Mineração WEB - Tarefa 1

Recuperação de Informação sobre fármacos

Aluno: João Lucas Gomes de Miranda (jlgm@cin.ufpe.br)

1. Descrição dos documentos

Os documentos foram montados a partir do site www.drugs.com, um site que é referência mundial na área de farmácia e que dispõe de informações detalhadas a respeito de um número muito grande de fármacos existentes.

Por conter um banco de dados muito imenso, o processo de captura das informações foi reduzido a um subconjunto do site. Especificamente, as informações que foram obtidas diz respeito às drogas utilizadas nos tratamentos das seguintes condições: acne/espinhas, insônia, depressão, dor, hepatite A, menopausa, diarréia e gripe.

A seguir tem-se um exemplo de conteúdo que consta em um dos documentos:

Generic Name: doxycycline (DOX i SYE kleen)Brand Names: Acticlate, Adoxa, Alodox, Avidoxy, Doryx, Mondoxyne NL, Monodox, Morgidox, Ocudox Convenience Kit, Oracea, Oraxyl, Targadox, Vibramycin
Doxycycline is a tetracycline antibiotic that fights bacteria in the body.
Doxycycline is used to treat many different bacterial infections, such as acne, urinary tract infections, intestinal infections, eye infections, gonorrhea, chlamydia, periodontitis (gum disease), and others.
Doxycycline is also used to treat blemishes, bumps, and acne-like lesions caused by rosacea. It will not treat facial redness caused by rosacea.
Some forms of doxycycline are used to prevent malaria, to treat anthrax, or to treat infections caused by mites, ticks, or lice.
You should not take doxycycline if you are allergic to any tetracycline antibiotic.
Children younger than 8 years old should use doxycycline only in cases of severe or life-threatening conditions. This medicine can cause permanent yellowing or graying of the teeth in children.

Doc1.txt - descrição de uma droga para acne

No total, foram montados 163 documentos. Um script em *Python* foi confeccionado para automatizar o processo; este encontra-se no repositório do projeto, na pasta *utility*.

2. Arquitetura do Sistema

O sistema foi desenvolvido em Java. Foram criados 4 packages — NoStopNoStem, NoStopWithStem, WithStopNoStem e, finalmente, WithStopWithStem —, cada um fazendo a indexação de acordo com o que o nome sugere (dessa forma, satisfazendo os requerimentos do projeto).

