УДК 004.912

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ ОБРАБОТКИ И АНАЛИЗА ТЕКСТОВ НА ПРИМЕРЕ ИССЛЕДОВАНИЯ ДИСКУРСА РУССКОЙ ИНТЕЛЛИГЕНЦИИ XIX ВЕКА

Рудалева Е. А. (Национальный Исследовательский Университет «Высшая Школа Экономики», Санкт-Петербург)

Доклад посвящен обзору ряда методов обработки и анализа текстов: тематического моделирования, сентимент-анализа, анализа векторов слов и кластеризации, — на примере рассмотрения дискурса русской интеллигенции XIX века на выборке из Национального Корпуса Русского Языка. На основе данных методов было проведено исследование ключевых для исследуемого дискурса слов.

Феномен русской интеллигенции широко обсуждается в сфере гуманитарных наук ввиду его противоречивости и размытости границ его определения. Применение методов компьютерной лингвистики для осмысления интеллигентского дискурса является актуальным направлением, поскольку позволяет получить представление об исследуемом объекте независимо от авторского отношения, а также переработать большой массив текстов с вычленением необходимой информации.

В рамках данного исследования на основе теоретической работы М. Ю. Лотмана «Интеллигенция и свобода» были вычленены важные для репрезентации интеллигенции ключевые слова: *совесть, самопожертвование, социальность, публичность* и сама *интеллигенция*. На их основе была сформирована выборка по вхождениям из Национального Корпуса Русского Языка (НКРЯ). К предобработанным данным (лемматизация через морфологический анализатор *MyStem*, удаление стоп-слов, токенизация) были последовательно применены методы анализа и обработки текстов.

1. Анализ векторов слов

Поиск векторов слов направлен на выявление сходства между словами в заданной выборке через создание вероятностных моделей, прогнозирующих совместную встречаемость слов. Реализация данного метода на исследуемой выборке проходила на основе алгебраического подхода расчету векторов слов через разложение матрицы номализованных вероятностей скип-грамм на сингулярные числа при помощи алгоритма SVD для дальнейшего поиска синонимов. Применение данного анализа позволило выявить сильную покровительственную связь интеллигенции с народом (ассоциированные леммы: *народ, любить, преданность*) и одновременно ориентацию на мыслительные практики приближения к нему (ассоциированные леммы: *думать, мысль, тяготение*). Более того, была выявлена традиция ассоциация ключевого слова *совесть* с леммами *самопожертвование* и *интеллигенция* выводит к идее об интеллигенте как носителе совести и мотивах жертвы как геройского подвига.

2. Сентимент-анализ

Сентимент-анализ производился при помощи словаря тональности от русскоязычного тезауруса *Карта Слов*, размеченный по положительной, отрицательной и нейтральной тональности с параметром силы оценочного заряда. В рамках выборки текстов интеллигенции были выявлены негативные саморепрезентации данной социальной группы как «отщепенца», демонстрирующие отстранённость и чуждость интеллигента и его сосредоточенность на «настоящем» *праве* и *правде.*

3. Тематическое моделирование

Тематическое моделирование (ТМ) направлено на решение тем в коллекции документов интеллигентского дискурса. Основные методы ТМ: Латентно-Семантический Анализ (*LSA*) и Латентное Размещение Дирихле (*LDA*). В рамка данного исследования выбран второй метод, поскольку позволяет словам относиться к нескольким темам одновременно и появляться в документах в разных сочетаниях.

4. Иерархическая кластеризация

В отличие от ТМ задача иерархической кластеризации — классификация документов через их разбиение на классы (кластеры). Был выбран агломеративный алгоритм по методу полной связи, где расстояние между кластерами определяется через расстояние между отдаленными элементами в разных кластерах.

Результаты тематического моделирования по методу LDA и иерархической кластеризации позволили выделить важную черту интеллигенции— ее обращенность к литературе и литературным образам, где слово принимается не только как деятельность и инструмент воздействия (*тематическое моделирование*), но и как стремление самой интеллигенции перенести текстовую реальность и черты персонажей в жизнь (*иерархическая кластеризация*).

Комплексный подход на основе методов обработки и анализа текста к анализу выборки позволил выделить основные черты интеллигенции XIX века: несмотря на характерную интеллигенции бесклассовость, она обладает классовым сознанием, а интеллигентский язык не избавлен от наименований классов (*анализ векторов слов, анализ тональности*). Вместе с тем, интеллигенция стремиться понять народ и, более того, ассоциироваться с ним (*анализ векторов слов*) через труд и право от любви и обращенности к истинности—через те положительные качества, которые интеллигент приписывает себе (*анализ тональности*). Такие идеалы во многом почерпнуты интеллигентом из литературы, а сама «литературность» жизни или творение жизни из текста является важнейшим атрибутом интеллигенции (*тематическое моделирование, иерархическая кластеризация*). При этом, интеллигент, устремленный к свободе (*анализ векторов слов*), не обладает однонаправленностью в политических убеждениях: в интеллигентском дискурсе равно наличествует ориентация на новое и европейское, а сам интеллигент представляется чужаком и отрешенным от России (*анализ тональности*) и ориентация на русское пространство вместе с озабоченностью общественными проблемами внутри государства и глубоким чувством любви к родине (*тематическое моделирование, иерархическая кластеризация*). Таким образом, современные методы обработки и анализа текстов позволяют решать довольно сложные аналитические задачи.