Секция «Историческая информатика»

Мир глазами образованного человека г. Минусинска конца 19 - начала 20 веков: распределение частотности географических названий в книгах Минусинской общественной библиотеки.

Меховский Вадим Александрович
Студент (бакалавр)
Сибирский федеральный университет, Гуманитарный институт, Красноярск, Россия
E-mail: mehovsky.zenit-champion@yandex.ru

В этом исследовании был проведен географический анализ цифрового литературного корпуса. В центре внимания находились книги из детской коллекции Минусинской общественной библиотеки конца 19-ого - начала 20-ого веков. В исследование было включено 121 произведение, написанное между 1719 и 1905 годами. Цель исследования - получить географическое распределение частоты упоминаний локаций в цифровых копиях книг раздела «Детская литература» сибирской общественной библиотеки конца 19 века. Чтобы получить географическое распределение, были проделаны следующие этапы: приведение текста к машиночитаемому виду, изменение дореформенной орфографии на современную, выявление географических именованных сущностей и создание карты.

Чтобы создать машиночитаемую копию произведений, была использована программа ABBYY FineReader 15[1]. Стоит отметить, что не все программы имеют возможность распознавать дореформенную орфографию, однако, у выбранного нами программного обеспечения, такой функционал присутствует. После проведения процедуры распознавания текста книг, были получены 119 текстовых файлов, 32 копии на английском и 87 на русском языке с дореформенной орфографией. К сожалению, по непонятным причинам, ABBYY FineReader 15 не сумел распознать 2 книги. Цифровые копии на иностранном языке переводиться не будут, будет изменен лишь язык распознавания на этапе определения именованных сущностей. Остальные книги готовы к изменению орфографии.

Чтобы реализовать перевод текста книг в машиночитаемый формат, была написана программа на языке Python[2], с помощью библиотеки prereform2modern[3]. Программа создает новый текстовый файл, в который записывает результат перевода текста исходного файла. Процесс перевода занимал непродолжительное время, в среднем 30 секунд на одну книгу. Таким образом мы получили цифровые копии книг, пригодные для последующего распознавания именованных сущностей (NER).

На этапе выявления именованных сущностей мы столкнулись с большим разнообразием библиотек Python, способных это делать. Была выбрана библиотека Spacy[4], так как большинство функций не требуют дополнительной настройки, и есть отличная документация. Еще одним критерием выбора Spacy стало то, что нам была необходима библиотека, поддерживающая не только русский язык, но и английский. Затем был написан код для нахождения именованных сущностей[5]. Для удобства работы, был настроен экспорт в виде Excel таблицы, со следующими столбцами: именованная сущность; частота ее употребления в тексте. В процессе работы было выявлено следующее ограничение, Spacy не поддерживает текст с общей длинной более 1 миллиона символов. Исходя из этого, некоторые книги пришлось делить на 2-3 разных файла. После того, как все книги были обработаны, Excel файлы были проверены вручную. Были удалены все именованные сущности, не связанные с географическими местами или не имеющие значения в рамках данного исследования, например, стороны света, наименования рек, гор и т.д. После этого все именованные сущности были объединены в общий Excel файл.

Перед построением частотной карты мира были построены круговые диаграммы распределения для первых десяти локаций по количеству употреблений стран и городов (Рис. 1,2).

Как мы можем видеть на рисунке 1, первые десять стран по употреблению в текстах занимают 61% от общего употребления. Не удивительно, что на первом месте оказалась Россия, однако, на втором месте оказалась Польша. Если вспомнить взаимоотношения между Россией и Польшей на протяжении 19-ого века, то такая частота употребления Польского государства обоснована. «Польский вопрос, наверное, был самым острым вопросом внутренней политики России на протяжении большей части XIX века»[6]. Данные отображенные на Рисунке 2 свидетельствуют о том, что употребление городов разнообразнее нежели стран, очевидно из-за большего количества городов. Первая десятка занимает 38% от общего употребления, а на первом месте внезапно оказался Киев, вместо предполагаемых Москвы или Санкт-Петербурга. Также весьма популярным стал индийский город Бомбей (4 место). Более подробный анализ будет представлен на конференции.

Для картирования полученных результатов была построена карта в Adobe Illustrator. Была найдена карта мира, в которой каждая страна отображена на отдельном слое, чтобы была возможность выбрать цвет для каждой области. Итоговая визуализация географического распределения будет представлена на конференции.

