Доклад

Content analysis of 150 years of British periodicals

2,3) Отправной точкой для нашего исследования было сравнить результаты для нашего корпуса с тем для Google книгах корпус (1), показывающие сходства и различия между использованием корпуса книг и в одной из газет, и выделив, что мы можем найти те же тенденции в нашем корпусе, но также, что анализ газеты могут быть более чувствительны к определенным культурным сдвигам, в частности из-за их тесной связи с текущими событиями, чем книги.

4) Коллекция газет Британской библиотеки является одной из лучших в мире и содержит большинство тиражей газет, издаваемых в Соединенном Королевстве с 1800 года. Масштабы газетной издательской индустрии начиная с начала XIX века были огромны: многие города и поселки издавали одновременно несколько газет, а другие газеты, имевшие целью обеспечить более широкий уездный тираж, представляли собой непревзойденную картину провинциальной жизни, охватывающую весь XIX и половину XX века.

4)В мае 2010 года FindMyPast начал сотрудничество с Британской библиотекой, чтобы оцифровать миллионы страниц этих исторических газет и сделать их доступными для общественности для поиска в интернете по адресу www.britishnewspaperarchive.co.uk.

4)Новые страницы все время сканируются в рамках 10-летнего проекта, который после его завершения будет содержать более 40 миллионов газетных страниц из газетной коллекции Британской

библиотеки. На сегодняшний день FindMyPast выпустил более 12 миллионов страниц из 535 различных газетных изданий, опубликованных между 1710 и 1959 годами, добавляя более 8000 новых страниц каждый день.

4) Затем изображения были переданы через процесс OCR для идентификации текста, используемого в каждом разделе страницы, в то время как соответствующие метаданные для каждой проблемы были переданы через проверку качества для исправления любых ошибок на этапе структурного извлечения.

4) Большая часть корпуса (78%) была вручную сегментирована на различные области по содержанию страницы, структурной информации или информации о заголовке. Ручная обработка потребовала бы слишком много времени и денег, поэтому большую часть корпуса анализировали при помощи специального программного обеспечения.

4) OCR был выполнен на цифровых изображениях в программном обеспечении CCS docWorks командой FindMyPast. Этот процесс выводит распознанный текст на изображении вместе со связанной информацией (например, расположение текста на странице), а также ряд других технических параметров.

Метаданные по каждой торговой точке вводились вручную во время оцифровки на основе каталога газет Британской библиотеки. Местоположение, присвоенное каждому изданию, было определено на основе местоположения исходной публикации. Сегментация страницы заголовка для материала, обрабатываемого в начале проект, вручную исправлялась операторами; позже эти шаги выполнялись уже без человеческого вмешательства. Редактор-человек использовался для

проверки качества структурных данных, извлекаемых программным обеспечением, а программное обеспечение выявляло систематические проблемы, которые затем исправлялись операторами. Этот процесс включал в себя проверку правильности названия, даты возникновения проблемы, а также сегментирование страниц на определенные типы.

4) После того, как процесс оцифровки был завершен, команда FindMyPast предоставила команде Bristol коллекцию документов, содержащих текстовое содержание газетных статей, а также связанные с ними метаданные, касающиеся названия статьи, даты публикации, названия, опубликовавшего статью, местоположения издателя и так далее. Документы были преобразованы из форматов METS, MODS и ALTO в документы объектной нотации JavaScript (31) и сохранены с соответствующими метаданными в коллекции MongoDB NoSQL (31). https://www.mongodb.com/).

4) Затем каждый документ в базе данных подвергался процедуре извлечения информации (описанной ниже), которая была направлена на то, чтобы позволить нам генерировать временные ряды любого n-грамма, извлекать ссылки на объекты в тексте и разрешать объекты и связывать объекты с внешними базами данных, где это возможно, для обогащения информации, содержащейся в каждом документе.

5) Основной целью данного исследования было показать подход к пониманию преемственности и изменений в истории, основанный на дистанционном чтении обширных новостных корпусов, что дополняет традиционное внимательное чтение историками. Мы показали, что изменения и континуиты, обнаруженные в содержании газет, могут

отражать свойства культуры, предвзятость в представлении или реальные события.

