

BAB III

METODOLOGI PENELITIAN

3.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah mengimplementasikan algoritma Smith-Waterman dalam membangun sebuah aplikasi yang dapat mendeteksi dugaan awal plagiarisme berdasarkan kemiripan isi teks dari dokumen-dokumen yang dibandingkan.

3.2 Tempat dan Waktu Penelitian

Penelitian ini dilakukan dilaboratorium multimedia jurusan Teknik Elektro Universitas Negeri Jakarta sejak bulan Oktober 2014 hingga Januari 2015

3.3 Definisi Kebutuhan (*Requirements Definition*)

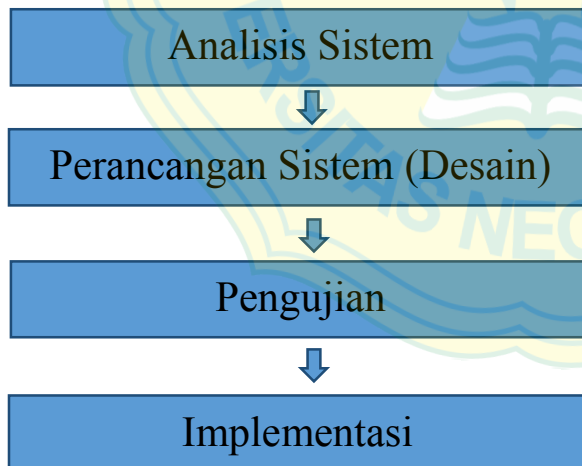
Mengimplementasikan Algoritma Smith-Waterman dalam penelitian ini untuk membantu *user* dalam mendeteksi dugaan awal plagiarisme berdasarkan tingkat kemiripan isi teks dokumen yang dibandingkan. Adapun kebutuhan perangkat keras dan perangkat lunak yang di gunakan dalam mengembangkan aplikasi dugaan awal plagiarisme adalah sebagai berikut:

- a. Processor Intel® Core™ i7-3630QM
- b. Memory 8 GB RAM
- c. Operation system windows 7 ultimate 32-bit
- d. NetBeans IDE 8.0
- e. Microsoft Visio 2010

3.4 Metode Penelitian

Metode penelitian yang digunakan pada penelitian ini adalah salah satu metode pengembangan perangkat lunak, yaitu metode pengembangan *software*. Pada metode waterfall, setiap tahapan harus diselesaikan terlebih dahulu secara penuh sebelum diteruskan ke tahap berikutnya dengan tujuan untuk menghindari terjadinya pengulangan tahapan (Kusrini, 2007). Proses ini lebih cocok untuk diterapkan dalam pengembangan sebuah aplikasi yang menggunakan algoritma Smith-Waterman dengan bahasa pemrograman *Java*. Dalam tahap uji coba menggunakan beberapa dokumen tugas mahasiswa untuk dibandingkan dan mencari nilai kemiripan pada teks dokumen mahasiswa. Semakin besar nilai kesamaan isi teks dokumen, semakin tinggi juga tingkat kemiripan isi teks dokumen. Untuk menentukan plagiarisme atau tidaknya adalah *user* yang menentukan, karena aplikasi ini hanya membantu *user* dalam mendeteksi dugaan awal kemiripan isi teks dokumen.

Berikut adalah gambar tahapan pengembangan aplikasi 3.1:



Gambar 3.1 Tahapan Pengembangan Aplikasi

3.4.1 Analisis Sistem

Tahap analisis sistem dalam pengembangan aplikasi dugaan awal plagiarisme adalah menentukan langkah-langkah rancangan aplikasi untuk mendapatkan nilai kemiripan teks antara dua dokumen yang berbeda untuk dibandingkan isi teks dokumen dengan memanfaatkan algoritma Smith-Waterman dalam perhitungannya.

