века и три – для рукописных текстов конца XIX-начала XX в.[8].

Другой популярный инструмент для распознавания исторических текстов – программа ABBY FineReader. В стандартную библиотеку FineReader входит дореволюционный русский язык, что может оказаться крайне удобно для исследователей, работающих с газетными или другими печатными текстами XIX – нач. XX в. Данная программа предоставляет возможность создания пользовательских шаблонов, добавляющих в библиотеки FR новые символы, обученные по инструкции исследователя. В случае с газетными материалами начала XX в. данная функция может существенно повысить качество распознавания сложных символов, не имеющих аналогов в стандартных словарях: например, специфических дробей или сокращений.



а уже разработаны.

Американскій флотъ.

Ь-ЯГО (Чили), 17 (30) января (П.А.). аніи конгресса при закрытыхъ две- ю прочитано предложеніе одной аме- й фирмы о продажѣ 6 чилійскихъ судовъ. Конгрессъ отклонилъ пред- отказавшись даже отъ принципіаль- уженія вопроса о продажѣ чилій- довъ. Цѣна, предложенная американ- мою, была очень высокая. (Рейтеръ).

Экспедиція въ Тибетъ.

ОНЪ, 16-го (29-го) января (Соб. . А.). Опубликованная сегодня ве- Синяя книга по поводу англійской ія въ Тибетъ указываетъ, что между имъ правительствомъ и полковни-

С.-Петербургъ, 17-го января (Отъ нашего биржевого хроникера).

Биржа для дивидендныхъ бумагъ прошла при неоживленныхъ оборотахъ, но при весьма твердомъ настроеніи для металлургическихъ бумагъ. Съ государственными фондами устой- чиво. Съ частными ипотечными цѣнностями слабѣе. Съ выигрышными займами безъ пе- ремѣнъ.

Государственные займы:

3,8 конв.	84¾—84⅝
4% рента	88⅜
Новый заемъ 1905 г.	95¼—95
I выигрыш. заемъ	402—403—402
II выигрыш. заемъ	236½—236
Двор. выигр. заемъ.	267½—267

Акціи коммерческихъ банковъ:

Част. б.	216

Рис. 1. Скан номера Биржевых ведомостей за (17 (30) янв.) 1905 года, № 8618. Красным цветом выделены символы, для которых исследователю вероятно придется создавать пользовательские шаблоны в любой из программ OCR

Еще один многообещающий проект – OCR-D – большая совместная разработка модуля Python немецкими специалистами в области Digital Humanities и прилагающейся к нему системы API для распознавания готического шрифта XVI -XVIII веков[9]. Интересен данный проект заявленной целью создания комплекса публично доступных пакетов, содержащих оптимизированные под исторические источники функции распознавания, а также предобученные модели, ключевой особенностью которых будет упор на качественное сегментирования текста на оцифрованной странице.

Сегодня проект находится на финальной стадии разработки, уже в 2024 году планируется масштабное применение технологии для перевода в машиночитаемый вид коллекции библиотеки Галле-Виттенбергский университета – ключевого партнера OCR-D. В России схожий проект реализуется, например, в Российской академии народного хозяйства и государственной службы (РАНХиГС) силами команды под руководством Р.Б. Кончакова [10]. Проект предполагает распознавание коллекции рукописных листов губернаторских отчетов с последующим созданием универсального приложения для оцифровки рукописных исторических источников. И хотя возможность создания подобного универсального приложения сегодня дискуссионна, данное начинание безусловно является шагом на пути дальнейших цифровизации исторических исследований.

* * *

Этот краткий обзор предваряет наше исследование. Дело в том, что любой исследователь, решающий воспользоваться инструментами распознавания отсканированных текстов, столкнется с рядом ограничений точности конвейера (последовательности операций распознавания). И даже наиболее качественно обученные модели могут давать существенную ошибку как минимум по причине неудовлетворительного состояния дошедшего до нас источника: порезы, изгибы, кляксы, стертые буквы – всё это мешает качественному распознаванию. Из-за этого, пытаясь распознать слово «акционер», исследователь может получить «ак_.онер», «акпионер», «акцноиер» и т.п. Данная статья предлагает одно из решений обозначенной проблемы. Стоит заметить, что предложенная методика подойдет только в том случае, если последующая обработка текста, предшествующая формализации, будет включать в себя лемматизацию и фильтрацию по критерию TF-IDF (или схожую с ним метрику оценки значимости слова в тексте и корпусе). (см. Приложение 1)

