УДК 004.93'14

Сравнительный анализ алгоритмов кластеризации текстовых данных

А.С. Холодова

ФГБОУ ВО «Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского»

г. Саратов, Российская Федерация

Стремительная цифровизация общества сопровождается быстрым ростом объемов неструктурированной текстовой информации. Для эффективного управления знаниями и автоматизации аналитических процессов возникает необходимость в применении методов машинного обучения, в частности, алгоритмов кластеризации, позволяющих выявлять скрытые тематические группы в массивах документов [1].

Целью данного исследования является проведение сравнительного анализа эффективности различных алгоритмов кластеризации при обработке русскоязычных новостных текстов, а также оценка влияния методов снижения размерности на качество выделения тематических кластеров.

В качестве экспериментальной базы выступил датасет новостных статей, содержащий тексты по шести основным тематикам: «Россия», «Мир», «Экономика», «Наука и техника», «Культура» и «Спорт». Предварительная обработка включала токенизацию, фильтрацию стоп-слов, лемматизацию и удаление специфического «информационного шума». Для числового представления документов применялся метод TF-IDF, формирующий разреженные векторы и позволяющий усилить значимость редких, но информативных терминов [2].

Учитывая высокую размерность полученных признаков, для снижения вычислительной сложности использовался метод сингулярного разложения (TruncatedSVD), после чего применялся алгоритм UMAP для визуализации и финального сжатия пространства признаков. В рамках исследования были

реализованы четыре подхода: K-means, DBSCAN, агломеративная иерархическая кластеризация и спектральная кластеризация. Оценка качества разбиения проводилась на основе метрики чистоты кластера путем сопоставления автоматически выделенных групп с исходными тематическими метками [3].

Эксперименты показали, что тематика «Спорт» обладает высокой семантической однородностью и стабильно выделяется в изолированный кластер всеми рассмотренными методами. Однако при обработке политико-экономических новостей методы продемонстрировали различную устойчивость. Алгоритм K-means в сочетании с UMAP обеспечил наилучший баланс между компактностью кластеров и их интерпретируемостью. Агломеративная кластеризация с косинусной метрикой также показала высокую результативность. В то же время основанный на плотности метод DBSCAN и спектральный подход оказались менее эффективными для данного типа высокоразмерных разреженных данных: DBSCAN классифицировал значительную долю объектов как шум, а спектральный метод сформировал размытые границы между тематически близкими группами.

На основе полученных результатов можно сделать вывод, что для задач тематического моделирования новостных лент оптимальным является связка TF-IDF с последующим понижением размерности и применением алгоритма K-means или иерархической кластеризации.

Список использованной литературы

1. Савченко И.Ю. Обработка естественного языка для использования в машинном обучении / И.Ю. Савченко // Аллея науки. – 2018. – № 6(22). - С. 1000-1002.
2. Гонтарев А.А. Выявление значимости слов в текстах на основе алгоритма TF-IDF / А.А. Гонтарев, Ю.М. Вишняков // Прикладная математика XXI века: современные проблемы математики, информатики и моделирования. — Краснодар, 2019. — С. 193-197.

3. Хорошилов А.А. Установление сходства текстовых документов / А.А. Хорошилов, А.В. Кан, Е.А. Евдокимова, С.Г. Пицхелаури // Моделирование и анализ данных. – 2023. – Т. 13, № 4. – С. 45-58.