# CHAPTER 4
# ANALYSIS AND DESIGN

## 4.1 Analysis

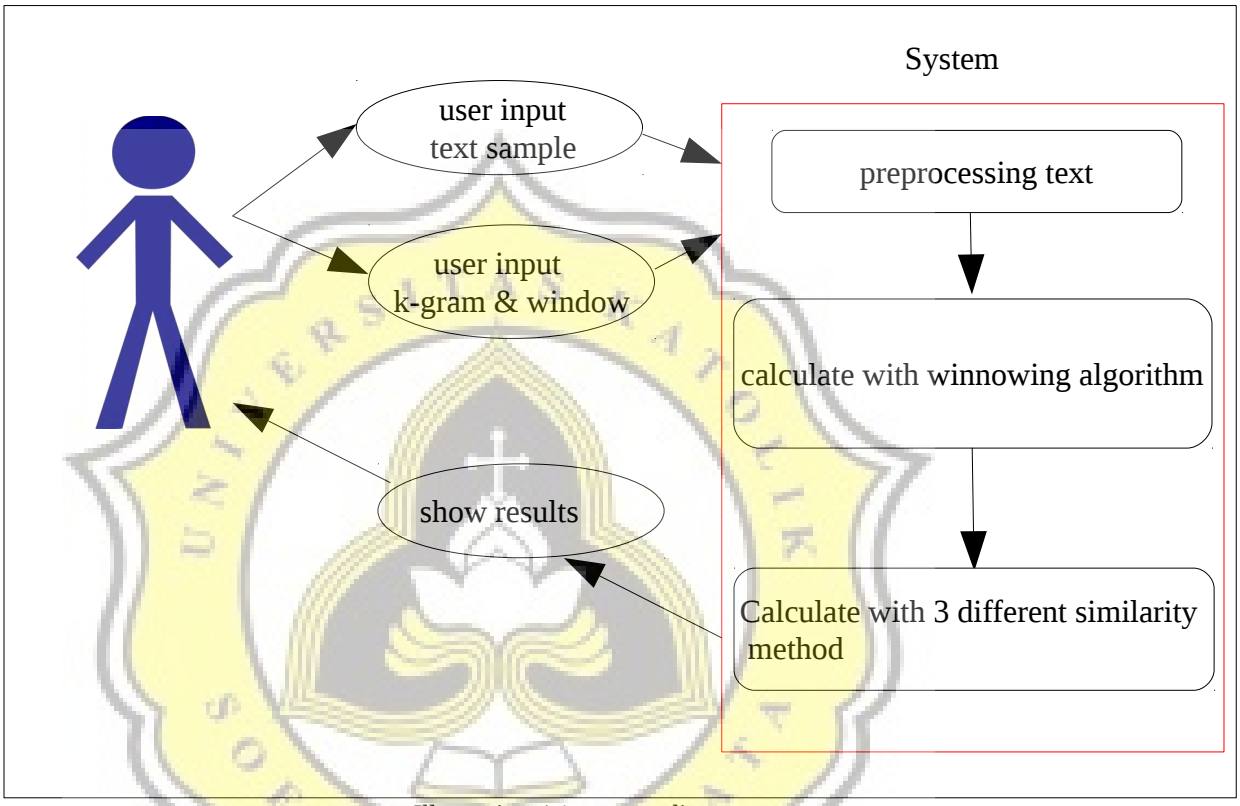


Illustration 4.1: usecasediagramsystem

In use case diagram above, first step is user input two text and input value of k-gram and window. After that the system will processing the text and calculate the similarity of the text using winnowing algorithm. This is the step of the winnowing algorithm :

1. Preprocessing whitespace insensitivity, removing unrelevant character such symbol, space and change uppercase into lowercase.

