Фунтов А. Ф. Из истории организации и деятельности Шатровского союза коммун в 1923–1925 гг. // Вопр. ист. Урала. Вып. 4. Свердловск: Урал. гос. ун-т, 1963. С. 103–122.

Фунтов А. Ф. Союз коммун Шатровского района Тюменского округа Уральской области в подготовке и проведении сплошной коллективизации // Вопр. агр. ист. Урала и Зап. Сибири. Свердловск, 1966. С. 98–116.

Центр документации общественных организаций Свердловской области (ЦДООСО). Ф. 4. Свердловский областной комитет КПСС.

УДК 930.22+004.912

И. Р. Соколовский
Институт истории Сибирского
отделения РАН, г. Новосибирск

РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ ЯЗЫКА PYTHON КАК СРЕДСТВО РАБОТЫ СО СЛАБОСТРУКТУРИРОВАННЫМИ ДАННЫМИ В ИСТОРИИ

В работе рассматривается неуклонный рост числа наборных копий исторических источников как причина необходимости разработать инструменты их автоматической обработки. Одним из таких инструментов являются регулярные выражения на языке Python. На примере ценного источника XVII в. показано, как можно получить информацию об источнике путем обработки указателя персоналий, составленного публикатором. Схожие принципы применимы к любому цифровому источнику.

Ключевые слова: Python, регулярные выражения, наборная копия, факсимильная копия, указатель, копийная книга, XVII в., Сибирь.

Одно из заявленных направлений настоящей конференции звучит как «Digital History: будущее исторической науки?». Нам кажется, что возможные ответы на этот вопрос можно найти как в самой истории исторической науки, так и в конкретных практиках историков в наши дни. В этой статье мы используем оба подхода, так как основным направлением наших исследований является попытка создания базы данных из слабоструктурированных сведений о служилых людях Сибири в XVII в. с помощью скриптов на языке программирования Python; в том числе, с помощью применения регулярных выражений – модуля этого языка программирования.

Известно, что историография как социальная практика не может оставаться неизменной и игнорировать все те социальные и техниче-

ские изменения, которые происходили в истории. Все отлично помнят, как технические изменения влияли на историографию прошлого.

Когда в первой четверти XV в. было изобретено книгопечатание, все бросились публиковать наборные копии исторических трудов, а затем исторических источников, так, что наборная публикация источников предшествовала факсимильной, теорию которой известный аббат Мабильон сформулировал только в конце XVII века.

> «Почти все хроники X–XV вв. изданы (...) Из огромного количества документов XVI–XVII вв. введено в научный оборот лишь небольшое число, прочая же масса еще ждет своих исследователей. Повествовательные источники этого периода в подавляющем числе были тогда же и напечатаны» [Люблинская, 1955].

Следующим революционным шагом стало появление в последней четверти XIX в. печатных машинок, которые, наравне с гектографией, позволили легко печатать учебные курсы, открывая историографию для работ университетских профессоров, таких, например, как В. О. Ключевский и М. К. Любавский (курс исторической географии 1897 г. [Петрова, 2013]) и др.

Совершенно очевидно, что появление электронно-вычислительных машин совершило ровно такую же революцию, как и предыдущие, принеся с собой новые способы исследовательской работы, тиражирования трудов историков и публикаций исторических источников. На первом этапе ЭВМ использовалась прежде всего для обработки источников с целью получения новой информации. В частности, пионерами в этой области выступили новосибирские ученые [Материалы переписи..., 1969].

С появлением персональных ЭВМ речь пошла не только о способе простого тиражирования, кстати, еще более эффективном, чем два предыдущие, но и о целом новом наборе способов обработки информации, теперь доступных буквально каждому историку, что ранее было немыслимым. [Ассоциация...; Баринова, 2024]. Можно сказать, что сейчас отдельный историк обладает таким набором вычислительных и сортировочных возможностей, которые сорок лет назад были доступны только крупным организациям.

Цифровизация издательского процесса привела к тому, что мы имеем цифровые копии для каждой книги, опубликованной после 2000 г. Множится число цифровых копий для книг и публикаций документов, которые вышли из печати до 2000 г. Такое большое число цифровых документов, доступных историку, требует перестройки некоторых приемов и методик исторических исследований.

Небольшая статья – не место для перечисления всех новых возможностей, поэтому остановимся только на одной. Это базы данных, которые являются, например, единственной формой обработки исторических данных, для которой можно четко указать количественные показатели эвристической ценности. Иначе говоря, для баз данных эта ценность счетна и исчисляется довольно легко. Поэтому базы данных являются одной из продуктивных форм обработки просопографической информации, особенно когда нам надо охарактеризовать какую-либо социальную группу. Или любую другую совокупность.

