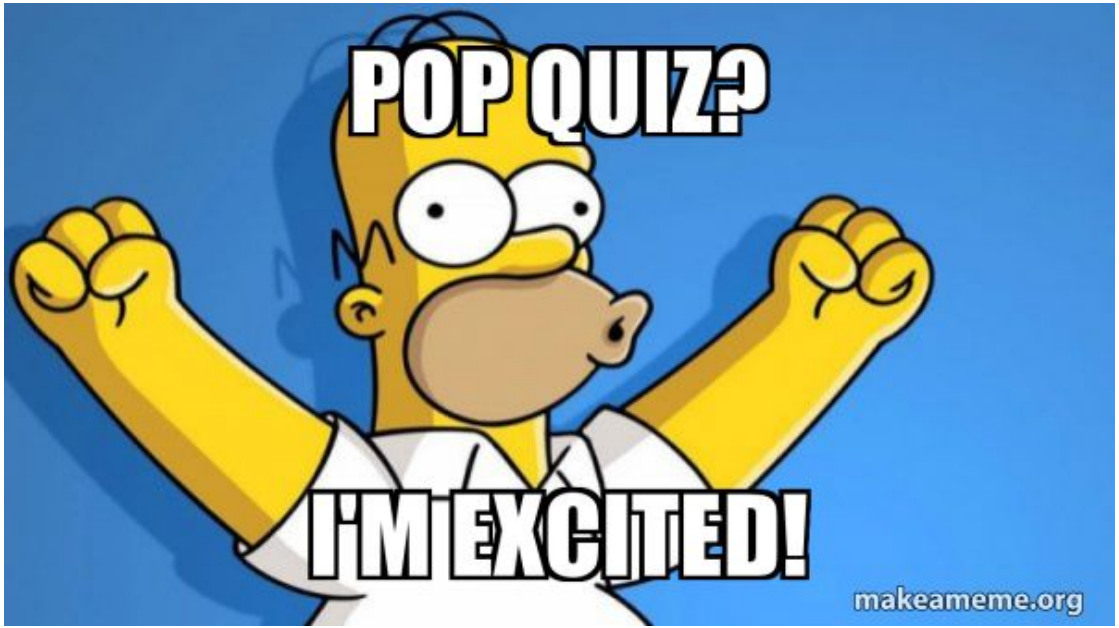
Natural Language Processing - Search

Natural Language Processing
Search

Unit 3.

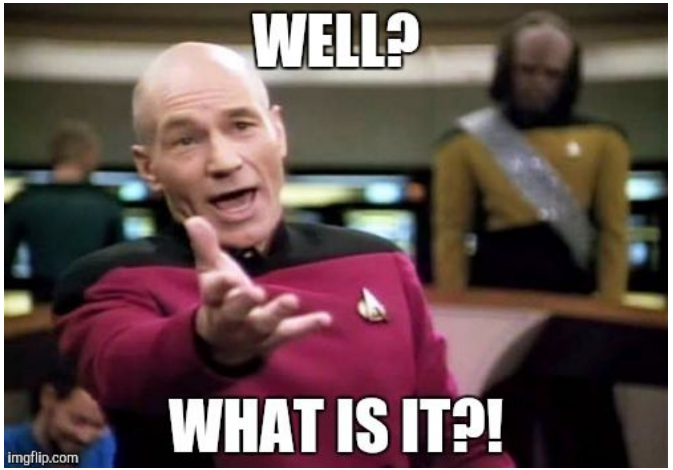
Search

Natural Language Processing - Search



Natural Language Processing - Search

- Content
- Context
- Part of Speech Tagging
- Word sense disambiguation
- Syntactic parsing
- Conference resolution
- Named Entity Recognition (NER)



