**Keywords** (drawn from Wikipedia):

Concept Search:

A **concept search** (or **conceptual search**) is an automated [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval) method that is used to search electronically stored [unstructured text](https://en.wikipedia.org/wiki/Unstructured_data) (for example, [digital archives](https://en.wikipedia.org/wiki/Digital_archive), email, scientific literature, etc.) for information that is conceptually similar to the information provided in a search query. In other words, the *ideas* expressed in the information retrieved in response to a [concept](https://en.wikipedia.org/wiki/Concept) search query are relevant to the ideas contained in the text of the query.

<https://en.wikipedia.org/wiki/Concept_search>

Elasticsearch:

**Elasticsearch** is a [search engine](https://en.wikipedia.org/wiki/Search_engine_(computing)) based on [Lucene](https://en.wikipedia.org/wiki/Lucene). It provides a distributed, [multitenant](https://en.wikipedia.org/wiki/Multitenancy)-capable [full-text search](https://en.wikipedia.org/wiki/Full-text_search) engine with an [HTTP](https://en.wikipedia.org/wiki/HTTP) web interface and schema-free [JSON](https://en.wikipedia.org/wiki/JSON) documents. Elasticsearch is developed in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) and is released as [open source](https://en.wikipedia.org/wiki/Open_source_software) under the terms of the [Apache License](https://en.wikipedia.org/wiki/Apache_License). Official clients are available in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), [.NET](https://en.wikipedia.org/wiki/.NET_Framework) ([C#](https://en.wikipedia.org/wiki/C_Sharp_(programming_language))), [PHP](https://en.wikipedia.org/wiki/PHP_(programming_language)), [Python](https://en.wikipedia.org/wiki/Python_(programming_language)), [Apache Groovy](https://en.wikipedia.org/wiki/Groovy_(programming_language)),Ruby and many other languages.[[2]](https://en.wikipedia.org/wiki/Elasticsearch#cite_note-offizsite-2) According to the [DB-Engines ranking](https://en.wikipedia.org/wiki/DB-Engines_ranking), Elasticsearch is the most popular enterprise search engine followed by [Apache Solr](https://en.wikipedia.org/wiki/Apache_Solr), also based on Lucene.[[3]](https://en.wikipedia.org/wiki/Elasticsearch#cite_note-3)

<https://en.wikipedia.org/wiki/Elasticsearch>

Information Retrieval:

**Information retrieval** (**IR**) is the activity of obtaining [information](https://en.wikipedia.org/wiki/Information) resources relevant to an information need from a collection of information resources. Searches can be based on [full-text](https://en.wikipedia.org/wiki/Full-text_search) or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for [metadata](https://en.wikipedia.org/wiki/Metadata) that describe data, and for databases of texts, images or sounds.

<https://en.wikipedia.org/wiki/Information_retrieval>

Lucene:

**Apache Lucene** is a [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source_software) [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval) [software library](https://en.wikipedia.org/wiki/Library_(computing)), originally written completely in [Java](https://en.wikipedia.org/wiki/Java_(programming_language)) by [Doug Cutting](https://en.wikipedia.org/wiki/Doug_Cutting). It is supported by the [Apache Software Foundation](https://en.wikipedia.org/wiki/Apache_Software_Foundation) and is released under the [Apache Software License](https://en.wikipedia.org/wiki/Apache_Software_License).

<https://en.wikipedia.org/wiki/Apache_Lucene>

Named-Entity Recognition:

**Named-entity recognition** (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of [information extraction](https://en.wikipedia.org/wiki/Information_extraction) that seeks to locate and classify [named entities](https://en.wikipedia.org/wiki/Named_entity) in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

<https://en.wikipedia.org/wiki/Named-entity_recognition>

Optical Character Recognition:

**Optical character recognition** (also **optical character reader**, **OCR**) is the [mechanical](https://en.wikipedia.org/wiki/Machine) or [electronic](https://en.wikipedia.org/wiki/Electronics) conversion of [images](https://en.wikipedia.org/wiki/Image) of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).[[1]](https://en.wikipedia.org/wiki/Optical_character_recognition#cite_note-1) It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitising printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as [cognitive computing](https://en.wikipedia.org/wiki/Cognitive_computing), [machine translation](https://en.wikipedia.org/wiki/Machine_translation), (extracted) [text-to-speech](https://en.wikipedia.org/wiki/Text-to-speech), key data and [text mining](https://en.wikipedia.org/wiki/Text_mining). OCR is a field of research in [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) and [computer vision](https://en.wikipedia.org/wiki/Computer_vision).

<https://en.wikipedia.org/wiki/Optical_character_recognition>

Relational Database:

A **relational database** is a digital [database](https://en.wikipedia.org/wiki/Database) based on the [relational model](https://en.wikipedia.org/wiki/Relational_model) of data, as proposed by [E. F. Codd](https://en.wikipedia.org/wiki/E._F._Codd) in 1970. A software system used to maintain relational databases is a [relational database management system](https://en.wikipedia.org/wiki/Relational_database_management_system) (RDBMS). Virtually all relational database systems use [SQL](https://en.wikipedia.org/wiki/SQL) (Structured Query Language) for querying and maintaining the database.

