Laporan Tugas Besar
IF2124 Teori Bahasa Formal dan Automata

HTML *Checker* dengan *Pushdown Automata* (PDA)

Disusun oleh:

Muhamad Rafli Rasyidin	13522088
Abdullah Mubarak	13522101
Christopher Brian	13522106

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
2023

Daftar Isi

Bab 1: Deskripsi Masalah

HyperText Markup Language (HTML) adalah bahasa markup yang digunakan untuk pembuatan struktur dan tampilan sebuah *website*. HTML digunakan untuk mengatur tampilan elemen-elemen web seperti teks, gambar, media, dan tautan. HTML menggunakan elemen-elemen (*tags*) untuk mengatur isi *website*. Umumnya, suatu blok kode akan memiliki *opening tag* dan *closing tag*, meski terdapat beberapa pengecualian. Contohnya, *tag* `<h1>` harus diikuti dengan `</h1>`. Kesalahan dalam penggunaan *tag* akan menimbulkan *error*. Teknologi *web browser* terbaru seperti Google Chrome dan Microsoft Edge cenderung mengacuhkan *error* pada HTML *website* yang mereka tampilkan, struktur dan sintaks HTML yang benar akan berdampak baik pada SEO (*Search Engine Optimization*), aksesibilitas, *maintenance*, dan *rendering speed*.

Untuk memastikan struktur dan sintaks dari sebuah kode HTML benar, diperlukan suatu program pendeteksi *error* yang dapat memeriksa penggunaan *tag* serta *attribute*. Pada tugas besar ini, kami membuat sebuah HTML *Checker* menggunakan konsep PDA (*Pushdown Automata*) yang diimplementasikan dalam bahasa pemrograman Python. Pengecekan dibatasi pada sejumlah *tag* dan *attribute* dan program akan memberikan hasil berupa *accepted* jika program tidak menemukan *error* dan *rejected* ketika program menemukan *error*.

Bab 2: Landasan Teori

2.1. *HyperText Markup Language* (HTML)

HyperText Markup Language (HTML) adalah suatu bahasa *markup* standar yang dirancang untuk ditampilkan di *web browser*. HTML menggambarkan struktur halaman web secara semantik dan isyarat awal yang disertakan untuk penampilan dokumen. Elemen HTML digambarkan oleh *tag*. Setiap *tag* ditulis menggunakan tanda kurung siku (<>). Beberapa *tag* seperti `<img>` langsung memperkenalkan konten ke dalam halaman. Sementara itu, *tag* lainnya seperti `<h1>` dan `</h1>` mengelilingi dan memberikan informasi tentang teks dokumen dan mungkin menyertakan *tag* lain sebagai sub-elemen. *Tag* seperti `<h1>` disebut sebagai *opening tag*, sementara *tag* seperti `</h1>` disebut sebagai *closing tag*.

Salah satu komponen penting dalam HTML adalah deklarasi tipe dokumen yang memicu *rendering* mode standar. Berikut adalah contoh program sederhana “Hello, World!” dalam HTML.

```
<!DOCTYPE html>
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <div>
      <p>Hello world!</p>
    </div>
  </body>
</html>
```

Teks antara `<html>` dan `</html>` mendeskripsikan halaman web, sedangkan teks antara `<body>` dan `</body>` adalah konten halaman yang terlihat. Teks *markup* `<title>This is a title</title>` mendefinisikan judul halaman web yang ditunjukkan pada *browser tab* dan judul *window*, sedangkan *tag* `<div>` mendefinisikan divisi halaman yang berguna untuk *styling*. Di antara `<head>` dan `</head>`, sebuah elemen `<meta>` dapat digunakan untuk mendefinisikan metadata halaman web. Deklarasi tipe dokumen `<!DOCTYPE html>` digunakan untuk HTML5. Jika deklarasi ini tidak ditambahkan, beberapa *browser* akan memilih mode “*quirks*” untuk *rendering*.

2.2. *Pushdown Automata* (PDA)

Pushdown Automata (PDA) adalah sebuah *finite automata* dengan sebuah memori ekstra yang disebut *stack*. *Stack* memungkinkan PDA untuk mengenali tidak hanya *regular languages*, tetapi juga *Context Free Languages* (CFL).

