Glossary

Information Retrieval

Terms and Definitions

A

Average Precision at K (AP@K): the sum of the precision at K of the values of K divided by the total number of relevant items in the top K results.

Advanced Search in google: Google advanced search operators are special commands and characters that filter search results. They do this by making your searches more precise and focused.

augmented matrix: is a matrix obtained by appending the columns of two given matrices, usually for the purpose of performing the same elementary row operations on each of the given matrices.

B

Boolean Retrieval: Queries are joined using AND, OR, NOT, etc. A document can be visualized as a keyword set. Based on the query a document is retrieved based on relevance.

bilingual search engine: it provides multiple language search pages when users type keyword in one language. This type of search engine gives more detailed information for people who speak more than one language.

C

Click-Through Rate (CTR): the number of clicks that your ad receives divided by the number of times your ad is shown, for example, if you had 5 clicks and 100 impressions, then your CTR would be 5%.

Cross-Validation: a resampling method that uses different portions of the data to test and train a model on different iterations.

Cross-Language Information Retrieval (CLIR): is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query.

Collaborative Filtering: a technique that can filter out items that a user might like on the basis of reactions by similar users.

D

Document: a record of some information that can be used as an authority or for reference, further analysis, or study.

Document Frequency (DF): the number of documents containing a particular term

Document Ranking: It presents retrieved documents in an order of their estimated degrees of relevance to query.

Discounted Cumulative Gain (DCG): It is based on non-binary relevance assessments of documents ranked in a retrieval result. It assumes that, for a searcher, highly relevant documents are more valuable than marginally relevant documents.

Document Preprocessing: Document Preprocessing applies a common sequence of preprocessing steps to clean and prepare text for subsequent analysis and comparison with other text.

E

Evaluation Metrics: used to measure the quality of the statistical or machine learning model

Elastic Search: it uses an inverted indexing method to store documents and texts along with their frequencies

F

F1-Score: a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset

Field-Based Retrieval: assigns a score or Retrieval Status Value (RSV) to a document D and a query Q by distinguishing the occurrences of query terms in the different field vectors, and by weighting the contribution of each field appropriately.

G

Geographic Information Retrieval (GIR): it is a specialized branch of traditional Information Retrieval (IR), which deals with the information related to geographic locations. One of the main challenges of GIR is to quantify the spatial relevance of documents and generate a pertinent ranking of the results according to the spatial information needs of user.

H

High-dimensional data: data in which the number of features (variables observed), p, are close to or larger than the number of observations (or data points), n.

I

Information Retrieval (IR): the process of retrieving information (usually unstructured text in documents) to satisfy user's information need.

Index: a systematic guide designed to indicate topics or features of documents in order to facilitate retrieval of documents or parts of documents.

Intra-document Similarity: Is the critical information which determines whether or not the cluster-based retrieval improves the baseline

Inverted Index: an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents.

Inverse Document Frequency (IDF): a weight indicating how commonly a word is used. The more frequent its usage across documents, the lower its score. The lower the score, the less important the word becomes.

Information Extraction: the automated retrieval of specific information related to a selected topic from a body or bodies of text.

J

Jaccard Similarity: it is a common proximity measurement used to compute the similarity between two objects, such as two text documents. Jaccard similarity can be used to find the similarity between two asymmetric binary vectors or to find the similarity between two sets.

K

K-Means Clustering: is used in information retrieval to enhance the retrieving process. It is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.

L

Latent Semantic Indexing (LSI): a method of analyzing a set of documents in order to discover statistical co-occurrences of words that appear together which then give insights into the topics of those words and documents.

Latent Dirichlet Allocation (LDA): A generative probabilistic model of a corpus.

linear algebra: Mathematical discipline that deals with vectors and matrices

M

Mean Average Precision (MAP): a popular metric used to measure the performance of models doing document/information retrieval and object detection tasks.

N

Normalized Discounted Cumulative Gain (NDCG): NDCG is calculated by dividing the discounted cumulative gain (DCG) of the ranked list by the DCG of the ideal ranked list, which is the list with the relevant items ranked in the most optimal order. NDCG ranges from 0 to 1, with higher values indicating better performance.

Named Entity Recognition (NER): a natural language processing (NLP) method that extracts information from text. NER involves detecting and categorizing important information in text known as named entities.

