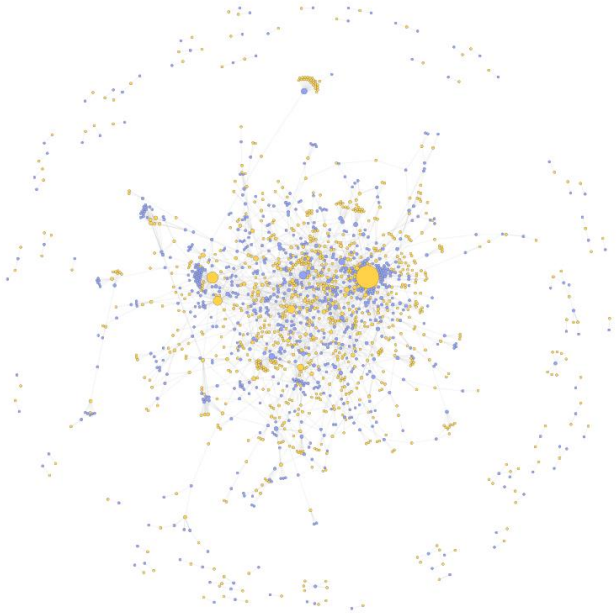
А.А. Пучковская, Л.В. Зимина, Д.А. Волков

ВВЕДЕНИЕ В ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ



Санкт-Петербург
2021

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

А.А. Пучковская, Л.В. Зимина, Д.А. Волков

ВВЕДЕНИЕ В ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ

УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлению подготовки 45.04.04 Интеллектуальные системы в гуманитарной среде
в качестве учебно-методического пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры

Санкт-Петербург
2021

Пучковская А.А., Зимина Л.В., Волков Д.А., Введение в цифровые гуманитарные исследования– СПб: Университет ИТМО, 2021. – 61 с.

Рецензент(ы):

Надежда Георгиевна Поврозник, к.и.н., отсутствует, доцент кафедры междисциплинарных исторических исследований, Пермский государственный национальный исследовательский университет;

Цель учебного пособия “Введение в Digital Humanities” — положить начало погружению студентов магистратуры в предметную область цифровых гуманитарных наук. Пособие состоит из четырех разделов, включающих теоретические материалы, тестовые задания на выявление знаний студента и закрепление материала, некоторые разделы снабжены практическими заданиями для получения студентами необходимых навыков по работе с полезными инструментами для проведения цифровых гуманитарных исследований. Вопросы, предложенные в конце разделов призваны не только проверить внимательность студентов, но и способствуют более глубокому пониманию предмета и формированию критического отношения к данным и методом, применяемым для их анализа.

Университет ИТМО – национальный исследовательский университет, ведущий вуз России в области информационных, фотонных и биохимических технологий. Альма-матер победителей международных соревнований по программированию – ICPC (единственный в мире семикратный чемпион), Google Code Jam, Facebook Hacker Cup, Яндекс.Алгоритм, Russian Code Cup, Topcoder Open и др. Приоритетные направления: IT, фотоника, робототехника, квантовые коммуникации, трансляционная медицина, Life Sciences, Art&Science, Science Communication. Входит в ТОП-100 по направлению «Автоматизация и управление» Шанхайского предметного рейтинга (ARWU) и занимает 74 место в мире в британском предметном рейтинге QS по компьютерным наукам (Computer Science and Information Systems). С 2013 по 2020 гг. – лидер Проекта 5–100.

Введение

У молодого исследователя, начинающего или только планирующего применять цифровые методы в своих научных изысканиях неизбежно возникает ряд непростых вопросов. Как сделать так, чтобы гуманитарное знание во всем своем многообразии и многоаспектности превратить в 1 и 0, понятные для компьютерных алгоритмов, и при этом оно не потеряло в разнообразии смыслов и интерпретации? Как сделать огромный массив гуманитарных данных доступным для широкой публики, прокладывая путь через тернии к звездам, а не, наоборот, запутывая пользователя? Есть ли, и если да, то каковы лимиты применения информационно-коммуникационных технологий к гуманитарному знанию? Этими и другими вопросами так или иначе задаются все ученые, кто решил попробовать свои силы на стыке гуманитарного знания и компьютерных технологий. Ответы на эти и другие вопросы можно найти в данном учебно-методическом пособии.

Цель пособия “Введение в Digital Humanities” — положить начало погружению студентов магистратуры в предметную область цифровых гуманитарных наук. Четыре раздела включают теоретические материалы, тестовые задания на выявление знаний студента и закрепление материала, некоторые из них снабжены практическими заданиями для получения студентами необходимых навыков по работе с полезными инструментами для проведения цифровых гуманитарных исследований. Вопросы, предложенные в конце разделов, призваны не только проверить внимательность студентов, но и способствуют более глубокому пониманию предмета и формированию критического отношения к данным и методом, применяемым для их анализа.

Каждый раздел посвящён одной из центральных тем, необходимых для дальнейшей исследовательской работы в области Digital Humanities. В первом вводном разделе уделяется внимание термину Digital Humanities, его определению, а также истории формирования исследовательского направления и сферам, которые могут быть в нем выделены. Второй раздел посвящен гуманитарным данным — их типологии, основным методам и принципам работы с ними, практические задания, приведённые в этом разделе, пошагово описывают процесс сбора данных и их подготовки к анализу. В третьем разделе приведена классификация DH проектов с примерами для формирования у студентов насмотренности, дан план работы над собственным DH проектом, а также предложен фреймворк для критической оценки DH проектов. В заключительной части перечислены основные методы, которыми студенты могут пользоваться для создания собственного исследования на стыке цифровых и гуманитарных наук.

Список дополнительной литературы и интернет-ресурсы:

1. Schreibman S., Siemens R., Unsworth J. (ed.). A companion to digital humanities. – John Wiley & Sons, 2008.[Электронный ресурс], способ доступа: http://www.digitalhumanities.org/companion/
2. Gold M. K. (ed.). Debates in the digital humanities. – U of Minnesota Press, 2021.[Электронный ресурс], способ доступа: https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities
3. Найхан Д., Ванхут Э., КИЖНЕР И. Цифровые гуманитарные науки Хрестоматия. – Siberian Federal University Press, 2017.[Электронный ресурс], способ доступа: https://www.pure.ed.ac.uk/ws/portalfiles/portal/46320044/Terraces_i_531505996.pdf
4. Thaller M. Controversies around the digital humanities: an agenda //Historical Social Research/Historische Sozialforschung. – 2012. – С. 7-23. [Электронный ресурс], способ доступа: https://www.jstor.org/stable/41636594?seq=1
5. Манифест Digital Humanities, способ доступа: https://tcp.hypotheses.org/501

Глава 1. Введение в Digital Humanities

К определению Digital Humanities

Говоря о междисциплинарном взаимодействии гуманитарных и компьютерных наук, используют термины Humanities Computing, eHumanities, computational linguistic, но чаще — Digital Humanities (DH). На русский термин переводят как цифровая гуманитаристика или гуманитарная информатика.

Существует множество точек зрения относительно определения Digital Humanities. Из многообразия определений можно извлечь не только различные мнения о том, чем занимаются цифровые гуманитарные науки и как донести их значимость до научного сообщества и широкой общественности, но и сложить представление о злободневных вопросах, которые занимают исследователей, и актуальных дискуссиях, спровоцированных этими вопросами. Так, в 2003 г. Уиллард Маккарти высказал мнение, что цифровым гуманитариям не нужно давать четкого ответа на вопрос «Что такое DH?», так как живая дискуссия важнее хрестоматийного определения[1].

[1] McCarty, W. (2003). “Humanities Computing”, Encyclopedia of Library and Information Science, New York: Marcel Dekker, p. 1224–1235.

С 2009 г. проект «День DH» предоставляет DH-сообществу ежегодную массовую исследовательскую площадку для такой дискуссии. Каждый год участникам — исследователям из разных стран и институций, кто так или иначе считает себя частью DH-сообщества — предлагается ответить на вопрос: «Что такое DH?». Разнообразие предложенных определений примечательно как по форме, так и по содержанию. Одни считают, что ключевая особенность DH — использование компьютеров; для других самое важное — это помощь в осмыслении практик применения новых технологий в гуманитарных науках. Для третьих DH – это в первую очередь проекты. Для четвертых главное в DH — это данные, их анализ и организация. Пятые считают, что в DH важнее всего критическое отношение к данным и новым медиа. По мнению шестых, DH — это прежде всего традиционные объекты исследования, такие как тексты и изображения, рассмотренные через оптику новых технологий. Седьмые убеждены, что за DH будущее традиционных гуманитарных наук, хотя есть и обратная точка зрения — что между “цифровыми” и “традиционным” гуманитарными науками нет существенной разницы и системный кризис гуманитарных наук может быть преодолен только возвратом к классическим методам. Наконец, некоторые сойдутся на возможности нечеткого и изменчивого определения, а кто-то вообще откажется его давать.

Приведем в пример несколько определений, данных различными учеными в различные дни цифровых гуманитарных наук в период с 2009 года.

“Цель DH — наладить связи между цифровыми технологиями и гуманитарными науками с их богатой историей. Это в то же время прагматический и философский посыл к созиданию и рефлексии. Это пристанище практиков, разбирающихся в Java не меньше, чем в постструктурализме; их в равной степени интересует iPhone и «Моби Дик»; они столь же оптимистичны по отношению к будущему, сколько скептичны к концепции постчеловека. В настоящее время DH – одна из самых поразительных отраслей гуманитарных наук – с растущим сообществом энтузиастов как студентов, так и профессоров”.

Стивен Рамси, Университет Небраски, США

“Так же, как и человек, согласно Ницше, «...есть животное, чей вид до сих пор не определен», DH есть вид гуманитарных наук, до сих пор не имеющий определения. Нам пока не по плечу полностью оценить характер влияния DH-институтов на будущее науки. Мы можем только предполагать, что последует за возрастанием доли цифровых публикаций в среде научных публикаций. Мы все еще не до конца проанализировали последствия применения информационных методов к таким «традиционным» отраслям знания, как история, литература,

театроведение. Для меня DH – это фундаментальная игровая площадка для экспериментов”.

Тома Кромбе, Университет Антверпена, Бельгия

“Для меня цифровые гуманитарные науки – это гуманитарные науки через призму доступных цифровых возможностей. Большинство моих программ и работ в меньшей степени посвящено определению новых цифровых парадигм гуманитарного дискурса. В основном они призваны «искоренить скучные и однообразные исследовательские процедуры, чтобы моя жизнь стала проще».

Тара Л. Эндрюс, Оксфордский университет, Великобритания

“Понятие, используемое для тактического преимущества”.

Мэтью Киршенбаум, Университет Мериленда, США

“Хотя я одобряю все то, что происходит под крылом DH, но сам термин меня не устраивает. Он кажется мне своего рода политическим ходом внутри сферы, подчеркивающим способность некоторых технологий и методик решать важнейшие гуманитарные эпистемологические вопросы. Некоторые мои коллеги, вероятно, с большой осторожностью прибегают к цифровым технологиям в преподавании, исследованиях, публикациях. Потому, мне кажется, DH неслучайно существенно отличаются от консервативных гуманитарных наук, чуждающихся всего незнакомого. В DH рассматривают постепенное проникновение технологий в науку. Совсем не нужно менять парадигму, чтобы напомнить всем, что цель науки – познание неизвестного любыми доступными средствами. Цифровой компонент термина не имеет никакого значения: нам нужно лишь подумать, как приспособить технологии к науке, к любому исследовательскому процессу. Очень полезная и важная наука, только название несколько неуместное. Едва ли я его вообще использую по отношению к моим трудам и к трудам моих коллег”.

Кимон Керамидас, Нью-Йоркский Университет

Итак, попытаемся и мы дать свое определение Digital Humanities, максимально широко охарактеризовав его как собирательный термин, обозначающий область исследования и разработок на стыке компьютерных и гуманитарных наук, объединяя исследовательские вопросы и методы традиционных гуманитарных наук (лингвистики, истории, искусствоведения, философии, культурологии и т.д.) с компьютерными технологиями.

Отметим, что даже если череда попыток выстроить (и истолковать) определения прекратится, то все же есть смысл проверять пульс DH-сообщества время от времени, чтобы понять текущие и будущие практические тенденции, определяющие облик дисциплины. DH-сообщество непрестанно борется за академический статус, налагающий определенные дисциплинарные ограничения в настоящее время.

Вопрос к обсуждению: *Какое определение цифровых гуманитарных наук из предложенных в разделе кажется вам наиболее близким по духу? Объясните свой выбор.*

Тест 1.

1. Как можно перевести Digital Humanities на русский язык? (возможно несколько ответов)
   - A. цифровая гуманитаристика
   - B. цифровые гуманитарные проекты
   - C. цифровые гуманитарные науки
2. Какой из предложенных ниже терминов никогда не применялся относительно DH и смежных областей?
   - A. eHumanities
   - B. Компьютерная лингвистика
   - C. Humanitarians Computing
3. Когда прошёл первый день DH?
   - A. 2003
   - B. 2009
   - C. 2018

История становления Digital Humanities

Как появилась данная область и что стало предпосылками для формирования ее научного дискурса?

Приступая к раскрытию вопроса истории развития предметной области цифровых гуманитарных наук, обратимся к принятой периодизации. Первый этап (1949 — 1970-е) характеризуется использованием более раннего термина Humanities Computing. Второй (1970-е — середина 1980-х гг.) примечателен появлением специализированных компьютерных программ для отдельных прикладных задач гуманитарных исследований. Третий этап (середина 1980-х — начало 1990-х гг.) отмечен появлением первого общедоступного интернет-браузера и развитием систем хранения гуманитарных данных, и наконец четвертый этап (1990-е гг. — по сей день), когда цифровые технологии стали наиболее широко применяться в гуманитарных исследованиях и получили широкую огласку. Как можно заметить из наиболее кратких описаний отмеченных периодов, данная периодизация связана, прежде

всего, с генеалогией развития информационно-коммуникационных технологий. В этой связи можно проследить тенденцию, что нововведения в DH соответствуют достижениям в точных науках и компьютерной инженерии соответствующих периодов. Давайте более подробно рассмотрим каждый период в отдельности.

1 этап (1949 г. — ок. 1970 г.)

Британская исследовательница Сьюзен Хокки в своей работе по истории цифровых гуманитарных исследований «The History of Humanities Computing» отсчитывает историю DH с 1949 года[2]. Согласно Хокки, в именнно в этом году итальянский иезуитский священник Роберто Буза поставил перед собой задачу разработать систему индексации для полного собрания сочинений Фомы Аквинского и комментариев к ним.

Итак, отец Буза сделал предположение, что именно новые технологии смогут помочь ему в реализации такой задачи. Тогда он обратился за помощью к основателю и главному исполнительному директору IBM Томасу Уотсону и смог его убедить проспонсировать издание Index Thomisticus и оказать поддержку в дальнейших исследованиях. Получив от IBM необходимую поддержку, отец Буза продолжал работать над усовершенствованием своего проекта до конца своей жизни в 2011 г. Получившаяся база данных включала в себя 11 миллионов слов на средневековой латыни, а его проект Index Thomisticus Treebank функционирует и по сей день[3].

Долорес Бертон — другая исследовательница истории DH — также отмечает, что именно в этот период начинаются эксперименты по использованию возможностей компьютера в европейской академической среде лингвистов, где компьютерные технологии главным образом использовались для создания и объединения словарей[4].

Позже, в 1960-х гг., компьютерные технологии и вычислительные методы находят применение в гуманитарных исследованиях с целью установления авторства произведений. Авторы книги «A Companion to Digital Humanities»[5] указывают, что первое использование компьютерных технологий для разрешения споров об авторстве приписывается Алвару Элигарду, шведскому лингвисту и пионеру DH. В 1962 г. исследователь

---

[2] Hockey S. The History of Humanities Computing. In A Companion to Digital Humanities / Susan Hockey // Mode of access: http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1

[3] Index Thomisticus Treebank. URL: https://itreebank.marginalia.it/.

[4] **Burton, D. M. Automated Concordances and Word Indexes: The Fifties. Computers and the Humanities 15. — 1981.**

[5] Schreibman S., Siemens R., Unsworth J. A. Companion to Digital Humanities. / Susan Schreibman, Ray Siemens, John Unsworth // http://www.digitalhumanities.org/companion/

использовал вычислительные методы для анализа “Писем Юния”[6] за 1769–1772 гг.

Одно из наиболее известных исследований авторства с применением вычислительных методов было проделано математиками Ф. Мостеллером и Д. Л. Уолессом и также относится к 1962 году[7]. В основе их исследования было изучение федеральных бумаг с целью установки авторства 12 спорных документов, опубликованных под собирательным псевдонимом “Публий”. Авторы исследования изобрели метод, позволяющий выявить индивидуальную закономерность подбора слов в тексте, который лежит в основе большинства современных компьютерных алгоритмов по установлению авторства.

В целом, развитие DH в 1950-е — середине 1970-х можно обозначить как период методологической консолидации и институализации, появления первых профессиональных сообществ. В 1962 году была основана “Ассоциация машинного перевода и компьютерной лингвистики”, позже сократившая своё название до “Ассоциации компьютерной лингвистики”(ACL), проводившая ежегодные встречи и положившая начало формированию международного профессионального сообщества для ученых, применяющих компьютерные методы в лингвистике. Стоит отметить, что журнал Ассоциации является одим из первых специализированных изданий для цифровых гуманитариев, с 1988 года он издается под редакцией MIT Press.

В 1966 г. был основан ещё один журнал, посвященный DH — Computers and the Humanities. Это издание стало основой для дальнейшего возникновения конференций и ассоциаций, объединяющих исследователей, практикующих цифровые методы в гуманитарных науках, что послужило дальнейшей консолидации сообщества и формированию новой дисциплины.

