Problem 1

Considering the problem of finding a short passage within a document which is a near duplicate of some passage in another document, even if the remainder of the two documents are not similar at all. My approach is use similarity hash technique(simhash of Charikar) to detect duplicate passages. The algorithmic complexity is c(64,b)*lgN. The steps of my approach is as following:
Firstly, divide the homework into passages, and then process each passage and rank its words based on the frequency.
Secondly, design the hash function for this algorithm and generate the hash value for each word in b bits.
Thirdly, initialize the b-dimensional vector V to 0, and initialize the b-bit binary number s to 0.
Fourthly, for the component of each passage, using hash function(MD5) to compute the value. For i = 1 to b, if the i-th bit of the value is 1, then add the weight to the i-th element of V, otherwise, minus the weight if 0.
Fifthly, if the i-th element of V is more than 0, then set the i-th bit of s to 1, otherwise, set the i-th bit of s to 0. Output s as the signature.
Lastly, compare the signature of each passage with the data from database (suppose all the data has been processed), if they have the same simhash signature, then the target homework has plagiarism problem. Otherwise, this homework is orginal.

For simhash approach, the more similar file has the more similar values and the smaller Hamming distance. Thus this approach can guaranteed to find matches.
However, the accuracy of this algorithm will be determined by the bits b, the greater b and the higher accuracy.
This system will identify passages of various lengths. The longest passage is 64-bits.
It doesn't matter to worry about spam, because this system will be use to detect plagiarism homework and the database will be provided by professor offline. For this system, it is not necessary to search online.