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

DIGITAL HUMANITIES: ИНСТРУМЕНТАРИЙ НАЧИНАЮЩЕГО ИССЛЕДОВАТЕЛЯ

Санкт-Петербург
2022

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

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

DIGITAL HUMANITIES: ИНСТРУМЕНТАРИЙ НАЧИНАЮЩЕГО ИССЛЕДОВАТЕЛЯ

УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

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

Санкт-Петербург
2022

Пучковская А.А., Волков Д.А., Зимина Л.В., Digital Humanities: инструментарий начинающего исследователя – СПб: Университет ИТМО, 2022. – 82 с.

Рецензент(ы):
Даниил Андреевич Скоринкин, к.ф.н., доцент факультета гуманитарных исследований, Национальный исследовательский университет «Высшая школа экономики»

Цель учебно-методического пособия «Digital Humanities: инструментарий начинающего исследователя» — положить начало погружению студентов магистратуры в предметную область цифровых гуманитарных наук. Пособие предназначено для использования в рамках курса «Введение в цифровые гуманитарные исследования», который реализуется на программе «Цифровые методы в гуманитарных исследованиях». Оно состоит из пяти разделов, включающих практические 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. Вопросы, предложенные в конце разделов призваны не только проверить внимательность студентов способствуют более глубокому пониманию предмета и формированию критического отношения к данным и методам, применяемым для их анализа.

Каждый раздел пособия описывает один из популярных low-code или no-code инструментов, необходимых для дальнейшей исследовательской работы в области Digital Humanities. В первом разделе уделяется внимание инструменту Voyant tools, его возможностям и способам исследования текстов. Второй раздел посвящен работе с инструментом Palladio, демонстрирующим подход к исследованию данных через создание визуализаций и работу с графиками. В третьем разделе говорится о подготовке данных к анализу с использованием инструмента OpenRefine и поднимаются вопросы об ответственности исследователя при работе с данными. В четвертом разделе обсуждаются базовые понятия анализа социальных сетей(SNA) и описывается процесс работы в инструменте Gephi. В заключительной части рассказывается об инструменте визуализации данных Tableau, который помогает пользователям строить понятные и убедительные визуализации.

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 не потерять все в разнообразие смыслов и интерпретаций? 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.

Цель учебно-методического пособия «Digital Humanities: инструментарий начинающего исследователя» — положить начало погружению студентов магистратуры в предметную область цифровых гуманитарных наук. Пособие предназначено для использования в рамках курса «Введение в цифровые гуманитарные исследования», который реализуется на программе «Цифровые методы в гуманитарных исследованиях». Оно состоит из пяти разделов, включающих практические 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. Вопросы, предложенные в конце разделов призваны не только проверить внимательность студентов, способствуют более глубокому пониманию предмета и формированию критического отношения к данным и методам, применяемым для их анализа.

Каждый раздел пособия описывает один из популярных low-code или no-code инструментов, необходимых для дальнейшей исследовательской работы в области Digital Humanities. В первом разделе уделяется внимание инструменту Voyant tools, его возможностям и способам исследования текстов. Второй раздел посвящен работе с инструментом Palladio, демонстрирующим подход к исследованию данных через создание визуализаций и работу с графиками. В третьем разделе говорится о подготовке данных к анализу с использованием инструмента OpenRefine и поднимаются вопросы об ответственности исследователя при работе с данными. В четвертом разделе обсуждаются базовые понятия анализа социальных сетей(SNA) и описывается процесс работы в инструменте Gephi. В заключительной части рассказывается об инструменте визуализации данных Tableau, который помогает пользователям строить понятные и убедительные визуализации.

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. Voyant tools

Введение

Глава дает базовые представления о работе Voyant tools, веб-платформы для обзора и анализа текста. Инструмент разработан Стефаном Синклером (Stefan Sinclair) и Джеффри Рокуэллом (Geoffrey Rockwell) для облегчения прочтения и интерпретации текстов.

Инструмент рассчитан на использование учеными, занимающимися цифровыми гуманитарными науками, и предлагает частотные списки слов, графики распределения частотности слов и определение контекста для ключевых слов.

Что вы можете делать с Voyant:

- Используйте его, чтобы узнать, как работает компьютерный анализ текстов.
- Используйте его для изучения текстов, которые вы найдете в Интернете, или текстов, которые вы тщательно отредактировали и храните на своем компьютере.
- Используйте его, чтобы добавить функциональность в свои онлайн-коллекции, журналы, блоги или веб-сайты, чтобы другие могли просматривать ваши тексты с помощью аналитических инструментов.
- Используйте его для добавления интерактивных доказательств в свои эссе, которые вы публикуете в Интернете.
- Используйте его для разработки собственных инструментов с использованием функций и кода.

Для взаимодействия с Voyant tools вы можете использовать веб-версию или установить сервер на ваш компьютер.

Форматы документов

Voyant Tools - это веб-среда для чтения и анализа текста. Он позволяет загружать для анализа различные текстовые форматы, включая документы TXT, HTML, XML, PDF, RTF и MS Word. Вы можете создать свою собственную коллекцию текстов или использовать один из корпусов, доступных в Voyant Tools.

Добавление корпуса для анализа

Всего есть четыре способа начать работу с текстом:

1. Voyant tools имеет встроенные корпуса текстов: сборник пьес Шекспира и романы Джейн Остин. Вы можете найти, нажав на кнопку «Open», и выбрать то, что интересно вам.
2. Также можно скопировать текст в окно, которое вы видите на экране.
3. Есть вариант с URL-ссылками, вы можете копировать ссылки страниц, с которых хотите собрать тексты. В каждой строке должно быть по одной ссылке.
4. Последний способ позволяет загружать файлы, находящиеся на вашем компьютере. Для этого нажмите на кнопку «Upload» и выберите все файлы, зажав кнопку «Shift».

Интерфейс

На экране (см. Рисунок 1) можно увидеть стандартное меню программы, разделенное на 5 отдельных инструментов. Оно появится после того, как вы нажмете кнопку «Reveal» в окне загрузки текстов.

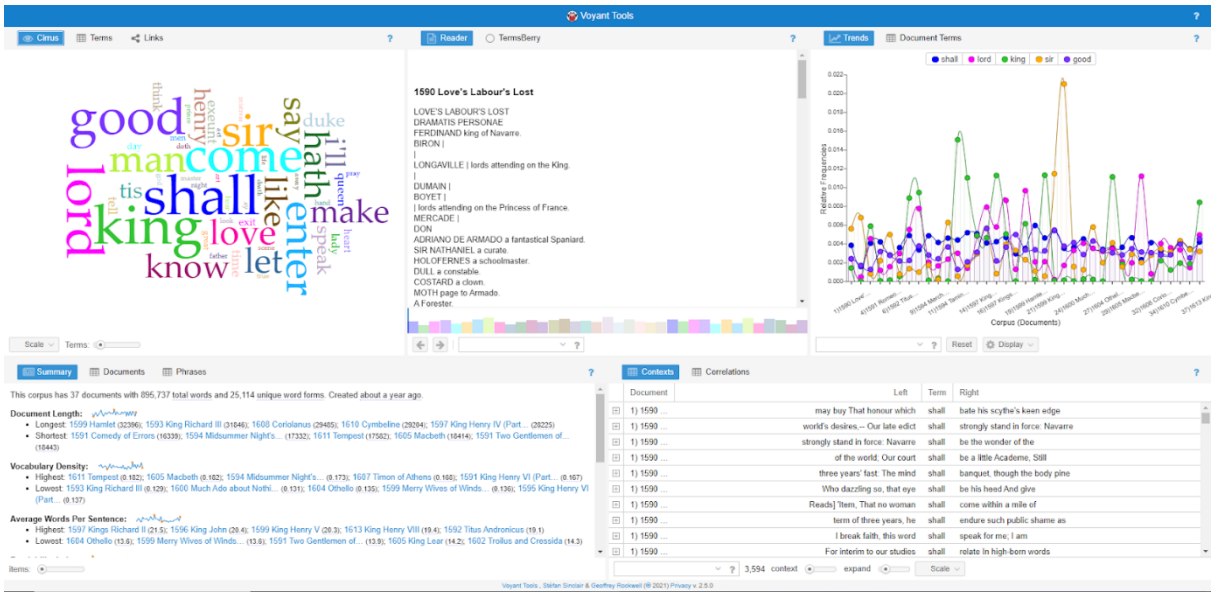

Рисунок 1 - Интерфейс Voyant tools

1. **Cirrus**
   Облако слов, которое визуализирует частоту употребления слов в корпусе. Чем чаще употребляется слово, тем больше оно становится на изображении, также чаще всего оно отображается в центре визуализации.
2. **Reader**
   Инструмент используется для отображения всего текста. Вы можете навести на слово и увидеть частоту его использования, а если нажмете на него, то частотность отобразится на графике справа.
3. **Trends**
   Инструмент показывает распределение используемых слов во всем тексте. Каждая линия раскрашена в цвет слова, которое она представляет. Сверху есть легенда, показывающая искомое слово и его

цвет. Если вы нажмете на определенное слово в легенде, то оно перестанет отображаться. Наведя на точку, вы можете увидеть информацию о слове в тексте.

4. **Summary**
   Инструмент показывает общую информацию о корпусе: количество документов, длину документов, количество слов в каждом документе, частоту употребления уникальных для каждого текста слов, среднее количество слов в предложениях и самые используемые слова для каждого текста.
5. **Contexts**
   Инструмент помогает определить контекст используемых слов. Таблица состоит из четырех колонок:
   - Документ, в котором находится слово
   - Слова, находящиеся слева от искомого слова
   - Интересующее пользователя слово
   - Слова, находящиеся справа от искомого слова

Чтобы поменять базовые инструменты, требуется нажать на кнопку (см. Рисунок 2). Она появляется при наведении на знак вопроса, который расположен в правом верхнем углу каждого инструмента. Voyant tools предоставляет большое количество инструментов для визуализации и анализа текста. Чтобы больше узнать об инструментах, необходимо нажать на знак вопроса и нажать на ссылку «More help…»

Рисунок 2 - Кнопка выбора инструментов

Облако слов и стоп-слова

Сначала мы разберем облако слов, как говорилось ранее, оно помогает определить самые распространенные слова. Чем чаще слово встречается, тем оно больше (см. Рисунок 3).

Рисунок 3 - Облако слов

Также можно увидеть слайдер «Terms» в левом нижнем углу, он отвечает за количество отображаемых слов. Максимальное количество слов 500, минимальное 25. Во вкладке «Scale» можно выбрать тексты, которые используются для анализа: можно переключаться между всем корпусом и отдельными текстами.

Стоп-слова

Работа со стоп-словами является важной частью при анализе текста, так как это помогает сделать более осмысленные выводы и исключить служебные части речи, для этого нужно перейти в опции инструмента (см. Рисунок 4).

Рисунок 4 - Кнопка для перехода в опции инструмента

В появившемся окне (см. Рисунок 5) можно выбрать несколько параметров: стоп-слова; белый список, отображающий только указанные слова; категории, шрифт, палитра цветов.

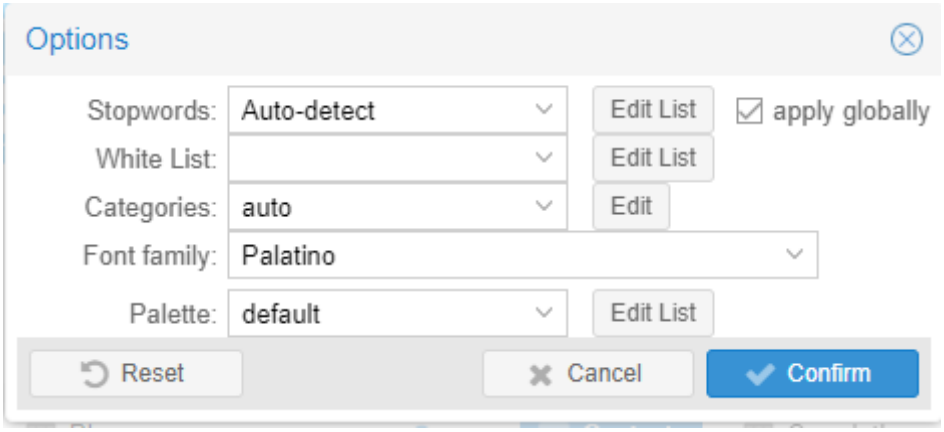

Рисунок 5 - Настройки отображения слов

Список стоп-слов редактируется, так что вы можете добавлять слова, которые вы бы не хотели видеть, нажав на кнопку «Edit List». После изменения списка стоп-слов визуализация изменится автоматически, но для изменений на графиках требуется нажать на кнопку «Reset», расположенную под используемым инструментом. Также самые частые в употреблении слова можно посмотреть в инструменте **Summary** (см. Рисунок 6).

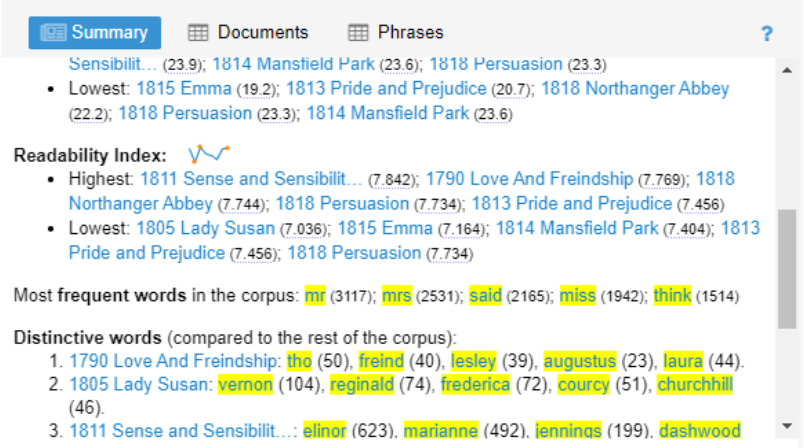

Рисунок 6 - Инструмент Summary

Тренды и частотность слов

Инструмент Trends (см. Рисунок 7), который также можно назвать графиком частотности слов, показывает употребление слов во всем корпусе. Каждая линия окрашена в определенный цвет, представляющий искомое слово. Сверху можно найти легенду, связанную с цветами, она интерактивна, так что можно убрать слова, если они вам неинтересны. При наведении на точку графика появляется статистика по употреблению слова в документе.

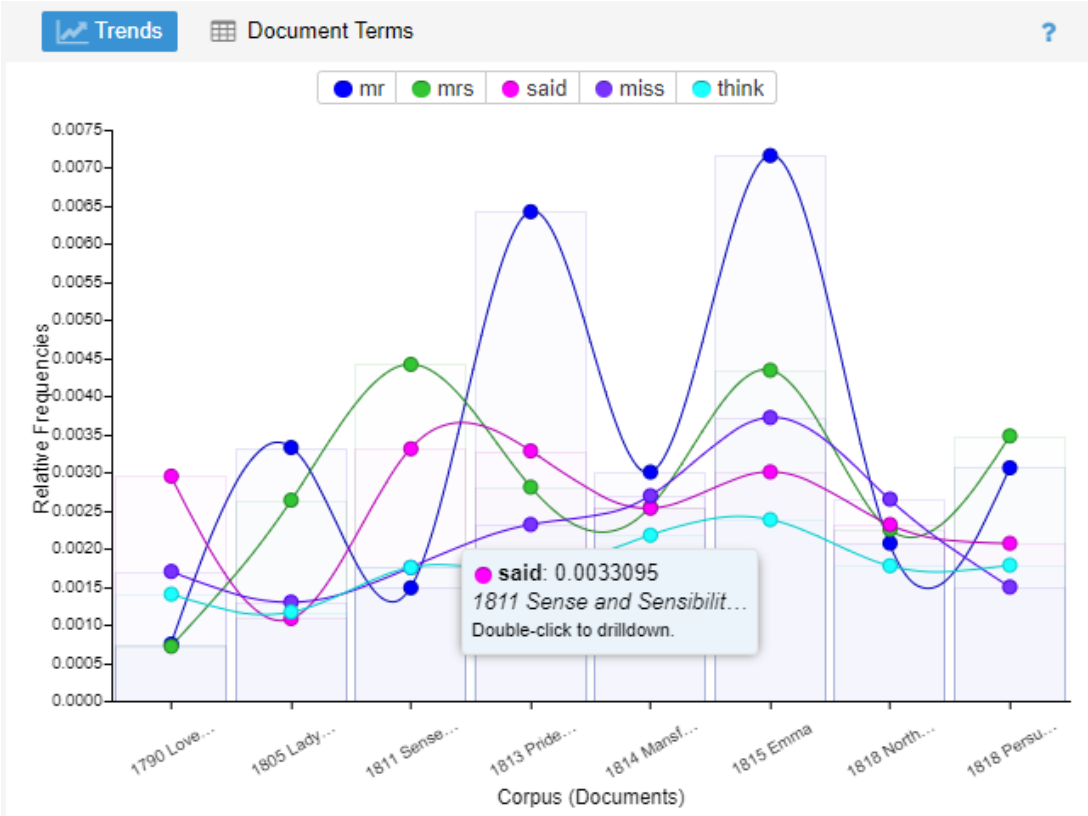

Рисунок 7 - Инструмент Trends

По умолчанию график показывает частоту употребления относительно всего корпуса, чтобы изменить ее на абсолютные значения, перейдите в настройки, нажав кнопку (см. Рисунок 8).

Рисунок 8 - Кнопка для включения окна настроек

В появившихся настройках (см. Рисунок 9) выберите «Raw», а затем нажмите «Confirm». Инструмент должен обновить графики автоматически.

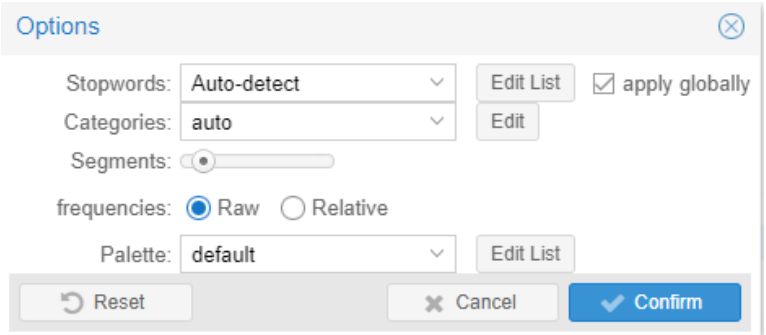

Рисунок 9 - Настройки инструмента Trends

Чтобы вернуться к первоначальным настройкам, требуется нажать кнопку «Reset». Сейчас можно обратить внимание на кнопку «Display», которая содержит несколько настроек отображения графика (см. Рисунок 10):

— Отображение названий (Show Labels) определяет отображение названия переменных на самом графике
— Тип отображения информации (Chart Mode)

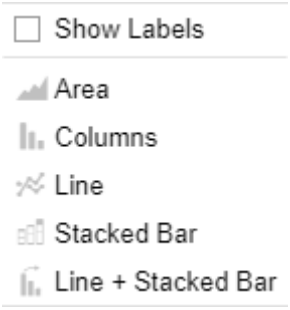

Рисунок 10 - Настройки графика

Давайте изменим отображение графика и выберем «Columns». Должен получится график вот такой график (см. Рисунок 11).

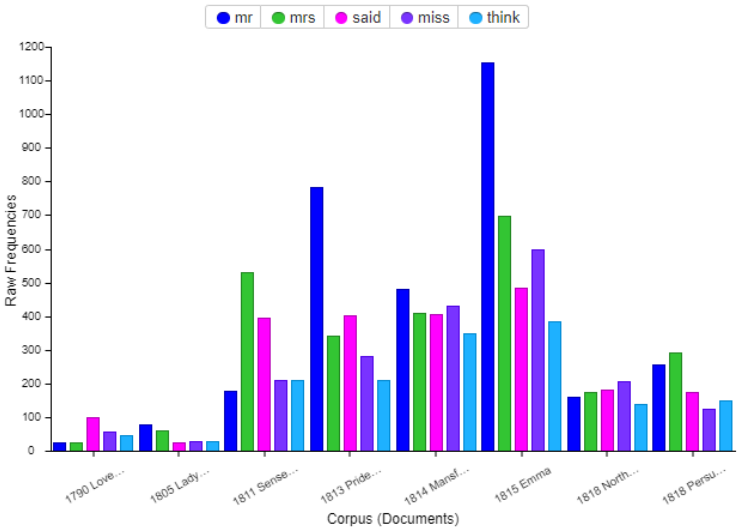

Рисунок 11 - График распределения слов в корпусе

Все инструменты в Voyant tools связаны между собой, поэтому вы можете выбирать термины в облаке слов, и они будут отображаться на графике. Если вас интересуют определенные слова, которых нет в облаке слов или они не представлены по умолчанию, можно воспользоваться поиском внизу инструмента. Например, мы хотим посмотреть, как отличается частота употребления слов «love» и «hate» в романах Джейн Остин. Для этого нам нужно ввести слова в поисковую строку и выбрать их из всего списка, график обновляется автоматически (см. Рисунок 12). Каждое слово в Voyant tools приведено к нижнему регистру, поэтому при поиске имен собственных необходимо писать слова с маленькой буквы.

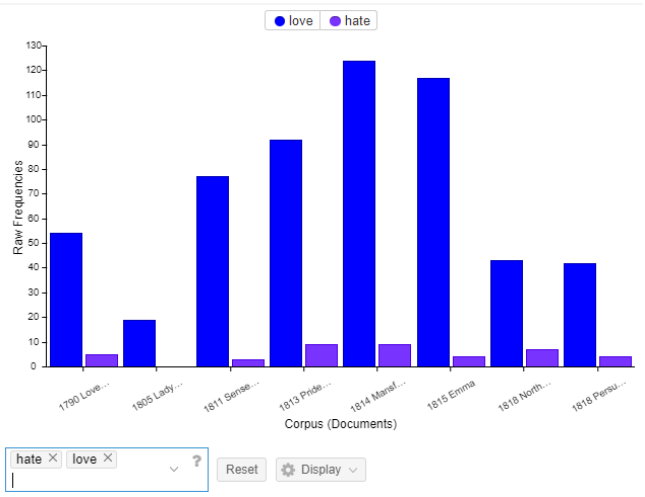

Рисунок 12 - Употребление слов "love" и "hate" в романах

Нам может быть недостаточно найти только употребление слов «love» и «hate», мы также хотим найти все слова, которые образовались от них. Для этого достаточно поставить «*» в конец слова (см. Рисунок 13).

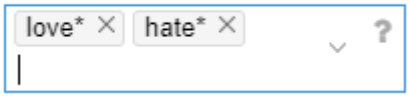

Рисунок 13 - Ввод интересующих слов

Таким образом, найдутся все слова, образованные от вышеуказанных.

Слова в контексте

Инструмент Contexts показывает каждое вхождение слова с текстом по обе стороны от него. Таблица состоит из четырех колонок:

1. Документ: какой текст используется при исследовании контекста
2. Слева (Left): слова слева от искомого слова
3. Термин (Term): запрос пользователя
4. Справа (Right): слова справа от искомого слова

По умолчанию показывается контекст для часто употребляемых слов (см. Рисунок 14). Чтобы уточнить свой запрос, можно использовать поисковую строку внизу инструмента. Кроме отображения слов, находящихся слева и справа, можно открыть целый текст, нажав на плюс в самой левой колонке.

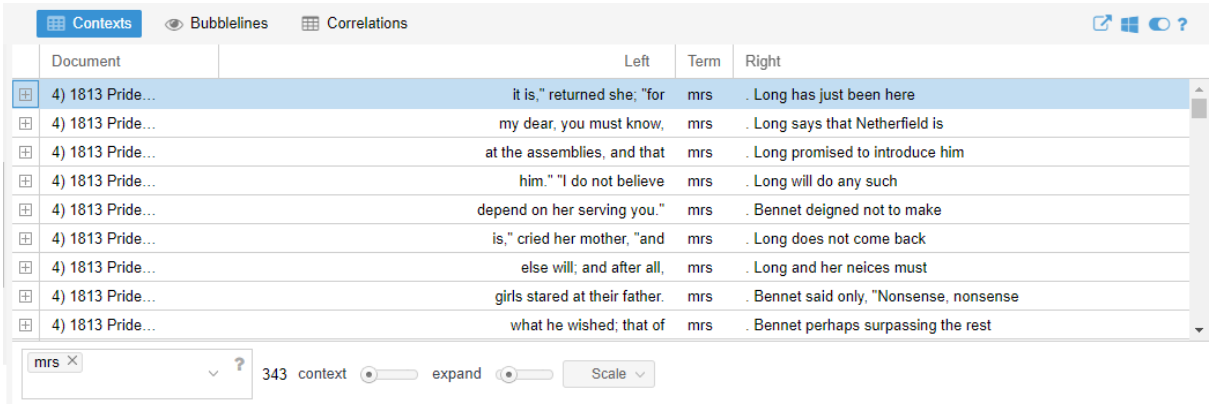

Рисунок 14 - Инструмент Contexts

Давайте вернемся к теме любви и посмотрим, в каком контексте используется это слово. Для этого введите его в поисковую строку. Вам может быть интересно посмотреть на контекст не во всем корпусе, а в отдельных документах. Для этого нажмите на кнопку «Scale», перейдя во вкладку «Documents» вы можете выбрать отдельные книги (см. Рисунок 15).