Большая часть исторических исследований, включающих компьютеризованный анализ текста, предполагает либо заданный, либо искомый набор категорий, отражающих семантическое ядро корпуса изучаемых текстов. В случае с контент-анализом такая система категорий задается исследователем; в случае с тематическим моделированием – система тем и слов, отражающих эти темы, определяется машинным способом, без участия исследователя. Обе методики предполагают количественную оценку наличия в текстах тех слов, которые маркируют тематики, интересующие исследователя. В ряде исследований такой набор задавался изначально еще на подготовительном этапе [11].

Наше исследование, для которого и была разработана описываемая в статье методика оптимизация распознавания текстов исторических источников, посвящено изучению доходности ценных бумаг на Санкт-Петербургской фондовой бирже в начале XX в. с позиции поведенческих финансов. Нас интересуют принципы стоимостной оценки публичных компаний – как определялись приемлемые или недостаточные уровни капитализации, можно ли выделить отраслевую специфику; вопросы выплаты купонов и дивидендов – какие уровни доходности считались достаточными и можем ли мы говорить о наличии какого-либо бенчмарка (бумаги-эталона с безрисковой ставкой) для деловой среды начала XX в. 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 её 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 отсканированных номеров за 1905 и 1913 года, утренние и вечерние выпуски).

Рис. 2. Скан номера «Биржевых ведомостей» за (25 окт. (7 нояб.)) 1913, № 13822 (25 окт. (7 нояб.)). Нас интересует колонка «Извлечение из протокола вторичного чрезвычайного собрания акционеров».

>РУ* ц-о гчфб-еще предсѣдатель собранія предложилъ на обсужденіе третій вопросъ повѣстки: «Объ избраніи новыхъ членовъ правленія и объ увеличеніи числа членовъ правленія». Послѣ продолжительныхъ преній предсѣдатель собранія поставилъ на разрѣшеніе общаго соб5'анія вопросъ: <...> По произведенному закрытому голосованію ' оказалось, цте за предложеніе подано было 783 голоса, а противъ предло-. женія 1.610 голосовъ. Въ виду отрицательнаго разрѣшенія этого вопроса, : вопросъ о выборѣ новыхъ членовъ правненія отпалъ, и потому предсѣдатель собранія не ставилъ его на баллотировку, <...> По произведенной закрытой баллотировкѣ оказалось, что за предложеніе подано 570 голосовъ, противъ 1.386 голосовъ. По разрѣшеніи этого послѣдняго вопроса повѣстки, многіе гг. акпци-неры предложили выразить правленію благодарность за его полезную дѣятельность, и полное довѣріе. Г. Предсѣдатель собранія заявилъ, что, если от гг. акціонеровъ не послѣдуетъ возраженій, то выраженіе благодарности будетъ занесено въ протоколъ.

Рис. 3. Результаты распознавания газетного текста. На рис. 3 выделены примеры неудачно распознанных слов, которые оказываются важны для реализации формализованного анализа текста в рамках нашего исследования. Также на первой строчке мы можем наблюдать «мусорные» последовательности символов, распознанных некорректно.

Наше предположение состояло в том, что, используя заранее заданный набор слов, маркирующих присутствие интересующей нас темы (доходность ценных бумаг), с помощью модуля нечетких множеств (Fuzzy sets) из NLP-библиотеки SpaCy, мы сможем восстановить по шаблонам те слова, которые по итогам процедуры распознавания оказались распознаны с ошибками [12]. Для того чтобы использовать весь мощный потенциал обученных моделей русского языка в SpaCy (в том числе для последующей лемматизации и бинаризации), мы перевели распознанный текст из дореволюционной орфографии в современную посредством использования библиотеки Russpelling для

языка программирования Python, разработанной И. Бёрном и Д. Бёрнбаумом в рамках проекта Софийского университета по применению методов NLP в области славянской филологии [13]. Здесь же стоит заметить, что используемая далее библиотека SpaCy доступна как для Python, так и для популярного в социально-гуманитарных науках языка программирования R [14].

