HTML parser

on C++

Nikita Bakaev

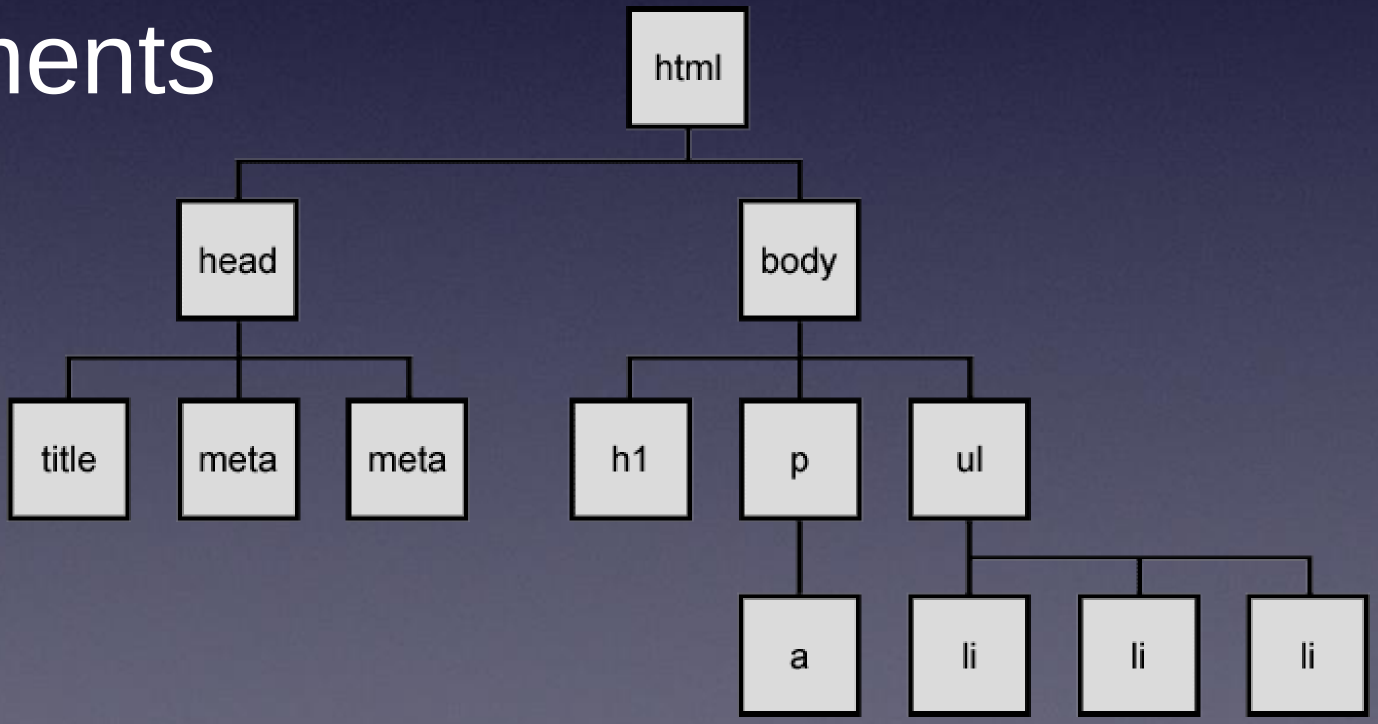
Parsing

- Process
- analysing a string of symbols
- according to the rules

DOM

- Contains elements
- relations: child, parent, subling
- easy selection elements

$.(“html body .btn”)



libXML, gumbo-parser

What can help

Problem

- Academic work
- No external libs likes <regex> , curl

Regular expressions

- Find matches, substrings

- Is input an integer?

```
string input;
    regex integer("(\\+|-)?[[:digit:]]+");

if(regex_match(input,integer))
            cout<<“yes"<<endl;
```