Example :

SANTI Membeli Buah!!! → santimembelibuah

2. Partition word based on k-gram (value k-gram should not more than text sample).

Example : k-gram = 9

santimembelibuah → santimemb antimembe ntimembel timembeli imembelib membelibu embelibua mbelibuah

3. Rolling Hash process to produce hash value.

Example : basic prime number = 3

santimemb antimembe ntimembel timembeli imembelib membelibu embelibua mbelibuah

H(C1) = 115*3^9 + 97*3^8 + 110*3^7 + 116*3^6 + 105*3^5 + 109*3^4 + 101*3^3 + 109*3^2 + 98*3

= 3263442 (santimemb)

H(C2) = 97*3^9 + 110*3^8 + 116*3^7 + 105*3^6 + 109*3^5 + 101*3^4 + 109*3^3 + 98*3^2 + 101*3

= 2999994 (antimembe)

H(C3) = 110*3^9 + 116*3^8 + 105*3^7 + 109*3^6 + 101*3^5 + 109*3^4 + 98*3^3 + 101*3^2 + 108*3

= 3272553 (ntimembel)

…

Do calculate rolling hash to each word, to find all word hash value.

4. Partition hash value based on window.

Example : window = 2

{3263442 2999994}, {2999994 3272553}, ….

5. Choose smallest value from each window to be a fingerprints.

Example : based step no 4 fingerprints is 2999994

6. Calculate percentage similarity with 3 different similarity method.

A.) Jaccard Similarity Coefficients

Formula : D(A,B) = $A \cap B$ / $A \cup B$ * 100

B.) Sorensen Dice Similarity Coefficients

Formula : D(A,B) = (2 * $A \cap B$ ) / (A + B) * 100

C.) Andberg Similarity Coefficients

Formula : D(A,B) = $A \cap B$ / ( $A \cup B$ + A $\triangle$ B) * 100

The process from winnowing algorithm is done and the result percentage similarity of text with 3 similarity method are appear.

## 4.2 Desain

To solve plagiarism cases, there is arranged a winnowing system. Here the flowchart of the system :