1. Menganalisa perancangan sistem aplikasi.
2. Melakukan *pre-processing* sebelum perhitungan algoritma Smith-Waterman seperti (*read file, tokenizing, case folding, sorting*).
3. Mengimplementasikan algoritma Smith-Waterman untuk perhitungan isi teks dokumen dalam menentukan nilai dugaan plagiarisme.
4. Mendesain *graphical user interface*, untuk mempermudah percobaan dan pemakaian aplikasi.
5. Melakukan uji coba aplikasi dalam menentukan kemiripan isi teks dokumen untuk mendapatkan presentasi nilai kemiripan antara dokumen yang dibandingkan.

3.4.2 Perancangan Sistem

Perancangan Sistem pada aplikasi ini sebagai penggambaran, perencanaan, dan pembuatan sketsa atau pengaturan dari beberapa elemen dalam proses pengolahan data dan prosedur untuk mendukung operasi sistem. Adapun tujuan utama perancangan sistem ini adalah untuk memenuhi kebutuhan para pemakai sistem (Dosen).

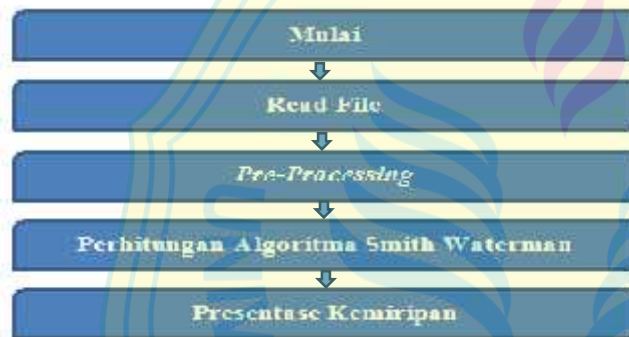
Dalam pengembangannya aplikasi ini menggunakan algoritma Smith-Waterman.

Adapun *input* dan *output* yang dihasilkan pada aplikasi ini adalah:

1. *Input* sistem dalam aplikasi ini adalah *user* meng-*upload* tugas mahasiswa yang ingin dibandingkan.
2. *Output* yang dihasilkan dalam sistem aplikasi ini berupa persentase kemiripan dokumen yang dibandingkan.

3.4.2.1 Tahapan Perancangan Sistem

Tahapan pengembangan aplikasi dalam penelitian ini terdiri dari empat tahap , yaitu : *read file*, *pre-processing*, perhitungan algoritma Smith-Waterman, dan persentase kemiripan. Berikut gambar 3.5 alur sistem:



Gambar 3.2 Tahapan Perancangan Sistem

Tahap pertama sistem membaca file dokumen yang telah di-*upload user*, file dokumen harus berektensi *txt*, *doc*, atau *docx*. Tahapan selanjutnya adalah *pre-proccesing* yang meliputi: *read file*, *tokenizing*, *case folding*, *analyzing* dan *sorting*. Kemudian dilakukan perhitungan isi teks dokumen menggunakan algoritma Smith-Waterman, dan tahap yang terakhir adalah nilai persentase kemiripan dokumen.

3.4.2.2 Pre-Processing

Pre-Processing adalah suatu proses atau langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas. Proses *pre-processing* dalam penelitian ini meliputi seperti: *readfile*, *tokenizing*, *case folding*, dan *sorting*.

1) *Read file*

Read file adalah tahap membaca *file* pada aplikasi setelah *user* melakukan *input* pada *file* tersebut. *File* yang bisa digunakan pada aplikasi ini harus *.txt*, *.doc*, *.docx*.

2) *Tokenizing*

proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen. Pada aplikasi ini proses *tokenizing* digunakan sebanyak dua kali, *tokenizing* pertama yaitu memecah paragraf menjadi satuan kalimat, kemudian kalimat dipecah kembali menjadi satuan kata.

3) *Case folding*

Case folding merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan.

4) *Analyzing*

Setelah proses *readfile*, *tokenizing*, dan *case folding*, maka dilakukan proses *analyzing*. Tahapan ini adalah untuk menyeleksi pasangan *token* kalimat antar dokumen yang dibandingkan untuk diproses ke tahapan selanjutnya.

