Papers Mapping Mutable Genres in Structurally Complex Volumes. (Картирование нестабильных жанров в текстовых объемах сложной структуры)

Данная работа посвящена обработке текстовых массивов с помощью самообучающихся алгоритмов. Авторы фокусируются на задаче классификации текстов, входящих в объем электронных библиотек, по жанру. Если текстовый объем библиотеки очень велик (а обычно это так), эта задача становится слишком трудоёмкой для ручной классификации. В то же время, для осмысленной работы с этими массивами, исследователи нуждаются в выделении из всего объема текстов меньших текстовых наборов определенных значимых категорий, например, различных жанров. Исследование текстов разных жанров — распространенная задача для филологов. Однако, при решении такой задачи возникают некоторые типичные проблемы. Одна из таких проблем — проблема выделения жанра в разные временные периоды.

Текстовые объемы электронных библиотек обычно охватывают несколько веков. На протяжении таких периодов времени многие жанры претерпевают достаточно сильные изменения, что значительно затрудняет электронное распознавание жанра. Задача по определению жанра произведения вообще часто вызывает трудности и безотносительно временных особенностей. Проблема жанровой классификации занимает исследователей в области гуманитарных наук еще со времен Аристотеля, и была проделана серьезная работа для ее решения. Но во многих случаях до сих пор отсутствует консенсус между исследователями по этому вопросу. В некоторых случаях, жанр можно выделить формально и однозначно — например, определить, что текст является сонетом. Но у таких жанров, как сатира или научная фантастика отсутствует фиксированная текстовая структура. Авторы утверждают, что в таких случаях электронная классификация подобных жанров позволяет решить эту задачу даже более эффективно, за счет более гибкого подхода к самому процессу категоризации.

При обработке текста электронными алгоритмами с целью выделения категорий, алгоритмы анализируют тексты по определенным параметрам, которые зависят от выбранной модели, представляющей ту или иную категорию. Так, к примеру, авторы используют так называемую модель «сумка слов», которая выделяет категорию текста с помощью подсчета частотности ключевых слов, отражающих данный жанр. Важно, что результат обработки количественный — алгоритм определяет вероятность, с которой можно отнести текст к определённому жанру. Это дает гибкий подход к определению жанра в сложных случаях. Биографии 18 века, в которых авторы, часто вставляли в текст вымышленные диалоги, часто напоминают собой повесть. Подобный текст в ходе электронной обработки получит некоторое положительное значение вероятности того, что жанр текста - художественная литература. Такое вероятностное определение жанра дает гибкий подход к различным пограничным случаям.

Авторы статьи приводят график, показывающий на выборке из 1000 случайных текстов 462200-текстового массива распределение вероятности того, что текст является художественным, полученное с помощью двух различных классификаторов, один из которых был натренирован на выборке текстов 19-го века, а другой — 18-го. Из графика видно, что в данном случае есть очень высокая корреляция между результатами, полученными разными классификаторами. Это значит, что особенности жанра «художественная литература» за этот период не слишком менялись со временем. Но не во всех случаях подобная проверка обнаруживает сходство результатов. Для жанра «готическая новелла» все попытки авторов

натренировать классификатор, который сумел бы распознать и готику эпохи романтизма (такую как «Дракула») и готику позднего 19 века как готику, потерпели поражение. Для таких случаев авторы предлагают тренировать несколько классификаторов для разных периодов времени с наложением (пересечением) выбранных временных интервалов. Такой подход позволяет эффективно преодолеть сложности, связанные с изменением характерных особенностей жанров с течением времени.

Далее, авторы демонстрируют, как можно использовать созданные ими алгоритмы не только для определения жанра, но и для изучения более узких особенностей этих жанров с учетом временной специфики. Так, например, они натренировали классификатор, который позволяет определить, с какой точки зрения ведется повествование — от первого или от третьего лица. Затем они выбрали с помощью своего алгоритма массив художественных текстов 18-19 веков и применили к ним вышеупомянутый классификатор.

