T: original document

T’: document without URL and special characters

t1: the words in “()”

t2: the words between “/”

t3: the uppercase phrases

t4: the hyphen phrases

wi(i=1,2,3,4): the manually generate wordlist based on ti

W1: the concept list based on w1, w2, w3 and w4

W2: the concept list based on wiki title

W3: the concept list based on part of English dictionary and international information

W: the concept list based on W1, W2 and W3

t’i: the documents after delete W from T’

ki: t’I after NLP with NLTK

W’: the wordlist based on ki

W4: the manually generate concept list based on W’

Wfina: the concept list based on W and W4

C: the document’s concept index based on Wfinal and T’