УДК 004.42

Обзор программных средств обработки текстов на русском языке

Ладанова Екатерина Олеговна
Мордовский государственный университет им. Н. П. Огарева
преподаватель

Аннотация
В данной статье приводится обзор программных средств обработки текстов на русском языке, рассматриваются их достоинства, недостатки и технологии применения.
Ключевые слова: семантический анализ, обработка, текст

Overview of software tools for word processing in Russian Language

Ladanova Ekaterina Olegovna
Ogarev Mordovia State University
lecturer

Abstract
This article provides an overview of software tools for word processing in Russian Language, their advantages, disadvantages and application technologies.
Keywords: semantic analysis, processing, text

Обработка текстовой информации, прежде всего, базируется на морфологическом и синтаксическом анализе в соответствии с правилами формальной грамматики. Работа программных продуктов подразумевает использование статистической базы – т.е., текстов, которые будут применяться для обучения данных программ, и алгоритмическое индексирование словарной базы, как правило, словаря с морфологическим модификатором. Обширно применяется вероятностный подход. Немаловажными задачами анализа считается машинная обработка больших размеров информации и обобщенное представление ее смысла в сжатой форме, выявление смысловых доминант и тематической структуры, определение стиля и жанра [1,5].

Абсолютно верен тот факт, что никакая программная обработка текстовой информации не сможет заменить анализ, который осуществляется человеком. возможность поменять собой тест, который имеет возможность реализовать человек – тем более специалист в что или же другой области. Однако те программные средства, которые будут описаны далее, дают возможность потратить на проведение обработки гораздо меньше времени. Также эти программы позволяют апробировать гипотезы на большом объеме информации. Именно с данных позиций будут рассмотрено существующее сейчас программное обеспечение для обработки текстов на русском языке.

Предложенные различными командами разработчиков программы могут применяться во всевозможных областях знания и с разными целями, мы сконцентрируемся на выявлении тех позиций, которые дают положительный результат при обработке текстов на русском языке.

Первая группа компьютерных программ предназначена для синтаксического и морфологического анализа русскоязычных текстов. Грамматический срез – один из важнейших при формировании целостного представления о системе языка, так что эти программы могут быть полезны в нашем исследовании.

Russian Morphological Dictionary, разработчиком которой является Сергей Сикорский работает с входным ASCII-текстом. Программа использует морфологический словарь А.Зализняка, включающий 120.000 слов. Данное решение быстро выявляет грамматические признаки слов с довольно быстрой скоростью. Однако при работе с ПО возникает проблема из-за ограниченности словаря А.Зализняка. В этом слова нет имен собственных, некоторых неологизмов, наречий, сложных слов, которые пишутся через дефис. Исходя из этого могут возникнуть трудности при определении грамматической принадлежности некоторых русских слов.

Mystem от компании Yandex является компактным, быстрым и бесплатным решением для парсинга текстов на русском языке. Данное ПО, как и вышеописанное, также работает со словарем А.Зализняка. Есть версии для ОС Windows и Linux. Программа представляет собой приложение и представляет результаты в различных режимах. Mystem производит морфологический анализ текста на русском языке. Если встречаются слова, которых нет в словаре, программа порождают гипотезы, основанные на частотности суффиксов. Однако, если проанализировать отзывы пользователей о программе, многие из них отмечают сложности при установке и введении нужных параметров для обработки.

Рабочее Место Лингвиста от компании Dialing анализирует текстовую информация для построения систем автоматического перевода с русского на английский язык и наоборот. Программное решение состоит из следующих модулей синтаксического анализатора русскоязычных текстов и морфологический анализатор англоязычных и русскоязычных текстов. Исследователи оставляют много хороших отзывов о данном ПО, однако в настоящее время программы нет на сайте и невозможно найти альтернативный вариант для того, чтобы попробовать ее в работе.

Морфологический анализатор разработанный С.А.Старостиным представляет собой онлайн-версию программы морфологического анализа текстов на русском и английском языках. ПО работает со словарем А.А. Зализняка для обработки русских текстов и со словарем В.К. Мюллера для обработки английских текстов. В морфологический анализатор можно вводить русские и английские слова в произвольной грамматической форме. Если введенное слово многозначно, программа выводит всевозможные варианты для анализа. Эта возможность программы является наиболее удобной.

Ко второй группе программных средств для обработки текстовой информации относятся продукты, которые дают представление о частоте выявленных лексических единиц и их группировке в тексте, а также дают основания для изучения семантических процессов.

TextAnalyst 2.0 разработанный в научно-производственном инновационном центре "МикроСистемы" анализирует символьные тексты и строит семантические сети. Имеется отдельный продукт TextAnalyst, а инструментарий для разработчиков TextAnalyst SDK, который включает возможности приведения слов к нормальной форме, построения частотных списков понятий, поиска слов в контексте и другое. Все модули предоставляются бесплатно. Результаты автоматической обработки текстовой информации данной программой позволяют выявлять семантические связи. Благодаря широким возможностям данное ПО находит применение в различных исследованиях.

