



UTFPR - CÂMPUS PONTA GROSSA - COCIC

Disciplina de Compiladores – 2016.2

Atividade Extra 02 – Análise Léxica, Tabela de Símbolos e Arquivos

Professor: Gleifer Vaz Alves

- **Regras para Atividade Extra:**

- **Atenção!** Leia todas as regras e informações descritas abaixo.
- Descrição: realizar a atividade listada no documento PDF, sendo que a atividade deve ser entregue via *moodle*.
- Equipe: as atividades podem ser realizadas *individualmente* ou em duplas (com *dois* alunos).
- Penalização & Bonificação: veja detalhes na definição da atividade (adiante) neste documento.
- Apresentação: na data marcada no *moodle*, algumas equipes (sorteadas *em sala de aula*) vão apresentar as suas soluções para a atividade em questão em sala de aula.

Observações:

- * Caso a equipe sorteada (*equipe β*) não apresente a solução do problema e a equipe β tenha recebido bônus, então β perderá (integralmente) o bônus recebido.

1. Descrição Geral

Técnicas de compilação podem ser aplicadas em diferentes cenários, não apenas em cenários relacionados a linguagens de programação.

Atualmente, há diversas pesquisas relacionadas a extração e reconhecimento de informações a partir de textos e dados obtidos em redes sociais. Por exemplo, uma empresa tem muito interesse em descobrir o que estão comentando a respeito de um novo produto lançado no mercado. Esse tipo de pesquisa caracteriza uma técnica da *Mineração de Dados*, especificamente conhecida como **Descoberta de Conhecimento em Textos**. Existe uma preocupação em coletar dados de redes sociais, para então tentar reconhecer informações que possam servir para inferir dados relevantes. Citando o exemplo acima, descobrir se os comentários a respeito do produto são positivos ou negativos.

A partir disso deseja-se implementar um programa que extraia informações de um conjunto de textos e possa retornar um relatório com informações significativas ao usuário.

Observação 1: cabe a cada equipe selecionar e escolher o cenário onde as informações serão obtidas. Por exemplo, textos de uma rede social, textos de um jornal, textos de vários jornais a respeito de um determinado tema, textos de vários blogs a respeito de um determinado tema, etc.

Ademais, cada equipe deverá determinar qual a meta principal deve ser atingida, ou seja qual informação alvo deve ser extraída a partir do cenário selecionado.

Por exemplo, poderíamos selecionar um conjunto de textos que discutem a respeito de um novo lançamento de uma empresa automotiva. Assim, a equipe deve cuidadosamente selecionar esse conjunto de texto, bem como selecionar um conjunto de palavras-chave que poderão apontar que fragmentos dos textos discutem o novo lançamento da referida empresa automotiva.

Observação 2: note que não é necessário implementar um mecanismo inteligente para descoberta de conhecimento. Aqui serão utilizadas apenas as ferramentas léxicas para “filtragem” dos textos e descoberta de uma informação útil. Porém, essa é uma informação inicial que posteriormente seria repassada para uma fase de mineração de dados ou descoberta de conhecimento.

2. Objetivos

Baseado na descrição da implementação de arquivos e **tabela de símbolos** com a ferramenta **Flex**, almeja-se implementar através dessas ferramentas um analisador de padrões em arquivos de texto, o qual deve ter algum resultado concreto a partir de uma análise léxica executada.

Especificamente os seguintes objetivos devem ser atingidos:

- (a) Selecionar um cenário para que sejam extraídas informações consideradas significativas.
- (b) A partir do cenário selecionado, extrair um conjunto de informações (no caso, um conjunto de textos, por exemplo). Há necessidade de que se tenham um número significativo de textos (ou arquivos).
- (c) Implementar um módulo (**Módulo A**) responsável pela extração de informações do conjunto de textos. Esse módulo necessariamente deve ter:
 - Manipulação de Arquivos (entrada/saída).
 - Tabela de Símbolos.
 - Analisador Léxico.
- (d) Implementar um módulo (**Módulo B**) responsável pelo reconhecimento daquelas informações extraídas pelo **Módulo A**.

Ou seja, o **Módulo B** necessariamente deve gerar um conjunto de informações relevantes e significativas que correspondam aos objetivos definidos na descrição do trabalho. Especificamente, em relação ao cenário escolhido na Etapa **a**.

Por exemplo, essas informações relevantes podem apontar quais palavras-chave são frequentes no conjunto de texto pré-selecionado. Retomando o exemplo supracitado, seria possível destacar que certas palavras frequentes no conjunto de texto apontam evidências *positivas* em relação ao produto lançado pela indústria automotiva.

- (e) **Sugestão:** informações estatísticas podem ser necessárias no momento de elaborar o relatório final do trabalho. O uso de dados estatísticos pode auxiliar na apresentação adequada dos resultados.

Note que saída do **Módulo A** servirá como entrada para o **Módulo B**. Ao passo que a saída do **Módulo B** deverá ter as informações relevantes e significativas referente aos resultados finais obtidos pelo programa.

3. Itens de avaliação:

Abaixo a descrição dos itens avaliados.

- (a) Implementação do **Módulo A** (**bônus** \Rightarrow *valor: 0,3*).
(b) Implementação do **Módulo B** (**bônus** \Rightarrow *valor: 0,4*).

Obs.: o bônus só será concedido ao aluno caso ele elabore o relatório e a respectiva defesa do trabalho.

4. Estrutura do documento/relatório:

- (a) Descrição (detalhada) do: cenário escolhido, conjunto de textos e objetivo(s).
(b) Definição da etapa de análise léxica. Utilizar ferramentas para representar as regras léxicas criadas:
i. REGEX: para descrever casos de teste (significativos) que exemplifiquem o comportamento esperado do **Módulo A**.
ii. REGEXPER ou DEBUGGEX: para ilustrar os respectivos diagramas de transição de cada expressão regular do **Módulo A**.
(c) Apresentação dos códigos referentes a:
i. Implementação do gerenciamento dos arquivos.
ii. Implementação da tabela de símbolos.
iii. Implementação das regras e padrões no FLEX.
(d) Análise e descrição dos resultados obtidos através do **Módulo B**.

5. Datas importantes:

- **Data:** 14/09/2016 (quarta-feira):
 - Entregar (via Moodle): Implementação, Relatório & Apresentação do Trabalho.
 - Apresentação das Equipes.

6. Formato de envio:

- Note que todos arquivos (texto, relatório, slides) devem ser enviados em formato **PDF**. Exceto os arquivos da implementação que devem ser enviados em seus respectivos formatos nativos.
- O nome do arquivo submetido via Moodle deve ter o seguinte formato:

COMP-Atividade2-Nome1-Nome2

- Trabalhos que não respeitem as restrições acima citadas estão sujeito a penalizações na nota final da equipe.