Relatório EP 2 - MAC 323

Julia Leite

20 de julho de 2020

Resumo

Esse exercício programa (EP) foi proposto na matéria MAC0323 do Instituto de Matemática e Estatística da Universidade de São Paulo, ministrada pelo professor Carlos Eduardo Ferreira. O objetivo é contruir um grafo de textos e realizar experimentos em dicionários de palavras, observando características como conexidade, distância média entre palavras, entre outras, em diferentes tipos de textos.

Sumário

1 Aluna:

- **Nome:** Julia Leite
- **NUSP:** 11221797

2 Implementação

Nesse exercício programa (EP) implementamos a classe Grafo, no arquivo **grafo.hpp**, inspirado na interface fornecida pelos monitores. Cada vértice do grafo pertence à classe Celula, também definida nesse arquivo. A lista de vértices do grafo é armazenada em um *vector* da biblioteca STL, do C++, assim como a lista de vértices adjacentes a um vértice do grafo.

No arquivo **main.cpp**, lemos o arquivo, inserimos as palavras no grafo, exibimos na tela uma análise das propriedade desse e realizamos testes iterativos.

Ao inserir as palavras no grafo removemos pontuação e números, como `.,()#0123456789`, com a função *limpa()*. Há, contudo, algumas exceções como “ ”.

A análise do grafo é feita pela função *analise()* que exibe na tela o número de vértices, arestas e componentes do grafo, além de informações como o grau médio dos vértices do grafo, tamanho médio das componentes, densidade do grafo, entre outras.

Observação: calculamos o grau médio dos vértices (g) com:

$$g = \frac{2A}{V}$$

E o a densidade do grafo d com:

$$d = \frac{2A}{V(V-1)}$$

Sendo A o número de arestas e V o de vértices.

É possível, então, realizar testes iterativos pelo prompt de comando.

3 Instruções:

3.1 Compilação:

Utilizamos um Makefile com as seguintes flags: -Wall -g -O0 . Para compilar, basta digitar no terminal:

```
make main
```

3.2 Execução:

Para executar o EP é necessário passar como argumento por linha de comando o número mínimo de letras de uma palavra do grafo (k) e o nome do arquivo (.txt)

```
./main <k> <nome do arquivo>
```

Caso o arquivo não esteja na pasta ou o usuário não digite um dos argumentos a execução é encerrada.

3.3 Testes iterativos:

Para testar o EP utilizamos o seguinte conjunto de instruções:

0. `manual`: exibe a lista de comandos dos testes iterativos
1. `show`: exibe o grafo no prompt de comando
2. `lista`: exibe a lista de palavras do grafo
3. `ciclo <palavra>`: informa se uma palavra está em um ciclo do grafo
4. `caminho <palavra 1> <palavra 2>`: informa se existe um caminho entre duas palavras
5. `dupla <palavra 1> <palavra 2>`: informa se duas palavras estão no mesmo ciclo
6. `distancia <palavra 1> <palavra 2>`: exibe a menor distancia entre 2 palavras do grafo
7. `END`: encerra o programa

Oberservação: nas instruções 3,4,5,6 será necessário digitar a(s) palavra(s) em seguida, exemplo:

```
caminho abacate abacaxi
```

4 Experimentos

Observação: o programa considera diferenças entre letras maiúsculas e minúsculas, dessa forma, palavras como *abacaxi* e *Abacaxi* são considerados vértices diferentes e adjacentes.

Testamos o programa nos textos:

- Conto Feliz Aniversário de Clarice Lispector[1]
- Romance romântico Iracema de José de Alencar[2]
- Reportagem sobre Feminismo do Jornal Nexo[3]
- Texto técnico da IBM[4]

[1] `https://cs.ufgd.edu.br/download/Lacos%20de%20Familia%20-%20Clarice%20Lispector.pdf`
[2] `http://objdigital.bn.br/Acervo_Digital/Livros_eletronicos/iracema.pdf`
[3] `https://www.nexojornal.com.br/explicado/2020/03/07/Feminismo-origens-conquistas-e-desafios-no-s%C3%A9culo-21`
[4] `http://www.gutenberg.org/cache/epub/27468/pg27468.txt`

4.1 Feliz Aniversário - Clarice Lispector

Testamos o programa com o conto **Feliz Aniversário**, parte da obra Laços de Família

Com um tamanho mínimo de palavra $\mathbf{k = 2}$, temos:

1. **Número de vértices**: 1339
2. **Número de arestas**: 539
3. **Grau médio dos vértices**: 0.8
4. **Densidade**: 0.0006
5. **Número de componentes**: 961
6. **Tamanho médio das componentes**: 1.39
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 97
10. **Distância média entre palavras**: 5.83

Com um tamanho mínimo de palavra $\mathbf{k = 3}$, temos:

1. **Número de vértices**: 1308
2. **Número de arestas**: 431
3. **Grau médio dos vértices**: 0.65
4. **Densidade**: 0.0005
5. **Número de componentes**: 966
6. **Tamanho médio das componentes**: 1.35
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 73
10. **Distância média entre palavras**: 6.08

