CHAPTER 3
RESEARCH METHODOLOGY

This project divided in several step, there is:

1. Finding data text sample and testing text

   Text sample is reference text for this project to find optimal value of kgram and window in range 1 till 10. The text sample in this project is taken from wikipedia with topic history computer. The text sample is rewritten from source to be txt format. While the testing text is reference text to find accuration of 3 similarity method. The testing text is one of the plagiarism cases in Indonesia that taken from news website and rewritten from source to be txt format.

2. Giving value k-gram and window

   Value k-gram is weight character that used as reference to split word. Partition word with k-gram is early step from winnowing algorithm. And value window is reference to split hash value from rolling hash step into window index. Range k-gram and window that will be used in this project between 1 until 10.

3. Implementation with winnowing algorithm

   a. Preprocessing whitespace insensitivity, removing irrelevant character such symbol, space and change uppercase into lowercase.
   b. Partition word based on k-gram.
   c. Rolling Hash process to produce hash value.
   d. Partition hash value based on window.
   e. Choose smallest value from each window to be a fingerprints.
   f. Calculate percentage similarity with jaccard similarity coefficients, sorensen dice similarity coefficient, and andberg similarity coefficients.
4. Testing and Analysing

Testing program with giving value k-gram range 1 until 10 and value window range 1 until 10 then analyse the percentage similarity results of 3 similarity method. After analysing the results then compare each method to find new conclusion of winnowing algorithm.