Математические и инструментальные методы машинного обучения

10. Интеллектуальный анализ текста

Определение текстового анализа

Текстовый анализ — это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных.

Суть текстового анализа

- Превращение неструктурированного текста в структурированные объекты

Статистическая обработка естественного языка

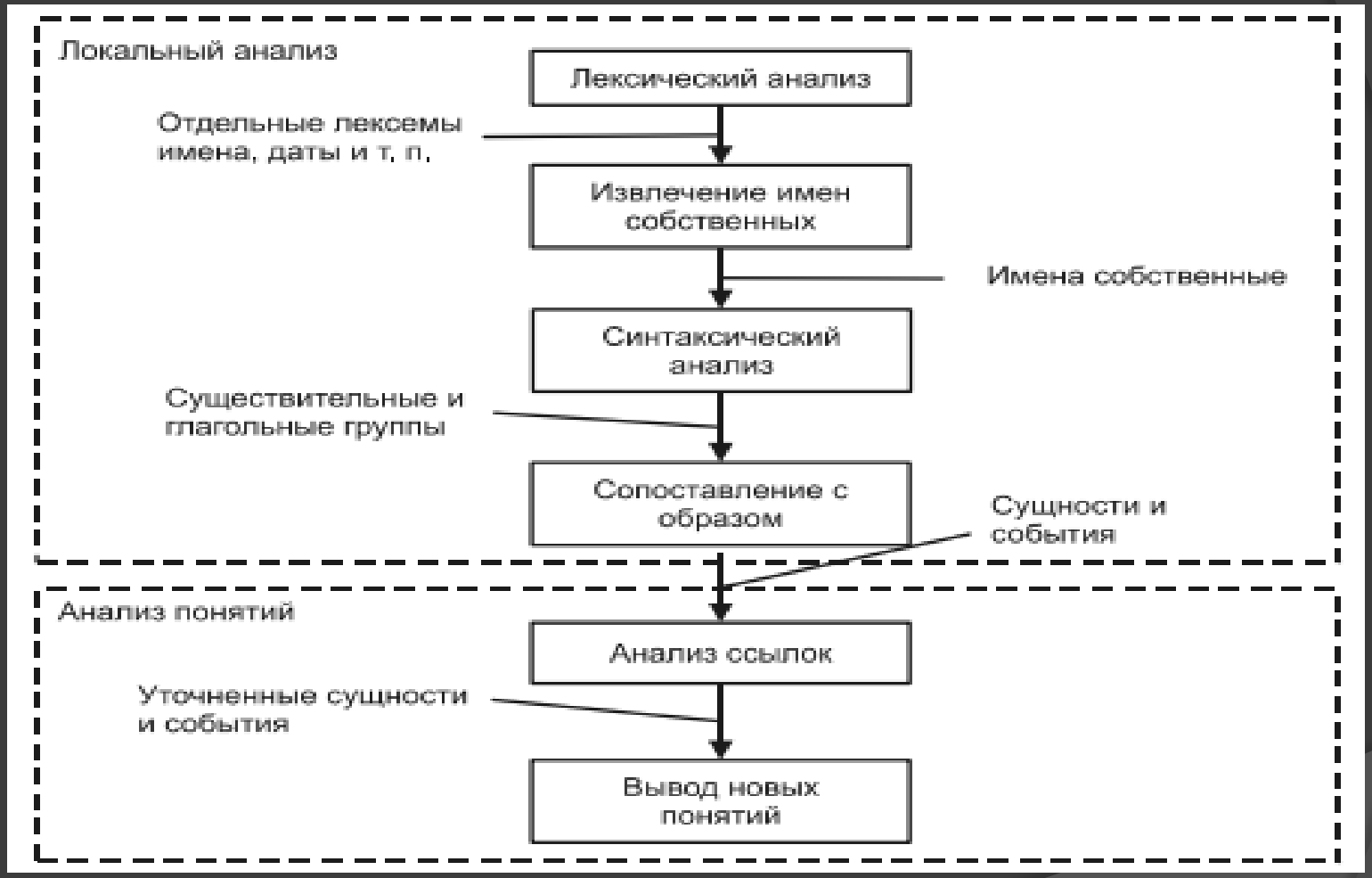
Интеллектуальный анализ данных

- Количественный анализ полученных объектов для получения знания

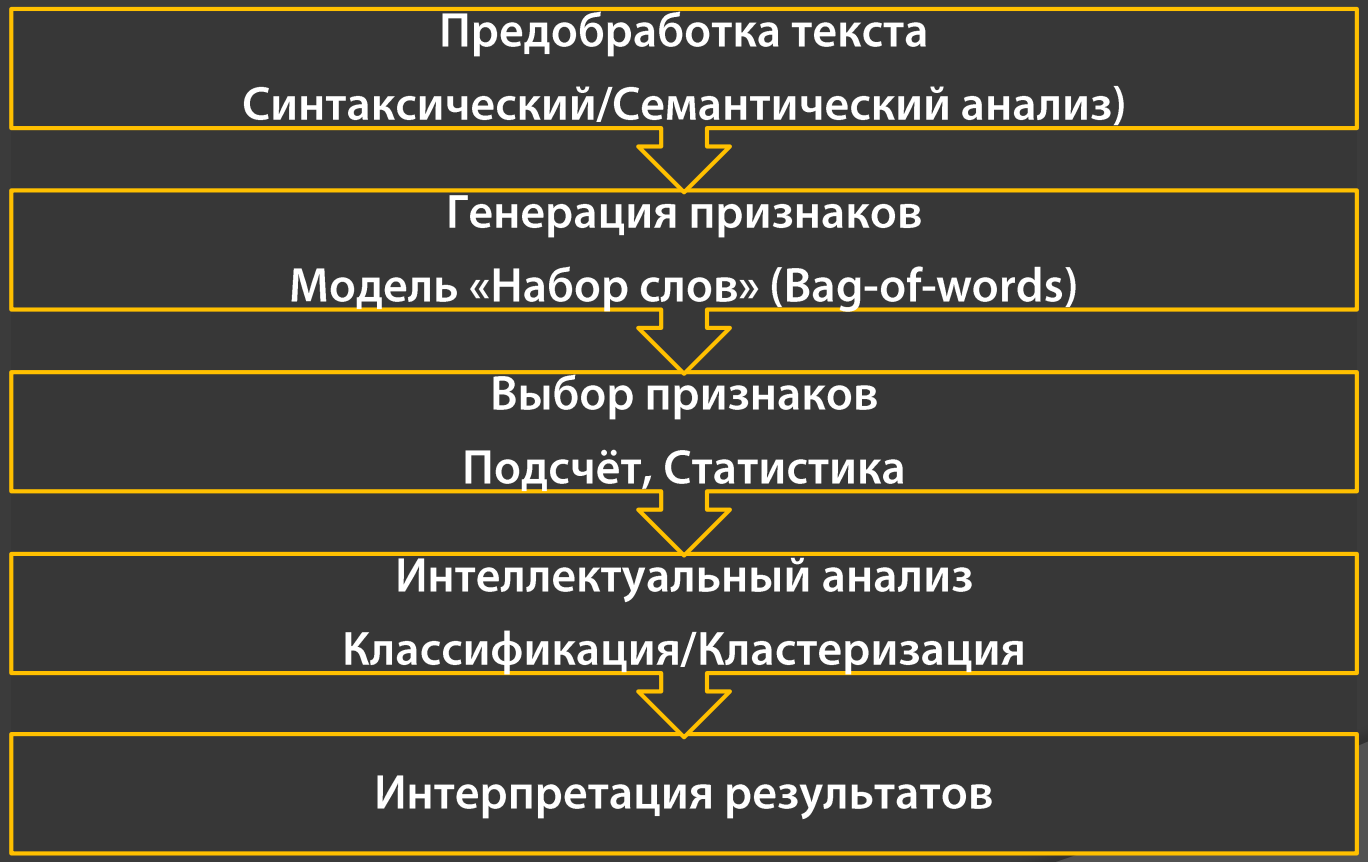
Виды текстового анализа

- Описательный
  - Анализ шаблонов и трендов
  - Создание базы знаний
  - Аннотирование
  - Визуализация
- Предсказательный
  - Классификация
  - Кластеризация
  - Прогноз шаблонов и трендов

Извлечение фактов из текста



Процесс текстового анализа



Генерация признаков

- Генерация признаков — это процесс и процедура создания и извлечения числовых признаков из сырых данных, которые можно подать на вход какой-либо модели для обучения.

Выбор признаков

- Извлечение признаков — это процесс построения информативных признаков из исходных, которые в будущем приведут к более быстрому обучению или могут лучше интерпретироваться.

Стемминг

Стемминг — это процесс нахождения основы слова для заданного исходного слова. Основа слова необязательно совпадает с морфологическим корнем слова.

Лемматизация

Лемматизация — процедура, в результате которой удаляются только флективные окончания и возвращается основная, или словарная, форма слова, называемая леммой.

Удаление стоп-слов

Стоп-слова — это слова, не несущие какой-либо самостоятельной смысловой нагрузки.

- Союзы и союзные слова