<https://en.wikipedia.org/wiki/Relational_database>

Structured Query Language (SQL):

**SQL** ([/ˌɛs.kjuːˈɛl/](https://en.wikipedia.org/wiki/Help:IPA/English) ([About this sound](https://en.wikipedia.org/wiki/File:En-us-SQL.ogg) [listen](https://upload.wikimedia.org/wikipedia/commons/5/5f/En-us-SQL.ogg)) *S-Q-L*,[[4]](https://en.wikipedia.org/wiki/SQL#cite_note-learningSQL-4) [/ˈsiːkwəl/](https://en.wikipedia.org/wiki/Help:IPA/English) "sequel"; **Structured Query Language**)[[5]](https://en.wikipedia.org/wiki/SQL#cite_note-Britannica-5)[[6]](https://en.wikipedia.org/wiki/SQL#cite_note-oed-US-6)[[7]](https://en.wikipedia.org/wiki/SQL#cite_note-IBM-SQL-7)[[8]](https://en.wikipedia.org/wiki/SQL#cite_note-MS-SQL-def-8) is a [domain-specific language](https://en.wikipedia.org/wiki/Domain-specific_language) used in programming and designed for managing data held in a [relational database management system](https://en.wikipedia.org/wiki/Relational_database_management_system) (RDBMS), or for stream processing in a [relational data stream management system](https://en.wikipedia.org/wiki/Relational_data_stream_management_system) (RDSMS). In comparison to older read/write [APIs](https://en.wikipedia.org/wiki/API) like [ISAM](https://en.wikipedia.org/wiki/ISAM) or [VSAM](https://en.wikipedia.org/wiki/VSAM), SQL offers two main advantages: first, it introduced the concept of accessing many records with one single command; and second, it eliminates the need to specify *how* to reach a record, e.g. with or without an [index](https://en.wikipedia.org/wiki/Database_index).

<https://en.wikipedia.org/wiki/SQL>

Term Frequency – Inverse Document Frency (tf-idf):

In [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), **tf–idf** or **TFIDF**, short for **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a [document](https://en.wikipedia.org/wiki/Document) in a collection or [corpus](https://en.wikipedia.org/wiki/Text_corpus). It is often used as a [weighting factor](https://en.wikipedia.org/wiki/Weighting_factor) in searches of information retrieval, [text mining](https://en.wikipedia.org/wiki/Text_mining), and [user modeling](https://en.wikipedia.org/wiki/User_modeling). The tf-idf value increases [proportionally](https://en.wikipedia.org/wiki/Proportionality_(mathematics)) to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Nowadays, tf-idf is one of the most popular term-weighting schemes; 83% of text-based recommender systems in the domain of digital libraries use tf-idf.

<https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

Text Mining:

**Text mining**, also referred to as ***text data mining***, roughly equivalent to **text analytics**, is the process of deriving high-quality [information](https://en.wikipedia.org/wiki/Information) from [text](https://en.wikipedia.org/wiki/Plain_text). High-quality information is typically derived through the devising of patterns and trends through means such as [statistical pattern learning](https://en.wikipedia.org/wiki/Pattern_recognition). Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a [database](https://en.wikipedia.org/wiki/Database)), deriving patterns within the [structured data](https://en.wikipedia.org/wiki/Structured_data), and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of [relevance](https://en.wikipedia.org/wiki/Relevance_(information_retrieval)), [novelty](https://en.wikipedia.org/wiki/Novelty_(patent)), and interestingness. Typical text mining tasks include [text categorization](https://en.wikipedia.org/wiki/Text_categorization), [text clustering](https://en.wikipedia.org/wiki/Text_clustering), [concept/entity extraction](https://en.wikipedia.org/wiki/Concept_mining), production of granular taxonomies, [sentiment analysis](https://en.wikipedia.org/wiki/Sentiment_analysis), [document summarization](https://en.wikipedia.org/wiki/Document_summarization), and entity relation modeling (*i.e.*, learning relations between [named entities](https://en.wikipedia.org/wiki/Named_entity_recognition)).

<https://en.wikipedia.org/wiki/Text_mining>

Unstructured Text:

**Unstructured data** (or **unstructured information**) is information that either does not have a pre-defined [data model](https://en.wikipedia.org/wiki/Data_model) or is not organized in a pre-defined manner. Unstructured information is typically [text](https://en.wikipedia.org/wiki/Plain_text)-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and [ambiguities](https://en.wikipedia.org/wiki/Ambiguities) that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or [annotated](https://en.wikipedia.org/wiki/Annotation) ([semantically tagged](https://en.wikipedia.org/wiki/Tag_(metadata))) in documents.

<https://en.wikipedia.org/wiki/Unstructured_data>