UB Social

CS50AI with Python

11.Parser

Mateus Schwede

Problemática

- Escreva IA para analisar frases e extrair sintagmas nominais;
- Parsing é tarefa comum em Processamento de Linguagem Natural (NLP), usada para determinar a estrutura de frase;
- Neste problema, será usada gramática livre de contexto (CFG) para analisar sentenças em inglês;
- Em CFG, aplicam-se regras de reescrita para transformar símbolos não-terminais em outros símbolos, até gerar sentença completa com símbolos terminais (palavras);
- Regras:
    - S -> N V: sentença é composta de substantivo seguido de verbo;
    - N -> "Holmes" e V -> "sat" gera frase "Holmes sat.".

Problemática

- Frases nominais podem ser mais complexas, como:
  - "my companion", "a country walk", "the day before Thursday";
- Para lidar com isso, criam-se regras como:
  - NP -> N | Det N: frase nominal (NP) pode ser apenas substantivo ou determinante seguido de substantivo;
- Símbolo | representa alternativas de reescrita;
- Para utilizar frases nominais como sujeito de sentença, regra S -> N V precisará ser modificada para usar NP no lugar de N;
- Regras adicionais também podem ser necessárias para lidar com frases nominais mais complexas.

Instruções

- Baixe código de https://cdn.cs50.net/ai/2023/x/projects/6/parser.zip e descompacte-o;
- Dentro do diretório parser, execute “pip3 install -r requirements.txt” para instalar dependência nltk (processamento de linguagem natural).

Funcionamento

- Arquivos na pasta 'sentences' contêm frases em inglês que deverão ser analisadas pelo parser a ser implementado;
- parser.py contém regras de gramática livre de contexto (CFG);
- Regras de terminais já estão definidas na variável TERMINALS, como:
  - Adj: adjetivos;
  - Adv: advérbios;
  - Conj: conjunções;
  - Det: determinantes;
  - N: substantivos;
  - P: preposições;
  - V: verbos.

Funcionamento

- Variável NONTERMINALS define regras de não-terminais, inicialmente com apenas regra:
  - S -> N V, que permite sentenças simples como "Holmes arrived.".
- Para analisar frases mais complexas, é necessário expandir regras em NONTERMINALS;
- Função main:
  - Lê frase de arquivo ou entrada do usuário;
  - Pré-processa frase com função preprocess;
  - Analisa frase segundo regras de CFG;
  - Exibe árvore(s) sintática(s) resultante(s);
  - Exibe trechos da frase identificados como "noun phrase chunks" usando função np_chunk.
- Funções preprocess e np_chunk devem ser implementadas, assim como as regras em NONTERMINALS.

Especificações

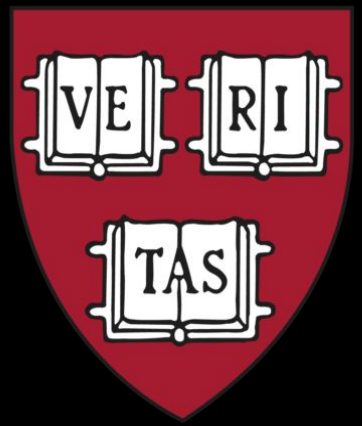
- Função preprocess(sentence):
  - Recebe string (frase em inglês);
  - Deve retornar lista de palavras em minúsculas, tokenizadas usando nltk.word_tokenize;
  - Excluir palavras que não tenham letras (ex.: números ou pontuação como ".", "28").
- Variável NONTERMINALS:
  - Deve conter regras de gramática livre de contexto (CFG) para analisar todas frases da pasta 'sentences/';
  - Deve começar com regra inicial S -> ... (representando sentença completa);
  - Pode incluir quantas regras e símbolos não-terminais forem necessários;
  - Use NP para representar "noun phrases";
  - Pode usar | para alternativas e cada regra deve estar em linha separada;
  - Pode remover regra inicial S -> N V se desejar.

Especificações

- Função np_chunk(tree):
  - Recebe árvore sintática (nltk.tree.Tree) com rótulo S;
  - Deve retornar lista de subárvores com rótulo NP, onde:
    - Cada NP não deve conter outra NP dentro de si (i.e., é menor unidade de NP possível).
  - Deve manipular objetos nltk.tree.Tree.

Submissão

- Visual Studio Code online: https://cs50.dev
- Testar precisão da lógica do algoritmo: check50 ai50/projects/2024/x/parser
- Testar estilização do código: style50 parser.py
- Para submissão:
  - Em https://submit.cs50.io/invites/d03c31aef1984c29b5e7b268c3a87b7b, entre com GitHub e autorize CS50;
  - Instale pacote Git, Python 3 (e pip), instalando pacotes: pip3 install style50 check50 submit50
  - Submeta o projeto: submit50 ai50/projects/2024/x/parser
- Verificar avaliação: https://cs50.me/cs50ai.



UB Social

Mateus Schwede
HBS ID 202400167108 - DCE ID @00963203