Make Tagged Vocabulary → enumerate the vocabulary list

word→ ID

Frequency Measure of the words in the whole document→

[(8, 324), (8, 293), (8, 262), (8, 177), (8, 148), (8, 105), (8, 53), (7, 437), (7, 367), (7, 229), (7, 17), (6, 387), (6, 362), (6, 341), (6, 257), (6, 223), (6, 202), (6, 87), (6, 83), (6, 28), (5, 444), (5, 425), (5, 424), (5, 413), (5, 411), (5, 396), (5, 372), (5, 368), (5, 271), (5, 258), (5, 252), (5, 239), (5, 232), (5, 227), (5, 187), (5, 156), (5, 145), (5, 134), (5, 130), (5, 79), (5, 69), (5, 61), (5, 60), (5, 40), (5, 31), (5, 29)

Number of Words matched with the ID (8, 324) → (Matching frequency, PageID)

Assumption: The searching term having high frequency in the searched Page

Then, Look for matching sequence within these documents,

Query (“I want to know about mobicontrol Settings”)

→ Feature extraction [“want”, “know”, “mobicontrol”, “settings”]

→ So I am looking for the term “mobicontrol settings”

in the highest ranked pages

Then I will get a rank list

[[4.5, 324],

[4.5, 293],

[4.5, 262],

[4.5, 177],

[4.5, 148],

[3.0, 105],

[3.0, 53],

[1.5, 437],

[1.5, 17],

[1.0, 229],

[0.5, 367]]

Then I will try to find Ranks of these high frequency ranked pages with Gensim Doc2Vec vectorizer with Title/Heading only

Then I will try to find Ranks of these high frequency ranked pages with BERT Doc2Vec vectorizer with Title/Heading only

Find Most high ranked pages in all the 3 scenarios.

Compare and find best ranked pages.