Natural Language Processing - Search

Information Retrieval

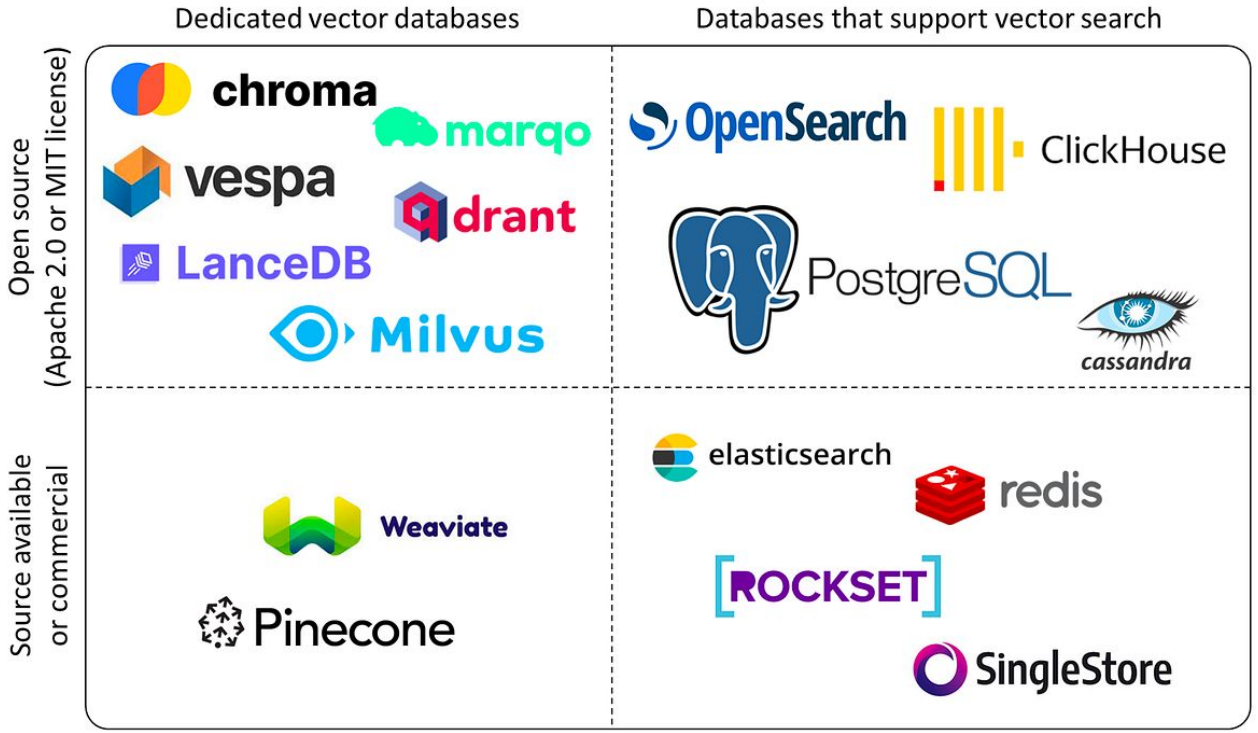
elasticsearch

Information Retrieval

elasticsearch

DuckDuckGo

Information Retrieval



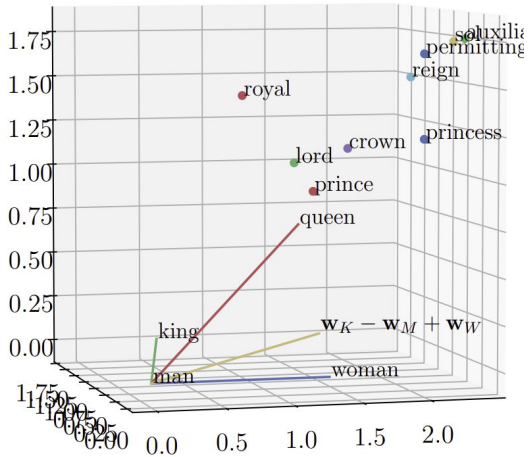
Information Retrieval

How do we search ?

Information Retrieval

How do we search ?

- Words exist in space
- Sentences are made of words
- Sentences exist in space
- Use distances to search



Information Retrieval

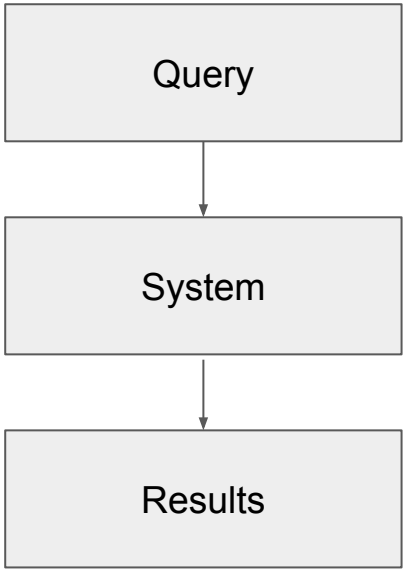
“In NLP and IR, "search" involves finding relevant information from a collection of documents based on a user's natural language query. This process includes tokenization, indexing, matching, ranking, and presentation of search results to help users locate the desired information efficiently.” - ChatGPT

