Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Computação

Recuperação de Informação na Web – 2017/1
Prof. Álvaro R. Pereira Jr.
Trabalho Prático – Parte 3

Datas de entregas:
Resultados: até 16/8/17 as 17hs (durante a aula).
Relatório e dados: até 17/8/17 as 23:55hs

O que deve ser feito:

1. Inicialmente, deve ser feito o que foi discutido em sala de aula, dia 7/8/17, a saber:
   a. Fazer uma seleção aleatória dos documentos coletados.
   b. Separar os dados selecionados em pelo menos dois grupos, que chamaremos de treino e teste.
   c. O grupo de treino pode ter quantos documentos a equipe queira. Estes documentos devem ser manuseados para que a equipe encontre padrões de estrutura, de linguagem, ou outro tipo de padrão que possa ser característico para se classificar o documento como sendo ou da classe “contrato” ou da classe “não-contrato”.
   d. A partir da análise realizada, derivar um algoritmo e seu programa para classificar os documentos de acordo com as classes apresentadas.
   e. Selecionar ao menos 100 documentos da base de dados de teste, ordená-los e rotular manualmente quanto às classes “contrato” ou “não-contrato”.
   f. Executar o algoritmo para os 100 documentos rotulados.
   g. Avaliar os resultados em termos de precisão, revocação e F1.
2. Formato de entrega:
   a. Resultados: apresentar os seus resultados e seu algoritmo (alto nível) em reunião que faremos na aula do dia 16/8/17. A presença nessa aula conta como parte da pontuação do trabalho, bem como a apresentação do algoritmo e dos resultados.
   b. Relatório deve conter qualquer configuração especial que tenha sido necessária para o experimento, como parâmetros usados ou considerações relevantes. Deve conter também os resultados e comentários sobre os resultados, tendo uma conclusão a partir dos dados. Pode ser breve, mas tem que ter as informações necessárias para compreensão do que foi feito e do resultado.
   c. Arquivo contendo os seus dados, separados em uma instância por linha. Cada instância deve conter, em ordem, a URL do documento avaliado, a classe real manualmente avaliada, e a classe escolhida pelo seu algoritmo. O formato será o .cvs, onde os campos são simplesmente separados pelo caractere ‘;’. Esse formato é mais simples e poderá ser usado com facilidade no Excel. Importante a equipe testar se abre corretamente no Excel.