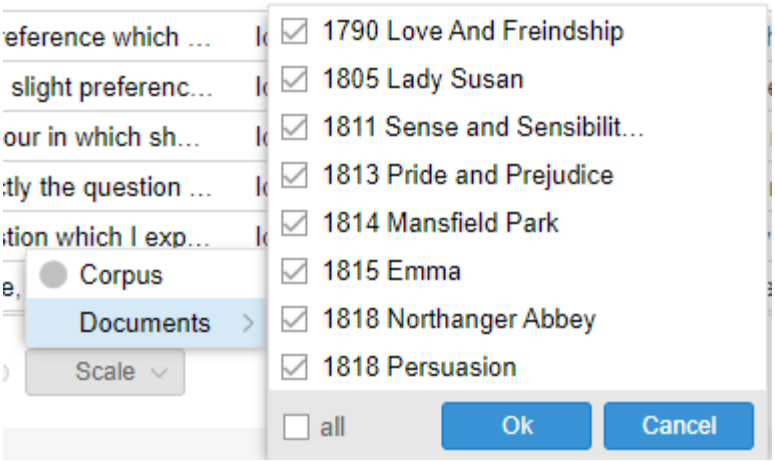

Рисунок 15 - Documents

В дополнение к этому можно использовать контекстный слайдер, который увеличивает отображение слов слева и справа, его можно найти внизу инструмента. Минимальный размер контекстного окна равен пяти, максимальный пятистам.

Давайте кликнем на первое слова в окне Contexts. Когда вы выбираете определенную строку, инструмент Reader автоматически находит эту строку в тексте. Инструмент Reader отображает весь корпус и состоит из двух компонентов: частотный график и текст.

В окне работы с текстом вы можете поближе познакомиться с текстом; узнать частоту употребления слова в документе, наведя на него курсор; кликнуть на слово, чтобы найти все его употребления. Частотный график демонстрирует информацию по всему корпусу, особенно полезно, когда он состоит из

нескольких документов. Каждый столбец иллюстрирует отдельный документ, также они расположены в порядке появления в корпусе (см. Рисунок 16).

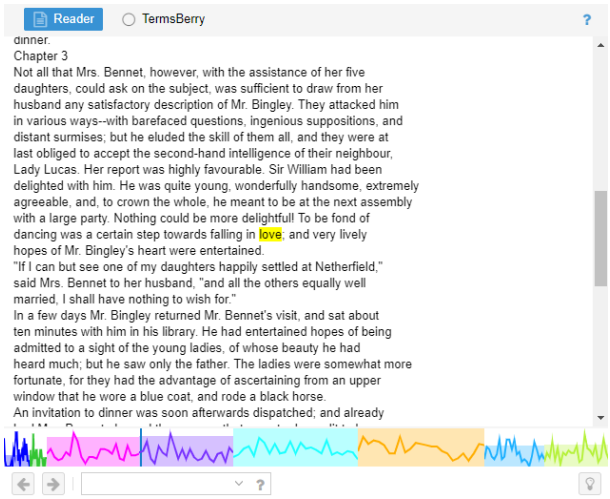

Рисунок 16 - Инструмент Reader

Относительная длина каждого документа представлена высотой и широтой столбца. Получается, чем шире столбец, тем длиннее сам документ. Когда вы используете инструмент для работы с ключевыми словами, он отображает частоту использования в каждом тексте, разделяя документ каждый документ на 25 равных частей.

Также на частотном графике отображается вертикальная синяя линия, которая показывает ваше текущее положение в корпусе. Вы можете щелкнуть в любом месте графика, чтобы перейти в другое место. В дополнение к этому для перехода вперед или назад можно использовать стрелки рядом с полем поиска.

Исследование коллокаций

В то время как инструмент Contexts позволяет исследовать контекст, в котором находится ключевое слово, инструмент Collocates показывает, какие термины чаще всего встречаются рядом друг с другом. Чтобы открыть инструмент, перейдите в меню инструментов (см. Рисунок 17) и выберите инструмент Collocates, находящийся во вкладке корпус «Corpus».


Рисунок 17 - Меню инструментов

Когда вы выберите его, то он должен автоматически заменить облако слов. Теперь вы можете вписать искомое ключевое слово.

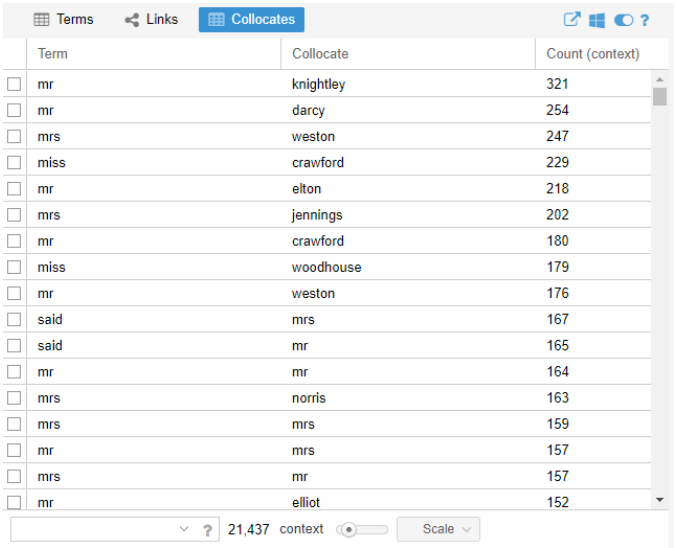

Рисунок 18 - Инструмент Collocates

По умолчанию таблица (см. Рисунок 18) состоит из трех колонок: термин (Term), коллокация (Collocates), количество (Count).

1. Термин определяет ключевое слово.
2. Коллокация показывает слово, которое часто встречается с ключевым словом.
3. Количество демонстрирует общее количество совместных вхождений в документе.

Вернемся к теме любви и посмотрим коллокации с этим ключевым словом. Нас интересуют две коллокации «love woman» и «love man», которые мы бы хотели рассмотреть во всем корпусе. Чтобы сделать это, вам нужно всего лишь поставить галочки напротив этих строк.

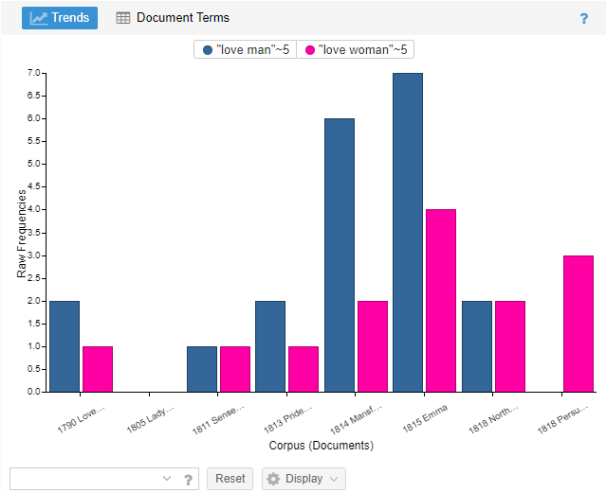

Рисунок 19 - График частоты употребления коллокаций

После этого график инструмента Trends должен измениться и показать частоту употребления коллокаций во всем корпусе (см. Рисунок 19).

Демонстрация результатов

Voyant предназначен для работы как отдельная среда (voyant-tools.org) или как набор независимых модулей, которые можно встроить в сайт. Функция экспорта позволяет встраивать инструменты Voyant в другие сайты. Наведите курсор на серую полосу, пока не появится меню значков.

Рисунок 20 - Кнопка для экспорта

В окне «Export» вам будет предложено от трех до четырех различных вариантов экспорта (см. Рисунок 21):

1. Фрагмент HTML для встраивания его в сайт.
2. Библиографическая ссылка на вашу сессию в Voyant tools
3. URL на эту сессию с данными и инструментами

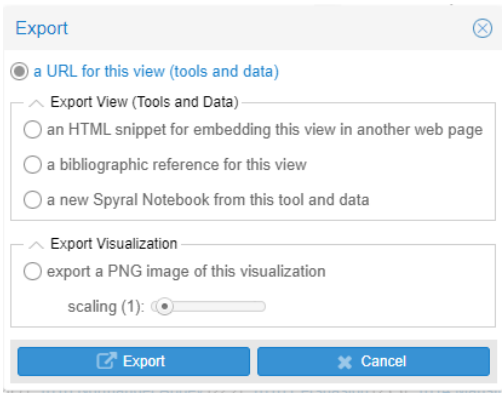

Рисунок 21 - Варианты экспорта результатов

Вам нужно будет создать фрагмент HTML для того, чтобы встроить результаты работы в сайт. Щелкните значок «Export», разверните раздел «Export View», выберите пункт «HTML snippet…», а затем нажмите кнопку «Export».

Вы должны получить всплывающее окно с фрагментом HTML, который вы можете использовать для встраивания выбранного инструмента и корпуса в другой веб-сайт. Скопируйте и вставьте «iframe», который появляется в поле (см. Рисунок 22).

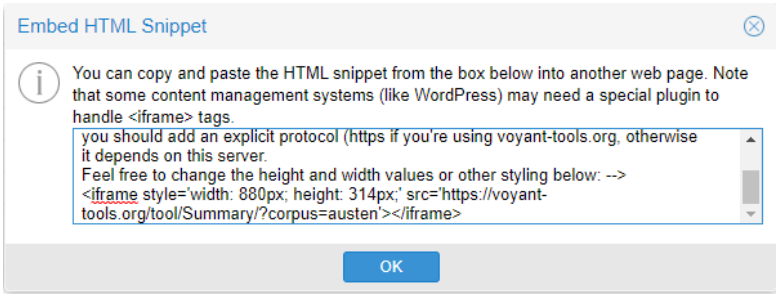

Рисунок 22 - Окно с HTML

Экспортировать PNG-изображение (см. Рисунок 23) можно в разделе «Export Visualization», он доступен только при работе с определенными инструментами: функция создает PNG-изображение текущего инструмента или создает фрагмент кода HTML, в котором содержится изображение.

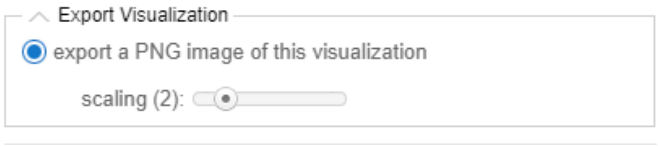

Рисунок 23 - Процесс экспорта в PNG

В главе лишь поверхностно рассмотрены возможности Voyant tools и типы анализа текстов, которые вы можете выполнять. Чтобы узнать больше о инструментах Voyant, вы можете обратиться к документации, ссылка на которую указана в источниках.

Источники

1. Официальная документация Voyant tools [Электронный ресурс]

Режим доступа: https://voyant-tools.org/docs/#!/guide

Тест 1.

1. Сколько существует способов для загрузки данных в Voyant tools?

A. 6

B. 2

C. 4

2. Для чего используется «белый» список?

A. Для удаления ненужных слов

B. Для поиска определенных слов

C. Для выявления самых частых слов

3. Что делает инструмент MicroSearch:

A. Визуализирует частоту и распределение слов в корпусе

B. Визуализирует слова, которые чаще всего встречаются вместе

C. Визуализирует повторяющиеся последовательности слов

Глава 2. Palladio

Введение

В этом разделе мы обсудим инструмент визуализации и исследования гуманитарных данных, разработанный в Стэнфордском университете. Целью проекта было осмысление разработки графических интерфейсов на основе запросов от исследователей-гуманитариев и создание универсального набора инструментов для визуализации и анализа. Нажмите «Start», чтобы начать работать в Palladio (см. Рисунок 24).

Рисунок 24 - Начальный экран

Загрузка данных

Для того чтобы загрузить датасет, с которым вы хотите работать, необходимо перетащить его в окно. Palladio принимает несколько типов файлов: csv, tab, tsv. В дополнение к этому можно использовать ссылки на датасеты и простое копирование содержания всего файла.

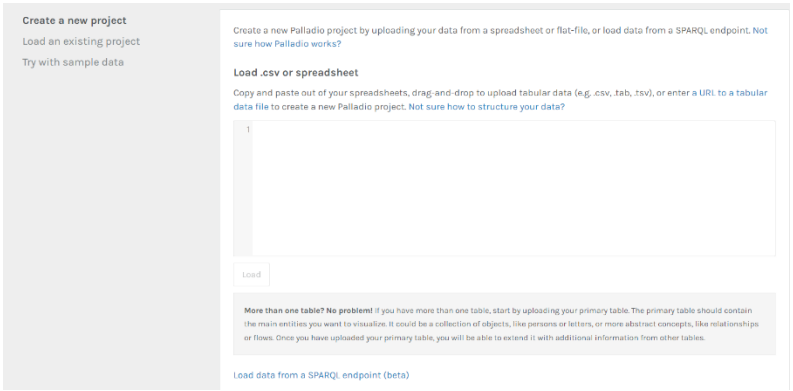

Рисунок 25 - Окно для загрузки данных

Если вы хотите загрузить несколько таблиц, Palladio позволяет это сделать. Функция будет доступна в разделе «Data» после того, как вы загрузите первый датасет. Для этой главы был использован тестовый датасет инструмента, который вы можете использовать, нажав кнопку «Try with sample data» (см. Рисунок 25).

Импортированные данные

После нажатия на кнопку «Try with sample data» у вас должно было открыться окно (см. Рисунок 26). Это основная рабочая область, здесь мы будем работать со всеми инструментами, которые предлагает Palladio. Мы можем использовать карты, графы, таблицы, и галерею, а также трансформировать сами данные в разделе «Data».

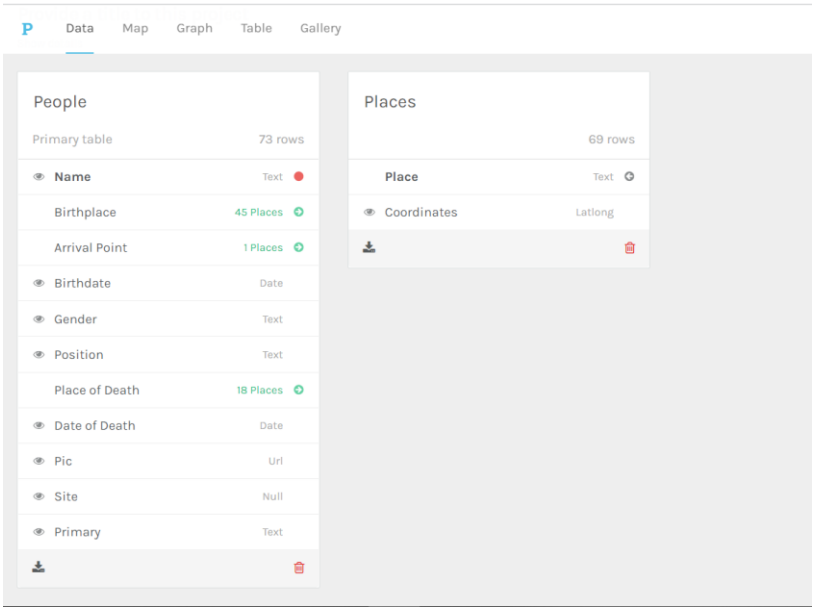

Рисунок 26 - Рабочая область

Каждый столбец в таблице имеет свой тип данных: Name — текст, Birthdate — дата, Pic — ссылка. Palladio автоматически их определил, но бывают случаи, когда инструмент делает ошибки, их можно исправить, выбрав интересующую вас колонку, вы увидите больше информации по колонке и возможности ее исправления (см. Рисунок 27). Если вы не хотите работать со всеми данными, то можно нажать на глаз, расположенный около названия колонки, это поможет Palladio работать быстрее.

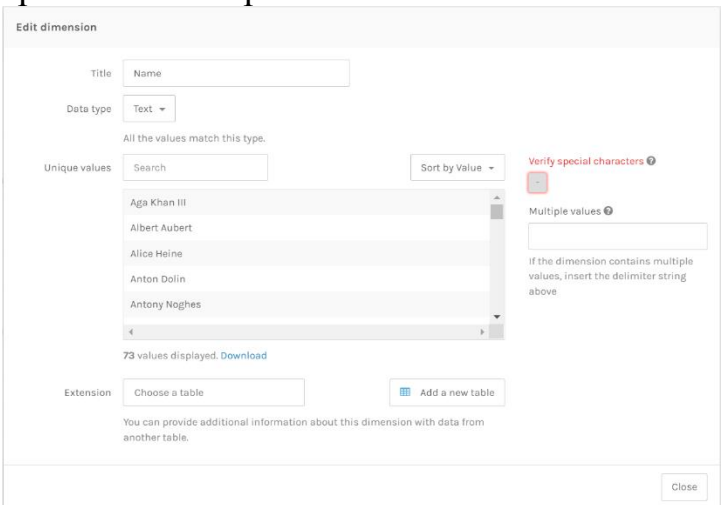

Рисунок 27 - Подготовка данных

Работа с картой

Щелкните вкладку «Map» в верхней части экрана, чтобы перейти в режим работы с картой для анализа ваших данных. Прежде чем мы продолжим, давайте поговорим о том, что вы видите на панели «Map layers».

Palladio предполагает, что вы будете использовать слои для нанесения данных на карту. Это означает, что вы можете нанести на карту не только что-то одно, например, место рождения; поверх этих данных можно создавать другие объекты. Например, при использовании слоя с местом рождения и слоя дорожных сетей можно предположить, как люди добрались до Монако.

Рисунок 28 - Map Layers

Выберите «New layer», чтобы открыть новое окно «Map layers» (см. Рисунок 28). В этом меню мы можем выбрать данные, которые хотим видеть на этом слое. Palladio предлагает два варианта: «Points» и «Point to point». Первый отвечает за простое расположение точек на карте, второй можно использовать для отображения путей (см. Рисунок 29). Слой можно назвать как угодно.

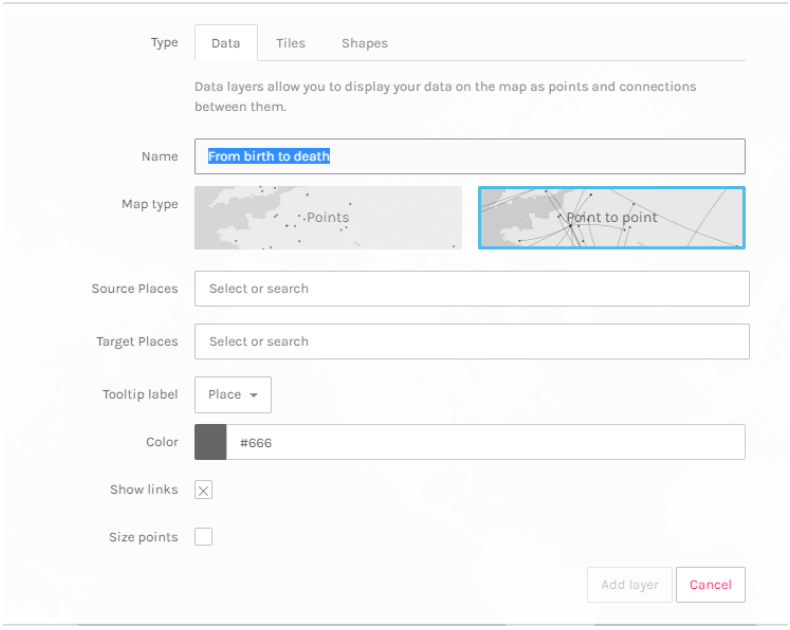

Рисунок 29 - Настройка слоя

Мы должны определить для Palladio следующие переменные: «Source place» — откуда человек едет, «Target place» — куда человек едет, «Tooltip label» — что отображать при наведении курсора на точку. Для первой переменной мы возьмем «Birthplace», для второй «Place of death», для третьей «Place». Когда вы настроите все, что хотите, то нажмите «Add Layer». Если вы все сделали правильно, то у вас должна получиться карта (см. Рисунок 30).

Рисунок 30 - Карта с местами рождения и смерти

Использование «Timespan» и «Facet»

Возможность наносить данные на карту - это прекрасно, но настоящая сила Palladio - это способность исследовать взаимосвязи различных данных с помощью «Facets», «Timelines», «Timespan». Начнем с использования «Timespan».

Начните с нажатия на вкладку «Timespan». Теперь вы можете увидеть распределение периодов жизни известных личностей во времени (см. Рисунок 31).

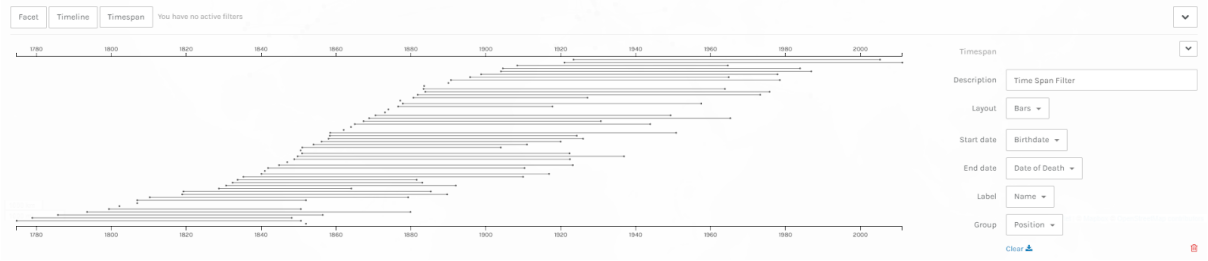

Рисунок 31 - Работа с "Timespan"

Бывает так, что нас могут интересовать отдельные случаи, например, мы хотим узнать про аристократов, о которых есть информация в этом датасете. С этой задачей нам поможет «Facet». Нажав на эту кнопку, вы увидите меню, позволяющее настраивать фильтры. Выберите «Position» во вкладке «Dimensions» (см. Рисунок 32).

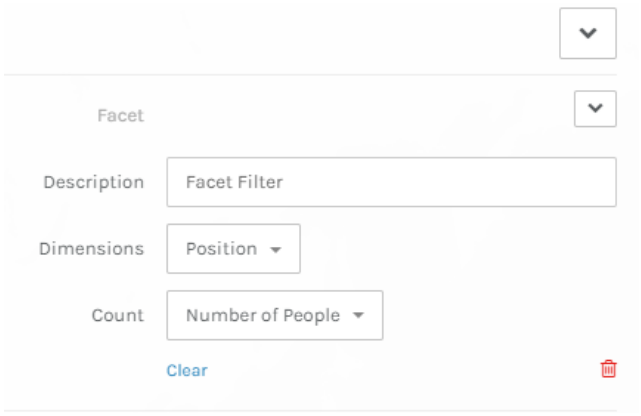

Рисунок 32 - вкладка "Dimensions"

После этого должно появиться меню со всеми категориями, которые присутствуют во вкладке «Position», дальше мы можем выбрать аристократов, так как нас интересуют только они.

Рисунок 33 - Выбор категорий

При использовании этого фильтра у нас изменятся отображаемые данные, посмотрите на карту или временную шкалу. Одновременно можно использовать несколько фильтров, просто вернитесь в меню «Facet» и измените выбор во вкладке «Dimensions». Если выбрать категорию Birthplace, то можно заметить, что большинство аристократов родились в Париже (см. Рисунок 34).