Information Retrieval

“In NLP and IR, "search" involves **finding relevant information** from **a collection of documents** based on a **user's natural language query**. This process includes tokenization, indexing, matching, ranking, and presentation of search results to help users locate the desired information efficiently.” - ChatGPT

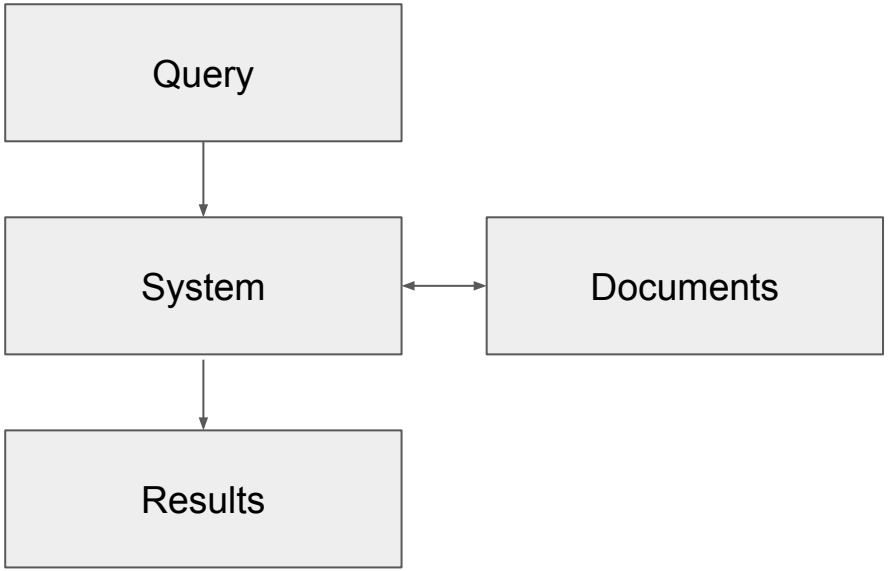
Information Retrieval

How do we search ?



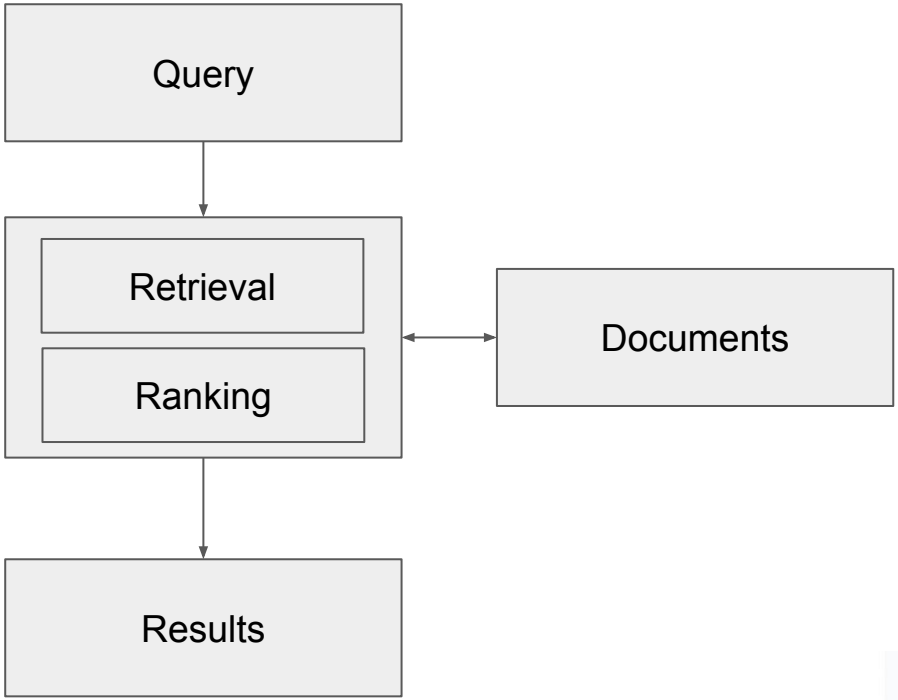
Information Retrieval

How do we search ?



