ISSN 2181-3187

ИСПОЛЬЗОВАНИЕ МАТРИЦ И ВЕКТОРОВ ДЛЯ АНАЛИЗА ТЕКСТА НА РУССКОМ ЯЗЫКЕ

М.Эрназарова преподаватель Джизакского филиала Национального университета Узбекистана
Ж.Райимбердийев студент Джизакского филиала Национального университета Узбекистана
Rayimberdiyevjaloliddin3@gmail.com

Аннотация: В данной работе рассматриваются способы применения математических понятий — матриц и векторов — для анализа текстов на русском языке. Особое внимание уделено методам векторного представления текста, таким как модель "мешка слов", TF-IDF и нейросетевые подходы (Word2Vec, FastText). Описаны особенности морфологической обработки русского языка, включая лемматизацию. Показано, как векторы и матрицы применяются в задачах классификации, тематического анализа и извлечения смысла из текстов. Работа демонстрирует, что математические методы играют ключевую роль в современной обработке естественного языка.

Ключевые слова: векторы, матрицы, анализ текста, русский язык, Word2Vec, TF-IDF, обработка естественного языка, классификация текстов

Современные технологии обработки текста всё чаще применяются в лингвистике, информационном поиске и системах искусственного интеллекта. Одним из ключевых подходов является использование математических понятий, таких как матрицы и векторы, для анализа текстов на русском языке.

Векторное представление текста — это способ перевода слов, предложений или целых документов в числовую форму. Одним из популярных методов является модель "мешка слов" (bag-of-words), где каждый текст преобразуется в вектор, элементы которого отражают количество вхождений каждого слова в словарь. Более продвинутым методом является TF-IDF (Term Frequency-Inverse

ISSN 2181-3187

Document Frequency), который учитывает не только частоту слов, но и их значимость в контексте всех документов.

Матрицы используются для представления больших объемов текстов. Например, можно построить **матрицу "документ-термин"**, где строки — это документы, а столбцы — термины (слова). Таким образом, каждый документ становится вектором в многомерном пространстве. Это позволяет применять методы линейной алгебры, такие как сингулярное разложение (SVD), для уменьшения размерности и выделения скрытых смыслов.

Для русского языка особую роль играет морфологическая и лексическая сложность. Перед тем как текст будет преобразован в вектор, часто используется лемматизация — процесс приведения слов к начальной форме. Это особенно важно для языков с богатым словоформообразованием, таких как русский.

Семантический анализ текста также опирается на векторы. Современные модели, такие как Word2Vec или FastText, позволяют каждому слову сопоставить вектор, обученный на больших корпусах текста. Эти векторы захватывают контекстное значение слова, что позволяет анализировать семантическое сходство и находить синонимы или тематические связи между словами и фразами.

В задачах машинного обучения и классификации текстов (например, определение темы, тональности или автора текста), векторные представления играют ключевую роль. Например, алгоритмы SVM, наивный байесовский классификатор или нейронные сети используют именно числовые признаки — векторы — для обучения и предсказания.

Таким образом, матрицы и векторы служат основой современного анализа текста. Их применение делает возможным глубокий, формализованный и автоматизированный подход к обработке естественного языка, включая русский. Это открывает широкие возможности в науке, бизнесе и технологиях.

Дополнительные подходы к векторизации текста включают использование трансформеров, таких как BERT и его русскоязычные аналоги (например,

ISSN 2181-3187

RuBERT). Эти модели обучаются на огромных корпусах и способны учитывать контекст слова в зависимости от его положения в предложении. Это особенно полезно для анализа сложных синтаксических конструкций и определения полисемии слов в русском языке.

Кроме того, анализ текста может быть интегрирован с визуализацией векторных представлений. С помощью методов снижения размерности, таких как t-SNE или UMAP, можно визуализировать семантические отношения между словами или документами на двумерной плоскости, что упрощает интерпретацию результатов и позволяет выявлять кластеры и тематические группы.

Использование векторов и матриц также находит применение в задаче суммаризации текста, где система автоматически извлекает ключевую информацию из больших текстов. Такие алгоритмы комбинируют статистические и нейросетевые методы, делая возможным генерацию краткого и информативного резюме на основе анализа содержимого.

Список литературы:

1. Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). — Pearson.
2. Мартынов, А. В. (2020). *Основы анализа текста на русском языке*. — Москва: Издательство МГУ.
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. — arXiv:1301.3781
4. Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. — Cambridge University Press.
5. Васильева, Е. А. (2019). *Машинное обучение и анализ текстов*. — СПб: Питер.