PLAGIARISM DETECTION BETWEEN TWO FILES

Longest Common Subsequence

Heriberto Espino Montelongo, ID: 175199
Universidad de las Américas Puebla
P24-LIS2012-2: Matemáticas Discretas
Dr. Luis Oswaldo Valencia Rosado
Mayo de 2024

PLAGIARISM VS LCS

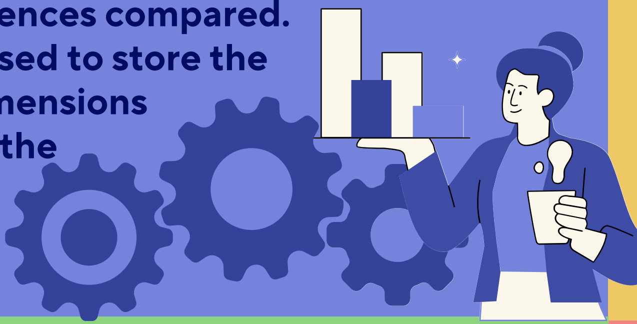
The problem of academic plagiarism is addressed through the use of the LCS, *Longest Common Subsequence* algorithm. With the aim of preserving academic integrity, addressing ethical concerns, and promoting academic excellence.

HOW LCS WORKS?

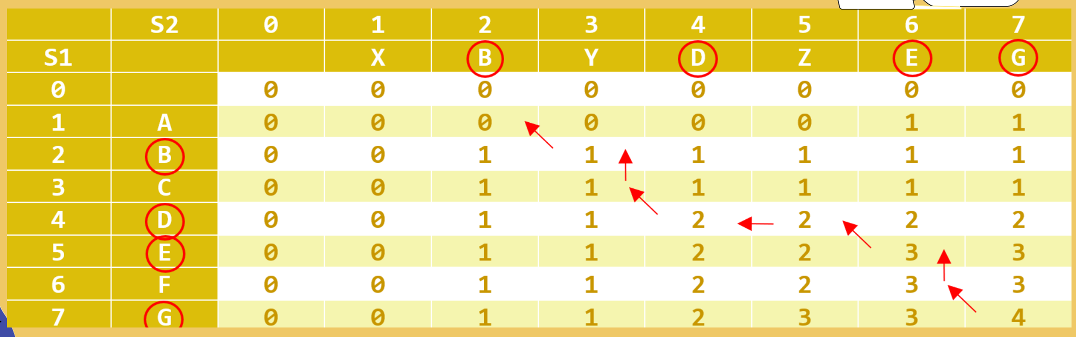
The LCS algorithm uses a matrix to compare characters from two sequences. It starts by creating an m*n matrix. where m and n are the lengths of the sequences. Then, it iterates through the matrix, comparing characters and updating lengths based on whether the characters are the same or different. .

TIME COMPLEXITY

The complexity is O(m*n), where m and n are the lengths of the two sequences compared. This is because the matrix used to store the intermediate results has dimensions m+1 by n+1, and each cell of the matrix requires constant time to be calculated.



VISUALIZATION



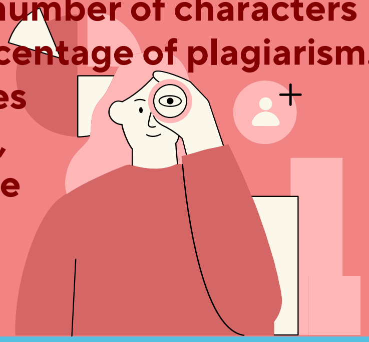
	S2	0	1	2	3	4	5	6	7
S1			X	B	Y	D	Z	E	G
0		0	0	0	0	0	0	0	0
1	A	0	0	0	0	0	0	1	1
2	B	0	0	1	1	1	1	1	1
3	C	0	0	1	1	1	1	1	1
4	D	0	0	1	1	2	2	2	2
5	E	0	0	1	1	2	2	3	3
6	F	0	0	1	1	2	2	3	3
7	G	0	0	1	1	2	3	3	4

LOOP INVARIANT

In each iteration of the main loop, the current submatrix contains the length of the longest common subsequence between the corresponding subsequences of the original sequences up to the current index.

APPLICATION

Only consecutive characters will be considered, with a minimum length requirement. The number of characters in the will be counted to get the percentage of plagiarism. Furthermore, different subsequences meeting these criteria will be stored, and their cumulative sum will provide the total plagiarism percentage.



RECURRENCE RELATION

LCS[i][j] = LCS[i-1][j-1] + 1 if X[i] == Y[j]
LCS[i][j] = max(LCS[i-1][j], LCS[i][j-1]) if X[i] != Y[j]

Where LCS[i][j] represents the length of the longest common subsequence between the first i elements of sequence X and the first j elements of sequence Y.

IN CONCLUSION...

The LCS algorithm offers an efficient method for determining the length of the longest common subsequence between two sequences. This algorithm is instrumental in various fields, including plagiarism detection, bioinformatics, and data compression, due to its ability to swiftly and accurately identify patterns of similarity.