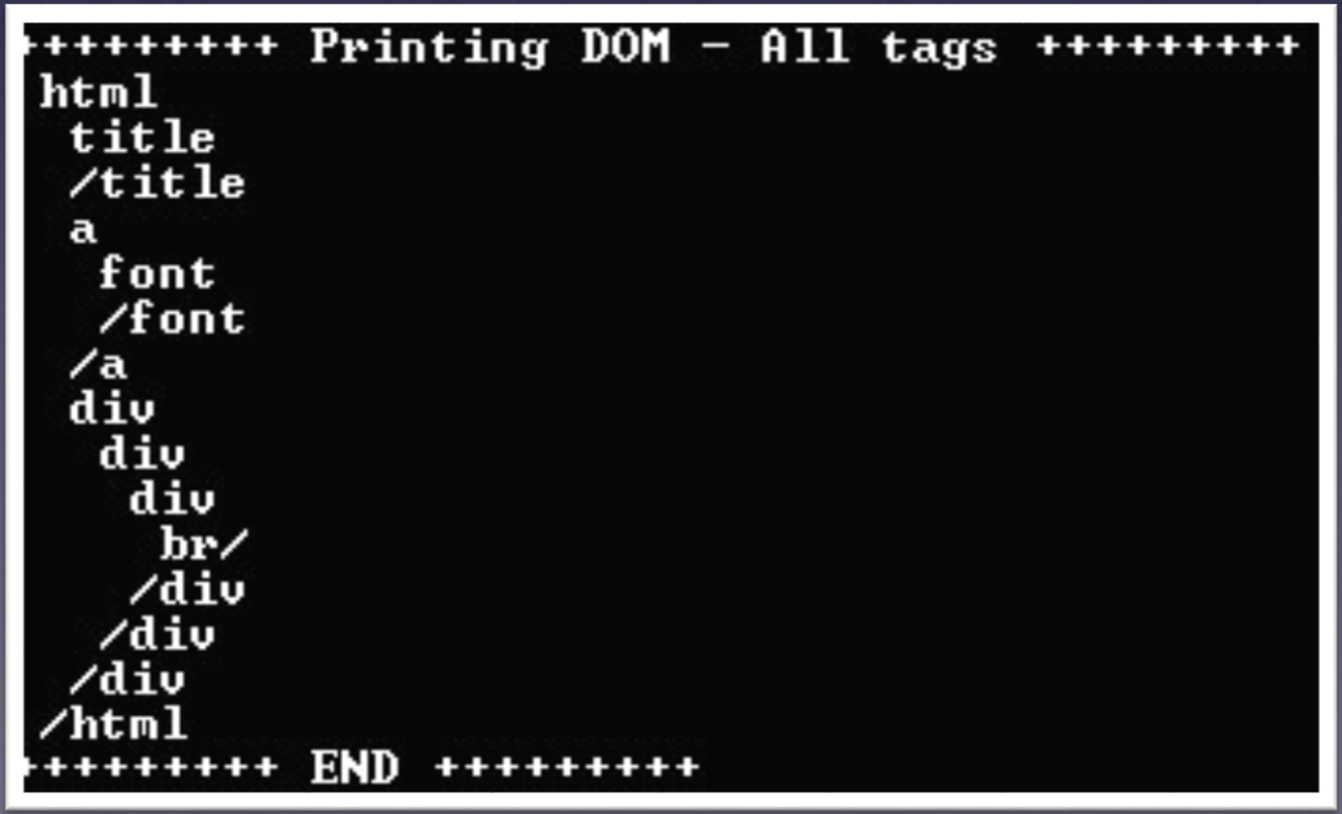
Need to analyze every symbol in a HTML

Where are we

- Tag name
- tag location
- tag attributes
- relations between tags
- space between tags

