MiniProjecto2 de Língua Natural
2012/2013

Hora/Data de entrega: até às 17 horas do dia 23 de Novembro de 2012

Objectivo do projecto:

O aluno deverá desenvolver um módulo capaz de decidir se uma palavra faz parte de um léxico pré-definido, se o faz a menos de uma alteração ortográfica ou se não o faz de todo. Esta proposta vai permitir que os alunos apliquem conhecimentos adquiridos na disciplina de LN a um problema real de PLN no quadro de um projecto, o FalaComigo.

Detalhes técnicos relativos ao módulo a implementar:

- Código em Java:
  - O código fonte deve ser colocado dentro da pasta *src*;
  - Os recursos usados e ficheiros de configuração devem estar na pasta *resources*;
  - Todos as classes devem pertencer à package *ist.ln.mp2*.

- Input do módulo:
  - Um ficheiro onde se encontra o léxico conhecido pelo sistema;
  - Uma palavra em teste.

- Output do módulo:
  - Uma lista contendo a palavra em teste se esta pertencer ao léxico conhecido pelo sistema;
  - Uma lista com até 5 palavras do léxico conhecido, ordenadas por relevância, representando possíveis alterações ortográficas a fazer à palavra em teste, caso esta não pertença ao léxico. De notar que esta lista só fará sentido se as palavras apresentadas tiverem em conta um número de correcções abaixo de um dado *threshold* para cada técnica usada. Estes thresholds são definidos pelos alunos, deverão ser parametrizáveis (ficheiro de configuração) e têm de ser explicados no relatório. Se todas as palavras encontradas no léxico implicarem alterações à palavras em teste acima do *threshold* de cada técnica, o sistema deverá assumir que a palavra é desconhecida de todo e deverá ser devolvida a lista vazia.

- Serão dadas, num ficheiro zip, as seguintes componentes do módulo descrito em java (com JavaDoc e comentários em todas as componentes java):
  - Classe abstracta LexicalTest (na pasta *src/ist/ln/mp2*, package *ist.ln.mp2*) – Classe abstracta base que representa um teste léxico. Esta classe não pode ser alterada e todos os testes léxicos desenvolvidos devem estender esta classe. Contém um *Set* de palavras conhecidas chamado *knownWords* acessível pelo método público *getKnownWords()* e uma função abstracta *List<String> test(String word)* que, por extensão desta classe, deve implementar a invocação das técnicas usadas e devolve a lista ordenada referida;
  - Classe MP2BaseLexicalTest (na pasta *src/ist/ln/mp2*, package *ist.ln.mp2*) – Exemplo de teste léxico que retorna a palavra recebida se esta pertencer às *knownWords*. Adicionalmente esta classe tem um método *main* que permite lançar o teste léxico numa consola dado um argumento contendo o caminho para o ficheiro de texto com o léxico conhecido;
  - Recurso exampleKnownWords.txt referente ao exemplo apresentado de seguida;
  - Recurso edgarKnownWords.txt referente a todo o léxico a reconhecer no âmbito de MP2.

Exemplo:

Supondo que do léxico fazem parte as palavras *quem*, *construiu*, *o*, *palácio*, *de*, *Monserrate*, *castelo*, *quadro*, *biblioteca* e *bibliotecas*, às palavras à esquerdas o módulo deverá dar as respostas à direita:

Palavra dada	Resposta
quem	[quem]
construíram	[construiu, construir]
Biliotecas	[biblioteca]
digestivo	[]

Entrega do projecto (entrega via Fénix):

Num pacote com o número do grupo (ex: 3.zip) deve ser entregue:

- todo o código relativo ao módulo implementado, incluindo o ficheiro de configuração, recursos utilizados e instruções necessárias (num README) para que o módulo possa correr em linha de comando dado: a) o nome do ficheiro contendo o léxico; b) uma palavra. De notar que o ficheiro de configuração deve conter variáveis para controlar o *threshold* de cada técnica e variáveis para poder suprimir ou ativar cada técnica individualmente (instruções sobre este assunto devem ter lugar no README).
- um relatório (pdf) até 8 páginas com o seguinte conteúdo:
    - Introdução
        - Descrição do problema em mãos e breve resumo da abordagem seguida
    - Descrição da abordagem seguida
        - Metodologia de trabalho;
        - Arquitectura do módulo;
        - Motivação e explicação das técnicas e *thresholds* escolhidas, bem como de ferramentas utilizadas (se aplicável)
    - Avaliação
    - Conclusões e trabalho futuro
        - Principais resultados alcançados;
        - Problemas encontrados;
        - Tarefas que ficaram por fazer
    - Bibliografia

Avaliação do Projecto (em 20):

- Relatório (8 valores)
- Execução (12 valores)

Dúvidas:

luisa.coheur@l2f.inesc-id.pt
sergio.curto@l2f.inesc-id.pt