УДК 81-139

АНАЛИЗ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ДЛЯ ИССЛЕДОВАНИЯ ХУДОЖЕСТВЕННОГО ТЕКСТА (НА ПРИМЕРЕ КНИГ ДЖ. К. РОУЛИНГ «ГАРРИ ПОТТЕР И …»)

Донина Ольга Валерьевна
Сигаева Татьяна Олеговна
Воронежский государственный университет
(г. Воронеж, Россия)

Данная работа посвящена активно развивающемуся в последние годы направлению тематического моделирования, суть которого заключается в создании семантических моделей корпуса текстов на основе разновидностей нечеткой кластеризации лексики. Проводится анализ эффективности работы инструмента тематического моделирования TopicModelingTool для индексации текстовых массивов, на основе книг Дж. К. Роулинг «Гарри Поттер и … ».

Ключевые слова: тематическое моделирование, веб-скрейпинг, анализ данных, классификация документов, извлечение информации, индексация текстов.

This article is devoted to the direction of thematic modeling, which has been actively developing in recent years, the essence of which is to create semantic models of the corpus of texts based on varieties of fuzzy clustering of vocabulary. The analysis of the effectiveness of the mathematical modeling tool Topic Modeling Tool for indexing text arrays, based on J. K. Rowling's books "Harry Potter and…", is carried out.

Keywords: topic modeling, web-scraping, data analysis, document classification, information extraction, text indexation.

Электронный архив библиотеки МГУ имени А.А. Кулешова

В рамках статьи будет рассмотрена возможность анализа художественного текста с использованием инструментария компьютерной лингвистики, а именно – метода тематического моделирования, который является способом построения модели корпуса текстов, отражающим переход от совокупности документов и совокупности слов в документах к набору тем, характеризующих содержание этих документов [1, С. 101–103]. Цель данного исследования – проанализировать эффективность работы приложения TopicModelingTool для анализа содержательного аспекта художественных текстов на основе книг Дж. К. Роулинг «Гарри Поттер и …» (всех семи частей). Для достижения поставленной цели нам был выработан следующий алгоритм действий:

• Первый этап – сбор массива текстовых данных из сети интернет [2] при помощи инструмента веб-скрейпинга OctoParse [4]. В результате работы программы было получено семь файлов в формате Excel, каждый из которых соответствует одной из частей саги. Всего созданный корпус насчитывает 1090739 словоупотреблений.

• Следующий этап – предобработка выгруженных данных. Это важный шаг, который помогает повысить качество данных. Мы воспользовались программой MyStem, разработанной И. Сегаловичем в компании «Яндекс», которая производит морфологический анализ текста на русском языке [3]. В результате работы программы все тексты были лемматизированы, то есть все слова в них приведены в начальную форму.

• На заключительном этапе мы приступили непосредственно к тематическому моделированию, используя инструмент TopicModelingTool [5]. В итоге проделанной работы автоматически были выделены списки слов для каждой из пяти тем (темы «0» - «4» в списке ниже):

0. глаз мистер палочка друг волшебный хвост профессор зал глядеть эльф мантия узнавать палец волшебник

1. профессор друг глаз палочка комната пойти рука волшебный школа голос скоро дверь нога открывать слово

2. палочка глаз смерть отвечать знать пожиратель волшебный понимать думать друг самый узнавать рука оставаться

3. профессор дядя камень словно мистер произносить глаз тетя письмо метла голос оказываться

4. свой знать рука спрашивать лицо дверь говорить голос голова видеть становиться отвечать думать увидеть ничто

Электронный архив библиотеки МГУ имени А.А. Кулешова

Далее мы визуализировали соотношение содержания текста каждой книги (на графике части книг обозначены значениями «1» - «7» по вертикали) и выявленных нами тем (рис. 1).

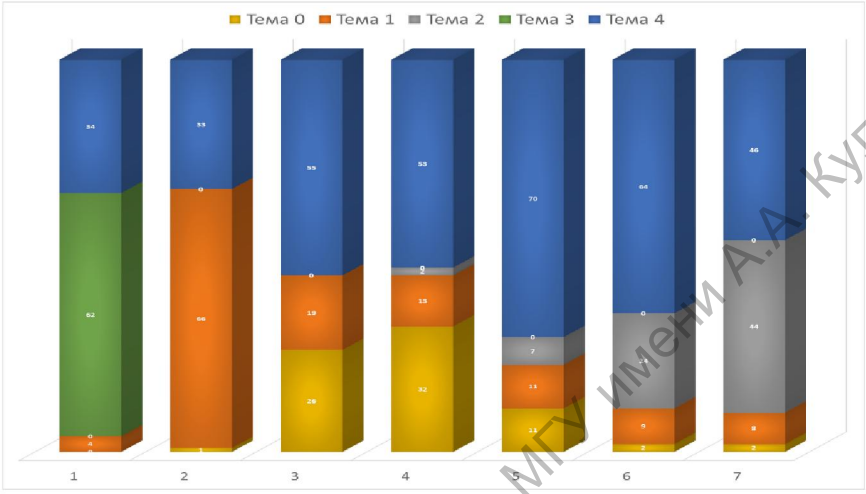

Рисунок 1. Соотношение содержания текста и тем

Как видно на графике, темы «1» и «3» отражают содержание конкретных книг, а темы «0», «2» и «4» являются общими для всех частей. Так, тема «3» встречается только в первой части «Гарри Поттер и Философский камень», что может быть обусловлено важностью лексемы *камень* для данной темы. Тема «1» в наибольшей степени содержится в части «Гарри Поттер и Тайная комната», что может быть связано с важностью для этой темы лексем *комната* и *дверь*.

Несмотря на то, что, по результатам нашего исследования, тематическое моделирование оказалось не очень эффективным для анализа содержательного аспекта художественного текста в связи с чрезмерным упрощением материала, оно может быть применено для выявления общих и уникальных тем в произведениях (например, при Distant Reading), а также выявляемые при тематическом моделировании ключевые слова и темы могут применяться при автоматическом рубрицировании и автоматическом поиске текстов.

Литература

1. Глушков, Н. А. Анализ методов тематического моделирования текстов на естественном языке / Н. А. Глушков // Молодой ученый. – 2018. – № 19 (205). – С. 101–103.

Электронный архив библиотеки МГУ имени А.А. Кулешова

2. Bib.bz [Электронный ресурс]. – Режим доступа: https://ru.bib.bz/author/4/. – Дата доступа: 09.12.2022.
3. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I. Segalovich // MLMTA-2003.
4. OctoParse [Электронный ресурс]. – Режим доступа: https://www.octoparse.com/. – Дата доступа: 09.12.2022.
5. TopicModelingTool [Электронный ресурс]. – Режим доступа: https://sourceforge.net/projects/topicmodeltool/. – Дата доступа: 09.12.2022.

Электронный архив библиотеки МГУ имени А.А. Кулешова