What are the classifications of Data?

1. Structured Data

Definition

Structured data refers to a type of data *organized in a highly predictable and organized manner*, typically in a tabular format, making it very easy to search, sort, and analyze using traditional information retrieval systems.

This type of data is typically quantitative in nature, meaning it can be represented numerically, and is often stored in a database or spreadsheet.

Examples of structured data include financial records, inventory lists, customer databases, addresses, and phone numbers. It's also easy to query with a structured query language like SQL.

Benefits & Drawbacks

Benefits	**Drawbacks**
Easy to use ● For machine learning ● Business users	Its use is limited
	Storage is inflexible ● Any changes require an update of all data.
	It's difficult to unify ● For merging data, normalization is required.

2. Unstructured Data

Definition

Unstructured data refers to a type of data that *does not have a predefined or organized format*. This type of data is text-based or non-text based and oftentimes is made by humans, although more and more unstructured data is now machine-generated. Unlike structured data, which is highly organized and predictable, unstructured data can take on a variety of forms, including text, images, audio, and video.

Due to its lack of structure, unstructured data can be more difficult to process and analyze using traditional data analysis methods. However, advances in natural language

processing and machine learning have enabled the development of tools and techniques to better handle unstructured data.

Examples of unstructured data include social media posts, emails, customer reviews, and news articles. By using an intelligent search engine that is powered by machine learning, companies can allow data scientists to gain valuable and actionable insights such as mining social media posts and product reviews and predictive analytics.

Benefits & Drawbacks

Benefits	Drawbacks
Flexible formats	Its use is limited
Easy Storage	Storage is inflexible ● Any changes require an update of all data.
Competitive advantage ● This data can be used effectively due to its volume and the ability to provide more substantial insights into customers,	It's difficult to unify ● For merging data, normalization is required.

3. Semi-structured Data

Definition

Semi-structured data refers to a type of data that has some structure, but not enough to fit into the traditional structure of a database or spreadsheet. Unlike structured data, which is organized into fields and tables, semi-structured data may have some organizational elements, such as tags or keys, but still requires more complex processing to extract meaningful information.

Examples of semi-structured data include XML and JSON files, as well as emails and social media posts that may include hashtags or other metadata. Semi-structured data is becoming increasingly important in fields such as web data analytics and big data, as it allows for more flexible data processing and analysis than traditional structured data.

Note - Emails and social media posts structure

Emails and social media posts are generally considered to be unstructured data, as they do not have a predefined format and can contain a variety of information types, such as text, images, links, and hashtags. However, some aspects of these data sources, such as email headers or social media metadata, may have a more structured format, which makes them semi-structured.

Overall, emails and social media posts can be considered a mix of structured, semi-structured, and unstructured data depending on the specific context and how the data is analyzed

Summary

Structured Data	Semi-structured Data	Unstructured Data
Predefined format	Has some structure - not a rigid schema	No predefined formats
Stored in a database or spreadsheet	May contain metadata or tags	Any format of data can be stored(text, images, audio, video)
Comprises numerical or categorical values	Able to be stored in a variety of formats(XML, JSON)	Able to be stored in a variety of formats(emails, social media posts, web pages)
Easy to search and analyze	Require complex processing to extract needed information	More difficult to analyze using traditional data processing tools(Natural language processes or machine learning techniques can provide valuable insights)

References

https://www.coveo.com/blog/structured-vs-unstructured-data/
https://chat.openai.com/

Natural Language Processing (NLP)

It is a subfield of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language. NLP is concerned with various aspects of language, such as grammar, syntax, semantics, and pragmatics. It involves the use of statistical models, machine learning algorithms, and deep neural networks to analyze and process human language data, including text, speech, and images. Some common NLP applications include sentiment analysis, machine translation, speech recognition, and text summarization.