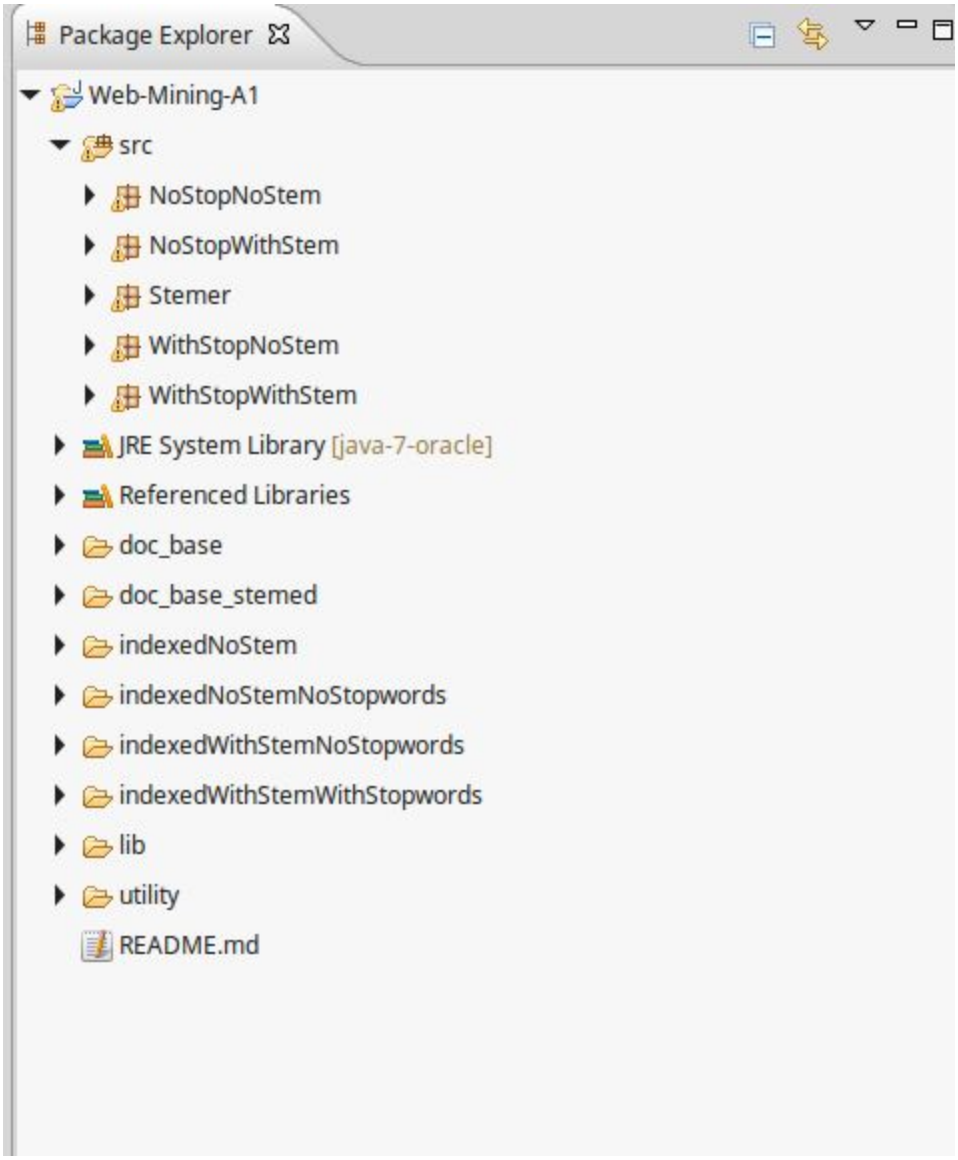

Fig.1 — estrutura de diretórios do projeto

O modelo de RI escolhido no projeto foi o modelo booleano, por questões de simplicidade.

Além da base extraída pelo processo automatizado (doc_base), também foi criada uma base nova chamada (doc_base_stemed) que contém os mesmos documentos, mas com todos os *tokens* stemizados (isso é, apenas os radicais das palavras foram mantidos). Para fins de testes mais elaborados, os arquivos nesse diretório tiveram seus nomes modificados — bem como a extensão, que passou a ser .stem.

Assim, sempre que o sistema for utilizar do processo de *Stemming*, a base utilizada para a indexação vai ser a base (doc_base_stemed), enquanto que quando não for, a base utilizada vai ser a (doc_base).

A seguir, um exemplo de documento na base (doc_base_stemed):

Gener Name sulfamethoxazol and trimethoprim SUL fa meth OX a zole and trye METH oh prim Brand Name Bactrim Septra SMZ TMP Septra contain a combin of sulfamethoxazol and trimethoprim Sulfamethoxazol and trimethoprim ar ar both antibiot that treat differ type of infect caus by bacteria Septra is us to treat ear infect urinari tract infect bronchiti traveler' diarrhea shigellosi and Pneumocysti jiroveci pneumonia Septra mai also be us for purpos not list in thi medic guid You should not us Septra if you have sever liver or kidnei diseas anemia caus by folic acid defici or a histori of low blood platelet caus by take trimethoprim or ani sulfa drug You should not us Septra if you ar allerg to sulfamethoxazol or trimethoprim or if you have sever liver or kidnei diseas

Doc26.stem - descrição da droga sulfamethoxazol

3. Criação das Bases Indexadas

Como pode ser observado na Fig.1, quatro bases indexadas foram criadas:

1. **indexedNoStem** — não faz o processo de *Stemming* nem elimina stopwords;
2. **indexedNoStemNoStopwords** — não faz o processo de *Stemming*, mas elimina as stopwords;
3. **indexedWithStemNoStopwords** — faz o processo de *Stemming* e também elimina as stopwords;
4. **IndexedWithStemWithStopwords** — faz o processo de *Stemming*, mas não elimina as stopwords.

4. Criação das Consultas

O sistema é capaz de processar consultas de palavras-únicas e também consultas com os operadores AND e OR. Também consegue recuperar palavras separadas por espaço (ex. “treatment for acne”).