Secara formal, sebuah PDA dapat didefinisikan oleh sebuah *tuple* yang terdiri atas tujuh komponen sebagai berikut.

1. Kumpulan *state* yang *finite* (Q)
2. *Input alphabet* ($\mathbf{\Sigma}$)
3. *Stack alphabet* (Γ)
4. Fungsi transisi ($\mathbf{\delta}$)
5. *Start state* (q_0 dalam Q)
6. *Start symbol* (Z_0 dalam Γ)
7. Satu atau lebih *final state* (F dalam Q)

Fungsi transisi dalam PDA menerima tiga argumen, yaitu sebuah *state* Q, sebuah simbol input dalam $\mathbf{\Sigma}$ atau $\boldsymbol{\varepsilon}$, dan sebuah *stack symbol* dalam Γ. $\mathbf{\delta}$(q, a, Z) adalah sebuah *set* dari nol atau lebih aksi dalam bentuk (p, $\boldsymbol{\alpha}$) di mana p adalah sebuah *state* dan $\boldsymbol{\alpha}$ adalah sebuah string yang terbentuk dari *stack symbol*. Jika $\mathbf{\delta}$(q, a, Z) memiliki (p, $\boldsymbol{\alpha}$) sebagai salah satu aksinya, maka salah satu hal yang bisa dilakukan oleh PDA dalam *state* q, dengan a di depan *input*, dan Z pada *top* dari *stack* adalah:

1. Ganti *state* menjadi p.
2. Hilangkan a dari depan *input* (a mungkin merupakan $\boldsymbol{\varepsilon}$).
3. Ganti Z pada *top* dari *stack* dengan $\boldsymbol{\alpha}$.

Terdapat bentuk notasi lain untuk PDA, yaitu *instantaneous description* (ID). Sebuah ID merupakan *triple* (q, w, $\boldsymbol{\alpha}$) di mana q adalah *state* sekarang, w adalah *input* yang tersisa, dan $\boldsymbol{\alpha}$ adalah konten *stack*, *top* di kiri.

Jika sebuah ID I dapat berubah menjadi ID J dalam satu gerakan dari PDA, kita dapat menulis $I \vdash J$. Secara formal, (q, aw, X$\boldsymbol{\alpha}$) $\vdash$ (p, w, $\boldsymbol{\beta\alpha}$) untuk setiap w dan $\boldsymbol{\alpha}$, jika $\mathbf{\delta}$(q, a, X) mengandung (p, $\boldsymbol{\beta}$). Kita dapat mengembangkan penggunaan tanda $\vdash$ menjadi $\vdash^*$ yang berarti “nol atau lebih gerakan” menggunakan basis dan induksi sebagai berikut.

- Basis: $I \vdash^* I$.
- Induksi: Jika $I \vdash^* J$ dan $J \vdash^* K$, maka $I \vdash^* K$.

Bab 3: Hasil PDA

P = (Q, Σ, Γ, δ, out, Z, F)

Finite State (Q)

Q = (out, inhtml, inhead, inbody, inh1, inh2, inh3, inh4, inh5, inh6, inbutton, inbody, inform, ina, intitle, inscript, inp, inem, inb, inabbr, instrong, insmall, inimage, intable, intr, intd, inth, inlink, inbr, outlink, inbr, inhr, ininput, fhtml, fhead, fbody, fh1, fh2, fh3, fh4, fh5, fh6, fbutton, fdiv, fform, fa, ftitle, fscript, fp, fem, fb, fabbr, fstrong, fsmall, ftable, ftr, ftd, fth, fbr).

Input Alphabet (Σ)

Σ = (<html, </html>, <body, </body>, <head, </head>, >, id="", class="", style="", <title, <link, <script, <h1, <h2, <h3, <h4, <h5, <h6, <p, <br, <em, <b, <abbr, <strong, <small, <hr, <div, <a, <img, <button, <form, <input, <table, <tr, <td, <th, Rel="", href="", src="", alt="", type="submit", type="reset", type="button", action="", method="GET", method="POST", type="text", type="password", type="email", type="number", type="checkbox").

Stack Alphabet (Γ)

Γ = (Z, a, h, H, b, c, d, e, f, i, z, l, 1, 2, 3, 4, 5, 6, t, y, x, z, s, p, u, k).

Transition Function (δ)

Terdapat 587 baris *transition function* yang kami implementasikan pada PDA.

Bab 4: Implementasi dan Uji Coba

4.1. Implementasi Program Utama

Untuk membangun program ini, ada dua hal utama yang perlu diimplementasikan, yaitu program main.py dan PDA.txt. File PDA.txt berisi *pushdown automata* yang akan digunakan untuk pengecekan file HTML. Program main.py berguna untuk memindahkan *pushdown automata* pada PDA.txt ke dalam program sehingga dapat digunakan dalam pemrosesan HTML. Selain itu, program main.py juga berfungsi untuk melakukan pengecekan pada file HTML.

4.2. Penjelasan Struktur Program

Program main.py berisi beberapa fungsi dan prosedur yang digunakan untuk pemrosesan file HTML. Prosedur ignoreBlank() berfungsi untuk mengabaikan spasi pada file HTML. Prosedur ignoreNewline() berfungsi untuk mengabaikan *newline* pada file HTML. Prosedur ignoreBoth() berfungsi untuk mengabaikan spasi dan *newline* pada file HTML. Fungsi readTag() berfungsi untuk membaca tag-tag yang ada pada file HTML. Fungsi readAttribute() berfungsi untuk membaca atribut yang ada pada tag. Prosedur createPDA() berfungsi untuk memindahkan *pushdown automata* dari PDA.txt ke dalam program.

4.3. Tata Cara Penggunaan Program

Berikut langkah-langkah untuk menggunakan program ini:

1. *Clone repository* ini ke *local repository.*
2. Pindahkan *file* html yang akan diperiksa ke folder *test.*
3. Jalankan program *main* yang ada di folder *src.*
4. Masukkan nama *file* html yang akan diperiksa.
5. Tunggu program menyelesaikan pengecekan. Setelah selesai, Anda dapat melihat hasil pengecekan tersebut. Jika *file* html *valid*, program akan menampilkan tulisan “Accepted”, tetapi jika *file* html tidak *valid*, program akan menampilkan tulisan “*Syntax Error*”.

4.4. Hasil Pengujian

4.4.1. Test Case 1

Input HTML:

```
<html>
  <head>
    <title>Simple Webpage</title>
  </head>
  <body>
    <h1>Hello, World!</h1>
    <p>This is a simple webpage.</p>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case1.html

Accepted, Mas.

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.2. Test Case 2

Input HTML:

```
<html>
  <head>
    <title>Simple Webpage</title>
  </head>
  <body>
    <h1>Hello, World!<h1>
    <p>This is a simple webpage.</p>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case2.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 6: <h1>Hello, World!<h1>

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.3. Test Case 3

Input HTML:

```
<html>
  <body>
    <h1>Hello, World!</h1>
    <p>This is a simple webpage.</p>
  </body>
  <head>
    <title>Simple Webpage</title>
  </head>
```

```
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case3.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 3: <h1>Hello, World!</h1>

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.4. *Test Case 4*

Input HTML:

```
<hmif>
  <head>
    <title>Simple Webpage</title>
  </head>
  <body>
    <h1>Hello, World!</h1>
    <p>This is a simple webpage.</p>
  </body>
</hmif>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case4.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 2: <head>

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.5. *Test Case 5*

Input HTML:

```
<html>
  <body>
    <h1>Hello, World!</h1>
    <p>This is a simple webpage.</p>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case5.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 3: <body>

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.6. *Test Case 6*

Input HTML:

```
<html>
  <head>
    <title>Simple Webpage</title>
  </head>
```

```
  <body>
    <h1>Hello, World!</h1>
    <h2>Welcome to my page</h2>
    <img src=”./welcome.jpeg”>
    <p>This is a <em>simple</em> webpage.</p>

    <div id=”footer” class=”footer”> This is the end of the page </div>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case6.html

Accepted, Mas.