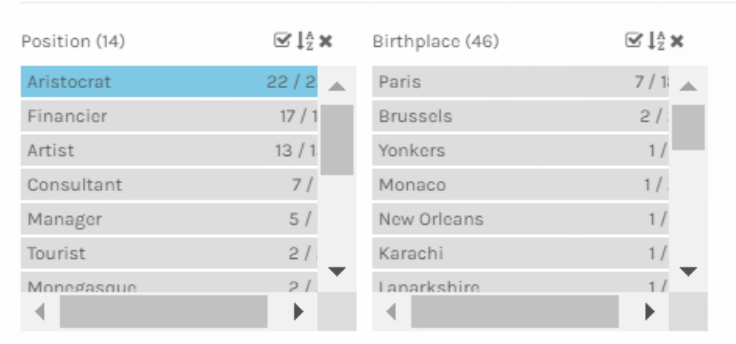

Рисунок 34 - Работа с несколькими фильтрами

Отображение данных в галерее

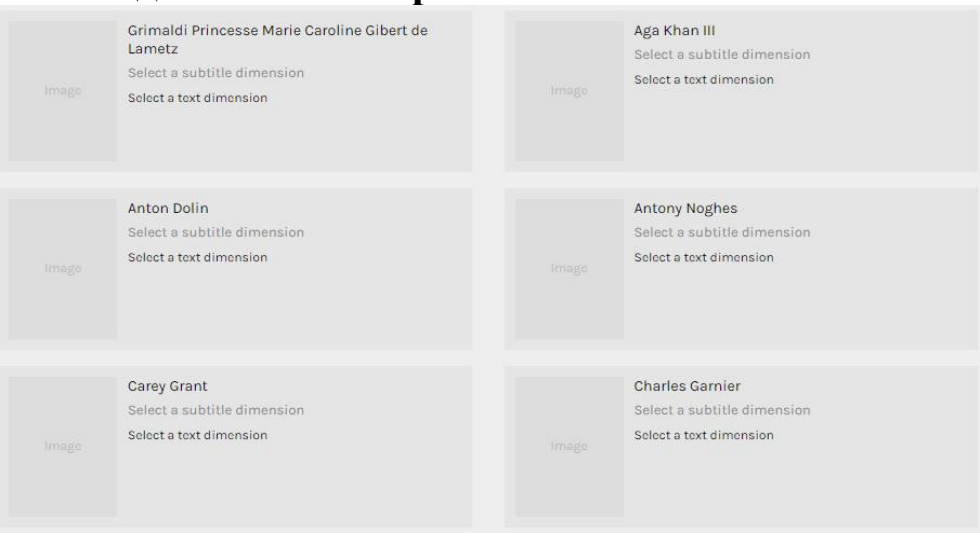

Рисунок 35 - Вид галереи

Использование галереи (см. Рисунок 35) может быть полезно, когда вы работаете с изображениями. Сначала удалите фильтры, которые остались после работы с «Timespan» и «Facet», щелкнув на мусорный бак, находящийся в правом нижнем углу каждой панели, или вы также можете удалить их, нажав на розовый крестик в верхней части панели фильтров (см. Рисунок 36).

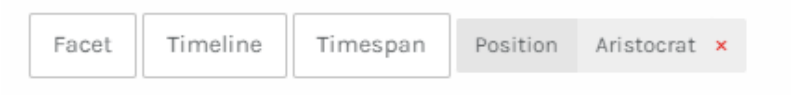

Рисунок 36 - Удаление фильтров

При работе с меню галереи вы можете выбрать параметры, которые должны в ней отображаться (см. Рисунок 37). В нашем датасете есть ссылки на картинки, так что, настроив вкладку «Image URL», мы подгрузим фотографии известных личностей.

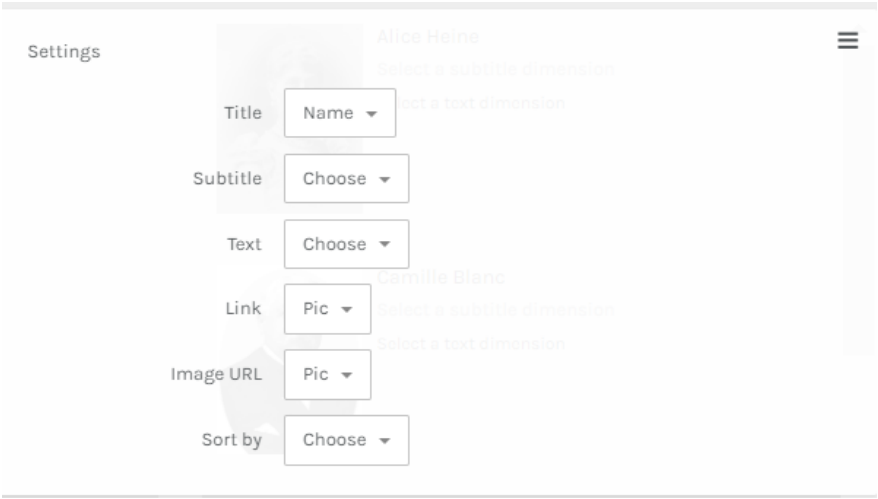

Рисунок 37 - Параметры галереи

По желанию вы также можете выбрать информацию, которую хотите отображать в галерее. Не забывайте, что здесь тоже можно пользоваться фильтрами, которые мы разобрали чуть раньше. После того как вы выберете все параметры, Palladio должен заполнить всю информацию для правильного отображения исторических личностей (см. Рисунок 38).

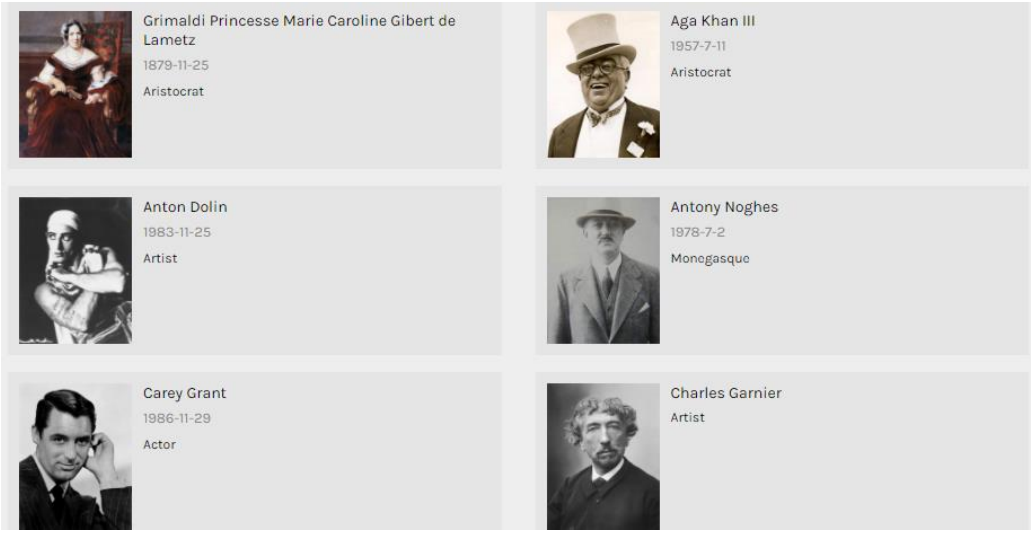

Рисунок 38 - Финальный вид галереи

Сетевые диаграммы

Сетевые диаграммы хороши для изучения отношений между разными сущностями. Часто такими сущностями являются люди, но мы можем использовать в качестве сущностей все что угодно, хоть это и не всегда нужно. Чтобы просмотреть данные в виде сетевой диаграммы, перейдите в раздел «Graph» (см. Рисунок 39).

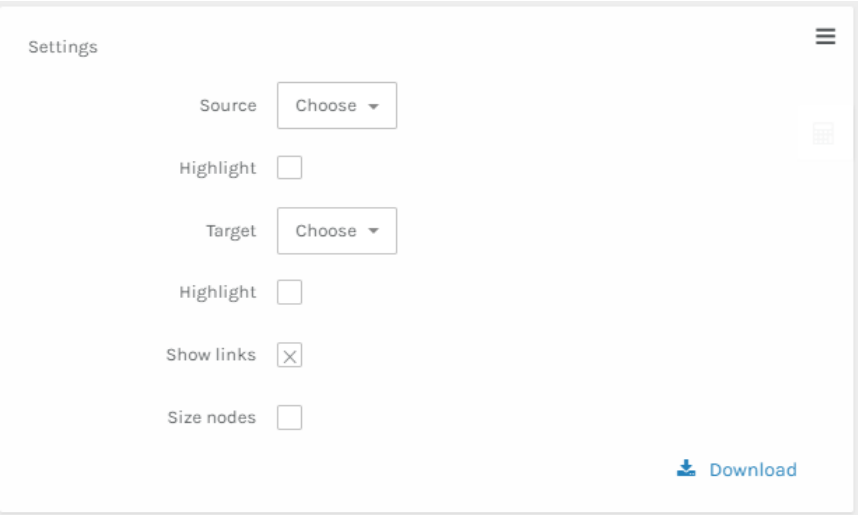

Рисунок 39 - Меню для настройки сети

Чтобы создать сетевую диаграмму, вам нужно указать две сущности в ваших данных, чьи взаимосвязи вы хотите исследовать. Можно заметить, что принцип работы схож с использованием «Point to point» при работе с картами.

В графе Source выберите Birthplace, в графе Target выберите «Place of death». Для выделения интересующих вас узлов можно использовать пункт «Highlight» (см. Рисунок 40).

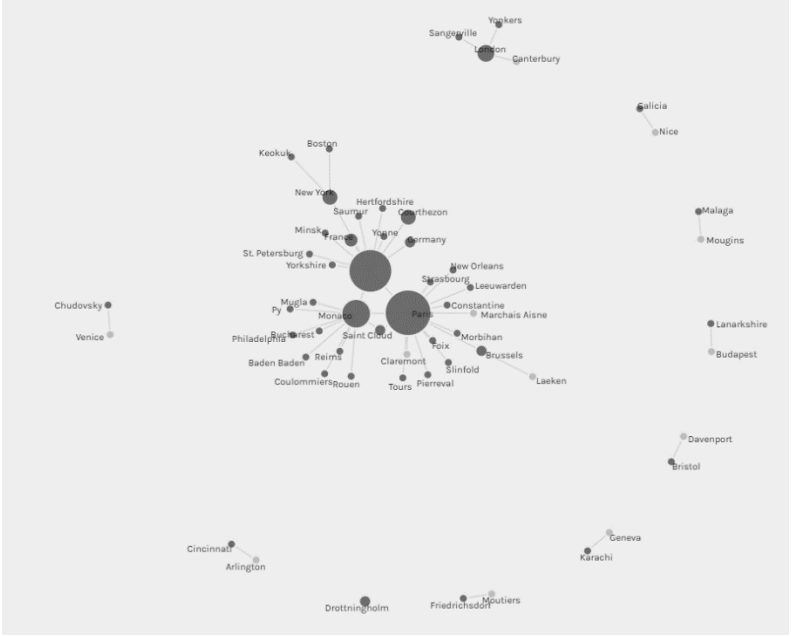

Рисунок 40 - Готовая сеть

Чтобы установить размер узлов в соответствии с количеством упоминаний в таблице, установите флажок «Size nodes». И вы можете использовать фильтры для диаграммы так же, как с картой и галереей. В дополнении к этому Palladio рассчитывает базовые статистики, использующиеся при анализе социальных сетей, их можно найти во вкладке под настройками диаграммы. Иконка похожа на таблицу (см. Рисунок 41).

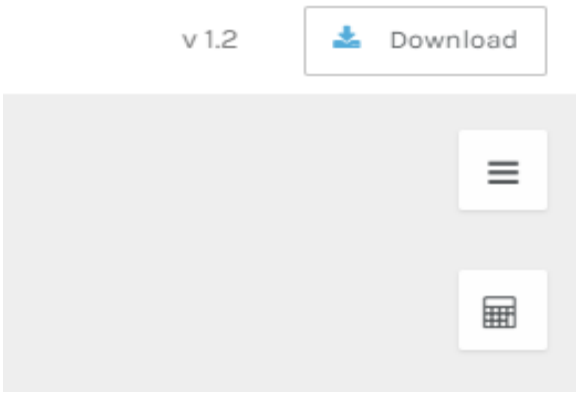

Рисунок 41 - Вкладка со статистиками

Сохранение своей работы

К сожалению, вы не можете встраивать интерактивные диаграммы Palladio в веб-страницы, но вы можете сохранять изображение, сделав снимок экрана или использовав функцию «Download», которая позволяет загрузить файл формата «svg» (см. Рисунок 42).

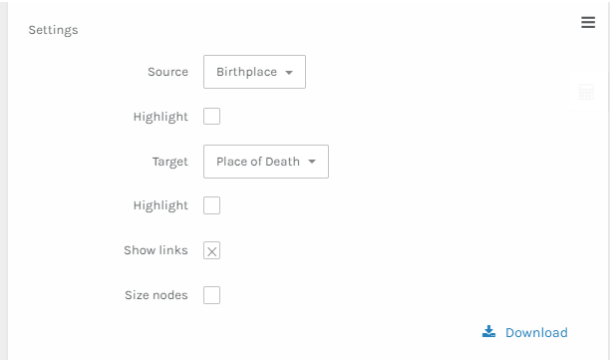

Рисунок 42 - Пример окна загрузки

Palladio не сохраняет ваши данные, но вы можете экспортировать свою модель данных и загрузить ее позже. Это избавит вас от необходимости настраивать все сначала, когда вы вернетесь к работе. Для этого нажмите на кнопку «Download», расположенную на вкладке с разными инструментами. На ваш компьютер будет загружен файл с расширением «json». В следующий раз, когда вы воспользуетесь Palladio, загрузите этот файл, чтобы открыть свой проект с того места, где вы остановились.

Источники

1. Официальное руководство Palladio [Электронный ресурс]
Режим доступа: https://hdlab.stanford.edu/palladio/help/

Тест 2

1. Сколько инструментов предлагает Palladio?
A. 1
B. 2
C. 4
2. Сколько слоев карт можно сделать в инструменте Map?
A. Пока компьютер не начнет замедляться
B. 3
C. 10
3. Как сохранить прогресс своей работы в Palladio?
A. Сохранить настройки в личном кабинете
B. Загрузить файл «.json» на свой компьютер, потом можно продолжить работать с ним
C. В Palladio нет такой функции

Глава 3. OpenRefine

Введение

Некоторые наборы данных, с которыми вы столкнетесь, могут быть беспорядочными. Часто существуют несоответствия в способе ввода данных: от орфографических ошибок до лишних пробелов, что может затруднить анализ данных в дальнейшем.

Очень важно очистить данные, прежде чем пытаться использовать их каким-либо образом. В этом руководстве мы узнаем, как очищать данные с помощью мощной программы OpenRefine. Хотя OpenRefine может выполнять множество задач по очистке, в этом руководстве будут рассмотрены только основы подготовки данных.

Загрузка и установка

Чтобы начать использовать OpenRefine, перейдите на страницу(https://openrefine.org/download.html), загрузите и следуйте инструкциям по установке. После установки запустите OpenRefine. Когда вы запускаете OpenRefine, он должен автоматически открывать новое окно браузера. (Примечание: OpenRefine не работает как настольное приложение, а вместо этого использует окно браузера.)

Загрузка файла и создание проекта

Теперь давайте попрактикуемся в очистке некоторых данных. Загрузите набор данных об объектах культурного наследия Санкт-Петербурга (https://classif.gov.spb.ru/irsi/7832000069-obuekty-kulturnogo-naslediya-na-territorii-sankt-peterburga/structure_version/436/) в формате «csv». В OpenRefine выберите вкладку «Create project» и выберите файл, который мы только что скачали (см. Рисунок 43).

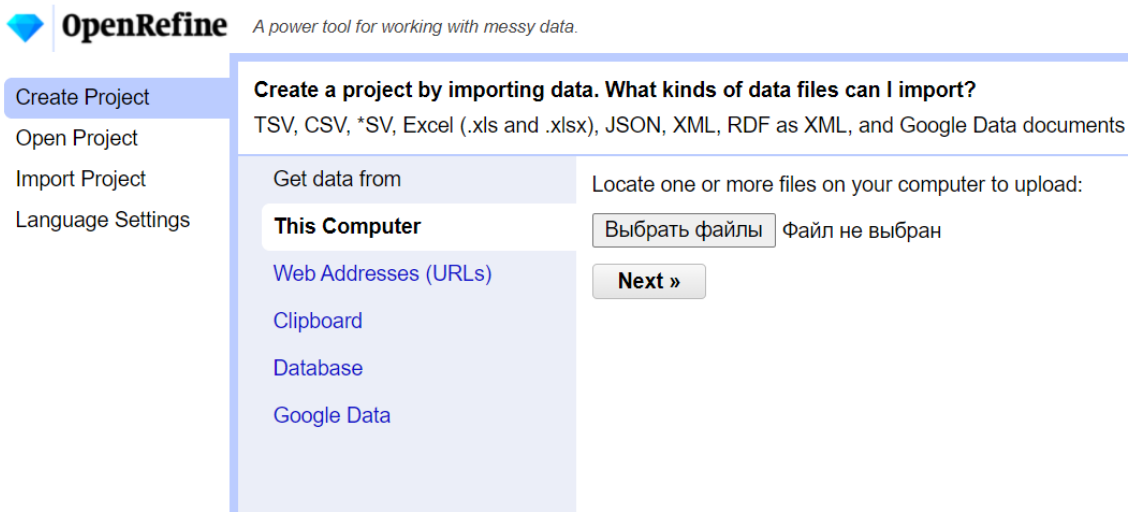

Рисунок 43 - Основное меню

Следующий экран, который вы увидите, — это экран предварительного просмотра. Он показывает, как OpenRefine видит данные, и позволяет вам изменять настройки перед их импортом (см. Рисунок 44).

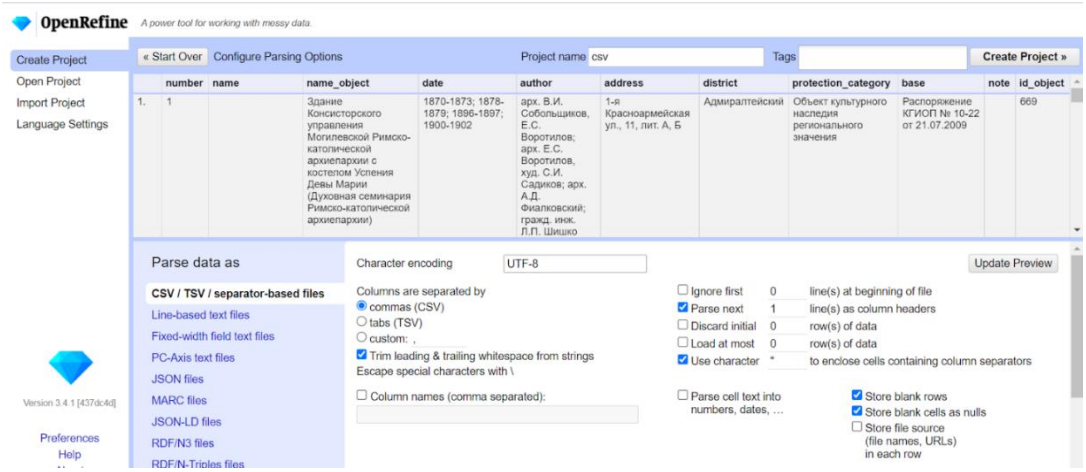

Рисунок 44 - Подготовка данных

Мы оставим настройки импорта нетронутыми за исключением одного небольшого изменения (см. Рисунок 45). В нижней части экрана обязательно установите флажок «Parse cell text into numbers, dates, …».

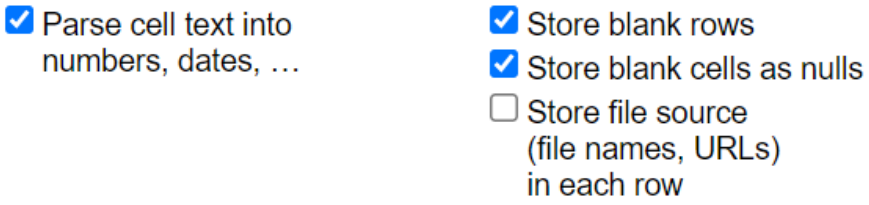

Рисунок 45 - Настройки импорта данных

Это позволяет OpenRefine классифицировать числовые переменные в ваших данных как числа или даты. Это не будет иметь большого значения в примере, который мы используем для этого руководства, но это хорошая привычка, чтобы двигаться вперед.

Теперь нажмите кнопку «Create project» в верхней правой части экрана, чтобы завершить импорт (см. Рисунок 46).

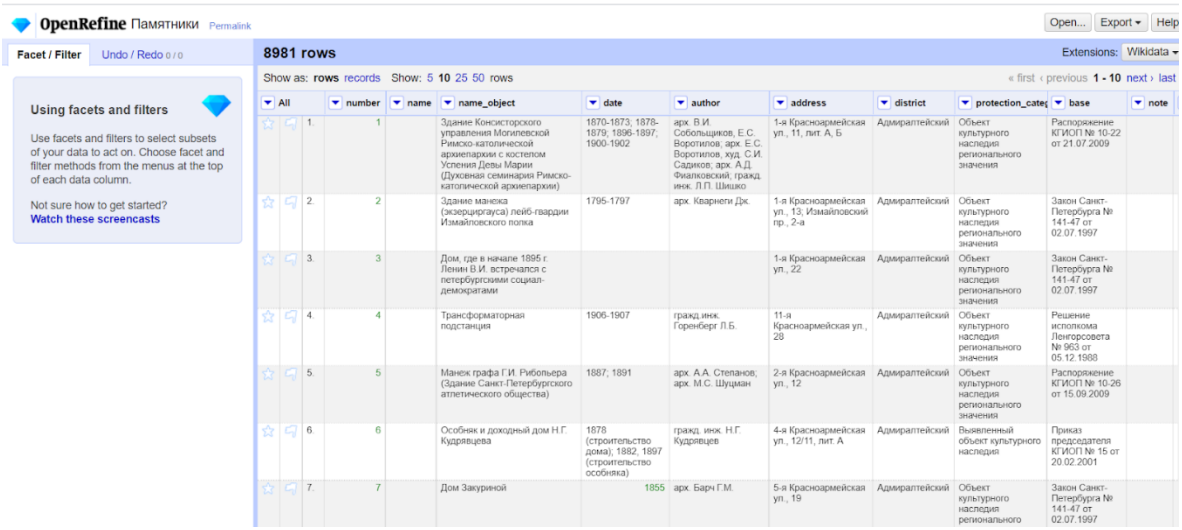

Рисунок 46 - Основная рабочая область

Чтобы увидеть больше данных, вы можете изменить количество отображаемых строк, изменив настройки в верхней части экрана, чтобы отобразить 50 строк вместо 10 по умолчанию.

Удаление и переименование столбцов

Можно заметить, что есть столбцы из исходных данных, которые нам не нужны, например, столбец «number», который дублирует автоматическую нумерацию OpenRefine. Мы можем удалить этот столбец, щелкнув на маленький треугольник у имени столбца и выбрав «Edit column» → «Remove this column» (см. Рисунок 47).

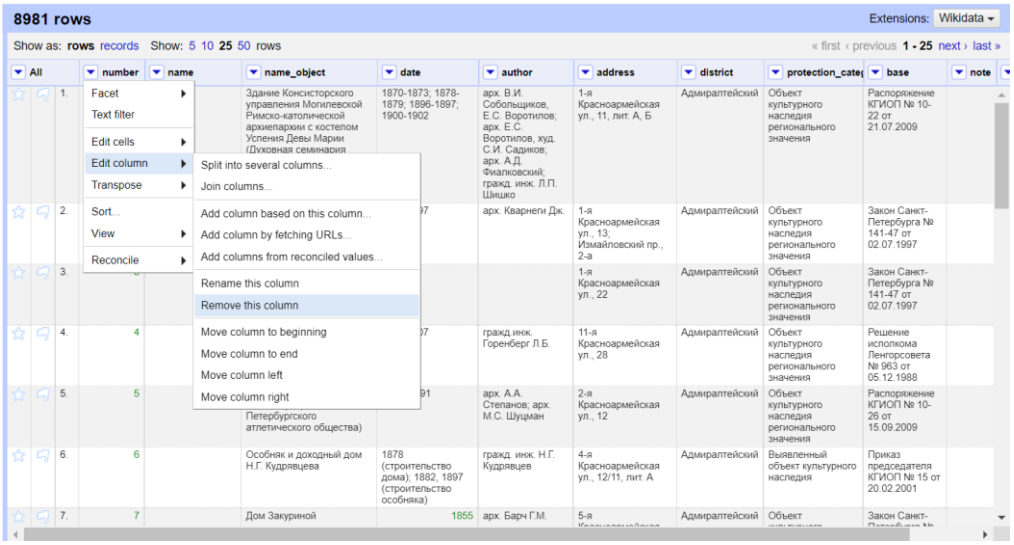

Рисунок 47 - Удаление колонок

Очистка через группировку и редактирование

Если мы пристально посмотрим на наши данные, то увидим, что исследователи, заполнившие данные, вносили имена архитекторов по-разному, что может привести к сложностям в дальнейшем. Если бы мы проверяли 8981 строку вручную, то это заняло бы много времени. Для

решения таких проблем есть мощный инструмент, помогающим в этой работе, который называется «Cluster and Edit». С помощью этой функции OpenRefine просматривает данные в выбранном вами столбце и использует алгоритмы, чтобы попытаться распознать значения, которые могут быть вариациями одного и того же. Затем он позволяет вам сгруппировать или объединить их вместе под одним согласованным именем по вашему выбору.