Перевод в машиночитаемую форму может осуществляться двояким образом. Во-первых, речь может идти о «машиночитаемых наборных копиях» исторических документов. И, во-вторых, о «машиночитаемых факсимильных копиях» исторических документов. В чем, собственно разница?

Возьмем какой-нибудь исторический документ, скажем, челобитную XVII в., которую необходимо перенести на электронные носители для создания копии. Тогда у нас есть два пути. Мы можем разбить текст челобитной на отдельные знаки, и по знакам «перенести ее в компьютер». Тогда каждому знаку письменного документа XVII в. можно будет сопоставить знак электронной копии. Поскольку в компьютере знаки обычно хранятся в виде нолей и единиц (в двоичной, восьмеричной, шестнадцатеричной и т. д. форме), то каждому знаку, который мы видим на экране или на печати, последовательность нолей и единиц сопоставляется согласно кодовой таблице (ANSCII, Utf-8, Utf-16 и т. д.).

Другой способ создания машиночитаемой копии – это аналог фотографирования. Вся поверхность исторического документа разбивается на небольшие точки, которых может быть около 300 на каждые 2,54 см линейных размеров документа (или больше, или меньше). И эти точки переносятся в электронную форму путем сканирования, цифрового фотографирования или иным способом. Таким образом, в компьютере мы получаем факсимильный образ документа, который отражает не только вид букв, из которых он составлен, но и пятна на бумаге, утраты и проч. Ровно так, как предлагали издавать документы в XVII в. и как иногда издают еще особо ценные памятники, например, летописные.

Почему необходимо так тщательно обсуждать все эти аспекты? Потому что дальнейшая машинная обработка копии исторического источника зависит от ее вида. В настоящий момент у нас нет хороших цифровых инструментов для машинной обработки факсимильных копий. Возможно, прогресс нейросетей в распознавании образов внесет

в эту область какие-то изменения. Но пока мы можем хорошо обрабатывать только наборные копии.

Оба вида электронных копий хранятся как нечто намагниченное, наэлектризованное, как измененное состояние полупроводника или иначе, но обязательно с использованием электрического тока. Поэтому строчные и прописные буквы, латинские буквы, а так же цифры могут быть объединены в однородные классы (на уровне синтаксиса языка программирования высокого уровня) или заданы простым перечислением с целью обработки их как электрических сигналов.

Это позволяет нам использовать при чтении электронной копии текста т. н. «регулярные выражения». Речь идет о своего рода языке, который, как мы уже сказали выше, позволяет описывать буквы текста по определенному шаблону. Мы можем искать отдельное слово, например, «челом». Можем искать все слова, которые начинаются с буквы «ч», или имеют такую-то длину (или больше, или меньше этой длины), можем искать все слова, которые начинаются с заглавных букв и проч. Данный инструмент был создан программистами для проверки криптостойкости паролей или для быстрого анализа логов и других задач.

Примером регулярных выражений будет «два слова с прописной буквы, идущие подряд и разделенные только пробелом» (1), «четыре цифры, идущие подряд» (2) и т. д. Очевидно, что с помощью регулярного выражения (1) мы можем обнаружить всех людей с именем и фамилией, упомянутых в тексте документа («Иван Иванов», «Петр Петров» и т. д.). К сожалению, в нашу выборку попадет часть географических названий: «Западная Сибирь» или «Северная Евразия». С помощью выражения (2) в поле нашего внимания попадут все даты после 1000 г., встречающиеся в тексте, но так же и любые четырехзначные числа, например, «2359» или «4456», которые придется исключать либо перечислением, либо заданием более точного шаблона.

В чем же отличие использования поиска через регулярные выражения от обычного чтения? Именно в возможности проделывать это относительно быстро, для большого массива текстов и с возможностью сохранить полученные результаты в отдельном месте. С помощью скрипта и регулярных выражений можно будет определить, имеет ли диссертация заявленные хронологические рамки «конец XVII – начало XVIII в.», если 80 % дат в тексте лежат после 1701 года. Так же можно исследовать глубину историко-юридической памяти.

Поскольку любой текст имеет какую-то структуру, то он может быть таким образом проанализирован. Как мы уже говорили выше, нашей задачей является такой анализ текстов источников XVII в. (окладных, таможенных и др. книг, например).

Так как большинство исторических данных, особенно до XVIII в., предстают в слабоструктурированном, а не в жестко структурированном виде вроде таблиц, пригодных для включения в базы данных, мы можем совершить перевод из слабоструктруированного вида в формализованный табличный вид без больших затрат труда и времени и без потери важной информации, если мы будем применять скрипты на языке Python к наборным копиям документов XVII в.