Стоит отметить, что полученные результаты не могут в полной мере дать ответы на все вопросы. Прежде всего, не стоит исключать определенные ошибки при переводе текста в машиночитаемый вид. Издания старые, и некоторые географические наименования могли распознаться неправильно из-за изношенности книги. Следующим немаловажным обстоятельством является проблема распознавания именованных сущностей, ведь большинство библиотек используют следующий принцип: сравнивают исходный текст с уже готовым списком географических мест, и при совпадении создается новый список с совпавшими наименованиями. При этом обеспечить полноту этого «эталонного» списка не всегда представляется возможным. Однако, не стоит недооценивать географический анализ текста, технологии развиваются, и в скором времени мы сможем с большей точностью сконструировать представления образованного жителя Центральной Сибири девятнадцатого века.

[1] https://pdf.abbyy.com/
[2] https://github.com/mehovsky-zenit-champion/children-s-books.git
[3] https://pypi.org/project/prereform2modern/
[4] https://spacy.io/
[5] https://github.com/mehovsky-zenit-champion/children-s-books.git
[6] Ширинянц А.А. и А.В. Мырикова «Внутренняя» русофобия и «польский вопрос» в России XIX в., 2015. С.19.

Иллюстрации

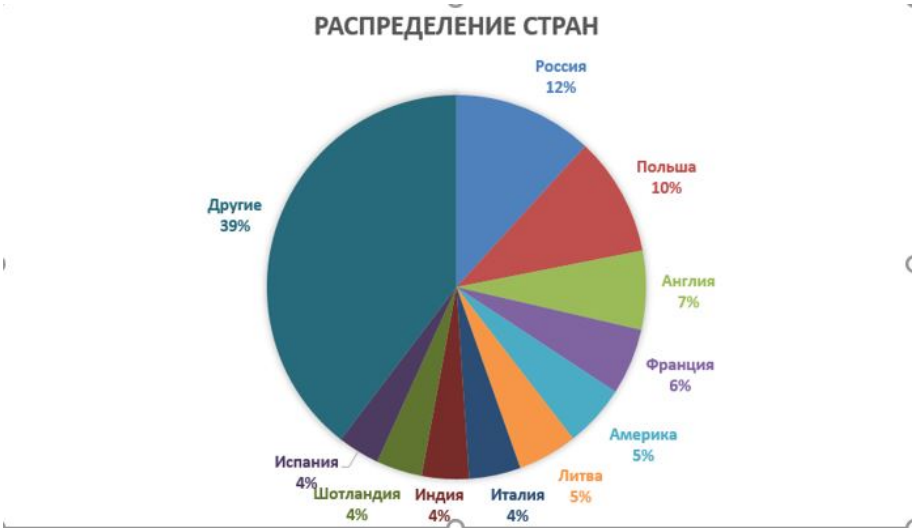

Рис. : Рисунок 1 – Диаграмма распределения стран

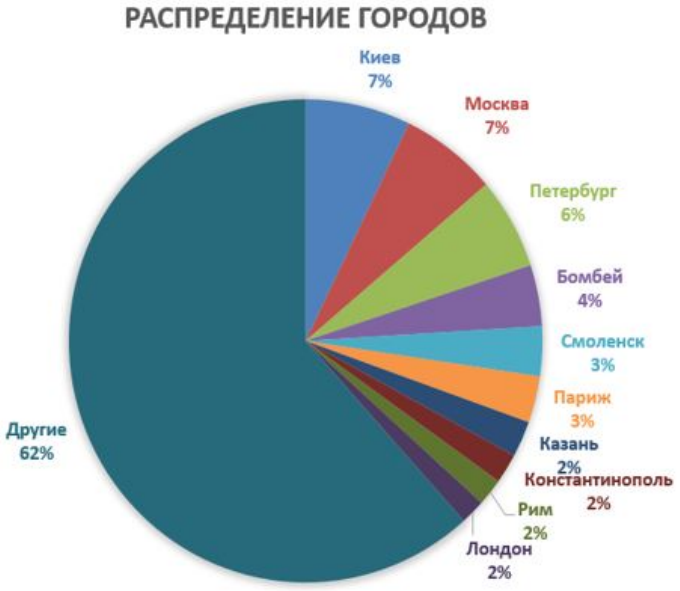

Рис. : Рисунок 2 – Диаграмма распределения городов