**Mineração da Web - 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 mais usada no mundo é a LUCENE da 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, incluindo interface de consulta SIMPLES para permitir a digitação da consulta. Medir a precisão e a cobertura do sistema com base em duas consultas (*queries*).

A base de documentos deve ter pelo menos 200 documentos. Uma base muito pequena não possibilita uma avaliação de precisão minimamente confiável...

A tarefa pode ser individual, em dupla ou em trio (no máximo!!!!). A quantidade de documentos na base dependerá do número de participantes. Para 1 ou 2 participantes, 200 documentos; para 3 participantes, 300 documentos.

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.

**Não dá pra adiar porque temos ainda muito trabalho pela frente. E esta tarefa é simples...**

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

Escrevam um relatório simples, com as informações solicitadas abaixo:

# Nome dos membros da equipe

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

* Temas/tópicos dos documentos selecionados
* Mostrar no relatório pelo menos 1 exemplo de documento do corpus
* Quantidade de documentos coletados (200 ou 300).

**Obs.:** Os documentos podem ser obtidos na Web. Vocês podem utilizar documentos oriundos de fontes variadas, bem como focar em sites dedicados. Em qualquer caso, seria bom escolher 2 ou 3 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.

Exemplos de bases de documentos:

* https://arxiv.org/list/cs/new
* 20newsgroups  <http://qwone.com/~jason/20Newsgroups/>

# Arquitetura do sistema

Informar a ferramenta usada para desenvolver o sistema, prover uma descrição breve dos módulos do sistema, e informar qual o modelo de RI implementado pelo sistema (vejam as aulas de modelos de RI).

# Criação das bases de documentos indexados (Preparação & Indexação dos documentos)

Eu quero que vocês criem, de forma automática, quatro bases de documentos indexados, cada base variando os processos utilizados na preparação dos dados:

* Base 1: documentos originais sem eliminar *stoplist* e sem usar *stemming;*
* Base 2: eliminar *stoplist* e sem usar *stemming;*
* Base 3: sem eliminar *stoplist* e usando *stemming;*
* Base 4: eliminar *stoplist* e usando *stemming.*
* Observações:
	+ Algumas ferramentas já oferecem *stoplist* e algoritmo de *stemming*.
	+ Não vou cobrar uso de tesauros nem *n-grams*, pra não complicar.

# Criação das consultas e preparação dos testes

Inicialmente, informem os tipos de consultas que o sistema consegue processar: palavras isoladas, com contexto (aspas), com operador booleano???

Passos para a preparação dos testes:

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 alguma palavra da lista de *stopwords*, para ver o efeito da retirada de *stopwords*
	* E.g, métodos **para** construir sistemas inteligentes
3. Incluam nas consultas algum termo que sofra variação de *stemming*, para ver o efeito da redução ao *stem* (radical).
	* E.g, método**s** constru**ir** sistema**s** inteligente**s**
	* Obs: se os textos forem em inglês, é preciso mais atenção na escolha de palavras que sofrem ação do algoritmo de *stemming*, pois alguns termos em inglês já aparecem na sua forma de *stem*.
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, a fim de possibilitar o cálculo automático da cobertura e da precisão do sistema. Exemplo:

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Doc 1 | Doc2 | ...... | Doc200 |
| Consulta1 | 1 | 0 |  | 1 |
| Consulta2 | 1 | 1 |  | 0 |

1. Mostrar parte da matriz no relatório, para eu poder ver o que está acontecendo.

# 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*.***

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

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

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

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*.***

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

# Conclusão

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