BAB 1: PENDAHULUAN

Bab ini akan membehas mengenai *overview* tentang aplikasi yang akan dikembangkan, hal yang dibahas termasuk latar belakang pemilihan judul, tujuan pengembangan aplikasi, penjelasan singkat mengenai konsep aplikasi dan struktur penulisan laporan. Selain itu di bab ini juga dibahas mengenai batasan – batasan aplikasi yang akan dikembangkan serta penjelasan istilah – istilah khusus yang muncul dalam laporan ini.

1.1. Latar Belakang

Saat ini perkembangan *internet* telah menjadikan *web* sebagai salah satu sumber informasi. Walaupun demikian banyaknya halaman *web* yang terdapat di *internet* sehingga *browsing* secara manual terkadang membuat beberapa informasi penting terlewatkan. Sedangkan melakukan *save* pada halaman - halaman tersebut juga tidak efektif, mengingat banyaknya *file* yang harus di *save*.

Karena alasan di atas maka penulis memilih untuk mengembangkan aplikasi *web* crawler yaitu sebuah aplikasi untuk melakukan penelusuran (*browsing*) halaman - halaman *web* secara otomatis.

1.2. Tujuan

Tujuan dari proyek tugas akhir ini adalah mengembangkan aplikasi *web crawler* bertipe *archiver crawler* menggunakan pendekatan *object oriented programming* dengan bahasa pemrograman *Java*.

1.3. Gambaran Umum

Aplikasi ini akan membaca suatu *URL* dari halaman *web* dan akan men*download file* yang terdapat pada *web server* tersebut dan menyimpannya di *local storage*. Program ini juga akan melakukan *parsing hyperlink* untuk meniru struktur *file* dan *folder* di *web server*, agar *file - file* yang sudah di*download* dapat di *browse* secara *offline*.

Aplikasi ini juga akan memiliki fitur untuk mendeteksi adanya *broken link* pada struktur *web* target juga dapat membuat gambaran *site map* dari *web* target dalam bentuk *tree*.

Selain hal - hal diatas, konfigurasi *default* aplikasi akan mematuhi *rules* dari *file robots.txt* yang mengatur *file* dan *folder* apa saja yang tidak diperbolehkan diakses di *web server*, tetapi *behaviour* ini dapat diubah karena aplikasi juga menyediakan opsi untuk tidak membaca isi *robots.txt* (*ignore robots rule*).

1.4. Pembatasan Masalah

Batasan - batasan pada pengembangan aplikasi ini adalah:

- *Link depth* yang dapat ditelusuri oleh aplikasi dapat dikonfigurasi oleh pengguna dan nilainya dibatasi maksimum 10. Ini dilakukan untuk membatasi jumlah *file* yang dapat diambil oleh *crawler*.

- Aplikasi tidak akan melakukan proses *crawling* untuk *link* mengarah ke *web server* atau *domain* yang berbeda dengan *reference URL*. Alasannya untuk membatasi hanya *site* target saja yang akan di *crawl*, tanpa pembatasan seperti ini aplikasi akan melakukan proses *crawling* ke seluruh *site* di *internet*,

- Aplikasi tidak akan melakukan *parsing* pada *link* yang *script generated.* Alasannya karena *link* yang *script generated* tidak mengarah langsung ke suatu *file HTML* atau *image* (*direct link*).

- Aplikasi ini akan dikembangkan untuk memiliki kemampuan deteksi *link* yang telah diproses, hal ini dilakukan untuk menghindari *black hole* (*infinite crawling*),

- Aplikasi ini tidak dapat melakukan proses *crawling* pada halaman *web* yang menggunakan sistem autentifikasi atau *session*. Hal ini karena aplikasi tidak mengimplementasikan fitur *HTTP cookies* yang terdapat pada kelas *java.net.HTTPCookie*,

- Sesi *TCP/IP* secara simultan ke *web server* target dibatasi maksimum 1000 koneksi. Hal ini dilakukan untuk membatasi beban koneksi yang dibuka oleh aplikasi ke *web server*, juga mengingat keterbatasan *resource processor* dan memori di mesin lokal yang menjalankan aplikasi (pada aplikasi ini setiap sesi *TCP/IP* akan dijalankan pada satu *Thread* tersendiri).

1.5. Detail Teknis

Berikut adalah beberapa detail teknis dari aplikasi ini:

- *Parsing* pada *file HTML* akan menggunakan *library* HTMLParser versi 1.6, pembuatan *site map* akan menggunakan kelas FileTreeNode dari *library* Flaminggo.
- Koneksi *internet* ke *web server* menggunakan *TCP/IP* yang implementasinya menggunakan *package java.net* dari *JDK*,

- Bersifat *multi-threaded*, berarti aplikasi ini dapat membuka lebih dari satu sesi koneksi *TCP/IP* ke *web server* target pada suatu waktu,
- Dapat berjalan dengan baik pada semua *operating system* yang mendukung *Java SE* dengan *JRE* minimal versi 1.6.0.

