



050032, Қазақстан Республикасы, Алматы қ., Медеу ауданы, Алатау ш.а., Ибрагимов к., 9
«Инновациялық технология паркі» АЗА
Тел./Факс: 8 (727) 375-25-55, E-mail: info@alem.kz

050032, Республика Казахстан, г. Алматы, Медеуский район, мкр. Алатау, ул. Ибрагимова, 9
СЗЗ «Парк Инновационных технологий»
Тел./Факс: 8 (727) 375-25-55, E-mail: info@alem.kz

Исх. № 256-20 от «26» октября 2020 года

**Генеральному директору РГП на
ПВХ Института информационных и
вычислительных технологий
д.ф.-м.н., проф. , академику НАН РК
г-ну Калимолдаеву М.Н.**

**Экспертное заключение на информационно-аналитическую
поисковую систему данных на казахском языке,
выполненную в рамках проекта AP05132950.**

Предварительные сведения о проблеме

На сегодняшний день проблемы, связанные с анализом начальных данных с целью выработки оптимального решения, являются очень серьёзными и поэтому для их решения создан специальный тип систем, имеющий название аналитические информационные системы (информационно-аналитические системы). Поисковая аналитика помогает владельцам веб-сайтов понять и улучшить свою работу в поисковых системах, например, выявлять очень ценных посетителей сайта или понимать намерения пользователей. Аналитика поиска включает в себя тенденции и анализ объема поиска, обратный поиск (ввод веб-сайтов для просмотра их ключевых слов), мониторинг ключевых слов, результаты поиска и др. Так же не мало важным аспектом является исследование и решение задачи вычислительной обработки казахского языка. Казахский язык относится к агглютинативной группе языков со сложной морфологии и богатой смысловой (семантической) структурой. А так же малое количество качественных языковых электронных ресурсов и цифровых инструментов для казахского языка, указывает на востребованность исследования и разработку прикладных систем, в том числе аналитической поисковой системы, как для научных работников так и для простых пользователей.

Практическая ценность решения проблемы

Для эффективного и ускоренного решения первичного анализа текста на казахском языке разработана полная система классификация аффиксов казахского языка. Разработан lexicon-free stemming алгоритм с использованием разработанной системой классификации окончаний казахского языка. Отличительными особенностями построенного алгоритма является его скорость и достаточно легкая воспроизводимость.

Разработан размеченный корпус, отличительными особенностями является разработанные модули (токенизация, лемматизация, морфологический анализ) обработки данных с учетом свойств казахского языка. Размеченный корпус казахского языка имеет достаточно большой объем, что может в дальнейшем применен для различных прикладных задач.

Разработан модуль информационно-аналитической обработки путем создания прикладного программного решения различного назначения с применением искусственного интеллекта для обработки и анализа как структурированных, так и неструктурированных больших данных. Алгоритмы и методы данного модуля в дальнейшем могут применяться как по отдельности, так и в комплексе для решения анализа больших текстовых данных.

В настоящее время работают следующие модули:

-модуль сбора и обработки текстовых данных на казахском языке. Информационная система классифицирует текстовые данные по жанру и обрабатывает по различным параметрам;

-модуль информационно-аналитической обработки, который включает в себя следующие ниже представленные разработанные алгоритмы и метод:

- алгоритм извлечения ключевых слов (фраз) из документов на казахском языке;
- алгоритм семантического анализа текста, с применением технологий машинного обучения (Machine Learning);
- метода summarization текста на казахском языке;

-модуль ранжирования и устранения дублирования выводов результата запроса.

Рекомендации по развитию системы

Рекомендации по дальнейшему развитию системы, в том числе предлагается следующее:

- интерфейс системы полностью на казахском языке, рекомендуем добавить русский и английский языки;

-немного улучшить интерфейс системы, это придаст больше привлекательности для пользователей.

В заключении, разработанные алгоритмы и методы системы могут быть применены в различных прикладных задачах искусственного интеллекта и обработки больших текстовых данных. В связи с актуальностью и востребованностью исследований и разработок технологических решений для анализа казахоязычного контента разработанная система может быть применена для внедрения и коммерциализации.

Генеральный директор
ТОО «ALEM RESEARCH»



Нарынов С.С.