2 этап

Второй этап датируется 1970-ми — серединой 1980-х. Этот этап стал периодом радикальных изменений, связанных с появлением прикладных программ для решения задач гуманитарных исследований. Теперь исследователи, применяющие цифровые методы для анализа гуманитарных данных, могли тратить сравнительно немного времени на изучение программных языков и пакетов программ, что позволяло реализовывать свои проекты гораздо продуктивнее, без привлечения дополнительного персонала. Так, в 1968 г. вышла первая версия пакета SPSS (*«Statistical Package for the Social Sciences»*) — это компьютерная программа для

[6] «Письма Юния» — собрание частных и открытых писем с критикой правительства британского короля Георга III от анонимного полемиста, под псевдонимом Юний.

[7] Mosteller, F. and Wallace D. L. Inference and Disputed Authorship: The Federalist. / Mosteller, F. and Wallace D. L. // Reading, MA, 1964: Addison-Wesley.

статистической обработки данных, предназначенная для проведения прикладных исследований в общественных науках. Это ПО развивалось в Чикагском университете. Первое руководство по использованию данной программы вышло в 1970 г.

Также в 1970-х в Великобритании была запущена серия первых регулярных конференций по тематике DH. Ежегодная конференция Literary and Linguistic Computing была основана Роем Висби и Майклом Феррингдоном в Кембриджском университете в марте 1970 года[8]. Программа конференции демонстрировала заинтересованность исследователей в использовании компьютерных технологий для работы с гуманитарными данными, в частности числе стиломтерии - количественном исследовании стилистики различных текстов с целью их атрибуции и датировки.

Следующим важным пунктом в истории развития DH стал 1973 год, когда была основана Association for Literary and Linguistic Computing (ALLC). Ассоциация с первого года своего существования запускает собственное периодическое издание, которое стало выходить трижды в год.

В США в середине 1970-х была основана ещё одна серия конференций под названием International Conference on Computing in the Humanities (ICCH), положившая начало Association for Computers and the Humanities (ACH). ICCH и ACH стали катализатором для появления широкого спектра исследований на стыке гуманитарных и технических наук. Вычислительные методы стали применяться в археологических исследованиях, анализе музыки и визуального искусства. В ретроспективе этот этап интересен тем, что заложил основы и методологию для квантитативных исследований в области гуманитарного знания.

3 этап

Третий этап датируется второй половиной 1980-х и началом 1990-х гг.. Одним из ключевых изобретений, повлиявшим не только на развитие DH, но и на нашу жизнь в целом, является появление первого веб-браузера — WorldWideWeb, который был разработан Тимом Бернерсом-Ли в рамках проекта по ядерным исследованием. Помимо всего прочего, создание и развитие всемирной паутины способствовало развитию горизонтальных связей и консолидации международного исследовательского сообщества, значительно ускорив коммуникацию исследователей между собой. Другой технологической доминантой этого периода стало широкое применение базы данных, решивших проблему формализации хранения больших объёмов гуманитарных данных.

---

[8] Hockey S. The History of Humanities Computing. In A Companion to Digital Humanities / Susan Hockey // Mode of access: http:// www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1

Таким образом, отличительной особенностью данного период является значительная консолидация ведущих DH институций. В 1987 году было организовано академическое сообщество Text Encoding Initiative (TEI), которое до настоящего времени поддерживает одноименный технический стандарт разметки текста. Инициатива была создана тремя уже упомянутыми институциями — Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL) и the Association for Literary and Linguistic Computing (ALLC).

К середине 1980-х отмечается рост заинтересованности в Digital Humanities во всем мире, совершаются значительные финансовые вложения в развитие DH проектов и научной области в целом. В результате появляются новые исследовательские центры при университетах Западной Европы и Северной Америки, ассоциации. Этот период также становится временем введения новых дисциплин и курсов по Digital Humanities и Humanities Computing в учебный процесс как для гуманитариев, желающих исследовать культурные феномены с помощью цифровых технологий, так и для студентов технических направлений, интересующихся возможностями применеия своих навыков в исследованиях гуманитарной направленности. Данный процесс дал начало дискуссиям о необходимости обучения студентов-гуманитариев основам программирования — так называемого minimal computing. Одни считали, что оно заменит в современной науке латынь, другие — что такие радикальные перемены в учебных планах нецелесообразны, так как изучение программирования может оказаться слишком сложным для гуманитариев и потребует много времени в ущерб собственно гуманитарным исследованиям.

4 Этап. Современность

Четвертый этап, который продолжается по сей день, представляет собой бурный расцвет DH и проектов в области цифровых гуманитарных исследований. Благодаря развитию OCR (Optical Character Recognition) - технологии оптического распознавания символов, позволяющей преобразовывать различные типы документов, например, отсканированных текстов или других изображений, в редактируемые форматы с возможностью поиска, стало возможно исследовать старинные рукописи и тексты при помощи машинного обучения.

В рамках современного периода стоит отметить другую, немаловажную инициативу, которая способствовала развитию DH в западном академическом коммьюнити - IIIF (The International Image Interoperability Framework) - стандартизированный метод описания изображений, который позволил упростить жизнь библиотекарям и спровоцировал бум DH проектов в области цифровых экспозиций.

Еще одной особенностью данного периода, как и области DH в целом, является разработка и развитие всевозможных открытых ПО, которые представляют собой готовые технические решения для создания DH проектов. Среди наиболее востребованных и открытых ПО следует отметить платформу OMEKA, с помощью которой можно создать цифровую коллекцию или экспозицию на готовых шаблонах, которые при необходимости могут быть изменены пользователем посредством кастомизации CSS и HTML файлов. Данную платформу широко используют многие культурные институции и организации, в частности, библиотеки (DIY History, Virginia Tech Special Collections Online, Isaiah Thomas Broadside Ballads и др.), архивы (Florida Memory, Digital History Archive of São Roque, Saint John's College Digital Archives и др.) и музеи (John J. Audubon's Birds of America, Marshall M. Fredericks Sculpture Museum и др.), а также небольшие исследовательские коллективы для визуализации собранных датасетов (From Farms to Freeways: Women's Memories of Western Sydney, New Roots: Voices from Carolina del Norte, Intemperance Archive и др.).

Стоит также упомянуть следующие платформы, предоставляющие готовые решения для исследователей в области DH: Collective Access, основной особенностью которой является удобная и простая в использовании цифровая каталогизация и хранение метаданных; CollectionSpace, которая в отличие от вышеперечисленных не имеет функционал по созданию цифровых коллекций, но прекрасно справляется с созданием кастомизированного словаря для описаний объектов культурного наследия и совместима с другими приложениями; Open Exhibits, схожую по своему функционалу с OMEKA и рассчитанную на пользователей как с опытом программирования, так и с полным его отсутствием и др. Среди примеров отечественного программного обеспечения, разработанного для работы с оцифрованными материалами, следует отметить платформу “Коллекция онлайн” — информационную систему на базе данных КАМИС, которую широко используют для виртуального экспонирования коллекций на сайтах российских музеев (Государственный исторический музей, Государственный Эрмитаж и др.).

Тест 2

Вопрос 1. В какую компанию обратился отец Роберто Буза за помощью в исследовании текстов Фомы Аквинского?

A. IBM
B. Microsoft
C. Apple

Вопрос 2. Ассоциации и журналы какой направленности стали первыми объединять ученых, осуществлявших исследования на стыке цифровых и гуманитарных наук?

A. история
B. лингвистика
C. культурология
D. философия

Вопрос 3. Что такое TEI?

A. ежегодная конференция, посвященная технологиям в исследовании англоязычных текстов

B. программное обеспечение для создания цифровых коллекций

C. консорциум, работающий над стандартом разметки текста

Вопрос 4. Какой этап характеризуется созданием готовых платформенных программных решений для DH, не требующих навыков программирования для работы с ними?

A. 1
B. 2
C. 3
D. 4

Области исследования цифровых гуманитарных наук

Итак, как было отмечено ранее, DH изначально базировалось на двух основаниях: с одной стороны, была твердая убежденность в том, что существуют более легкие автоматизированные способы обработки больших объемов информации; с другой стороны, сложившееся практика показывала, что различные методологические традиции гуманитарных и естественных наук могут и должны сближаться. Итак, какие основные области могут быть выделены в рамках цифровых гуманитарных наук?

Мы приводим разделение, предложенное немецким исследователем М. Таллером в его лекции “Дискуссии вокруг Digital Humanities”, который предлагает разделять Digital Humanities на четыре основных области или парадигмы.

1. Область, ориентированная на анализ текста как такового

В этой области можно выделить по меньшей мере три направления исследований. Во-первых, это компьютерная лингвистика. Компьютерные лингвисты в своих исследованиях предпринимают попытки моделировать формальные правила, по которым живет язык и строится живая речь, или анализируют лингвистические явления в корпусной лингвистике. Во-вторых, в литературоведении применяются различные цифровые средства

анализа: от тривиального индексирования для построения словаря отдельно взятого автора до более амбициозных, например решения задачи формализованного определения авторского стиля или стиля определённой «школы». В третьих, это направление editorial philology — близкая к текстологии область, ориентированная на реконструкцию «правильных» или «изначальных» текстов, которые могли быть утрачены в процессе их длительного бытования, выявление различных слоев в оригинальном тексте и презентацию результатов. Такие исследования проводят на древних текстах - например, хрониках, произведениях античных авторов и религиозных текстах.

2. Вторая область фокусируется на анализе фрагментов информации, извлеченных из различных источников: текстов, описаний изображений, социальных сетей и так далее. В рамках данной области исследователи не занимаются анализом текстов как таковых, а используют тексты или другие источники для получения некоторых фрагментов информации, условно называемых «фактами», для дальнейшего анализа совокупности этих «фактов». Данный подход наиболее широко применяется в дисциплинах, которые связаны с изучением социальных явлений, например, в истории, антропологии, или материальных объектов — например, в археологии и истории искусства.

В подходах разных гуманитарных наук к анализу «фактов» можно выделить две основных группы подходов. Во-первых, это математическое моделирование, которое пользовалось наибольшей популярностью в 1980х гг., ставит целью сравнение результатов развития некоторого явления, полученных в ходе вычислений и представленных в виде математической модели, с результатами, наблюдаемыми в реальности.

Во-вторых, это организация и анализ больших объёмов данных. Под организацией данных чаще всего предполагается создание баз данных. В данном случае извлеченные «факты» могут служить, например, для построения визуализаций или для репрезентации информации об объектах.

3. Далее, существует большая область Digital Humanities, в рамках которой идет работа с нетекстовыми ресурсами. С одной стороны, данная область включает оцифровку больших коллекций изображений и дальнейшую работу с ними в таких дисциплинах, как археология или история искусства; с другой стороны, она связана с использованием трехмерных моделей артефактов и других средств визуализации в гуманитарных дисциплинах. Данная область включает в себя также построение разного рода визуализаций для репрезентации данных — визуализации достаточно широко используются в DH, так как существует множество результатов исследований, которые легче понять, если они представлены графически, а не в виде таблиц.

4. Наконец, М. Таллер выделяет область, которую он называет Humanities Computer Science (русский аналог — информатика в гуманитарных науках). В рамках данной области Таллер отмечает два подхода: формализованный и эпистемологический. **Формализованный подход** к гуманитарной информатике предполагает, что существуют некоторые фундаментальные отличия в типах информации, которыми оперируют гуманитарные дисциплины, от тех данных, что используются в других прикладных сферах. Эти отличия настолько велики, что требуют специальной адаптации программных средств под специфичные нужды исследователей в DH. Если интересы исследователей не могут быть удовлетворены средствами существующего программного обеспечения, Humanities Computer Science развивает знания и навыки решения таких задач. **Эпистемологический подход**, возникающий в данном контексте, рассматривает вопрос о том, как изменяются гуманитарные исследования под влиянием методов новых методов компьютерного анализа.

Тест 3

Совместите термины с их краткими определениями:

1. IIIF	A. Упорядоченный набор структурированной информации
2. Стилометрия	B. Набор технических спецификаций для расширенного использования изображений в интернете
3. Корпус	C. Технология оптического распознавания символов
4. OCR	D. Статистический анализ применительно к письменным текстам
5. База данных	E. отобранная и обработанная совокупность текстов, используемых в качестве базы для исследования языка

Рекомендуемая литература

1. Nyhan J., Flinn A. Computation and the humanities: towards an oral history of digital humanities. – Springer Nature, 2016. – С. 285.
2. Sula C. A., Hill H. The Early History of Digital Humanities //DH. – 2017.

3. Kirschenbaum M. G. What is digital humanities and what's it doing in English departments? — 2010. [Электронный ресурс], способ доступа: https://mkirschenbaum.files.wordpress.com/2011/03/ade-final.pdf

Глава 2. Гуманитарные данные: от сбора к анализу

Что такое гуманитарные данные?

Если мы соглашаемся с наиболее общим определением DH, то цифровые гуманитарные науки представляют собой взгляд на гуманитарные области знания в их классическом понимании сквозь оптику цифровых технологий или применение цифровых инструментов для поиска ответов на исследовательские вопросы гуманитарных областей. В этой связи неизбежно встает вопрос - к чему мы можем приложить те самые цифровые инструменты? Ответ кроется в трансформации взгляда исследователя: исследовательский материал должен стать исследовательскими данными. Но о каких данных может идти речь применительно к гуманитарным наукам? Ответу на этот вопрос будет посвящен следующий раздел.

Итак, практически все проекты цифровой гуманитаристики построены с опорой на данные в том или ином виде: будь то визуализация, интерактивная карта, мультимедиа публикация и тд. Таким образом, закономерными вопросами при проектировании DH проекта любой сложности и масштаба будут:

- Какие данные использовать?
- В каких форматах эти данные должны быть представлены (в зависимости от задействованных платформ и инструментов).
- Где собирать данные и как их хранить?
- Как собранные данные следует репрезентировать, чтобы они работали на достижение цели проекта или исследования?

Типология гуманитарных данных

Что такое данные в целом? Как и в случае попыток определения, что такое Digital Humanities, данные могут быть определены совершенно разными способами. Но, в общих чертах, можно сказать, что данные - это поддающаяся обработке и интерпретации информация. Данные не всегда подразумевают сложную техническую обработку кодом, данные могут представать перед исследователем во множестве форматов: это могут быть изображения, аудио, видео, текстовая и табличная информация в различных форматах. Если вы занимались научным исследованием, то наверняка вы уже имели дело с данными, даже если не осознавали этого.

Основные примеры данных, которые можно использовать для гуманитарных исследований, включают:

- текстовые файлы, извлеченные из корпусов текстов с помощью программного обеспечения оптического распознавания символов (OCR) и возможной последующей разметки;
- данные, собранные посредством парсинга или с помощью API из открытых источников в сети интернет;
- цифровые изображения музейных предметов или произведений искусства;
- геопространственные данные, включая растровые и векторные файлы;
- аудио файлы и расшифровки устной речи;
- архивные или музейные метаданные.

Говоря об исследованиях, в идеальном мире работа над Digital Humanities проектом начинается именно с данных: анализируете какие данные вам доступны, насколько они готовы к обработке и анализу, и достаточно ли их для ценных научных заключений и выводов. Однако чаще всего идея приходит раньше, чем понимание возможности ее реализации. Соответственно, перед каждым исследователем стоит задача поиска и сбора данных под свою конкретную идею.

В таком случае данные станут для вас либо ресурсом вдохновения для проекта, либо фильтром, сквозь который вы будете смотреть на свою идею. Они помогут сузить грандиозный неподъемный замысел до реализуемого, будучи объективным мерилом ваших возможностей. Если вы не можете найти данные под свой исследовательский вопрос, значит, его нужно сузить или посмотреть под другим углом.

Этические принципы работы с гуманитарными данными

В сфере Digital Humanities, как и в любой другой научной области, очень серьезно относятся к корректности данных и заключений, построенных на этих данных. К сожалению, на практике иногда появляются исследования, которые преувеличивают или, наоборот, преуменьшают какие-то показатели для того, чтобы их вывод звучал свежим и новым для сообщества. Более того, известны случаи фальсифицирования данных для исследований в области цифровой гуманитаристики. Один из анти-примеров - исследование профессора политологии Колумбийского университета Дональда Грина и Майкла ЛаКура, который на момент скандала являлся аспирантом политологии Калифорнийского университета. Их исследование демонстрировало возможность изменения мнения людей о проблеме однополых браков. Статья, опубликованная в академическом

журнале Science в декабре 2014 года, была признана мошеннической и отозвана полгода спустя после того как выяснилось, что данные были сфальсифицированы Майклом ЛаКуром.

Иногда исследования невозможно проверить из-за закрытого доступа к данным, но их заключения оказывают большое влияние на науку и жизнь в целом. Однако такая ситуация встречается редко, особенно в сферах, не имеющих отношения к государственным и корпоративным тайнам. Все чаще можно встретить просьбы предоставить вместе с выводами исследования данные, на которых они основываются. Для предотвращения обмана и проверки правдивости и честности исследователя научное сообщество требует полное описание данных, само их наличие, а также порядок их обработки для того, чтобы весь процесс анализа можно было повторить и прийти к тем же выводам, что и автор статьи/исследования.

Тест 4

Вопрос 1. Выберете верные утверждения о данных.

A. Данные - это информация, представленная в формализованном виде, доступная для анализа

B. Существует только два типа данных - текстовые и визуальные

C. Существует множество типов данных

D. В DH сообществе очень спокойно относятся к фальсифицированную данных

Вопрос 2. Чем известен Майкл ЛаКур?

A. Фальсификацией данных для исследования

B. Он дал определение данных, которым все пользуются

C. Создатель основной типологии данных

Сбор гуманитарных данных и их подготовка к анализу

Где же искать данные? Понятно, что данные окружают нас, и практически все, что мы воспринимаем в течение дня, так или иначе можно назвать данными. Однако, особенно для начинающего исследователя, есть несколько полезных типов ресурсов данных. В них входят открытые данные правительств и независимых организаций, а также веб-сайты, удобно структурированные для сбора данных с их страниц.

