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, atualizando-o para trabalhar 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:

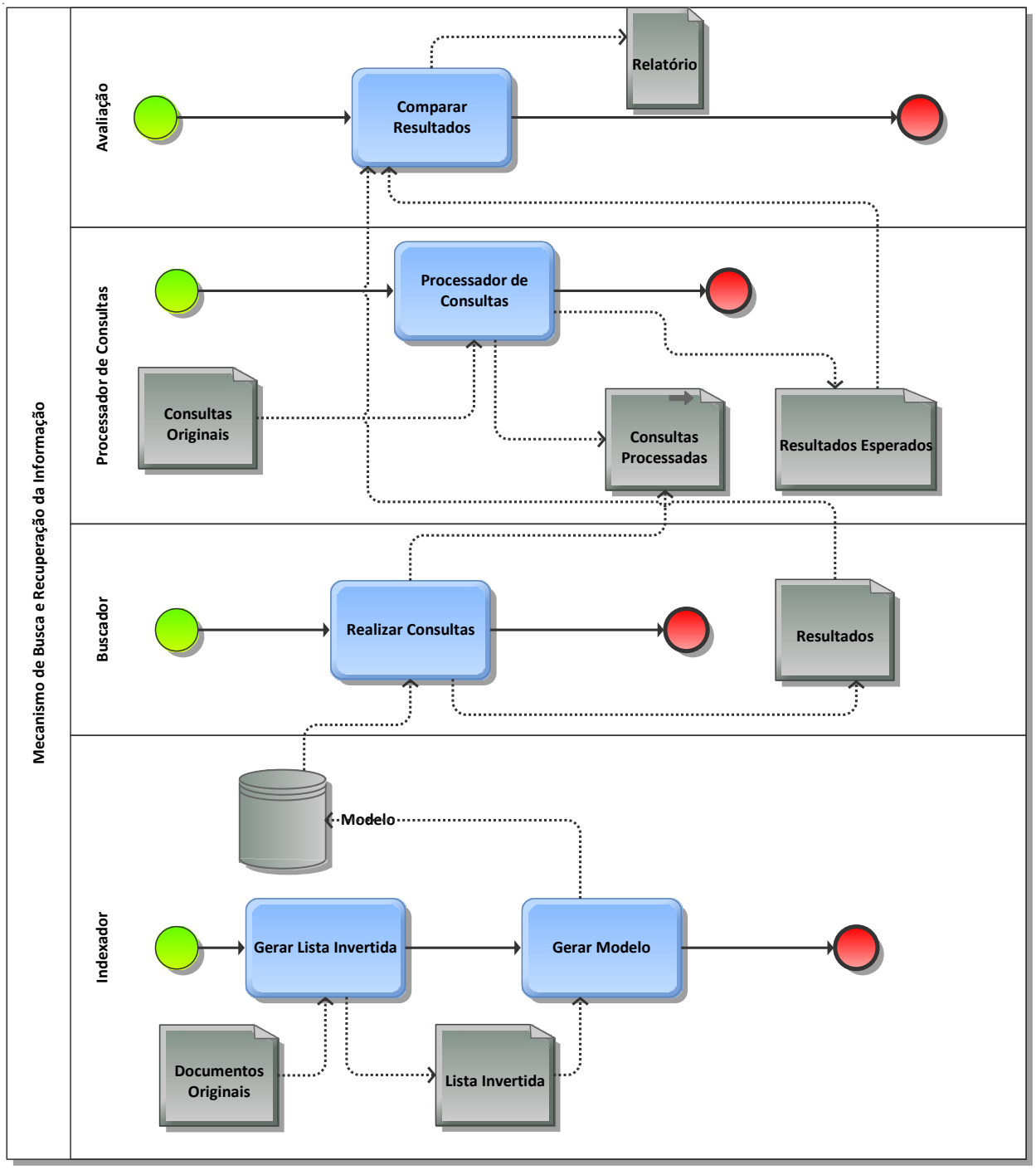
1. Gráfico de 11 pontos de precisão e recall
2. F_1
3. Precision@5
4. Precision@10
5. Histograma de R-Precision (comparativo)
6. MAP
7. MRR
8. Discounted Cumulative Gain (médio)
9. Normalized Discounted Cumulative Gain

Se alguma decisão do limite para o rank for importante, use 10.

Os resultados devem todos ser entregues em um arquivo RELATORIO.MD.

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.cvs 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/

4. A palavra STEMMER ou NOSTEMMER deve ser somada a palavra RESULTADOS para formar o nome do arquivo de resposta, resultando em RESULTADOS-STEMMER ou RESULTADOS-NOTEMMER.

ENTREGA

Os alunos devem entregar como uma atualização do GitHub do exercício anterior. Os resultados da avaliação devem ficar em um diretório AVALIA. O arquivo README.MD deve ser atualizado.
No Moodle deve ser colado o link para o repositório.