Далее все слова в распознанных источниках были проверены на орфографию через модуль *pyenchant* на основе публично доступного словаря русского языка от LibreOffice. Мы предположили, что все неправильно распознанные слова данный тест на правописание не пройдут. Сохраняя все слова на своих местах, мы отобрали все непрошедшие тест слова в отдельную коллекцию. Для восстановления оригинального слова методом нечетких множеств по расстоянию Левенштейна (через эту метрику реализованы все операции нечеткого сравнения в SpaCy, см. Приложение 1) нам было необходимо задать набор шаблонов. Так, мы выбрали следующие слова, которые маркируют интересующие нас тему:

Таблица 1. Перечень паттернов Fuzzy-match и соответствующих им форм слов.

Индекс	Паттерн	Формы
0	облигация	['облигация', 'облигацией', 'облигаций', 'облигации', 'облигацию', 'облигациею', 'облигациях', 'облигациям', 'облигациями']
1	процент	['процента', 'проценте', 'процентах', 'процентом', 'процент', 'процентами', 'проценту', 'проценты', 'процентам', 'процентов']
2	проц	['проц']
3	заем	['займа', 'заём', 'займы', 'займам', 'займ', 'займе', 'займов', 'займу', 'займами', 'займах', 'займом']
4	государственные	['государственная', 'государственных', 'государственно', 'государственной', 'государственным', 'государственною', 'государственными', 'государственны', 'государственный', 'государственное', 'государственен', 'государственные', 'государственную', 'государственном', 'государственному', 'государственного']
5	фонды	['фондов', 'фонда', 'фонд', 'фондами', 'фонды', 'фонде', 'фондам', 'фондах', 'фонду', 'фондом']
6	бумага	['бумаге', 'бумагою', 'бумагу', 'бумагах', 'бумага', 'бумаг', 'бумагами', 'бумаги', 'бумагам', 'бумагой']
7	дисконт	['дисконте', 'дисконт', 'дисконта', 'дисконты', 'дисконтов', 'дисконту', 'дисконтам', 'дисконтах', 'дисконтами', 'дисконтом']
8	купон	['купонам', 'купона', 'купоне', 'купонах', 'купон', 'купонами', 'купоном', 'купону', 'купонов', 'купоны']
		['рентою', 'рент', 'рентами', 'рентам', 'рентой',

9	рента	['ренною', 'рент', 'ренами', 'ренам', 'реной', 'рентах', 'рента', 'ренту', 'ренты', 'ренте']
10	ипотечный	['ипотечный', 'ипотечными', 'ипотечною', 'ипотечному', 'ипотечных', 'ипотечном', 'ипотечное', 'ипотечную', 'ипотечного', 'ипотечна', 'ипотечная', , 'ипотечен', 'ипотечным', , 'ипотечной', 'ипотечные']
11	обязательство	['обязательству', 'обязательства', 'обязательство', 'обязательствами', 'обязательстве', 'обязательством', 'обязательств', 'обязательствах', 'обязательствам']
12	доходность	['доходность', 'доходностями', 'доходностях', 'доходности', 'доходностям', 'доходностью', 'доходностей']
13	котировка	['котировкам', 'котировку', 'котировкою', 'котировке', 'котировка', 'котировках', 'котировок', 'котировки', 'котировками', 'котировкой']
14	эмиссия	['эмиссии', 'эмиссиями', 'эмиссий', 'эмиссию', 'эмиссией', 'эмиссия', 'эмиссиею', 'эмиссиям', 'эмиссиях']
15	баланс	['балансом', 'балансах', 'баланса', 'балансам', 'балансе', 'баланс', 'балансами', 'балансы', 'балансов', 'балансу']
16	русский	['русская', 'русскую', 'русское', 'русскими', 'русского', 'русском', 'русским', 'русский', 'русские', 'русских', 'русскому', 'русскою', 'русской']
17	ценности	['Ценность', 'ценности', 'ценностях', 'ценностью', 'ценностями', 'ценностям', 'ценностей', 'ценность']
18	денежный	['денежною', 'денежную', 'денежным', 'денежное', 'денежных', 'денежными', 'денежному', 'денежный', 'денежном', 'денежного', 'денежные', 'денежная', 'денежной']
19	рынок	['рынком', 'рынок', 'рынками', 'рынкам', 'рынках', 'рынка', 'рынков', 'рынке', 'рынки', 'рынку']
20	дивиденд	['дивидендом', 'дивидендами', 'дивиденда', 'дивиденде', 'дивидендов', 'дивидендах', 'дивидендам', 'дивиденду', 'дивиденды', 'дивиденд']
21	прибыль	['прибылям', 'прибылью', 'прибылей', 'прибылями', 'прибыли', 'прибыль', 'прибылях']
		['акционерную', 'акционерною', 'акционерное', 'акционерные', 'акционерными', 'акционерном', 'акционерный', 'акционерного', 'акционерной', 'акционерных', 'акционерным',

