12 January 2024 19:32

RAKE => Rapid Automatic Keyphrase extraction

SAARC is consist of South Asian countries, India is supporting SAARC.

① Candidate generation

Splits data → Stopword
punctuation

List of candidates

[SAARC, consist, South Asian countries, India, Supporting SAARC]

② Candidate scoring

$$\text{Score} = \frac{\text{Degree}}{\text{Frequency}}$$

unique words	Frequency	Degree	Score
SAARC	2	2	2/2 = 1
consist	1	1	1/1 = 1
South	1	0	0/1 = 0
Asian	1	1	1
countries	1	1	1
India	1	1	1
supporting	1	0	0

	SAARC	consist	South	Asian	countries	India	supporting
SAARC	1	0	0	0	0	0	0
consist	0	1	0	0	0	0	0
South	0	0	0	1	0	0	0
Asian	0	0	0	0	1	0	0
countries	0	0	0	0	0	0	0
India	0	0	0	0	0	1	0
supporting	1	0	0	0	0	0	0
Degree	2	1	0	1	1	1	0

[SAARC, consist, South Asian countries, India, supporting SAARC]

"SAARC SAARC"

"SAARC"

"SAARC consist"

"consist consist"

"consist"

"South South"

"South"

Score of Keyphrases:

Score(South Asian countries) = Score of South + score of Asian + score of countries

= 0 + 1 + 1

= 2

Score (Supporting SAARC) = Score of supporting + score of SAARC

$= 0 \quad +1$

$= 1$

RAKE $\Rightarrow$ Keyphrase Score $\uparrow \Rightarrow$ Keyphrase important $\uparrow$

YAKE :-> Yet Another Keyphrase extracter.

SAARC is consist of South Asian countries,

India is supporting SAARC.

① Candidate generation

Splits data -> Stopword ____ stopword

Punctuation ____ stopword

Candidates -> SAARC, South Asian countries, supporting SAARC

② Candidate scoring

YAKE -> For every word 5 scores

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix}$$

a - score

ⓐ casing $(w) = \dfrac{\max(\text{count}(w \text{ is capital}), \text{count}(w \text{ is acronym}))}{1 + \log(\text{count}(w))}$

Acronym = SAARC, NGO

Capital = Title case count(w) = 4 + 3 + 2 = 9

SAARC = 4 times

Saarc = 3 times

saarc = 2 times

$$casing(SAARC) = \frac{masc(3,4)}{1+\log(9)} = \frac{4}{1+\log(9)}$$

b - score

(b) word position = $\log(\log(3 + median(Sen(w))))$

Sen(w) = list of positions of word

1 SAARC

10 SAARC SAARC 19

25 SAARC SAARC 36

Sen(SAARC) = [1, 10, 19, 25, 36]

word position (SAARC)

$= \log(\log(3 + median([1, 10, 19, 25, 36])))$

$= \log(\log(3 + 19))$

c - score

(c) word frequency score = count of word

$$\text{mean(count)} + \text{std.dev(count)}$$

$$\text{word frequency (SHARC)} = \frac{\text{count of SHARC}}{\text{mean}([c_1 \cdots c_{100}]) + \text{std.dev}([c_1 \cdots c_{100}])}$$

1 document = 100 words = $w_1, w_2, w_3, \cdots w_{100}$

count of word = $c_1, c_2, c_3 \cdots c_{100}$

count = $[c_1, c_2, c_3, \cdots c_{100}]$

d-score

(d) Word Relatedness

$$= 1 + (WR + WL) \times \frac{\text{count}(w) + PL + PR}{\text{max count}}$$

$$WR = \frac{\text{No. of unique words on Right}}{\text{Total words on Right}}$$

$$WL = \frac{\text{No. of unique words on left}}{\text{Total words on left}}$$

Max count = Total words in the document

$$PR = \frac{\text{Total words on Right}}{\text{Max count}}$$

$PL = \frac{\text{Total words on left}}{\text{Max count}}$

We are learning NLP and NLP is consist of both Machine learning and Deep learning.

Word Relatedness (consist) $= 1 + \left(\frac{6}{7} + \frac{6}{7}\right) \times \frac{1}{15} + \frac{7}{15} + \frac{7}{15}$

$WR = \frac{6}{7}$, $WL = \frac{6}{7}$

Max count = 15

$PR = \frac{7}{15}$, $PL = \frac{7}{15}$

e - score

(e) word different score = $\frac{\text{No. of sentences containing that term}}{\text{Total no. of sentences}}$

word different (SAARC) = $\frac{10}{100}$

$$\text{Score of } (w) = \frac{d \times b}{a + \left(\frac{c}{d}\right) + \left(\frac{e}{d}\right)}$$

Here, a = casing score

b = Position score

c = frequency score
d = word relatedness
e = Word different score

$$\text{Keyphrase score} = \frac{\text{Product}(\text{score}(A), \text{score}(B))}{1 + (\text{sum of score of } A, B) \times \text{count}(\text{Keyphrase})}$$

(A, B)

Supporting SARC

↳ "ss"

"ss"

"ss"

"ss"

count(Keyphrase) ↓ 3

YAKE => Keyphrase score ↓ => Keyphrase important ↑

RAKE => Keyphrase score ↑ => Keyphrase important ↑