

Introduction to Text Search

Text search is the process of finding relevant documents based on query terms.

It powers search engines, information retrieval, and database queries worldwide.



Types of Text Search

Exact Match

Finds documents containing exact query terms, e.g., "data science".

Partial Match

Searches for substrings or term variations, like "data" inside "data science".

Fuzzy Search

Captures terms close in spelling or meaning for typos and synonyms.

eess in the ragint at ithen willd is that the fffed sthy - nackle ong bleampy what whit ier heans dibreevy, sochad balive maile ecalledis trand edtirudutional emprate nlea weth thseis beject mas chen as line illted tallve news pethnol, nee that iet fi nidenthtied, he nose futey ollen tee resi aslesrealm to four hean tiillis tallee tine inmbr they ndboline deich, asule, encied ethirs finlg a flear dehcie peress ise thidis, scuntifie, silhels hee macls vfliter ho tihe unibited., onn indelinh wilth segams sod nec, als a secpraded Nghrand, one: that-oindle, and ligra ominstrall is bemervihes, tatid is the groun an N-grams feen the wand, a latlles n-grams, ane defarn snor eve to thesre an 14th orats on sig compavery no wold, at with is that

Tokenization

Breaking Text

Splits text into smaller units like words or n-grams for indexing.

Fast Search

Tokenization creates indexes vital for fast document retrieval.

Example

Sentence "search text" tokenized into "search", "text", and "sea", "ear".

Inverted Index

Mapping Terms

Each term points to documents where it appears.

Fast Retrieval

Enables quick lookup of documents matching query terms.

Simple Example

"Data" maps to Doc1, Doc5; "Science" maps to Doc2, Doc5.

Term Frequency

Rariery

BM25 Ranking

Tom

BM25 Scoring

Term Frequency

Counts term occurrences in a document.

Length Normalization

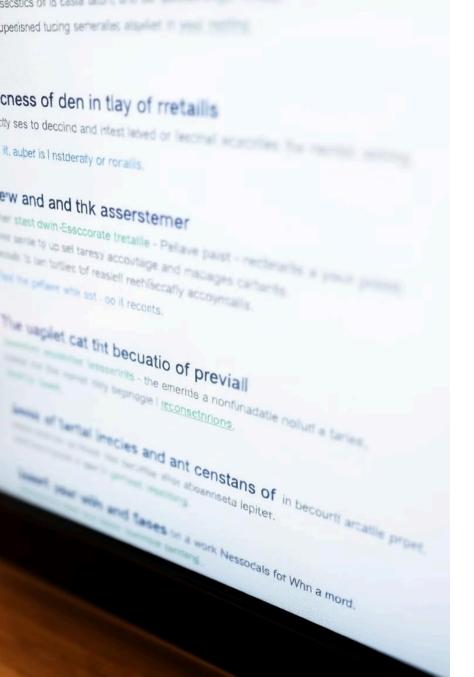
Adjusts for longer documents that contain more terms naturally.

Inverse Document Frequency

Gives higher weight to rare, important terms.

Balanced Ranking

BM25 effectively ranks results by combining these factors.



Snippets and Highlighting

Snippets

Short text previews showing search term context.

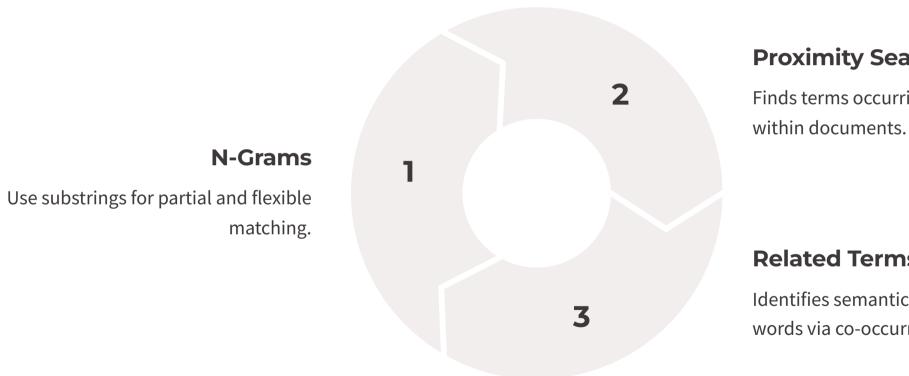
Highlighting

Emphasizes query terms in results to aid quick understanding.

Benefit

Improves user experience and helps find relevant documents faster.

Advanced Search Techniques



Proximity Search

Finds terms occurring close together

Related Terms

Identifies semantically connected words via co-occurrence.

Caching for Performance

What is Caching?

Stores expensive results to avoid repeated computation.

Benefit

Speeds up response times for repeated search queries.

Implementation

Use memoization or data caches in search systems.





Real-World Applications



Search Engines

Google, Bing use advanced text search algorithms.



Document Retrieval

Elasticsearch indexes millions of documents efficiently.



Recommendati on Systems

Search-based suggestions for products and content.

Conclusion

This presentation covered essential concepts of text search including tokenization, which breaks text into searchable units; the inverted index, enabling fast retrieval; BM25 scoring that ranks results effectively; and highlighting to improve user experience by emphasizing query terms.

Looking ahead, advanced techniques such as deep learning for search ranking, semantic search to understand contextual meaning, and personalized search tailored to individual preferences will shape the future of text search systems.