В рамках вводного курса мы будем рассматривать в основном структурированные данные. Неструктурированные данные, например, неразмеченный текст, фотографии и видеоматериалы, могут содержать огромное количество полезной информации, однако их обработка требует

дополнительных продвинутых навыков в таких сферах, как компьютерное зрение, анализ данных, машинное обучение и так далее.

Структурированные данные, то есть данные, которые уже были собраны и частично подготовлены другими исследователями, организациями и проектами, являются наиболее комфортными для работы начинающего исследователя в сфере Digital Humanities. Примерами структурированных данных могут быть размеченные тексты, табличные данные и различные структуры данных. Мы остановимся подробнее на структурированных данных и рассмотрим, как их найти и собрать, а также как их адаптировать и подготовить для своего проекта.

Открытые данные

Открытые данные являются универсальным ресурсом материала для начинающего исследователя. Само понятие открытых данных означает данные, которые любой пользователь может свободно получить, использовать и распространять - на эти данные не наложено ограничение авторского права, и их можно использовать в своих исследованиях и проектах. Чаще всего, это уже подготовленные к анализу структурированные данные, большинство ресурсов предоставляют простейшую визуализацию вместе с самими датасетами.

К сожалению, не всегда термин открытые данные связан с понятием бесплатного доступа к данным. Когда мы говорим об открытых данных, мы всегда имеем в виду свободу того, как мы данные используем. При этом получение доступа к данным может быть ограничено с финансовой точки зрения, так как сбор, обработка и публикация данных в открытом доступе всегда связана с определенными затратами. Пожалуй, самый популярный платный ресурс данных - Statista (https://www.statista.com), которая содержит уникальную статистику и отчеты по 150 странам и 600 отраслям. В нашем курсе мы рассмотрим только бесплатные ресурсы данных.

В современном мире открытые данные предоставляет большая часть государств и международных организаций. Кроме того, многие исследователи делятся собранными данными. Как мы обсудили ранее, доступ к данным какого-либо исследования в современной науке считается необходимым, так как “открытость” позволяет проверить истинность заключений и высказать конструктивную критику исследованию, если это необходимо.

Для исследователей в сфере Digital Humanities открытость данных является одним из самых важных принципов работы. “We call for open access to data and metadata, which must be documented and interoperable, both technically and conceptually[9]” - первое требование манифеста

[9] https://tcp.hypotheses.org/411 - манифест DH

исследователей Digital Humanities. Открытость данных не только позволяет исследователям осуществлять свои проекты без бесконечных запросов на доступ к данным, но и позволяет осуществлять проверку этих проектов и их выводов и заключений.

Ресурсов с открытым доступом к данным достаточно много, их можно использовать как ресурсы вдохновения и понимания, какие проекты можно реализовать, основываясь на доступных данных и осуществленных проектах. Мы рассмотрим несколько ключевых ресурсов открытых данных и какие именно данные можно там найти.

Одними из ключевых ресурсов являются данные всемирных организаций, например, The World Bank Data (https://data.worldbank.org) - данные Всемирного Банка, UN Data (http://data.un.org/Default.aspx) - данные ООН, OECD Stat (https://stats.oecd.org) - данные Организации экономического сотрудничества и развития. Всемирный банк, являясь крупнейшей организацией по организации финансовой и технической поддержки государств по всему миру, предоставляет открытый доступ к данным о развитии стран. Ресурс содержит статистические сведения по 570 показателям мирового развития для 208 стран во временной промежуток начиная с 1960 года. Охвачены экономические, социальные, финансовые показатели, данные по природным ресурсам и окружающей среде. ООН также предоставляет данные о своих странах-участницах — население, уровень образование, состояние рынка труда, уровень преступности, и так далее.

Кроме организаций, свой вклад в открытость данных вносят и обычные пользователи. Например, существует открытый ресурс по data science и машинному обучению (https://www.kaggle.com) с разделом датасеты. Любой человек может опубликовать свой датасет, дополнить чужой и в целом поделиться своим мнением по поводу собранных данных. Также есть открытое сообщество data.world (https://data.world/datasets/) где также можно найти много датасетов на различные темы.

Помимо всемирных организаций и общественных начал, многие страны публикуют в открытом доступе свои данные. Например, открытые данные Великобритании (https://data.gov.uk), США (https://www.data.gov), Германии (https://www.govdata.de) и так далее.

В России, по сравнению с другими странами, единый ресурс государственных открытых данных до конца не налажен, однако отдельные министерства и отдельные города обладают своими собственными порталами открытых данных. Например, портал открытых данных министерства культуры РФ (https://opendata.mkrf.ru), министерства труда РФ (https://mintrud.gov.ru/opendata), и многие другие.

К сожалению, несмотря на доступность, открытые данные могут иметь некоторый ряд проблем или не отвечать запросам конкретного

исследования. Часто такие данные нужно обрабатывать дополнительно или дополнять из других источников.

API — это контракт, который предоставляет программа. «Ко мне можно обращаться так и так, я обязуюсь делать то и это»

Яркие примеры компаний, которые предоставляют доступ к API, а соответственно и к своим данным, это Google, ВКонтакте, Вики данные. Проект DH центра St.Retrospect тоже имеет открытое API.

Веб-скрейпинг

На просторах интернета есть бесконечное количество сайтов, с которых удобно собирать информацию. Это могут быть как онлайн энциклопедии вроде Википедии, так и обычные сайты, изначально выполняющие другие задачи, например, маркетинговые. Для начала давайте рассмотрим некоторые плюсы и минусы сбора данных с веб-сайтов.

В чем плюс такого подхода? Однозначно, удобство при сборе данных, которых, возможно, не существует в уже подготовленном формате. Как мы говорили ранее, идеи для проектов и исследований, особенно в таком широком междисциплинарном поле, как Digital Humanities, зачастую приходят раньше, чем понимание того, как эти идеи реализовывать. Найти готовые данные не всегда представляется возможным, особенно если идея проекта или исследования новая и до этого не разрабатывалась.

Есть ли минусы у такого подхода? К сожалению, да. Например, вы легко можете столкнуться с защитой сайта от парсинга. К такому могут прибегнуть владельцы веб-ресурса, если они опасаются кражи своего уникального контента. Парсинг данных, помимо исследователей и аналитиков, активно используют, например, маркетологи — чтобы не заполнять описание товаров и услуг из раза в раз, они могут спарсить описания с сайтов смежной тематики и переиспользовать в своих целях. Во избежание таких ситуаций, многие ставят защиту от парсеров. В таком случае вы сталкиваетесь с ситуацией, когда большую часть времени, выделенного на свое исследование, вы будете пытаться обойти защиту одного ресурса. В этой ситуации важно понимать, оправданы ли такие затраты, и не можете ли вы найти похожие или те же самые данные на другом сайте.

Помимо этого, сайт может быть в целом непригодным для парсера с точки зрения структуры. Иногда вам, с точки зрения человеческого понимания информации, кажется, что сайт понятен и доступен. Однако, зачастую, структура таких сайтов для машинной обработки может не иметь смысла. И в таких случаях парсер не распознает часть данных как важные и либо теряет их, либо записывает в итоговую таблицу не так, как вам бы хотелось. Случаи таких сайтов также показывают еще один минус - если вы пишете код, простейший скрипт всегда будет заточен под конкретный

ресурс, вы не сможете переиспользовать его несколько раз на разных сайтах. И вновь встает вопрос, оправдано ли затраченное время.

И напоследок, проблема может быть в самих данных. На первый взгляд, вам может казаться, что данные, содержащиеся на ресурсе, вам очень помогут. Однако при сборе окажется, что, например, часть нужных вам данных вовсе отсутствует, что может быть незаметно на первый взгляд. Или данные в целом являются ошибочными. От такого, к сожалению, никто не застрахован, и иногда, в случае очевидной ошибки в одном месте, вы не застрахованы от перепроверки правильности данных вручную на протяжении всего спарсенного датасета.

Важно также уделить отдельное внимание правовому аспекту парсинга сайтов. Парсинг — не то же самое, что API. Например, компания может открыть доступ к API, чтобы позволить другим системам взаимодействовать с ее данными. Однако API дают далеко не все ресурсы, так как это подразумевает дополнительные затраты и усилия со стороны компании. Яркие примеры компаний, которые предоставляют доступ к API, а соответственно к своим данным, это Google или ВКонтакте. При этом у этих компаний также есть ограничение на доступ к данным - какое-то количество запросов в день или какой-то фиксированный объем полученных данных.

Что касается парсинга, периодически он становится центром судебных разбирательств. Например, компании Facebook и eBay уже подавали в суд на тех, кто собирал данные с их сайта. В случае, если вы пользуетесь парсингом, вам необходимо убедиться в том, что вы не собираете контент, защищенный авторским правом, что вы не затрагиваете персональные данные пользователей сайта. В случае научного исследования вам необходимо указать ресурс, откуда данные были получены и каким образом.

Таким образом, учитывая все тонкости парсинга, рассмотренные в этом блоке, мы можем сказать, что для начинающего исследователя парсинг сайта будет скорее дополнением к уже имеющимся собранным данным. Если вам нужно расширить уже существующий датасет, дополнить его данными о новых объектах исследования или получить дополнительную информацию о уже имеющихся объектах, парсинг поможет вам сделать это достаточно быстро. В следующем разделе мы разберем, как в дальнейшем множественные датасеты можно будет объединить и привести к общему виду.

Как же все богатство данных собрать и использовать для своего проекта или исследования? Конечно, есть вариант вручную копировать и вносить все в таблицу, однако на это уйдет не день и не два. Поэтому в этом блоке мы рассмотрим с вами несколько путей автоматизации процесса сбора данных, которые подойдут людям с любым уровнем знания программирования. Готовые парсеры выполняют тот же механизм, что и

написанный вручную код, но делают это, не требуя от вас понимания исходного кода страницы и функций извлечения данных.

Давайте рассмотрим, какие варианты готовых парсеров у нас с вами есть в доступе. Самым простым решением являются парсеры - расширения для браузеров. Парсеры также бывают самостоятельными программами и сервисами, не требующими установки, но они обычно не бесплатные. Мы будем для примера использовать парсер Web-scraper. Web-scraper (https://webscraper.io) - один из самых доступных ресурсов получения данных с веб-страницы без необходимости писать длинный код, он является плагином, доступным для Chrome и Firefox. Устанавливаем расширение и начинаем работу.

Парсинг является довольно распространенной задачей, поэтому для облегчения работы сбора данных с помощью кода, программисты создали множество решений для упрощения задачи. Например, существует ряд библиотек для языка программирования Python. Мы рассмотрим BeautifulSoup в рамках практического занятия. (https://www.crummy.com/software/BeautifulSoup/). Библиотека BeautifulSoup является универсальной.

Подготовка данных к анализу

Полученные перечисленными способами данные не всегда бывают готовыми для анализа. Неподготовленные данные часто называют сырыми. Такие данные могут иметь различные недостатки: в них может быть лишняя и избыточная информация, а может наоборот чего-то не хватать; данные могут быть неоднородными или быть представленными в неподходящем формате и так далее. Для того, чтобы подготовить их к анализу и дальнейшей визуализации, нам необходимо превратить сырые данные в чистые — данные, приведенные к единому формату, легко поддающиеся дальнейшему анализу и преобразованиям.

В зависимости от проблем, которые исследователь видит в первоначальном «сыром» датасете, чистка данных может разбиваться на несколько последовательных этапов:

Первый этап — фильтр данных. На этом этапе вам необходимо понять, все ли данные, полученные в предыдущем этапе, нужны для вашего исследования. И все ли необходимые вашему исследованию данные вы нашли. Этот этап не всегда становится ключевым, но его необходимо держать в голове. Кроме того, иногда недостающие данные становятся очевидными на этапе обработки и анализа данных, от этого никто не застрахован.

Второй этап — структурирование данных. Если вы сочетаете данные из разных ресурсов, например, часть вы нашли в открытых данных, а часть спарсили с сайта, вам необходимо привести их к общей форме, например, к

общей таблице с разбитыми категориями. Также вас может не устраивать изначальная категоризация данных. Если вам необходимо ввести свою категоризацию, ее важно продумать и обосновать.

Третий этап — унификация данных. Для того, чтобы в дальнейшем проводить анализ полученных данных, вам нужно убедиться, что все данные приведены к единому формату. Например, даты записаны одним форматом, названия записаны все в едином стиле (с большой буквы / с маленькой, содержат знаки препинания / не содержат, и так далее).

Помимо технических характеристик, конечно же, у нас есть требования к содержанию данных. Во-первых, нам важно убедиться, что данные точные, как минимум не содержат устаревшую информацию, например, в базе данных организаций указан старый адрес или номер телефона. Параллельно с этим нам важно понимать надежность данных, насколько мы доверяем ресурсу, откуда данные взяты.

По итогу, уделив внимания всем перечисленным этапам, вы получите данные для дальнейших манипуляций - анализа и визуализации. В практической части мы предлагаем несколько рекомендаций по предобработке данных и приведём примеры решений по обработке данных с помощью готовых инструментов и с помощью кода.

Тест 5

Вопрос 1. Выберете верные утверждения об API (возможно несколько правильных вариантов).

A. Это описание способов, которыми одна компьютерная программа может взаимодействовать с другой программой.

B. Многие компании и интернет-сервисы, например, Facebook, Google, Vk, имеют свой API.

C. Часто используется при написании приложений

Вопрос 2. С какими проблемами вы можете столкнуться при парсинге сайтов? (возможно несколько правильных вариантов)

A. Сайт может быть защищен от парсинга.

B. При обновлении дизайна сайта парсер может сломаться.

C. Это очень сложно. Для парсинга нужно обязательно писать большой код.

D. Сайт может заблокировать запросы с вашего IP адреса.

Вопрос 3. Что такое BeautifulSoup?

A. Это библиотека Python, с помощью которой создаются сайты

B. Это готовая программа для парсинга сайтов

C. Библиотека Python для парсинга сайтов

D. Это веб-сайт, на котором можно найти готовые наборы данных

Практика

Вариант 1: Web Scaper

Теперь мы создаем парсер для веб-сайта, данные которого хотим собрать. В этом случае нас интересует, какие игры представлены на первой странице Metacritic и какие у них пользовательские оценки. Более конкретно, мы хотим последовательно собрать три переменные:

а. Название игры

б. Аннотация

в. Оценка пользователей

Откройте интересующий вас веб-сайт, чтобы спарсить данные, используя плагин для Chrome. В качестве примера для этого руководства мы будем использовать список игр с портала Metacritic. (https://www.metacritic.com/browse/games/score/metascore/all/all/filtered)

Мы хотим получить URL-адрес на этой странице (см. Рисунок 1).

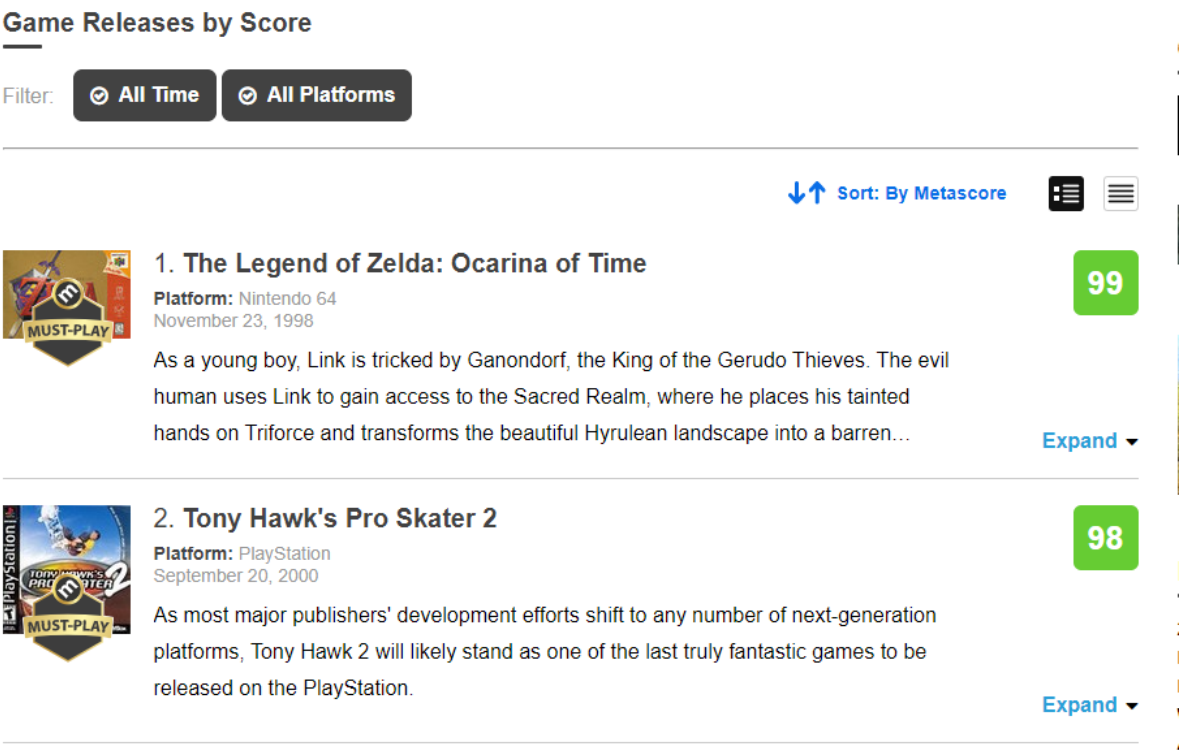


Рисунок.1 — Сайт для выполнения практики

Откройте URL-адрес, данные с которого вы хотите собрать (в данном случае https://www.metacritic.com/browse/games/score/metascore/all/all/filtered) и откройте Web Scraper, щелкнув страницу правой кнопкой мыши и выбрав «Проверить» или «Посмотреть код».