22	акционерный	'акционерная', 'акционерному']
23	капитал	['капитале', 'капиталы', 'капиталу', 'капиталам', 'капиталов', 'капитала', 'капитал', 'капиталом', 'капиталами', 'капиталах']
24	акция	['акцией', 'акциями', 'акциях', 'акций', 'акцию', 'акциею', 'акциям', 'акция', 'акции']
25	лимитированный	['лимитированный']
26	приказ	['приказов', 'приказом', 'приказе', 'приказа', 'приказах', 'приказам', 'приказу', 'приказы', 'приказ', 'приказами']
27	общество	['обществами', 'обществ', 'общества', 'общество', 'обществом', 'обществу', 'обществе', 'обществам', 'обществах']
28	собрание	['собрания', 'собрание', 'собрании', 'собранию', 'собраниях', 'собраниям', 'собраниями', 'собранием', 'собраний']
29	акционер	['акционерам', 'акционер', 'акционеров', 'акционерах', 'акционерами', 'акционера', 'акционеры', 'акционеру', 'акционере', 'акционером']
30	член	['члена', 'член', 'членов', 'члены', 'членах', 'членом', 'члену', 'членами', 'членам', 'члене']
31	правление	['правлений', 'правлениях', 'правлениям', 'правлением', 'правлению', 'правления', 'правление', 'правлениями', 'правлении']

На следующем этапе все слова, непрошедшие тест на правописание, сравниваются с приведенным набором шаблонов. Если по расстоянию Левенштейна (использовался порог в 1 или 2 перестановки) из поданного в конвейер слова невозможно восстановить шаблон – оно удаляется из текста. Здесь может возникнуть вопрос, насколько правомерным может быть такое удаление. Поскольку в нашем исследовании и так применяется TF-IDF фильтр, взвешивающий относительную значимость слов для отдельного текста и корпуса в целом, можно резонно предположить, что большая часть плохо распознанных слов в любом случае бы подлежала удалению как фоновая лексика. Если по нашему тесту нечеткого совпадения из слова можно восстановить заданный шаблон – мы заменяем в распознанном тексте слово с ошибкой на слово-шаблон, определенный методом fuzzy-match. Так, слово «акпионер» заменится на «акционер», «.ронды» на «фонды» и т.д.

председатель собрания предложил на обсуждение третий вопрос повестки: «об избрании новых членов правления и об увеличении числа членов правления» после продолжительных прений председатель собрания поставил на разрешение общего **собрания** вопрос: <…> по произведенному закрытому голосованию оказалось, за предложение подано было 783 голоса, а против 1.610 голосов в виду отрицательного разрешения этого вопроса, вопрос о выборе новых членов **правления** отпал, и потому председатель собрания не ставил его на баллотировку, <…> по произведенной закрытой баллотировке оказалось, что за предложение подано 570 голосов, против 1.386 голосов по разрешении этого последнего вопроса повестки, многие **акционеры** предложили выразить правлению благодарность за его полезную деятельность, и полное доверие председатель собрания заявил, что, если кого из акционеров не последует возражений, то выражение благодарности будет занесено в протокол

Рис. 4. Результат работы конвейера нечетких множеств. Текст, представленный на рис.3, был успешно очищен и восстановлен.

На рис. 4.можно наблюдать результат проведения операций по преобразованию искаженного текста, который теперь удобно поддается семантическому анализу всеми доступными средствами компьютеризированной обработки текста.