Various types of NLP systems

- Rule-based systems: These systems rely on a set of pre-defined rules and patterns to analyze and process language data. They are often used for tasks such as spell-checking and grammar correction. Patterns are detected based on pre-defined rules.
- Statistical systems: These systems use statistical models and algorithms to analyze language data. They are often used for tasks such as language identification and machine translation. Patterns are detected based on the statistical relationships found in a specific text.
- Machine learning systems: These systems use machine learning algorithms to learn from language data and improve their performance over time. They are often used for tasks such as sentiment analysis and named entity recognition. Patterns are detected based on how machine learning is trained.
- Deep learning systems: These systems use deep neural networks to process and analyze language data. They are often used for tasks such as speech recognition and natural language understanding.
- Hybrid systems: These systems combine multiple approaches, such as rule-based and statistical or machine learning and deep learning, to improve performance on specific NLP tasks. Patterns are detected based on both a rule-based and statistical analysis.

References

https://www.rchnfoundation.org/?p=5349#:~:text=Auto%2Dindexing%20is%20one%20of,from%20free%20text%20or%20speech

IR stages

1. Information Need Identification:
This involves understanding the user's information needs and determining what kind of information they are looking for.

2. Document Collection:
This involves gathering a set of documents that may be relevant to the user's information needs.

3. Preprocessing and Indexing:

This involves cleaning and processing the document collection to prepare it for efficient searching. This includes tasks such as tokenization, stopword removal, stemming, and building an index to support fast searching.

4. Query Processing:
This involves processing the user's query to identify relevant documents from the collection. This may involve techniques such as query expansion, relevance feedback, and ranking.

5. Retrieval and Ranking:
This involves retrieving the relevant documents from the collection and ranking them based on their relevance to the query.

6. Presentation of Results:
This involves presenting the search results to the user in a meaningful and easy-to-understand format.

7. Evaluation:
This involves measuring the effectiveness of the search system through various metrics such as precision, recall, and F1-score, and improving the system based on the evaluation results.

Important Terms in IR

These are the important terms that should be known in IR.

What is an IR Model?

An Information Retrieval (IR) model selects and ranks the document that is required by the user or the user has asked for in the form of a query. The documents and the queries are represented in a similar manner, so that document selection and ranking can be formalized by a matching function that returns a retrieval status value (RSV) for each document in the collection.

Many Information Retrieval systems represent document contents by a set of descriptors, called terms, belonging to a vocabulary V.

An IR model determines the query-document matching function according to four main approaches:

- **Acquisition**: During this step, we gather text-based documents and other relevant content from different sources on the web. Web crawlers help us find and collect the necessary data, which is then saved in a database for future use.
- **Representation**: It consists of indexing that contains free-text terms, controlled vocabulary, manual, and automatic techniques as well. Example: Abstracting contains summarizing and Bibliographic description that contains author, title, sources, data, and metadata.
- **File Organization**: There are two types of file organization methods. i.e.
    - Sequential: It contains documents by document data.

    - Inverted: It contains term by term, a list of records under each term. Combination of both.
- **Query**: An IR process starts when a user enters a query into the system. Queries are formal statements of information needs, for example, search strings in web search engines. In information retrieval, a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

What is Query?

Definition

A query can either be a request for data results from your database or for action on the data, or for both. A query can give you an answer to a simple question, perform calculations, combine data from different tables, and add, change, or delete data from a database. The query is typically entered by a user.

Query Language

For a machine to understand a request for information in the first place, the query must be written according to a code known as query language.

SQL represents one of the standard languages used for database management purposes, while MySQL, instead, is the software using that specific language. Although SQL is a fairly universal query language, other commonly used ones include DMX, Datalog, and AQL.

Types of Queries

SQL represents one of the standard languages used for database management purposes, while MySQL, instead, is the software using that specific language. Although SQL is a fairly universal query language, other commonly used ones include DMX, Datalog, and AQL.

Keyword Queries	• Simplest and most common queries. • The user enters just keyword combinations to retrieve documents. • These keywords are connected by logical AND operator. • All retrieval models provide support for keyword queries.
Boolean Queries	• Some IR systems allow using +, -, AND, OR, NOT, ( ), Boolean operators in a combination of keyword formulations. • No ranking is involved, either document satisfies such a query or does not satisfy it. (0, 1) • A document is retrieved for a boolean query if it is logically true as an exact match in a document.