1. System Flowchart

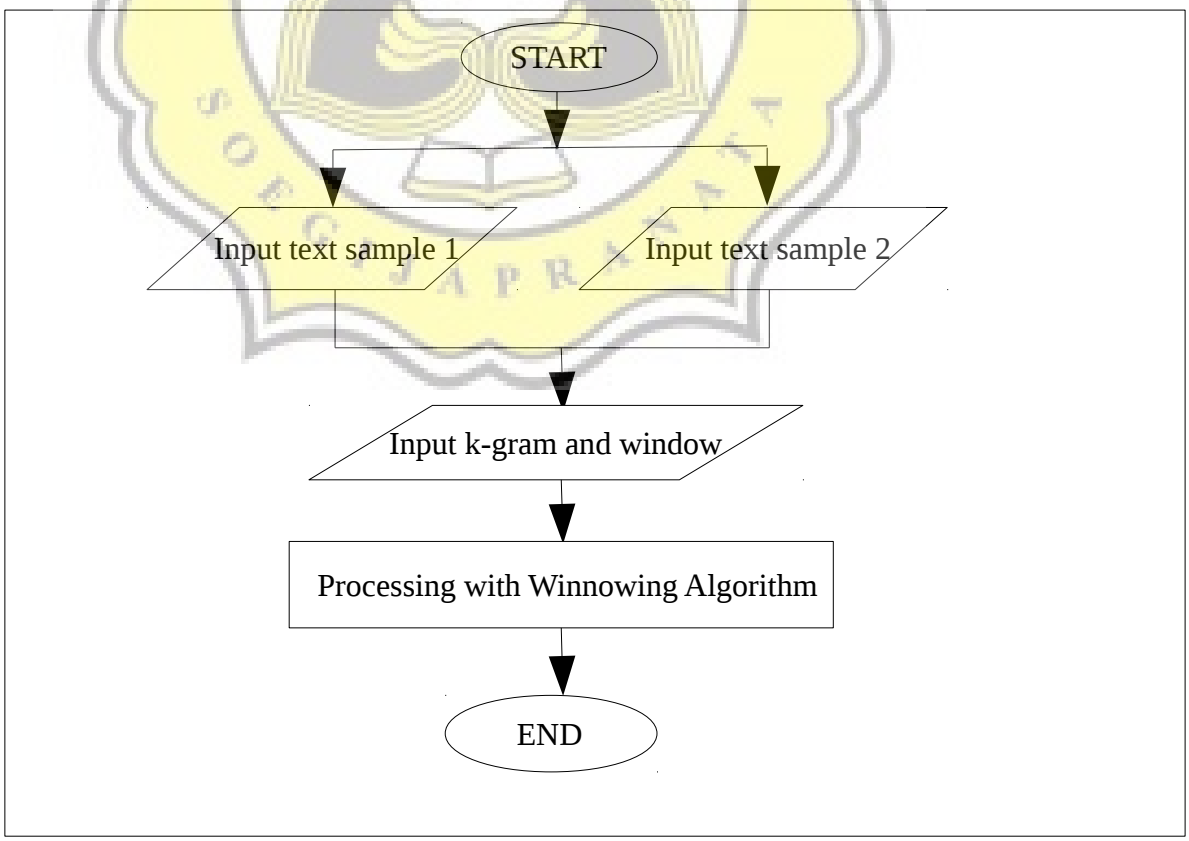


Illustration 4.2: systemflowchart

As seen in system flowchart, the first step from the system is input the text. After text has been added then input k-gram and window. Value k-gram and window is value that will be used as reference value on calculation in winnowing algorithm. Next step is calculate the similarity in text with winnowing algorithm using 3 different similarity method. Then do that step continuously with different input of value k-gram and window in range 1 until 10. After gain percentage of similarity in some testing with 3 different similarity method. Analysing the result and compare each method into another method.