1.6. Definisi dan Singkatan

Berikut adalah penjelasan dari singkatan - singkatan maupun istilah yang terdapat pada laporan ini:

- *JRE (Java Runtime Environment)*

  Merupakan *software* yang memungkinkan aplikasi *Java* berjalan pada berbagai *operating system* maupun *hardware*, berperan menjadi perantara antara kode program dan *operating system*. Aplikasi ini menggunakan *Java SE* (*Standard Edition*), yaitu *java runtime* untuk aplikasi *desktop*.

- *JDK (Java SE Software Development Kit)*

  Compiler beserta kumpulan *library* yang dikembangkan oleh Sun Microsystem untuk membuat aplikasi *desktop* berbasis *Java*. Aplikasi *web crawler* ini dikembangkan menggunakan *JDK* versi 1.6.0.

- *Web Crawler*

  Aplikasi yang berfungsi untuk menelusuri halaman - halaman *web* secara otomatis. Perbedaan *web crawler* dengan sebuah *search engine* atau *bot* adalah pada kemampuan *information retrieval*. *Search engine* merupakan sebuah *web crawler* dengan kemampuan mengambil isi dan membuat *index* dari halaman - halaman *web* yang ditelusurinya.

Aplikasi yang dikembangkan merupakan sebuah *archiver crawler*, yaitu *web crawler* yang berfungsi melakukan proses *crawling* dan menyimpan *file - file web* hasil *crawling* tersebut ke *local storage*. *Archiver crawler* juga dapat melakukan pengubahan *link* dari *file - file* hasil *crawling* atau meniru struktur *file* dan *folder* yang terdapat di *web server*, sehingga halaman - halaman *web* yang telah disimpan dapat di *browse* secara *offline*.

- *Link Depth / URL Depth*

  Karena halaman *web* memiliki struktur yang menyerupai *tree*, nilai ini merupakan representasi dari berapa dalam lokasi sebuah *file* atau *folder* pada sebuah *web server*. Dalam bekerja, sebuah *web crawler* akan mempunyai nilai *link depth* maksimum sebelum proses *crawling* dinyatakan selesai atau dihentikan.

- *Black Hole*

  Istilah untuk halaman *web* yang tidak akan pernah selesai jika ditelusuri oleh *web crawler*. *Web crawler* sendiri bekerja dengan menelusuri halaman - halaman pada *web server* sampai *link depth* tertentu tercapai. Pada halaman yang merupakan *black hole*, *link depth* nya selalu tetap (biasanya satu) sehingga *crawler* akan terus menerus mengambil *file* dari *web server* tanpa pernah berhenti. Hal ini biasanya terjadi pada *file - file web* yang *hyperlink*-nya *script generated* atau *hyperlink* dari *server side scripting* seperti jsp, php atau asp.

- *Robots.txt*

  File teks pada *root node web server* yang berisi daftar *file - file* atau *folder* tertentu yang dilarang diakses oleh program - program penjelejah

halaman *web* seperti *web crawler* atau *search engine*. Mematuhi *file* ini merupakan aturan tak tertulis yang dijalani oleh *search engine* maupun *crawler*.

- *Root node*

  Root node merupakan node paling atas atau node pertama pada struktur *file - file* yang terdapat pada sebuah *web server*. Pengertian *root node* pada *web server* sama dengan *root* (/) pada hirarki *system file UNIX* atau struktur *DNS*.

- *TCP/IP (Transport Control Protocol / Internet Protocol)*

  Standar protokol komunikasi yang dipakai oleh *internet* dan jaringan komputer di seluruh dunia. Protokol ini memungkinkan komunikasi secara transparan antar *host* yang berbeda *hardware* maupun *software*.

1.7. Struktur Laporan

Laporan tugas akhir ini terdiri dari enam bab dengan penjelasan sebagai berikut:

- Bab 1, Pendahuluan

  Gambaran umum tentang aplikasi yang dikembangkan pada tugas akhir ini. Menjelaskan latar belakang pemilihan topik, istilah - istilah khusus serta pembatasan cakupan masalah yang akan diimplementasikan pada aplikasi.

- Bab 2, Gambaran Keseluruhan

  Gambaran teknis dari aplikasi, berisi gambaran keseluruhan aplikasi (*overview*) dan spesifikasi produk.

- Bab 3, Desain Perangkat Lunak

  Berisi pemodelan rancangan aplikasi dalam bentuk diagram - diagram beserta penjelasannya.

- Bab 4, Implementasi Desain

  Berisi implementasi pengembangan aplikasi dalam bentuk bahasa pemrograman, dengan mengacu pada diagram yang terdapat di bab 3.

- Bab 5, Pengujian

  Membahas metode pengujian yang dilakukan terhadap aplikasi untuk kemudian mengambil kesimpulan mengenai fungsionalitas produk akhir dan hasil yang dicapai serta perbandingannya dengan gambaran yang tercantum pada bab 1 maupun bab 2.

- Bab 6, Kesimpulan dan Saran

  Berisi kesimpulan dari pengembangan aplikasi serta saran - saran untuk pengembangan lebih lanjut di masa yang akan datang.