Berikut adalah rumus perhitungan pada tahap *analyzing*:

$$\text{Nilai analyzing} = \left\{ \frac{\text{Total Kesamaan Kata}}{\max(TkD1, TkD2)} \right\} \quad \text{Rumus (3.1)}$$

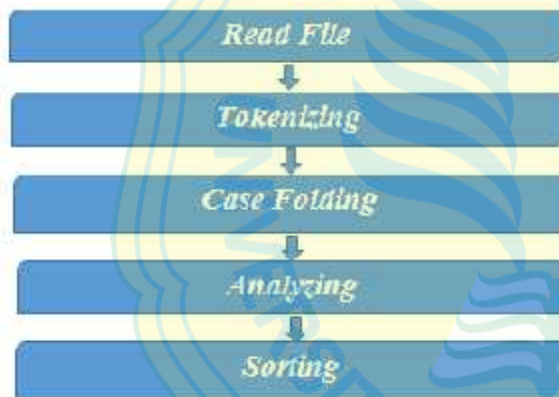
*ket:

TkD1 = Total kata pada token kalimat dokumen 1

TkD2 = Total kata pada token kalimat dokumen 2

5) *Sorting*

Tahap *sorting* yaitu proses pengurutan kata, Pada dasarnya ada dua macam urutan yang biasa digunakan dalam suatu proses *sorting* namun yang digunakan pada aplikasi ini adalah secara *ascending*. Tahapan ini dilakukan pada pasangan *token* kalimat yang lolos *threshold*. Setelah tahap ini, maka dilakukan proses akhir, yaitu *processing*.



Gambar 3.3 Alur Pre-processing

3.4.2.3 Cara mendapatkan Nilai Akhir

Berikut adalah contoh cara mendapatkan nilai akhir:

a. Dokumen uji 1

Frans mengerjakan tugas bahasa indonesia di malam hari. Karena di pagi hari harus membantu ibu berjualan. Dengan sekuat tenaga ia bekerja.

b. Dokumen Uji 2

Frans mengerjakan tugas bahasa indonesia di pagi hari. Karena di malam hari harus membantu ibu berjualan. Dengan penuh tenaga ia bekerja.

Pemecahan paragraf menjadi kalimat dokumen:

a. Dokumen uji 1

1. Frans mengerjakan tugas bahasa indonesia di malam hari
2. Karena di pagi hari harus membantu ibu berjualan
3. Dengan sekuat tenaga ia bekerja

b. Dokumen uji 2

1. Frans mengerjakan tugas bahasa indonesia di pagi hari
2. Karena di malam hari harus membantu ibu berjualan
3. Dengan penuh tenaga ia bekerja

Perbandingan kalimat dokumen uji 1 dan dokumen uji 2:

Dokumen A1 : frans mengerjakan tugas bahasa indonesia di malam hari

Dokumen B1 : frans mengerjakan tugas bahasa indonesia di pagi hari

Jumlah kata yang sama = 7

Total banyaknya kata maksimal antara kalimat A1 dan B1 = 8

Jumlah kata yang sama dibagi total banyaknya kata maksimal kalimat A1 dan B1 $\frac{7}{8}$

Hasil *analyzing* = 0.875

Dokumen A1 : Frans mengerjakan tugas bahasa indonesia di malam hari

Dokumen B2 : Karena di malam hari harus membantu ibu berjualan

Jumlah kata yang sama = 3

Total banyaknya kata maksimal antara kalimat A1 dan B2 = 8

Jumlah kata yang sama dibagi total banyaknya kata maksimal kalimat A1 dan B2 $\frac{3}{8}$

Hasil *analyzing* = 0

Dokumen A1 : frans mengerjakan tugas bahasa indonesia di malam hari

Dokumen B3 : dengan penuh tenaga ia bekerja

Tidak ada kata yang sama

Hasil *analyzing* = 0

Dokumen A2 : karena di pagi hari harus membantu ibu berjualan

Dokumen B1 : frans mengerjakan tugas bahasa indonesia di pagi hari

Jumlah kata yang sama = 3

Total banyaknya kata maksimal antara kalimat A2 dan B1 = 8

Jumlah kata yang sama dibagi total banyaknya kata maksimal kalimat A2 dan B1 $\frac{3}{8}$