2. Winnowing Algorithm Flowchart

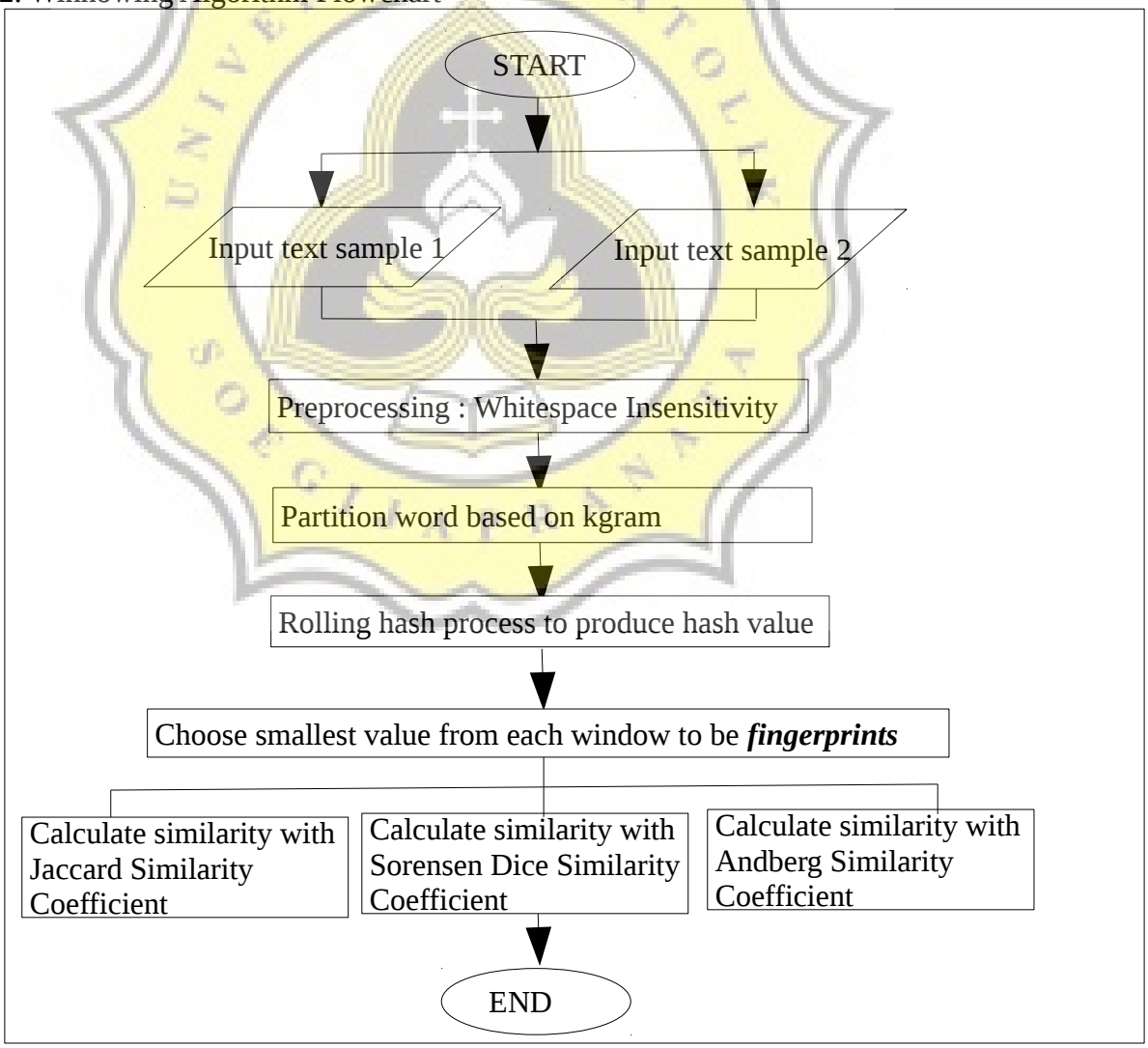


Illustration 4.3: winnowingalgorithmflowchart

In winnowing algorithm flowchart, explain the step by step algorithm. The first step is preprocessing the text, removing the symbol,space and change uppercase into lowercase in text. Second step is split the text into words based input value k-gram. Third step is finding hash value in each words using rolling hash function. Fourth step is split the hash value in each word into index window based input value window. Fifth step is check each index window and select smallest hash value to be a fingerprints. The last step after gain fingerprints is calculate similarity of the fingerprints using 3 different similarity method which is Jaccards Similarity Coefficient , Sorensen Dice Similarity Coefficient, Andberg Similarity Coefficient.