Нажмите на маленькую стрелку рядом со столбцом «author» и в меню выберите «Edit cells», затем «Cluster and edit» (см. Рисунок 48).

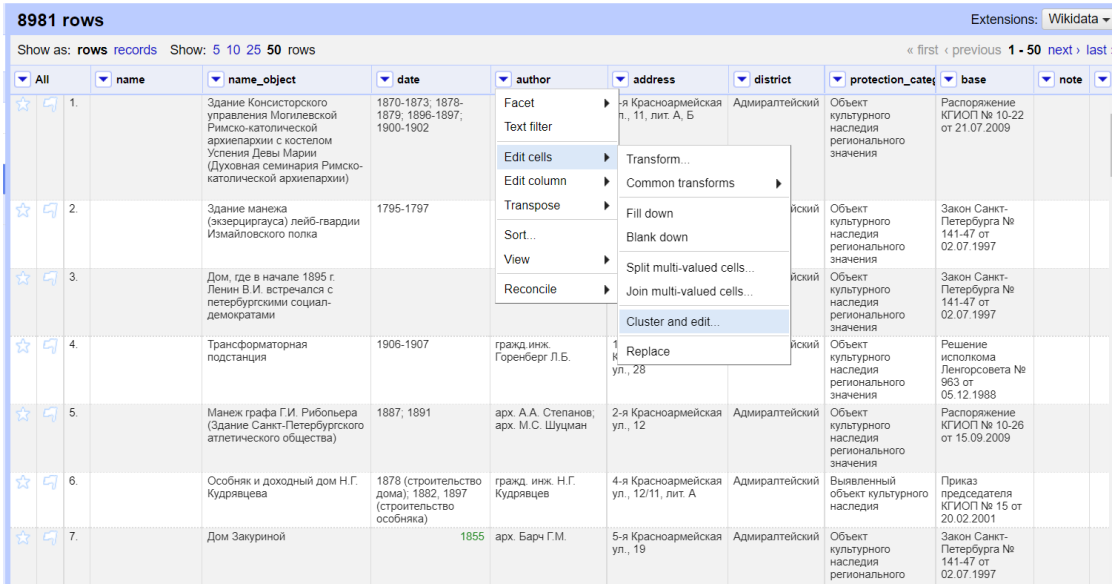

Рисунок 48 - Cluster and edit

Должен появиться следующий экран (см. Рисунок 49).

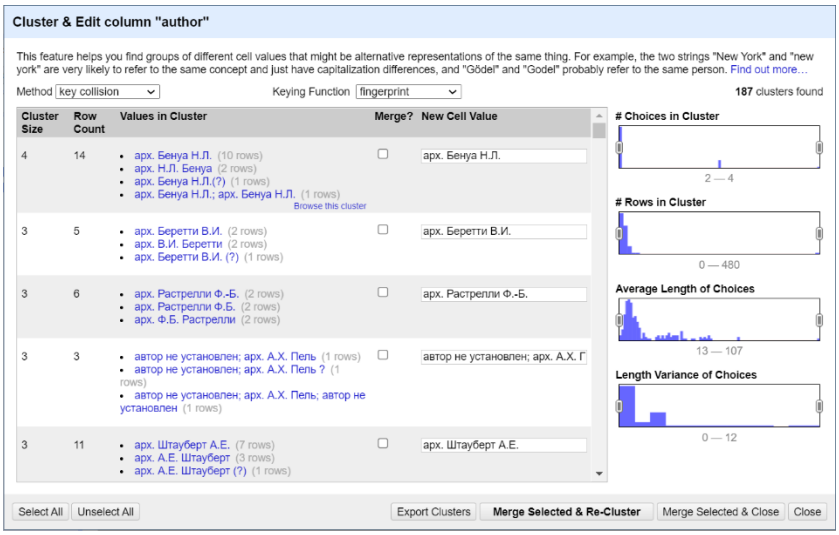

Рисунок 49 - Инструмент Cluster and edit

Прежде чем приступить к очистке, давайте удостоверимся, что мы понимаем, на что мы смотрим в окне «Cluster and Edit».

Давайте посмотрим на столбец «Values in Cluster». Здесь мы видим все вариации имени, которые подбирает выбранный алгоритм (см. Рисунок 50).

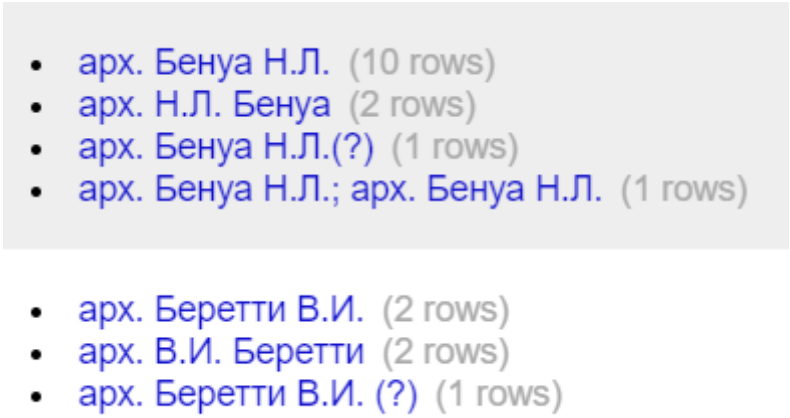

Рисунок 50 - Переменные в Cluster and edit

Теперь посмотрим на столбец «New Cell Value». Он содержит текстовое поле с предложением OpenRefine для согласованного имени данных (см. Рисунок 51). Важно всегда обращать внимание на это предложение и при необходимости редактировать его, чтобы получить данные в нужном вам формате.

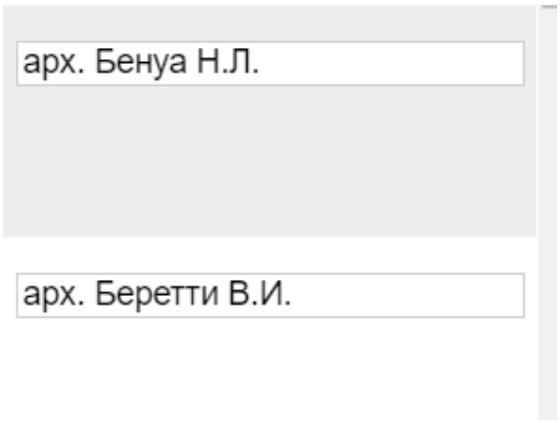

Рисунок 51 - Строка редактирования имен

Чтобы обновить любое имя, все, что нам нужно сделать, это установить флажок под надписью «Merge?» и нажать кнопку «Merge Selected and Re-Cluster». Как только мы это сделаем, варианты имени в столбце «Values in Cluster» объединятся под новым именем, которое мы выбрали в столбце «New Cell Value».

Еще одна настройка окна «Cluster and Edit», которую необходимо понять, — это настройки алгоритма. В верхней части окна вы увидите два

раскрывающихся меню под названием «Method» и «Keying Function» (см. Рисунок 53).

Рисунок 52 - Метод поиска похожих сущностей

Не беспокойтесь о том, что означают эти термины, но знайте, что настройки в этом меню определяют алгоритм, который OpenRefine использует для распознавания вариаций среди ваших данных.

Некоторые алгоритмы более консервативны — это означает, что в нашем случае OpenRefine по-прежнему распознает небольшие различия между именами в наших данных, на всякий случай, если это могут быть разные люди. Другие менее консервативны, что означает, что OpenRefine делает более широкие предположения о том, какие варианты имен, по его мнению, принадлежат одному и тому же человеку.

В общем, лучше всего очищать данные в порядке от наиболее консервативных к наименее консервативным алгоритмам, чтобы можно было быть уверенным, что мы случайно не сгруппируем неверные данные вместе.

Очистка имен

Давайте посмотрим на имя и фамилию архитектора: Бенуа Н.Л.

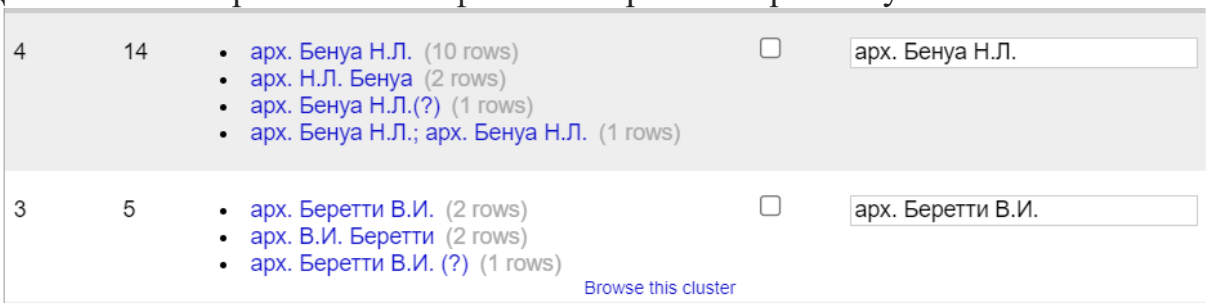

Рисунок 53 - Переменные по Бенуа Н.Л.

Мы видим, что есть четыре варианта этого имени в столбце «Values in Cluster» и предложение о том, как мы можем форматировать имя в будущем «New Cell Value» (см. Рисунок 53).

Если вы хотите изменить текст в столбце «New Cell Value», то сейчас самое время сделать это. Работа с кластерами может зависеть от ваших задач, здесь предлагаю оставить имя, которое предлагает OpenRefine. Теперь поставим галочку рядом с «Merge». Ваш экран должен выглядеть так, как показано ниже(см. Рисунок 54).

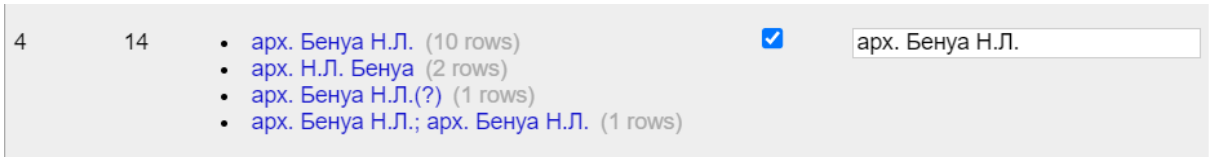

Рисунок 54 - Изменение переменной

Для обновления таблицы нажмите «Merge Selected and Re-Cluster».

Вы заметите, что имя исчезло из нашего окна, это произошло потому, что OpenRefine просто переименовал варианты, которые мы видели слева, на новое значение ячейки, которое мы выбрали справа, то есть мы только что очистили данные.

Давайте сделаем то же самое для нашего следующего имени, Беретти В.И. Текст в столбце «New Cell Value» должен быть «арх. Беретти В.И.». Нажмите «Merge Selected and Re-Cluster» (см. Рисунок 55).

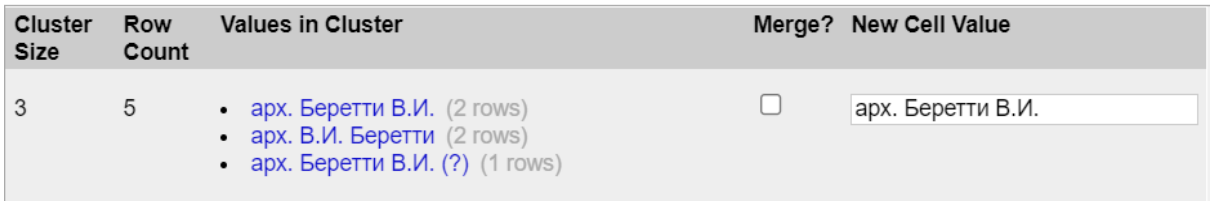

Рисунок 55 - Изменение переменной 2

Если вы уверены, что все предположения, которые сделал OpenRefine, верны, то можно выбрать все элементы кнопкой «Select All» и задать им новое значение (см. Рисунок 56).

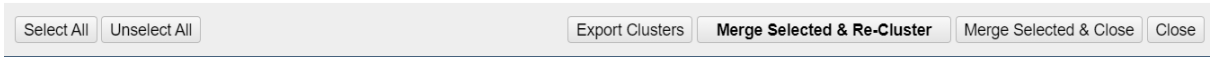

Рисунок 56 - Нижнее меню Cluster and edit

Когда вы закончите с этим набором имен, вы должны увидеть следующий экран (см. Рисунок 57).

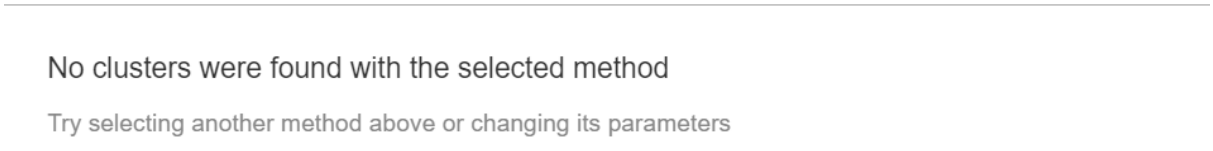

Рисунок 57 - Меню после завершения изменения переменных

Экран выше означает, что мы удалили все имена, которые выбрал данный алгоритм. Дальше вы можете проверить не только столбец с архитекторами,

но и название культурных объектов, используя инструмент «Cluster and Edit». Также вы можете посмотреть на разницу работы алгоритмов и поэкспериментировать с другими столбцами.

Разделение ячеек с множественными значениями

При работе с данными можно заметить, что в одной ячейке может находится несколько переменных (см. Рисунок 58).

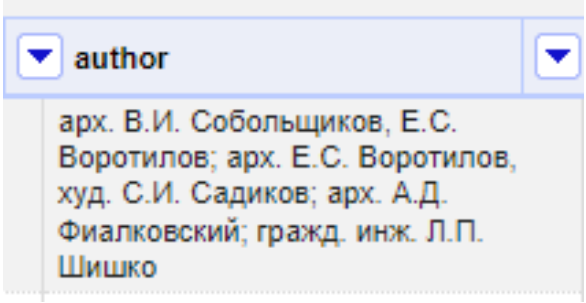

Рисунок 58 - Пример переменной

Чтобы разделить сущности, можно использовать функцию, которая называется «Split multi-valued cells». Для выбора этой функции перейдите во вкладку «Edit cells» (см. Рисунок 59).

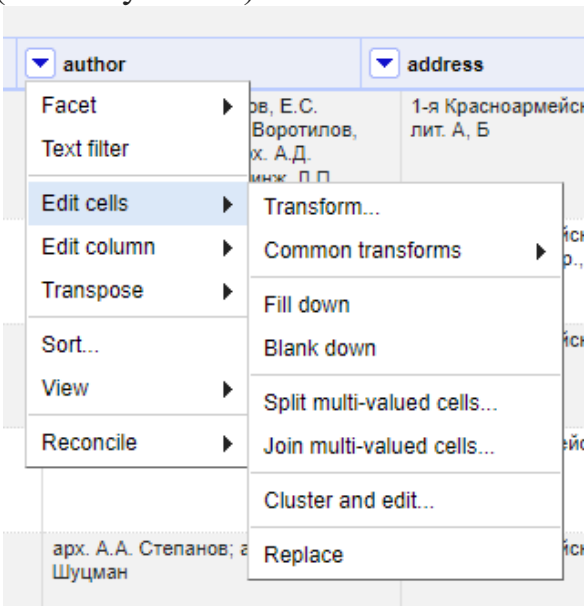

Рисунок 59 - Функция Edit cells

В меню вы сможете выбрать разделитель для множественных значений. Для этого нужно посмотреть на данные и выбрать подходящий для вас. После определения разделителя необходимо нажать «OK» (см. Рисунок 60).

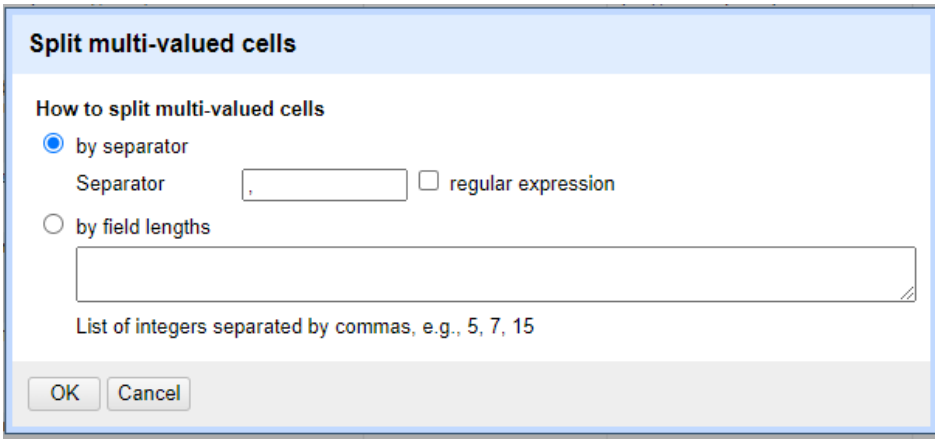

Рисунок 60 - Экран выбора разделителей

После разделения сущностей можно увидеть авторов, которые все равно находятся в одной ячейке, потому что кто-то использовал в качестве разделителя точку с запятой. Мы можем воспользоваться функцией ещё раз, поменяв значение разделителя на точку с запятой (см. Рисунок 61).

name	name_object	date	author
edit	Здание Консисторского управления Могилевской Римско-католической архиепархии с костелом Успения Девы Марии (Духовная семинария Римско-католической архиепархии)	1870-1873; 1878-1879; 1896-1897; 1900-1902	арх. В.И. Собольщиков
			Е.С. Воротилов
			арх. Е.С. Воротилов
			худ. С.И. Садиков
			арх. А.Д. Фиалковский
			гражд. инж. Л.П. Шишко

Рисунок 61 - Результат разделения

Теперь каждый архитектор или скульптор имеет свою собственную строку. Благодаря этой функции мы можем улучшить качество подготовки данных, потому что мы будем редактировать не кластеры имен, а отдельных личностей. После работы с отдельными сущностями их также можно обратно совместить, выбрав функцию «Join multi-valued cells», после чего выберете нужный разделитель и нажмите «ОК»

Очистка отдельных записей

Давайте вернемся и посмотрим на наши данные, чтобы увидеть, сколько еще манипуляций нам нужно сделать. Бывают случаи, когда легче работать с данными точечно. Для этого можно использовать функцию «Facet». Выберите интересующую вас колонку и нажмите «Facet» → «Text Facet» (см. Рисунок 62).

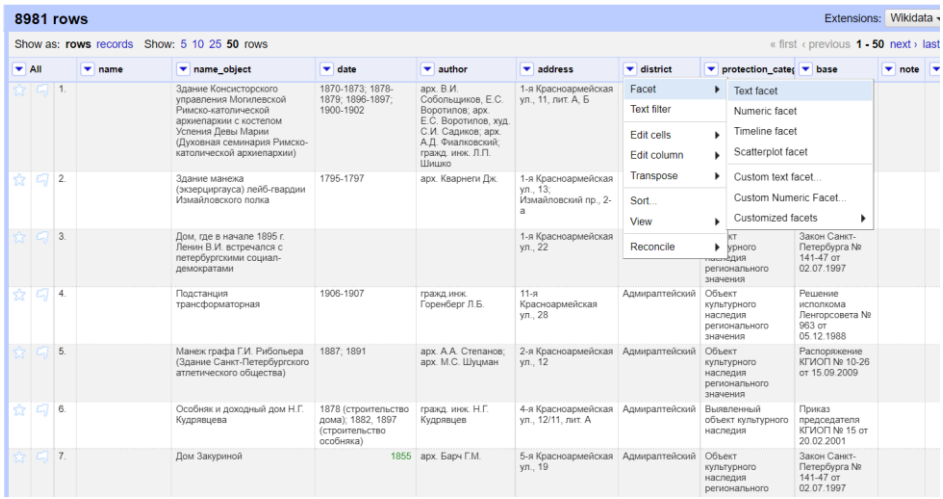

Рисунок 62 - Text Facet

Слева появиться частотный список значений, которые есть в вашей колонке. Мы можем очистить их вручную, просто щелкнув «Edit» рядом с именем в окне «Text Facet» и переименовав названия районов, которые мы хотим изменить (см. Рисунок 63). К счастью, в данных примера нам менять ничего не надо, да и оптимальным такой способ не назовешь, учитывая количество данных.

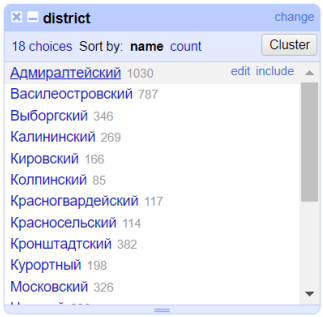

Рисунок 63 - Редактура отдельных сущностей

Фильтрация

Чтобы найти определенные слова или объекты, которые вас интересуют, можно использовать функцию «Text Filter». Например, я хочу найти все статуи, которые упоминаются в этом датасете (см. Рисунок 64).

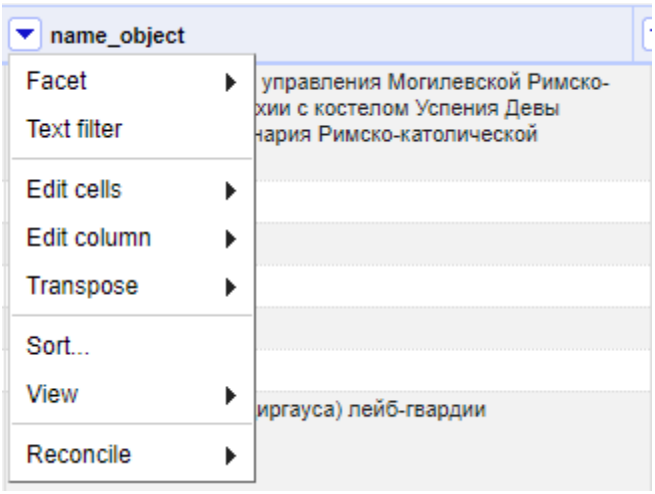

Рисунок 64 - Text filter

Для этого достаточно просто ввести слово в фильтр, и он отберет все подходящие объекты. Важно сказать, что фильтров может быть несколько. Таким образом, мы можем выбрать все статуи в Центральном районе (см. Рисунок 65).

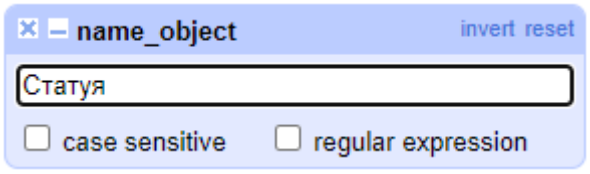

Рисунок 65 - Ввод слова для работы фильтра

Изменение переменных

Если вы обратили внимание, то в столбце «author» есть пустые ячейки, в которых не указаны архитекторы/авторы памятников, но есть отдельные ячейки, где использована форма «автор не установлен». Давайте заменим все пустые ячейки на текст «автор не установлен».

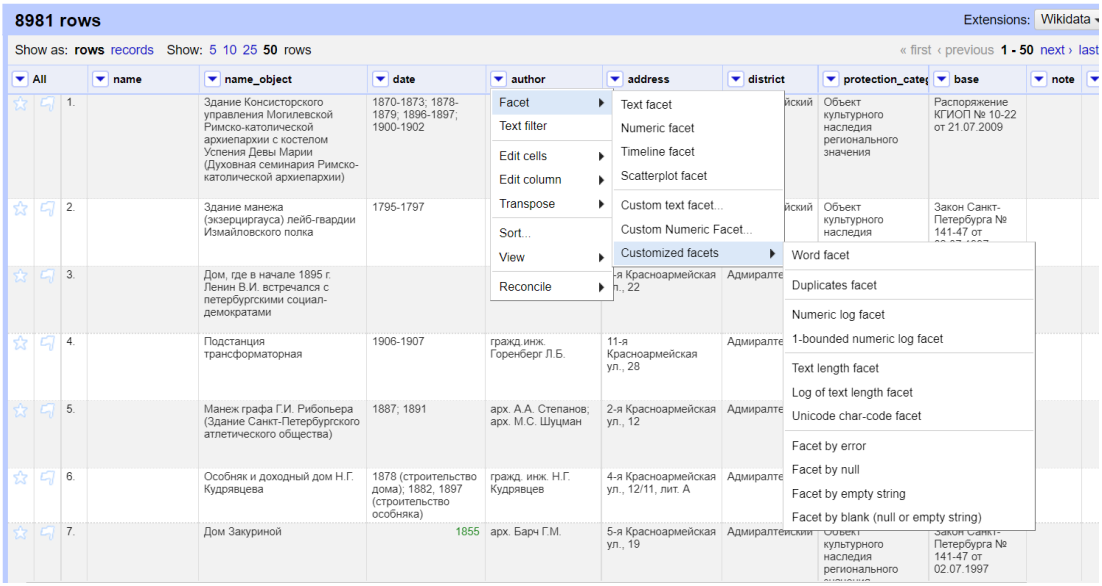

Рисунок 66 - Facet by blank(null) or empty string

Для этого необходимо выбрать все пустые значения, выберите «Facet» → «Customized facets» → «Facet by blank(null) or empty string» (см. Рисунок 66). Слева на экране появится следующее окно (см. Рисунок 67).

