Processamento de Linguagem Natural (PLN) em artigos IEEE

Universidade Estadual de Maringá
Ciência da Computação - Inteligência Artificial I
Prof. Dr. Wagner Igarashi

Ricardo Henrique Brunetto (94182)
Thiago Kira (78750)

Problema

- Identificar os seguintes aspectos de um artigo IEEE aplicando técnica de Processamento de Linguagem Natural.
  - Autor
  - Objetivo
  - Termos mais citados
  - Referências
  - etc

Solução Proposta

- Construir um analisador léxico e um interpretador a partir de uma gramática flexível para a estrutura do artigo, tendo como base o uso de expressões regulares.

Modelagem

- Modelou-se o artigo como uma sequência de regras gramaticais predefinidas.
- Cada regra pode ser composta por outra regra e/ou um determinado padrão de expressão regular.
- O analisador léxico é responsável por identificar os ***tokens*** que compõe as regras gramaticais.
- O interpretador é responsável por compor as regras conforme recebe os ***tokens*** e construir a sintaxe especificada pela gramática.

Gramática

- 28 regras
- 11 tokens

```
Rule 0     S' -> article
Rule 1     article -> header content references
Rule 2     header -> code publication titleandauthor abstract keywords
Rule 3     header -> publication code titleandauthor abstract keywords
Rule 4     code -> YEAR
Rule 5     code -> NUMBER
Rule 6     publication -> IEEE text MONTH YEAR
Rule 7     abstract -> ABSTRACT text INDEX
Rule 8     keywords -> text TERMS text
Rule 9     content -> chapter_seq REFERENCES
Rule 10    chapter_seq -> chapter chapter_seq
Rule 11    chapter_seq -> chapter
Rule 12    chapter -> CHAPTER_MARK ctext
Rule 13    ctext -> YEAR ctext
Rule 14    ctext -> NUMBER ctext
Rule 15    ctext -> REFERENCE_B ctext
Rule 16    ctext -> GENERAL ctext
Rule 17    ctext -> INDEX ctext
Rule 18    ctext -> IEEE ctext
Rule 19    ctext -> <empty>
Rule 20    references -> reference_seq ctext
Rule 21    reference_seq -> REFERENCE_B reference_seq
Rule 22    reference_seq -> text reference_seq
Rule 23    reference_seq -> <empty>
Rule 24    titleandauthor -> text
Rule 25    text -> GENERAL text
Rule 26    text -> code text
Rule 27    text -> GENERAL
Rule 28    text -> <empty>
```

Principais regras

article -> header content references

header -> code publication titleandauthor abstract keywords

publication -> IEEE text MONTH YEAR

abstract -> ABSTRACT text INDEX

keywords -> text TERMS text

content -> chapter_seq REFERENCES

references -> reference_seq ctext

Ambiente e Recursos

- Sistema Operacional Ubuntu 16.04 LTS
- Python 2.7

- PLY - Python Lex Yacc
- PDFMiner
- NLTK

Etapas

1. Converter PDF para texto (PDFMiner)
2. Pré-processamento do texto
    a. Remove/trata caracteres especiais
    b. Remover palavras pouco significantes
3. Processamento de Linguagem Natural
    a. Parser (gramática)
    b. Top 10 termos
4. Salvar o resultado

1 - Conversão PDF-texto

- Biblioteca PDFMiner
- Extrai o texto do PDF mantendo a formatação
  - Quebra de linha
  - Hífens
  - Caracteres desconhecidos
- Utilizou-se o **UTF-8** para codificaçãode caracteres

2 - Pré-processamento

- Biblioteca NLTK
- Remover alguns caracteres especiais
- Quebra a linha de publicação para reconhecimento da gramática

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 10, NO. 6, SEPTEMBER 2016

<->

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 10, NO. 6,

SEPTEMBER 2016

3 - PLN

- Biblioteca PLY
- Aplica o ***parser*** com as regras determinadas no texto pré-processado
- Seciona o artigo em
  - Header
    - Artigo
    - Título/Autor
    - Publicação
  - Content
    - Lista de Chapter
  - References
    - Lista de Reference
- Retorna um objeto Article

Identificação dos termos mais citados

- Identifica cada um dos termos utilizados no artigo e incrementa um contador associado a este termo para cada repetição.
- Utiliza um conjunto pré definido de palavras para filtrar o artigo
- NLTK - stopwords
    - Desconsidera conectivos, conjunções e artigos.

4 - Salvar o resultado

- Salva em um arquivo ***.out***

- Informações consideradas
  - Título
  - Publicação
  - Top 10 termos
  - Lista de capítulos
  - Lista de referências