Авторы приводят результаты такой обработки - среднюю вероятность того, что художественный текст написан от первого лица для каждого пятилетнего периода. Линия тренда на графике показывает, что эта вероятность постепенно снижалась в течении 18 века, оставаясь приблизительно на уровне 50 процентов и резко пошла вниз на рубеже веков, стабилизировавшись примерно на 20-ти процентах. Также авторы приводят ключевые слова, которые помогали алгоритмам определять тексты от первого и от третьего лица. Выяснилось, что в этом периоде в художественных текстах от третьего лица часто встречаются термины, имеющие отношение к семье (такие слова как дочь, муж, брак), частям тела (глаза, лицо, губы, голос), чувствам и эмоциям (дрожащий, улыбка, бледный, любимый). Для текстов от первого лица оказались характерны числительные и слова, относящиеся к измерению (количество, тонна, часть). Также в текстах от первого лица распространены слова имеющие отношение к мореплаванию, такие как корабль, берег, юго-запад, остров, путешествие, провизия.

Авторы статьи приводят свои гипотезы о том, как можно интерпретировать полученные результаты. Возможно, что повествование от третьего лица в этом периоде наиболее характерно для определенного жанра (например, для куртуазного романа, возможно, что в третьем лице преимущественно писали авторы-женщины. Авторы намерены проверить обе гипотезы в ближайшее время. При интерпретации лексических особенностей текстов от первого лица авторы проявляют бОльшую уверенность. Даже беглое знакомство с заголовками позволяет отнести наиболее яркие тексты-представителей категории «от первого лица» к жанру Робинзонады. Авторы отмечают, что, конечно, популярность *Робинзона Крузо* (1719) и произведений по его мотивам общеизвестна. Но только получив количественные данные, авторы поняли, насколько устойчивой и продолжительной была эта популярность. Судя по всему, произведения Робинзонады регулярно и широко перепечатывались разными издательствами вплоть до середины 19 века. Авторы полагают, что эти лексические особенности — сочетание повествования от первого лица, колониального сеттинга и акцента на точных количественных описаниях и экономике — распространяются далеко за пределы Робинзонады и включают в себя самые разные приключенческие жанры, от рассказов типа *Копей царя Соломона* (1885) до вестернов и описаний этнографических путешествий.

Вторая проблема, возникающая при распознании жанров алгоритмами — проблема гетерогенности изучаемых текстовых объемов. Так, при обработке типичной новеллы 19 века, мы столкнемся с тем, что повествование предваряется реальным, не выдуманным описанием жизни автора и заканчивается 20 страницами рекламных объявлений. Для отделения такого паратекста от основного повествования можно использовать специальные электронные

алгоритмы. Те же алгоритмы позвляют сегментировать текст, присвоив разным его частям различные классификационные категории. Если в тексте романа встречаются стихотворения или письма, классификатор сможет отделить их и добавит эти отрывки к соответствующим массивам. Для такой обработки авторы статьи тренируют классификаторы на основе модели Маркова. Авторы показывают, что для задачи по определению жанра многослойная обработка текстов с использованием таких алгоритмов позволяет сгладить шумы и улучшить качество итоговой обработки. При сравнении с другими алгоритмами их подход демонстрирует высокое качество классификации. Впрочем, они отмечают, что итоговые значения вероятности в 0.8 — 0.9 даже с использованием такого сглаживания все-таки недостаточно хороши, поскольку в большинстве исследований необходима большая точность. Это ставит вопрос о поиске новых стратегий обработки, которые позволят усовершенствовать существующие алгоритмы. Авторы указывают на некоторые возможные ходы — например, взаимообучение разных алгоритмов и тренировка классификаторов на бОльших объемах текстов с заранее вручную определенным жанром. Также авторы подчеркивают, что для последовательного развития их подхода следует относиться к алгоритмам не как к «черным ящикам», результат работы которых - некое окончательное и однозначное заключение, а, скорее, как к гибкому способу электронной обработки больших массивов данных, включающему в себя неопределенность, как неотъемлемую часть своей внутренней логики. В идеале, они хотели бы, чтобы исследователи в области гуманитарных наук относились к алгоритмам машинного обучения не как к набору инструментов, а как к статистическому по своей природе языку, который позволяет формулировать и решать исследовательские задачи в вероятностной логике.