Keyword Extraction

Keyword extraction is a natural language processing technique that involves automatically identifying the most important words and phrases in a text document. It is useful in a variety of applications such as information retrieval, document classification, and summarization.

The process of keyword extraction involves several steps. First, the text is preprocessed to remove stop words (such as "the", "and", "a") and to perform stemming (reducing words to their root form). Then, the remaining words are assigned a score based on their frequency and relevance to the overall document. Finally, the top-scoring words are selected as the keywords.

Keyword extraction can be performed using several algorithms such as TF-IDF (term frequency-inverse document frequency), RAKE (Rapid Automatic Keyword Extraction), and TextRank. TF-IDF is a popular method that assigns a score to each word based on its frequency in the document and its rarity in the corpus (collection of documents). RAKE, on the other hand, uses a combination of heuristics such as word frequency, co-occurrence, and degree of word separation to identify candidate keywords. TextRank is a graph-based method that ranks words based on their centrality in a network of co-occurring words.

Keyword extraction is useful in a variety of real-life applications. For example, in e-commerce, keyword extraction can be used to extract product features from customer reviews to help improve product descriptions and customer satisfaction. In legal documents, keyword extraction can be used to identify key topics and arguments, which can help lawyers prepare cases more efficiently.

Overall, keyword extraction is an important technique in natural language processing that helps to automatically identify the most important words and phrases in a text document. It can be performed using various algorithms, and is useful in a wide range of applications.