

ОТКРЫТАЯ СИСТЕМА АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТОВ

*канд. техн. наук, доц. О.В. Касилов, магистр В.В. Скороход,
Национальный технический университет "Харьковский
политехнический институт", г. Харьков*

Непрерывный рост объемов информации сделал крайне актуальными задачи поиска и обработки информации в огромных массивах данных. Обработка текстовых данных невозможна без использования лингвистических методов.

Целью работы является создание системы автоматизированного анализа текста, предоставляющей пользователю возможность гибкой настройки на решаемую задачу путем выбора оптимального набора инструментов и создания новых инструментов.

Концептуальная модель системы анализа текста содержит три уровня обработки текста: лингвистический, статистический, аналитический и включает в себя: подход к обработке данных, структуру системы, принципы развития и наполнения системы.

Качественные и количественные характеристики составляют основу для получения интегральных показателей. Лингвистические и статистические методы обработки необходимы для получения исходных данных для применения аналитических методов.

Структурные элементы системы (система базовой обработки, система анализа результатов, аналитическая система накопления поступающей информации, система хранения, пользовательская система) взаимодействуют между собой.

Основой для развития и наполнения системы является аналитическая система накопления поступающей информации, построенная в соответствии с моделью адаптивно-динамического преобразования информации. Информация, получаемая в результате базовой и аналитической обработки исходных данных, остается в системе, помещается в базу данных и может быть использована при обработке исходной информации. Это делает систему динамической и расширяет внутренние "знания" системы, которые также включают в себя устоявшиеся алгоритмы решения задач (сценарии).

Язык сценариев используется для создания собственных алгоритмов анализа. Использование "шаблонных" алгоритмов позволяет гибко настраивать систему обработки. Система реализована средствами Java Script с использованием набора инструментов NPM, технологии Angular и набора библиотек анализа текстов.