Testamos, então, aumentando o k:

Com um tamanho mínimo de palavra $\mathbf{k = 4}$, temos:

1. **Número de vértices**: 1234
2. **Número de arestas**: 302
3. **Grau médio dos vértices**: 0.49
4. **Densidade**: 0.0004
5. **Número de componentes**: 969
6. **Tamanho médio das componentes**: 1.27
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 30
10. **Distância média entre palavras**: 6.34

Com um tamanho mínimo de palavra $\mathbf{k = 5}$, temos:

1. **Número de vértices**: 1115
2. **Número de arestas**: 189
3. **Grau médio dos vértices**: 0.34
4. **Densidade**: 0.0003
5. **Número de componentes**: 934
6. **Tamanho médio das componentes**: 1.19
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 12
10. **Distância média entre palavras**: 1.79

Observamos, então, que aumentar o tamanho mínimo das palavras impacta no número de vértices e, principalmente arestas, além da progressiva redução do grau médio dos vértices e da densidade.

O tamanho médio das componentes, assim como o tamanho da maior componente diminuiu com o aumento do k.

Já o número de componentes e a distancia média entre as palavras aumentaram quando $k \leq 4$ e reduziram quando $k = 5$. Isso pode indicar que as componentes dos grafos gerados anteriormente tinham uma proporção razoável de palavras com menos que 5 letras.

4.2 Iracema - José de Alencar

Testamos, também, com a obra romântica **Iracema** escrita por José de Alencar.

Com um tamanho mínimo de palavra $\mathbf{k = 2}$, temos:

1. **Número de vértices**: 4402
2. **Número de arestas**: 2971
3. **Grau médio dos vértices**: 1.34
4. **Densidade**: 0.0003
5. **Número de componentes**: 2426
6. **Tamanho médio das componentes**: 1.81
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 960
10. **Distância média entre palavras**: 9.25

Com um tamanho mínimo de palavra $\mathbf{k = 3}$, temos:

1. **Número de vértices**: 4362
2. **Número de arestas**: 2789
3. **Grau médio dos vértices**: 1.27
4. **Densidade**: 0.0003
5. **Número de componentes**: 2428
6. **Tamanho médio das componentes**: 1.79
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 922
10. **Distância média entre palavras**: 9.6

Com um tamanho mínimo de palavra $\mathbf{k = 4}$, temos:

1. **Número de vértices**: 4208
2. **Número de arestas**: 2340
3. **Grau médio dos vértices**: 1.11
4. **Densidade**: 0.00026
5. **Número de componentes**: 2442
6. **Tamanho médio das componentes**: 1.72
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 731
10. **Distância média entre palavras**: 10.53

Com um tamanho mínimo de palavra $\mathbf{k = 5}$, temos:

1. **Número de vértices**: 3879
2. **Número de arestas**: 1720
3. **Grau médio dos vértices**: 0.88
4. **Densidade**: 0.00022
5. **Número de componentes**: 2463
6. **Tamanho médio das componentes**: 1.57
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 368
10. **Distância média entre palavras**: 10.46

Observamos que houve uma redução do número de arestas e vértices com o aumento do k e, consequentemente, uma diminuição o grau médio dos vértices e da densidade.

Podemos perceber que a quantidade de componentes aumenta e o tamanho da maior componente diminui conforme o elevamos o k.

O aumento da distância média entre as palavras evidencia que, comparando com o exemplo anterior (Feliz Aniversário), as componentes conexas dos grafos desse texto são formadas, em maior proporção, por palavras com mais de 4 ou 5 letras.

4.3 Reportagem

Testamos, também na reportagem do jornal Nexo sobre Feminismo.

Com um tamanho mínimo de palavra $\mathbf{k = 2}$, temos:

1. **Número de vértices**: 2899
2. **Número de arestas**: 949
3. **Grau médio dos vértices**: 0.65
4. **Densidade**: 0.0002
5. **Número de componentes**: 2177
6. **Tamanho médio das componentes**: 1.33
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 160
10. **Distância média entre palavras**: 6.95

Com um tamanho mínimo de palavra $\mathbf{k = 3}$, temos:

1. **Número de vértices**: 2862
2. **Número de arestas**: 840
3. **Grau médio dos vértices**: 0.58
4. **Densidade**: 0.0002
5. **Número de componentes**: 2177
6. **Tamanho médio das componentes**: 1.31
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 94
10. **Distância média entre palavras**: 5.3

Com um tamanho mínimo de palavra $\mathbf{k = 4}$, temos:

1. **Número de vértices**: 2773
2. **Número de arestas**: 689
3. **Grau médio dos vértices**: 0.49
4. **Densidade**: 0.0001
5. **Número de componentes**: 2176
6. **Tamanho médio das componentes**: 1.27
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 24
10. **Distância média entre palavras**: 2.11

Com um tamanho mínimo de palavra $\mathbf{k = 5}$, temos:

1. **Número de vértices**: 2594
2. **Número de arestas**: 519
3. **Grau médio dos vértices**: 0.4
4. **Densidade**: 0.0001
5. **Número de componentes**: 2114
6. **Tamanho médio das componentes**: 1.22
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 16
10. **Distância média entre palavras**: 1.65

Assim como nos textos anteriores, a elevação do k se refletiu na redução do número de vértices, da quantidade de arestas, do grau médio dos vértices, da densidade do grafo e do tamanho médio das componentes.