5, 6, 7) Простой контент-анализ этого корпуса позволил нам с высокой точностью выявлять конкретные события, такие как войны, эпидемии, коронации или конклавы, в то время как использование более совершенных методов искусственного интеллекта позволило нам выйти за рамки подсчета слов путем обнаружения ссылок на названные сущности. Эти методы позволили нам наблюдать как систематическую недопредставленность, так и неуклонный рост числа женщин в новостях в течение 20-го века и изменение географической направленности различных концепций. Мы также оцениваем даты, когда электричество обогнало пар, а поезда обогнали лошадей как средство передвижения, как примерно в 1900 году, так и наблюдая другие культурные переходы. Мы считаем, что эти подходы, основанные на данных, могут дополнить традиционный метод близкого чтения в обнаружении тенденций непрерывности и изменений в исторических корпуса

8,9) Этот результат подтверждает-с более высокой степенью сложности-результаты, полученные с использованием трендов n-грамм (Рис. 4 e и F), что свидетельствует о том, что в течение всего рассматриваемого периода женщины неизменно представлены в меньшей степени, чем мужчины, и позволяет нам изучить нюансы и характер различных предположений, сделанных в отношении Пола. Этот более совершенный подход также свидетельствует о медленном, но неуклонном увеличении числа женщин после 1900 года. Эти результаты можно прочесть в сочетании с аналогичными результатами для современных новостей (17), показывающими, что гендерная предвзятость в средствах массовой информации, по-видимому, не сильно изменилась, и в современных газетах примерно в три раза больше мужчин, чем женщин.

8)Кроме того, возвращаясь к концепциям, которые мы исследовали с помощью трендов N-грамм, мы составили географические карты Соединенного Королевства для каждого термина, отражающие постепенное увеличение или снижение (а не всплеск активности) (рис. 6). Мы извлекли все местоположения, найденные в статьях, в которых упоминается одно из понятий, снова сняли с них неоднозначность, используя DBpedia (19), и восстановили их географические координаты.Мы отмечаем, что термины "британский" и "английский" были достаточно широко распространены на большей части территории Великобритании в 1854 году. К 1940 году использование английского языка сократилось, а британский стал национальным идентификатором по умолчанию (рис. 6A).

11 (10)) В ценностях и убеждениях мы проверяем гипотезу, выдвинутую Гиббсом и Коэном (3) о снижении так называемых “викторианских ценностей” в течение исследуемого периода. Мы видим, что упоминания о некоторых ключевых викторианских ценностей (3) в целом снижаться, хотя такие термины, как “долг”, “мужество” и “выносливость” новый импульс во время войны, в то время как другие ключевые условия, в частности, “бережливость” и “терпение” не проявляют тенденцию к снижению, отборочный простой счетах предполагаемую кончину викторианских ценностей (рис. 2A и B).

11 (10)) В целом, консервативные и либеральные партии получили примерно одинаковый уровень охвата в течение 19-го века, хотя они оба затмеваются с 1920-х годов и далее Лейбористской партией (рис. 2D). Это изменение, конечно, не может считаться отражением уровня политической

поддержки, но оно предполагает, что появление и рост Лейбористской партии определяли повестку дня региональной и местной прессы с 1920 по 1950 годы (особенно после первого правительства Лейбористской партии в 1924 году).

11 (10))Наши результаты также указывают на очень четкую временную линию появления "британскости"как популярной идеи, с термином "британский", обгоняющим термин "английский" в конце 19-го века (рис. 2 E и F). После этого мы видим значительный рост использования термина "британец" в первой половине 20-го века, с резким увеличением во время обеих мировых войн. Термин английский сократился в тот же период (и действительно, страдает небольшими провалами во время Первой Мировой Войны и Второй мировой войны)— до такой степени, что термин “шотландский” обгоняет его в конце 1940-х годов, предполагая, что британцы заменили английский язык в качестве национального идентификатора по умолчанию. Хотя ученые предполагают, что развитие британскости предшествует этому росту (16), эти данные свидетельствуют о том, что доминирование британскости в народном воображении является феноменом 20-го века.

12 (11)) В области технологий мы отслеживаем распространение инноваций в энергетике, транспорте и связи. На первом участке мы наблюдаем устойчивый спад пара и постоянный рост электричества, с точкой перехода в 1898 году (Рис. 3A). В области транспорта мы наблюдаем, как поезда обогнали лошадей по популярности в 1902 году, после рассвета железнодорожной эпохи, которая началась в 1840-х годах, показывая культурное значение лошадиных сил на протяжении всего 19-го века (рис. 3B).

12 (11)) В области коммуникаций, мы исследуем скорость распространения телеграфа, телефона, радио и телевидения, подтверждающие предыдущие выводы (1), что наблюдается возрастающими темпами освоения новых технологий, что привело к быстрому росту телевидения (рис. 3C).