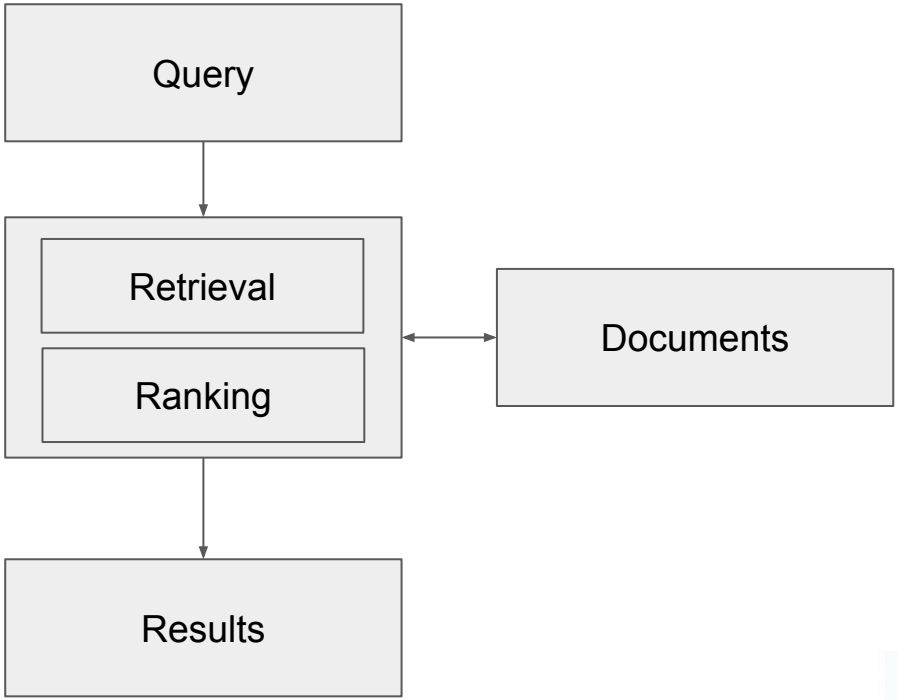
Information Retrieval

How do we search ?



Information Retrieval

How do we search ?



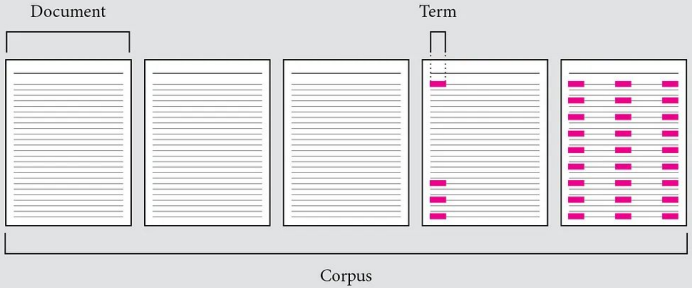
Natural Language Processing - Search

Relevant concepts

Natural Language Processing - Search

Relevant concepts

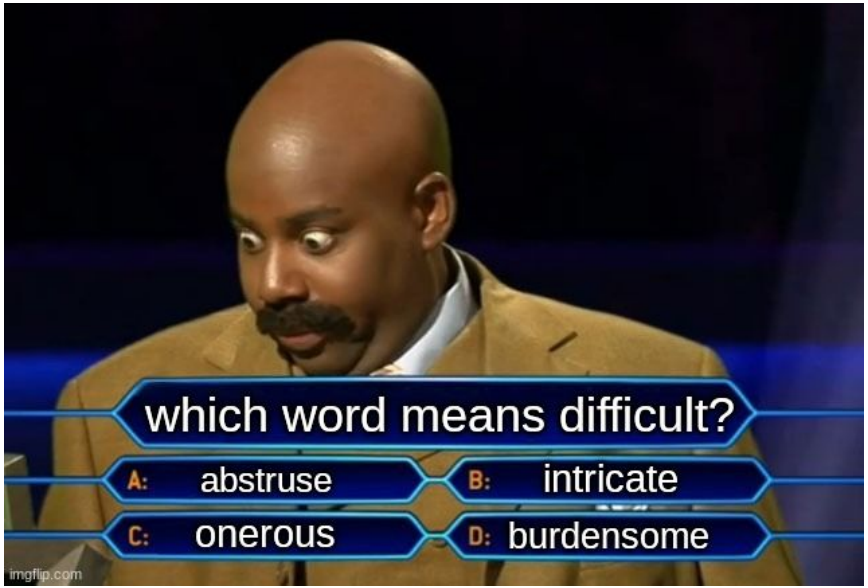
- Corpus



Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary

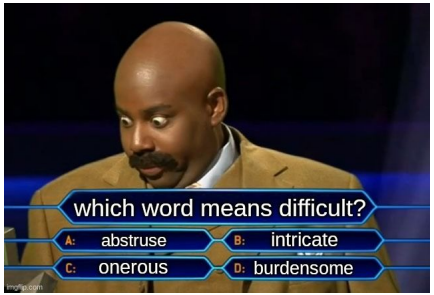


Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary

Vocabulary is hard.

