УДК 004.624, 004.912

ИЗВЛЕЧЕНИЕ СТРУКТУРИРОВАННОГО ТЕКСТА ИЗ ДОКУМЕНТОВ В ФОРМАТЕ PDF

Ткешелашвили А. М., федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», Санкт-Петербург
Научный руководитель – старший преподаватель Клименков С. В., федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», Санкт-Петербург

В ходе работы был предложен подход к извлечению текста из PDF с восстановлением исходной структуры документа. Для определения элементов структуры используется нейронная сеть, анализирующая выбранные критерии. Предложен список критериев и разработано обучающее множество.

Одной из актуальных и приоритетных областей датамайнинга является обработка документов неструктурированных форматов и извлечение информации из них. В настоящее время данные хранятся в документах множества офисных форматов с различной структурой, в частности в распространенном формате PDF. Поскольку данные внутри документа в формате PDF хранятся блоками, порядок хранения которых не всегда совпадает с порядком следования их в документе, извлечение данных и восстановление логической структуры документа данного формата представляют наибольшую сложность.

Целью работы является разработка алгоритма извлечения структуры PDF документа и преобразование его во внутренний формат системы семантического анализа текстов, с сохранением семантически значимой информации при помощи обучения нейронной сети.

Извлечение структуры текста происходит путем конвертации документа из формата PDF во внутренний формат системы семантического анализа, уточняя и сохраняя семантически значимую информацию и связи в документе, в первую очередь высокоуровневые сущности, такие как строки и абзацы, а также заголовки и относящуюся к ним информацию и др. Для автоматизации данного процесса была выбрана технология нейронных сетей.

В ходе работы был произведен анализ существующих параметров текста и выбран набор наиболее семантически значимых из них. При помощи модуля конвертации документов из формата ODT во внутренний формат системы семантического анализа было сформировано обучающее множество документов. Используя полученное множество документов и набор выбранных параметров текста, была обучена нейронная сеть для получения структурированного документа во внутреннем формате системы семантического анализа текстов. Реализация представляет собой программный модуль на языке Java, который был интегрирован в систему семантического анализа.