BAB 2
TINJAUAN PUSTAKA

2.1 Kesalahan Penulisan

Ketika mengetik teks, terkadang terjadi kesalahan penulisan yang dapat mengubah makna kata atau kalimat [13]. Kesalahan semacam itu bisa mengakibatkan ketidakjelasan informasi kepada pembaca, bahkan dapat menimbulkan pemahaman yang keliru terhadap informasi yang disampaikan

2.2 Kata Majemuk

Kata majemuk merupakan hasil gabungan morfem dasar yang secara keseluruhan menjadi sebuah kata dengan pola bunyi, gramatikal, dan semantis tertentu sesuai dengan aturan bahasa yang relevan. Pola khusus ini membedakannya dari kombinasi morfem dasar yang bukan termasuk dalam kategori kata majemuk. Dengan kata lain, kata majemuk terbentuk melalui proses penggabungan dua unsur kata yang membawa makna atau konsep baru. Kata majemuk berbeda dengan frasa. Jika kata majemuk merupakan gabungan dua unsur kata sehingga membawa makna baru, maka frasa merupakan gabungan dua unsur kata tetapi tidak membentuk makna baru[14].

Adapun karakteristik, pola pembentuk serta tata cara penulisan dari kata majemuk, yaitu sebagai berikut.

- Karakteristik Kata Majemuk[7]
  - Membentuk makna baru
    contoh: Matahari, rumah sakit, rumah makan
  - Tidak dapat disisipi kata lain
    contoh: pancaindra, kacamata, tinggi hati
  - Kata dasar tanpa imbuhan
    contoh: air mata, panjang tangan, otak udang
  - Kata yang membentuk tidak dapat dibalik
    contoh: alam semesta, orang tua, kaki tangan
- Pola Pembentuk Kata Majemuk[14]

  - Dapat terbentuk dari gabungan kata benda + kata sifat, begitu sebaliknya.
  - Dapat terbentuk dari gabungan kata benda + kata benda
  - Dapat terbentuk dari gabungan kata benda + kata kerja
  - Dapat terbentuk dari gabungan kata bilangan + kata benda
  - Dapat terbentuk dari gabungan kata kerja + kata kerja
  - Dapat terbentuk dari gabungan kata sifat + kata sifat

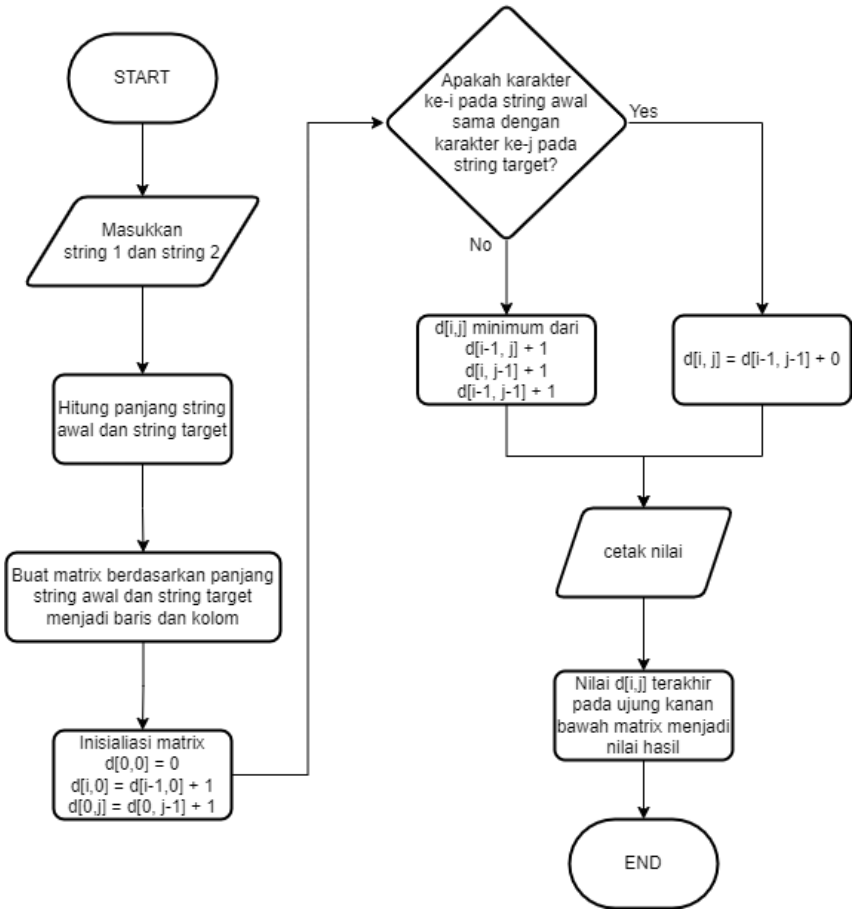
- Tata Cara Penulisan Kata Majemuk[15]
  - Kata Majemuk Senyawa