3. Rolling Hash Flowchart

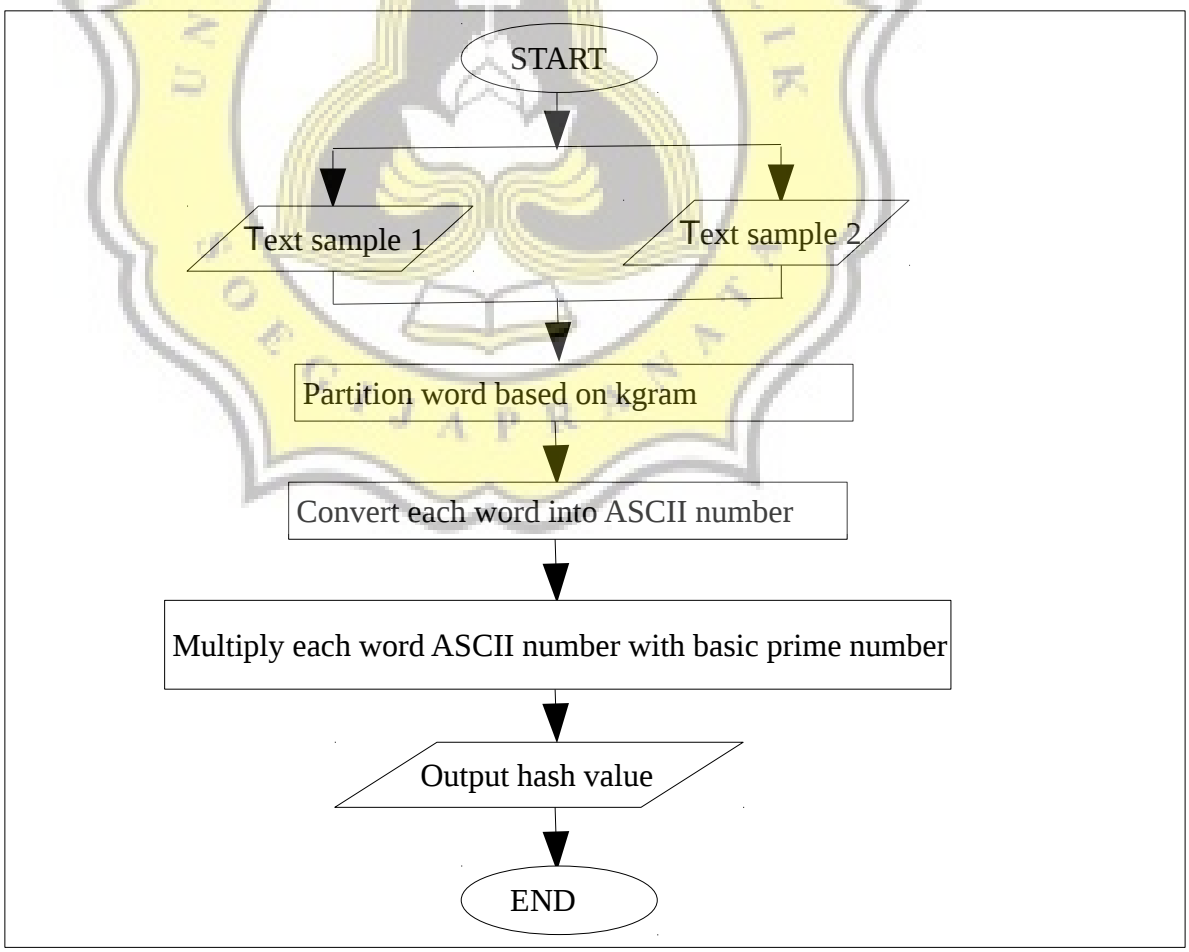


Illustration 4.4: rollinghashflowchart

In rolling hash flowchart is describe rolling hash function step. Rolling hash function is convert string into hash value. The first step is convert the word into ASCII number. After the word is change to ASCII, then multiply the ASCII with basic prime number. Basic prime number is part of rolling hash function formula. The results from the process is hash value each word.

