Parser

PARSER–LEXER COMBINATION:

1. **Lexical Analysis(词法分析)**

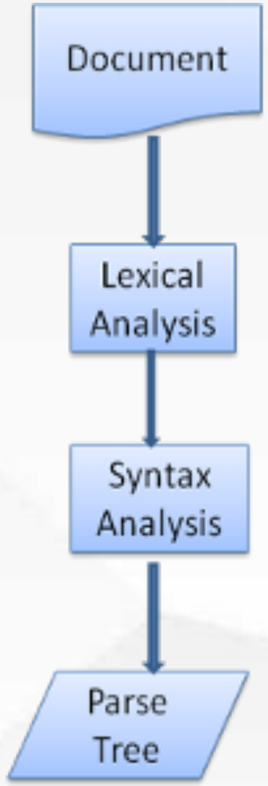
   — The process of breaking the input into tokens. Tokens are the language vocabulary: the collection of valid building blocks.

   — the Lexer: Responsible for breaking the input into valid tokens and it knows how to strip irrelevant characters like white spaces and line breaks.

2. **Syntax Analysis(语法分析)**

   — Syntax analysis is the applying of the language syntax rules.

   — the Parser: Responsible for constructing the parse tree by analyzing the document structure according to the language syntax rules.



MAIN FLOW OF PARSE

AN EXAMPLE:

1. **consider an example: 2 + 3 - 1**
2. **Vocabulary and RE**

   Vocabulary is usually expressed by regular expressions.

   - INTEGER: 0|[1-9][0-9]*
   - PLUS: +
   - MINUS: -

3. **Syntax and BNF**

   Syntax is usually defined in a format called BNF.A language can be parsed by regular parsers if its grammar is a context free grammar that can be entirely expressed in BNF.

   - expression := term operation term
   - operation := PLUS | MINUS
   - term := INTEGER | expression

Two types of parsers:

— top down parsers and bottom up parsers.

— Top down parsers examine the high level structure of the syntax and try to find a rule match.

— Bottom up parsers start with the input and gradually transform it into the syntax rules, starting from the low level rules until high level rules are met.

HTML Parser

1. **What's the job of HTML parser?**

   Parse the HTML markup into a parse tree.

2. **How does HTML parser Defined?**

   — HTML cannot easily be defined by a context free grammar to be parsed using the regular top down or bottom up parsers. The reasons are:

   - The forgiving nature of the language.
   - The fact that browsers have traditional error tolerance to support well known cases of invalid HTML.
   - The parsing process is reentrant(可折返的). In HTML, dynamic code (such as script elements containing document.write() calls) can add extra tokens, so the parsing process actually modifies the input.

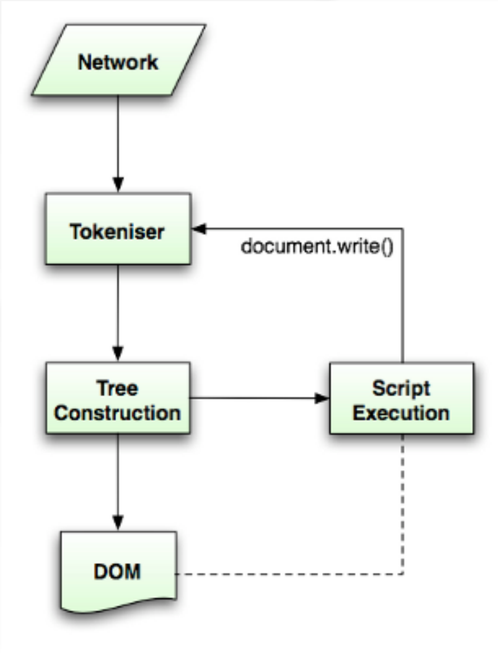
   — A formal format for defining HTML is DTD.

   — HTML DTD(Document Type Definition):

   The format contains definitions for all allowed elements, their attributes and hierarchy. There are a few variations of the DTD. The strict mode conforms solely to the specifications but other modes contain support for markup used by browsers in the past. The purpose is backwards compatibility with older content.

3. **How does HTML parser work?**

   - Tokenization is the lexical analysis, parsing the input into tokens. Among HTML tokens are start tags, end tags, attribute names and attribute values.
   - The tokenizer recognizes the token, gives it to the tree constructor, and consumes the next character for recognizing the next token, and so on until the end of the input.
   - DOM(Document Object Model):The output tree (the "parse tree") is a tree of DOM element and attribute nodes. It is the object presentation of the HTML document and the interface of HTML elements to the outside world like JavaScript.



THE HTML PARSING ALGORITHM