Текст майнинг (Text Mining)

> Синонимы: Text data mining, Text analytics, Интеллектуальный анализ текстов, Анализ текстов
> Разделы: Бизнес-задачи

Text Mining — это технология получения информации из неструктурированных текстовых данных путем их преобразования в пригодный для дальнейшей работы набор структурированных данных, представленных в удобном для машинной обработки виде. То есть, посредством методов Text Mining пользователь может извлекать знания из огромных массива данных, лишенной «понятной» компьютеру структуры.

Text Mining обычно включает в себя процесс структурирования исходного текста с применением синтаксического анализа, использования некоторых лингвистических функций с последующей загрузкой в базу данных и интерпретацией результатов. Главной целью является превращение текста в структурированные данные, пригодные для анализа методами интеллектуального анализа данных.

Результаты анализа текста оцениваются с точки зрения некоторых критериев качества, включающих актуальность, новизну и интерес. Типичные задачи анализа текста включают:

- категоризацию,
- кластеризацию,
- извлечение концептов (сущностей),
- разработку таксономий,
- обобщение документов,
- моделирование отношений между сущностями,
- тематическое индексирование,
- поиск по ключевым словам,
- изучение частотных распределений слов,
- аннотирование и т.д.

Следует отметить, что в сфере аналитических технологий имеет место некоторая несогласованность терминологии в отношении понятия Text Mining. Некоторые источники переводят его как интеллектуальный анализ текста, по аналогии с Data Mining (интеллектуальный анализ данных), другие же ограничиваются просто «анализом текста».

Под анализом текста в настоящее время понимают набор лингвистических, статистических процедур и методов машинного обучения, которые моделируют и структурируют информационный контент текстовых источников для бизнес-аналитики и

интеллектуального анализа данных. В последнее время термин «анализ текста» чаще используется в бизнес-среде, в то время как «интеллектуальный анализ текста» относится к ранним этапам применения технологии (1980-е годы).

Термин «анализ текста» также описывает реагирование на проблемы бизнеса, независимо или в сочетании с анализом данных. Действительно, 80% деловой информации поступает в неструктурированной форме, в основном в виде текста. Методы и процессы анализа текстов обнаруживают и представляют знания и бизнес-правила, которые оказываются «заблокированными» в текстовой форме, недоступной для автоматической обработки.

Процесс Text Mining обычно содержит следующие этапы:

- сбор и идентификация набора текстовых источников из Интернета, файлов документов, баз данных и т.д.;
- распознавание именованных объектов — это использование справочников или статистических методов для идентификации именованных текстовых объектов: людей, организаций, географических названий, товаров, брендов и т. д.
- устранение неоднозначностей — использование контекстных подсказок для интерпретации неоднозначных понятий (например, машина — это и транспортное средство, и компьютер, и механизм);
- распознавание объектов, идентифицированных по шаблону — номеров телефонов, адресов обычной и электронной почты, количества (с единицами измерения) можно распознать с помощью регулярного выражения или другого соответствия шаблону;
- кластеризация документов: идентификация наборов похожих текстовых документов;
- идентификация имен существительных и других терминов, относящихся к одному и тому же объекту (кореферентность).
- обнаружение фактов и событий, взаимосвязей между ними, выявление ассоциаций между сущностями;
- анализ настроений включает в себя распознавание субъективного аспекта и извлечение различных форм поведенческой информации: настроения, мнения, эмоций.

Технологии Text Mining в настоящее время широко применяется для решения различных задач в области бизнеса, научных исследований, государственного управления, разведки и безопасности.

О том, как автоматизировать процесс категоризации текстовых данных с помощью Loginom, можно узнать в статье.