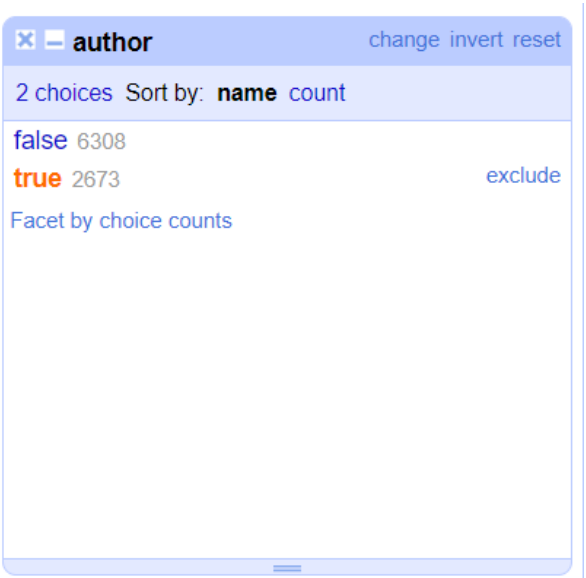

Рисунок 67 - Пустые значения

Вам необходимо выбрать значение «true», так как OpenRefine использует условные выражения для проверки запросов. В нашем случае мы нашли 2673 пустых значения, с которыми мы будем работать в дальнейшем. Дальше необходимо перейти выбрать «Edit cells» → «Transform» в столбце «author» (см. Рисунок 68).

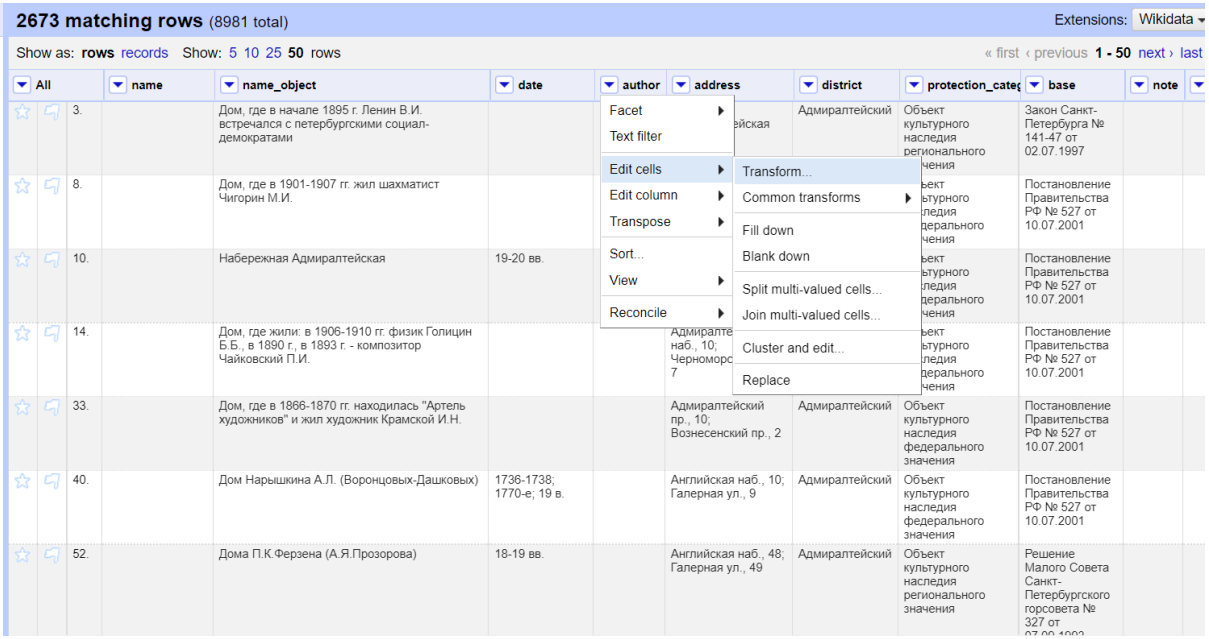

Рисунок 68 - Transform

Перед вами откроется меню для редактирования значений внутри ячеек, как было заявлено ранее, мы будем использовать текст «автор не установлен». OpenRefine поддерживает разные языки, в том числе Python, так что если вы будете писать более сложные выражения, то можете сменить язык в строке «Language». Для того, чтобы изменить значение переменных, напишите в окне «автор не установлен» и нажмите OK. Окно «Preview» позволит вам посмотреть на результат до принятия изменений (см. Рисунок 69).

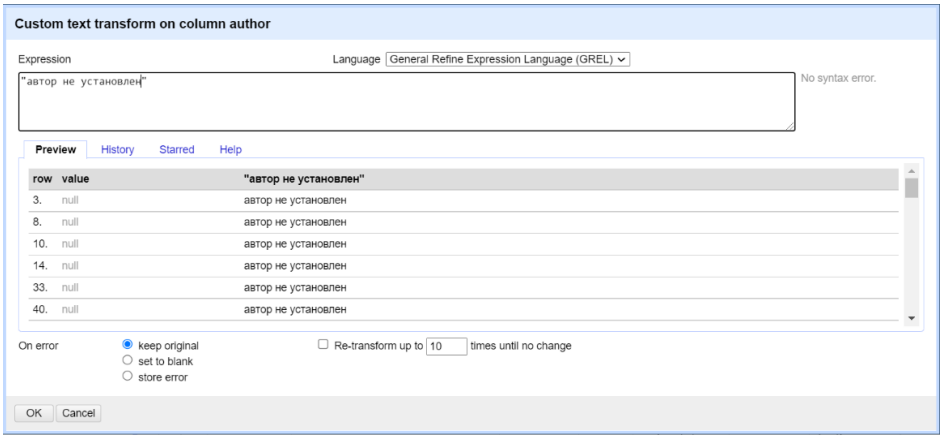

Рисунок 69 - Окно изменения переменных

В дальнейшем мы увидим, что число пустых значений станет равняться 0 (см. Рисунок 70).

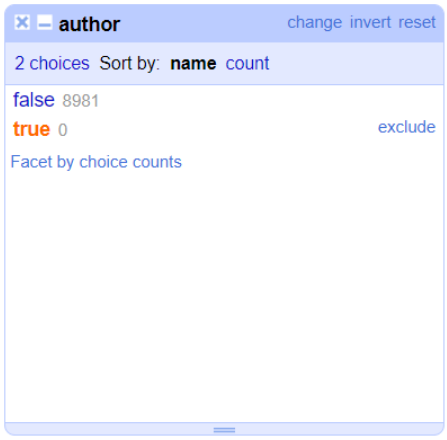

Рисунок 70 - Результаты преобразований

Теперь можно закрыть это окно и продолжать работу с частично исправленным датасетом.

На этом наша глава заканчивается, но вы можете видеть, что осталась задача по работе с датами. Если вы хотите разобраться с этим, то советую вам обратиться к курсу по работе с OpenRefine (https://openrefine.org/documentation.html). Также важно понимать, что, работая с социально-значимыми данными, мы должны нести ответственность за их изменение и дальнейшую интерпретацию. К примеру, можно заметить знаки вопроса около имен некоторых архитекторов при работе в инструменте «Cluster and Edit», что означает неоднозначность в определении авторства некоторых памятников и зданий. И вам, как исследователям, предстоит принять решения по кластеризации и обновлению этих данных, то же самое касается и дат, с которыми вы поработаете самостоятельно.

Экспорт файла

Когда вы закончите, вы можете экспортировать очищенный набор данных в формате CSV, нажав «Export» в верхней части экрана и выбрав «Comma-separated value».

Ход работы

Если вы поймете, что сделали что-то неправильно, то можно перейти в меню «Undo / Redo» и отменить изменения (см. Рисунок 71). Меню также помогает следить за изменениями, которые вы совершаете с данными.

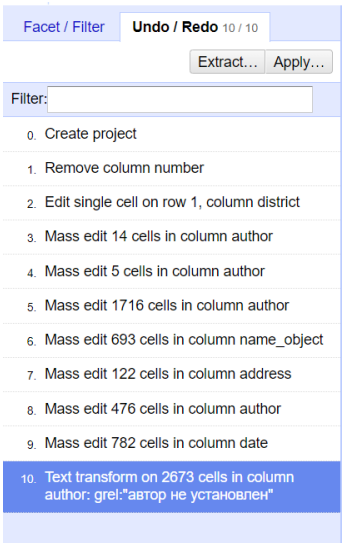

Рисунок 71 - Окно прогресса преобразований

Источники

1. Официальная документация OpenRefine [Электронный ресурс]

Режим доступа: https://openrefine.org/documentation.html

Тест 3

1. Для чего используется OpenRefine?

A. Для построения социальных сетей

B. Для подготовки данных и первичного анализа

C. Для создания визуализаций

2. Как называется функция для унификации похожих данных?

A. Merge

B. Cluster and Edit

C. Filter

3. Какой инструмент подойдет для изучения текстовых данных?

A. Numeric facet

B. Timeline facet

C. Text facet

Глава 4. Gephi

Введение

Знакомство с анализом социальных сетей (Social Network Analysis) следует начать с определения самой области. Анализ социальных сетей — это область анализа данных, в которой для понимания социальных структур используются сети и теория графов. Методы также могут применяться к сетям за пределами социальной сферы.

Для построения графиков SNA требуются два ключевых компонента: участники и отношения между ними. Обычно методы SNA применяются в Интернете: веб-страницы в Интернете часто ссылаются на другие веб-страницы - либо на их собственном веб-сайте, либо на другом веб-сайте. Эти ссылки можно рассматривать как отношения между участниками (веб-страницами). Фактически, это ключевой компонент архитектуры поисковой системы.

Сети окружают нас повсюду - например, дорожные сети, интернет-сети и социальные сети, такие как Facebook. Хотя в этой главе основное внимание уделяется анализу социальных сетей (ключевое слово: социальные), изучение этих методов добавит ценные инструменты в ваш арсенал, что позволит получить новое представление о различных источниках данных.

Основные понятия

Для знакомства с дисциплиной нам требуется определиться с главными понятиями. И предлагаю снова начать понимание того, что такое сеть. **Сеть** — сложная закрытая система, которая состоит из сущностей (узлы/вершины) и отношений (ребра/связи) между этими сущностями. Как можно заметить, определение достаточно общее, и под него могут подходить любые закрытые системы, в которых существуют связи между элементами сети. Используя это определения, сетью можно назвать взаимоотношения сотрудников компании, любую социальную сеть в интернете и даже железнодорожную сеть, перечисление можно продолжать достаточно долго.

Сеть состоит из двух элементов — узел и грань.

Узел — сущность, представляющая множество акторов (может быть чем угодно в зависимости от ваших задач). Может быть персоной, книгой, персонажем пьесы и т.д.

Атрибуты узла — качества, присущие акторам. Под качествами следует понимать свойства, которыми можно описать акторов. Для персоны это может

быть место получения образования, для книги - количество страниц, для персонажа - сцены, в которых он появляется.

Типы узла — количество множеств разного типа в рамках одной сети. Сеть может состоять не только из книг, но и авторов, которые их написали. В таком случае граф называется бимодальным или мультимодальным в зависимости от количества типов узлов в рамках одной сети.

Бимодальный/мультимодальный граф — граф, в котором присутствует несколько типов узлов.

Ребро — связь между узлами, описывающая формат отношений, или то, как они взаимодействуют. Гранью может быть коммуникация между пользователями,

Бывают ситуации, когда между узлами может быть несколько ребер, представляющих разный формат отношений, в таком случае граф определяется как **мультиграф**.

Метрики

Если вы забудете, что определяет каждая из метрик, то можете обратиться к этому разделу. Более подробно они будут разобраны в практической части, когда мы начнем работать с Gephi.

Степень узла — количество связей узла с другими узлами сети.

Центральность по посредничеству — это количество кратчайших путей, в которые входят узел, деленное на общее количество кратчайших путей.

Центральность по близости — это среднее количество переходов через узел, необходимых для достижения каждого другого узла в сети.

Модулярность — используется для поиска сообществ.

Вес ребра — количества определенных связей между двумя узлами.

Размер сети — количество узлов в сети.

Плотность сети — это количество ребер, деленное на общее количество возможных ребер.

Длина пути — количество переходов между начальным и конечным узлами.

Клика — три узла и больше, все связанные между собой (обычно сообщество).

Практическая часть

В этой главе объясняется, как визуализировать набор данных в Gephi и получить интересную информацию из этой визуализации. Набор данных для практики вы можете найти по этой ссылке: https://dracor.org/rus/chekhov-tri-sestry, однако не стесняйтесь позже опробовать знания на своем собственном наборе данных.

Gephi — это интерактивный программный инструмент для визуализации сетей, которые в дальнейшем можно исследовать. Его разработали Матье Бастиан, Эдуардо Рамос Ибаньес, Матье Жакоми, Сезари Бартусиак, Джулиан Бликке, Патрик Максуини, Андре Паниссон и др. (2017 г., версия 0.9.2). На своем веб-сайте они заявляют: Gephi — это инструмент для аналитиков данных и ученых, стремящихся исследовать и понимать графы. Цель инструмента состоит в том, чтобы помочь аналитикам данных проверять гипотезы, интуитивно обнаруживать закономерности, изолировать структурные особенности или неисправности во время поиска данных. Это дополнительный инструмент к традиционной статистике, поскольку теперь признано, что визуальное мышление с интерактивным интерфейсом облегчает рассуждение. Это программное обеспечение для анализа данных, парадигма которого появилась в области исследований Visual Analytics, об этом можно прочитать на сайте gephi.org в разделе «Features». Gephi визуализирует ваши данные в режиме реального времени, а это значит, что вы можете видеть, как график обретает форму прямо у вас на глазах. Инструмент используется для исследования и визуализации реляционных данных или сетей.

Важные моменты

1. В Gephi нет кнопки «отменить» или «повторить». CTRL + Z не работает! (Обещали, что появится в версии 1.0)
2. Gephi не перечисляет и не сохраняет ваши действия и настройки. Таким образом, вам нужно где-то записывать ход работы, чтобы вы могли вспомнить, что вы делали, если позже откроете свой проект.
3. Часто сохраняйте свой проект, используйте опцию «Сохранить как»! Также сохраните промежуточные результаты, так как кнопки отмены нет. Если вы внесете в свою визуализацию большие изменения, которые не принесут желаемых результатов, у вас все равно останется старая версия.
4. Когда вы запускаете алгоритм на огромном наборе данных, дайте ему поработать подольше.
5. В Gephi вы можете использовать различные алгоритмы расположения узлов. ForceAtlas2 — наиболее используемый алгоритм. Если вы используете Gephi для академических исследований, вам необходимо понимать, как работает алгоритм, хотя бы на базовом уровне. Для этого вы можете прочитать об алгоритмах в статьях, написанных разработчиками.

6. Чтобы другие могли понять визуализацию Gephi, вам потребуется предоставить дополнительную контекстную информацию.
7. При наведении указателя мыши на функции в Gephi появится небольшое желтое поле, в котором кратко объясняется, за что она отвечает.
8. Не бойтесь пробовать разные варианты и смотреть, как соответственно изменяются ваши визуализации. Чтобы освоить Gephi, вам просто нужно попробовать разные вещи.

Gephi визуализирует сети. Но что такое сеть? Давайте посмотрим на следующий пример (см. Рисунок 72).

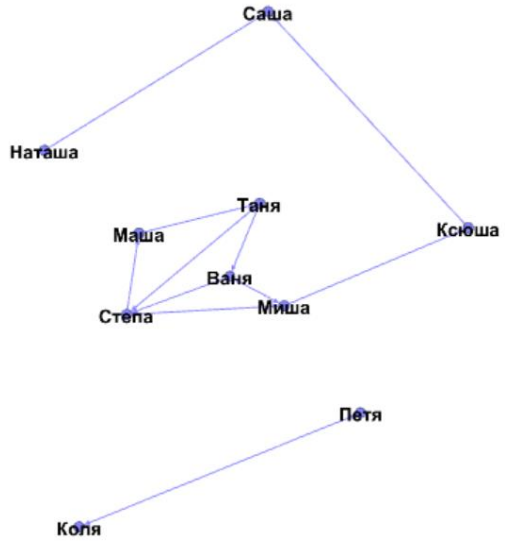

Рисунок 72 - Пример сети

Эта визуализация показывает взаимодействия между людьми. Например, Степа общался с Машей, Таней, Мишей. Миша, в свою очередь, также общался со Степой, Ваней и Ксюшей. Когда эти люди взаимодействуют с другими, они образуют сеть. Однако Коля и Петя общались только между собой. Они не взаимодействовали с другими участниками сети. Таким образом, это означает, что они не связаны напрямую с основной сетью, а только друг с другом.

В Gephi участников сети представляют узлами (nodes), а связи между этими узлами — ребрами (edges). Иногда у ребер также есть вес. Этот вес указывает на силу связи между узлами. Эта сеть не имеет взвешенных ребер, но связи между узлами могут быть двусторонними. Например, вы можете видеть на рисунке две стрелки на линии, идущей от Степы к Тане. Это означает, что Таня взаимодействовала со Степой, а Степа также общался с Таней. Это ориентированный граф, но графы также могут быть неориентированными. В неориентированных графах по умолчанию рёбра имеют двустороннюю связь, поэтому не имеет значения, в каком направлении они движутся. В

ориентированных графах ребра могут иметь односторонние и двусторонние связи, в этом случае направление имеет значение.

Для этой главы мы используем готовый набор данных, подготовленный в рамках проекта DraCor, однако вы также можете выполнить большинство шагов со своим собственным набором данных. Прежде чем импортировать набор данных в Gephi, вам необходимо подготовить его. Чтобы сделать визуализацию в Gephi, ваш набор данных всегда должен содержать по крайней мере источник (source) и цель (target). Вы также можете указать вес (weight), который показывает силу соединения. Более того, вы можете выбрать тип графа: ориентированный или неориентированный. Для подготовки данных вам потребуется программное обеспечение для работы с электронными таблицами, такое как Microsoft Excel или OpenOffice Calc. Вам необходимо сохранить вашу электронную таблицу как файл «.csv», чтобы Gephi мог ее прочитать. Помимо .CSV, Gephi также может читать: GEXF, GDF, GML, GraphML, Pajek NET, GraphViz DOT, CSV, UCINET DL, Tulip TPL, Netdraw VNA и электронные таблицы.

Импорт данных

Во-первых, давайте откроем Gephi и начнем новый проект! Щелкните значок рабочего стола Gephi на рабочем столе. Перейдите в «File» (верхний левый угол) → «New Project» или щелкните «New project» на появившемся экране.

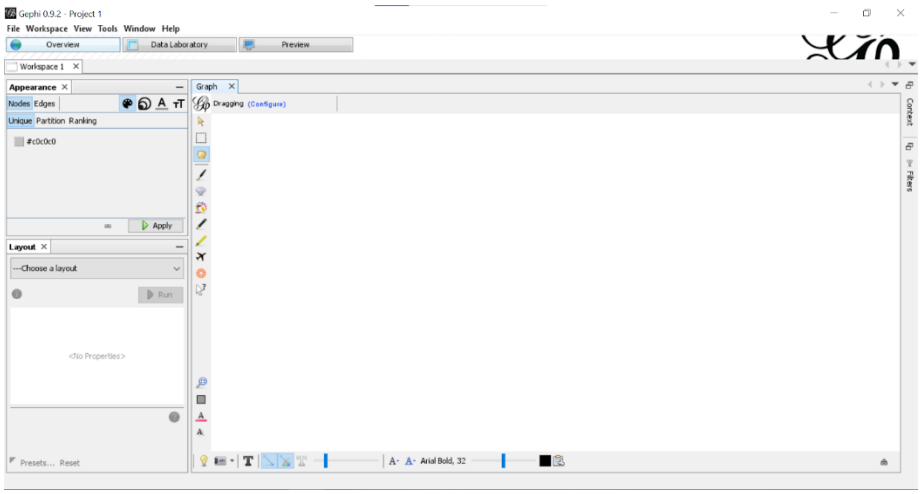

Рисунок 73 - Вид интерфейса

В зависимости от вашей операционной системы может выглядеть немного иначе, но все должно быть там (см. Рисунок 73). Если это не так, попробуйте нажать «Window» в строке меню, чтобы активировать недостающие элементы.

Чтобы импортировать набор данных, перейдите в «File» → «Import spreadsheet» → Выбрать файл.

Теперь вам нужно убедиться, что ваш набор данных импортирован правильно. Выберите разделитель. Разделителем часто является табуляция или точка с

запятой. В нашем случае формат файла «.csv», так что разделителем является запятая. Просто убедитесь, что в предварительном просмотре ваши столбцы аккуратно расположены. Файл содержит набор связей между персонажами пьесы. Таким образом, выберите для импорта таблицу ребер. Наконец, кодировка - UTF-8 (см. Рисунок 74).

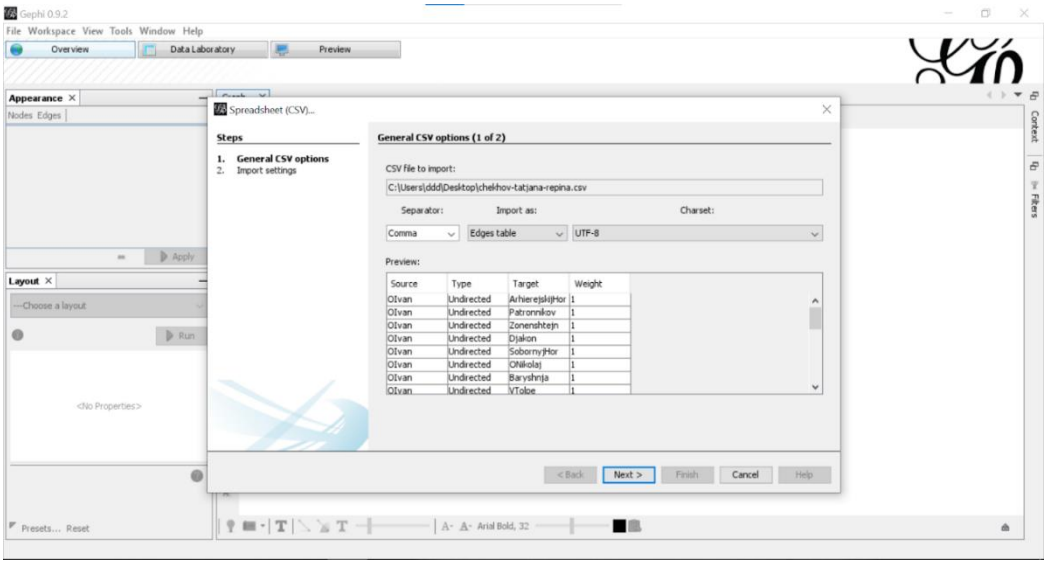

Рисунок 74 - Процесс импорта

Далее идут настройки импорта. Мы уже определили «Target» и «Source» в нашем файле, и Gephi считывает это автоматически. Теперь нам нужно удостовериться, что мы хотим импортировать все колонки, но этот шаг уже зависит от целей вашего исследования, а сейчас мы просто нажмем «Finish», чтобы завершить импорт.

Затем вы увидите отчет об импорте. В этом отчете перечислены проблемы, если они есть, и вы также можете увидеть, сколько ребер и узлов будет создано. Мы используем тип графика «Undirected» (ненаправленный). Если у вас возникнут проблемы при импорте набора данных, Gephi перечислит проблемы и строки, в которых возникает проблема. Иногда вам, возможно, придется вернуться к своим данным, чтобы исправить ошибки. Обязательно сделайте это перед визуализацией.

Отлично, теперь ваши данные импортированы, и мы можем продолжить заниматься полезными делами! Вам будет представлен экран, который будет выглядеть примерно так (см. Рисунок 75).