Hasil *analyzing* = 0

DokumenA2 : karena di pagi hari harus membantu ibu berjualan

DokumenB2 : karena di malam hari harus membantu ibu berjualan

Jumlah kata yang sama = 7

Total banyaknya kata maksimal antara kalimat A2 dan B2 = 8

Jumlah kata yang sama dibagi total banyaknya kata maksimal kalimat A2 dan B2 $\frac{7}{8}$

Hasil *analyzing* = 0.875

DokumenA2 : karena di pagi hari harus membantu ibu berjualan

DokumenB3 : dengan penuh tenaga ia bekerja

Tidak ada kata yang sama

Hasil *analyzing* = 0

DokumenA3 : dengan sekuat tenaga ia bekerja

DokumenB1 : frans mengerjakan tugas bahasa indonesia di pagi hari

Tidak ada kata yang sama

Hasil *analyzing* = 0

DokumenA3 : dengan sekuat tenaga ia bekerja

DokumenB2 : karena di malam hari harus membantu ibu berjualan

Tidak ada kata yang sama

Hasil *analyzing* = 0

DokumenA3 : dengan sekuat tenaga ia bekerja

DokumenB3 : dengan penuh tenaga ia bekerja

Jumlah kata yang sama = 4

Total banyaknya kata maksimal antara kalimat A2 dan B2 = 5

Jumlah kata yang sama dibagi total banyaknya kata maksimal kalimat A2 dan B2 $\frac{4}{5}$

Hasil *analyzing* = 0.80

Dari perhitungan diatas, perbandingan pasangan *token* kalimat yang nilai *analyzing*-nya melebihi *threshold* 60 adalah:

Dokumen A1 : frans mengerjakan tugas bahasa indonesia di malam hari

Dokumen B1 : frans mengerjakan tugas bahasa indonesia di pagi hari

DokumenA2 : karena di pagi hari harus membantu ibu berjualan

DokumenB2 : karena di malam hari harus membantu ibu berjualan

DokumenA3 : dengan sekuat tenaga ia bekerja

DokumenB3 : dengan penuh tenaga ia bekerja

Maka selanjutnya, pasangan *token* kalimat ini akan melalui tahap *sorting*.

Dokumen A1 : bahasadi frans hari indonesiamalam mengerjakan tugas

Dokumen B1 : bahasadi frans hari indonesiamengerjakan pagitugas

Nilai Smith = 0.797

Dokumen A2 :berjualandihariharusibukarenamembantupagi

Dokumen B2 :berjualandihariharusibukarenamalammembantu

Nilai Smith = 0.821

Dokumen A3 :bekerjadenganiasekuattenaga

Dokumen B3 :bekerjadenganiapenuhtenaga

Nilai Smith = 0.777

Setelah *sorting*, maka akan dicari nilai perhitungan Smith-Waterman. Hasil perhitungan dari pasangan *token* yang lolos *threshold* kalimat diatas adalah = 0.797, 0.821, 0.777. Untuk mencari persentase akhir, maka nilai perhitungan dari pasangan *token* kalimat yang lolos *threshold* diakumulasi.

Dari contoh diatas, karena ada tiga pasangan *token* kalimat yang lolos *threshold*, maka total akumulasinya adalah 79.90. Rumus untuk mendapatkan persentase nilai akhir adalah

$$\text{Persentase akhir} = \left\{ \frac{\text{Akumulasi nilai SmithWaterman}}{\max(\text{lolos}, \max(\text{TkD1}, \text{TkD2}))} \right\} \times 100$$

$$= \left\{ \frac{0,797+0,821+0,777}{3} \right\} \times 100$$

$$\text{Persentase akhir} = 79,9\%$$

Rumus (3.2)

*ket:

TkD1 = Total kalimat pada token kalimat dokumen 1

TkD2 = Total kalimat pada token kalimat dokumen 2

3.4.3 Desain Aplikasi Sistem

Dalam aplikasi ini user interface dirancang untuk memudahkan user dalam penggunaan aplikasi. Berikut adalah rancangan *user interface* aplikasi:



Gambar 3.4 User Interface 1

1. *Upload*: hanya untuk mengunggah file yang berektensi txt, doc, dan docx
2. *Threshold* : untuk menentukan nilai *threshold*
3. *Bandingkan* : untuk membandingkan kemiripan isi file



Gambar 3.5 *User interface 2*

1. Dokumen 1 dan dokumen 2 : dokumen uji yang dibandingkan
2. Kemiripan (%) : Presentase kemiripan isi file
3. Reset : Mengulang uji coba dengan file berbeda
4. Nilai kemiripan diurutkan dari presentase kesamaan file paling besar.

