1. Анализ естественного языка. Мотивация. Направления.

Обработка естественного языка (Natural language processing - NLP) – область, находящаяся на пересечении computer science, искусственного интеллекта и лингвистики. Цель заключается в обработке естественного языка для перевода текста и ответа на вопросы.

**Где применяется NL**

* • поиск (письменный или устный);
* автоматический (или при содействии) перевод;
* распознавание речи и чат-боты,
* голосовые помощники.

**Машинный перевод**

Преобразование текста на одном естественном языке в эквивалентный по содержанию текст на другом языке. Делает это программа или машина без участия человека. В машинном переводе использутся статистика использования слов по соседству.

**Голосовые помощники**

Программы для поддержания разговора с собеседником. Например, исследователи из Microsoft и Facebook создали нейросеть, способную создавать чувствительные к контексту ответы в разговоре. Эта система может тренироваться на большом количестве неструктурированных диалогов в Twitter.

**Вопросно-ответные (QA) системы**

Идея систем заключается в извлечении информации непосредственно из документа, разговора, онлайн поиска или любого другого места, удовлетворяющего потребности пользователя. Вместо того, чтобы заставлять пользователя читать полный текст, QA системы предпочитают давать короткие и лаконичные ответы.

**Краткое изложение текста (Text Summarization)**

Извлечение краткого содержания (Text Summarization) – инструмент для помощи в интерпретации текстовой информации.

Выделяют два фундаментальных подхода к сокращению текста: извлекательный и абстрактный.

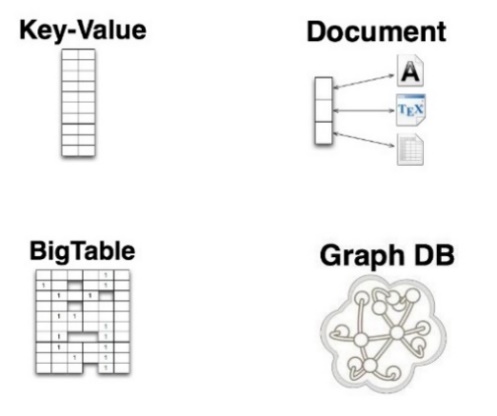
Первый извлекает слова и фразы из оригинального текста для создания резюме.

Последний изучает внутреннее языковое представление, чтобы создать человекоподобное изложение, перефразируя оригинальный текст.

Методы в извлекательном сокращении работают на основе выбора подмножества. LexRank и TextRank — хорошо известные представители этого подхода, которые используют вариации алгоритм сортировки страниц Google PageRank.

1. Классификация и различия NoSQL хранилищ

NoSQL использует разные хранилища данных для оптимизации для конкретных целей в зависимости от требований к системе, на рисунке 4 категории хранилищ данных NoSQL:



**1. key-value store - Berkeley DB, MemcacheDB, Redis, Riak, Amazon DynamoDB**

Сохраняет данные (значения) и применяет к ним метку (ключ) и сохраняет их либо в памяти, либо в системе хранения, оптимизированной для быстрого извлечения данных.

Значение обычно можно получить, только сославшись на его ключ

Базы данных типа "ключ-значение" отлично подходят для случаев, когда вам нужно хранить большие объемы данных, но вам не нужно выполнять сложные запросы для их получения.

**2. document store - CouchDB, Couchbase, MongoDB, Berkeley DB XML**

Предназначен для хранения, поиска и управления документальной информацией. Хранилища документов сохраняют всю информацию об объекте как экземпляр в базе данных, а не распределяют данные по разным таблицам.

Может масштабироваться по горизонтали для размещения больших объемов данных.

**3. wide column database - Apache HBase, Apache Cassandra, Apache Accumulo, Amazon SimpleDB**

Использует таблицы, строки и динамические столбцы, отличные от реляционных баз данных, поскольку для каждой строки не требуется один и тот же столбец, что позволяет горизонтальное масштабирование базы данных.

Пример использования - хранение профилей пользователей.

**4. graph database - Neo4j, MarkLogic, Ontotext, OrientDB**

Хранит данные в виде узлов и ребер. Узлы хранят данные объекта, а ребра хранят информацию о взаимосвязи узлов.

Используется для просмотра взаимосвязей и закономерностей в данных. Варианты использования - социальные сети, искусственный интеллект и механизм рекомендаций.

Graph отлично подходят для тех случаев, когда вам нужно изучить отношения для поиска паттернов, таких как социальные сети, обнаружение мошенничества и механизмы рекомендаций.