**Tarefa 1**

**Construção de um sistema e RI para indexação e busca de documentos**

**Descrição geral:**

Com auxílio de alguma ferramenta pré-existente (ver slides no site da aula), criar um sistema para indexação e busca de documentos.

A ferramenta sugerida é o Solr do LUCENE Apache:

<https://lucene.apache.org/>

<https://lucene.apache.org/solr/>

Alguns sites com sugestões de outras ferramentas:

* https://en.wikipedia.org/wiki/List\_of\_information\_retrieval\_libraries
* http://www.emse.fr/~mbeig/IR/tools.html

Usar o sistema para indexar uma base de documentos, permitindo consultas.

* A quantidade de documentos na base dependerá do número de participantes. Para 1 ou 2 participantes, 100 documentos; para 3 participantes, 200 documentos.
  + Uma base muito pequena não possibilita avaliação de desempenho (precisão x cobertura).
* Medir a precisão e a cobertura do sistema com base em duas consultas (*queries*).

A tarefa pode ser individual, em dupla ou em trio (no máximo!!!!). Equipes com 3 alunos terão uma atividade extra, para justificar o tamanho da equipe.

A data de entrega está no cronograma do curso. O relatório deve ser enviado por email até esta data.

Quero ver o sistema em execução.

**Projeto & Relatório** (tudo misturado):

Escrevam um relatório simples, com as informações solicitadas abaixo:

# Nome dos membros da equipe

# (0,5) Descrição dos documentos (corpus) que serão indexados pelo sistema

* Temas/tópicos dos documentos selecionados
* Mostrar no relatório 2 exemplos de documentos do corpus (um para cada tema escolhido).

**Obs.:** Os documentos podem ser obtidos na Web ou de bases particulares. Vocês podem utilizar documentos oriundos de fontes variadas, bem como focar em sites dedicados. Em qualquer caso, seria bom escolher 2 temas de interesse e refletir isso nas consultas ao Google. Caso contrário, correrão o risco de montar uma base muito diversificada, prejudicando a precisão do sistema.

# (1,0) Arquitetura do sistema

Informar a ferramenta usada para desenvolver o sistema. Prover uma descrição breve dos módulos do sistema: opções de pré-processamento, tipos de consultas permitidas (palavras isoladas, com contexto (aspas), com operador booleano, etc), funções de ranking, etc. Informar qual modelo de RI está disponível pelo sistema (vejam as aulas de modelos de RI).

Obs.: atenção com a função de ordenação... o Solr aceita várias funções diferentes. O Solr padrão utiliza o algoritmo BM25 para ordenar os documentos, não é o cosseno. Quem quiser utilizar cosseno precisa modificar isso no arquivo de configuração.

# (3,0) Criação das bases de índices invertidos (Preparação & Indexação dos documentos)

Vocês devem criar, de forma automática, quatro bases de índices invertidos a partir do mesmo corpus de documentos, porém variando os processos utilizados na preparação (pré-processamento) dos documentos e das consultas:

* Base 1: documentos originais, sem usar nenhum tipo de pré-processamento (nenhum filtro)*;*
* Base 2: eliminar *stopwords* e não usar *stemming;*
* Base 3: não eliminar *stopwords* e usar *stemming;*
* Base 4: eliminar *stopwords* e usar *stemming.*
* Observações:
  + Algumas ferramentas já oferecem *stoplist* e algoritmo de *stemming*.
  + Só vou cobrar uso de tesauros ou *n-grams* para equipes com 3 participantes

# (2,5 – divididos entre duas tarefas) Criação das consultas (1,0) e preparação dos testes – matriz de relevância (1,5)

**Passos:**

1. Definir 2 consultas diferentes, levando em conta os temas cobertos pelos documentos da base
   * caso contrário, as consultas poderão não recuperar nenhum documento relevante
2. Incluam nas consultas pelo menos 1 *stopword*, para ver o efeito da retirada de stopwords
   * artigo, preposição, conjunção...
3. Incluam nas consultas algum termo que sofra variação de *stemming*, para ver o efeito da redução ao stem (radical).
   * Plural, verbo flexionado, etc
4. Avaliar manualmente a relevância de cada documento da base em relação a cada consulta.
   * A relevância pode ser binária – é bem mais simples.
5. Guardar o resultado da sua avaliação manual em uma matriz (pode ser uma planilha Excel), a fim de possibilitar o cálculo automático da cobertura e da precisão do sistema. Exemplo:

**Exemplo de Matriz de relevância “*Consultas x Documentos”.***

|  |  |  |
| --- | --- | --- |
|  | Consulta1 | Consulta2 |
| Doc 1 | 1 | 0 |
| Doc2 | 1 | 1 |
| Doc3 |  |  |
|  |  |  |
| Qtd de docs relevantes | 35 | 42 |

Obs.: A última linha deve informar a quantidade de documentos julgados relevantes por vocês (avaliação manual) para cada consulta. Assim fica mais fácil de entender os valores de precisão e cobertura.

# (2,5) Testes

Submeter as 2 consultas definidas a cada base do sistema, e avaliar cada uma separadamente; i.e., calcular separadamente a precisão e a cobertura de cada consulta em relação a cada base criada. Usar as fórmulas vistas em aula: precisão, cobertura e *F-measure*.

Incluir no relatório uma matriz de resultados para CADA consulta. Assim podemos ver a influência do pré-processamento dos documentos no resultado final do sistema.

**Matriz de resultados para a Consulta 1*.***

***Obs.: incluam aqui o texto da consulta. Assim fica mais fácil de interpretar os resultados. Incluam Tb a quantidade de docs relevantes manualmente identificados (seção 5 acima).***

***Exemplo:***

**Consulta: Information retrieval for news about sports**

**Total de docs relevantes: 35**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure | Qtd de documentos retornados |
| Base 1 |  |  |  |  |
| Base 2 |  |  |  |  |
| Base 3 |  |  |  |  |
| Base 4 |  |  |  |  |

**Matriz de resultados para a Consulta 2*.***

***Igual à anterior.***

As medidas de precisão, cobertura e *F-meause* do sistema serão obtidas calculando-se a média entre os resultados obtidos com cada consulta em relação a cada base criada.

**Matriz de resultados para o Sistema (média entre as duas matrizes anteriores)*.***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão média | Cobertura média | F-measure  média |
| Base 1 |  |  |  |
| Base 2 |  |  |  |
| Base 3 |  |  |  |
| Base 4 |  |  |  |

# (0,5) Conclusão

Um texto curto para explicar o que vocês concluem a partir do resultado dos experimentos.

**Critérios de nota da apresentação oral:**

Cada aluno da dupla deve apresentar a execução de UMA das duas consultas submetidas ao sistema:

* Explicar rapidamente como utilizar a ferramenta (2 pontos).
* Mostrar a etapa de indexação de cada um dos 4 índices invertidos (CORE) (3 pontos).
* Explicar o os efeitos do pré-processamento de cada uma das bases nos resultados obtidos com a consulta escolhida pelo aluno (3 pontos).
* explicar como se deu o pré-processamento da consulta (2 pontos).