Метод обработки текстов с помощью регулярных выражений хорош еще и тем, что в XVII в. не было строгих орфографических правил, в том числе и правил для передачи имен собственных. Поэтому мы легко можем создавать списки вариантов, в том числе для человеческого анализа или машинной обработки, тем более, что функции модуля регулярных выражений принимают переменные точно так же, как и всякие другие функции этого языка.

Вопросы различного написания имен и фамилий, как правило разрешены в указателе к уже опубликованному источнику (своего рода базы данных персоналий текста). Подобные задачи нам, например, предстоит решить для справочника по корпусу переводчиков Посольского приказа. Но пока воспользуемся другим примером, «копийной книгой», тщательно изданной екатеринбургскими коллегами [Копийная книга..., 2014].

Алфавитный указатель (база данных персоналий) может быть весьма объемным. В опубликованной А. В. Полетаевым копийной книге о ссыльных в Сибирь в 1614–1624 гг. указатель имеет 1 909 строк. В нем 101 598 знаков и около 11 041 слов. Ручной подсчет в таком массиве был бы трудоемким и ненадежным. Однако, если мы применим регулярные выражения, то легко получим базу из 1 789 уникальных персоналий, выявленных в публикуемых документах, причем список этот будет сохранен в отдельное место (переменную). Для 1 720 персоналий мы можем отследить упоминание социального статуса, поскольку он чаще всего отделен двумя запятыми.

Однако мы знаем, что среди этих персоналий есть не только деятели прошлого, но и историки (даже некоторые живущие ныне, например, ваш покорный слуга). Как отделить их от исторических персоналий? Очень просто. Историки записаны по схеме: фамилия и два инициала. Фамилия начинается с заглавной буквы русского алфавита и отделяется от инициалов пробелом. Инициалы записываются с помощью одной заглавной буквы русского алфавита, с последующей точкой. Данной информации нам вполне достаточно для того, чтобы с помощью регулярного выражения выявить в тексте указателя 70 коллег.

Дальнейший анализ показывает, что 160 упоминаемых в работе персоналий были «воеводами», которых упоминали в разных документах, причем, в это число включены и «полковые воеводы». Ссыльными были 519 чел., что показывает информационную ценность источника. Перед нами не просто сухой перечень ссыльных, который сам по себе не имел бы ценности, но и данные по управленческой деятельности, и сведения еще о почти 1 100 других людях, которые жили в тот период и так или иначе соприкасались со ссыльными. Кроме того, видно, что автор собрал информацию почти во всех опубликованных работах сибиреводов (кажется, что список из 70 персоналий охватывает почти всех историков. кто когда либо занимался первой четвертью XVII в. на русском языке и по отечественным источникам).

Кроме того, изучение указателя с помощью регулярных выражений открывает возможности и для филологических исследований. Например, словарь социальных терминов включает в себя 341 нелемматизированную словоформу (4 330 знаков). Мы пропустили этот список через бесплатный лемматизатор (https://arsenkin.ru/tools/lemma/) и получили 294 лемматизированных словоформ. Данный список может быть основой для какого-то исследования русского языка, основанного на данном источнике.

Таким образом, мы видим, что регулярные выражения в языке Python – это простой, гибкий, легко настраиваемый инструмент, который позволяет нам возвращать из слабоструктруированного текста любой объем необходимой нам информации, который может быть напрямую использован для понимания истории или, переведенный в табличную форму, может послужить для дальнейшего анализа средствами систем управления базами данных.

Ассоциация «История и компьютер». URL: https://aik-hisc.ru/ (дата обращения: 24.11.2024).

Баринова Е. П. Цифровая история (digital history) : уч. пособие. Самара: Самар. ун-т, 2024. URL: https://repo.ssau.ru/bitstream/Uchebnye-izdaniya/Cifrovaya-istoriya-digital-history-110030/1/978-5-7883-2046-5_2024.pdf (дата обращения: 24.11.2024).

Копийная книга «о опалных людех», сосланных в Сибирь в 1614–1624 гг. / подг. текста, вступит. ст. и комм. А. В. Полетаева. Екатеринбург ; Верхотурье: Изд-во Верхотур. гос. ист.-арх. музея-заповедника, 2014.

Люблинская А. Д. Источниковедение истории средних веков Л.: Изд. ЛГУ, 1955.

Материалы переписи 1916 года по Томской губернии (из опыта обработки на ЭВМ) / А. М.Бауфал, Л. М. Горюшкин, В. С. Золототрубов, И. В. Островский, А. М. Рябоконев. Новосибирск: Наука, 1969.

Петрова О. С. Историческая география России как учебная дисциплина в XIX – начале XX вв. // Genesis: исторические исследования. 2013. № 1. С. 30–49. URL: https://nbpublish.com/library_read_article.php?id=617 (дата обращения: 24.11.2024).