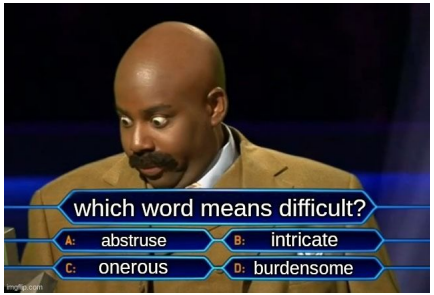


Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary

We need to have an explicit defined vocabulary.
A vocabulary is a set of terms (or tokens) known to the system.



Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
-

Natural Language Processing - Search

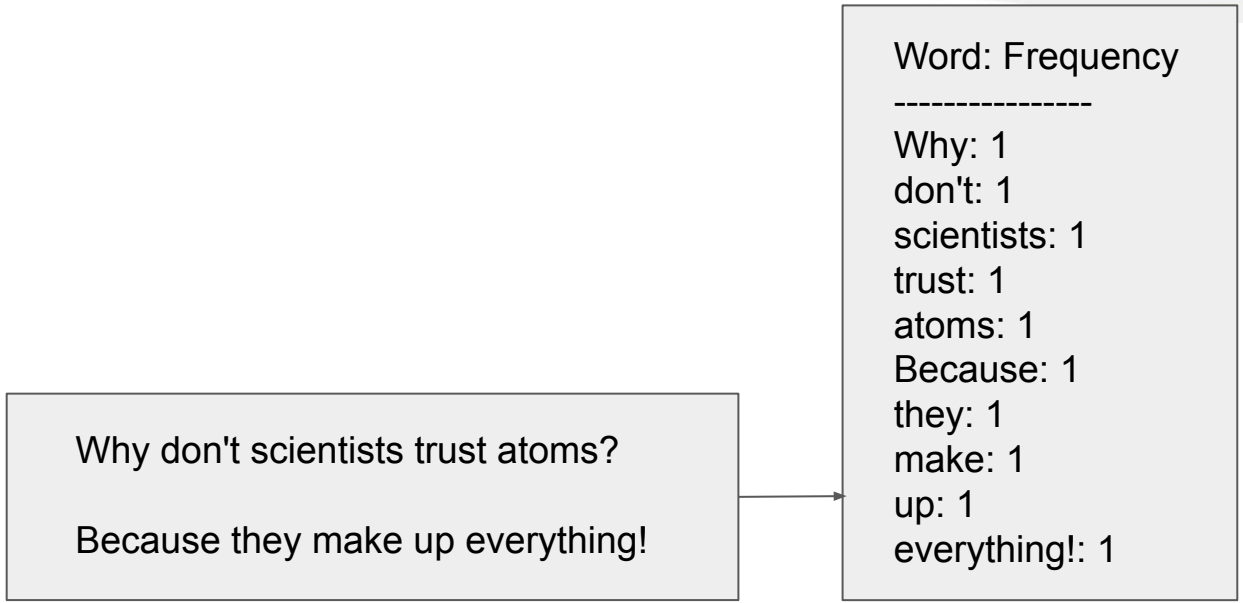
Relevant concepts

- Corpus
- Vocabulary
- Bag of words

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words



Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures

Natural Language Processing - Search

Relevant concepts

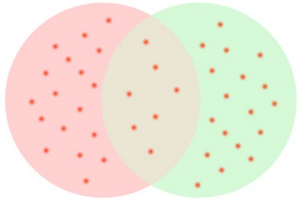
- Corpus
- Vocabulary
- Bag of words
- Similarity Measures