4. Jaccards Similarity Coefficient Flowchart

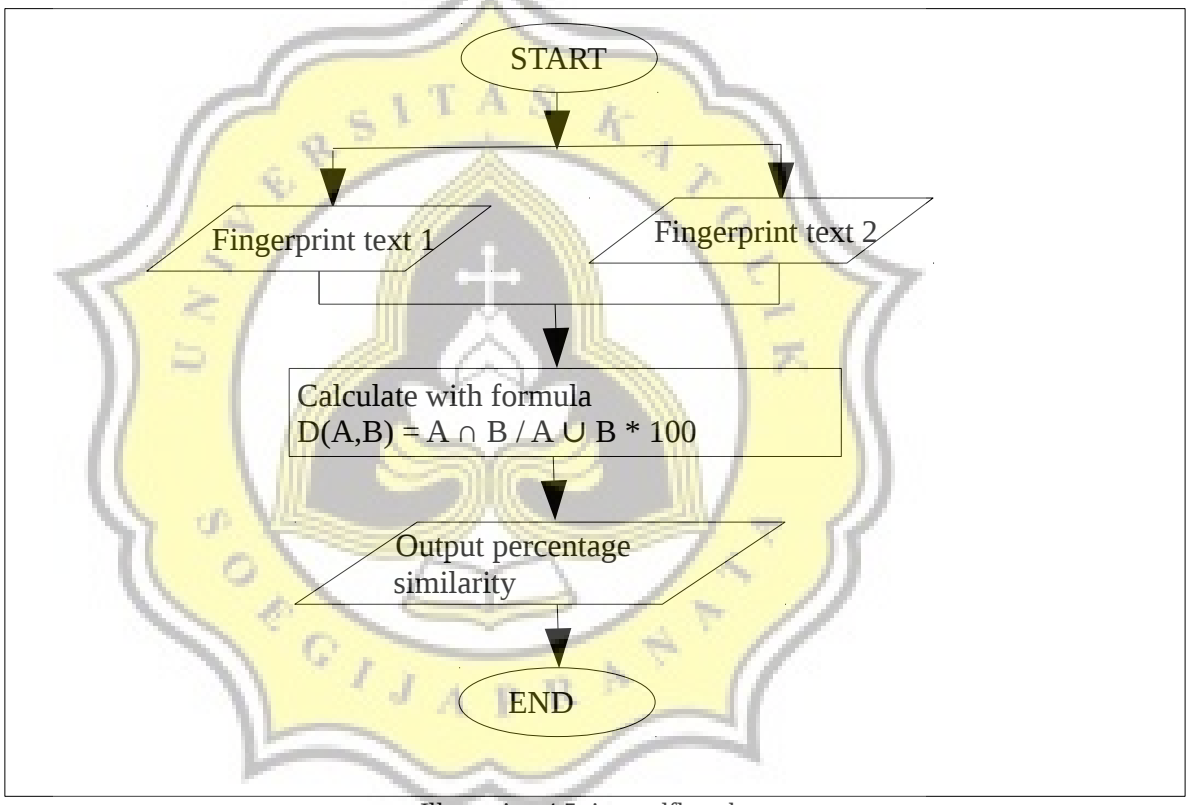


Illustration 4.5: jaccardflowchart

In jaccards similarity coefficient flowchart is discussed jaccard coefficient to calculate similarity fingerprints from text. The formula is amount of intersection in fingerprint text 1 and text 2 is divided with amount of union in fingerprint text 1 and text 2 and multiply with 100. The results is percentage of similarity in 2 text based on fingerprints.

5. Sorensen Dice Similarity Coefficient Flowchart

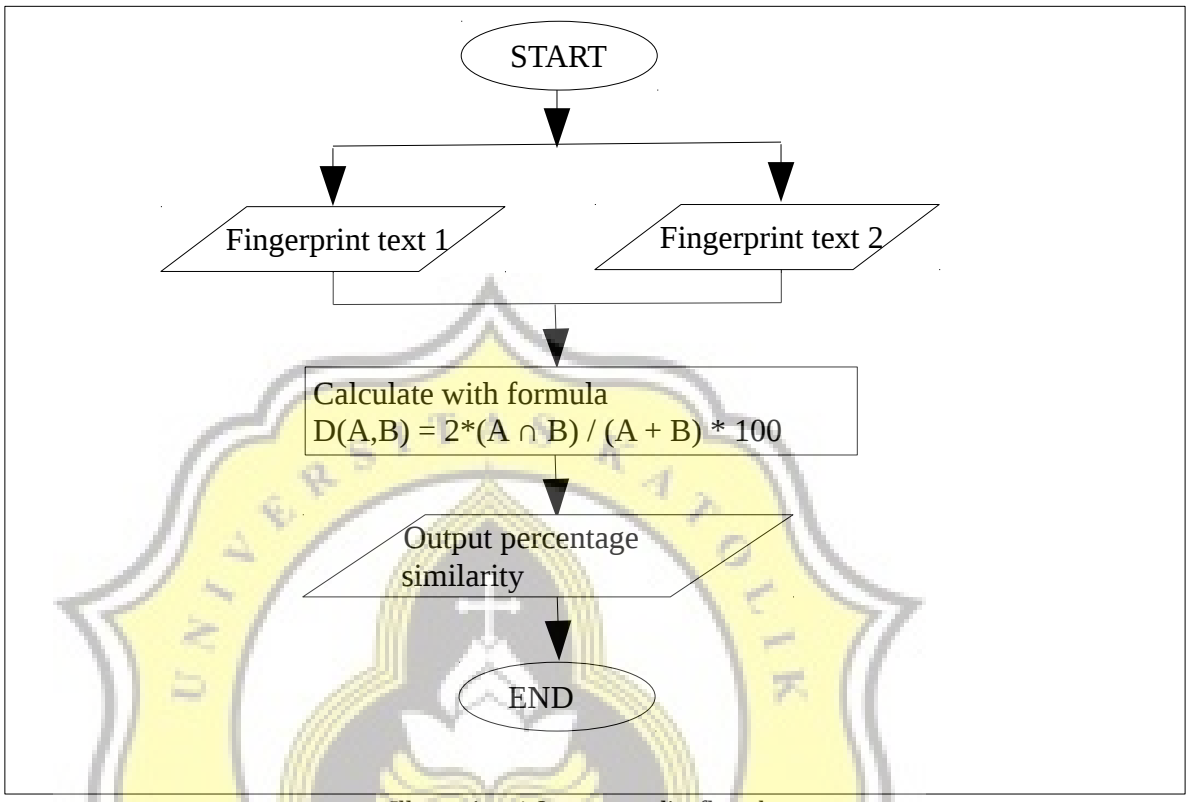


Illustration 4.6: sorensendiceflowchart

In sorensen dice similarity coefficient flowchart is discussed sorensen dice to calculate similarity fingerprints from text. The formula is amount of intersection in fingerprint text 1 and text 2 is multiply with 2 and then divided with the sum of fingerprint text 1 and fingerprint text 2 and multiply with 100. The results is percentage of similarity in 2 text based on fingerprints.

