

Parser

Um "parser" é um programa ou parte de um programa de computador que é usado para analisar e interpretar texto, normalmente em um formato específico, como uma linguagem de programação, uma linguagem de marcação (como HTML ou XML) ou um formato de dados estruturado (como JSON).

O objetivo principal de um parser é transformar o texto em uma estrutura de dados que possa ser facilmente manipulada e compreendida pelo programa. Isso envolve a identificação de elementos individuais no texto (como palavras-chave, símbolos, tags ou estruturas de dados) e a construção de uma representação interna desses elementos que o programa pode usar para realizar tarefas específicas.

Por exemplo, em uma linguagem de programação, um parser pode ler o código-fonte do programa e convertê-lo em uma árvore de sintaxe, o que permite ao compilador ou interpretador entender a estrutura e a semântica do código para compilação ou execução.

Em resumo, um parser desempenha um papel fundamental na interpretação e processamento de texto estruturado em formatos específicos, tornando-o acessível e utilizável para programas de computador.

```
] import re

def parse_expression(expression):
    # Remove espaços em branco
    expression = expression.replace(" ", "")

    # Usando expressões regulares para dividir a expressão em números e operadores
    tokens = re.findall(r'(\d+|\+|\-)', expression)

    stack = []
    result = 0
    operator = '+'

    for token in tokens:
        if token.isdigit():
            num = int(token)
            if operator == '+':
                result += num
            elif operator == '-':
                result -= num
            else:
                operator = token

    return result

# Exemplo de uso do parser
expressao = "10 + 5 - 3"
resultado = parse_expression(expressao)
print("Resultado:", resultado)
```

Neste exemplo, o parser recebe uma expressão matemática como entrada e a divide em números e operadores. Em seguida, ele percorre a lista de tokens e executa as operações de adição e subtração, mantendo o resultado atual em uma variável chamada "result". O parser lida com números inteiros e operadores de adição e subtração simples.

Sobre o Querido Diário

o Querido Diário quer centralizar os dados e metadados de diários oficiais de municípios do Brasil. A gente mapeia onde as cidades publicam os diários oficiais e o modo como fazem isso. Também desenvolvemos raspadores para coletar esses arquivos diariamente em cada site de município cadastrado no projeto. Além disso, buscamos deixar o conteúdo dos diários coletados em um formato aberto, uma vez que as prefeituras normalmente publicam em formato fechado (pdf). Os objetivos principais são justamente centralizar as informações dos diários das cidades (que não seguem um padrão), tornar isso acessível por meio do buscador (site do Querido Diário) e utilizável por outras ferramentas / iniciativas / projetos (por meio da API) para os mais diversos fins.

Parser no Querido Diário

Parser, em computação, é um conceito que aparece em várias áreas, daí uma resposta mais robusta, eu sugeriria pesquisar (ainda que mais ou menos na mesma ideia do que vou dizer em seguida)... mas pro QD, parser é uma etapa de analisar texto, explorar a estrutura dele e processar o que é relevante no meio do texto. No caso de raspadores, é como o coração do raspador, responsável achar metadados (data, numero da edição, etc) no meio do conteúdo de uma página web sendo raspada. Um exemplo tá aqui: https://github.com/okfn-brasil/querido-diario/blob/main/data_collection/gazette/spiders/am_manaus.py. (poderia ser literalmente qualquer outro raspador, todos eles tem um parse())

Parser é sempre algo assim, analisar dados textuais e explorar seus padrões com algum objetivo. Dei um exemplo em raspadores, mas se você tiver explorando padrões no conteúdo do diário em si (não no site onde ele está), isso também seria criar um parser pro texto do diário.

Juliana Trevine – Administradora e Criadora do QD