

Introdução ao Apache Lucy

Rafael F. Ruske¹

¹Departamento de Ciência da Computação – Universidade do Estado de Santa Catarina (UDESC)
89.219-710 – Joinville – SC – Brazil

Abstract. *This paper introduces Apache Lucy, a library for C and Perl used for fast text searching and indexing.*

Resumo. *Este artigo introduz Apache Lucy, uma biblioteca para C e Perl usada para busca rápida e indexação textual.*

1. O que é o Apache Lucy

O Apache Lucy é uma biblioteca de busca de código aberto focada em busca textual. Baseada no Apache Lucene, e chamada pelos criadores de uma adaptação ‘relaxada’ do Apache Lucene; o Apache Lucy propõe fornecer as mesmas funcionalidades que o Apache Lucene, com algumas sutis diferenças; uma troca por uma inicialização mais rápida, por uma velocidade de indexação um pouco mais lenta.[HUMPHREY 2012]

2. Como funciona

A utilização do Apache Lucy ocorre em duas principais fases, a indexação e a busca.

2.1. Indexação

Durante a indexação o Apache Lucy quebra o texto a ser processado nas palavras que o compõe, em seguida utiliza um algoritmo de ‘stemming’ para encontrar a raiz das palavras, e indexa essas raízes; tornando possível encontrar palavras parecidas com a procurada ao invés de apenas palavras iguais.[McCandless et al. 2010]

2.2. O Processo de Stemming

Para realizar o processo de Stemming, o Apache Lucy utiliza a biblioteca de código aberto Snowball. Trata-se de uma biblioteca que implementa uma linguagem própria, específica para ‘stemming’ de palavras. Atualmente o Snowball possui regras implementadas somente para o português de Portugal. O processo de ‘stemming’ converteria por exemplo a palavra ‘casa’ em ‘cas’ podendo então encontrar ‘casa’, ‘caseiro’, etc.[McCandless et al. 2010]

2.3. Busca

De posse do índice, o processo de busca simplesmente procura pelas palavras desejadas e indica o documento onde se encontram mais dessas palavras rankeando pelo maior número de acertos.

References

HUMPHREY, M. e. a. (2012). Apache lucy. Apache Software Foundation.

McCandless, M., Hatcher, E., and Gospodnetic, O. (2010). *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA.