Analisador Léxico de Vocabulário Básico de Recursos Naturais e Meio Ambiente

Danilo Yudi Futata Kasssuya
Universidade Estadual de Londrina Arapongas, Brasil
danilo.yudi.futata@uel.br

Felipe Alves Barusso
Universidade Estadual de Londrina, Londrina, Brasil
felipe.barusso@uel.br

Guilherme Henrique Gonçalves. Silva
Universidade Estadual de Londrina Londrina, Brasil
guilherme.henriique.silva@uel.br

ABSTRACT

In computer science, lexical analysis is the process of converting a string of characters (as in a computer program or web page) into a sequence of tokens (strings with an assigned meaning and therefore identified). This study aims to create a lexical analyzer where token sequences represent the basic vocabulary of natural resources and the environment. The Basic Vocabulary of Natural Resources and Environment is aimed at all those who are dedicated to this scientific field, serving a wide range of users from different areas of knowledge and interests. And it brings together the entries considered the most relevant in environmental studies and which were selected among those in most current use. To create the lexical analyzer, a Hash Table was implemented, where each Hash Table item has two parts, the key and the object. The key is the word itself to be searched for in the lexical analyzer and the object contains the information for that word. In conclusion, it was possible to observe that the creation of this lexical analyzer allows the spread of concepts in an agile way, facilitating sometimes time-consuming research. And understand the functioning of a lexical analyzer implemented through a specific Hash Table with a natural language dictionary.

RESUMO

Na ciência da computação, análise léxica é o processo de converter uma sequência de caracteres (como em um programa de computador ou página da web) em uma sequência de tokens (strings com um significado atribuído e, portanto, identificado). Este estudo tem como objetivo a criação de um analisador léxico onde as sequências de tokens são representação do vocabulário básico de recursos naturais e meio ambiente. O Vocabulário Básico de Recursos Naturais e Meio Ambiente está voltado a todos aqueles que se dedicam a este campo científico, atendendo a um leque amplo de usuários de diferentes áreas de conhecimento e interesses. E reúne os verbetes considerados os mais relevantes em estudos ambientais e que foram selecionados entre os de uso mais corrente. Para a criação do analisador léxico foi realizada a implementação de uma Hash Table, onde cada item da Hash Table tem duas partes, a chave e o objeto. A chave é a própria palavra a ser pesquisada no analisador léxico e o objeto contém as informações daquela palavra. Na conclusiva foi possível observar que a criação desse analisador léxico permite difundir conceitos de forma ágil, facilitando pesquisas por vezes demoradas. E entender o funcionamento de um analisador léxico implementado através de uma Hash Table em específico com um dicionário de língua natural.

Palavra-chave: Análise léxica. Vocabulário. Meio ambiente. Hash Table.

Palavras-chave ACM
Processamento de Linguagem Natural.

INTRODUÇÃO

Um dicionário tem a função de facilitar o entendimento de novas palavras, mas existem diversos significados que uma palavra pode ter dependendo do contexto, para isso alguns dicionários trabalham especificamente com alguns campos léxicos do conhecimento.

O objetivo do projeto era desenvolver o conhecimento e entendimento dos membros tanto sobre campos léxicos, como os métodos para implementar a ferramenta do dicionário digital. Devido a isso o grupo escolheu função hash disponível no artigo "Visual TAHS: Ferramenta Para Analisar a Eficácia de Buscas das Funções Hash em um Léxico Para Língua Natural".

O dicionário funciona de modo que uma pessoa escreva uma palavra relacionada ao campo léxico e como resultado o dicionário responda com o significado da palavra em questão, sendo sua maior utilidade seria facilitar a comunicação de pessoas de áreas mais diversas, podendo se usar dicionários de diversas áreas diferentes para traduzir expressões específicas da área de maneira simples.

Desenvolvimento

Para o desenvolvimento o grupo precisou estudar tanto maneira de ligar as palavras do dicionário aos seus significados usando funções de procura hash como o campo léxico do meio ambiente, usando esta função hash foi possível criar uma chave ligando as duas informações permitindo consultas rápidas sobre o significado de algumas das palavras implementadas no livro sobre o campo léxico.