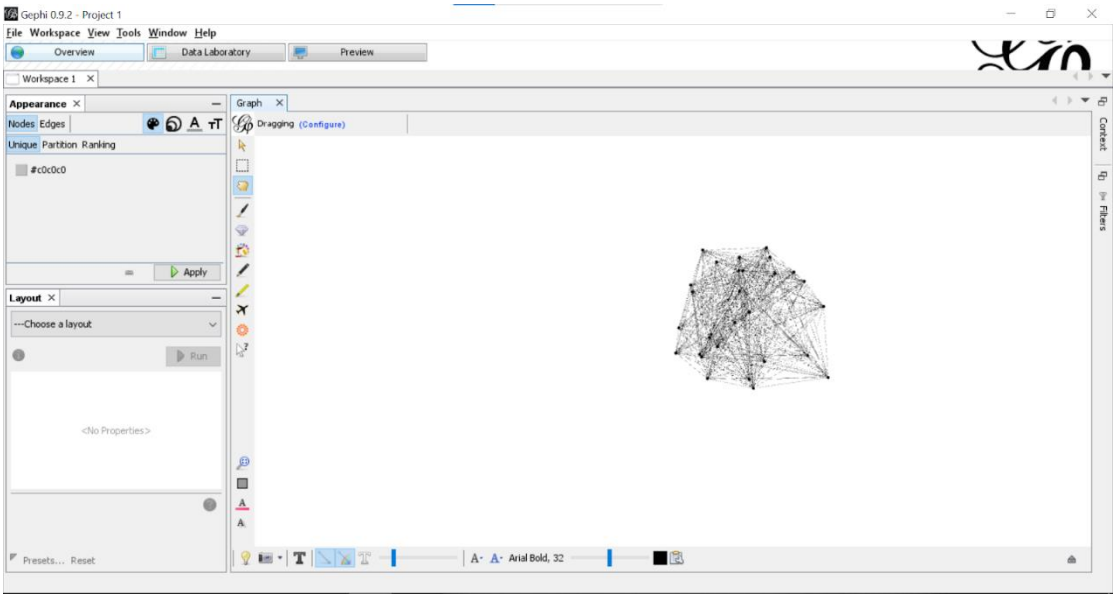

Рисунок 75 - Рабочая область после импорта

В окне «Graph» появились данные, которые вы только что импортировали. Вкладка «Overview» — это наша основная рабочая область. Здесь мы проанализируем и визуализируем наши данные.

Наконец, сохраните ваш проект. Выберите «File» → «Save» → Придумайте понятное название→ Сохраните его как файл Gephi.

В следующем разделе мы получим некоторые практические знания для исследования и визуализации сети в Gephi. Если вы забыли, что означает какой-либо из терминов, ознакомьтесь с глоссарием в начале главы.

Вторая часть. Работа с графом

В этом разделе мы собираемся визуализировать нашу сеть и опробовать различные функции Gephi для исследования. Сначала мы обсудим лабораторию данных и обзорную панель.

В верхнем левом углу вы увидите три разные вкладки: «Overview», «Data Laboratory» и «Preview» (см. Рисунок 76).

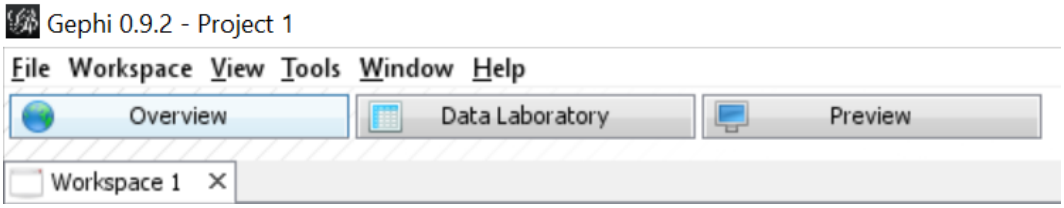

Рисунок 76 - Три рабочих окна

Лаборатория данных (Data Laboratory)

Начнем с лаборатории данных, в которой находится ваш набор данных. Чуть ниже основных вкладок вы увидите вкладки «Nodes» и «Edges». Щелкните по ним и обратите внимание на разницу между этими вкладками. Мы импортировали таблицу ребер, однако Gephi также автоматически создаст таблицу узлов, эта таблица содержит всех персонажей, которые были перенесены из таблицы ребер (см. Рисунок 77).

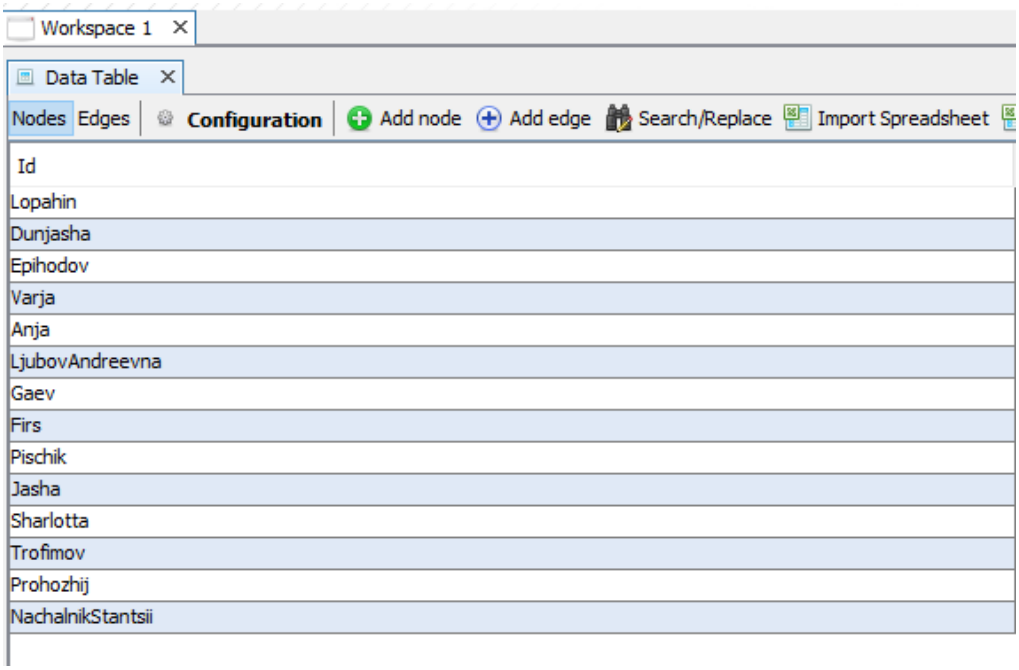

Рисунок 77 - Узлы сети

Затем взгляните на таблицу ребер, здесь вы снова можете увидеть «Source» и «Target». Gephi автоматически создал столбцы «Type» и «Id». Мы видим столбцы «Weight», а также два других пустых столбца «Label» и «Interval». В нижнем меню есть различные варианты работы с колонками. Здесь вы можете добавить новый столбец, удалить и дублировать столбцы и т. д.

Панель обзора (Overview)

Перейдите на вкладку «Overview» в верхнем левом углу, это основное рабочее пространство для визуализации нашей сети. Давайте разберем различные панели на этой вкладке, чтобы понять, что делает каждая из них.

1. «Appearance»

Здесь вы можете определить цвет и размер узлов, ребер и меток. Внешний вид разделен на две вкладки: «Nodes» и «Edges». На этих вкладках есть несколько вариантов преобразования: одиночные узлы, группировка и ранжирование (см. Рисунок 78).

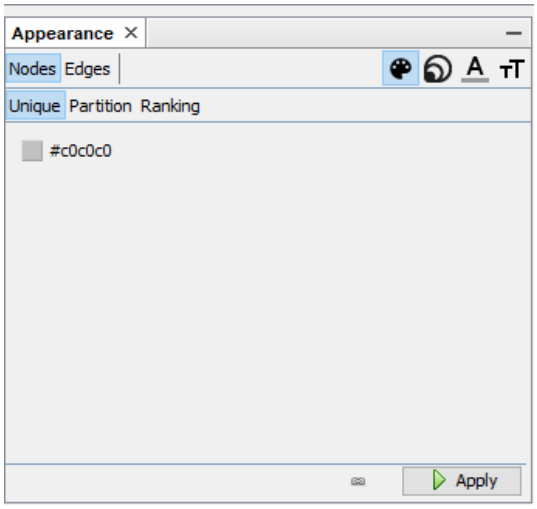

Рисунок 78 - Appearance

2. «Layout»

Здесь вы определили алгоритм, который вы будете использовать при работе с сетью, он придаст сети ей форму (см. Рисунок 79).

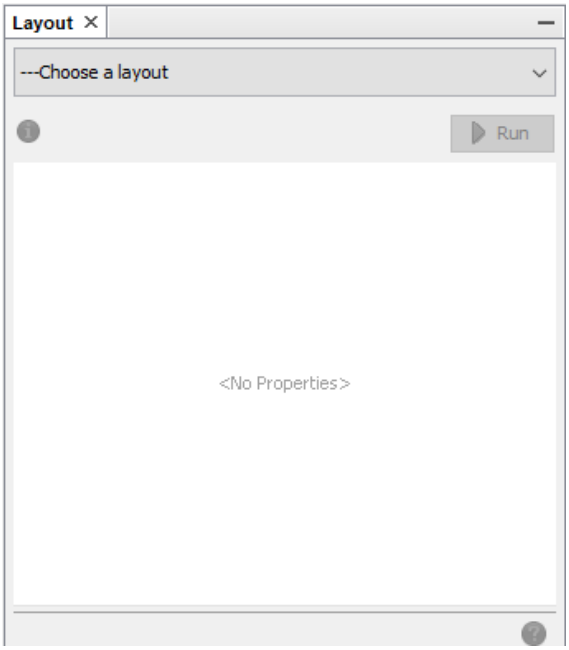

Рисунок 79 - Layout

3.«Graph»

Здесь отображается ваша сеть. Кроме того, левая и нижняя строки меню позволяют вносить изменения в визуализацию. Удерживая правую кнопку мыши, перетаскивайте мышь, чтобы перемещаться по графику. Увеличение можно выполнить с помощью колеса мыши. Если вы потеряли график, используйте значок лупы в нижнем левом углу, чтобы снова отцентрировать визуализацию (см. Рисунок 80).

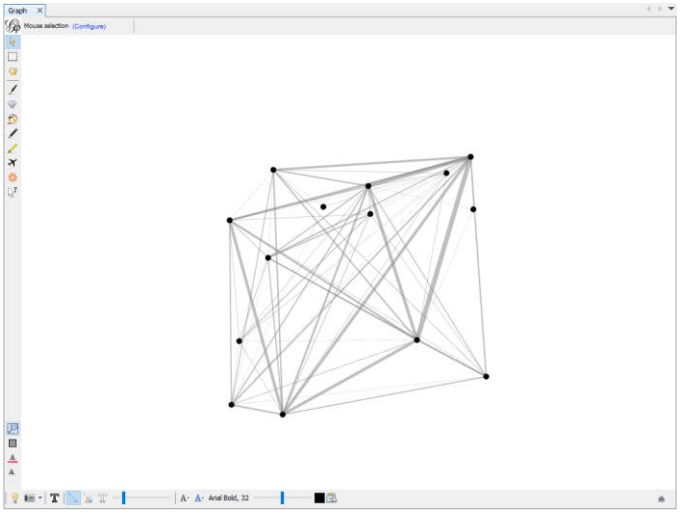

Рисунок 80 - Graph

4. «Context»

Здесь отображается количество узлов и ребер. В огромных сетях или при использовании фильтров это также покажет вам процент узлов и ребер, отображаемых в текущей визуализации. Пока показаны все.

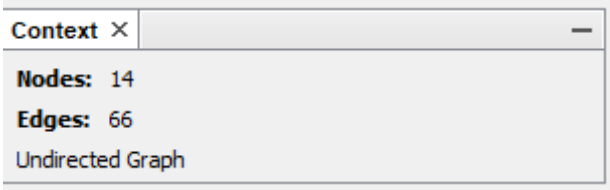

Рисунок 81 - Context

5. «Filters, Statistics»

Вы можете использовать фильтры для работы с определенными узлами, ребрами или атрибутами. На вкладке «Statistics» вы можете применить алгоритмы для анализа набора данных (см. Рисунок 82).

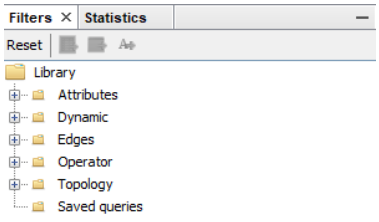

Рисунок 82 - Filters, Statistics

6. Queries

Здесь отображаются использованные фильтры, и здесь вы, кроме того, можете настроить все фильтры и применить их к визуализации.

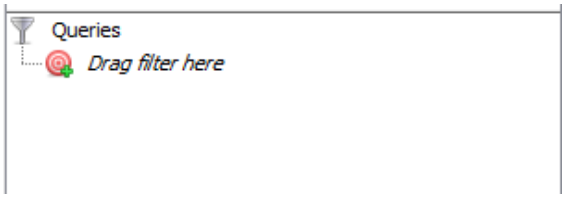

Рисунок 83 – Queries

Визуализация и изучение нашего набора данных

Начнем с применения алгоритма компоновки ForceAtlas2 к набору данных. На панели алгоритмов компоновки выберите ForceAtlas2 и нажмите «Выполнить». Все узлы притянутся к друг другу, но это можно исправить, скопировав параметры с картинки ниже (см. Рисунок 84). В правом нижнем углу видно, что алгоритм работает. Вы сразу увидите изменение визуализации после того, как нажмете «Run».

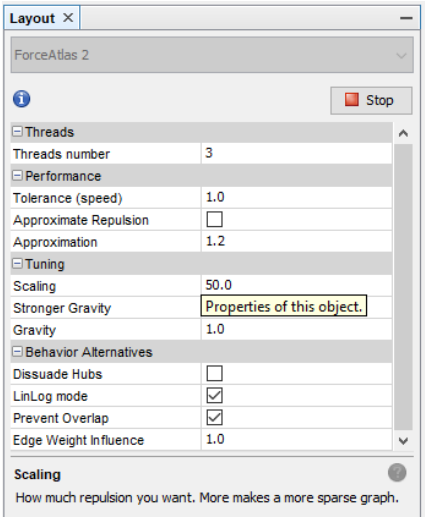

Рисунок 84 - Настройки Layout

Этот алгоритм определяет положение узлов в сети относительно других узлов. Узлы с большим количеством связей располагаются ближе друг к другу.

Под выбором алгоритма, как уже было показано выше, находится меню, в котором вы можете настроить свойства визуализации, такие как масштабирование и гравитация. Наведите указатель мыши на конфигурации, чтобы увидеть, что они могут делать.

Оставьте ForceAtlas2 запущенным, отметьте поля «Lin Log» и установите «Scaling» на 50, и посмотрите, как график меняется прямо на ваших глазах. Опять же, дайте алгоритму поработать пару минут. Помните, что вы можете

использовать лупу в панели «Graph», чтобы отцентрировать график, если он потерялся.

Теперь остановите алгоритм. Также пора сохранить свой проект. Сделайте это, перейдя в «File» → «Save as». Не забывайте почаще сохранять в разные версии! В моем случае график получился выглядит таким (см. Рисунок 85).

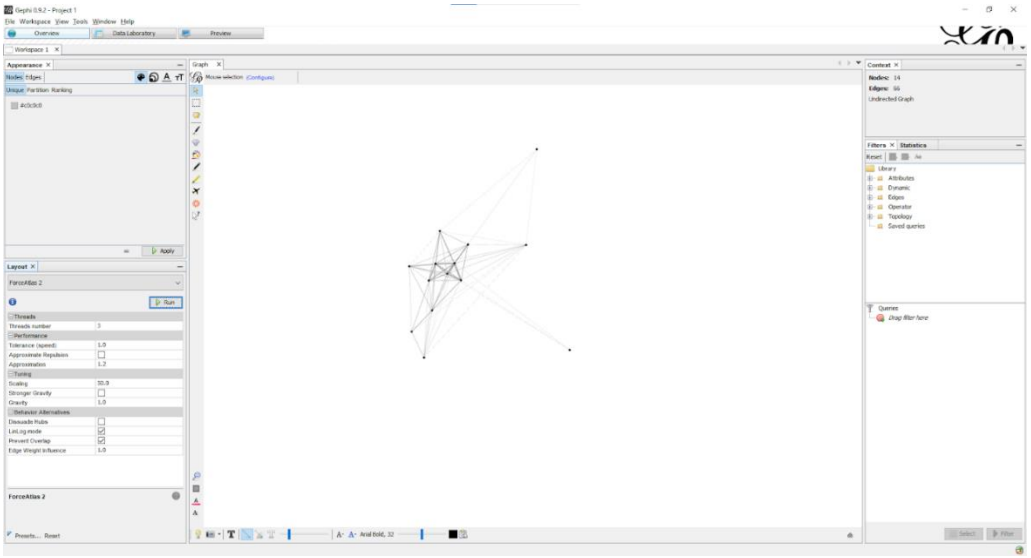

Рисунок 85 -График после применения алгоритма компоновки

Основы: статистика, внешний вид и фильтры

Вы можете видеть, что на визуализации уже виден один большой кластер. Теперь мы можем провести дополнительные вычисления. Перейдите на панель статистики (в правом верхнем углу) и выберите «Average Degree». Вы сразу получите отчет с результатами, его можно закрыть или сохранить на своем компьютере.

Обратите внимание, что на вкладке узла в лаборатории данных автоматически добавилась колонка для измеренной статистики.

Теперь мы можем поменять размер узлов в соответствии с степенью, используя ранжирование. В нынешнем виде нашего графика трудно увидеть отдельные узлы, и все они имеют одинаковый размер. Допустим, нас интересуют наиболее активные участники пьесы, и мы хотим выделить их на графике (см. Рисунок 86). Вы можете изменить размер этих узлов в зависимости от их активности. Перейдите в «Appearance» → «Nodes» → «Size» (вторая вкладка в правом меню) → «Ranking» → Выберите степень → выберите «Мин.: 10, Макс: 50 → Apply». Некоторые узлы стали больше, потому что это узлы с наивысшей степенью, а, следовательно, они чаще появляются в пьесе.

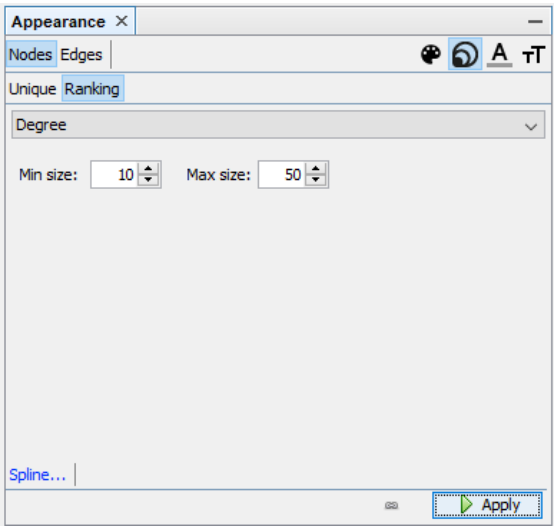

Рисунок 86 - Настройки параметров

Если вы хотите оставить на графике только важных персонажей, то на правой панели перейдите в «Filters» → «Attributes» → «Range» → перетащите «Degree» в запросы (см. Рисунок 87). Текущий диапазон теперь установлен от 3 до 12.

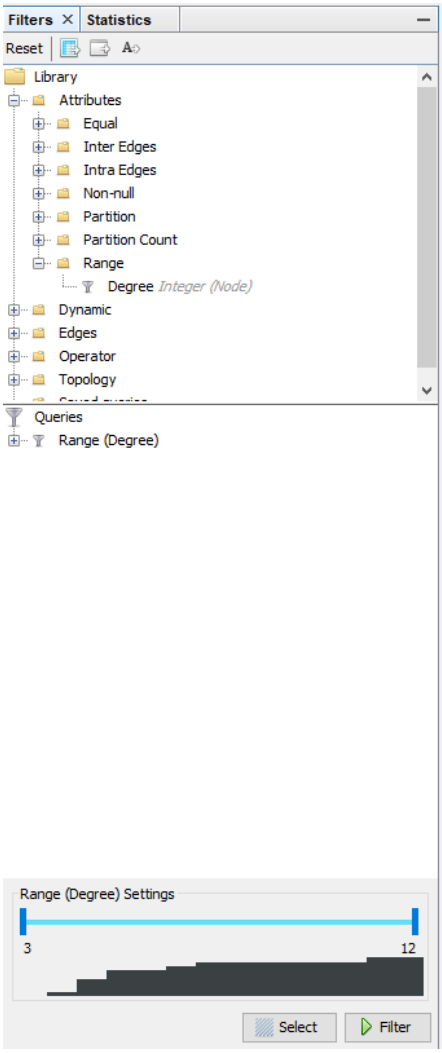

Рисунок 87 - Настройка фильтров

Допустим, нас интересуют персонажи только со степенью 4 или выше. Установите диапазон от 4 до 12 (вы можете перетащить или дважды щелкнуть число и ввести его), нажмите «Filter» и убедитесь, что вы используете кнопку «Enter» на клавиатуре, когда набираете число, чтобы его подтвердить. Визуализация снова изменилась, теперь видно меньше узлов, поскольку мы отфильтровали некоторые из них. Обратите внимание, что в меню «Контекст» теперь показано, сколько узлов и ребер все еще видно. При желании можно отключить фильтр, нажав «Стоп» (см. Рисунок 88).

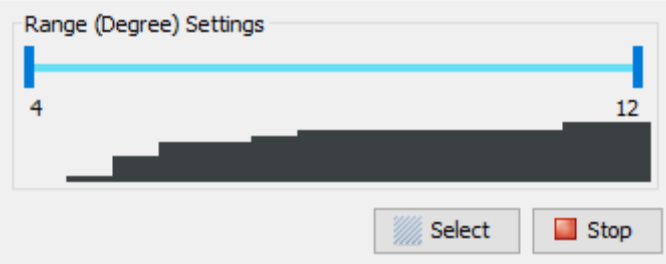

Рисунок 88 - Диапазон значений

Фильтры полезны для обеспечения удобочитаемости визуализации и для фильтрации любых данных, которые вам не интересны, или для более внимательного изучения интересующих вас данных. С помощью фильтров вы можете определить информацию, которую хотите представить в визуализации. Кроме того, Gephi позволяет применять отфильтрованный контент к новому рабочему пространству, выбирая второй значок прямо под вкладками «Filters» и «Statistics».

Однако вы также можете работать в одной рабочей области, но использование большего количества рабочих пространств может быть полезно при визуализации больших наборов данных. После применения фильтра (к новому рабочему пространству) вы снова можете запустить статистику только для отфильтрованного содержимого.

При наведении указателя мыши на узел в рабочем пространстве выделяется узел и его соединения.

Чтобы увидеть, кого представляет узел, мы можем добавить метки с именами, а также отобразить степень узла. На экране графика щелкните значок, похожий на дом, в правом нижнем углу (красный квадрат на снимке экрана). Откроется меню с параметрами «Global», «Edges» и «Labels» (см. Рисунок 89).

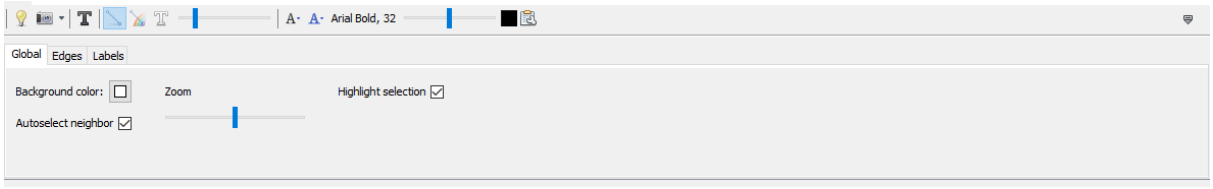

Рисунок 89 - Параметры на нижней панели экрана

Перейдите во вкладку «Labels» и выберите настройки справа, поставьте галочку около «Id» и уберите у «Label», если вы зайдете в лабораторию данных, то поймете, почему мы выбрали именно такие параметры. Также во вкладке «Labels» нажмите галочку около «Node», определите размер шрифта и поставьте цвет, который вам нравится. Теперь, если вы наведете курсор на узлы, вы увидите, кого он представляет.

Более того, если вы хотите видеть подписи всех узлов, не ставьте галочку «Hide non-selected». Это может быть полезно для небольших графиков, но для больших графиков визуализация становится нечитаемой.

Если вы хотите узнать больше про отдельный узел, щелкните на его правой кнопкой мыши и выберите «Выбрать в лаборатории данных». Затем перейдите на вкладку «Data Laboratory», чтобы просмотреть информацию о нем в наборе данных. Попробуйте изучить некоторые другие узлы, которые выделяются, наведя на них курсор и исследуя их также в лаборатории данных.

Поиск сообществ

Одним из способов исследования сетей является поиск сообществ, для этого Gephi реализует алгоритм обнаружения сообщества с измерением модулярности. Перейдите к вычислениям в правой части экрана и запустите подсчет модулярности (см. Рисунок 90).

Рисунок 90 - Модулярность

Не меняйте настройки (см. Рисунок 91), но нажмите ОК. Отчет о модульности показывает, что алгоритмы нашли 3 сообщества.

