Problemas de Programação - Projeto 2

Recuperação da Informação - IF962

Douglas Soares Lins **(dsl)**
Jônatas de Oliveira Clementino **(joc)**
Valdemiro Rosa Vieira Santos **(vrvs)**

CIn - UFPE
Recife - 5 de Julho de 2018

1.

Criar Arquivo Invertido

Douglas Soares - dsl

Pré-Processamento

- NLTK
  - Tokenização
  - Stop-words
  - Porter Stemmer
- Filtros adicionais
  - Regex - ‘[0-9]*?\.’ | ‘[a-z]*?\.[a-z]*?’
  - Ignorar palavras no formato ascii

Palavras em geral e time limit.

- Atributos
  - Título
    - Title
  - Descrição do problema
    - Description
  - Descrição do input
    - Input, Input Descritpion, Input Format, INPUT
  - Descrição do output
    - Output, Output Descritpion, Output Format, OUTPUT
  - Tempo limit
    - Time Limit
  - Corpo do problema
    - Problem

- **Discretizar Time Limit**

  - Foi utilizado primeiro, o Octil para dividir o conjunto

  - Conjunto dividido por senso dos membros do grupo

0-0.5 0.5-1 1-2 2-3 3-4 4-5

5-6 6-7 7-8 8-9 9-10 10-15

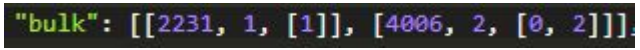
15-20 20-30 30-

Tipos de índice invertido

- Básico

- Frequência

- Posicional

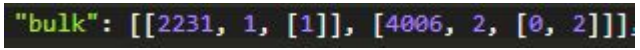


Versões de Compressão

- **Sem compressão**
  - Básico

  - Frequência

  - Posicional



Versões de Compressão

- **Com compressão**
  - Básico

```
"differ": [1231, 81, 1703, 249, 1032, 117, 359],
```

  - Frequência

```
"reduct": [[3275, 1], [1085, 2]],
```

  - Posicional

```
"bulk": [[2231, 1, [1]], [1775, 2, [0, 2]]],
```

Comparação

Arquivo	Tipo de Index	Sem Compressão	Com Compressão
Title	Basic	118 KB	105 KB
	Frequency	178 KB	165 KB
	Positional	239 KB	225 KB
Description	Basic	1.58 MB	1.12 MB
	Frequency	2.77 MB	2.31 MB
	Positional	5.71 MB	5.25 MB
Input Description	Basic	655 KB	435 KB
	Frequency	1.13 MB	940 KB
	Positional	2.19 MB	1.97 MB
Output Description	Basic	463 KB	325 KB
	Frequency	804 KB	665 KB
	Positional	1.37 MB	1.23 MB
Time Limit	Basic	36.7 KB	19.7 KB
	Frequency	68 KB	51 KB
	Positional	99.3 KB	82.3 KB
Problem	Basic	3.39 MB	2.35 MB
	Frequency	5.87 MB	4.83 MB
	Positional	11.8 MB	10.8 MB

Codificação de Tamanho Variável

docIDs	824	829	215406
gaps		5	214577
VB code	00000110 10111000	10000101	00001101 00001100 10110001

