Keyword extraction

Для разрабатываемой системы поиска релевантных документов, обладающей cross-language функциональностью, входными данными будет являться текст на некотором языке либо адрес веб страницы. Язык исходного текста/веб страницы будет принадлежать множеству языков, равному пересечению множества языков, для которых наша система умеет извлекать ключевые слова, и множества языков, поддерживаемых многоязычным тезаурусом, который будет использоваться при реализации системы.

Выходными данными реализованной системы будут наборы запросов для поисковой системы Google. Язык поисковых запросов будет принадлежать множеству языков, поддерживаемых многоязычным тезаурусом, который будет использоваться при реализации системы.

Рассмотрим существующие сервисы/инструменты популярные в IT, позволяющие извлекать ключевую информацию из текста.

OpenCalais представляет собой web-сервис, разработанный компанией Thomson Reuters, который позволяет извлекать из текстов на естественном языке семантические метаданных. Он является бесплатным и также доступен для коммерческого использования Семантическими метаданными являются именованные сущности вместе с относящимися к ними фактами. В основе OpenCalais лежат методы обработки естественного языка, а также заранее подготовленные онтологии для различных предметных областей и машинное обучение. Первоначально над входным текстом выполняется графематическая и морфологическая разметка, затем полученные в ходе разметки словосочетания проходят идентификацию посредством обученной модели классификации именованных сущностей, между которыми осуществляется поиск семантических отношений. Полученный в результате граф сущностей и отношений между ними конвертируется в набор RDF-троек. Web-сервис. Поддерживаемые языки: английский, французский и испанский. Ограничения на передаваемый размер файла 100кб, 50000 запросов в сутки, до 4 запросов в секунду по одному ключу.

IBM's Watson Developer Cloud(ранее [**AlchemyAPI**](http://www.alchemyapi.com/)) предоставляет возможность анализа текста на естественном языке для извлечения семантических метаданных. Входными данными для данного сервиса может являться как обычный текст, html, так и url-адрес некоторого веб сайта. Сервис предварительно очищает HTML перед анализом, удаляя большинство рекламных объявлений и другой нежелательный контент. Поддерживает языки: английский, французский, немецкий, итальянский, португальский, русский, испанский. Ограничения для бесплатной версии: 1000 баллов в сутки, 1 балл – 1 запрос размером до 10000 символов и до 50 кб. Запросы превышающие по размеру ограничения стоят большее количество баллов.

Yahoo Content Analysis(ранее Yahoo! Term Extraction Web Service) — сервис, задействованный в работе поисковой системы Yahoo! Search. Имеет возможность обнаруживать ключевые фразы из текста на естественном языке. Подход к извлечению терминов в документации не описан. Обмен данными с пользователем осуществляется в форматах XML и JSON. Ограничение 5000 запросов в сутки и не доступен для коммерческого использования.

Extractor набор инструментов разработчика для автоматического извлечения терминов. Предназначен для обработки естественного языка. В основе системы Extractor, согласно документации, лежит машинное обучение, генетические алгоритмы, а также статистические методами обработки естественного языка. Перед использованием систему нужно обучить на корпусе текстов, который предварительно был размечен.

Mining Cloud (ранее Text Analytics) — сервис, предназначенный для поиска информации и анализа содержания текстов, в основе которого лежат методы обработки естественного языка, а также методы машинного обучения. Mining Cloud позволяет пользователям встраивать текстовую аналитику и семантическую обработку в любое приложение или систему, достаточно простым благодаря облачной инфраструктуре, с которой легко интегрироваться. Mining Cloud предоставляет следующую функциональность: извлечение темы, посредством распознавания именованных сущностей тексте; классификация текстов через присваивание им одной или нескольких категорий в предопределенной таксономии (сервис включает несколько стандартных таксономий классификации из коробки); определение эмоциональной окраски (положительная, отрицательная, нейтральная) документа или его отдельных частей. Сервис также предлагает расширенные API-интерфейсы, такие как дополнительные тезаурусы, таксономии и т.п., оптимизированные для разных отраслей и сценариев приложений. Большинство данных API доступны на следующих языках: английском, испанском, французском, итальянском, португальском.

Stanford's Core NLP Suite предоставляет набор инструментов анализа естественного языка. Система поддерживает английский, китайский, французский, немецкий и испанский языки и включает в себя инструменты для разметки текста (разбиение текста на слова), определение базовой формы слова, части речи, извлечение именных сущностей, ключевых слов и т.д. Stanford CoreNLP предназначен для того, чтобы очень легко применить большое число инструментов лингвистического анализа к фрагменту текста, написав несколько строк кода, CoreNLP является достаточно гибким и расширяемым. Stanford CoreNLP объединяет многие инструменты Stanford’s NLP, включая частеречную разметку, распознавание именованных сущностей, синтаксический анализатор, определение эмоциональной окраски фрагмента текста и т.д.

Natural Language Toolkit - пакет библиотек и программ, предназначенный для анализа естественного языка в приложениях, разработанных на языке Python. Он предоставляет возможность выполнять следующие операции над исходным текстом: классификации, токенизации, стемминг, тэгирование и т.д. Существует подробная документация по данному пакету, в том числе объясняющая основные концепции, встречающиеся в задачах обработки естественного языка, которые можно решить с помощью данного пакета.

Apache OpenNLP - интегрированный пакет инструментов, предназначенных для обработки текста на естественном языке и работающих на основе машинного обучения. Пакет работает на платформе Java и поддерживает наиболее распространенные задачи обработки естественного языка, такие как токенизация, сегментирование предложений, частеречная разметка, извлечение именованных сущностей, интаксического разбора предложения и т.д. Эти задачи часто встречаются при реализации систем обработки текста. Работать с данным пакетом можно посредством прикладного программного интерфейса или через командную строку. Apache OpenNLP можно использовать на условиях лицензии Apache License. Исходный код данного пакета присутствует на официальном сайте проекта.

В качестве сервиса/инструмента для извлечения ключевой информации из текста был выбран IBM's Watson Developer Cloud. При выборе учитывались следующие параметры:

Простота интеграции и отсутствие необходимости в поднятия собственного сервера;

В качестве входных данных можно передавать, как обычный текст, так и url-адрес веб сайта, в последнем случае сервис на этапе предварительной обработки очистит веб страницу от рекламы и другого нежелательного контента;

Поддерживает языки: английский, русский, французский, немецкий, итальянский, португальский, испанский;

Присутствует бесплатная версия API с ограниченным числом запросов в сутки.

Реализация.

Учитывая все выше перечисленное, в дипломной работе была поставлена задача по разработке мобильного приложения под операционную систему Android, которое будет простым в использовании, обладать удобным и приятным интерфейсом, минималистическим, поддерживать весь спектр существующих устройств с версией API 14+(Android 4.0), тем самым, согласно официальной статистике от google, будет поддерживать 99.1% активных девайсов.

Интерфейс для работы с приложение будет реализован в соответствии с концепцией material design.

При реализации данной системы для извлечения ключевой информации из поступающего на вход текста/url-адреса веб сайта воспользуемся IBM's Watson Developer Cloud, для разрешения лексической многозначности используем многоязычный тезаурус BabelNet.