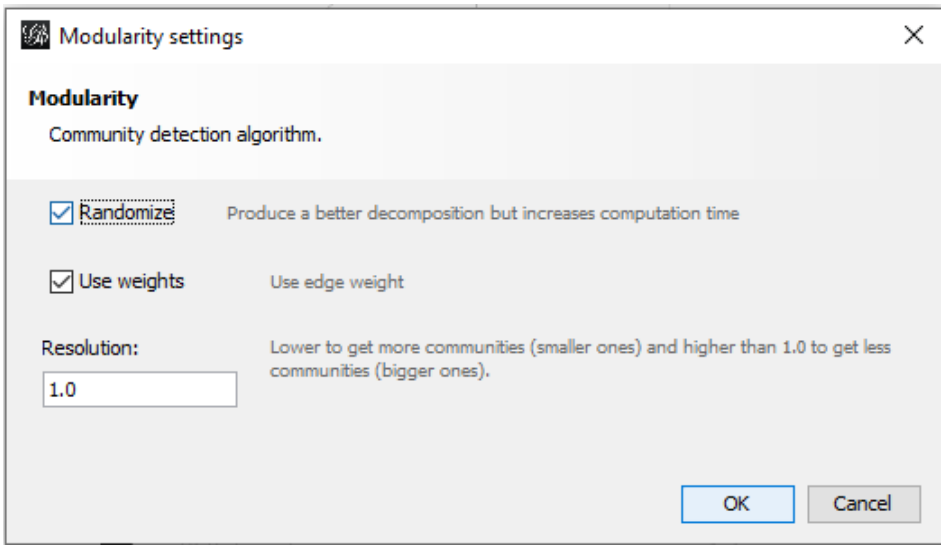

Рисунок 91 Настройки вычислений

Затем перейдите к внешнему виду в верхнем левом углу панели → «Nodes» → «Color» (первый значок) → «Partition» → Выберите атрибут «Modularity» → Нажмите «Apply». Ваши настройки должны выглядеть так (см. Рисунок 92).

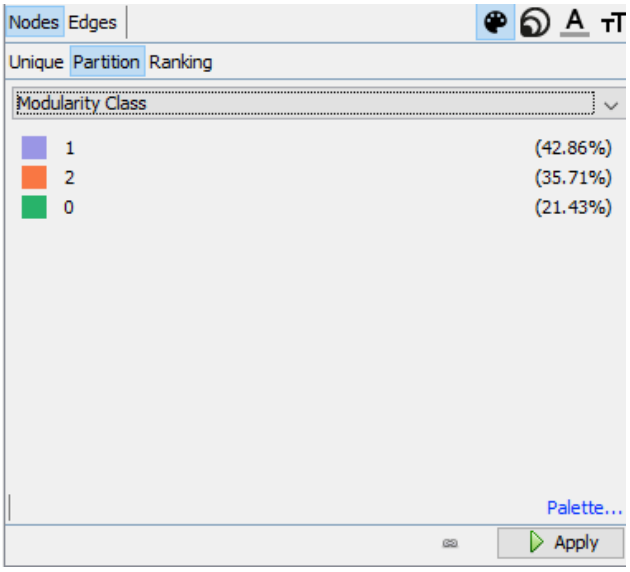

Рисунок 92 - Параметры настроек

Наша визуализация, безусловно, стала более красочной (см. Рисунок 93).

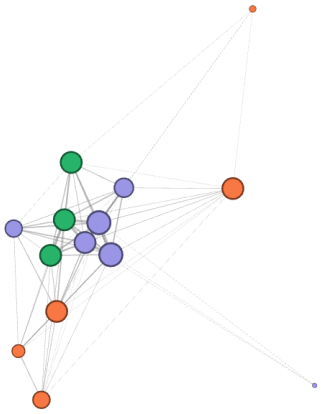

Рисунок 93 - График после применения настроек цвета

Принцип работы алгоритма по поиску сообществ можно описать так: с помощью модулярности вы можете находить сообщества в сети и измерять их силу. Высокая модулярность означает, что группа узлов прочно связана, создавая плотное сообщество, в котором пользователи в основном взаимодействуют только друг с другом, а не с другими сообществами в сети. Сообщества с низкой модулярностью менее плотны и больше взаимодействуют с другими сообществами.

Левое число - это случайное число, присвоенное сообществам. По сути, это число мало что говорит, однако процент важнее. Например, класс модулярности «2» имеет процент 35,71. Это означает, что этот класс модулярности или кластер сообщества охватывает 35,71% сети. Таким образом, это сообщество довольно большое. Имейте в виду, что если вы применяете подсчет модулярности к отфильтрованному графику, то расчеты

проводятся только с учетом видимых узлов. В таком сценарии, когда вы запускаете модулярность на отфильтрованном графе, Gephi будет вычислять модулярность этого графа, а не всей исходной сети!

Поиск информации

Что из этого можно извлечь? Gephi - это инструмент, позволяющий исследовать сети в вашем наборе данных различными способами.

Давайте запустим еще одну статистику в новом рабочем пространстве, которое мы только что создали. Перейдите в статистики, запустите «Network Diameter». Выберите «Undirected» и отметьте «Normalize Centralities».

Опять же, вычисления будет добавлены в лабораторию данных. Здесь измеряются три статистики: центральность по посредничеству, центральность по близости и центральность по эксцентричности.

Центральность по посредничеству: это измерение основано на количестве кратчайших путей между двумя узлами. Это показатель центральности или важности узла в сети. Высокое значение предполагает, что узел соединяет части сети вместе. Меньшее значение означает, что узлы не являются центральными в сети.

Центральность близости: эта мера указывает на близость узла к другим узлам. Более высокое значение означает, что среднее расстояние от узла до других узлов в сети больше. Меньшее значение означает, что среднее расстояние короче. Это может быть индикатор скорости, с которой информация течет по сети.

Центральность по эксцентричности: эта мера подразумевает расстояние от узла до самого дальнего от него узла. Высокое значение означает, что это расстояние большое, тогда как низкий эксцентриситет означает, что расстояние небольшое.

Чтобы увидеть, какие узлы являются центральными в этой сети, мы можем ранжировать узлы по размеру. Перейдите во вкладку «Appearance» → «Nodes» → «Size» → «Ranking» → выберите «Betweenness Centrality» и примените изменения. Центральные узлы в нашей сети теперь больше. Вместо того, чтобы определять размеры узлов в соответствии с их степенью, они теперь имеют размер в соответствии с центральностью по посредничеству (см. Рисунок 94).

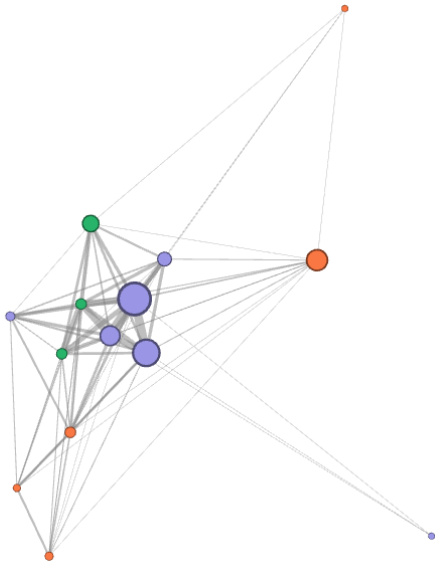

Рисунок 94 - Применение центральности по посредничеству

Диаметр сети измеряет длину самого длинного пути между двумя узлами в сети. В данном случае диаметр сети равен 2. Это означает, что самый длинный путь от одного узла к другому в нашей сети состоит из 2 шагов.

Предварительный просмотр

Во-первых, нам нужно подготовить нашу визуализацию в меню «Preview», прежде чем мы сможем экспортировать ее, чтобы показать миру.

Откройте вкладку предварительного просмотра, не касаясь каких-либо настроек, нажмите кнопку «Refresh». Отлично, теперь у нас есть первая версия для работы.

В меню «Preview Settings» вы можете настроить параметры, касающиеся узлов, ребер и названий. Также есть несколько предустановок, которые в настоящее время стоят по умолчанию. Попробуйте поиграться с настройками и нажмите кнопку «Refresh», чтобы увидеть, как это меняет график. Однако по умолчанию ваш график будет выглядеть примерно так (см. Рисунок 95).

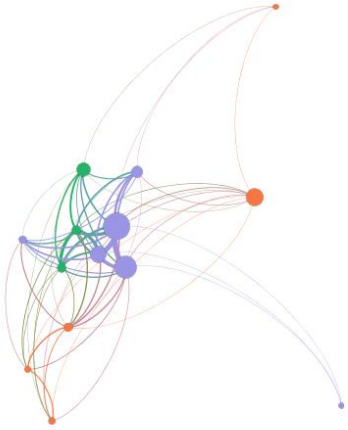

Рисунок 95 - визуализация в меню Preview

Теперь важно понимать, что вы должны настроить отображение графика так, чтобы ваша визуализация могла передать информацию, которую вы бы хотели подчеркнуть. Таким образом, вам нужно подумать о цели визуализации, и на чем вы хотели бы сделать акцент.

Сначала в настройках ярлыков узлов отметьте «Show Labels». Ярлыки не появятся, потому что мы их еще не определили, а наш столбец «Label» все еще пуст. Для этого зайдите в лабораторию данных. В меню ниже нажмите «Copy data to other column». Теперь мы скопировали данные в столбец «Label»! Вам только осталось обновить график и подобрать подходящий размер шрифта.

Экспорт

Пока закончим этим графиком, мы можем экспортировать его, щелкнув в нижней части экрана: SVG означает масштабируемая векторная графика. Я бы посоветовал вам сохранять файл во всех трех типах файлов. Сохранение его как PDF полностью сохранит все ребра и узлы (см. Рисунок 96). Более того, он позволяет увеличивать масштаб. Файл PNG плохого качества, поэтому его лучше использовать для краткого обзора.

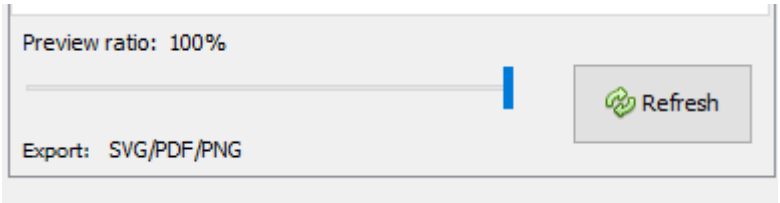

Рисунок 96 - Кнопка экспорта

Лучшее качество - PDF или SVG. Сохраните визуализацию под понятным названием.

Примеры проектов

1. Russian Drama Corpus(dracor.org/rus)
2. Six degrees of Francis Bacon
3. Star Wars Social networks
4. Social Networks in movies
5. Early African American Films

Источники:

1. How to get started with Social Network Analysis, Mitchell Telatnik [Электронный ресурс]. Режим доступа:https://towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374
2. Demystifying Networks, Scott Weingart [Электронный ресурс]. Режим доступа: http://www.scottbot.net/HIAL/index.html@p=6279.html
3. Introduction to Social Networks methods, Robert A. Hanneman and Mark Riddle [Электронный ресурс]. Режим доступа: http://www.faculty.ucr.edu/~hanneman/nettext/
4. Network Analysis – Digital Humanities 101
5. Gephi - The Open Graph Viz Platform[Официальное руководство]

Тест 4

1. Что может представлять узел сети? (несколько ответов)

A. Книгу

B. Персонажа пьесы

C. Отправителя письма

2. Какая статистика рассчитывает степень узла?

A. Average Degree

B. Network Diameter

C. Weighted Degree

3. Для чего используется подсчет модулярности?

A. Поиск сообществ

B. Определение длины пути между узлами

C. Подсчет количества узлов

Глава 5. Tableau

Tableau — инструмент для анализа и визуализации данных, позволяющий работать с несколькими источниками данных одновременно. Простота в использовании и красивые визуализации сделали его одним из популярных инструментов аналитиков и дата-журналистов наравне с Power BI.

Tableau предлагает несколько версий на выбор:

1. Tableau Public, с которой мы будем работать
2. Tableau Online, появившаяся не так давно
3. Tableau Desktop

Последние две имеют бесплатный период, так что вы можете проверить, нужен ли вам весь функционал Tableau для решения ваших задач. Важные отличия Tableau Public от остальных это то, что все ваши работы хранятся в открытом доступе на платформе, где каждый может смотреть ваши визуализации, а вы можете смотреть визуализации остальных. Также есть ограничения по доступу к данным с использованием отдельных API. Установить Tableau Public можно по ссылке(https://public.tableau.com/en-us/s/)

Импорт данных в Tableau

Теперь мы импортируем наш очищенный набор данных про памятники, с которым мы работали во время главы по работе с OpenRefine. Запустив Tableau, перед вами появится следующий экран (см. Рисунок 97).

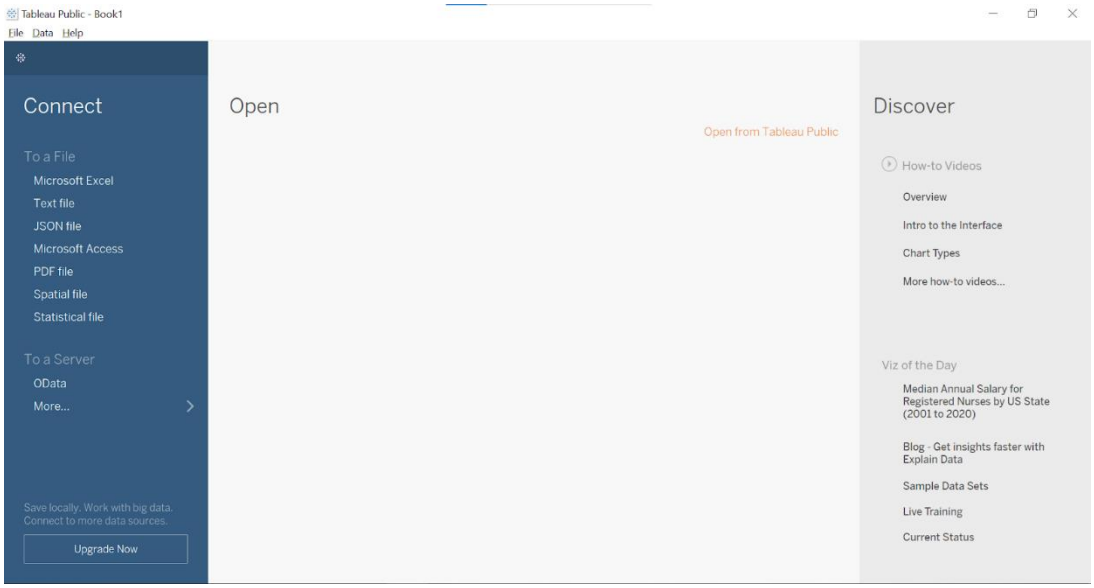

Рисунок 97 - Пример интерфейса

Слева вы увидите форматы файлов, с которыми вы можете работать, а справа полезные ссылки на обучающие материалы и «Визуализацию дня», которая демонстрирует возможности Tableau лучше всего.

Для того, чтобы открыть файл, выберите в меню «Connect» пункт «Microsoft Excel». Также можно импортировать другие типы файлов, такие как CSV или даже PDF, но пока мы будем использовать книгу Excel. После того, как вы откроете файл, вы увидите экран «Источник данных».

Tableau позволяет работать сразу с несколькими листами Excel. Слева вы можете выбрать листы, которые хотите использовать. Здесь также отображаются электронные таблицы в вашей книге Excel. Сейчас в нашем файле только один лист, поэтому он переносится в рабочую область автоматически.

Примечание: Если вы хотите добавить ещё файлы, то в меню «Connections» нажмите на «Add» и выберите формат файла, который хотите добавить. Новые листы появятся в меню «Sheets». Перетащив листья в верхний правый раздел, где написано «Drag tables here to relate them», вы сможете работать с несколькими источниками одновременно.

Экран будет выглядеть примерно так, как показано на скриншоте ниже. В правом нижнем углу отображаются данные только что размещенного листа (см. Рисунок 98).

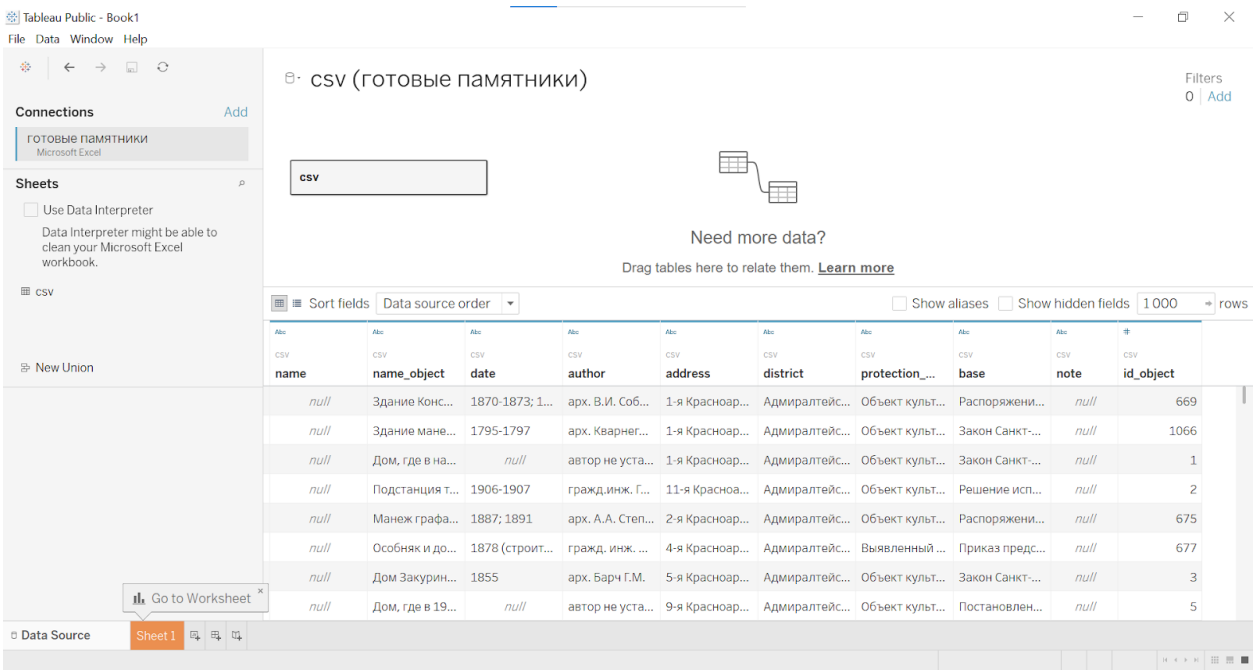

Рисунок 98 - Пространство по работе с данными

Вы можете видеть различные переменные в правом нижнем углу, где отображается предварительный просмотр данных, Tableau автоматически распознает формат этих переменных на основе форматирования ячеек Excel. Хэштег «#» зеленого цвета обозначает числовое значение, а синий цвет «abc» обозначает строку (которая представляет собой простой текст). Кроме того, вы увидите «null», что означает, что ячейка пуста и, следовательно, не содержит

данных. Типы переменных можно изменять, если это необходимо, то, кликнув на тип переменной, можно выбрать подходящий вариант, как это показано ниже (см. Рисунок 99). К счастью, в нашем случае Tableau все понял правильно, и ничего не нужно исправлять.

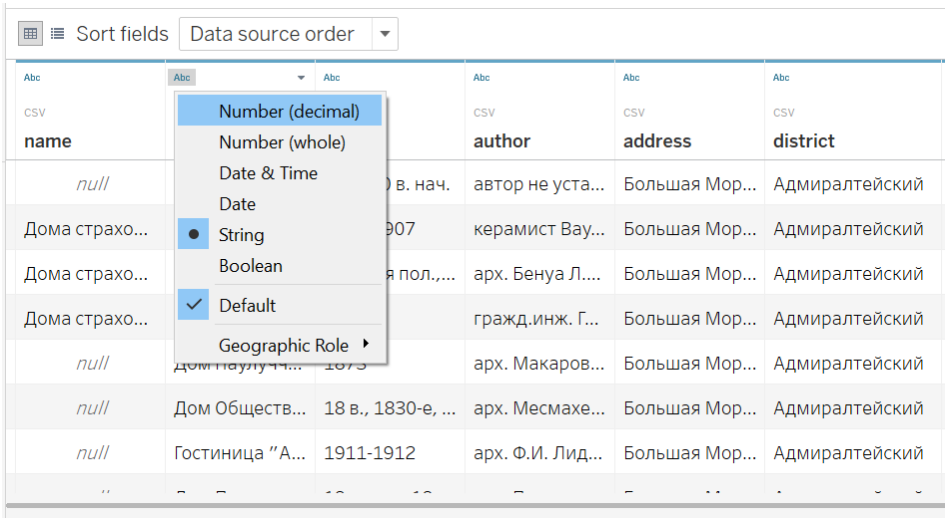

Рисунок 99 - Изменение типа переменной

Теперь ваши данные готовы, и мы можем приступать к визуализации. Для этого вам понадобится меню, находящееся внизу экрана. Вы увидите этот экран, выбрав вкладку «Sheet 1» (см. Рисунок 100).

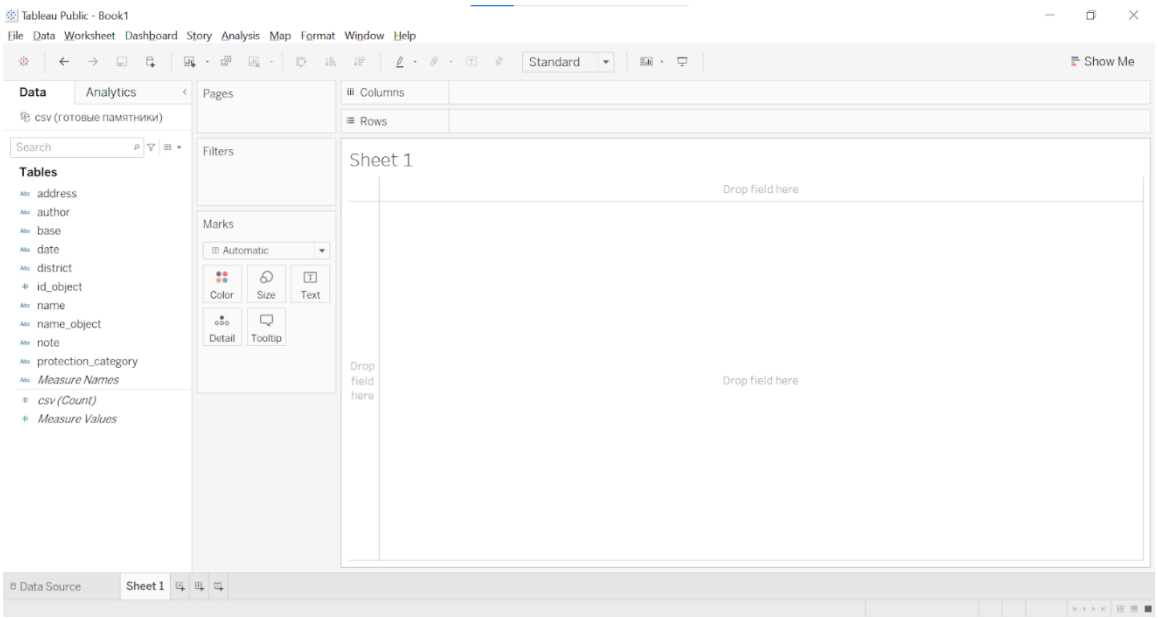

Рисунок 100 - Рабочая область

Это ваше основное рабочее пространство. Здесь ваши данные могут быть проанализированы и визуализированы. В левом меню вы можете увидеть пространственные измерения (Dimensions) и числовые (Measures). Важно понимать, в чем разница между ними.

Пространственные измерения (отображаются с синими значками) отвечают на вопрос "Что?" → Это в основном текст (строки), единицы времени или географические данные. Грубо говоря, пространство и время.

Числовые измерения (отображаются с зелеными значками) отвечают на вопрос «Сколько?» → Это числовые значения, с его помощью мы можем проводить вычисления.