Затем щелкните вкладку Web Scraper на панели Inspect (вкладка в верхней части панели). Примечание. Если вы его не видите, вы можете

нажать кнопку «>>» на панели, чтобы найти вкладку «Web Scraper», и, следуя инструкциям, переместите панель вниз(см. Рисунок 2).

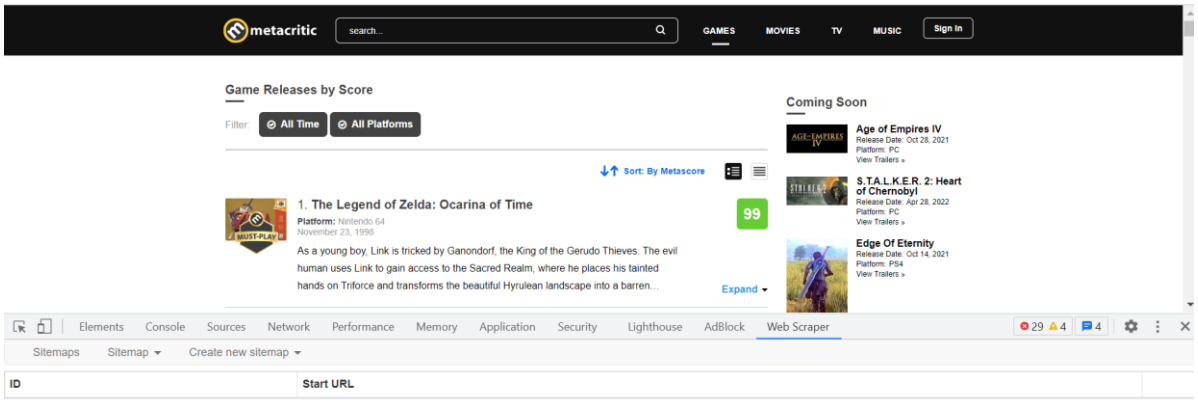


Рисунок 2 — Вид рабочего окна Web Scraper

Теперь нам нужно создать новую карту сайта, которая буквально представляет собой «карту», которую мы разрабатываем для «сайта», данные с которого мы хотим собрать. Для этого щелкните вкладку «Создать новую карту сайта» (на вкладке Веб-парсер) и выберите «Создать карту сайта» (см. Рисунок 3).

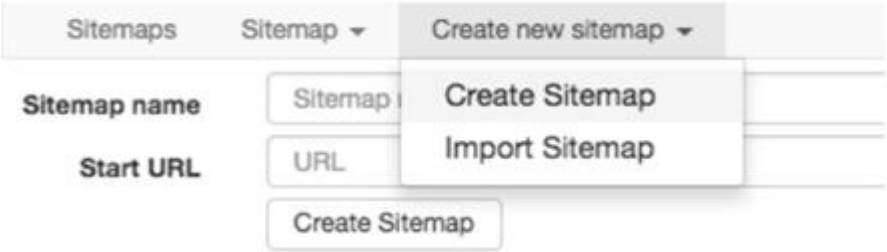


Рисунок 3 — Окно для создания «карты сайта»

Выберите имя для нашей карты сайта, здесь мы назовем его «metacritic_games». Также заполните URL стартовой страницы (скопируйте и вставьте), в нашем случае: https://www.metacritic.com/browse/games/score/metascore/all/all/filtered(см. Рисунок 4).

Нажмите «Создать карту сайта».

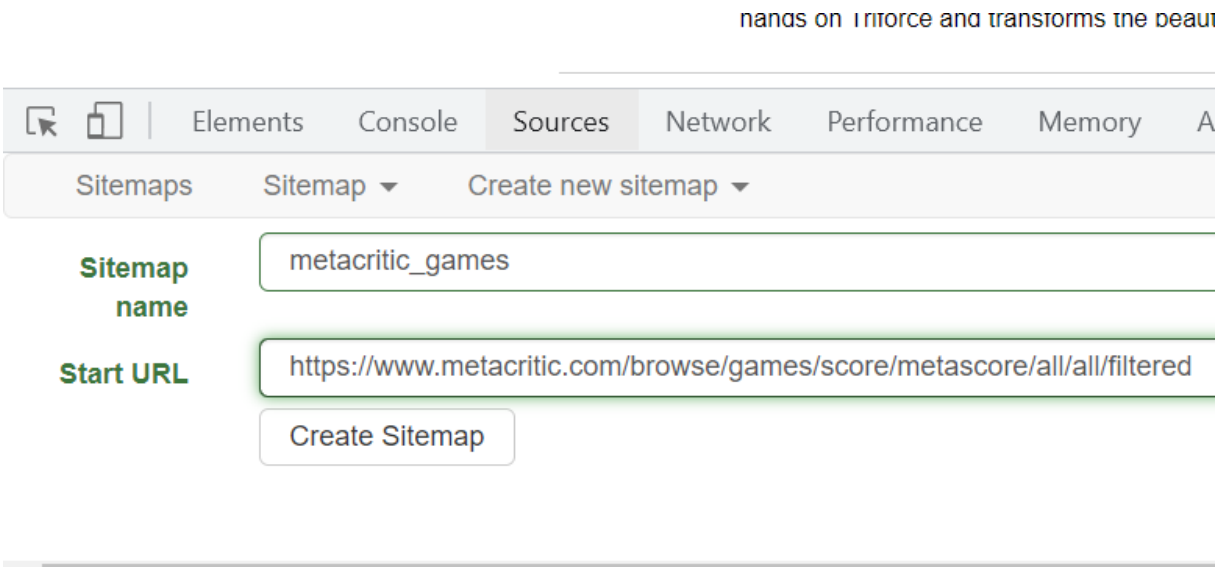


Рисунок 4 — Параметры при создании парсера

После того, как мы создали карту сайта, мы получим новый пустой каталог с именем _root (см. Рисунок 5).

Рисунок 5 — Основная рабочая область при создании

Теперь, когда у нас есть пустая карта, нам нужно указать парсеру, что выбрать. Мы добавляем новый селектор для нашего парсера, нажав кнопку «Добавить новый селектор» (см. Рисунок 6). Мы назовем наш первый селектор «Game», так как мы будем работать с играми. Установите для типа значение «Element» (нам нужны отдельные данные внутри этого элемента), затем нажмите кнопку «Select» под селектором.

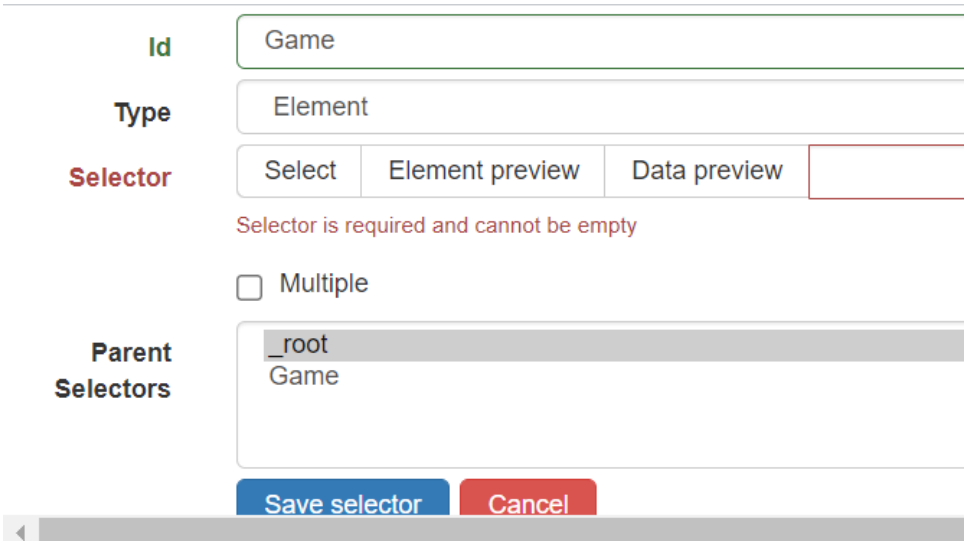


Рисунок 6 — Создание селектора

Теперь нам нужно выбрать то, что мы хотим собрать. Обратите внимание, что, когда вы наводите указатель мыши на ссылки с веб-страницы, они меняют цвет (вам может потребоваться уменьшить окно управления парсером). Наведите курсор мыши на первую игру. Затем щелкните (см. Рисунок 7). Цвет должен измениться от желтого к красному.

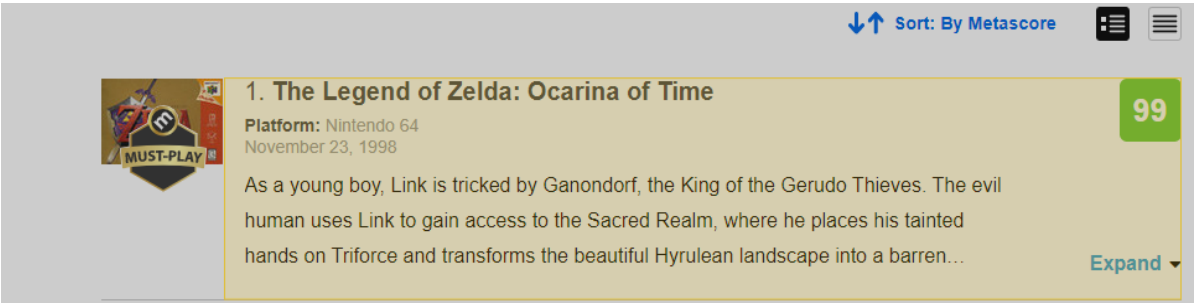


Рисунок 7 — Пример выбора элементов

Теперь мы успешно нашли первую игру для нашего парсера. Следующим шагом будет сбор всех похожих элементов (все ссылки заголовков). Здесь происходит волшебство, за исключением того, что это не волшебство, а веб-парсер, использующий архитектуру веб-страницы. Щелкните вторую игру, и парсер распознает все интересующие нас элементы на странице и закрасит их в красный цвет (см. Рисунок 8). Парсер понял, что все эти элементы имеют одинаковую структуру, и пришел к выводу, что мы хотели бы получить их все.

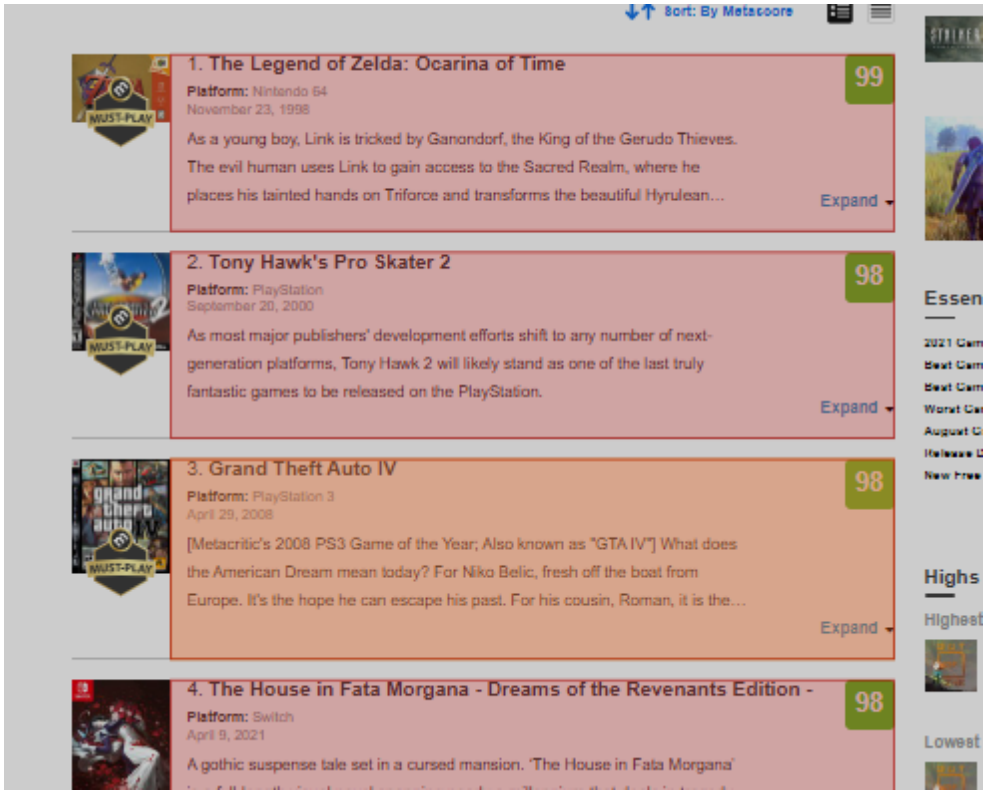


Рисунок 8 — Пример автоматического выбора элементов

Теперь вам нужно сохранить этот выбор, нажав кнопку «Done selecting». Затем щелкните галочку для «Multiple», поскольку мы хотим, чтобы парсер просматривал все игры. Затем нажмите «Save selector». Теперь мы успешно создали наш первый селектор (см. Рисунок 9), он собирает «элементы» — игры.

ID	Selector	type	Multiple	Parent selectors
Game	.one td.clamp-summary-wrap	SelectorElement	yes	_root

Рисунок 9 — Пример готового селектора

Дальше вы переходите в селектор Game и выбираете все интересующие вас элементы в рамках одной игры (см. Рисунок 10), а Web Scraper переведет эту схему на другие игры. Например, мне нужны название, аннотация и пользовательская оценка.

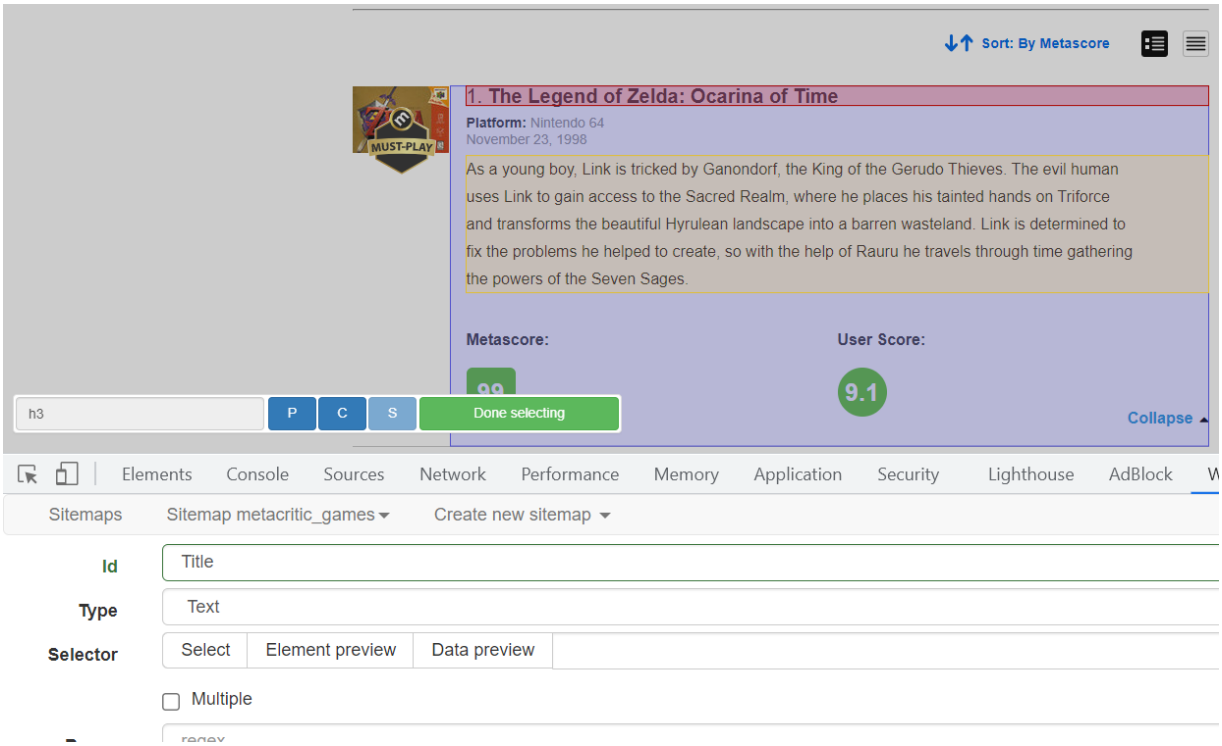


Рисунок 10 — Создание нового селектора внутри селектора Game

Таким образом, селекторы внутри игры будут выглядеть так (см. Рисунок 11) .

Рисунок 11— Готовые селекторы

Чтобы проверить логику вашего парсера, можно выбрать sitemap metacritic_games и перейти в selector graph. Вы должны увидеть результат, похожий на рисунок 12. Все кружки интерактивны.

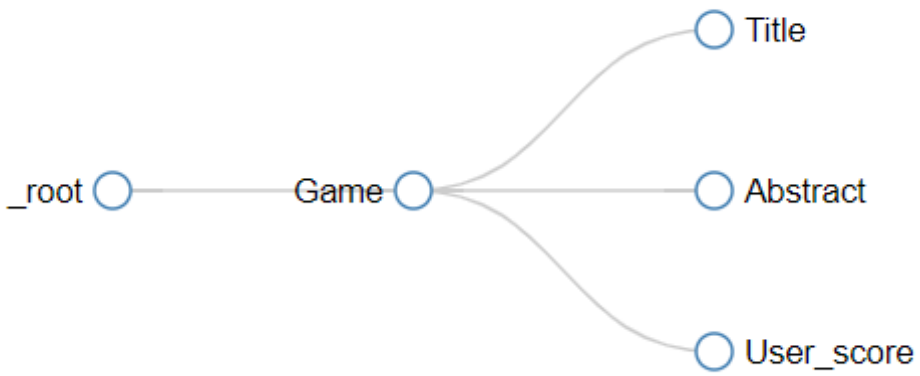


Рисунок 12 — Граф связей внутри парсера

Если все в порядке, можно переходить к парсингу. Запустите парсер в разделе «Sitemap metacritic_games» и нажмите кнопку «Scrape». Затем нажмите «Start scraping» (см. Рисунок 13).

