

AVALIAÇÃO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO

Esse exercício está dimensionado para ser feito em Python com a biblioteca NLTK.

O exercício usará a base CysticFibrosis2, disponível no Moodle e o resultado do exercício anterior.

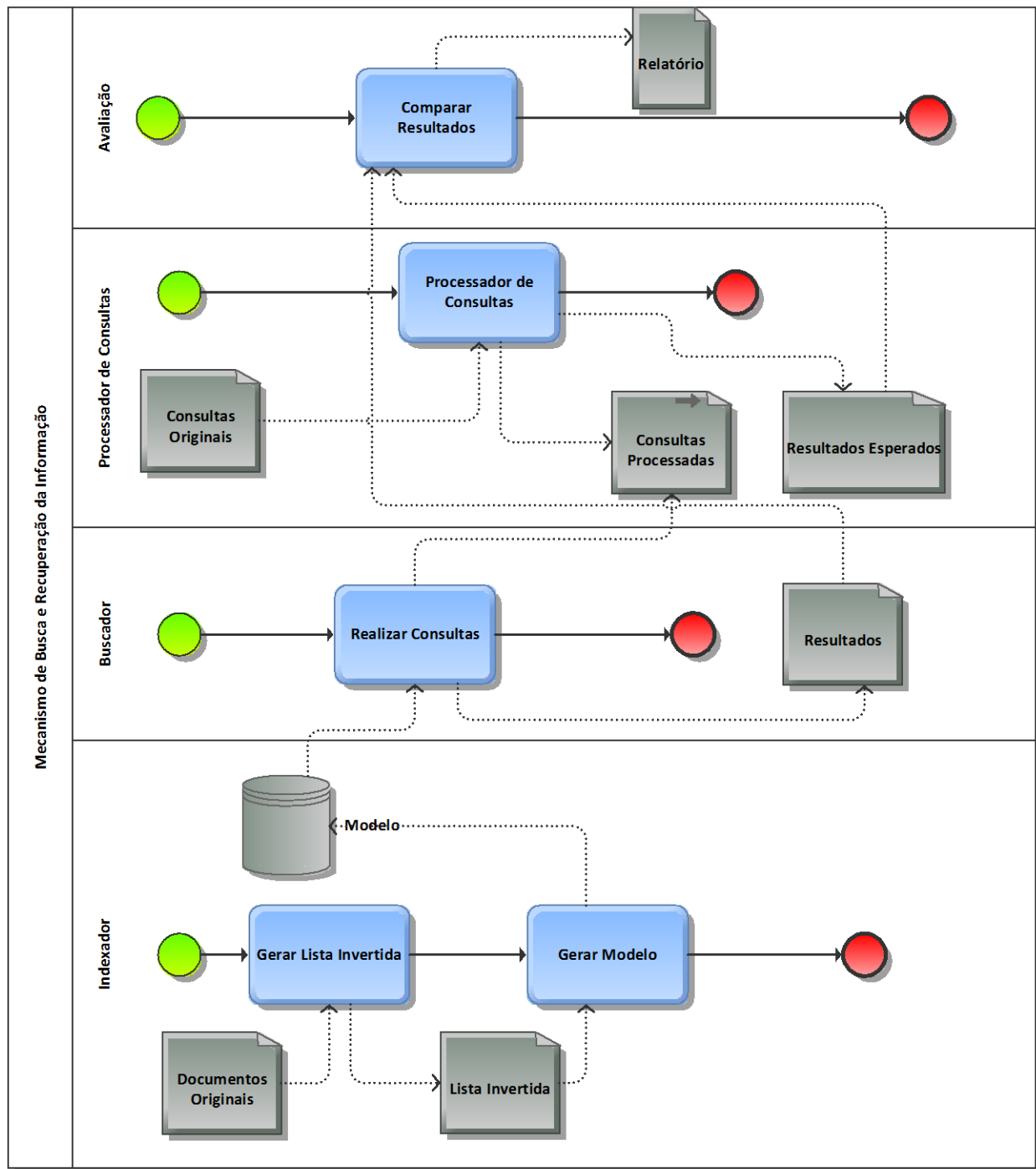
Você deve avaliar o seu sistema de recuperação de informação, com e sem o uso do stemmer de Porter, usando os arquivos RESULTADOS.CSV e RESULTADOS ESPERADOS.CSV para obter as seguintes medidas e diagramas:

1. Gráfico de 11 pontos de precisão e recall
2. F_1
3. Precision@10
4. MAP
5. Discounted Cumulative Gain
6. Normalized Discounted Cumulative Gain

Os resultados devem todos ser entregues em um arquivo RELATORIO.TXT, incluindo os pontos dos gráficos.

Os diagramas também devem ser entregues (cada um) na forma de um arquivo CVS e de um ou mais arquivo gráfico (PDF ou outro), ambos no formato <tipo de gráfico>-<dado usado>-<sequencial>.<formato do arquivo>. Por exemplo 11pontos-nostemmer-1.csv ou 11pontos-stemmer-2.pdf

Segue o novo modelo:



FAZENDO O STEMMING

1. Você deve incluir no arquivo de configuração dos programas que analisam diretamente o texto a possibilidade de usar ou não um stemmer
2. A opção é uma linha no início com a palavra STEMMER ou NOSTEMMER
3. Você deve usar o Stemmer de Porter disponível em <http://tartarus.org/martin/PorterStemmer/>

ENTREGA

Os alunos devem entregar em um arquivo ZIP com o nome do aluno (formato <nomedoaluno>.zip”:

1. Todo o código fonte
2. Um arquivo README.TXT com qualquer instrução adicional para uso do código entregue
3. Todos os arquivos criados por sua execução.