UFU/FACOM
Curso de Bacharelado em Sistemas de Informação
GSI024 - Organização e Recuperação da Informação
Lab01

- Escreva um programa que calcule o número de ocorrências de cada palavra de um conjunto de textos, ou *corpus*, armazenados em html
- Apresente a lista de palavras em pares (palavra, frequência de ocorrência) em ordem decrescente de ocorrências
- Considere o tratamento de algumas *anomalias* no arquivo, por exemplo, acentuação, pontuação, hífen, letras maiúsculas/minúsculas, tags html, etc.
- Teste seu programa com o conjunto de arquivos html em: http://www.clul.ul.pt/sectores/linguistica_de_corpus/corpus_oral_pf_publicado.zip
- O programa pode ser escrito na linguagem de sua preferência
- À partir do arquivo com os pares (palavra, frequência), trace um gráfico usando uma ferramenta gráfica de sua preferência, onde o eixo X é a ordem da palavra na lista classificada por frequência (1, 2, 3, etc..) e o eixo Y é a frequência da palavra. Descreva o comportamento da curva gerada.