COP290 ASSIGNMENT7

CHIRAG MOHAPATRA
2018CS50403

In this assignment , we had to develop a plagiarism checker which shows similarity between documents in a corpus directory versus a target document .

The method I have used for detecting the similarity percentage is the

Term_frequency-Inverse_document_frequency(TF-IDF) + Cosine similarity .

To do this , I maintain a vocabulary which keeps a track of words that have occured and their respective frequencies for each document and a count of the no of documents which contain this word(This count is used to determine the importance of the word) since words like ‘i’ , ‘and’ which occur in almost every document are comparitively less important .

For the vocabulary , I have maintained a hash table of words which follows direct hashing with a single hash function .

For my hash table , both search and insert are amortized O(1) . I have assumed that maximum 1,000,000 unique words would occur in the vocabulary and set this as table size , though this is easily modifiable .

The time complexity of my algorithm would be O(n) where n is number of words in vocabulary . This algorithm is very fast .

Once this is done , I make vectors for each document(represented as array of floats) with size as the number of words in our vocabulary and the values as TF*IDF where:

TF = no of times the word occurs in the document / total number of words in doc
IDF = log_e(Total number of documents/ No of documents containing the word)

Once the vectors are made , we then return cosine of the angle between the target vector and the plagiarized vector as the similarity value between the two documents .

Overall the results for the sample corpus with input file catchmeifyoucan.txt were satisfying . The three plagiarized files show similarity levels of 34% , 42% and 53% while other files have levels less than 3 % .

ecu201.txt : 34.13 %
hal10.txt : 42.3 %
tyc12.txt : 53.33 %

For my algorithm:
Time Complexity = O(n) amortized
Space Complexity = O(n) where n is number of words in vocabulary
Implementation used : Hash table
Algorithm used : TF-IDF + cosine similarity