Для проверки качества процедуры восстановления текста на выборке из 50 номеров газеты «Биржевые ведомости» мы посчитали оценки количества слов, которые бы не вошли в семантический анализ из-за неправильного распознавания. Все метрики были посчитаны также с использованием паттернов нечетких множеств. Оказалось, что в среднем на номер «Биржевых ведомостей» приходится 938.9 слов, маркирующих тему нашего исследования – торговые и финансовые операции с ценными бумагами. Из них изначально правильно распознаются в среднем 87.2% слов. Примерно 119.6 слов (в среднем на 50 номеров) содержат опечатки, связанные с некорректным распознаванием. Благодаря использованию алгоритмов нечетких множеств нам удалось эти слова восстановить и включить в семантический анализ. Мы считаем, что восполнение 12.8% слов, потенциально относящихся к изучаемой теме (перечисленных в Таблице 1) – это хороший результат, существенно повышающий качество дальнейшего семантического анализа текста методами компьютерного моделирования. С помощью предложенной методики мы смогли обработать все вошедшие в нашу коллекцию номера газет.

Для интересующихся читателей мы оставляем ссылку на доступ к предложенному автором программному коду, исходным материалам и распознанным текстам [15].

***

Таким образом, можно констатировать, что применение методов нечетких множеств позволяет существенно повысить качество распознания текста в тех случаях, когда распознавание не рассматривается как конечная цель задачи исследователя, работающего с документов. Безусловно, описанная методика является частным случаем для задачи тематического моделирования и/или контент-анализа, когда исключение некоторых слов документа является частью комплекса операций по предобработке текста. Мы считаем, что применение алгоритмов нечетких множеств в проектах по распознаванию текстов, находящихся на архивном хранении, потребует принципиально иной архитектуры конвейера. Не имея возможности исключать слова, а также следуя необходимости соблюдать грамматическую согласованность текста, исследователь, намеренный применить алгоритмы нечетких множеств, будет решать задачу совершенно

иного порядка, требующую детального учета контекста каждого слова. Однако в то же время представленный в данной статье пример показывает, что область Fuzzy-технологий обладает значительным потенциалом в задачах оптимизации текстового распознавания.

Приложение 1: Словарь определений

Расстояние Левенштейна. В теории информации, лингвистике и информатике расстояние Левенштейна (LEV) — это строковая метрика для измерения разницы между двумя последовательностями. Неформально расстояние Левенштейна между двумя словами можно трактовать как минимальное количество односимвольных правок (вставок, удалений или замен), необходимых для замены одного слова другим. Например, LEV(«Крт», «Кот») = 1, так как потребуется провести одну замену "Р" на "О". Названо в честь советского математика Владимира Левенштейна, предложившего эту метрику в 1965 году.

Нечеткое множество — понятие, введённое Лотфи Заде в 1965 году в статье «Fuzzy Sets», в котором он расширил классическое понятие множества, допустив, что характеристическая функция множества (названная Заде функцией принадлежности для нечёткого множества) может принимать любые значения в интервале (0,1), а не только значения 0 или 1. В нашей статьи мы предполагаем, исходя из нечеткой логики, что хотя слово «акп-онер» написано неверно, сам объект (слово в изучаемом тексте) приводим к форме «акционер», а допустимой мерой «нечеткости» является расстояние Левенштейна, равное 2.

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.

Библиография

1. Солощенко Н.В. Многотиражная газета «Бабаевец» как источник по истории пищевой промышленности СССР в годы первой пятилетки (опыт контент-анализа и сетевого анализа) // Историческая информатика. — 2021.-№ 2.-С.1-23.
2. Kale, Sunil Digamberrao and Rajesh Shardanand Prasad. "A Systematic Review on Author Identification Methods." Int. J. Rough Sets Data Anal. 4 (2017): 81-91.
3. Гарскова И.М. Международная научная конференция «Аналитические методы и информационные технологии в исторических исследованиях: от оцифрованных данных к приращению знаний» // Историческая информатика. — 2018.-№ 4.-С.143-151.
4. Tze-I Yang, A.J.Torget, R.Mihalcea. Topic modeling in historical newspapers. 2011
5. Assael, Y., Sommerschield, T., Shillingford, B. et al. Restoring and attributing ancient texts using deep neural networks. Nature 603, 280–283 (2022).
6. Lopresti, Daniel. (2009). Optical character recognition errors and their effects on natural language processing. IJDAR. 12. 141-151.
7. Papers with Code. URL: https://paperswithcode.com/sota
8. Transkribus. Public models. URL: https://readcoop.eu/transkribus/public-models/
9. OCR-D. URL: https://ocr-d.de/en/
10. Доклад Р.Б. Кончакова (РАНХиГС) и С.В. Боловцова (РАНХиГС) «Распознавание