Galaktika-ZOOM от корпорации Галактика является автоматизированной системой поиска и обработки текстовой информации. ПО позволяет извлекать необходимые данных из больших объемов данных. Решение ведет профессиональный поиск информации по запросу. «Galaktika-ZOOM» формирует список документов, в которых есть информация, которую ищет пользователь и формирует информационный портрет объекта. Программа дает возможность получать качественные результаты за короткий период времени [2-3,6]. Вышеописанные особенности программы часто применяются в исследованиях по обработке текстов.

Система Пропись 4.0 от АО «Агама» представляет собой набор средств для лингвистической обработки текстов на русском языке, который включает проверку орфографии, расстановку переносов, построение списка синонимов и антонимов, толкование слова и статистический анализ текстов. Данная программная система широко в обучении. Также наличие функций соотнесения со словарем и статистический анализ текстов может делает это ПО полезным при обработке информации.

NetXtract разработан корпорацией Relevant Software Inc. ПО подключается к браузеру и позволяет быстро получить упорядоченный индекс слов в загружаемом HTML-документе. Индекс может упорядочить по алфавиту или частоте. При помощи данного ПО можно быстро отыскать необходимые данных на web-страницах и в документах, а также сохранять их. NetXtract позволяет автоматически индексировать документы, отображаемый в браузере, и выделять контекст для каждого термина. Можно сделать вывод, что применение данного компонента очень эффективно для обработки текстовых и web-документов.

WordStat от Yandex позволяет подсчитывать частоты встречаемости слов в текстовых и web-документах. Распознает русскую кодировку и может использоваться для быстрого извлечения и анализа информации в большом количестве документов. С помощью данное ПО можно делать контент-анализ, извлекать информации и данные, выявлять авторство. Программа является самым используемым сервисом для отображения статистики по

ключевым словам и помогает в прогнозировании трафика. Однако данный продукт требует определенного уровня компьютерной грамотности, в отличие от аналогов.

В третью группу входят системы, предназначенные для сбора данных и необходимые для определения принадлежности текстов к определенным стилям, а также выявления оригинальности текстов [4] с помощью использования нейронных сетей [7], Ансамбль-систем [8] и других методов.

Свежий взгляд, разработанный Д.Кирсановым, является DOS-утилитой, которая реализует стилистическую проверку текстов на русском языке. ПО позволяет найти места, где близко расположены фонетически и морфологически схожие. Данная функция программы является очень продуктивной для использования.

Технологии поиска и анализа текстовой информации от компании «Гарант-Парк-Интернет» представляет собой сайт, позволяющий автоматическое реферирование, морфологический, синтаксический и семантический анализ текста, а также производить навигацию по большому объему текста.

Худломер, разработанный Л. Делицыным позволяет автоматически классифицировать стили текстов на основе эмпирических кривых распределения длин слов. Программа может анализировать, к какому стилю относится входной текст (разговорная речь, художественная литература, газетная или научная статья). Данное решение широко применяется для определения стилистической принадлежностью текстов.

Лингвоанализатор, разработанный Д.В.Хмелевым является онлайн-версией программы математического анализа структуры текста. Цель анализа – определить близость введенных пользователем слов. Решение анализирует входной текст и выдает имена трех писателей, которые могли бы быть его наиболее вероятными авторами. Также программа находит три произведения авторов, которые более близки к этому тексту.

Библиографический список

1. Егунова А. И. Проектирование развивающего сайта молодёжных квестов / А. И. Егунова, Е. О. Ладанова, С. А. Ямашкин и др. // Образовательные технологии и общество. 2017. Т. 20. № 3. С. 292-298.
2. Афонин В. В. Методы моделирования и оптимизации с примерами на языке C/C++ и MATLAB. Том. Часть 1. Методы моделирования / В. В. Афонин, В. В. Никулин. Саранск : ИП Афанасьев Вячеслав Сергеевич, 2017. 188 с.
3. Афонин В. В. Методы моделирования и оптимизации с примерами на языке C/C++ и MATLAB. Том. Часть II. Методы безусловной оптимизации / В. В. Афонин, В. В. Никулин. Саранск : ИП Афанасьев Вячеслав Сергеевич, 2017. 232 с.
4. Вдовин С. М. Получение, хранение и распространение геоданных как единый информационный процесс / С. М. Вдовин, С. А. Федосин, А. А.

Ямашкин, С.А. Ямашкин // Природные опасности: связь науки и практики: материалы II Международной науч.-практ. конф. / отв. ред. С. М. Вдовин. Саранск, 2015. С. 82–90.
5. Ямашкин, С.А. Структура регионального геопортала, как инструмента публикации и распространения геопространственных данных / С. А. Ямашкин // Научно-технический вестник Поволжья. 2015. № 6. С. 223–225.
6. Афонин В.В. Моделирование систем / В.В. Афонин, С.А. Федосин. М.: Интуит, 2016. 231 с.
7. Ямашкин А.А., Ямашкин С.А. Использование нейронных сетей прямого распространения для ландшафтного картографирования на базе космических снимков // Геодезия и картография. 2014. № 11. С. 52-58.
8. Ямашкин С.А., Радованович М.М., Ямашкин А.А., Вукович Д.В., Фролов А.Н. Использование ансамбль-систем для картографирования ландшафтов // Геодезия и картография. 2016. № 7.