Mineração Web
Relatório

RI-System
2018.2

Equipe:
Fillipe de Menezes ()
Franclin Cabral (fcmo)

1. **Descrição dos documentos Indexados pelo Sistema**
    - **Tema:**
        - Foi escolhido o tema med. O tema em questão contém documentos relacionados a casos de medicina e conversas de sintomas ou eventos ocorridos a alguma praga.
    - **Exemplo de documento:**

```
From: wright@duca.hi.com (David Wright)
Subject: Re: Name of MD's eyepiece?
Organization: Hitachi Computer Products, OSSD division
Lines: 21
NNTP-Posting-Host: duca.hi.com

In article <19387@pitt.UUCP> geb@cs.pitt.edu (Gordon Banks) writes:
>In article <C4IHM2.Gs9@watson.ibm.com> clarke@watson.ibm.com (Ed Clarke) writes:
>>|> |It's not an eyepiece.  It is called a head mirror.  All doctors never
>>
>>A speculum?
>
>The speculum is the little cone that fits on the end of the otoscope.
>There are also vaginal specula that females and gynecologists are
>all too familiar with.

In fairness, we should note that if you look up "speculum" in the
dictionary (which I did when this question first surfaced), the first
definition is "a mirror or polished metal plate used as a reflector in
optical instruments."

Which doesn't mean the name fits in this context, but it's not as far
off as you might think.

  -- David Wright, Hitachi Computer Products (America), Inc.  Waltham, MA
     wright@hicomb.hi.com  ::  These are my opinions, not necessarily
     Hitachi's, though they are the opinions of all right-thinking people
```

    - **Quantidade de documentos coletados:**
        - Ao todo foram coletados e indexados 200 documentos.

2. **Arquitetura do sistema**

Utilizamos para o desenvolvimento do sistema a linguagem Java, com a IDE Eclipse e as bibliotecas do Lucene 7.4.0.

O programa é dividido em indexação dos documentos, onde se resume em indexar os documentos coletados sobre o tema, pesquisa é referente a busca e recuperação dos documentos baseados em uma string de busca e o módulo de performance que provê toda informação com relação a precisão.

Foi Utilizado um RI booleano por ser mais simples e o sistema não ser tão complexo.

3. **Criação das Bases de Documentos Indexados.**

A preparação da base foi feita, mas utilizamos uma lista extra de stopwords além das ja utilizadas pelo Lucene. Para Stemming, foi utilizado os do próprio Lucene.

Os documentos originais são armazenados em "Documents/docs_00". A indexação normal dos arquivos é feita em "index/index_00" e simultaneamente aos documentos originais, serão armazenados nas pastas "docs_01", "docs_10" e "docs_11" com stemming, stopwords e stemming&stopwords respectivamente. Com uma pasta de index para cada.

Seguindo o planejamento, obtivemos quatro bases distintas e separadas que serão úteis na hora de fazer uma pesquisa.

4. Criação de consultas e preparação de testes

O sistema é capaz de processar pesquisas do tipo: term, wildcard searches, fuzzy searches, proximity search, boosting term, boolean operators e context search.

As queries selecionadas foram:

- High fever
- Biggest case of tumor
- Cases of cancer in the last year

Abaixo, segue uma imagem da matriz de relevância usada, foi criada manualmente.

```
private static int[][] relevance = {
        {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
        {0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
};
```

5. Testes

Para os testes foi utilizado o arquivo logs.txt que contém a matriz de avaliação.

A partir dessa matriz, é possível obter a precisão do sistema em relação as queries e as bases.

```
System accuracy:
Q1      Q2      Q3
0.375   0.06    0.36
0.315   0.03    0.125
0.48    0.465   0.66
0.475   0.395   0.6
Media:
0.411   0.238   0.436
```

```
Precision:
Q1      Q2      Q3
0.101   0.021   0.112
0.093   0.025   0.089
0.119   0.019   0.16
0.118   0.017   0.146
Media:
0.108   0.021   0.127

Recall:
Q1      Q2      Q3
1.0     0.8     0.941
1.0     1.0     1.0
1.0     0.4     0.706
1.0     0.4     0.765
Media:
1.0     0.65    0.853

f-measure:
Q1      Q2      Q3
0.183   0.041   0.2
0.17    0.049   0.163
0.213   0.036   0.261
0.211   0.033   0.245
Media:
0.194   0.04    0.217

---------------------
System accuracy:
Q1      Q2      Q3
0.375   0.06    0.36
0.315   0.03    0.125
0.48    0.465   0.66
0.475   0.395   0.6
Media:
0.411   0.238   0.436
```

6. Conclusão

Com os resultados dos testes, chegamos a conclusão que a baixa precisão é dada pelo baixo número de documentos aleatoriamente indexados, mesmo que pertencentes ao mesmo tema. Diante do problema, ficou difícil obter uma query que retornasse documentos de fato relevantes. Com o crescimento da base de dados, a precisão poderia ser melhor e maior.