13 (12)) В социальных изменениях мы наблюдаем резкие временные границы в таких явлениях, как суфражистское движение и период анархической активности; мы наблюдаем пики волнений, которые соответствуют известным периодам забастовочных действий в 1912 и 1919 годах, тогда как выражение восстание соответствует напряженности в британских колониях, в частности, низкому канадскому восстанию 1837-1838 годов и "Индийскому мятежу" 1857 года (рис. 4A).Частота суфражистки имеет четко разграниченный временной интервал (1906-1918) (рис. 4B), что соответствует периоду от популяризации термина в ответ на срыв публичных собраний до достижения избирательного права для многих, хотя и не всех, взрослых женщин в 1918 году. Несмотря на предшествовавшую ей многолетнюю политическую кампанию, мы видим резкий рост освещения суфражисток (и суфражистов) после драматической смерти Эмили Уайлдинг Дэвидсон, которую затоптала до смерти королевская лошадь в Аскоте. Этот резкий рост освещения, возможно, является примером начала 20-го века важности "медиа-события" для политической кампании и его способности захватить журналистское воображение.

13 (12)) В популярной культуре ученые-медийщики зафиксировали рост интереса к новостям (и пропорциональное снижение в общественных делах), при этом эти данные указывают на четкий график возрастающей важности массовой культуры в освещении новостей. Например, мы видим, что ссылки на "актеров", “певцов” и “танцоров” начинают увеличиваться в

1890-х годах, значительно увеличиваясь после этого, тогда как ссылки на "политиков", напротив, постепенно снижаются с начала 20-го века (рис. 4F). Мы видим ту же картину в увеличении охвата n-граммами "футбола" и "крикета", причем футбол стал более заметным, чем крикет, начиная с 1909 года (рис. 4G).

14) Этот шаг приближает нас к уровню понятий и семантики, а также позволяет обойти многие риски, связанные с подбором ключевых слов (материалов и методов). Кроме того, можно автоматически связывать именованные сущности с существующими базами данных недавно ставших доступными сущностей, которые предлагают авторитетный список людей, местоположений и организаций. Эти списки с открытым исходным кодом включают Yago (18) и DBpedia (19), и они позволяют автоматизировать включение внешней информации о различных сущностях, которых нет в самом корпусе, таких как пол и род занятий человека или координаты местоположения. Парсинг текста, таким образом, в результате извлечения 263,813,326 упоминает в 1,009,848 разных сущностей в корпус.

14) Обнаружение каждый раз, когда человек, упомянутый в корпусе, также присутствует в DBpedia (19) или другой базе знаний, часто позволяет нам сопоставить их с типом занятий. Эта процедура позволяет автоматизировать исследования (1) славы для людей различных профессий в течение жизни (рис. 5A).время.Среди прочего, мы подтверждаем их вывод о том, что политики и писатели, скорее всего, добьются известности в течение своей жизни, в то время как ученые и математики с меньшей вероятностью добьются известности; однако мы также наблюдаем снижение для политиков и писателей в новостях, которые не наблюдались в книгах, в то время как время, кажется, добрее к ученым и математикам.

14) Мы также извлекаем каждое упоминание о человеке в корпусе (независимо от того, присутствуют ли они во внешних ресурсах) и выводим пол с помощью плагина ANNIE GATE, стандартного инструмента для NLP (20). Этот процесс дал нам более 35 миллионов ссылок на людей с разрешенным полом, что позволяет рассчитать общую вероятность того, что человек, упомянутый в новостях, является мужчиной (или женщиной), и, наконец, изучить, как эта вероятность изменяется с течением времени (рис. 5B).

Наблюдались также географические фокусы технологических достижений с течением времени, которые мы показываем для перехода от пара к электричеству (рис. 6C) и от лошадей до поездов (рис. 6D). Для пара, мы можем видеть, что упоминания во время его самого высокого использования года в 1854 широко распространены, с концентрацией сосредоточены вокруг крупных портов. Однако, принятие электричества заменяет пар к 1947, с электричеством будучи упомянутым особенно в ссылке на Лондон, Лидс, и зоны Зюйдвеста (смоквы. 6C). Во время самого раннего пика внимания к лошади в 1823 году мы видим, что упоминания в основном рассеяны по всей стране без отличительного рисунка, что указывает на их использование в сельских общинах, и есть только нечетное упоминание о поезде, которое при более близком чтении было выявлено, как правило, в другом контексте (речь идет о дрессировке животных или шествиях). К 1948 году снижение лошади явно вступило в силу, все, кроме исчезновения с этой карты, в то время как поезд сильно упоминается, особенно вокруг крупных городов, показывая аналогичную картину электричества.