CHAPTER 3
RESEARCH METHODOLOGY

3.1 Collecting Data

This project used database Kamus Besar Bahasa Indonesia that contains 1 column with 28526 rows.

3.2 Processing Data

1. Choose 3 words from database that is similar with the wrong word.
2. And then calculate the wrong word with each 3 suggestion word with Levensthein and Jaro Winkler algorithm.
3. The application of the Levenshtein Distance formula in correcting words is as follows:

d=(i-1,j)+1

d=(i,j-1)+1

d=(i-1,j-1)+*cost*

Where:

d = distance

i = *index* i

j = index j

cost = if the alphabet is same then the cost value is 0, if not same the value is 1.

From the result of 3 formula, take the lowest value of the distance.

4. The application of the Jaro Winkler Distance formula:

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \quad (1)$$

$$\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1. \quad (2)$$

$$d_w = d_j + \left(l \times p(1 - d_j)\right) \quad (3)$$

Where:

$|S1|$ = the length of the string 1

$|S2|$ = the length of the string 2

m = the number of matching char

t = the number of transposition

l = Prefix length (the same character length before found inequality) max 4 characters

p = Constant scaling factor (standard value for this constant according to Winkler is p = 0.1).

After each words are calculated by both algorithm, then analysis the results of the two algorithms which one is the best, where the distance values often indicate the most relevant word suggestions

3.3 Final Result Data

The result of Jaro Distance 0 indicates no similarity and 1 to indicate similarity. The result of Levensthein Distance 0 indicates similar and 1 to indicate no similarity

3.4 Report Writing

Write conclusion about the result calculation using both Levensthein and Jaro algorithm, including how the process determines the the final results of distance. After the conclusion, the author also add suggestion for the net study