Phrase Queries	• When documents are represented using an inverted keyword index for searching, the relative order of items in a document is lost. • To perform exact phase retrieval, these phases are encoded in an inverted index or implemented differently. • This query consists of a sequence of words that make up a phase. • It is generally enclosed within double quotes.
Proximity Queries	• Searching for words that are close to each other, but not necessarily next to each other. • Most commonly used proximity search option is a phase search that requires terms to be in exact order. • Other proximity operators can specify how close terms should be to each other. Some will specify the order of search terms. • Search engines use various operator's names such as NEAR, ADJ (adjacent), or AFTER. • However, providing support for complex proximity operators becomes expensive as it requires time-consuming pre-processing of documents and so it is suitable for smaller document collections rather than for the web.
Wildcard Queries	• It supports regular expressions and pattern matching-based searching in text. • For locating a specific item when you can't remember exactly how it is spelled. • Wildcards are special characters that can stand in for unknown characters in a text value and are handy for locating multiple items with similar, but not identical data. • Retrieval models do not directly support this query type. • In IR systems, certain kinds of wildcard search support may be implemented. Example: usually words ending with trailing characters.
Natural Language Queries	• There are only a few natural language search engines that aim to understand the structure and meaning of queries written in natural language text, generally as questions or narratives. • The system tries to formulate answers for these queries from retrieved results. • Semantic models can provide support for this query type.

Note - inquiry vs enquiry vs query

Inquiry is chiefly the North American spelling, the British spelling is enquiry. The plural forms are inquiries and enquiries. The word inquiry comes from the Anglo-French word enqueren, meaning to inquire. Remember, query may be used as a noun or a verb, inquiry is only used as a noun.

Refrences

- https://www.techopedia.com/definition/5736/query
- https://www.geeksforgeeks.org/types-of-queries-in-ir-systems/

- https://support.microsoft.com/en-us/office/introduction-to-queries-a9739a09-d3ff-4f36-8ac3-5760249fb65c#:~:text=A%20query%20can%20either%20be,delete%20data%20from%20a%20database.

What is Document?

Definition

A unit of information that can be searched and retrieved by an IR system. This can include web pages, emails, books, articles, and more.

What is Corpus?

Definition

In natural language processing, a corpus contains text and speech data that can be used to train AI and machine learning systems. A corpus can be made up of everything from newspapers, novels, recipes, and radio broadcasts to television shows, movies, and tweets.

Features of a good corpus & Challenges

Features of a good corpus	Challenges
Large corpus size	Deciding the type of data needed to solve the problem statement
High-quality data	Availability of data
Clean data	Quality of the data
Balance	Adequacy of the data in terms of the amount

Annotation Methods in NLP

- *Gold standard annotation(GSC)* is the process of manually annotating a dataset by multiple human annotators, typically experts in the domain, who work independently to identify and label relevant information. The goal is to achieve high inter-annotator agreement and produce a reliable, high-quality dataset for training and evaluation.
- *Silver standard annotation(SSC)*, on the other hand, involves using automated methods to annotate data, such as using heuristics, rules, or statistical models. While this approach is less accurate than gold standard annotation, it can be

useful when large amounts of data need to be annotated quickly and cost-effectively.

Note - Corpora vs Datasets

- A corpus is a representative sample of actual language production within a meaningful context and with a general purpose.
  A large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based.
- A dataset is a representative sample of a specific linguistic phenomenon in a restricted context and with annotations that relate to a specific research question.
  A collection of any kind of data is a dataset.