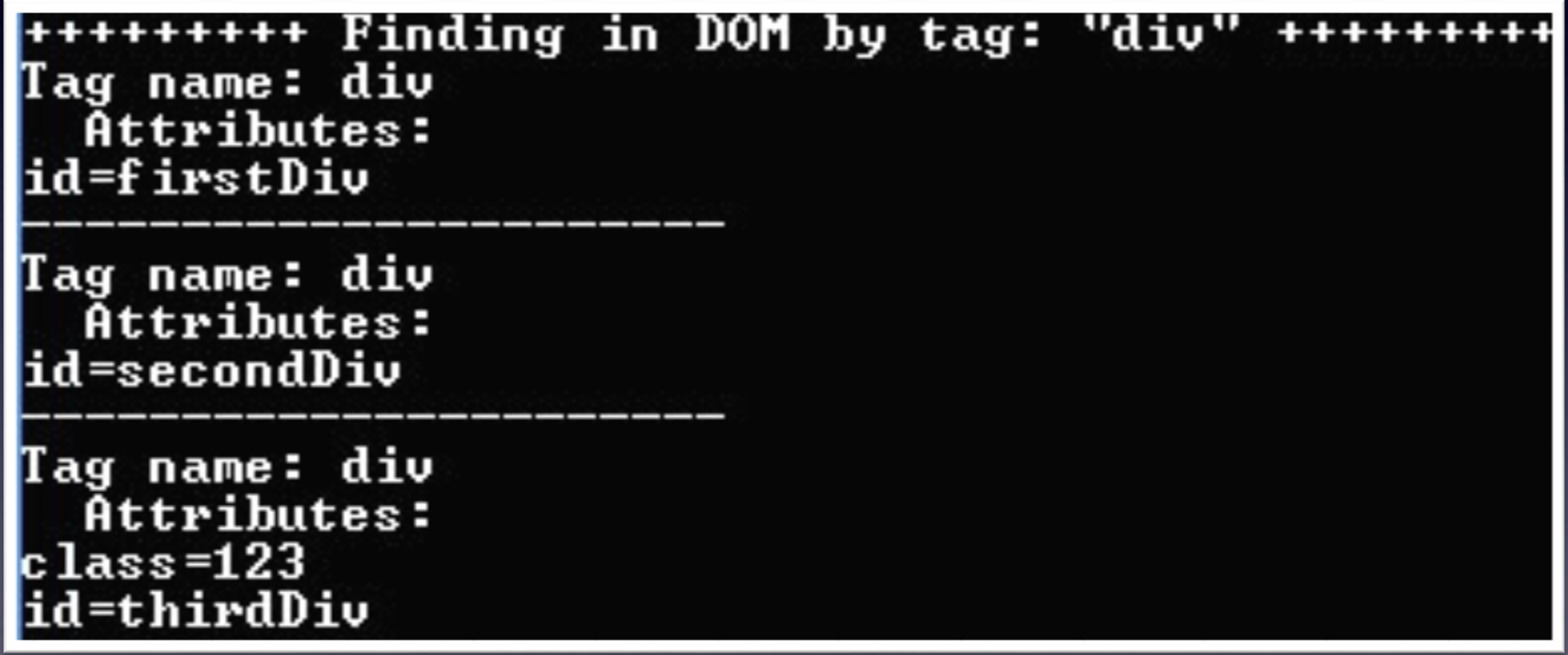
CParser

```
<html>
    <title>myTitle</title>
    hello page begin
        <a href="localhost">
            <font color="red">LocalHost</font>
            LinkName</a>
    <div id="firstDiv">
        Some interesting text inside first div
            <div id="secondDiv">
                Some text inside second div
                <div class="123" id="thirdDiv">
                    text inside third div<br/>text new line
                </div>
            </div>
        </div>
    hello end
</html>
```



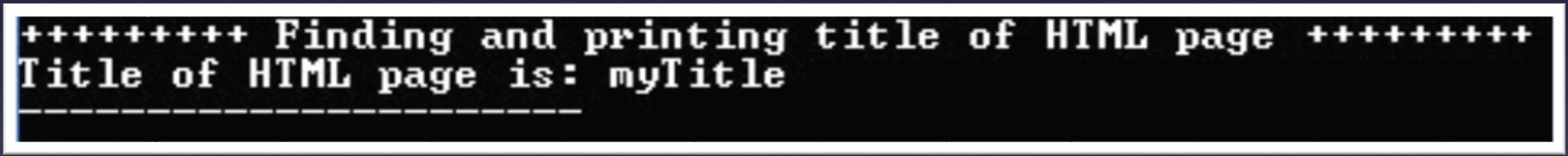
```
+++++++++ Printing DOM - All tags +++++++++
html
 title
 /title
 a
  font
  /font
 /a
 div
  div
   div
    br/
   /div
  /div
 /div
/html
+++++++++ END +++++++++
```

Find element



```
++++++++++ Finding in DOM by tag: "div" +++++++++
Tag name: div
  Attributes:
id=firstDiv
----------------------
Tag name: div
  Attributes:
id=secondDiv
----------------------
Tag name: div
  Attributes:
class=123
id=thirdDiv
```

Print inner text



```
+++++++++ Finding and printing title of HTML page +++++++++
Title of HTML page is: myTitle
----------------------
```

Errors in HTML

- You can find error in HTML (see in text of academic work)

That’s all!

Thank’s ☺

Questions ???