- Местоимения
- Предлоги
- Частицы
- Междометия
- Указательные слова
- Цифры
- Знаки препинания
- Вводные слова

N-граммы

N-граммы — это альтернатива морфологическому разбору и удалению стоп-слов. N-грамма — это часть строки, состоящая из N символов. Например, слово "дата" может быть представлено 3-граммой «_да», «дат», «ата», «та_» или 4-граммой «_дат», «дата», «ата_», где символ подчеркивания заменяет предшествующий или замыкающий слово пробел.

Применение текстового анализа

- Классификация (определение спама, организация документов)
- Кластеризация (анализ трендов, определение тематики)
- Веб-анализ (анализ трендов, извлечение мнений, создание онтологий)
- Классическая обработка естественного языка (аннотирование текстов, ответы на вопросы, извлечение информации)

Применение текстового анализа

- Большинство алгоритмов кластеризации требуют, чтобы данные были представлены в виде модели векторного пространства (*vector space model*).
- В этой модели каждый документ представляется в многомерном пространстве, в котором каждое измерение соответствует слову в наборе документов.
- Набор измерений конструируется при помощи исключения редких слов и слов с высокой частотой.

TF-IDF

- *TF* (term frequency — частота слова) — отношение числа вхождения некоторого слова к общему количеству слов документа. Таким образом, оценивается важность слова ti в пределах отдельного документа.
- *IDF* (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт *IDF* уменьшает вес широкоупотребительных слов.
- *TF-IDF* — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

$$TF = \frac{n_i}{\sum_k n_k} \qquad IDF = \log \frac{|D|}{|(d_i \supset t_i)|}$$