FACOM- UFU
Professor: Wendel Melo

Terceiro trabalho de Organização e Recuperação da Informação 2018-02

Descrição

Este trabalho tem como foco a avaliação de um sistema de busca através de uma coleção de referência com consultas conhecidas. Assim, os passos principais são:

1. Ter em mãos uma implementação própria de modelo de recuperação de informação para avaliar. Você pode usar (adaptar) a **sua** implementação do modelo vetorial desenvolvida no trabalho 2, ou implementar algum modelo de RI especificamente para este trabalho. Aqui, você é livre para propor e experimentar suas próprias modificações no modelo usado ou no esquema de ponderação.

2. Desenvolver um código para rodar a sua implementação de modelo de RI para **todas** as consultas da coleção de referência e construir um gráfico de precisão por revocação média usando a biblioteca *matplotlib*. As métricas de precisão e revocação são discutidas na aula "Avaliação da Recuperação" (veja os slides).

3. Entregar, além do código desenvolvido, um relatório explicando qual o modelo utilizado; possíveis modificações no modelo; valores adotados para os parâmetros (por exemplo, se usar o modelo vetorial, informar qual o valor do patamar mínimo de similaridade para relevância foi usado), o gráfico de precisão por revocação e comentários adicionais que se mostrem pertinentes. Se quiser impressionar o professor, você pode propor alguma modificação no algoritmo original do modelo ou da ponderação de termos e apresentar um gráfico comparando o modelo original com o modelo modificado.

Deve ser entregue apenas um **único** programa desenvolvido em Python 3. O programa deve usar apenas as bibliotecas padrão Python 3, isto é, as bibliotecas que já vem com a instalação padrão do interpretador da linguagem, com exceção da biblioteca *nltk*, que deve ser utilizada para remoção de *stopwords* e extração de radicais, e das bibliotecas *numpy* e *matplotlib* que podem ser utilizadas para cálculos algébricos em geral e plotagem de gráfico. Como a coleção de referência possui documentos na língua inglesa, será preciso considerar *stopwords* e extração de radicais para esta linguagem. Para a lista de *stopwords* em inglês pode se considerar:

```
import nltk
sw = nltk.corpus.stopwords.words('english')
```

Para a extração de radicais na língua inglesa, pode-se usar

```
st = nltk.stem.SnowballStemmer("english")
st.stem("sleeping")
'sleep'
```

Neste trabalho não é necessária a classificação gramatical de termos (etiquetação) para remoção de stopwords por classe gramatical. Assim, não é necessário utilizar nenhum dos etiquetadores presentes no pacote nltk. Todavia, a remoção de stopwords pela lista disponibilizada pelo nltk ainda deve ser realizada.

O trabalho deve ser feito em grupo de **um ou dois alunos,** e o código e relatório gerados devem ser entregues por e-mail ao professor (wendelmelo@ufu.br) até a data 13/12/2018.

Aviso importante: se for detectado cópia ou qualquer tipo de trapaça entre diferentes grupos, todos os grupos serão punidos com a nota zero. Portanto, pense bem antes de pedir para copiar o trabalho do seu coleguinha, pois ele poderá ser punido também!

A base de dados

A base de dados é a coleção de referência CF disponível na Medline e pode ser baixada no site da disciplina. Ao todo, são 1239 documentos semiestruturados que representam resumos de artigos sobre fibrose cística em inglês. Note que, no pacote que traz a coleção de referência, há um arquivo denominado *base.txt* com os caminhos de todos os documentos que compõe a base, seguindo o padrão estabelecido nos trabalhos anteriores. Cada um desses documentos que compõe a base de dados pode possuir os seguintes campos:

1. TITLE: título do documento;
2. AUTHORS: lista de autores
3. SOURCE: fonte do documento
4. ABSTRACT: resumo do artigo, que no nosso caso, é o corpo do documento propriamente dito
5. MAJOR SUBJECTS: tópicos principais do artigo
6. MINOR SUBJECTS: tópicos secundários (não centrais) do artigo
7. REFERENCES: lista de referências artigo, isto é, lista de outros artigos citados por este documento
8. CITATIONS: lista de citações, isto é, lista de outros artigos que citam este documento.

Observe que a semi-estruturação dos documentos permite a construção de funções de ponderação de termos e de ranqueamento de documentos mais apuradas. Por exemplo, um documento com número maior de citações poderia vir a ser beneficiado de alguma forma no processo de ranqueamento.

O modelo de Recuperação de Informação

Neste trabalho, há liberdade para propor um novo modelo de recuperação de informação. É permitido usar também algum dos modelos já implementados em trabalhos anteriores como base, ou implementar algum outro modelo visto em aula ou da literatura. No entanto, a criatividade na adaptação do modelo ao considerar a base de dados proposta, que, sendo, semi-estruturada provê oportunidades de melhorias, será bonificada com pontuação extra. Modificações gerais nos modelos que não se valham das características da base também são bem-vindas, assim como modificações no tratamento das consultas. Tudo o que for usado ou proposto deverá constar no relatório a ser entregue junto com o trabalho. Você pode discutir sobre possíveis modificações, ou mesmo pedir sugestões, com o seu professor.

As consultas

As consultas de referência estão especificadas no arquivo *queries.txt*. Ao todo, este arquivo contempla 100 consultas formuladas e respondidas por especialistas. Dessa forma, para cada uma das 100 consultas, já se conhece o seu respectivo conjunto de documentos relevantes. Adicionalmente, é fornecido um valor de ranqueamento entre 1 e 8 para cada documento na resposta. Quanto maior o valor, maior a similaridade entre o respectivo documento e a consulta. Observe que seu modelo de RI deve adotar sua própria função de ranqueamento que não precisa (e não deve!) produzir os mesmos valores de ranqueamento presentes para as consultas do arquivo *queries.txt*.
Cada uma dessas consultas possui os seguintes campos:

1. NUMBER: número da consulta;
2. TEXT: texto (corpo) da consulta. Note que o corpo da consulta é composto apenas de termos. **Assim considere que todos os termos das consultas são ligados por conectores AND**;
3. NUMBER OF RELEVANT DOCS: número de documentos relevantes à consulta segundo a avaliação dos especialistas;
4. RELEVANT DOCS AND SCORES: uma lista com os documentos relevantes e seus respectivos escores de similaridade. Por exemplo, se a lista é composta por

   324,8 268,8 449,7

   isto significa que temos 3 documentos relevantes. O documento 324 é relevante a consulta em questão e possui escore de similaridade 8, assim como o documento 268. O documento 449 também é relevante à consulta em questão e possui escore de similaridade 7

O(s) gráfico(s) de precisão por revocação

O relatório entregue deverá contemplar gráficos de médias precisão por revocação conforme visto em sala aula. Para a geração desse gráfico, será preciso contemplar as 100 consultas de referência da coleção em testes no sistema implementado. **Observe que não devem ser gerados 100 gráficos distintos!** Para cada configuração de algoritmo avaliada, deve ser gerado apenas um gráfico contendo as médias de precisão por revocação para as 100 consultas nos 11 níveis de revocação de referência vistos em sala de aula. Assim, se o seu trabalho apenas testar uma configuração de algoritmo de RI, apenas um gráfico deve ser apresentando no relatório considerando as médias de precisão por revocação para as 100 consultas. **É importe frisar que todas as 100 consultas sempre devem ser levadas em conta na geração de cada gráfico**.