Test :-

① Similarity of token in each basket.
↓ word   ↓ line

⇒ Hamming distance (match baskets of same length)

② longest match of tokens between two baskets

⇒ longest common subsequence

③ frequency of token plagiarism

⇒ Levenshtein distance

④ wordNGT on each subsequent token.

⑤ Stylometric by storing the vocabs strength

⇒ Source code:

① No. of lines of code.

② matching keywords on each line to complete document.

③ logical operation of matching keyword line.

④ number of methods, params for each method.