УДК 004.415.25

РАЗРАБОТКА ФОРМАТА ХРАНЕНИЯ И ВЗАИМОДЕЙСТВИЯ С СЕМАНТИЧЕСКОЙ ИНФОРМАЦИЕЙ

Есаков К.И. (Университет ИТМО)
Научный руководитель – старший преподаватель Цопа Е. А. (Университет ИТМО)

В работе автором предложена универсальная семантическая структура, выделяемая из документов различного формата и разработан ряд программных модулей для взаимодействия с ней. Работа подразумевает изучение и усовершенствование существующих решений и создание функционирующего программного кода.

Введение. Семантический анализ текста является одним из способов работы с информацией. Извлечение семантической информации – смыслового содержания логической единицы в конкретном тексте, позволяет работать как с глубинным, так и поверхностным смыслом текста, а автоматизация этого процесса посредством вычислительных машин позволяет обрабатывать большой объем информации за короткие сроки.
Для хранения и взаимодействия с визуально-значимой информацией необходимо предложить формат и разработать программный модуль для него.

Основная часть. Формат включает в себя работу с визуально-значимой информацией, в частности с текстовыми и табличными файлами. Для получения доступа к тексту и информации о форматировании и структуре документа был выбран Apache OpenOffice API. Из полученной информации выделяются значимые данные, такие как форматирование текста, на какие страницы, абзацы и списки он разбит, какова структура таблицы. Каждая такая единица информации - токен. Токены могут быть вложенными или образовывать группу. Порядок токенов также имеет значение.

Выводы. Проблема является актуальной с точки зрения науки, а ее решение будет иметь практическую пользу. Разрабатываемая библиотека будет применяться в рамках существующего проекта семантического анализа документов.

Есаков К.И. (автор) Подпись

Цопа Е. А. (научный руководитель) Подпись