	Prototypical corpus	**Prototypical dataset**
Language	Unrestricted production	Specific phenomenon
Context	Wide	Restricted
Purpose	General	Research question

References

- Filannino, M. and Di Bari, M., 2015. Gold standard vs. silver standard: the case of dependency parsing for Italian. CLiC it, p.141.
- Wissler, L., Almashraee, M., Díaz, D.M. and Paschke, A., 2014. The Gold Standard in Corpus Annotation. IEEE GSC, 21.
- Ménard, P.A. and Mougeot, A., 2019, September. Turning silver into gold: error-focused corpus reannotation with active learning. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 758-767).
- https://www.hypersenseai.com/aiglossary/corpus/#:~:text=A%20corpus%20can%20be%20made,AI%20and%20machine%20learning%20systems
- https://stats.stackexchange.com/questions/85930/difference-in-meaning-of-these-terms-dataset-vs-corpus#:~:text=In%20contrast%2C%20dataset%20appears%20in,of%20data%20is%20a%20dataset.&text=%22Corpus%20is%20a%20large%20collection,a%20linguistic%20analysis%20is%20based.%20%22
- https://corpuslinguisticmethods.wordpress.com/2013/12/28/corpora-versus-datasets/

What is Benchmark?

Definition

A benchmark refers to a standard point of reference against which things can be compared.

A benchmark as it is used in ML or NLP typically has several components:

- One or multiple datasets
- One or multiple associated metrics
- A way to aggregate performance (Collect and analyze data on different aspects of performance and create an overall assessment or evaluation)

A benchmark sets a standard for assessing the performance of different systems that is agreed upon by the community.

Metrics

Accuracy or F-score: In fine-grained sentiment analysis, it may mix up very positive and very negative → Errors may be costly.

MLPerf: MLPerf is a benchmarking organization and covers a range of tasks, such as image classification, object detection, language translation, and recommendation systems, among others. The metrics used are typically related to the [1]time required to perform a given task, the [2]energy consumption of the system, or the [3]accuracy of the results obtained.

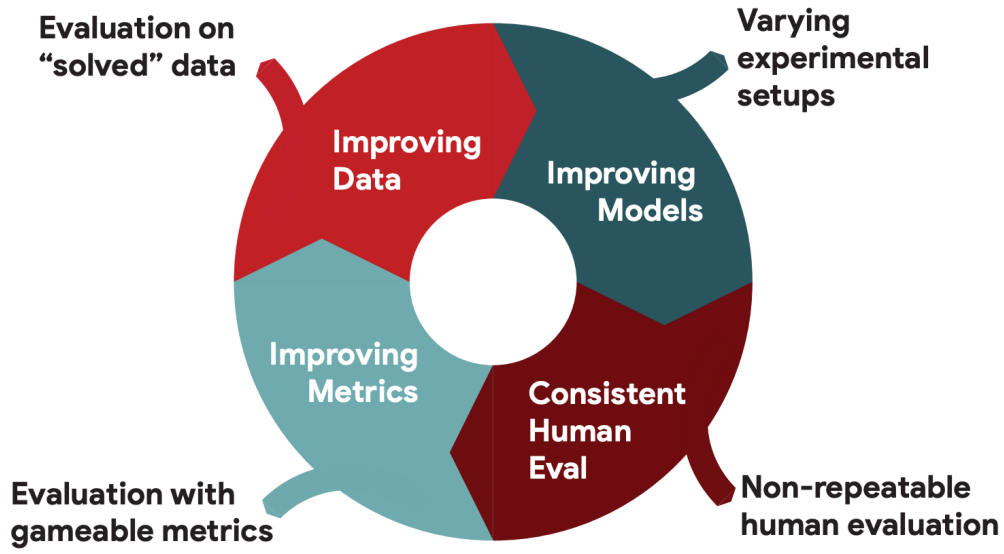
The percentage of correctly transcribed words: In ASR (Automatic Speech Recognition, which is a technology that enables computers to transcribe spoken language into text), only the percentage of correctly transcribed words (akin to accuracy) was initially used as the metric.

Word error rate (WER): It measures the difference between the original speech and the transcribed text by calculating the number of errors or mismatches between them.