As consultas criadas nesse trabalho foram as seguintes:

1. “Acne”
2. “In the”

A primeira consulta é suposta de retornar os documentos doc1.txt - doc25.txt — esses documentos foram recuperados a partir do tema “acne”, sendo portanto os mais prováveis de ser recuperados nessa query. A palavra “acne” também sofre modificação no processo de *Stemming*, passando a ser “acn”.

A segunda query — que visa observar o funcionamento da aplicação quando a query contém stopwords — deve retornar os documentos que contém as palavras “in” e “the” como parte do conteúdo (ou seja, basicamente todos quando não ocorrer eliminação de stopwords).

A seguir, parte da matriz de relevância:

	doc1	doc2	doc3	doc4	doc5	doc6	doc7
“acne	1	1				1	
“In the”	1	1	1	1	1	1	1

Tabela 1: parte da matriz de relevância

A matriz completa está no repositório do projeto, na pasta *utility*.

5. Testes das Consultas

Da matriz de relevância, sabemos que:

1) Para a query 1 (“acne”), há 19 páginas relevantes.
2) Para a query 2 (“in the”), há 163 páginas relevantes.

Com base nisso, podem-se formar as matrizes de resultados, analisando-se, pois, os resultados dados pelo sistema nas suas diferentes implementações. Os testes são descritos a seguir:

5.1 NoStopNoStem

Teste na aplicação quando não é usado *stemming* mas é removido stopwords.

“Acne”:

```
163 File indexed, time taken: 278 ms
18 documents found. Time :17
File: /home/joao/workspace/Web-Mining-A1/doc_base/doc6.txt
File: /home/joao/workspace/Web-Mining-A1/doc_base/doc24.txt
File: /home/joao/workspace/Web-Mining-A1/doc_base/doc13.txt
(...)
```

“In the”:

```
163 File indexed, time taken: 277 ms
0 documents found. Time :15
```

5.2 NoStopWithStem

Teste na aplicação quando é usado *stemming* e também são removidas as stopwords.

“Acne”:

```
163 File indexed, time taken: 272 ms
18 documents found. Time :24
(...)
```

“In the”:

```
163 File indexed, time taken: 297 ms
0 documents found. Time :13
```

5.3 WithStopNoStem

Teste na aplicação quando não são removidas as stopwords e também não é utilizado *Stemming*.

“Acne”:

```
163 File indexed, time taken: 257 ms
10 documents found. Time :19
File: /home/joao/workspace/Web-Mining-A1/doc_base/doc13.txt
File: /home/joao/workspace/Web-Mining-A1/doc_base/doc6.txt
(...)
```

“In the”:

```
163 File indexed, time taken: 264 ms
161 documents found. Time :19 (...)
```

5.4 WithStopWithStem

Por fim, teste na aplicação quando não são removidas as stopwords mas é utilizado o processo de *stemming*.

“Acne”:

```
163 File indexed, time taken: 237 ms
18 documents found. Time :27
File: /home/joao/(...)/doc_base_stemed/doc103.stem
(...)
```

“In the”:

```
163 File indexed, time taken: 253 ms
161 documents found. Time :29
(...)
```

Com esses resultados, pode-se formar duas matrizes de resultados, uma para cada consulta.

“Acne”:

	Precisão	Cobertura	F-measure
NoStopNoStem	1	0,947368421	0.972973
NoStopWithStem	1	0,947368421	0.972973
WithStopNoStem	1	0,526315789	0.689655
WithStopWithStem	1	0,947368421	0.972973

Tabela 2: matriz de resultado da consulta 1 (“acne”)

“In the”:

	Precisão	Cobertura	F-measure
NoStopNoStem	0	0	nan
NoStopWithStem	0	0	nan
WithStopNoStem	1	0,987730061	0.993827
WithStopWithStem	1	0,987730061	0.993827

Tabela 3: matriz de resultado da consulta 2 (“in the”)

6. Conclusão

Esse projeto serviu como um primeiro contato (prático) ao mundo dos buscadores web.

Inicialmente, criou-se uma ferramenta para obtenção de informação, montando-se uma base de documentos a partir disso. Após essa fase, os arquivos foram indexados seguindo diferentes metodologias discutidas em aula, e um estudo comparativo foi realizado entre elas.