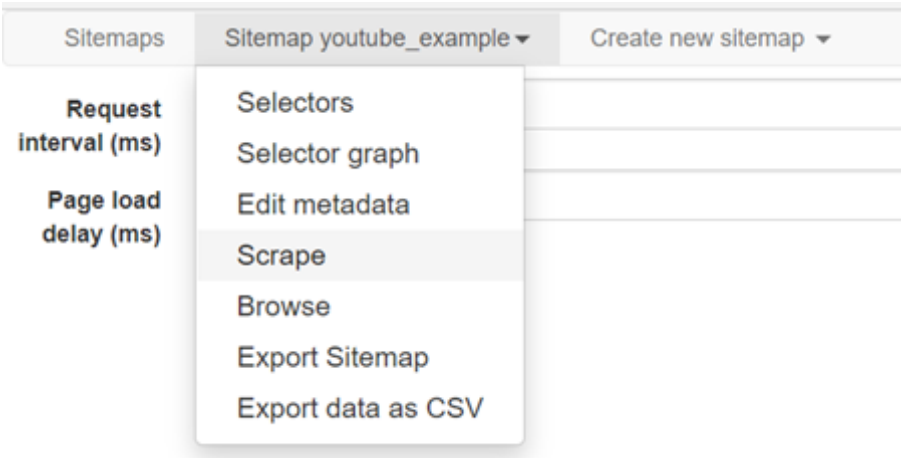


Рисунок 13 — Вид окна для начала парсинга

Появится новое окно, в котором будут отображаться все игры, которые вы собрали. Хорошо, что это не нужно делать вручную! Здесь вы видите, что ваш алгоритм выполняет свою работу за вас. Сбор данных выполняется, когда новое окно закрывается. Нажмите «Refresh» (который появится), чтобы увидеть данные.

После завершения парсинга мы можем вывести данные в CSV-файл, который мы можем открыть в MS Excel, таблицах Apple iWork Numbers, Google Sheets или LibreOffice и т.д. Нажмите «Export data as CSV», чтобы перейти на страницу загрузки. Нажмите «Download now», чтобы скачать файл.

Это конец туториала, теперь ваша очередь собирать данные с интересующих вас сайтов.

Вариант 2: Парсинг данных с помощью кода (Beautifulsoup4)

Для того, чтобы освоиться с парсингом через код желательно выбрать страницу с простой структурой. Для того, чтобы взаимодействовать со страницей, нам понадобиться несколько дополнительных библиотек (requests, beautifulsoup4). Если вы работаете в локальной среде разработки, то используйте код с рисунка 14. Мы же советуем использовать Google Colab, так как эти библиотеки уже встроены, и вам ничего не нужно скачивать на ваш компьютер — все действие происходит в облаке.

```
pip install requests
pip install beautifulsoup4
```

Рисунок 14 — Пример установки библиотек

Для работы с этими библиотеками нам нужно их импортировать, с этим поможет код на рисунке 15.

```
[1] import requests
    from bs4 import BeautifulSoup
```

Рисунок 15 — Импорт необходимых библиотек

Чтобы проверить работу библиотеки ***requests,*** можете скопировать код с рисунка 16. В качестве примера мы взяли страницу про цифровые гуманитарные исследования (https://ru.wikipedia.org/wiki/%D0%A6%D0%B8%D1%84%D1%80%D0%BE%D0%B2%D1%8B%D0%B5_%D0%B3%D1%83%D0%BC%D0%B0%D0%BD%D0%B8%D1%82%D0%B0%D1%80%D0%BD%D1%8B%D0%B5_%D0%BD%D0%B0%D1%83%D0%BA%D0%B8).

```
[3] url = "https://ru.wikipedia.org/wiki/%D0%A6%D0%B8%D1%84%D1%80%D0%BE%D0%B2%D

    response = requests.get(url)
    print(response.status_code)

    200
```

Рисунок 16 — Пример запроса к сайту Wikipedia.org

В переменной ***url*** мы сохраняем адрес страницы, который нам интересен. А функция ***requests.get*** помогает нам получить информацию со страницы, передающуюся в переменную ***response***. Последняя строчка проверяет успешность запроса, если появляется число 200, то все сделано правильно.

Теперь перейдем к работе с самой страницей и библиотекой Beautifulsoup4. Для разметки текста используется язык разметки HTML, который использует теги и атрибуты. Для изучения интересующих вас элементов страницы, обозначенных тегами, можно использовать сочетание клавиш Ctrl+Shift+C, они вызовут инструмент Inspect. Мы можем выделить полезный нам элемент страницы, а Inspect покажет нам всю информацию о теге, которая его описывает. Например, мы хотим достать название страницы, как это показано на рисунках 17, 18.

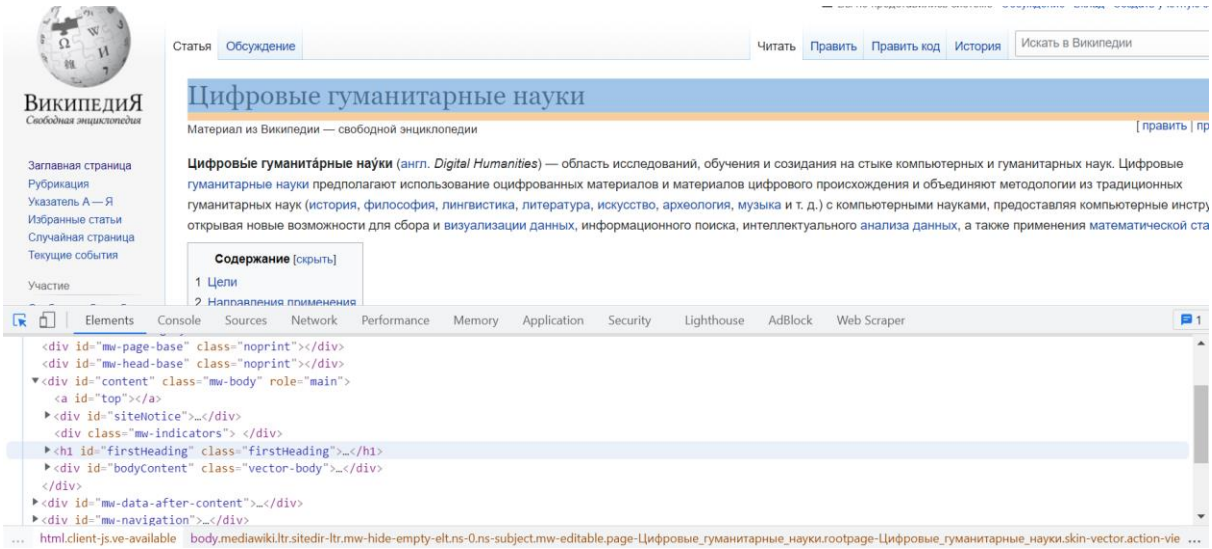


Рисунок 17 — Пример работы с функцией инспектора

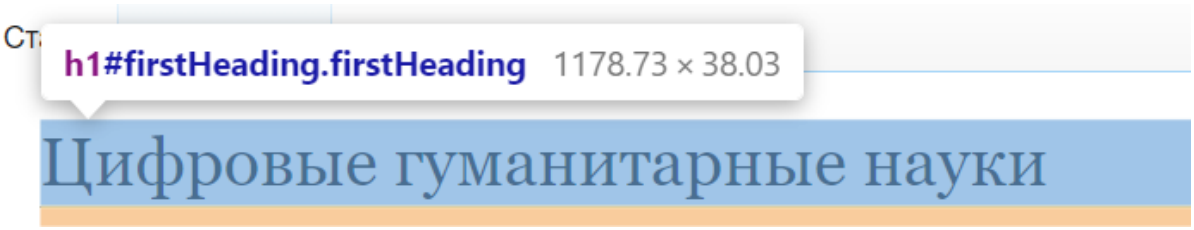


Рисунок 18 — Пример работы с функцией инспектора

Beautiful soup позволяет достать элемент по ID (см. Рисунок 19), и необходимый код будет выглядеть так:

```
[ ]  title = soup.find(id="firstHeading")
```

Рисунок 19 — Выбор элемента по ID

Чтобы программа заработала, нам нужно все соединить (см. Рисунок 20).

```
response = requests.get(url)
soup = BeautifulSoup(response.content, 'html.parser')

title = soup.find(id="firstHeading")
print(title.string)

Цифровые гуманитарные науки
```

Рисунок 20 — Готовый код для сбора названия

Важно понимать, что веб-страницы имеют древовидную структуру, а библиотека Beautifulsoup умеет разбирать ее на отдельные элементы. Для доступа к некоторым страницами вам потребуются дополнительные

параметры вроде useragent и proxy, их описание вы можете найти в документах к библиотеке requests.

А теперь советую вам попрактиковаться самостоятельно и собрать со страницы все ссылки в главе «Направления и применения». Для поиска решения вы можете использовать документацию (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) и дополнительные ресурсы.

Рекомендуемая литература

1. Митчелл, Р. Скрапинг веб-сайтов с помощью Python : руководство / Р. Митчелл ; перевод с английского А. В. Груздев. — Москва : ДМК Пресс, 2016. — 280 с. — ISBN 978-5-97060-223-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/100903 (дата обращения: 16.09.2021). — Режим доступа: для авториз. пользователей.
2. Kirschenbaum M. G. The remaking of reading: Data mining and the digital humanities //The National Science Foundation symposium on next generation of data mining and cyber-enabled discovery for innovation, Baltimore, MD. – 2007. – Т. 134.

Глава 3. DH проект: постановка проблемы и основные этапы

DH проекты: какими они бывают?

Сфера Digital Humanities привлекает к себе огромное количество исследователей своей инклюзивностью и многообразием. Как было отмечено ранее — нет единого определения, что такое цифровая гуманитаристика, в том числе потому, что нет единого метода или подхода к реализации DH проекта. Существует мнение, что лучше всего определять, что такое DH, как раз через разного рода проекты в этой области, потому что именно сквозь призму реализованных проектов становится понятен весь спектр задач, которые может поставить перед собой исследователь, и весь арсенал доступных методов, подходов и инструментов. Именно поэтому для начинающего исследователя в сфере DH так важно тренировать “насмотренность” и быть в курсе актуальных DH проектов.

Прошлый раздел был посвящен методам и подходам, которые наиболее часты для области цифровой гуманитаристики. Выбор метода напрямую зависит от выбранной исследователем темы, исходного набора данных, которыми он или она располагает или хочет собрать, а главное, в каком виде планируется представить полученный результат, кто будет целевой аудиторией данного проекта. Совокупность этих факторов составляет уникальность конкретного проекта. Предлагая некоторую кластеризацию DH проектов, следует иметь в виду, что не существует

единой выверенной системы классификации: разделять проекты на группы можно, руководствуясь самыми разными принципами.

- По принципу разных способов визуализации, примененных в проекте: мэппинг проекты, проекты в виде графов, таймлайнов, инфографик и пр.
- По принципу типов данных, задействованных в проекте - текстовые, визуальные, пространственные данные или мультимедиа проекты, гибридные по своей сути.

Мы предлагаем следующую классификацию DH проектов, основанную на исследовательской мотивации:

- data driven проекты, в основе которых лежит анализ корпуса гуманитарных данных;
- проекты, решающие прикладные задачи исследователей, которые создаются и развиваются для всего коммьюнити исследователей в определенной области, фокусируясь на обогащении базы данных по конкретной тематике, делая эти данные более доступными для использования;
- проекты для общества, основная цель и миссия которых - популяризировать ту или иную узкую тему гуманитарного знания, сделав ее более доступной для понимания широкой публики, а именно всех тех, кто интересуется данной темой, но непосредственно не занимается ей в узко профессиональном ключе. На наш взгляд, данные проекты наиболее сложны в реализации, так как должны соблюдать в себе баланс между глубиной и точностью исследуемой темы и доступностью ее репрезентации в схематичном и более упрощенном формате. Поиск такого баланса - целое искусство.

Давайте более подробно рассмотрим 3 группы DH проектов, а также приведем примеры подобных проектов от DH центра Университета ИТМО и коллег из других исследовательских центров.

Data driven проекты

Одной из возможных мотиваций к созданию DH проекта могут стать сами данные. В таких случаях зачастую данные сами начинают диктовать потенциальное развитие проекта и провоцировать вас на новые и новые исследовательские вопросы. В зависимости от природы данных - текстовые, пространственные, визуальные или гибридные - будет выстраиваться исследовательский вопрос и строится гипотеза. Далее будут понятны методы, которые будут применяться для их анализа: уже упомянутые нами стилометрия, если интересно выяснить подлинность авторства того или иного произведения, анализ тональности, если важно посмотреть эмоциональную окраску исследуемых высказываний, и т.д.

Примером data-driven проекта может служить проект “Цифровые методы исследования семантической структуры англоязычных текстов национальных гимнов”[10], реализованный в рамках DH центра в Университете ИТМО.

В начале работы у команды проекта было главное предположение о взаимосвязи стран и народов между собой через общие темы в гимнах. Кроме того, было предположение о возможном наличии проблематичных тем в тексте гимнов, например, дискриминации, насилия и так далее.

Для исследования вручную был собран корпус англоязычных гимнов. Дальше, на базе собранного корпуса, начали выдвигаться более конкретные гипотезы. Например, сразу стало заметно, что многие гимны упоминают бога. Появилась гипотеза о том, что упоминание бога в гимне зависит от наиболее популярной религии в стране. Наши студенты провели корпусный анализ гимнов с дополнительным исследованием эмоциональной окраски каждого текста, подсчитывая частотность слов. Анализ и визуализация (в виде облака слов) проводились при помощи языка R. Следующим этапом работы стало создание датасета, в который исследователи добавили главенствующую религию, политический режим, упоминание бога и отсылки к национальным характеристикам в гимнах.

Также при работе с текстами использовались корпус-менеджеры AntConc и LancsBox. Данные на карте визуализировались в Tableau Desktop. Завершением работы стала публикация карт на сервер и интеграция на сайт в виде JS-элементов. Финальный результат можно посмотреть на нашем сайте dh.itmo.ru в разделе проектов.

В процессе сбора корпуса текстов англоязычных гимнов команда заметила тенденцию, что каждое государство через национальный гимн рассказывает часть своей истории, а также особенности и черты своих народов. Впоследствии все эти наблюдения были подтверждены анализом и внесены в финальную визуализацию.

Данный проект начинался как data-driven, а в результате стал проектом для широкой аудитории, который легко можно использовать в рамках курсов по истории, географии, культурологии, политологии и других гуманитарных дисциплин, на которых поднимаются вопросы сравнительного анализа разных культур. Красочная визуализация и репрезентация данных (см. Рисунок 21), в свою очередь, делает проект более интерактивным, позволяя студентам самостоятельно поработать с корпусом текстов и найти интересные тенденции и тренды. Например, на рисунке представлена визуализация превалирующей характеристики нации, отмеченной в текстах гимнов.

[10] http://anthemtalk.tilda.ws/

Рисунок 21 — Основные характеристики наций из текстов гимнов

Множество примеров data-driven проектов можно найти, полистав периодические издания, специализирующиеся на журналистике данных, например, Pudding. Часто подобные исследования выходят в New York Times или РБК. В данном случае отдельная статья представляет собой целостный data driven проект. Хотя методология таких исследований может отличаться от научной и допускать больше спекуляций, а сформулированные к данным вопросы могут быть не столь глубокими, как в полноценных научных исследованиях, подобные проекты зачастую отличаются качественными и выразительными визуализациями и безусловно заслуживают внимания.

Примером data driven журналистского исследования может считаться статья Джулии Силж “She Giggles, He Gallops”, в которой она анализирует женские образы в американском кинематографе[11]. К слову, автор этого исследования - создатель библиотеки tidytext для языка программирования R, предназначенной для майнинга текстовых данных.

Набор данных для проекта включал 1966 сценариев фильмов, выпущенных в период с 1929 по 2015 год; большинство из них относятся к 1990 г. и позже. Каждый сценарий был обработан для извлечения только описаний действия и ремарок, исключая диалоги. Затем были определены все биграммы, в которых первым словом было «он» или «она».

Затем было рассчитано логарифмическое отношение шансов, чтобы найти примеры, которые демонстрируют наибольшую разницу в характере действий, которые совершают мужские и женские персонажи. После чего была предпринята попытка исследовать зависимость в описании действий мужских и женских персонажей авторами разного пола. Пол писателей определялся по биографиям, фотографиям и именам с сайта IMDB.

11

Вот к каким выводам пришла исследовательница. Во-первых, на киноэкранах женские персонажи чаще всего прижимаются, хихикают, визжат и всхлипывают, в то время как мужские персонажи закрепляют что-то ремнём, скачут, стреляют и издают вопли. На рисунке 22 показаны по 10 самых популярных “женских” и “мужских” глаголов.