(substitutions + deletions + insertions) / number of words in reference
number of words in the reference = substitutions + deletions + correct words

Automatic metrics: A recent trend in natural language generation (NLG) is towards the development of automatic metrics such as BERTScore that compare a reference sentence with a generated sentence and considers the subtle differences in the language and meaning of the sentences, showing how much the two compared sentences overlap.

→ It is important to update and refine metrics, in order to replace efficient simplified metrics with application-specific ones



References

- https://www.ruder.io/nlp-benchmarking/#what-is-a-benchmark

What is Indexing?

Definition

A data structure that enables fast searching of documents based on their content. An index typically includes a list of terms and the documents that contain those terms.

Grep

Allows searching one or two files for lines that contain a pattern in a small amount of data very efficiently.

Drawbacks

- It is unable to find the document that has the highest frequency of query words appearing.
- The performance of the command becomes sluggish or delayed when processing or handling a large volume of data.

Document Term Matrix

DTM is a mathematical representation of text data in the form of a matrix, where each row represents a document and each column represents a term. The values in the matrix denote the frequency of occurrence of each term in each document.

	Word1	Word2
Doc1	0	2
Doc2	1	0
Doc3	3	1

Drawbacks

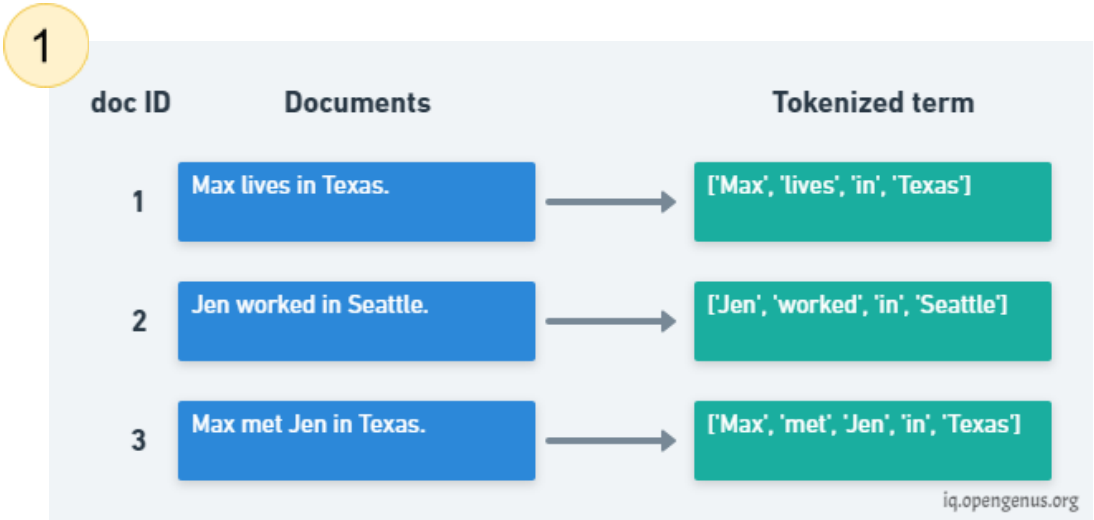
- The size of the document-term matrix grows very quickly and It ends up occupying a lot of space.

