Perbandingan Metode Pemeriksa Ejaan antara SymSpell dan Kombinasi Damerau-Levenshtein Distance dengan Struktur Data Trie = A Spell Checker Method Comparison Between SymSpell and a Combination of Damerau-Levenshtein Distance With the Trie Data Structure

Hanif Arkan Audah, author

Abstrak

Non-word error merupakan kesalahan ejaan yang menghasilkan kata yang tidak ada dalam kamus. Tujuan dari penelitian ini adalah membandingkan dua metode pemeriksa ejaan non-word error, yaitu SymSpell dan kombinasi Damerau-Levenshtein distance dengan struktur data trie. Kedua metode tersebut melakukan isolated-word error correction terhadap non-word error. Dalam implementasi, SymSpell dibedakan menjadi dua, yaitu weighted dan unweighted. Proses perbandingan metode dimulai dengan penyusunan kamus menggunakan entri kata dari KBBI V yang diperkaya dengan kata-kata tambahan dari Wiktionary. Kamus yang dihasilkan memuat 91.557 kata. Selanjutnya, disusun dataset uji yang dibuat secara sintetis dengan memanfaatkan modifikasi dari candidate generation Peter Norvig. Dataset uji sintetis yang dihasilkan memuat 58.532 kata salah eja. Dilakukan perbandingan antara Weighted SymSpell, Unweighted SymSpell, dan kombinasi Damerau-Levenshtein distance dengan struktur data trie menggunakan dataset uji sintetis tersebut. Perbandingan tersebut mengukur best match accuracy, candidate accuracy, dan run time. Hasil perbandingan menyimpulkan bahwa SymSpell memiliki performa yang lebih baik dibandingkan dengan metode kombinasi Damerau-Levenshtein distance dan struktur data trie karena unggul dari aspek best match accuracy dan run time serta memperoleh candidate accuracy yang setara dengan metode-metode lain. Implementasi SymSpell yang unggul, yaitu Weighted SymSpell memperoleh best match accuracy 66,79%, candidate accuracy 99,33%, dan run time 0,39 ms per kata.

.....Non-word errors are errors during writing where the resulting word does not exist in the dictionary. The objective is to compare non-word error spell checker methods, which are SymSpell and a combination of Damerau-Levenshtein distance with the trie data structure. Both methods handle non-word errors using isolated-word error correction.

During implementation, SymSpell is divided into two types: weighted and unweighted.

The comparison process starts by compiling a dictionary from word entries in KBBI V and Wiktionary. The resulting dictionary contains 91,557 words. The next step

is to synthetically generate a test dataset using a modified version of Peter Norvig’s candidate generation method. The resulting test dataset contains 58,532 misspellings.

A comparison is made between Weighted SymSpell, Unweighted SymSpell, and a

combination of Damerau-Levenshtein distance with the trie data structure using the synthetic test dataset that was generated. The comparison measures the best match accuracy, candidate accuracy, and run time.

The results found that SymSpell performed better than the method that used a combination of Damerau-Levenshtein distance with the trie data structure because it obtained a higher best match accuracy, lower run time, and

an equivalent candidate accuracy compared to the other methods. The best performing

SymSpell implementation is Weighted SymSpell which obtained a best match accuracy of 66.79%,

candidate accuracy of 99.33%, and a run time of 0.39 ms per word.