6. Andberg Similarity Coefficient Flowchart

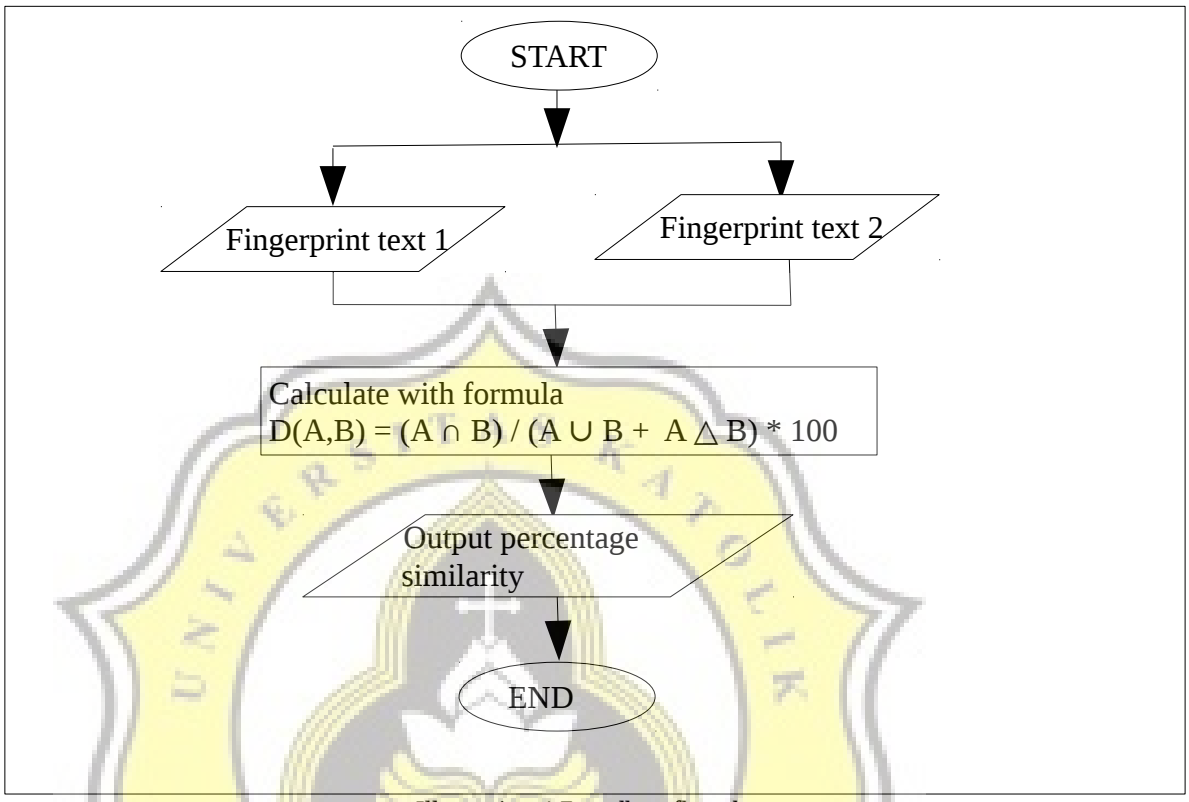


Illustration 4.7: andbergflowchart

In andberg similarity coefficient flowchart is discussed andberg similarity to calculate similarity fingerprints from text. The formula is amount of intersection in fingerprint text 1 and text 2 is divided with the sum of union fingerprint text 1 ,fingerprint text 2 and symmetric difference fingerprint text 1, fingerprint text 2. Symmetric difference is describe all elements in A or B, but not in both. The results is percentage of similarity in 2 text based on fingerprints.