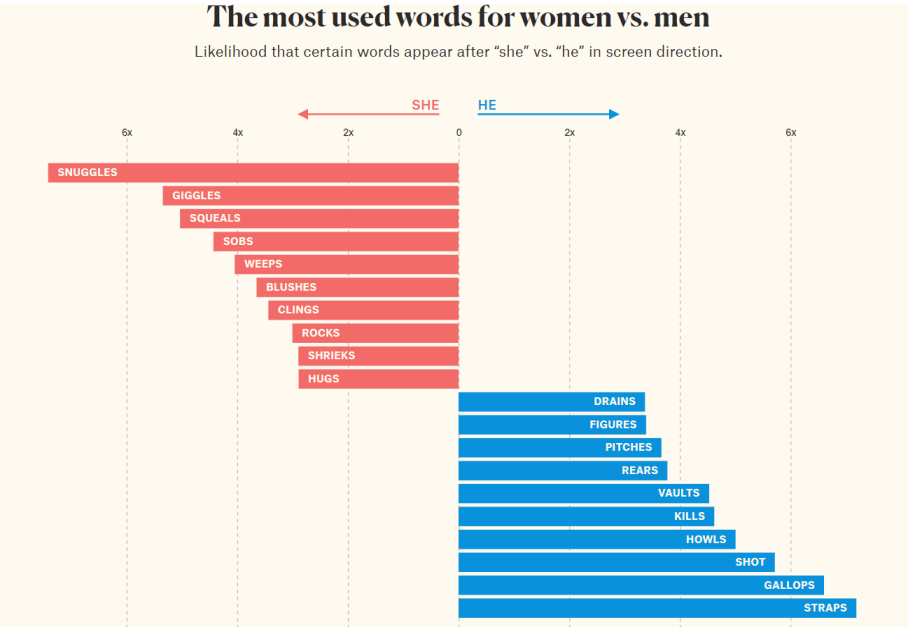


Рисунок 22 — Вероятность появления определённых глаголов в описаниях женских и мужских действий.

Во-вторых, что касается анализа зависимости в описании действий мужских и женских персонажей от пола автора сценария, удалось прийти к следующим заключениям. С одной стороны, в большинстве случаев пол писателя не имеет значения. По сравнению с мужчинами женщины чаще ахают, торопятся, улыбаются, сомневаются и что-то мешают (в основном во время готовки), независимо от того, мужчина или женщина пишет писатель. Мужчины же с большей вероятностью будут ломать предметы, доставать оружие, усмехаться, подмигивать, указывать и говорить. Однако при описании персонажей противоположного пола и мужчины, и женщины используют больше романтически и сексуально окрашенных слов, например «целовать», «поглаживать», «обнимать» и т.д.

Более подробное описание методов исследования и использованных инструментов приведено в справочном пособии за авторством Джулии Силж[12]. В этой книге вы также можете найти больше примеров data-driven исследований.

[12] Silge J., Robinson D. Text mining with R: A tidy approach. – " O'Reilly Media, Inc.", 2017.

Исследователи для исследователей

Из огромного ряда проектов в сфере DH также ярко выделяется еще одна группа проектов — это прикладные проекты, ставящие своей целью сделать данные более доступными для последующих исследований. Можно назвать такую мотивацию “проекты от исследователей для исследователей”.

В каком-то смысле подобные проекты тоже являются data driven - исследователи берут еще не обработанные и не пригодные к анализу данные, структурируют их, приводят в пригодный для анализа вид и предоставляют доступ другим ученым для их дальнейшей работы. Ключевое различие от предыдущей группы проектов заключается в том, что в большинстве случаев гипотезы в таких проектах служат не сужением фокуса проекта, а наоборот, расширением исследуемого поля. Чем больше гипотетических исследовательских проблем команда сможет затронуть своим проектом, тем полезнее ее ресурс станет для дальнейших исследователей.

Часто такие проекты осуществляются совместными усилиями исследовательских центров и учреждений, которые осуществляют хранение этих данных - например, это могут быть культурные институции, которые в англоязычных странах принято обозначать аббревиатурой GLAM: это галереи, библиотеки, архивы и музеи.

Например, проект Digital Panopticon[13], который стал возможным благодаря совместным усилиям исследователей из университетов Оксфорда, Ливерпуля, Шеферда, а также при поддержке Национальных архивов Великобритании.

Этот проект объединяет наборы данных, связанных с темой уголовного правосудия и переселения заключённых из Великобритании в Австралию. Главной целью исследователей было изучить влияние различных видов уголовных наказаний на жизнь 90 000 человек, осужденных в Олд-Бейли в период с 1780 по 1925 год.

Используя методы организации и визуализации данных, этот проект объединяет уже широко известные среди исследователей большие наборы данных (Old Bailey Online, London Lives и Founders and Survivors) с вновь оцифрованными данными, которые не были доступны ранее. Проект охватывает данные о всех лицах, осужденных в Олд-Бейли в период между уходом Первого флота в Австралию (1787 г.) и до смерти последнего переселенца из Лондона в Австралии в начале 1920-х годов.

В результате удалось собрать и опубликовать данные, которые показывают относительное влияние различных видов наказания на сопротивление преступным действиям, состояние здоровья осужденных,

[13] https://www.digitalpanopticon.org/

рассмотреть возможности трудоустройства и семейную жизнь осужденных в долгосрочной перспективе. С помощью инструментов визуализации, доступных на платформе, пользователи могут создавать проекты, на интересующие их темы. Например, визуализация на рисунке 23 показывает популярные темы татуировок заключенных в зависимости от пола.

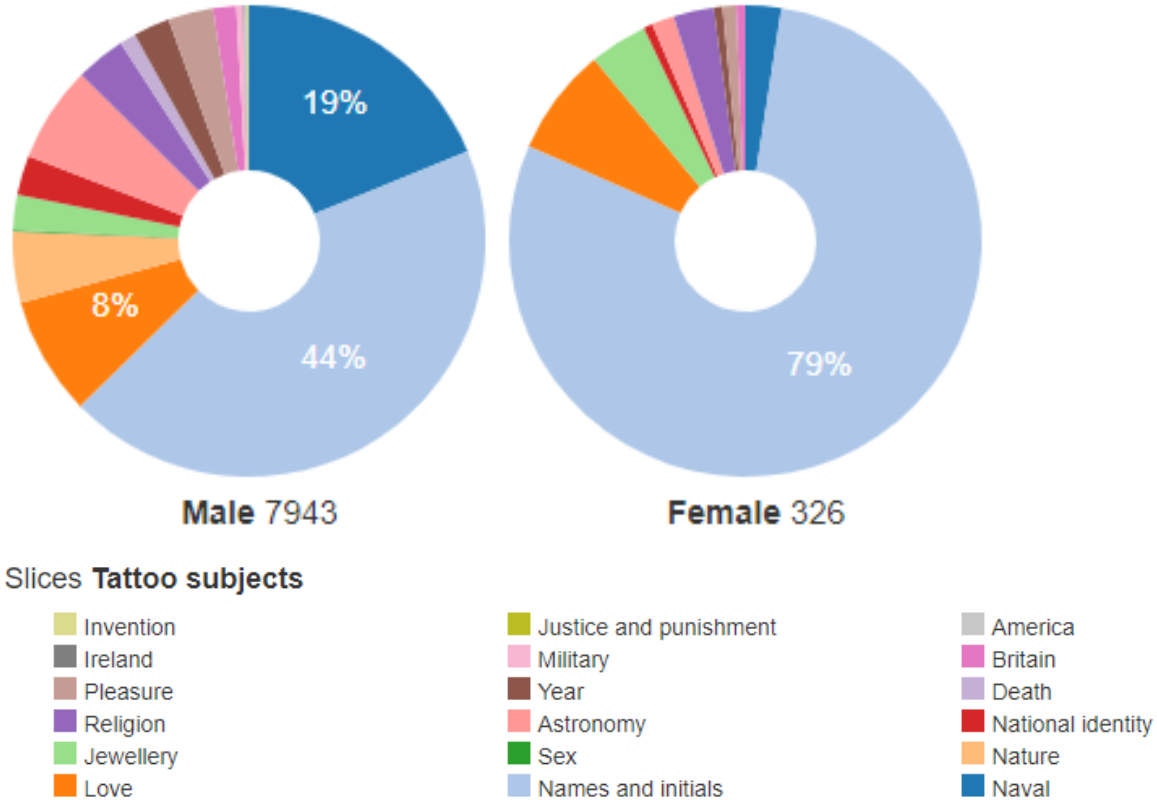


Рисунок 23. — Тематика женских и мужских татуировок осужденных

Суть проекта Dracor.org[14](см. Рисунок 24) заключается в том, что он объединяет корпуса драмы (текстов пьес) на разных языках, в т.ч. на русском. Все собранные тексты снабжаются общей стандартизированной разметкой в формате XML, согласно стандарту TEI, который используется и во многих других проектах, в которых задействованы корпуса текстов. Размеченные тексты становятся пригодными для “удалённого чтения”. На рисунке показана информация о самых обширных корпусах драмы, собранных на сайте.

[14] https://dracor.org/

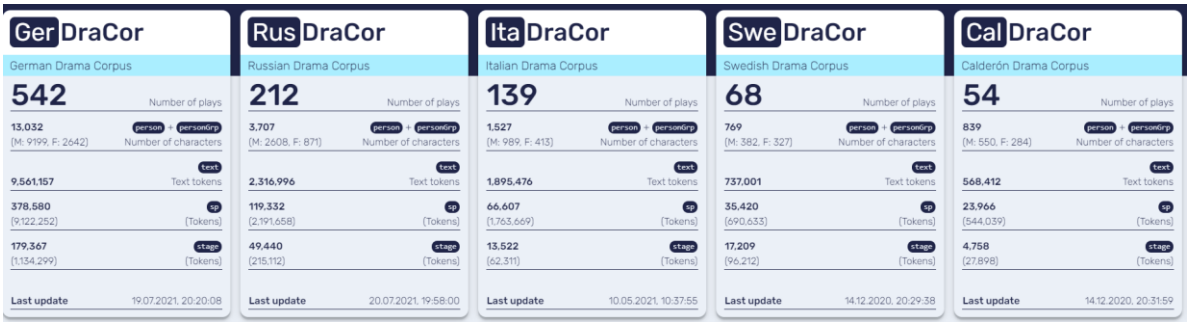


Рисунок 24 — Основная информация о самых объёмных корпусах драмы в архиве.

Конечно же, работу исследователей упрощает сам формат жанра, так как в некотором роде пьесы уже размечены: реплики следуют за именем персонажа, описания и авторские ремарки выносятся отдельно и так далее, однако для машиночитаемости текста необходимо произвести стандартизированную XML разметку.

Стандарт TEI подразумевает строго определённые теги для разных элементов драматического текста. Например, говорящий размечается тегом <speaker>, реплика — тегом <sp> (то есть speech), сценическая ремарка (e.g. «входит») — тегом <stage>:. Другим несомненным плюсом XML-разметки является древовидная иерархия — она используется для кодирования общей структуры пьес: действия, в них явления, в них сцены и так далее.

Полученные данные можно использовать для различных исследований. Например, применения сетевого анализа — ведь каждая пьеса может быть представлена в виде социальной сети персонажей, объединенных одними сценами. В ходе данного исследования были выделены разные типы персонажей. Одни из наиболее интересных типов — персонажи-посредники. Это персонажи, у которых не слишком много прямых контактов, но через которых идет много связей между отдельными группами персонажей (или говоря в терминах теории графов, высокая betweenness centrality при не самой высокой степени (degree centrality) узла). Оказывается, что такими персонажами становятся преимущественно всевозможные “двойные агенты”, тайные посланники и т.п.

Помимо прочего, платформа также оснащена инструментом визуализации, с помощью которого пользователи могут рассматривать пьесы под новым углом. Например, на рисунке показан граф социальных связей комедии А. С. Грибоедова “Горе от ума”(см. Рисунок 25).

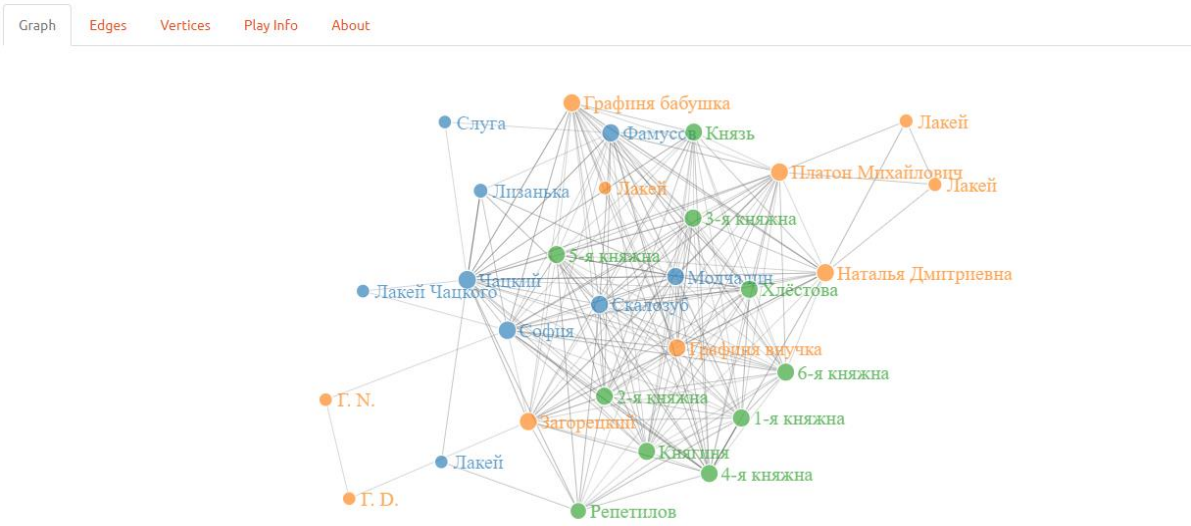


Рисунок 25 — Граф социальных связей пьесы “Горе от ума”

Другим примером проекта, направленного на сбор и подготовку данных, можно считать проект, реализованный DH центром Университета ИТМО - St. Retrospect. Данная инициатива ставит своей задачей создание единой интерактивной онлайн-среды, содержащей историко-культурологические метаданные о культурно значимых локациях Санкт-Петербурга. Проект сам по себе состоит из двух частей - веб-платформы с базой данных, снабженной пользовательским интерфейсом, открытым API и веб-интерфейса с возможностью поиска и визуализации информации о связи исторических персон и локаций на карте города; а также мобильного приложения, которое призвано популяризовать собранную информацию и представить её в виде интерактивных маршрутов широкому кругу пользователей. В данном случае мы говорим с вами про веб-платформу, которая призвана систематизировать и визуализировать краеведческие данные об истории развития Санкт-Петербурга.

Главная цель платформы, которую поставили разработчики, — стать площадкой, на которой исследователи и культурные институции смогут транслировать знание о городе широкому кругу пользователей в интересном и интерактивном формате. Платформа объединяет обширную базу данных об исторических локациях города с удобным веб-интерфейсом и позволяет осуществлять интеллектуальный поиск по локации, персоне или определенной тематике.

Одной из центральных задач в рамках работы над платформой является конструирование репрезентативной базы данных, на основании которой можно было бы объединить разрозненные и неоднородные данные о локациях, персоналиях, событиях, и другие знания как о материальной, так и нематериальной культуре, и о городе в целом в единый гипертекст с проработанной и понятной системой навигации. Альфа-версия сайта с ограниченным функционалом была запущена еще в декабре 2019 и сейчас

активно дорабатывается. Так как проект сочетает в себе и исследовательскую, и просветительскую составляющие, он может служить мостиком к следующей группе проектов, нацеленных на популяризацию гуманитарного знания с помощью цифровых средств.

Итак, все перечисленные выше проекты в первую очередь становятся полезными для будущих исследователей, поскольку предоставляют доступ к ранее недоступным данным, систематизируя их и делая более открытыми для работы исследователей. Следующий шаг - это трансформация исследовательских данных в общественно значимые проекты, которые делают гуманитарное знание понятным и доступным широкой публике. Именно о такого рода проектах поговорим далее.

Просветительские проекты

Третья возможная мотивация, которую можно выделить, пожалуй, встречается наиболее часто и ставит своей задачей помочь обществу или наладить коммуникацию с обществом. Такие проекты ориентируются на потребности небольших коммьюнити, городов, стран. Чаще всего цель проекта близка к образовательной - исследователь хочет объяснить обществу важные открытия академического мира, которые влияют на повседневную жизнь человека и о которых важно знать, в том числе, и людям, далеким от академии.

Одним из примеров такого проекта может быть уже упомянутый проект DH Center Университета ИТМО St. Retrospect, а точнее, его часть в виде мобильного приложения “Que.St”[15]. Мобильное приложение "Que.St" направлено на геймификацию процесса изучения истории Санкт-Петербурга. Помимо маршрутов и квестов, приложение также будет включать популярные в компьютерных играх механики: очки опыта, коллекционные предметы, достижения и др.

Одна из главных задач проекта – найти баланс цифрового и аналогового, поставить информационные технологии на службу культуре, говоря с пользователем на одном языке, но в то же время, не забывая о просветительской составляющей и помогая сформировать новый запрос на качественную культурную информацию.

Современные технологии, положенные в основу проекта «St. Retrospect», открывают совершенно новую перспективу для музеев, библиотек, архивов и других культурных институций города - распространить свои нарративы в городское пространство. Пользователи смогут также активно участвовать в процессе производства нового знания, создавая собственные маршруты, используя информацию, накопленную в нашей базе данных. Единая платформа будет способствовать консолидации

[15] https://quest.dh-center.ru/

с сотрудничеству культурных организаций города и заинтересованных жителей, складыванию совместно накопленных знаний в единую картину, формируя и дополняя онлайн-облик Санкт-Петербурга. В перспективе платформа может быть масштабирована на другие города России и мира. На рисунке 26 представлены два пользовательских экрана приложения.