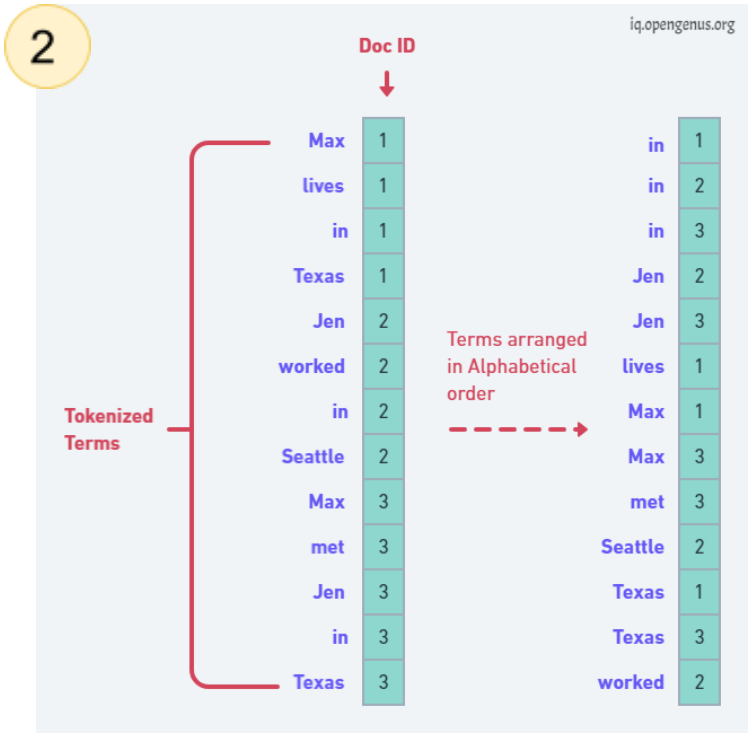
Inverted index

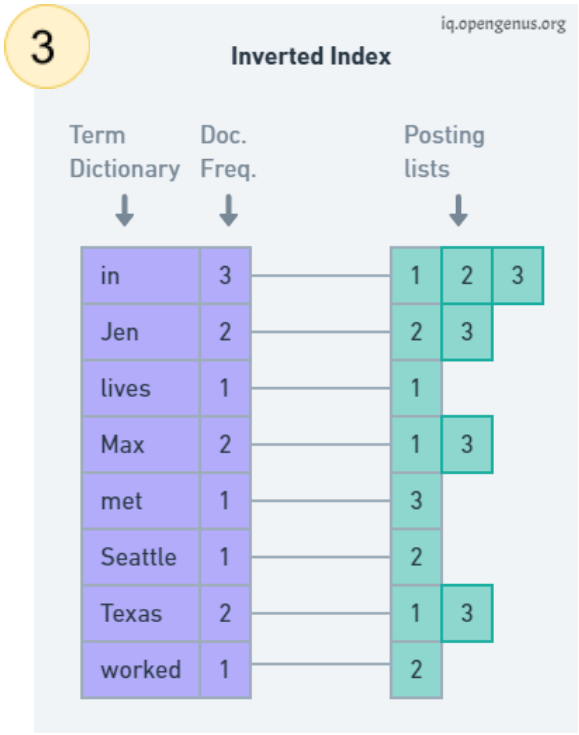
In an inverted index, the index is organized by terms (words), and each term points to a list of documents or web pages that contain that term.

Steps to building an inverted index

- Fetch the Document
- Removing Stop Words: Common words that are often removed from queries and document text, as they are not useful for information retrieval (e.g. "the", "and", "of").
- Stemming of Root Word: The process of reducing words to their base or root form to increase the recall of a search result.
- Record Document IDs: Either add the relevant document ID or create a new list. Also, it is possible to add additional information such as the frequency and location of the word







The major steps in building inverted index are

- Collection of documents that need to be indexed
- Tokenization of the text and converting each text document into a list of tokens
- Linguistic preprocessing of the data, making a list of normalized tokens that are the indexing terms.
- Indexing the documents and sorting the list alphabetically.

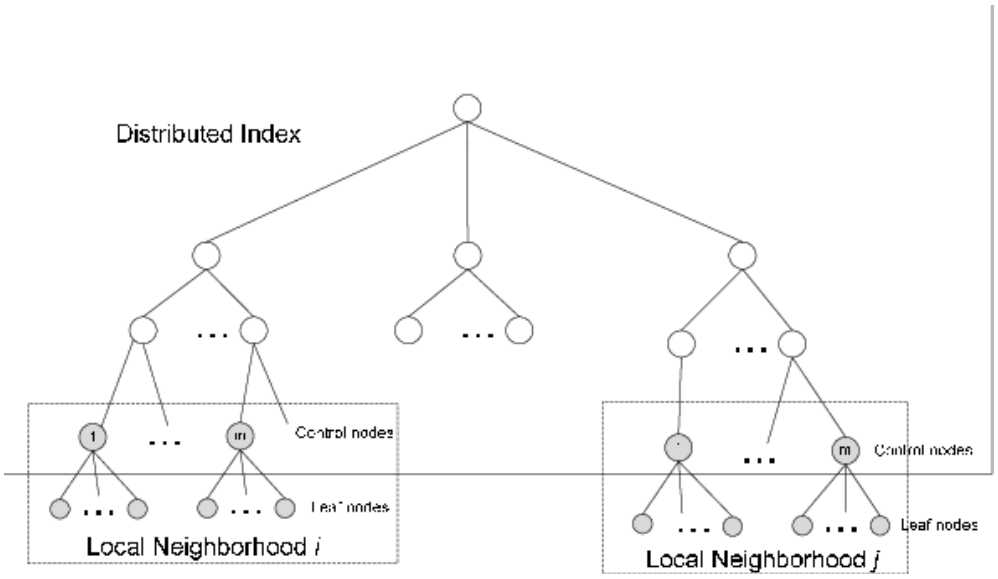
Note - Static Indexing

The Document-term matrix and the inverted index both work on the condition that the corpus is static and the data fits into the hard disk of a single machine.

Distributed Indexing

Web search engines generally use this for index construction. This method is used for index construction in large computer clusters. The general idea of the cluster is to perform the tasks on general computing machines instead of supercomputers.

A master node assigns the task to each computer or node in the cluster. The computing job is distributed in chunks to complete the tasks in a short duration of time. First of all, the input data is split into n splits and the size of each split is chosen such that the work can be done evenly and efficiently. When a machine completes its work on one split then a new split is assigned. If the machine dies before completion then the split is assigned to another machine.



Dynamic Indexing

Most of the data collections in the real world get updated with new data and documents being added or old ones being removed from it. Therefore new terms need to be added and posting lists need to be updated with existing terms.

One way of doing this periodically reconstructing the index from scratch. This is feasible only when the number of changes over time is small and the delay in making searchable documents is acceptable.

If we need the documents to be included quickly, one of the best ways is to maintain two indexes, one for the main index and the other for the auxiliary index for new documents. Every time the Auxillary index becomes very large we merge it with the main index.

Although the time taken by this method is less than reconstructing it from scratch the process is very complex therefore many large search engines prefer reconstructing from scratch.

References

https://www.analyticsvidhya.com/blog/2021/07/indexing-in-natural-language-processing-for-information-retrieval/

What is Retrieval?

Definition

The process of selecting and retrieving relevant documents from a collection based on a user's query.

What is Ranking?

Definition

The process of ordering retrieved documents based on their relevance to the user's query.

IR Metrics(Definition- Success(hit) at K- MRR- NDCG - MAP - F1 score-

Translation metrics: rouge - BLEU - exact match)

What is Precision?

Definition

A measure of the accuracy of a search result, typically defined as the number of relevant documents retrieved divided by the total number of documents retrieved.

Formula

Precision = TruePositives / (TruePositives + FalsePositives) = TP / (TP+ FP)

What is Recall?

Definition

A measure of the completeness of a search result, typically defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection.

Formula

Precision = TruePositives / (TruePositives + FalseNegative) = TP / (TP+ FN)

What is Relevance?

Definition

The degree to which a document is related to a user's information need or query.

Results in any given results list can be listed in order of "Relevance". Relevance in Dimensions is defined by searching for the keyword(s) in the whole corpus of data, and then assigning relative scores to results that match those keywords.
This score is based on a group of criteria:

- How many times the keywords are found, as a proportion of the total text,
- Whether they are found in the title or abstract, or both.

The higher the score, the more "relevant" the document will be in the results list.

References

- https://dimensions.freshdesk.com/support/solutions/articles/23000022475-what-is-relevance-and-how-is-it-calculated-#:~:text=Results%20in%20any%20given%20results,a%20match%20for%20those%20keywords