- Cosine Similarity
- Euclidean Distance
- Jaccard Similarity

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

total elements in intersection

total elements in union i.e. Universal Set

$J(A,B)$ is thus probability of picking a random element from the universal set and finding that it is present in both the participating sets

similar to chances that you throw a dart and it hits the intersection

Jaccard Similarity Coefficient as Probability

Natural Language Processing - Search

Relevant concepts

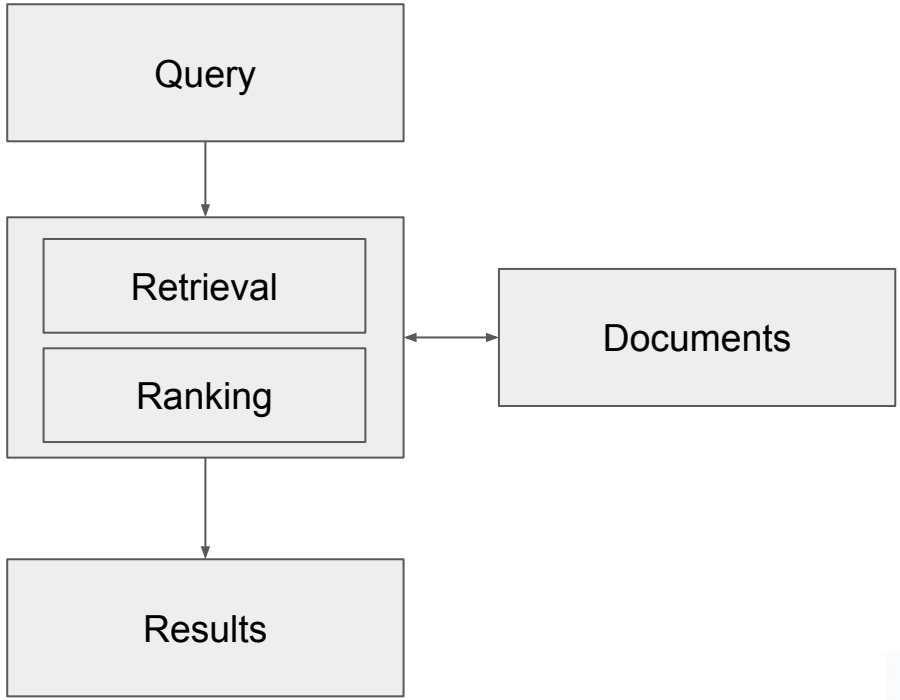
- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing

"... refers to the process of creating a searchable index or catalog of data"

Natural Language Processing - Search

Relevant concepts

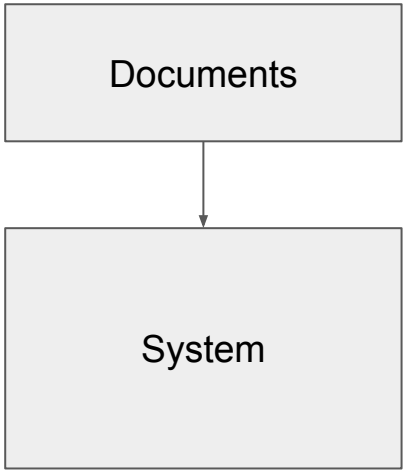
- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing



Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing



Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

Frequency of term t in document d

Total number of terms

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

$$idf\,(t, D) = log\,(\frac{N}{count\,(d \in D{:}t \in d)})$$

Total number of documents

Number of documents d with term t.

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

$$idf\,(t, D) = log\,(\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

$$tf\,idf\,(t, d, D) = tf\,(t, d)\ .\ idf\,(t, D)$$

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

Evaluates the importance of a term (word) within a document relative to a collection of documents (a corpus)
I.e relative bag of words

Natural Language Processing - Search

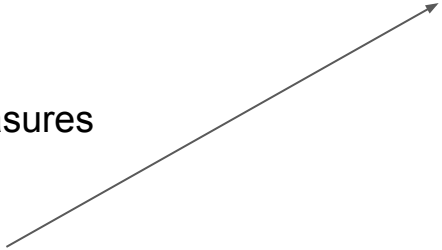
Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25



Best Matching 25

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25

Best Matching 25

… is a ranking function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity).

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25

Best Matching 25

… is a ranking function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity).

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

k1 and b are free parameters

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25

Documents

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

average document length

Query

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

is the number of times that q_i occurs in the document D

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25
- Vector Space Model

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25
- Vector Space Model

“Vector space model or term vector model is an algebraic model for representing text documents as vectors of identifiers (such as index terms)” - Wikipedia

Natural Language Processing - Search

Relevant concepts

- Corpus
- Vocabulary
- Bag of words
- Similarity Measures
- Indexing
- TF-IDF
- Okapi BM25
- Vector Space Model