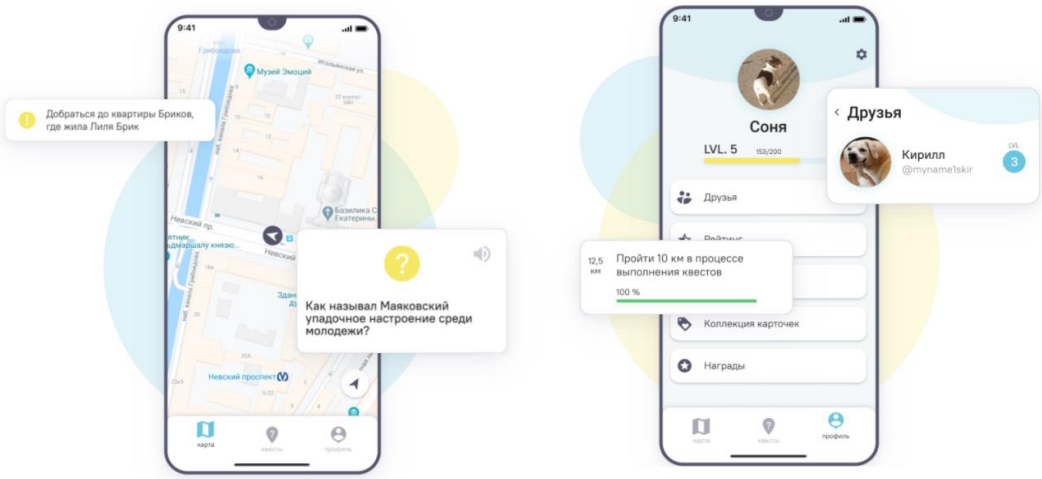


Рисунок 26 — Основные экраны приложения Que.St

Другой проект, реализованный в рамках DH Center, ИТМО, Ретроспектива ГУЛАГа, также может быть ярким примером образовательного проекта для общества. Суть проекта – освещение исторических событий с помощью коротких, но емких историй в формате ретроспективной визуализации уникальных данных, предоставленных по запросу музеем истории ГУЛАГа. Пользователь получает целостную картину репрессивной системы посредством визуализации микро- и макроуровня истории.

Главная задача при подаче материала заключалась в создании эмоционального отклика путем использования современных практик дизайна, игровых механик и репрезентативности информации. Всё это делается ради создания ощущения сопричастности у пользователя, чтобы человеческие судьбы не оставались просто пикселями на экране. Например, в процессе раскрытия конкретной темы пользователь может взаимодействовать с тщательно отобранным материалом в нужной последовательности - документы, повлиявшие на судьбы миллионов людей, умозаключения лидеров, приведшие к серьезным изменениям в политике, и многое другое. Таким образом выстраивается логическая последовательность событий, демонстрирующая, как тот или иной факт повлиял на отдельного человека и страну в целом, а также наше настоящее.

Не менее важный образовательный проект PubHistory[16], посвященный сравнению репрезентаций событий в медиа-пространстве с реальной историей, был реализован исследовательской группой из Пермского филиала Высшей школы экономики. Авторы проекта выбрали отечественные и зарубежные видеоигры, кинофильмы, сериалы и провели анализ конструкции прошлого. Все истории собраны на портале PubHistory, а также транслируются в соцсетях.

Исторические фильмы и книги — важнейший источник представлений о прошлом, событиях, исторических деятелях и просто об образе жизни в ту или иную эпоху. Но они же становятся источником многих исторических мифов и заблуждений. В рамках проекта эти мифы разобраны студентами-историками с точки зрения исторических источников и исследований об этих периодах и событиях. На текущий момент на сайте проекта собраны 25 таких разборов, и сайт постоянно пополняется новыми проектами.

Тест 6

Совместите название типа проекта, утверждение, характеризующее данный тип проектов и пример подобного проекта.

I. Data-driven	A. подобные проекты часто призваны стимулировать больше исследований по данной теме	1. Dracor.org
		2. PubHistory
II. Для исследователей	B. такие проекты делаются, чтобы привлечь внимание широкой общественности к какой-то теме	3. She giggles, he gallops
		4. National Anthems
III. Просветительские	C. В таких проектах гипотезы строятся на основании уже собранных	5. Que.St

[16] http://pubhistory.tilda.ws/

	данных	
		6. Digital Panopticon

Этапы работы над проектом

Ключевым результатом исследования в сфере DH чаще всего становится готовый проект. Даже самая простая визуализация анализа данных, представленная с контекстным сопровождением и опубликованная в открытом доступе, может считаться готовым DH проектом.

Давайте рассмотрим ключевые стадии разработки проекта: поиск проблемы, планирование, поиск информации, создание проекта, презентация продукта. Давайте более подробно рассмотрим каждый из них.

1. **Поиск проблемы.** На стадии поиска проблемы нужно хотя бы в общих чертах понять тему и цели будущего проекта. В такой междисциплинарной сфере, как Digital Humanities, исследователю важно понимать, какие технические методы и какие подходы гуманитарных наук понадобятся в ходе исследования. Если один исследователь не обладает достаточной квалификацией, ему понадобится собрать рабочую группу - найти специалистов в сферах, которые будут задействованы в будущем проекте, или обратиться за консультационной поддержкой к культурной институции или в образовательное учреждение.
2. **Планирование.** Следующий этап разработки проекта - планирование. На данном этапе важно определить возможные источники информации для будущего проекта и исследования. Помимо этого, необходимо определить, каким образом собранная информация будет анализироваться, будут ли это готовые программные решения low code/no code), или этого будет недостаточно для достижения поставленной цели и потребуется писать код для создания более кастомизированного решения. Кроме того, с самого начала нужно определиться, в каком виде планируется представлять результаты исследования, как именно будет выглядеть проект. Будет ли это сайт, мобильное приложение, научная статья, и так далее. Также важно определить критерии успеха и временные рамки различных стадий проекта, оценить риски и так далее. Более того, если проект командный, на данном этапе также важно распределить сферы ответственности, обязанности и задачи, а также режим работы и промежуточные встречи.

3. **Сбор данных.** После планирования проекта можно приступать к, пожалуй, самому энергозатратному шагу - исследованию и сбору информации. На данном этапе важно, во-первых, воспользоваться максимальным количеством доступных инструментов поиска данных для исследования. К этому этапу нужно подходить очень ответственно - ведь от полноты собранных данных напрямую зависит качество исследования. Существуют ли готовые наборы данных? Если да, достаточно ли полная информация в них представлена? Существуют ли открытые API, с помощью которых вы можете найти необходимые вам данные? Может быть, данные можно собрать самостоятельно с одного или нескольких веб-сайтов? Достаточно ли собранных данных для исследования? Если да - переходите к следующему шагу.

4. **Создание проекта** включает в себя такие шаги, как анализ собранных данных, подготовка визуализации данных, а также формирование выводов. Выводы могут касаться не только непосредственного исследования, но и вашего исследовательского опыта в рамках данного проекта: все ли у вас получилось, что именно не получилось реализовать и почему, и так далее. В результате каждого выполненного вами проекта вы развиваетесь не только как исследователь, но и как проджект-менеджер или участник команды. Если что-то не получилось реализовать в полной мере - не отчаивайтесь, ведь многие DH проекты - это компромисс между изначальной задумкой и объективными ограничениями, которые накладываются недостаточностью или неполнотой оцифрованных данных или возможностями инструментов.

5. **Презентация проекта** - это финальный этап, который включает в себя подготовку всех отчетов о проведении исследования, о его результатах. Кроме того, в случае проекта в сфере Digital Humanities вам важно понимать, в каком виде вы его будете презентовать и где. В зависимости от специфики проекта вы можете выбрать как академическую презентацию (конференция, грант, проект для защиты диплома или диссертации), так и коммерческое продвижение.

Задание

Итак, мы рассмотрели несколько примеров DH проектов различной направленности. Для начинающего исследователя очень важно научиться критически оценивать работы коллег в своей предметной области, поэтому вашим следующим заданием будет написать критическое эссе с разбором DH проекта на интересующую вас исследовательскую тематику. При рассмотрении проекта необходимо уделить особое внимание следующим вопросам:

1. Основная информация о проекте: кем он осуществляется, главные цели и задачи проекта и т.д.
2. Какие данные представлены в проекте, являются ли они исчерпывающими?
3. Какова методология исследования? Отвечают ли методы поставленным задачам?
4. Сложности и ограничения: с какими сложностями столкнулись авторы исследования, какие ограничения и недостатки вы можете выделить в представленном исследовании?
5. Какие ещё исследовательские вопросы можно поставить, основываясь на приведенном исследовании, чем его можно дополнить?

Минимальный объем эссе - 500 слов.

Рекомендуемая литература

1. Project Management for the Digital Humanities[Электронный ресурс], способ доступа: https://scholarblogs.emory.edu/pm4dh/
2. Development for the Digital Humanities[Электронный ресурс], способ доступа: http://devdh.org/

Глава 4. Методы

Как мы уже отмечали, развитие DH идет рука об руку с развитием технологий. К примеру, технологии обработки естественного языка сегодня шагнули далеко вперед, и немалая заслуга в этом принадлежит машинному обучению, применяемому, в частности, для понимания текстов. Это в свою очередь дало огромный толчок развитию DH проектов в области анализа и представления текстовых данных.

Обработка естественного языка

Обработка естественного языка или NLP (Natural language processing) — область, находящаяся на пересечении компьютерных наук и лингвистики. Ее цель заключается в обработке и “понимании” естественного языка. В нее входят такие задачи, как извлечение знаний из текстов, классификация, распознавание и генерация речи, машинный перевод, определение тематик и другие операции, направленные на понимание текстов и составление баз знаний для дальнейшей манипуляции с данными. Первым, кто рассмотрел задачи, связанные с обработкой

текстов и грамматикой естественного языка, был американский лингвист Ноам Хомский. Он описал ключевую парадигму для компьютерной лингвистики — контекстно-независимую грамматику. Ее использовали для обработки текстов и создания деревьев разбора, которые в дальнейшем применяли для составления логического представления знаний с использованием свода правил и заранее подготовленного лексикона. Дальше логическое представление могло быть использовано для разного рода операций вроде проверки утверждений или ответов на вопросы.

Ряд работ по статистической лингвистике был сделан в начале 90-ых годов, что позволило активно развиваться методам машинного обучения. Появились алгоритмы тематического моделирования, IBM занялась проектом по статистическому машинному переводу. Наконец, были заложены основы глубокого обучения (deep learning), которое только недавно стало активно применяться в исследованиях и индустрии. Интерес к использованию технологии обусловлен прогрессом в области высокопроизводительных систем и появлением больших объемов данных, используемых для обучения. Впоследствии появилась модель лексикализованной вероятностной грамматики, что увеличило точность грамматического разбора до 93%. Сейчас алгоритмы доступны широким массам исследователей.

Отметим, однако, что в DH более популярны не столько методы обработки естественного языка, сколько то, что называется анализом данных. Это широкий спектр инструментов, помогающий увидеть закономерности и тренды в данных. Анализ данных включает в себя расчет корреляций, регрессии, кластерный анализ и прочее. Нельзя обходить стороной и визуализацию, которая позволяет наглядно показать выводы и результаты исследования.

NER \\ Извлечение именованных сущностей и отношений

Одной из самых сложных и неоднозначных проблем, которая может встретиться во время работы с текстовыми данными, является извлечение именованных сущностей (Named-entity recognition, NER) – слов, обозначающих предмет или явление определенной категории.

Допустим, у нас есть какой-то текст (или набор текстов), и данные из него нужно ввести в базу данных (таблицу). Классические именованные сущности, такие как имена, локации, отношения, могут соответствовать строкам такой таблицы или же служить содержанием каких-то ячеек. Соответственно, чтобы правильно заполнять таблицу, нужно перед этим выделить в тексте те данные, которые мы будете в нее вносить. При этом в текстах у нас могут встречаться просто объекты мира: стол, стул, дерево, корабль, которые могут либо как-то взаимодействовать между собой в тексте, либо просто упоминаться. Это сущности в целом. Именованными сущностями называются такие объекты, у которых есть конкретное

индивидуальное обозначение: имя и фамилия человека, адрес, название компании, имя корабля. Когда мы говорим просто “корабль”, то это просто сущность, когда говорим “корабль Мария”, то это уже — именованная сущность. Пример работы алгоритма можно видеть на рисунке 27 ниже, а также можно опробовать модель самостоятельно, используя ссылку:

https://explosion.ai/demos/displacy-ent

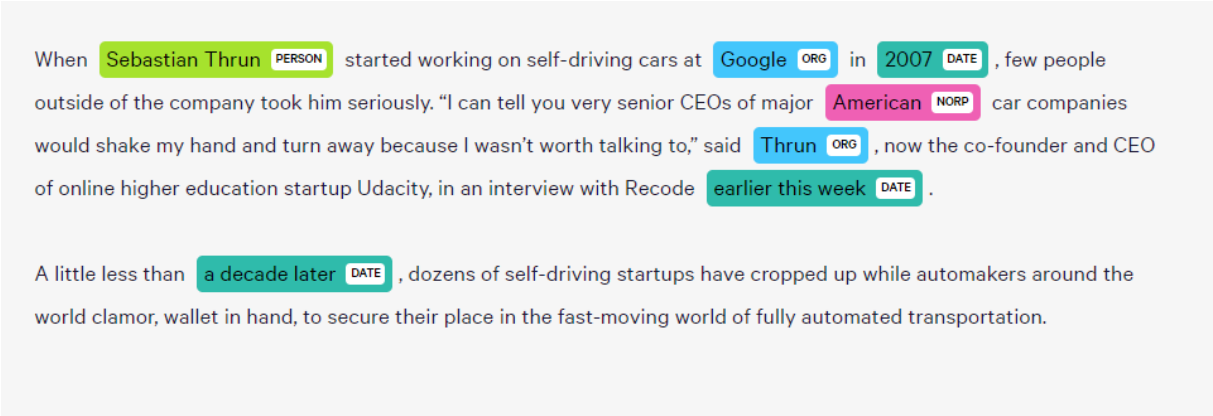


Рисунок 27 — Пример работы алгоритма по извлечению именованных сущностей

Извлечение именованных сущностей необходимо для решения нескольких задач:

1. Технология позволяет лучше понять текст. Так, мы можем выделять и анализировать отдельные абзацы текста с интересующими нас именованными сущностями.
2. Обычно сущности являются надежными и устойчивыми коллокациями, что может быть важным для определенных задач. Допустим, у вас есть название именованной сущности, и, какой бы она ни была, скорее всего, она непрерывна, и для сохранения смысла все действия с ней нужно совершать как с единым блоком. Например, перед вами стоит задача перевести название сказки «*Красная шапочка*» на французский язык. Лучше перевести его единым куском, а не разбивать на несколько не связанных друг с другом фрагментов. Умение определять коллокации полезно и для многих других задач — например, для синтаксического парсинга.
3. Наконец, извлечение именованных сущностей используется для разрешения проблемы местоименной анафоры, которая позволяет нам понять, на какой элемент текста ссылается местоимение.

Важно отметить, что задача NER традиционна и хорошо изучена, особенно для английского языка. Существует большое количество как коммерческих, так и открытых решений.

Для русского языка таких программ меньше, да и сделать анализ сложнее из-за более сложной семантики и морфологии русской речи. К тому же, если искусствоведу или культурологу нужно найти все адреса художника в его автобиографии, изданной в 1937 году, то проблема становится еще сложнее.

Схожую проблему решала наша интердисциплинарная команда DH центра ИТМО, когда пыталась проанализировать старинные тексты для последующей репрезентации полученной базы данных на карте Санкт-Петербурга в рамках работы над масштабным проектом St. Retrospect.

Для того, чтобы извлечь необходимые нам именованные сущности - имена знаменитых петербуржцев, локации и то, что их связывало (отношения), вначале мы взяли более или менее все крупные существующие решения по извлечению именованных сущностей русского языка (включая библиотеку «Наташа») и применили для наших исторических текстов. Выяснилось, что если для современных текстов эти модели дают до 95% качества, то на наших текстах (книгах, статьях, заметках), которым было 60 и более лет, результат составил в районе 70-72%. Причем чем старше текст, тем хуже был результат. Как оказалось, главная проблема для алгоритма, нетренированного на современных текстах, заключается в старых именах вроде Феодоры, Февронии, Иоланты или Мазепы. Именно на них точность работы проседала сильнее всего. Мы предложили дополнительные эвристики, которые смотрят на то, что извлекалось, проводят частотный анализ и на основе анализа улучшают характеристики извлечения. Фактически мы создали дополнительный блок постобработки, который смотрит на частотность определения тех или иных слов и принимает решение относительно того, не сделал ли ошибку основной алгоритм. В результате точность распознавания повысилась до 78-79%. И это только начало пути, на котором предстоит огромная работа по увеличению точности распознавания и поиска оптимального соотношения ручного труда и применения компьютерных алгоритмов.

Дистрибутивная семантика

Дистрибутивная семантика представляет область лингвистики, занимающейся оценкой степени семантической близости между лингвистическими единицами на основании их распределения в корпусе текстов. Из определения выше мы выводим, что значение слова — это в каком-то смысле просто сумма всех тех контекстов, в рамках которых мы его встречали в тексте. Таким образом, чтобы научить компьютер или искусственный интеллект "понимать" семантику, нам нужно построить модель этих контекстов на достаточно большом текстовом корпусе. Получается, что если два слова в корпусе постоянно встречаются в одном и том же контексте, то эти слова означают ровно одно и то же.

Лексическая единица определяется через вектор, в котором значения показывают частоту совместного употребления интересующей нас единицы с другими словами определенного корпуса. Если у каждой лексической единицы есть вектор, то следующий шаг в работе — это оценка степени семантической близости этих слов. К сожалению, в традиционной дистрибутивной семантике размер векторов получается весьма большим.

Сейчас же стали популярны нейронные модели, в которых представлением слова выступает сжатый вектор (по-английски он называется *embedding*). Для него максимизируется сходство с ближайшими векторами и минимизируется с теми, кто соседями не является. Это помогает быстро получать компактные репрезентации слов, которые демонстрируют отличное качество на стандартных семантических метриках.

Word2vec — библиотека, включающая в себя модели на основе нейронных сетей, используемые для трансформации слов в векторные представления. Пакет программ был представлен сотрудниками (Mikolov) Google в 2013 году. Инструмент интересен тем, что каждое слово представляется вектором из чисел в маленьком пространстве. Сначала векторам даются случайные значения, но чем дольше происходит процесс обучения, тем точнее подбирается вектор для слова, который будет схож векторами других слов, используемых в схожем контексте (см. Рисунок 28). Для работы с контекстом существует два алгоритма: Skip-gram и CBoW. Первый использует контекст для определения текущего слова, второй использует слово для определения контекста вокруг.

