VIII CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

PIBIC/CNPq/UFCG-2011

Concepção e Desenvolvimento de uma Ferramenta para Extração Periódica e Automática do Sistema e-Pol

Catharine Quintans Bezerra[1], Dalton Dario Serey Guerrero[2]

RESUMO

Os termos que compõem os identificadores utilizados em um código fonte constituem o vocabulário do *software*. As informações extraídas do vocabulário do software podem revelar outros aspectos não capturados pelas métricas estruturais, aprimoram a qualidade do *software* e facilitam a compreensão e a manutenção do sistema. Para viabilizar estudos relacionados ao Vocabulário de Software é necessário um ferramental para tokenização dos identificadores e normalização dos termos. Nesse trabalho nós comparamos dois algoritmos de tokenização para lingua inglesa: O INTT (*Identifier Name Tokenisation Tool*) desenvolvido por Butler et a, e o um nosso CamelCase e Underscore; e três algoritmos de normalização para língua portuguesa: Orengo, Savoy e Porter. Também evoluímos o nosso ferramental de extração, o VocabularyExtractor, e adaptamos os algoritmos de normalização para o contexto da pesquisa. O resultados apontam que o Camelcase e Underscore tem melhor desempenho para tokenização de identificaodres em inglês e o Orengo apresentou melhores resultados para extração de normalização de termos em português.

Palavras-chave: Vocabulário de Software, Recuperação da Informação, Compreensão de Software

Design and Development of a Tool for Periodic and Automatic Extraction of the e-Pol System

ABSTRACT

The terms that comprise the identifiers presents in source code are part of the software vocabulary. The information extracted from the software vocabulary may reveal other aspects not captured by structural metrics, improves software quality and facilitates understanding and maintenance systems. To facilitate studies on Software Vocabulary is necessary a tool for identifiers tokenization and terms standardization terms. In this paper we compare two algorithms for tokenization English: The INTT (Tokenisation Name Identifier Tool) developed by Butler et al, and our one CamelCase and Underscore; and three normalization algorithms to Portuguese: Orengo, Savoy and Porter. We evolved our extraction tool, the VocabularyExtractor, and adapted the algorithms to normalize to the research context. The results reveal that the CamelCase and Underscore has better performance for the tokenization identificaodres English and Orengo showed better results to the terms standardization in portuguese.

Keywords: Software Vocabulary, Information Retriever, Software Comprehension.

[1] Aluna do Curso de Ciência da Computação, Centro de Engenharia Elétrica e Informática, UFCG, Campina Grande, PB, E-mail: catharine.bezerra@ccc.ufcg.edu.br * Autor para correspondências.

[2] Ciência da Computação, Professor. Doutor, Unidade Centro de Engenharia Elétrica e Informática, UFCG, Campina Grande, PB, E-mail: dalton@dsc.ufcg.edu.br