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.7. Test Case 7

Input HTML:

```
<html>
  <head>
    <title>Simple Webpage</title>
  </head>
  <body>
    <!-- Bagian utama web -->
    <h1>Hello, World!</h1>
    <h2>Welcome to my page</h2>
    <hr>
    <img src="./welcome.jpeg" alt="Welcome Banner">
    <p>This is a <em>simple</em> webpage.</p>

    <!-- Custom element -->
    <div id=”footer” class=”footer”> This is the end of the page </div>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case7.html

Accepted, Mas.

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.8. Test Case 8

Input HTML:

```
<html>
  <head>
    <title>Simple Webpage</title>
  </head>
  <body>
    <!-- Bagian utama web -->
```

```
    <h1>Hello, World!</h1>
    <h2>Welcome to my page</h2>
    <img alt="Welcome Banner">
    <p>This is a <em>simple</em> webpage.</p>

    <!-- Custom element -->
    <div id="footer" class="footer"> This is the end of the page </div>
  </body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case8.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 10: <img alt="Welcome Banner">

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.9. Test Case 9

Input HTML:

```
<html>
<head>
  <title>Simple Webpage</title>

</head>
<body>

<h2>HTML Forms</h2>

<form action="/action_page.php" method="TEMBAK">
  <div id="label">First name:</div><br>
  <input type="text" id="fname"><br>
  <div id="label">Last name:</div><br>
  <input type="text" id="lname"><br><br>
  <button type="submit">Submit</button>
</form>

<p>If you click the "Submit" button, the form-data will be sent to a page
called "/action_page.php".</p>

</body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case10.html
Syntax Error.
Terjadi kesalahan ekspresi pada line 11: <form action="/action_page.php" method="TEMBAK">

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.10. Test Case 10

Input HTML:

```
<html>
<head>
  <title>Simple Webpage</title>

</head>
<body>

<h2>HTML Forms</h2>

<form action="/action_page.php" method="POST">
  <h5 class="label">First name:</h5><br>
  <input type="text" id="fname"><br>
  <h5 class="label">Last name:</h5><br>
  <input type="text" id="lname"><br><br>
  <button type="submit">Submit</button>
</form>

<p>If you click the "Submit" button, the form-data will be sent to a page
called "/action_page.php".</p>

</body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case9.html

Accepted, Mas.

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

4.4.11. Test Case 11

Input HTML:

```
<html>
<head>
  <title>Simple Webpage</title>
  <script>
    document.getElementById("demo").innerHTML = "Hello JavaScript!";
  </script>
</head>
<body>

<h1>The script element</h1>

<p id="demo"></p>

</body>
</html>
```

Output program:

```
PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
ah/Semester 3/IF2124 Teori Bahasa Formal dan Otomata/Tubes/Tubes-TBFO/src/main.py"
Masukkan nama file: test_case11.html

Accepted, Mas.

PS C:\Users\Rafli\Documents\Miruu's Stuffs\Kuliah\Semester 3\IF2124 Teori Bahasa Formal dan Otomata\Tubes\Tubes-TBFO\src>
```

Bab 5: Deliverables

Link repository tugas besar : https://github.com/MRafliRasyiidin/Tubes-TBFO.git

Link diagram :
https://www.figma.com/file/dpsJDep5tRh6V9Wb7iJ7sQ/HTML-Parser-PDA-Diagram?type=whiteboard&node-id=0%3A1&t=ixzY9srvGbW6TszA-1

Bab 6: Pembagian Tugas

Nama	NIM	Tugas
Muhamad Rafli Rasyiidin	13522088	Membuat main.py
Abdullah Mubarak	13522101	Membuat PDA
Christopher Brian	13522106	Membuat diagram

Bab 7 : Daftar Pustaka

"HTML". https://en.wikipedia.org/wiki/HTML. Accessed 26/11/2023.

Hopcroft, John E., Rajeev Motwani, Jeffrey D. Ullman. 2003. Introduction to Automata Theory, Languages, and Computation. Addison Wesley. Page: 225-260.