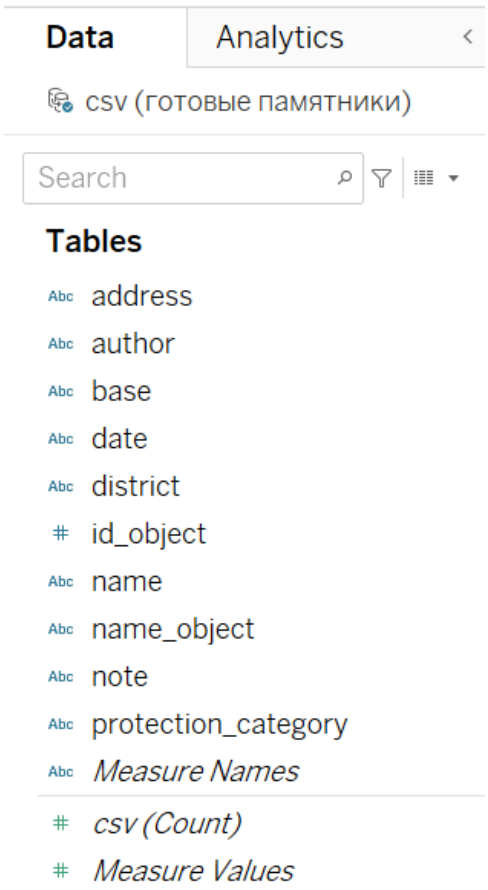

Рисунок 101 - Типы переменных

Примечание: На этом экране также можно изменить тип данных. Вам не обязательно обращаться во вкладку «Data Source». Щелкните значок типа данных рядом с именем переменной и при необходимости выберите другой.

Создание таблицы

Справа вы найдете меню под названием «Show Me». Здесь отображаются различные типы визуализации. Tableau также укажет, что вам нужно для этой конкретной визуализации (см. Рисунок 102).

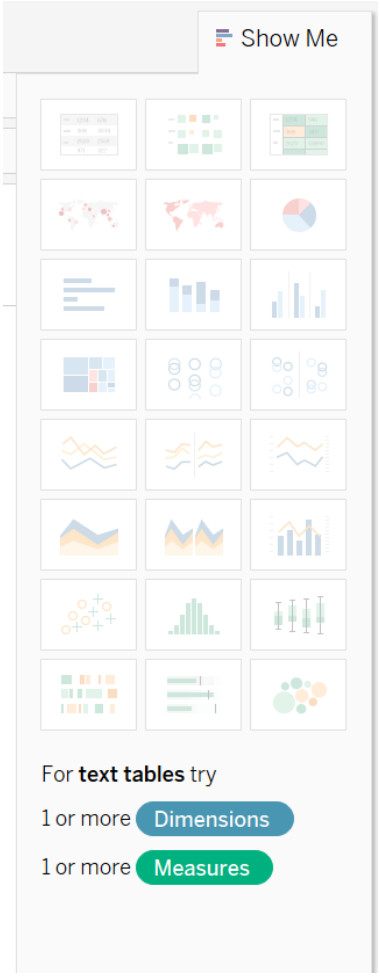

Рисунок 102 - Меню с выбором визуализаций

Как видите, для таблицы нам понадобится 1 одна переменная типа «Dimensions» и 1 типа «Measures». Для ее создания нам необходимо перетащить переменные из левого столбца в строки «Rows» и «Columns» в верхней части экрана (см. Рисунок 103).

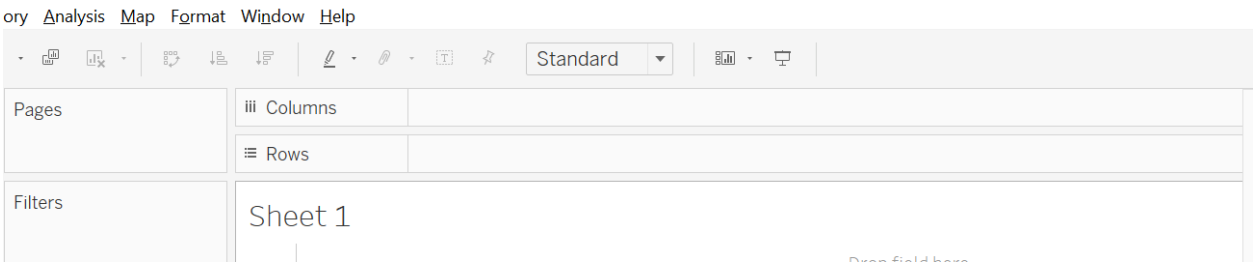

Рисунок 103 - Строки и столбцы

Если вы все сделали правильно, то вы должны увидеть график, расположенный ниже (см. Рисунок 104). Для превращения графика в таблицу мы можем пойти двумя путями.

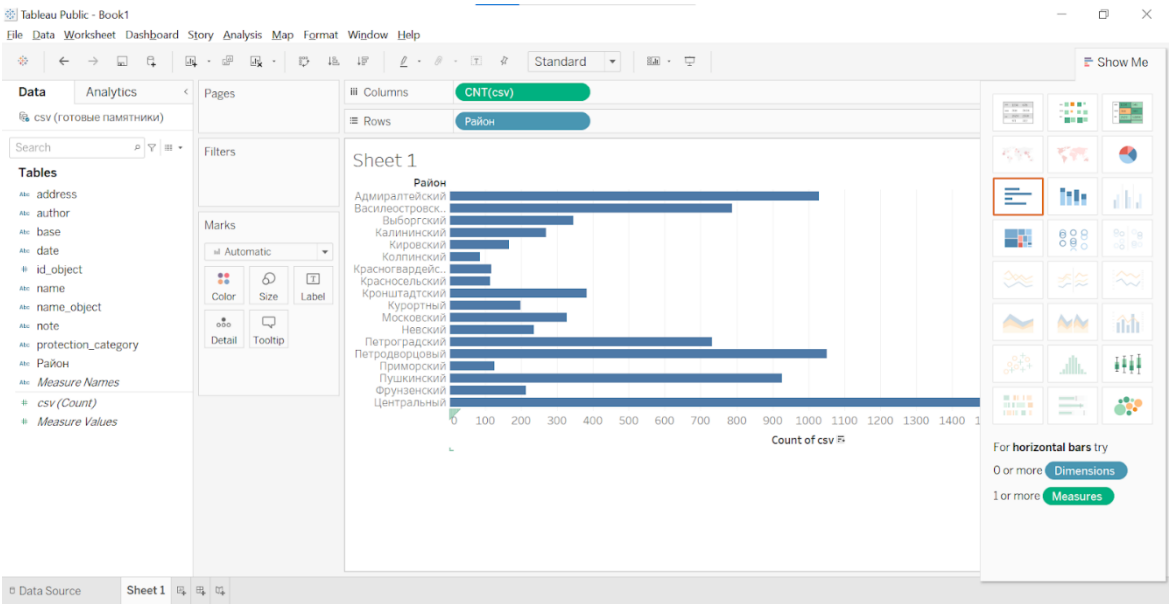

Рисунок 104 - Пример графика

Первый вариант — мы можем просто выбрать ее в меню «Show me», второй — перенести CNT (csv) в блок «Marks» и выбрать отображение текстом (см. 105). Для того, чтобы выбрать способ отображения, нажмите на четыре точки около переменной и выберите «Text»

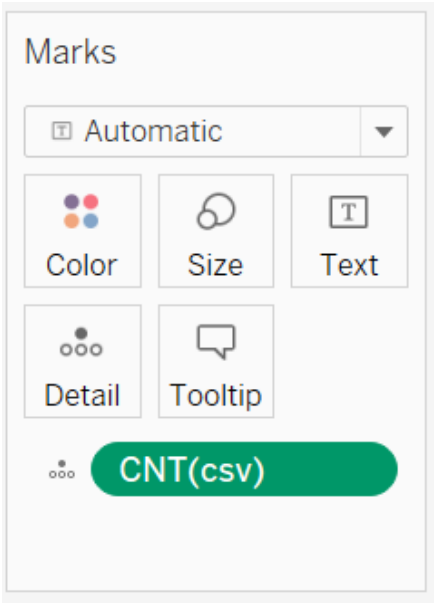

Рисунок 105 - Marks

В финале у нас получится таблица с количеством единиц культурного наследия по районам. Осталось только переименовать таблицу, достаточно два раза кликнуть на название «Sheet 1», и, возможно, раскрасить таблицу, обвести границы и т.д., это можно сделать, выбрав пустое пространство в таблице → Правая кнопка мыши → Format… Вы также можете заметить, что я переименовал переменную «district» в «Район», это можно сделать в меню с переменными слева (см. Рисунок 106).

Sheet 1

Район	
Адмиралтейский	1 030
Василеостровск..	787
Выборгский	346
Калининский	269
Кировский	166
Колпинский	85
Красногвардейс..	117
Красносельский	114
Кронштадтский	382
Курортный	198
Московский	326
Невский	236
Петроградский	732
Петродворцовый	1 052
Приморский	126
Пушкинский	926
Фрунзенский	214
Центральный	1 875

Рисунок 106 - Готовая таблица

Создание гистограммы

В Tableau можно редактировать несколько графиков одновременно, для создания новой рабочей области нажмите первую иконку после «Sheet 1» (см. Рисунок 107).

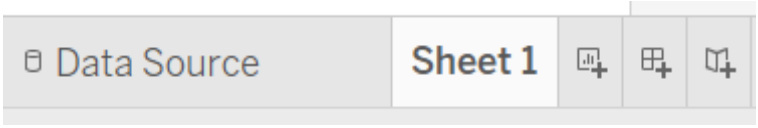

Рисунок 107 - Нижнее меню

Мы уже видели, как создаются графики, но давайте попробуем взять другую переменную. Посмотрим, кто из архитекторов и скульпторов создал больше всего объектов. Для этого нам понадобятся переменные «author» и «csv (Count)». Если вы перетащите автора в графу «Rows», то у вас появится оповещение, что значений слишком много, сейчас мы попробуем с этим разобраться, а пока нажимаем «Add all members». Мы увидим бесконечный список имен, который не даст нам никакой информации, и количество, созданных объектов. Чтобы отсортировать их по убыванию, нам нужно нажать кнопку «Sort author descending by Count of csv» (см. Рисунок 108).

Рисунок 108 - Сортировка по убыванию

Что-то стало проясняться, во-первых, мы видим, что у многих объектов не установлен автор (мы заполняли пустые значения в туториале по OpenRefine), эту переменную можно исключить из графика, выбрав переменную на графике, а потом нажать «Exclude» (см. Рисунок 109).

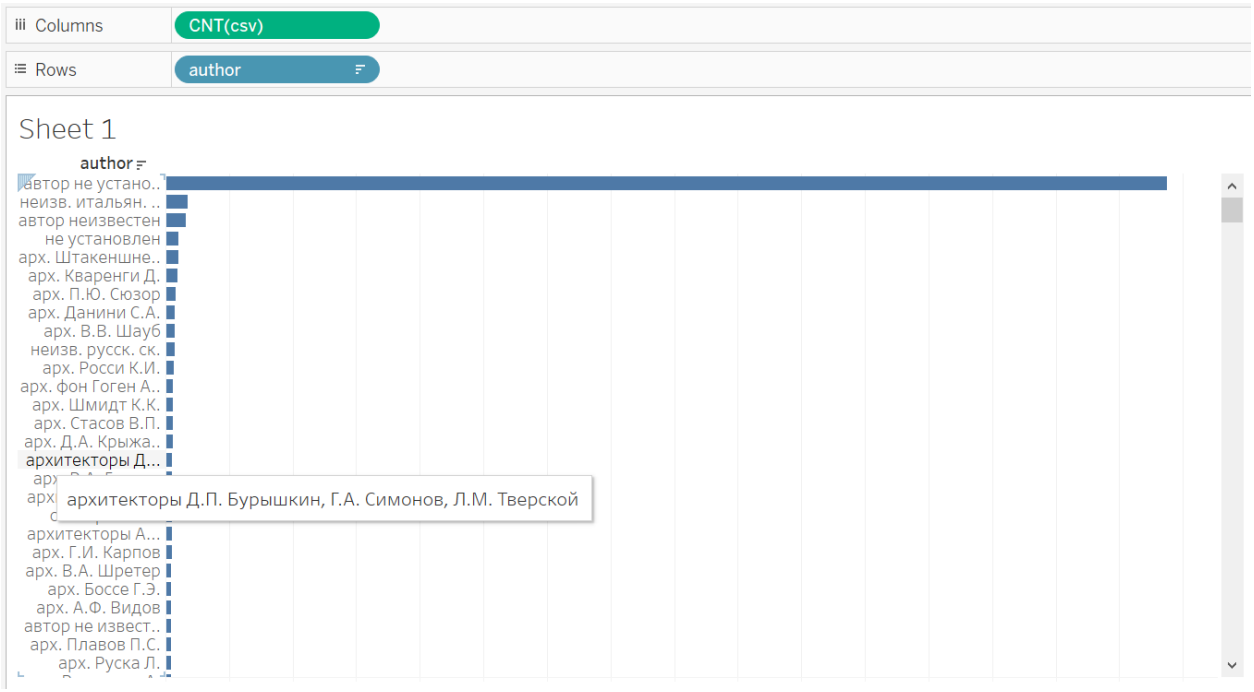

Рисунок 109 - Пример графика

Во-вторых, мы обнаружили, что исследователи использовали разные форматы записи для неизвестных объектов: «не установлен», «автор не установлен», «автор неизвестен» и т.п.. Давайте разберемся с количеством отображаемой информации, используя блок «Filters». Перенесите «CNT (csv)» в «Filters» и выберите диапазон от 20. Таким образом, должно получится следующее (см. Рисунок 111).

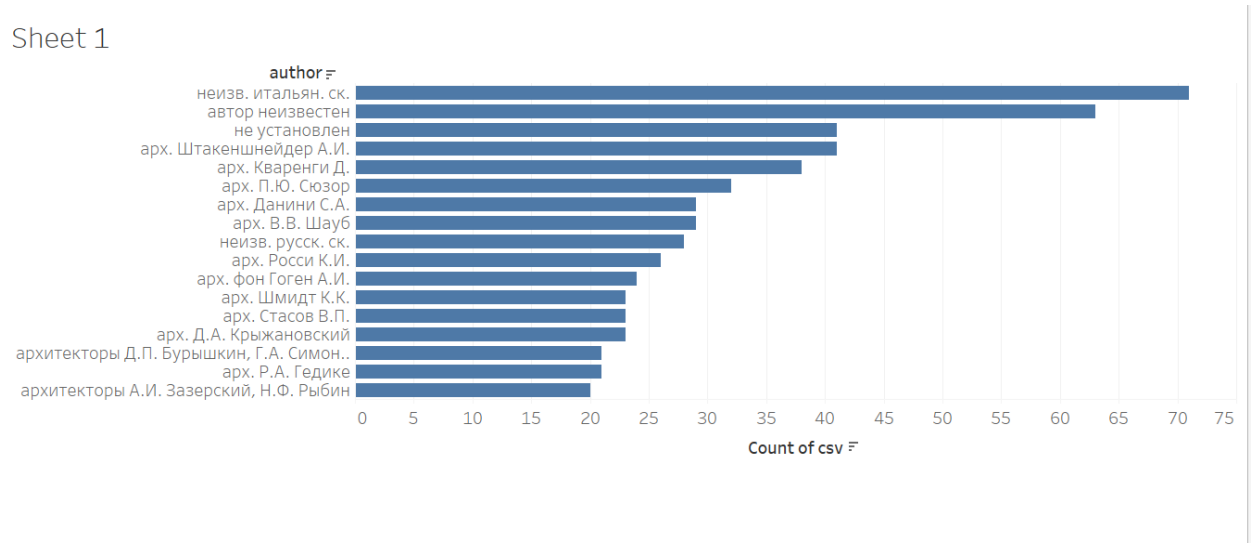

Рисунок 110 - Применение фильтров

Также, как и в прошлом примере, можно изменить название, цвета, изменить название осей.

Создание карты

Для построения карты нам необходимо знать координаты объектов, в этом разделе мы будем использовать датасет, связанный с музеями Санкт-

Петербурга (https://classif.gov.spb.ru/irsi/7808025993-musems_showrooms/structure_version/343/). При добавлении этих данных не забудьте убрать датасет, связанный с культурным наследием. Если посмотреть на переменную «coord», то мы увидим, что широта и долгота находятся в одном столбце, нам нужно разделить их.

	Sheet1 www	Sheet1 email	Sheet1 description	Sheet1 description...	Sheet1 work_time	Sheet1 for_disabled	Sheet1 coord	Sheet1 ogrn	Sheet1 inn	author
)-...	anichkov.ru	anichkovmuz...	Старейшее з...	The oldest bui...	пн-сб 10:00-1...	0	59.932723,30...	21474836471...	7840325991	null
:-...	вноводевичи...	priest-denis...	Действующи...	Active Ortodo...	09:00–18:00	0	59.89853,30....	10378580026...	7810387493	null
	нет данных	нет данных	В здании раб...	null	Круглосуточно	0	59.935402,30...	null	null	null
)-...	hermitagemu...	форма на сай...	Зимний двор...	null	вторник, чет...	1	59.939741,30...	10378430318...	7830002416	null
;-...	visit-petersbu...	null	Здание Смол...	Smolny Instit...	11:00–18:00, ...	0	59.946163,30...	10478440269...	7825453621	null
;-...	cathedral.ru	office@cathe...	Музей-памят...	St. Isaac's Cat...	пн, вт, чт-вс 1...	0	59.934039,30...	10278102848...	7812025107	null
I-...	kazansky-spb....	sobor.go@ma...	Казанский Ка...	Kazan Cathed...	пн-сб 08:30-2...	0	59.934387,30...	10278102848...	7812025107	null

Рисунок 111 - Переменные в датасете

Для этого выберите переменную «coord» → нажмите «Split» → измените тип переменных в «Geographic Role» на «Latitude» и «Longitude» (см. Рисунок 112). К концу манипуляций у вас должны появиться следующие переменные (см. Рисунок 113).

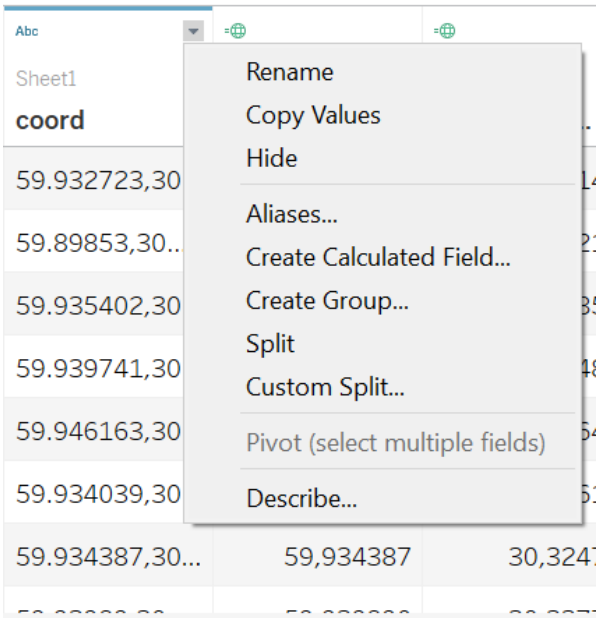

Рисунок 112 - Разделение переменной

Calculation coord - Spli...	Calculation coord - Spli...
59,932723	30,34146
59,898530	30,32210
59,935402	30,33857

Рисунок 113 - Финальный результат

Дальше останется только перенести все переменные в правильные поля. Предположим, что нам интересно посмотреть на плотность музеев в Центральном районе Санкт-Петербурга и их доступность для инвалидов. В поле «Marks» необходимо внести переменные «for disabled» и «name» для отображения информации при наведении на точку. В итоге должна получится следующая карта (см. Рисунок 114).

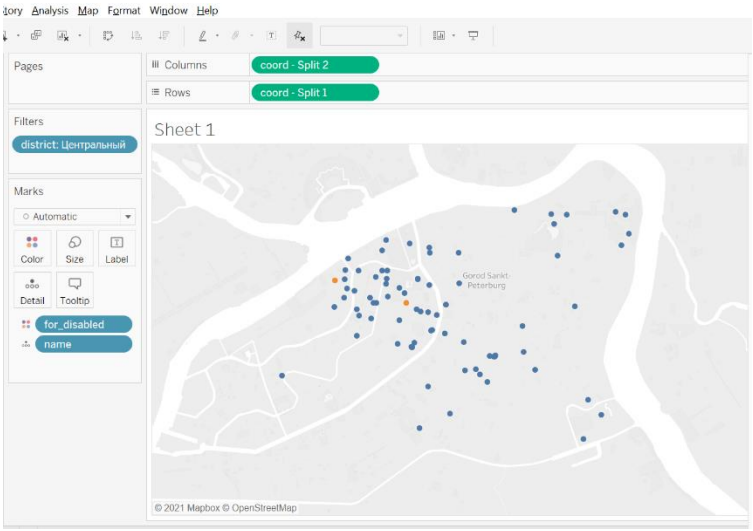

Рисунок 114 - Пример карты

Завершение и экспорт вашей визуализации

В Tableau есть несколько вариантов сохранения результатов:

1. Отдельный график
2. Дашборд

Если с сохранением графика все понятно: его можно сохранить, находясь в рабочей области нужного вам графика, то с дашбордом немного сложнее. Создать дашборд можно, используя нижнее меню, от вас требуется нажать иконку, следующую за созданием новой рабочей области (см. Рисунок 115).

Рисунок 115 - Создание дашборда

Преимущества дашборда перед графиками в том, что он позволяет объединить в одном пространстве несколько графиков, добавить иллюстраций, пояснительный текст, ссылки на полезные материалы. Ниже можно увидеть примитивный дашборд, составленный из графиков, которые были созданы на основе данных по музеям Санкт-Петербурга (см. Рисунок 116).

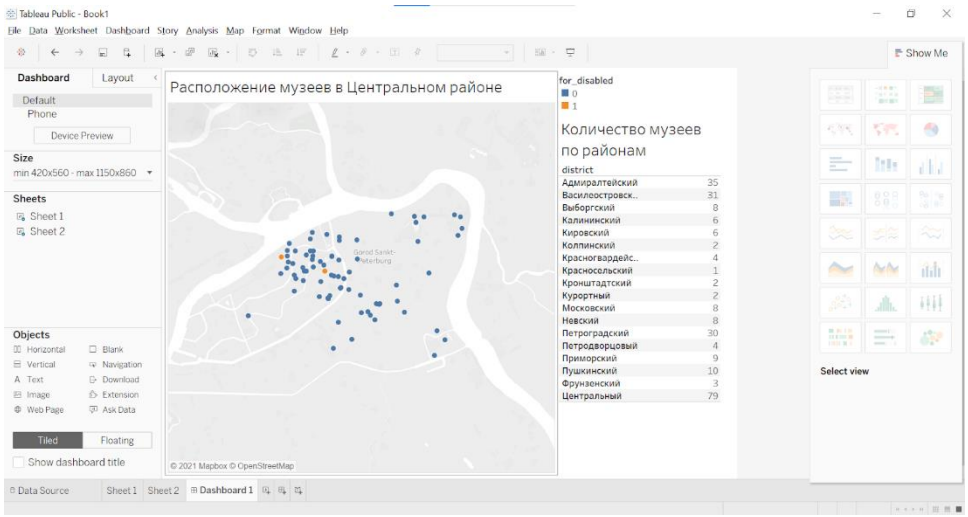

Рисунок 116 - Пример дашборда

Чтобы сохранить ваши графики, нужно перейдите в «File» → «Save to Tableau Public» (см. Рисунок 117). После этого ваш график или дашборд загрузится на сервер и будет доступен в вашем личном кабинете.

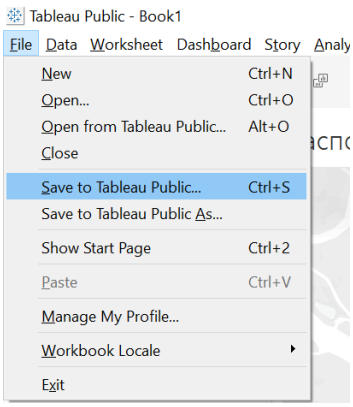

Рисунок 117 - Сохранение работы

Это конец главы, но могу порекомендовать вам портал Kaggle в качестве источника данных и будущей практики ваших навыков.

Источники:

1.Официальные руководства по Tableu [Электронный ресурс],

Режим доступа: https://onlinehelp.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-home.html

3. Официальные обучающие видео по Tableau:
   Режим доступа: https://www.tableau.com/learn/training

Тест 5

1. Какие виды измерений есть в Tableau? (несколько вариантов)
A. Dimensions
B. Measures
C. Digits

2. Какой раздел отвечает за подсказки при создании визуализаций?
A. Show me
B. Marks
C. Sheet

3. Как можно представлять информацию в Tableau? (несколько вариантов)
A. Отдельный график
B. Дашборд
C. Снимок экрана

Ответы к тестам

Тест 1

1. C

2. B

3. A

Тест 2

1. C

2. A

3. B

Тест 3

1. B

2. B

3. C

Тест 4

1. A, B, C

2. A

3. A

Тест 5

1. A, B

2. A

3. A, B, C

Оглавление

Пучковская Антонина Алексеевна
Зимина Лада Владимировна
Волков Дмитрий Алексеевич

Digital Humanities: инструментарий начинающего исследователя

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
Университет ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49, литер