P

Precision: the ratio of the number of relevant and retrieved documents to the number of total retrieved documents from the query.

Precision at K (P@K): the proportion of recommended items in the top-k set that are relevant.

Precision-Recall (PR) Curve: A precision-recall curve can be calculated in scikit-learn using the precision_recall_curve () function that takes the class labels and predicted probabilities for the minority class and returns the precision, recall, and thresholds.

Precision-Recall Gain: plots Precision Gain on the y-axis against Recall Gain on the x-axis in the unit square

Precision-Recall Break-Even Point (BEP): it is an evaluation measure originally introduced in the field of information retrieval to evaluate retrieval systems that return a list of documents ordered by their supposed relevance to the user's information need

Q

Query: user's free text to express desires

Query Expansion: is a process in Information Retrieval which consists of selecting and adding terms to the user's query with the goal of minimizing query-document mismatch and thereby improving retrieval performance.

Query Performance Prediction: to predict the retrieval quality of a search system for a query without relevance judgments.

Query Intent: the identification and categorization of what a user online intended or wanted to find when they typed their search terms into an online web search engine for the purpose of search engine optimization or conversion rate optimization.

Query Log: Each data source has a **query log** that lists all the running and completed queries that were run against it in the last 30 days.

Query Log Analysis: is a tool in the Azure portal to edit and run log queries from data collected by Azure Monitor logs and interactively analyze their results.

R

Ranked Retrieval: appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each such set, precision and recall values can be plotted to give a precision-recall curve.

Ranking Algorithm: a procedure that ranks items in a dataset according to some criterion. Ranking algorithms are used in many different applications, such as web search, recommender systems, and machine learning.

Recall-Precision Curve: a plot of the precision (y-axis) and the recall (x-axis) for different thresholds, Relevance Ranking: assigns ranking scores to results based on its predetermined criteria, such as the frequency of a user's query terms in the result text.

Recall: Total number of documents retrieved that are relevant/Total number of relevant documents in the database.

Relevance Feedback: The idea behind relevance feedback is to take the results that are initially returned from a given query, to gather user feedback, and to use information about whether or not those results are relevant to perform a new query.

Relevance Judgment: A standard interface for relevance feedback consists of a list of titles with checkboxes beside the titles that allow the user to mark relevant documents.

Relevance Judgment Scale: The parameters speed, accuracy and error rate require relevance judgements, while satisfaction and condense can be measured with questionnaires

S

Spam Filtering: identify emails that attackers or marketers use to send unwanted or dangerous content.

Sentiment Analysis: the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral

Stemming: the process of reducing a word to its stem that affixes to suffixes and prefixes or to the roots of words known as "lemmas".

Stop Words: words in any language which do not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on.

Spelling Correction: two types of spelling correction, context-sensitive and isolated word. The goal of spelling correction is to find the correction, out of all possible candidate corrections (including the original query), that has the highest probability of being correct.

T

Term Frequency (TF): is the number of times the term appears in a document compared to the total number of words in the document

Term Frequency-Inverse Document Frequency (TF-IDF): a handy algorithm that uses the frequency of words to determine how relevant those words are to a given document.

Text Summarization: is the most challenging task in information retrieval systems. Data reduction helps the user to find required information quickly without wasting time and effort in reading the whole document. Automated information retrieval systems are used to reduce "Information Overload".

Term Weighting: a procedure that takes place during the text indexing process in order to assess the value of each term to the document.

Term Frequency Normalization: handles the case when a long document has higher values of term frequencies just because of the length of the document and it will have same terms repeated again and again.

Text Mining: an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

U

User Experience (UX): is how a user interacts with and experiences a product, system, or service. It includes a person's perceptions of utility, ease of use, and efficiency.

User-Centric Information Retrieval: Design is based upon an explicit understanding of users, tasks, and environments; is driven and refined by user-centered evaluation; and addresses the whole user experience. The process involves users throughout the design and development process, and it is iterative.

V

Vector Space Model (VSM): is an algebraic model for representing text documents as vectors of identifiers

W

Web Crawling: a computer program that's used to search and automatically index website content and other information over the internet. These programs, or bots, are most commonly used to create entries for a search engine index.