O dicionário funciona de maneira simples apenas informando um significado pré determinado para chaves específicas, mas com esse método podemos ver que é possível desenvolver um sistema mais complexo que pode permitir com que um dicionário possa identificar o significa de uma palavra no meio de uma sentença por exemplo e nos dizer o significado dela em seu contexto geral e não apenas em um campo léxico específico

Metodologia

Para a criação do analisador léxico, o grupo implementou uma tabela *hash*. Esta estrutura de dados funciona através da associação de chaves de busca com valores respectivos. Assim, ao utilizar uma palavra do dicionário como chave, podemos realizar uma busca sobre suas propriedades com baixo custo computacional.

O algoritmo desenvolvido cria a tabela *hash*, insere o dicionário nela e em seguida permite ao usuário realizar consultas.

Na etapa de inserção, cada palavra é enviada, junto de suas propriedades, à função denominada "inserir". Nela, o programa calcula o índice em que esta palavra deve ser inserida em um vetor através de uma função denominada "função_hash".

Neste trabalho em particular, adotamos o método *one at a time* criado por Bob Jenkins [1]. Ela realiza uma série de operações *bitwise* para cada letra da chave (neste caso um termo do dicionário) e retorna um índice. O grupo optou por este método pois ele foi desenvolvido para trabalhar com *Strings* (textos).

```
def funcao_hash(chave):
    indice = 0
    i = len(chave) - 1
    while i >= 0:
        indice += ord(chave[i])
        indice += (indice << 10)
        indice ^= (indice >> 6)
        i -= 1
    indice += (indice << 3)
    indice ^= (indice >> 11)
    indice += (indice << 15)
    return indice % TAMANHO
```

Figura 1: **Função *hash one at a time***

Porém, não é possível garantir que o índice gerado seja único para todas as chaves. Por isso, empregamos o conceito de *bucketing*. Cada posição do vetor é uma lista e caso duas palavras sejam inseridas no mesmo índice o programa anexa o termo novo no final da lista.

Por fim, com a flexibilidade da linguagem *python*, podemos inserir objetos complexos no vetor da tabela *hash* com facilidade. Assim, o grupo escolheu inserir uma tupla formada pela própria palavra (chave) e seus atributos.

A experiência do usuário é um menu onde ele pode inserir uma palavra. Se esta palavra for "sair", o programa é finalizado. Caso contrário, é realizada uma busca (através de um processo de calcular o índice similar a etapa de inserção) e as propriedades daquela palavra são exibidas pelo programa.

Resultados

Com uma tabela de tamanho 2048, o programa inseriu 1382 termos do dicionário em 1003 índices distintos. Além disso, o índice com mais objetos teve 5 palavras inseridas.

A média do tempo de execução na etapa de inserção de 1382 termos no dicionário foi de 6,0012 milisegundos.

A média do tempo de execução da etapa de busca foi complexa de obter, devido ao intervalo minúsculo de tempo que uma busca leva. Assim, desenvolvemos um teste para executar 1000 buscas 100 vezes e calcular a média de tempo de 1000 buscas. O resultado obtido foi 0.19977 milissegundos para cada 1000 buscas, demonstrando a eficiência da tabela *hash* em conjunto com a função escolhida.
Todos os cálculos de tempo de execução foram realizados utilizando a biblioteca *time* da linguagem *python*.

CONCLUSÔES

Este trabalho abordou uma análise da criação do analisador léxico a qual é necessária como primeira etapa em Sistemas de Processamento de Linguagem Natural. O grupo escolheu o campo léxico do meio ambiente.

O trabalho passou por várias etapas, desde a aquisição dos dados pela internet, sua manipulação para separar as palavras e armazenar cada informação. Para a realização da análise foi construído um sistema em Python e com a utilização de uma função hash para a criação de chave que liga as informações, com isso, permitindo consultas sobre o significado das palavras que foram implementadas no campo léxico.

Todo esse processo nos forneceu uma grande visão sobre ampliar o leque de abrangência deste vocabulário relacionado ao meio ambiente e difundir conceitos de forma ágil, facilitando pesquisas por vezes demoradas.

Como trabalho futuro também é possível continuar a análise desse corpus em outros níveis de análise da Linguagem Natural, como nas análises sintática e semântica dos textos. Além disso, é possível também realizar a análise da mesma forma em um outro campo léxico e então estudar os resultados de forma comparativa.

REFERENCES

[1] Bob Jenkins, S. 01, & Jenkins, B. (n.d.). *Algorithm alley*. Dr. Dobb's. Retrieved December 2, 2021, from https://www.drdobbs.com/database/algorithm-alley/184410284.