

# Ficha 5

## Scripting no Processamento de Linguagem Natural

22 de Março de 2015

### Árvore de prefixos

Uma árvore de prefixos armazena todos os prefixos das palavras introduzidas. Uma das grandes vantagens desta estrutura é que serve para armazenar uma lista de palavras de uma forma compacta (cf. Figura 1). Quando uma lista de palavras é muito grande (por exemplo, um dicionário) é comum que as palavras partilhem prefixos. Neste caso, uma árvore de prefixos permite armazenar a informação de uma forma compacta.

Crie um objeto chamado **PrefixTree** com os seguintes métodos:

**new** O construtor que poderá receber vários argumentos, sendo cada argumento o nome de um ficheiro. Caso estes ficheiros tenham a extensão **.gz** ou **.bz2** então o ficheiro está compactado e deverá ser aberto usando os comandos **gunzip** ou **bunzip2** respetivamente

**save** Este método deverá permitir armazenar (existe um módulo chamado *Storable*<sup>1</sup>) o objeto em disco

**load** Este método deverá permitir consultar o objeto guardado previamente com o método **save**

**add\_dict** Este método deverá permitir adicionar mais ficheiros (deve-se usar mais uma vez a estratégia acima)

**add\_word** Este método deverá permitir adicionar mais uma palavra

**rem\_word** Este método deverá permitir remover uma palavra

**get\_words\_with\_prefix** Este método deverá devolver uma lista de todas as palavras que contenham um dado prefixo

**prefix\_exists** Este método deverá devolver verdadeiro ou falso caso o prefixo exista no dicionário

**word\_exists** Este método deverá devolver verdadeiro ou falso caso a palavra exista no dicionário

Segue-se um exemplo de utilização que demonstra algumas das potencialidades do sistema.

```
1 use Data::Dumper;
2 use PrefixTree;
3
4 # Cria um objecto e inicializa-o com dois ficheiros sendo o primeiro em texto e o segundo compactado com o
  comando bzip2
5 my $pt = new PrefixTree(qw {smalldict.txt dic.bz2});
6
7 # Adiciona algumas palavras
8 $pt->add_word($_) for (qw {aba abaco abeto abrir aberto abertura});
9
10 print Dumper($pt);
11
12 # Imprime todas as palavras que tenham os prefixos ab e ap
13 printf("%s: %s\n", $_, join(" ", $pt->get_words_with_prefix($_))) for (qw {ab ap});
```

---

<sup>1</sup><http://perldoc.perl.org/Storable.html>

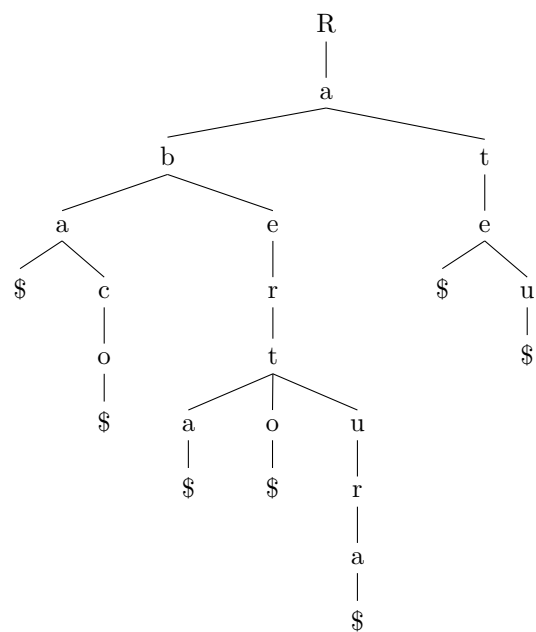


Figura 1: Árvore de prefixos para as palavras *aba*, *ábaco*, *aberta*, *aberto*, *abertura*, *ate* e *ateu*.