

ASSIGNMENT 1: Information Retrieval & Indexing

Prepared By
Prethika Acharjee

Roll No: 002500802005

Paper Name: Information Retrieval-II

Paper Code: ML-10

Session: 2025-2026

Under the supervision of

Dr. Avik Roy

Librarian of Derozio Memorial College

Guest Faculty, DLIS, Jadavpur University

and WBSU, Barasat

1. What is Automatic Indexing?

Automatic indexing is the process of generating index terms for documents using computer-based systems, without direct human intervention. Instead of relying on professional indexers to read and analyze a document, automatic indexing employs algorithms to examine the text and identify significant words, phrases, or concepts that accurately represent its content. These systems typically use statistical techniques, such as term frequency analysis and TF-IDF weighting, along with natural language processing, to determine which terms are most relevant. The selected terms are then assigned as index entries, allowing the document to be effectively organized and retrieved within an information retrieval system.

Automatic indexing has become increasingly important in the digital era, where the volume of electronic documents far exceeds the capacity of manual indexing processes. As the number of documents exponentially increases with the proliferation of the Internet, automatic indexing has become essential to maintaining the ability to find relevant information in a sea of irrelevant data. Natural language systems are used to train a system based on seven different methods to help process this information: Morphological, Lexical, Syntactic, Numerical, Phraseological, Semantic, and Pragmatic methods.

2. How is Automatic Indexing Related to Information Retrieval?

Automatic indexing is foundational to information retrieval (IR). Automatic indexing operates within an information retrieval system by systematically analyzing document content and converting it into searchable index terms that can be matched to user queries. Without a reliable index, an IR system cannot efficiently locate relevant documents within a large collection.

In the face of the ever-increasing document volume, libraries around the globe are more and more exploring automated approaches to subject indexing. This helps sustain bibliographic objectives, enrich metadata, and establish more connections across documents from various collections, effectively leading to improved information retrieval and access. Variations of TF-IDF weighting schemes, generated through automatic indexing, are often used by search engines in scoring and ranking a document's relevance given a query. In university digital repositories, indexing allows students to search through thousands of academic papers in seconds rather than manually examining each document.

3. What is IDF (Inverse Document Frequency)?

IDF stands for Inverse Document Frequency. It is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. IDF reduces the weight of terms that appear commonly across many documents and increases the weight of terms that appear rarely, making those rarer terms stronger discriminators.

Mathematically, IDF is calculated as the logarithm of the total number of documents divided by the number of documents containing the term. For lower values of document frequency, the IDF is very high, suggesting a good discriminator. As document frequency increases, IDF smoothly descends and reaches zero when the term appears in every document in the corpus. IDF is always combined with TF (Term Frequency) to produce the TF-IDF score, which is one of the most popular measures used in information retrieval and text mining.

4. What is DF (Document Frequency)?

Document Frequency (DF) is the number of documents in a corpus that contain a specific term. It is a raw count of how widely a term is distributed across a document collection. A term with a high document frequency appears in many documents and is generally considered a poor discriminator (e.g., common words like 'the', 'and', 'is'). A term with a low document frequency appears in few documents and is considered a strong indicator of specific topic content.

DF is directly used in the calculation of IDF. The relationship is inverse: as DF increases for a term, its IDF value decreases, thereby reducing its weight in TF-IDF scoring. DF is used in term selection approaches such as TF-DF and TF-IDF, which are applied during automatic indexing to reduce attributes and find effective term selection methods for better clustering accuracy and retrieval precision.

5. Steps of Automatic Indexing Techniques

Automatic indexing follows a systematic process comprising the following steps:

Document Collection: Gathering the documents that need to be indexed. This is the starting point, where the corpus or document set is assembled.

Text Preprocessing: Cleaning the text by removing punctuation, converting to lowercase, and eliminating stop words (common words like 'the,' 'and,' 'is') that do not contribute meaningful information.

Stemming and Lemmatization: Reducing words to their root or base form using algorithms such as the Porter Stemmer, so that variations of the same word are treated as one term.

Tokenization: Breaking down the text into individual terms or tokens so that each unit of meaning can be processed separately.

Term Selection: Identifying significant terms that best represent the document's content using statistical approaches such as TF-IDF, TF-Df, or frequency-based thresholds.

Thesaurus / Vocabulary Control Application: Using tools such as WordNet thesaurus to maintain relationships between important terms and standardize vocabulary to improve retrieval consistency.

Index Creation: Building a data structure (typically an inverted index) that maps terms to their locations in documents, enabling efficient query matching.

Weight Assignment: Assigning weights to selected terms using methods like TF-IDF to reflect the relative importance of each term within a document and across the collection.

References

Anderson, J. D. (1997). Guidelines for indexes and related information retrieval devices. NISO Press.

Birger, L. (2004). References and citations in automatic indexing and retrieval systems: Experiments with the boomerang effect. Department of Information Studies, Royal School of Library and Information Sciences. <https://www.researchgate.net/publication/289520637>

Hlava, M. (2011). The taxobook: Principles and practices of building taxonomies. Morgan & Claypool.

LIS Education Network. (2026, February 14). Automatic indexing: Definition, methods, and applications. Library & Information Science Education Network. <https://www.lisedunetwork.com/automatic-indexing/>

LIS Academy. (2025, November 9). How indexing and information representation drive information retrieval. LIS Academy. <https://lis.academy/information-processing-retrieval/how-indexing-information-representation-retrieval/>

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. McGraw-Hill.

Savoy, J. (2010). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, 60(1), 1–29. <https://www.tandfonline.com/doi/full/10.1080/01639374.2021.2012311>

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>