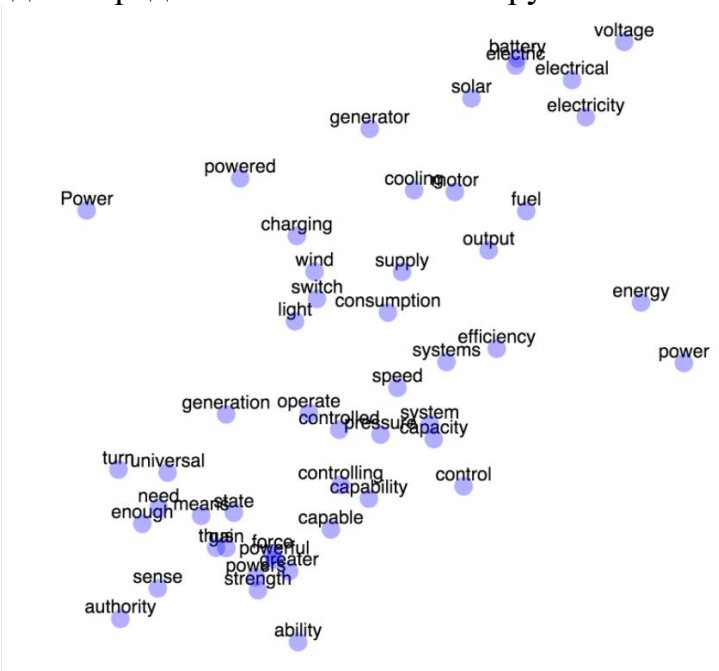


Рисунок 28 — Визуализация работы алгоритма Word2Vec

Векторные представления используют в большом количестве задач, где требуется сравнивать семантику слов. Это может быть машинный перевод, классификация и кластеризация текстов, система информационного поиска, определение тональности, оценка семантической близости

Тематическое моделирование

Еще одним популярным методом в DH является тематическое моделирование (topic modeling) — это способ построения модели текстовых корпусов, которая определяет, к каким темам относится каждый из документов.

Для чего это нужно? Прежде всего переход из пространства терминов в пространство найденных тематик помогает разрешать многовариантность значений — терминов, а также эффективнее решать такие задачи, как тематический поиск, классификация, суммаризация и аннотация текстовых корпусов и новостных потоков.

Тематическое моделирование как вид статистических моделей для нахождения тем, присутствующих в текстовом корпусе, нашло свое применение в таких областях, как машинное обучение и уже упомянутом нами обработке естественного языка, и конечно, цифровой гуманитаристике. В зависимости от поставленного исследовательского вопроса ученые используют различные тематические модели для анализа текстов, текстовых архивов документов, а также для понимания и последующей интерпретации трансформации тем в наборах документов. Напомним, что одной из самых важных задач в рамках тематического моделирования является подготовка данных или пре-процессинг. Почему он так важен? Как и в целом при работе с текстовыми корпусами, чем лучше и чище будут ваши данные, тем точнее будут результаты применения компьютерных алгоритмов. Скажем, вы хотите поработать с текстовым корпусом научных статей на русском языке и понять, какие темы наиболее часто затрагиваются учеными в определенной научной области в определенный период времени. Для этого вам нужны будут тексты статей - обычно это pdf-файлы, которые, во-первых, представлены в машиночитаемом формате, а во-вторых, не содержат в себе то, что в последующем может усложнить работу алгоритма - например, ссылки, сноски, представление текста статьи в двух-трех колоночном виде, аннотацию на английском языке и пр. Так как почти все современные научные статьи содержат ссылки, сноски, вкрапления английского языка и пр., то будьте готовы провести качественную предобработку данных, иначе тематическое моделирование будет работать очень плохо.

Принцип работы алгоритмов тематического моделирования можно описать следующим образом: алгоритм предполагает, что тема состоят из набора определенных терминов, которые встречаются в ней чаще в других

темах, а документы, в свою очередь, состоят из набора тем. Например,: «ученый» и «статья» встречаются чаще в документах про научную сферу деятельности, «программист» и «код» будут встречаться в документах о сфере IT, предлоги «и» и «в» будут встречаться в обеих тематиках. Обычно документ касается нескольких тем в разных пропорциях, таким образом, о документе, в котором 10 % темы составляют ученые, а 90 % темы — программисты, можно предположить, что слов про программистов в 9 раз больше.

Тематическое моделирование отражает эту интуицию в математической структуре, которая позволяет на основании изучения текстовых корпусов и исследования частотных характеристик слов в каждом документе сделать вывод, что каждый документ — это некоторый набор тем (см. Рисунок 29).

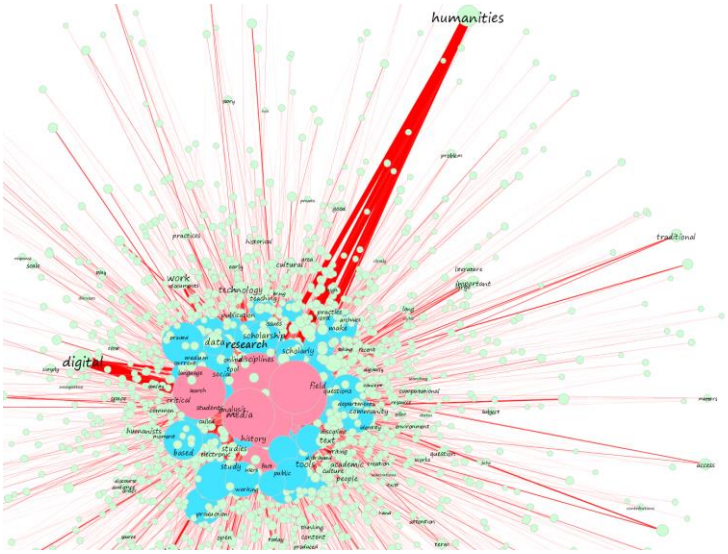


Рисунок 29 — Пример работы алгоритма тематического моделирования

Интересным примером исследования в области тематического моделирования на примере цифровых гуманитарных исследований был проект Клэя Темплтона, который сделал обзор работ по тематическому моделированию в гуманитарных науках, сгруппированных по синхроническому и диахроническому подходу. Синхронический подход позволяет выделить темы в некоторый момент времени, например, таким образом, получилось узнать, на какие темы писали блогеры в День цифровых гуманитарных наук в 2010 году.

Диахронический подход, в свою очередь, позволяет взглянуть на интересующую тему в ретроспективном ключе и проследить изменение тематик в интересующем корпусе текстов на определенном промежутке времени. Так, Томас Грифитс и Марк Стейверс использовали тематическое моделирование для обзоров аннотаций статей в журнале *Proceedings of the National Academy of Sciences или* PNAS, определив изменения в

популярности тем с 1991 по 2001 год. Дэвид Мимно, в свою очередь, использовал тематическое моделирование для анализа 24 журналов по классической филологии и археологии за 150 лет, чтобы определить изменения популярности тем и узнать, насколько сильно изменились журналы за это время.

Стилометрия

Мы уже ранее упоминали примеры DH проектов в рамках поиска авторства, акцентируя внимание на том, что их инициаторы были одними из родоначальников всего направления. Теперь давайте поговорим подробнее о том, как это делается.

За долгие годы ученые выработали научный метод изучения лингвистического стиля печатного текста и почерка – стилометрию, буквально измерить или выявить стиль. При исследовании стиля автора словарный срез текста изучается через два подхода.

Исследователи пытаются определить уникальные слова, используемые автором, которые могут быть явным признаком его неповторимого стиля. Человек с богатым словарным запасом обычно выражает свои мысли более емкими словами и фразами, наиболее близкими к описываемой ситуации, его речь четко построена и выразительна. Люди, обладающие небольшим словарным запасом, используют повторение одних и тех же слов, поэтому их устная и письменная речь выглядит проще.

Также по словарю текста можно судить об времени написания документа, так как все его слова уже должны были существовать и, скорее всего, входили в активный словарь. Характерные слова могут означать принадлежность автора к определенной группе: можно судить о профессии автора, уровне образования, о географическом месте, в котором он вырос или жил, об уровне культуры человека и т.д.

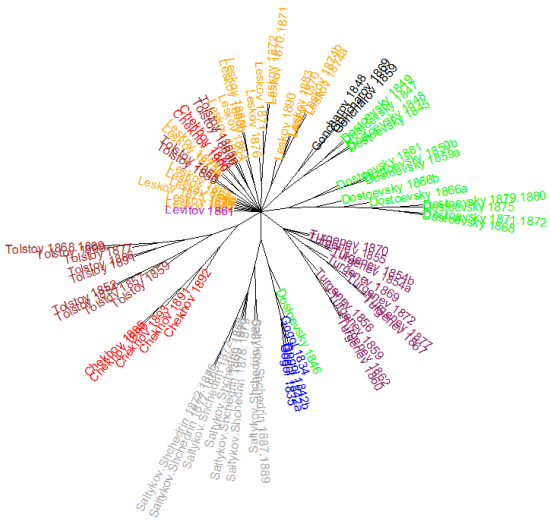


Рисунок 30 — Пример определения авторского стиля с помощью пакета stylo

Данный прием автороведческой экспертизы более других соотнесен с практикой исследования идиостиля автора в филологических работах (см. Рисунок 30). Однако при таком подходе возникает ряд проблем и вот каких:

1. У автора, как и у текста, может не быть ярких характерных особенностей лексикона.
2. Некоторые особенности лексикона могут быть намеренно введены в текст автором для осуществления художественного замысла.
3. Поиск отличительных черт авторского лексикона во многом носит субъективный характер и в большой степени зависит от интерпретации исследователя.
4. Набор лексикона сильно зависит от темы и жанра, что приводит к сложностям при проведении экспертизы разножанровых текстов одного автора.
5. Необходимость привлечения к статистическому анализу значительного по объему корпуса текста. Из-за этого подход чаще используется при определении авторов художественных текстов.

Однако с приведенными выше ограничениями прием может быть использован и используется в столеметрии. Главным плюсом такого подхода является тот факт, что корпус анализируемых текстов не требует особой лингвистической разметки.

Вернемся ко второму подходу автороведческой экспертизы, в основе которого лежит обращение к лексическому уровню текста для выявления частотности слов. В этом случае слова, сгруппированные по частоте встречаемости в текстах, используются в качестве признаков. Группы слов, встречающиеся в тексте один или два раза, обычно являются знаменательными. К примеру, термины из группы служебных слов отличаются высокой повторяемостью в тексте, они не так важны для исследователей, хотя бывают случаи, когда стоит их учитывать.

Несмотря на то, что при работе с этими методами не требуется специальная разметка, определенную предобработку текстов все равно приходится проводить. Словоизменение, как правило, не учитывается при анализе словаря, но стемминг или лемматизация, которая особенно актуальна для русского языка, помогут добиться лучших результатов автороведческой экспертизы. Также можно проводить морфологический анализ с помощью библиотеки pymorphy2 и других, они помогут определить грамматический класс, нормальную форму слова и т.д.

Таким образом, задача стилометрии крайне интересна, но лучше решается только с опорой на разработанные специальные корпуса текстов и требует тщательной предобработки текстов для улучшения точности результатов, полученных при использовании компьютерных алгоритмов.

Анализ тональности

Анализ тональности – это определение полярности эмоциональной оценки автора текста к сущностям, личностям, вопросам, событиям, темам и их атрибутам. Проще говоря, он отвечает на вопрос “Как относится к этой теме автор?”, разделяя результаты на три группы: положительное, нейтральное, отрицательное.

Применение инструментов анализа тональности в культурной аналитике интересно для понимания и последующей интерпретации реакции сообщества на определенную ситуацию (будь то выборы, проведение массового мероприятия и пр.)

В ходе анализа тональности в тексте выявляются слова и выражения с положительной, нейтральной, отрицательной окраской. Для того, чтобы распознать тональность текста, человек использует не только свои лингвистические знания, но и контекст, в котором он находится. Для компьютера все немного сложнее, так как его интерпретация не зависит от контекста, а только от языковой модели, которую он с легкостью распознает, так что точность анализа тональности текста, сильно зависящего от контекста, оставляет желать лучшего. К примеру, слово “дешевый” может иметь несколько смыслов: цена, что несет положительный контекст, или качество, что говорит о неудовлетворенном потребителе. Вдобавок слово может не нести оценочного суждение вовсе, а просто констатировать факт. В связи с этим инструменты были доработаны и уже умеют работать с текстами разной тематики или конкретного домена. Другой показатель, который нас интересует – это степень эмоциональной окраски, чтобы понимать, как сильно автору что-то нравится или не нравится. Возможность определять степень выраженности той или иной тональности в текстах имеет большое значение в процессе принятия решений.

Какие основные подходы выделяют при анализе тональности?

Мы уже обсудили, что обычно тексты классифицируют на три категории: позитивные, нейтральные и негативные. Следующим шагом в развитии технологии стали попытки определить эмоциональное состояние, связанное с текстом, в частности, счастье, печаль, страх и злость. Аспектный анализ тональности — это подвид анализа тональности, задача которого связана с определением отношения к отдельным аспектам общей темы. Так, подходы к анализу тональности можно разделить на три вида:

1. **Подходы на основе правил (rule-based).**

   Обычно исследователи сами пишут правила, по которым определяется тональность текста. В основу этих правил ложаться ключевые слова, выражающие эмоции, и их

взаимодействие с другими словами в тексте. Проблема этого подхода заключается в том, что они работают с текстами отдельных тематик, но перенос правил на другие темы может понизить эффективность.

2. **Подходы на основе машинного обучения**

Они используют автоматическое извлечение признаков из текста и применение алгоритмов машинного обучения. Автоматическое извлечение признаков из текста помогает ускорить работу, обрабатывать большое количество данных и переносить результаты обучения на другие тематики. Впрочем, для обучения модели также требуются размеченные данные, а часть работы исследователям придется делать вручную

3. **Гибридные подходы.**

В них объединяют подходы двух предыдущих видов. Комбинация методов позволяет добиться лучших результатов, но эти подходы наследуют ограничения алгоритмов, которые они в себя включают.

Ниже приведен пример работы алгоритма (см. Рисунок 31), также можно попробовать самому, перейдя по ссылке: https://demo.allennlp.org/sentiment-analysis/roberta-sentiment-analysis

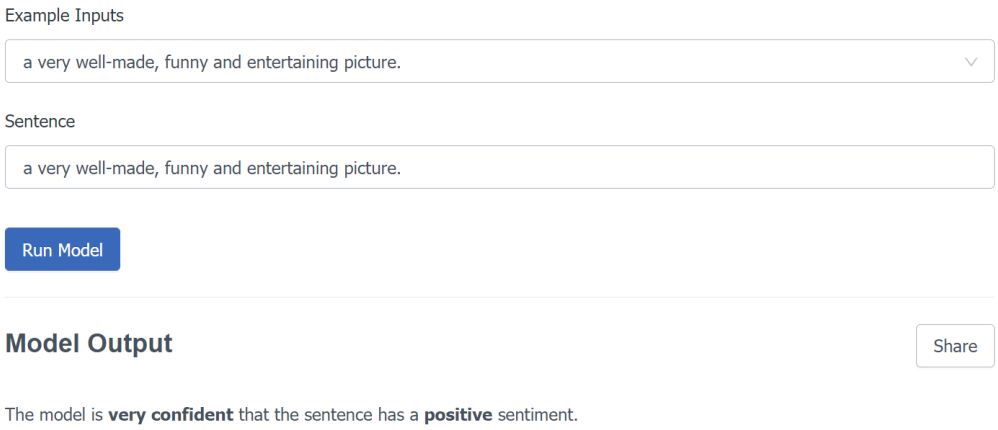


Рисунок 31 — Пример работы алгоритма по определению тональности

Словари настроений

1. RuSentiLex
2. LINIS Crowd
3. SenticNet
4. SentiWordNet
5. SentiWords

Тест 7

Вопрос 1. Какой алгоритм компьютерного анализа стал активно разрабатываться после предложенной американским лингвистом Ноамом Хомски парадигмы конктекстно-независимой грамматики?

A. BERT
B. OCR
C. NLP

Вопрос 2. Является ли Царскосельский лицей именованной сущностью?

A. Да
B. Нет

Вопрос 3. Что описывает вектор в традиционной дистрибутивной семантике?

A. Предложение
B. Лингвистическую единицу
C. Контекст

Вопрос 4. Какое "отношение" не принято выделять в качестве “окраски” при анализе тональности текстов?

A. Грубое
B. Положительное
C. Нейтральное

Вопрос 5. Какую задачу решает тематическое моделирование?

A. Распознавания
B. Оптимизации разметки текстов
C. Классификации

Рекомендуемая литература

1. Gibbs F., Owens T. Building better digital humanities tools //DH Quarterly. – 2012. – Т. 6. – №. 2.
2. Bradley A. J. et al. Visualization and the digital humanities //IEEE computer graphics and applications. – 2018. – Т. 38. – №. 6. – С. 26-38.
3. Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 16.09.2021). — Режим доступа: для авториз. пользователей.

Ответы к заданиям

Тест 1

1. B
2. B
3. B

Тест 2

1. A
2. B
3. C
4. D

Тест 3

1. B
2. C
3. A
4. D

Тест 4

1. A, C
2. A

Тест 5

1. A, B, C
2. A, B, D
3. C

Тест 6

I - C - 3,4
II - A - 1,6
III - B -2,5

Тест 7

1. C
2. A
3. B
4. A
5. C

Содержание

Пучковская Антонина Алексеевна
Зимина Лада Владимировна
Волков Дмитрий Алексеевич

Введение в цифровые гуманитарные исследования

Учебно-методическое пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49, литер А