# Rapid Automatic Keyword Extraction

A keyword is a word or a short phrase that concisely describes the context of a document / larger text. Using them, we are able to fetch relevant documents without having to fully parse them

NOTE: USUALLY KEYWORDS DON'T CONTAIN PUNCTUATION OR STOP WORDS

RAKE IS A FAST AND UNSUPERVISED ALG. THAT AUTOMATICALLY EXTRACT THE KEYWORDS OF A SINGLE DOCUMENT

## INPUT

- set of word delimeters
- set of phrase delimeters
- list of stopwords (stoplist)
- The document to parse

# STEPS

## (1) FIND THE CANDIDATE KEYWORDS :

- Split the document into an array of words, using the specified word delimiters
- Merge the array into sequences of contigous words at phrase delimiters and then at stop word
- The resulting sequences are candidate keywords

## (2) Scoring candidate keywords

- Compute the matrix of co-occurrences M

  ( EACH ENTRY COUNTS THE NUMBER OF CANDIDATE KEYWORDS WHERE BOTH WORDS ARE PRESENT )

- Compute the frequency of each word

  ( $freq(w) = M_{ij}$ where $i == j$ )

  and the degree of each word

  ( $deg(w) = \sum_j M_{ij}$ = sum of freq. over its row )

- Compute the score for each word

  ( $score(w) = \frac{deg(w)}{freq(w)}$ )

- The final score of a keyword is the sum of the score of its words

## (3) Merge keywords

- Identifies keywords that contain stop words
- Looks for pairs of keywords that adjoin one another at least twice in the same document and in the same order
- The final score of the new keywords is the sum of its member keywords final score

## (4) Selecting keywords

Return top one-third