UNIT V INFORMATION RETRIEVAL AND WEB SEARCH 9

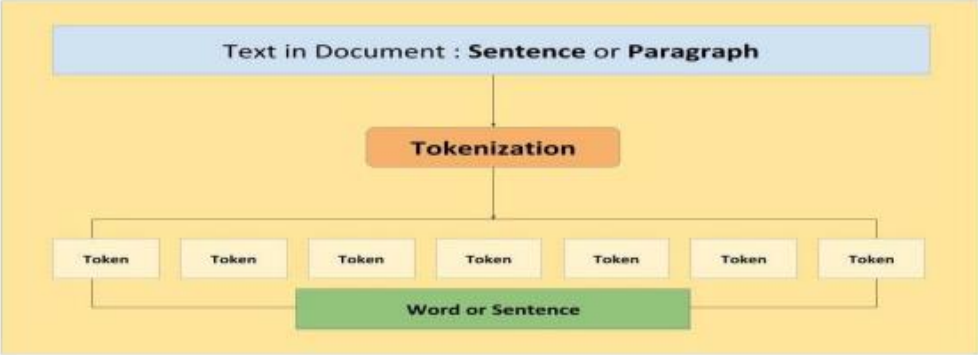
IR concepts – Retrieval Models – Queries in IR system – Text Preprocessing – Inverted Indexing – Evaluation Measures – Web Search and Analytics – Current trends.

TEXT PREPROCESSING

Text preprocessing is an initial phase in text mining. There are various preprocessing techniques to categorize text documents. These are filtering, splitting of sentences, stemming, stop words removal and token frequency count. Filtering has a set of rules for removing duplicate strings and irrelevant textThe various text preprocessing steps are:

1. Tokenization.
2. Lower casing.
3. Stop word removal.
4. Stemming.
5. Lemmatization.

The purpose of tokenization is to protect sensitive data while preserving its business utility. This differs from encryption, where sensitive data is modified and stored with methods that do not allow its continued use for business purposes. If tokenization is like a poker chip, encryption is like a lockbox.



Stemming and Lemmatization are Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing.

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative

computers	compute	computers
feet	feet	foot

Stop words removal

Stop word removal is one of the most commonly used preprocessing steps across different NLP applications. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

The preprocessing of the text data is an essential step as there we prepare the text data ready for the mining. If we do not apply then data would be very inconsistent and could not generate good analytics results.

Text Pre-processing is used to clean up text data: Convert words to their roots (in other words, lemmatize). Filter out unwanted digits, punctuation, and stop words. Some of the common text preprocessing / cleaning steps are:

- Lower casing.
- Removal of Punctuations.
- Removal of Stop words.
- Removal of Frequent words.
- Removal of Rare words.
- Stemming.
- Lemmatization.
- Removal of emojis.

Evaluation measure

Evaluation measures for an information retrieval system are used to assess how well the search results satisfied the user's query intent. The field of information retrieval has used various types of quantitative metrics for this purpose, based on either observed user behavior or on scores from prepared benchmark test sets. Besides benchmarking by using

this type of measure, an evaluation for an information retrieval system should also include a validation of the measures used, i.e. an assessment of how well the measures what they are intended to measure and how well the system fits its intended use case. [1] Metrics are often split into two types: online metrics look at users' interactions with the search system, while offline metrics measure theoretical relevance, in other words how likely each result, or search engine results page (SERP) page as a whole, is to meet the information needs of the user.

Online metrics

Online metrics are generally created from search logs. The metrics are often used to determine the success of an A/B test.

Session abandonment rate

Session abandonment rate is a ratio of search sessions which do not result in a click.

Click-through rate

Click-through rate (CTR) is the ratio of users who click on a specific link to the number of total users who view a page, email, or advertisement. It is commonly used to measure the success of an online advertising campaign for a particular website as well as the effectiveness of email campaigns.[2]

Session success rate

Session success rate measures the ratio of user sessions that lead to a success. Defining "success" is often dependent on context, but for search a successful result is often measured using dwell time as a primary factor along with secondary user interaction, for instance, the user copying the result URL is considered a successful result, as is copy/pasting from the snippet.

Zero result rate

Zero result rate (ZRR) is the ratio of Search Engine Results Pages (SERPs) which returned with zero results. The metric either indicates a recall issue, or that the information being searched for is not in the index.

Offline metrics

Offline metrics are generally created from relevance judgment sessions where the judges score the quality of the search results. Both binary (relevant/non-relevant) and multi-level (e.g., relevance from 0 to 5) scales can be used to score each document returned in response to a query. In practice, queries may be ill-posed, and there may be different shades of relevance.

WEB SEARCH

A web search engine is a specialized computer server that searches for data on the Web. The search results of a user query are restored as a list (known as hits). The hits can

include web pages, images, and different types of files. There are various search engines that also search and return data available in public databases or open directories. Search engines differ from web directories in that web directories are supported by human editors whereas search engines work algorithmically or by a combination of algorithmic and human input.

Web search engines are large data mining applications. There are several data mining techniques are used in all elements of search engines, ranging from crawling (e.g., deciding which pages must be crawled and the crawling frequencies), indexing (e.g., selecting pages to be indexed and determining to which extent the index must be constructed), and searching (e.g., determining how pages must be ranked, which advertisements must be added, and how the search results can be customized or create "context aware").

ANALYTICS

Analytics is the systematic computational analysis of data or statistics.[1] It is used for the discovery, interpretation, and communication of meaningful patterns in data. It also entails applying data patterns toward effective decision-making. It can be valuable in areas rich with recorded information; analytics relies on the simultaneous application of statistics, computer programming, and operations research to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business performance. Specifically, areas within analytics include descriptive analytics, diagnostic analytics, predictive analytics, prescriptive analytics, and cognitive analytics.[2] Analytics may apply to a variety of fields such as marketing, management, finance, online systems, information security, and software services. Since analytics can require extensive computation (see big data), the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics

CURRENT TRENDS IN WEB SEARCH

1. Voice search will become even more relevant

Voice search is already an integral part of our daily lives: we ask Siri where the closest gas station is or say “Hey Google, which Thai restaurant is the highest rated in my town?“ At the moment, optimizing for these kinds of voice searches is recommended especially for ecommerce or websites whose users are likely to have their hands full. For example, if you run a recipe blog, you want your users to find the answer on how long to let the dough rest without having to type with their potentially dirty hands on the phone.

2. Your site search can no longer offer zero results pages

A zero result page for your user means a lost client for you. But what seems like a problem can be a great opportunity to increase your revenue. Let’s go back to our example. In this case, you cannot offer your user Ralph Lauren winter shoes. But you can show them results for other relevant products such as summer shoes by Ralph Lauren or winter shoes by other brands.

3. Search will become more personalized than ever

With personalization, you can offer relevant results for each user based on their preferences and prior search behavior. Going back to our example, an HR person might have already downloaded a pdf targeted towards HR managers on the website. Based on their behavior, they would get assessed as a B2B user and can get more B2B oriented results in their search.

4. Site search will feel less like search and more intuitive

A good site search is the one you do not even think about as a user. You use it so intuitively that you don’t need to assess what you are doing – you just do it. In 2022, site search will look even less like classical search.