Arquivo	Tipo de Index	Sem Compressão	Com Compressão	Variant Size		Serializado	
				Sem Compressão	Com Compressão	Sem Compressão	Com Compressão
Title	Basic	118 KB	105 KB	119 KB	114 KB	132 KB	127 KB
	Frequency	178 KB	165 KB	131 KB	126 KB	252 KB	246 KB
	Positional	239 KB	225 KB	143 KB	138 KB	360 KB	355 KB
Description	Basic	1.58 MB	1.12 MB	896 KB	706 KB	1.11 MB	943 KB
	Frequency	2.77 MB	2.31 MB	1.11 MB	949 KB	3.48 MB	3.29 MB
	Positional	5.71 MB	5.25 MB	1.87 MB	1.68 MB	6.45 MB	6.27 MB
Input Description	Basic	655 KB	435 KB	340 KB	257 KB	441 KB	356 KB
	Frequency	1.13 MB	940 KB	441 KB	358 KB	1.41 MB	1.33 MB
	Positional	2.19 MB	1.97 MB	680 KB	597 KB	2.59 MB	2.51 MB
Output Description	Basic	463 KB	325 KB	264 KB	211 KB	333 KB	279 KB
	Frequency	804 KB	665 KB	332 KB	279 KB	1010 KB	959 KB
	Positional	1.37 MB	1.23 MB	461 KB	408 KB	1.72 MB	1.67 MB
Time Limit	Basic	36.7 KB	19.7 KB	12.7 KB	6.61 KB	18.8 KB	12.9 KB
	Frequency	68 KB	51 KB	18.9 KB	12.9 KB	80.8 KB	74.8 KB
	Positional	99.3 KB	82.3 KB	25.2 KB	19.1 KB	137 KB	131 KB
Problem	Basic	3.39 MB	2.35 MB	1.97 MB	1.56 MB	2.45 MB	2.05 MB
	Frequency	5.87 MB	4.83 MB	2.46 MB	2.05 MB	7.39 MB	6.99 MB
	Positional	11.8 MB	10.8 MB	4.11 MB	3.71 MB	13.4 MB	13 MB

2.

Processamento da Consulta

Valdemiro Vieira - vrvs

Ranking

- **Modelo Espaço Vetorial**
  - Com e sem tfidf
- **Proximidade de Termos**
  - Considera as menores distâncias dois a dois das palavras da consulta no texto
  - Segunda prioridade foi utilizado o Modelo Espaço Vetorial
- **Ler postings**
  - Document-at-a-time

Consultas

- 1 - Graph Search
- 2 - Fibonacci Sum
- 3 - Dynamic Programming
- 4 - Card Game
- 5 - Bit Manipulation

Resultados - Spearman

Domínio	Sem tfidf - Proximidade	Com tfidf - Sem tfidf	Proximidade - Com tfidf
Consulta 1	0.7479	0.7488	0.9999
Consulta 2	0.9721	0.9721	0.9999
Consulta 3	0.9898	0.9898	0.9999
Consulta 4	0.9744	0.9744	0.9997
Consulta 5	0,8251	0.8251	0.9999

Resultados - Kendal Tau

Domínio	Sem tfidf - Proximidade	Com tfidf - Sem tfidf	Proximidade - Com tfidf
Consulta 1	0.7265	0.7270	0.9995
Consulta 2	0.9738	0.9738	0.9994
Consulta 3	0.9885	0.9886	0.9999
Consulta 4	0.9678	0.9672	0.9965
Consulta 5	0.8390	0.8388	0.9998

3.
Composição da Resposta

Jônatas Clementino - joc

Arquitetura Website

- Web Framework Django de python para criar o website;
- Dados armazenados utilizando sqlite3 para rápido acesso;
- 4 telas principais:
  - Home
  - Problems
  - Search
  - About us
- Interface feita utilizando em sua maior parte Bootstrap

Interface do Search

- 6 Campos de Busca:
  - Título
  - Descrição do Problema
  - Input
  - Output
  - Time Limit
  - Campo Livre (Pesquisa em todo o problema)
- Em cada problema as informações apresentadas foram: Título (âncora), Time Limit e a Descrição do Problema resumida em no máximo duas linhas.

Mutual Information

- Calculado mutual information, baseado numa janela de tamanho 5;
- Escolhido 3 palavras para cada atributo, exceto Time Limit, que possuem maior mutual information;

Mutual Information

Atributo	Palavras	Mutual-Information
Title	'stage'	0.0091
	'game'	0.0066
	'numbers'	0.0058
Input	'line'	0.0537
	'contains'	0.0408
	'number'	0.0351

Mutual Information

Atributo	Palavras	Mutual-Information
Output	'output'	0.0545
	'number'	0.0325
	‘line’	0.0315
Description	'one'	0.0135
	'number'	0.0113
	'two'	0.0064

Obrigado!

Douglas Soares Lins **(dsl)**
Jônatas de Oliveira Clementino **(joc)**
Valdemiro Rosa Vieira Santos **(vrvs)**