3.Parsing and DOM tree construction

Parser

PARSER–LEXER COMBINATION:

1. **Lexical Analysis(词法分析)**

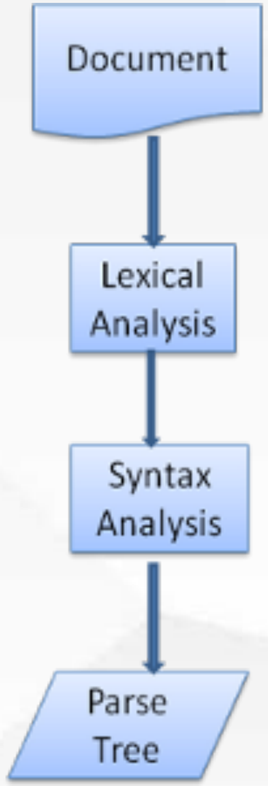
   — The process of breaking the input into tokens. Tokens are the language vocabulary: the collection of valid building blocks.

   — 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.