    Kata majemuk senyawa merupakan kata majemuk yang cara penulisannya dirangkaikan atau digabung. Contoh: matahari, narasumber, tunanetra
  - Kata Majemuk Non-senyawa

    Kata majemuk tak senyawa adalah istilah untuk kata majemuk yang dalam penulisannya keberadaan morfem-morfem dasarnya tetap terpisah. Contoh: harga diri, rumah sakit, kepala negara

2.3 Fuzzy-Wuzzy

FuzzyWuzzy adalah *library* Python yang menyediakan algoritma pencocokan *string* (*string matching*) memungkinkan pencocokan kata secara kasar. Dalam penelitian ini, digunakan metode *ratio* untuk mencocokan kata. Metode *ratio* ini sendiri merupakan metode yang dimiliki oleh *library fuzzywuzzy* yang berguna untuk menghitung *ratio* kemiripan antara dua *string*. *FuzzyWuzzy* merupakan *library* berbasis algoritma *Levenshtein Distance*. Algoritma ini berguna ketika Anda ingin mencocokkan kata yang mungkin memiliki kesalahan pengetikan, perbedaan karakter, atau variasi lainnya. Algoritma *Levenshtein Distance* bekerja dengan mengukur seberapa berbeda dua *string* dengan cara menghitung jumlah minimum operasi yang diperlukan untuk mengubah satu string menjadi string lainnya sehingga *string* yang memiliki nilai operasi paling sedikit saat dibandingkan dengan *string* lain dianggap sebagai *string* paling dekat atau paling cocok[16]. Operasi tersebut antara lain melibatkan penyisipan (*insertion*), penghapusan (*deletion*), dan penggantian (*substitution*) dari satu karakter.



Gambar 2.1. Flowchart Algoritma Levenshtein Distance

Gambar 2.1 menunjukkan langkah-langkah algoritma *Levenshtein Distance* bekerja yang akan dijelaskan berikut[12].

- Pertama memasukkan string awal dan string target
- Menghitung panjang string awal dan string target
- Membuat matrix MxN dengan M panjang string awal dan N panjang string target
- Inisialisasi matrix baris pertama dengan 0...M dan kolom pertama dengan 0...N
- Masuk ke dalam proses perhitungan, Periksa S[i] untuk $1 < i < M$ dan Periksa T[j] untuk $1 < j < N$. Jika S[i] = T[j], maka nilai yang dimasukkan ke dalam matrix adalah nilai yang terletak pada tepat didiagonal atas sebelah kiri, yaitu

d[i,j] = d[i-1,j-1]. Jika S[i] $\neq$ T[j], maka nilai yang dimasukkan ke dalam matrix adalah d[i,j] minimum dari:

  - Nilai yang terletak tepat diatasnya, ditambah satu, yaitu d[i,j-1]+1
  - Nilai yang terletak tepat dikirinya, ditambah satu, yaitu d[i-1,j]+1
  - Nilai yang terletak pada tepat didiagonal atas sebelah kirinya, ditambah satu, yaitu d[i-1,j-1]+1

- lakukan proses sebelumnya hingga *cell* matrix paling akhir hingga *cell* matrix d[i,j] ditemukan

Tidak berhenti disitu, setelah mendapat nilai operasi berdasarkan hasil perhitungan *Levenshtein Distance* selanjutnya adalah menghitung nilai atau bobot kemiripan. Formula untuk menghitung nilai atau bobot kemiripan adalah sebagai berikut[17].

$$\text{Sim} = 1 - \left(\frac{\text{Dis}}{\text{MaxLength}} \right) \tag{2.1}$$

- Dis merupakan hasil nilai operasi *Levenshtein Distance*
- MaxLength merupakan jumlah kata terpanjang dari string awal atau string target

Jika nilai *similarity* sama dengan 1 atau 100%, maka kedua *string* yang dibandingkan itu sama atau tidak ada operasi yang diperlukan. Sebaliknya, jika nilai *similarity* dibawah 1 atau dibawah 100%, maka kedua *string* yang dibandingkan tidak sama.

2.4 NLP-Id

NLP-Id merupakan *library* yang mampu memahami sebuah kata atau kalimat terutama dalam bahasa Indonesia. Salah satu fitur yang digunakan dalam program ini adalah *PosTag*. Fitur ini merupakan fitur yang dapat memberi tahu status kata dalam sebuah kalimat. Dengan berdasarkan fitur ini, program berbasis *rule based* dapat dibuat. Selain fitur *PosTag*, terdapat fitur lain yang digunakan, yaitu *lemmatizer*. Fungsi dari fitur ini adalah untuk mengekstrak suatu kata menjadi kata dasarnya itu sendiri. Tujuan dari hal itu adalah untuk mengetahui suatu kata tersebut memiliki imbuhan atau tidak karena salah satu ciri dari kata majemuk adalah tidak dapat diperluas yang berarti tidak memiliki imbuhan.