Abstract of DTI project

RAKE stands for "Rapid Automatic Keyword Extraction." It's a simple and efficient algorithm used for extracting keywords or key phrases from a body of text. RAKE relies on statistical measures like word co-occurrence and word frequency to identify important words and phrases in a text document.

In this algorithm, the text is pre-processed to remove any noise or unnecessary characters like punctuation marks, numbers, and stop-words. The text is split into individual words or tokens. Then, candidate phrases are generated by grouping adjacent words together. For each candidate phrase, two scores are calculated:

Degree: The degree of a candidate phrase is the sum of the number of words it contains.

Frequency: The frequency score of a candidate phrase is the sum of the frequencies of its constituent words.

Finally, a score is assigned to each candidate phrase by dividing its frequency by its degree. The candidate phrases with the highest scores are considered the most significant keywords or key phrases.

This algorithm can be further modified by enhancing the score calculation mechanism. This can be implemented by integrating Rake with text-rank algorithm. For each candidate phrase, two scores are obtained. First will be based on RAKE’s frequency based approach and another will be based on text-rank algorithm’s importance based approach. Both these obtained scores will be combined to obtain a single enhanced score by calculation of harmonic mean using the formula-

Combined Score = 2 * (Importance Score * Frequency Score) / (Importance Score + Frequency Score)

This method ensures that the combined score reflects both high importance and high frequency, rather than being dominated by one or the other.