**Laporan Tugas Besar “Text Similarity”**

Laporan Ini Disusun untuk Memenuhi Tugas Mata Kuliah Struktur Data dan Algoritma

PROGRAM STUDI DIPLOMA IV TEKNIK INFORMATIKA

Jurusan Teknik Komputer dan Informatika



**Muhammad Azra Sabila (171524020)**

**Mumuh Kustino Muharram (171524025)**

**Satrio Muliatama Sudrajat (171524028)**

POLITEKNIK NEGERI BANDUNG

2018

Deskripsi Program

Pada kesempatan kali ini, kami akan membuat program untuk memeriksa kemiripan dari dua buah file teks. Output yang dihasilkan adalah statistik kata dari masing-masing file teks dan presentasi plagiarisme dari kedua file tersebut, presentasenya dapat dihitung dengan rumus jumlah frekwensi kata yang sama pada BST dikali jumlah file yang di bandingkan dibagi jumlah frekuensi kata pada setiap BST x 100%.

Untuk memproses kedua file ini ada beberapa tahapan yang harus dilewati, yaitu Case Folding, Tokenizing, Stopword Removal, dan Stemming. Keempat proses tersebut dilakukan agar hasil dari pemeriksaan kedua teks memiliki hasil yang akurat. Aplikasi yang akan kami buat akan membandingkan kata yang telah diolah dengan keempat proses tadi lalu mencari frekuensi jumlah kata yang sama.

Case Folding

Case Folding adalah proses untuk mengubah isi file teks ke bentuk lowercase pada semua karakter yang ada, hal ini dilakukan agar membuat kata yang sama dianggap sama jika semua karakter dalam kata tersebut sama. Contoh, kata “saya” dan “Saya” akan dianggap berbeda karena ascii dari huruf ‘s’ dan ‘S’ berbeda.

Tokenizing

Tokenizing adalah proses pemisahan kata pada teks, cara kerjanya adalah dengan membaca masing-masing karakter dan dimasukkan kedalam array hingga ditemukan karakter spasi ‘ ’. jika spasi ditemukan maka menandakan karakter yang sudah dibaca merupakan satu kata.

Stopword Removal

Stopword removal adalah proses penghapusan kata-kata yang tidak memberikan informasi penting seperti kata hubung, proses ini berguna agar kata yang diproses menjadi lebih efektif dan jumlahnya tidak terlalu banyak. Metode yang digunakan pada proses Stopword Removal ini adalah dengan cara memperoleh stopword dari kamus stopword yang sudah ada lalu membandingkan kata yang diperiksa dengan kamus yang sudah ada. Jika kata tersebut termasuk kedalam kategori Stopword maka kata tersebut akan dibuang.

Stemming

Stemming adalah proses untuk mengubah suatu kata kedalam bentuk dasarnya, dengan cara merubah imbuhan agar mendapatkan bentuk dasar dari kata tersebut. Proses ini dilakukan agar mengurangi jumlah kata dan mencegah adanya duplikasi kata pada statistik.

Alur Program

1. Pertama-tama program akan membaca dua teks file, dimana kedua teks tersebut adalah teks yang akan diperiksa tingkat similaritasnya.
2. Terlebih dahulu dilakukan case folding dimana masing-masing karakter akan dibuat lowercase dan jika karakter yang dibaca bukan karakter alfabet maka karakter tersebut tidak akan dimasukkan
3. Proses kedua yang dilakukan adalah tokenizing, file akan dibaca per char nya dan memasukkan char yang dibaca kedalam array of char. Apabila char yang dibaca adalah spasi maka pembacaan char pada teks akan ditunda, spasi menandakan akhir dari satu kata.
4. Sebelum kata yang sudah didapat diinput kedalam statistik, dilakukan proses Stopword Removal terlebih dahulu. Kata yang ada dalam array of char akan dibandingkan dengan isi dari binary tree stopword yang merupakan kamus dari stopword yang disimpan dalam bentuk binary tree, struktur binary tree digunakan agar proses pengecekan kata dilakukan lebih cepat karena menggunakan Binary Search Tree dan struktur tree yang digunakan adalah AVL.
5. Proses selanjutnya adalah stemming, kata yang sudah melewati proses Stopword removal selanjutnya masuk ke proses stemming. Dimana, kata yang memiliki imbuhan diawal atau diakhir akan diubah kebentuk kata dasarnya. Proses dilakukan dengan cara membandingkan awal atau akhir kata dengan kamus stemming yang sudah ada, cara menyimpan kamusnya sama dengan proses stopword removal. Setelah dibandingkan, ternyata kata yang disimpan itu berimbuhan maka imbuhan akan diganti dengan huruf yang seharusnya atau imbuhannya dihilangkan.
6. Jika kata sudah berbentuk kata dasarnya, kemudian kata akan disimpan dalam struktur AVL. Semua proses tersebut dilakukan hingga mencapai End Of File bagi kedua file, kemudian langkah yang dilakukan adalah cek similaritasnya.

Fitur program :

* Terdapat file kamus dari stopword yang dapat kita tambahkan dan isinya melalui program.
* Tampilan program hasil dari membandingkan dua teks ialah table table kata per file teks.
* Hasil statistik dan persentase juga akan disimpan pada file.