Nesse texto, quando aumentamos o k de 2 para 3, observamos que, apesar da redução da quantidade de vértices e arestas o número de componentes não se altera. As demais elevações de k resultam na redução do número de componentes.

Observamos uma progressiva redução da distância média entre as palavras, isso indica que uma parte significativa das componentes conexas do grafo são formadas por palavras com poucas letras, um possível reflexo da linguagem formal, porém não rebuscada característica de textos jornalísticos.

É importante ressaltar que nesse texto ocorrem algumas “ ”, nesses casos, o programa as interpreta como letras, então os resultados com $k \leq 3$ podem apresentar palavras como: “A” ou “A.

4.4 Texto técnico - IBM

Observamos o comportamento do EP no texto técnico sobre o computador da IBM

Com um tamanho mínimo de palavra $\mathbf{k = 2}$, temos:

1. **Número de vértices**: 523
2. **Número de arestas**: 173
3. **Grau médio dos vértices**: 0.66
4. **Densidade**: 0.001
5. **Número de componentes**: 382
6. **Tamanho médio das componentes**: 1.36
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 28
10. **Distância média entre palavras**: 2.68

Com um tamanho mínimo de palavra $\mathbf{k = 3}$, temos:

1. **Número de vértices**: 500
2. **Número de arestas**: 123
3. **Grau médio dos vértices**: 0.49
4. **Densidade**: 0.0009
5. **Número de componentes**: 388
6. **Tamanho médio das componentes**: 1.28
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 10
10. **Distância média entre palavras**: 1.86

Com um tamanho mínimo de palavra $\mathbf{k = 4}$, temos:

1. **Número de vértices**: 468
2. **Número de arestas**: 99
3. **Grau médio dos vértices**: 0.42
4. **Densidade**: 0.0009
5. **Número de componentes**: 373
6. **Tamanho médio das componentes**: 1.25
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 9
10. **Distância média entre palavras**: 1.8

Com um tamanho mínimo de palavra $\mathbf{k = 5}$, temos:

1. **Número de vértices**: 396
2. **Número de arestas**: 72
3. **Grau médio dos vértices**: 0.36
4. **Densidade**: 0.0007
5. **Número de componentes**: 352
6. **Tamanho médio das componentes**: 1.21
7. **Conexo?** não
8. **Tamanho da menor componente**: 1
9. **Tamanho da maior componente**: 8
10. **Distância média entre palavras**: 1.6

Observamos, que houve uma redução no número de vértices, arestas, grau médio, densidade do grafo e do tamanho médio das componentes com o aumento do k. Além de uma diminuição progressiva da distância mínima entre as palavras.

Quando elevamos o k de 2 para 3, observamos um aumento na quantidade de componentes do grafo, as demais elevações do k resultaram na diminuição da quantidade de componentes.

Percebemos que esse grafo é significativamente mais denso que os exemplos anteriores, além disso a distância média entre as palavras é consideravelmente menor que os anteriores.

5 Comparando os tipos de texto

Comparamos os grafos gerados pelos 4 textos selecionados, utilizando um tamanho mínimo de palavra, k = 4:

	Feliz Aniversário	Iracema	Reportagem	Texto Técnico
Nº vértices	1234	4208	2773	468
Nº arestas	302	2340	689	99
Grau medio dos vértices	0.49	1.11	0.49	0.42
Densidade	0.0004	0.0002	0.0001	0.0009
Nº componentes	969	2442	2176	373
Tamanho Médio Componentes	1.37	1.72	1.27	1.25
Distância média entre palavras	6.34	10.53	2.11	1.8

Tabela 1: Comparação entre os textos

Podemos perceber que o romance Iracema possui um grau médio mais elevado entre os textos. Logo em seguida, temos o conto Feliz Aniversário e a reportagem com o mesmo grau médio, destacamos, contudo, que o grau médio desses significativamente menor que o do romance. O texto técnico, então, apresenta o menor grau médio.

Analisando a densidade dos grafos, observamos que o texto técnico tem maior densidade, seguido pelo conto, romance e reportagem, respectivamente.

Já em relação ao tamanho médio das componentes e distância média entre palavras, observamos que Iracema apresenta os índices mais elevados, logo após temos o Feliz aniversário, a Reportagem e o texto técnico.

Observamos, então, que o romance resultou em um grafo com um grau médio, tamanho de componentes e distância média mais elevados que os outros tipos textuais, um reflexo, talvez, da linguagem mais rebuscada empregada no texto.

Já o grafo formado pelo texto técnico apresenta a menor distância média entre palavras, e uma maior densidade, reflexo, talvez, do vocabulário mais técnico empregado.

Concluímos que a linguagem empregada no texto, reflexo do tipo textual, pode interferir nas propriedades do grafo, como densidade e grau médio dos vértices.