3.4.4 Pengujian Aplikasi

3.4.4.1 Kasus Uji Coba

1. Pada contoh kasus 1, dokumen yang dibandingkan adalah dokumen yang memiliki isi teks yang sama, namun telah dirubah struktur kalimatnya. Berikut adalah dokumen yang digunakan pada contoh kasus 1:

- a. Dokumen A

Manfaat tomat (*Solanum lycopersicum*) untuk tubuh dapat diketahui dari berbagai publikasi ilmiah. Misalnya, Edward Giovannucci di dalam publikasi berjudul “Tomato Products, Lycopene, and Prostate Cancer: A Review of the Epidemiological Literature di American Society for Nutritional Sciences”, 2005 berpendapat bahwa masih ada kontroversial seputar manfaat tomat dalam mencegah berbagai penyakit salah satunya kanker prostat. Manfaat tomat sebagai anti kanker prostat ini disebabkan oleh adanya kandungan lycopene dalam tomat. Lycopene juga terdapat di berbagai produk olahan tomat dan variasinya, seperti: pizza, sup tomat, kecap, jus, salad, saus spaghetti, salsa, pasta tomat. Berbagai produk olahan tomat ini merupakan sumber lycopene yang bioavailability-nya lebih baik daripada buah tomat segar.

- b. Dokumen B

Manfaat tomat sebagai anti kanker prostat ini disebabkan oleh adanya kandungan lycopene dalam tomat. Lycopene juga terdapat di berbagai produk olahan tomat dan variasinya, seperti: pizza, sup tomat, kecap, jus, salad, saus spaghetti, salsa, pasta tomat. Berbagai produk olahan tomat ini merupakan

sumber lycopene yang bioavailability-nya lebih baik daripada buah tomat segar. Manfaat tomat (*Solanum lycopersicum*) untuk tubuh dapat diketahui dari berbagai publikasi ilmiah. Misalnya, Edward Giovannucci di dalam publikasi berjudul “Tomato Products, Lycopene, and Prostate Cancer: A Review of the Epidemiological Literature di American Society for Nutritional Sciences”, 2005 berpendapat bahwa masih ada kontroversial seputar manfaat tomat dalam mencegah berbagai penyakit salah satunya kanker prostat.

2. Contoh kasus² membandingkan dokumen yang memiliki makna yang mirip dan telah dirubah struktur kata dalam kalimatnya. Berikut adalah dokumen yang digunakan pada kasus 2:

- a. Dokumen A

Media adalah segala bentuk dan saluran yang digunakan untuk menyampaikan informasi atau pesan. Kata media berasal dari kata latin, merupakan bentuk jamak dari kata “medium”. Secara harfiah kata tersebut mempunyai arti "perantara" atau "pengantar", yaitu perantara sumber pesan (*a source*) dengan penerima pesan (*a receiver*). Jadi, dalam pengertian yang lain, media adalah alat atau sarana yang dipergunakan untuk menyampaikan pesan dari komunikator kepada khalayak.

- b. Dokumen B

Media adalah saluran yang digunakan untuk memberikan pesan atau informasi. Kata media berasal dari kata latin, yaitu bentuk jamak dari kata “medium”. Secara harfiah kata tersebut memiliki arti "perantara" atau

"pengantar", yaitu sebagai sumber pesan (*a source*) dengan penerima pesan (*a receiver*). Dengan demikian, media dapat dikatakan sebagai sarana atau alat yang digunakan untuk memberikan pesan dari komunikator kepada seluruh masyarakat.