отчетов начальников губерний Российской империи: вызовы и подходы» был представлен на семинаре «Искусственный интеллект в исторических исследованиях: автоматизированное распознавание текстов рукописных исторических источников», организованном ассоциацией «История и компьютер» и РАНХиГС на площадке РАНХиГС 11 февраля 2023 г.: https://ion.ranepa.ru/news/budushchee-istorii-kak-tsifrovye-navyki-otrazhayutsya-na-rabote-istorikov/

11. Солощенко Н.В. Многотиражная печать как источник по изучению процесса формирования «нового человека» в советской промышленности первых пятилеток // Исторический журнал: научные исследования. — 2019.-№ 3.-С.106-117.
12. SpaCy. URL: https://spacy.io/
13. Russpelling. URL: https://github.com/ingoboerner/russpelling
14. SpaCyR. URL: https://cran.r-project.org/web/packages/spacyr/vignettes/using_spacyr.html
15. GitHub. URL: https://github.com/iodinesky/Fuzzy-sets-in-historical-sources-OCR

Результаты процедуры рецензирования статьи

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

Рецензируемая статья относится к чрезвычайно актуальному направлению исследований, связанному с созданием электронных версий и библиотек исторических источников путем распознавания их сканированных изображений. В статье предложена оригинальная методика оптимизации процесса распознавания текста исторических источников на основе методов NLP (natural language processing).
Методология статьи основана на современных представлениях о предметных областях взаимодействия гуманитарных наук (прежде всего исторической) с возможностями информатики, главным образом с теми ее направлениями, которые связаны с интеллектуальным анализом данных, в частности реализацией методов нечетких множеств.
Актуальность исследования определяется огромным и продолжающим постоянно возрастать интересом научного сообщества к созданию полноценных библиотек исторических источников, позволяющих иметь дело непосредственно с цифровыми их версиями, дающими возможность применить целый спектр методов и технологий их содержательного изучения, в частности контент-анализа.
Научная новизна рецензируемой статьи заключается в изложении новой методики распознавания текста и представлении результатов ее использования, которые однозначно указывают на то, что это шаг вперед в направлении дальнейшей оптимизации распознавания исторических текстов.
Построение и содержание статьи полностью соответствует современным представлениям об изложении результатов научных исследований. После постановки проблемы определяются основные задачи работы, дается краткая характеристика ряда платформ для распознавания текста исторических источников и реализации проектов на их основе. В следующем разделе статьи на примере газеты «Биржевые ведомости» изложена собственно методика оптимизации распознавания для корректировки неудачно распознанных слов. Представленные примеры очищения и восстановления текстов показывают полезность и результативность методики с использованием нечетких множеств. Статья содержит удачные иллюстрации, позволяющие лучше раскрыть ее содержание, а также дополняется небольшим словарем терминов, использованных при

изложении материала. Особо хочется отметить язык и стиль рецензируемого текста, который, несмотря на очевидную сложность содержания, излагает основные моменты проведенного исследования понятно и вполне доступно не только для специалистов по распознаванию текстов, но и для широкого круга потенциальных читателей статьи.

Библиография статьи невелика по объему, соответствуя достаточно специфической проблематике исследования. В ней приведены ссылки на рассмотренные платформы распознавания исторических текстов, а также на некоторые исследования с применением контент-анализа.

Статья не содержит положений, связанных с дискуссионными моментами в силу ее специфического методического содержания.

Хотя представленное исследования посвящено достаточно частным методическим вопросам, оно вносит свою лепту в сложную и крайне важную проблему распознавания исторических текстов. Статья не содержит выраженных недостатков, она полностью соответствует формату журнала и, безусловно, найдет своего читателя, поэтому она рекомендуется к публикации.