Trabalho Prático 2 - AED

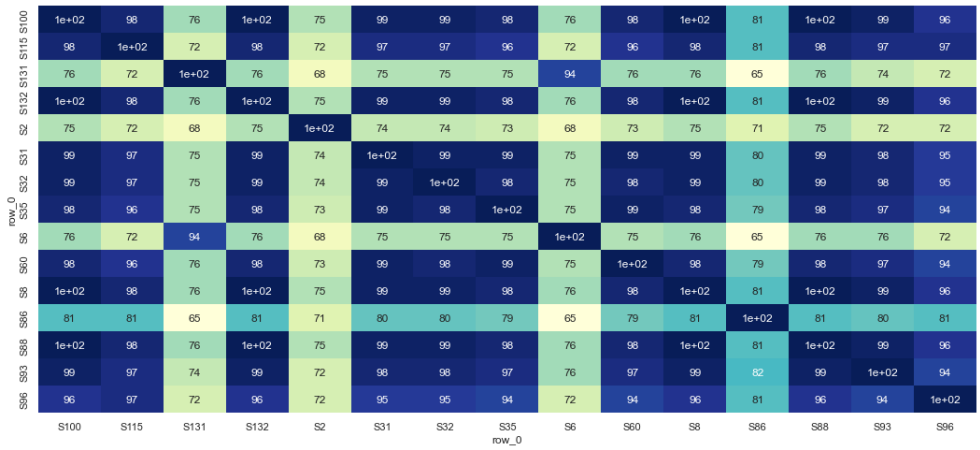
Este trabalho prático tem como objetivo desenvolver um sistema para a deteção de cópias, partindo de um *dataset* com correlações de código fonte e utilizando para a análise, os grafos como estrutura de dados.

1-Descrição do problema

Os estudantes submetem código fonte para a resolução de problemas de programação em plataformas Web. A análise ao código submetido, através de algoritmos que detetam a correlação entre submissões, pode permitir detetar cópias e identificar grupos com código semelhante.

A partir de um *dataset* com informação sobre submissões, é pretendido encontrar as correlações e detetar grupos com elevada probabilidade de terem copiado.

Observe a seguinte figura, que mostra a correlação entre submissões, utilizando como visualização, o mapa das temperaturas.



Uma distância de *Levenshtein* adaptada foi utilizada para calcular a correlação entre o código fonte das submissões. O ordem cronológica das submissões permite a utilização de grafos orientados, com identificação temporal entre submissões.

Após a construção de um grafo orientado, contendo os dados apresentados, deve ser realizado o seguinte passo, que é a análise do grafo para a identificação de grupos com elevada probabilidade de terem copiado. Considere a colocação no grafo apenas, das submissões com mais de 80% de correlação entre si.

2-Dados fornecidos

Para o desenvolvimento deste trabalho prático, são fornecidas as seguintes tabelas e um exemplo de utilização.

1. Correlação entre submissões, `correl.csv`:

```
row_0,S100,S115,S131,S132,S2,S31,S32,S35,S6,S60,S8,S86,S88,S93,S96
S100,100,98,76,100,75,99,99,98,76,98,100,81,100,99,96
S115,98,100,72,98,72,97,97,96,72,96,98,81,98,97,97
S131,76,72,100,76,68,75,75,75,94,76,76,65,76,74,72
S132,100,98,76,100,75,99,99,98,76,98,100,81,100,99,96
S2,75,72,68,75,100,74,74,73,68,73,75,71,75,72,72
S31,99,97,75,99,74,100,99,99,75,99,99,80,99,98,95
S32,99,97,75,99,74,99,100,98,75,98,99,80,99,98,95
S35,98,96,75,98,73,99,98,100,75,99,98,79,98,97,94
S6,76,72,94,76,68,75,75,75,100,75,76,65,76,76,72
S60,98,96,76,98,73,99,98,99,75,100,98,79,98,97,94
S8,100,98,76,100,75,99,99,98,76,98,100,81,100,99,96
S86,81,81,65,81,71,80,80,79,65,79,81,100,81,80,81
S88,100,98,76,100,75,99,99,98,76,98,100,81,100,99,96
S93,99,97,74,99,72,98,98,97,76,97,99,82,99,100,94
S96,96,97,72,96,72,95,95,94,72,94,96,81,96,94,100
```

2. Ordem cronológica de submissões, `time.csv`:

```
id,time
S2,10:37:00
S6,10:56:00
S8,11:14:00
S31,11:36:00
S32,11:39:00
S35,11:45:00
S60,12:09:00
S86,12:27:00
S88,12:29:00
S93,12:33:00
S96,12:35:00
S100,12:37:00
S115,12:47:00
S131,12:59:00
```

```
S132,12:59:00
```

3. Exemplo de utilização da aplicação:

```
Caso de uso, copy-detection:

copy-detection --correlation=correl.csv --time=time.csv
    --threshold=80

O resultado dever ser:

1: S2 (100%)
2: S6 S131 (97%)
3: S8 S31 S32 S35 S60 S86 S88 S93 S96 S100 S115 S132 (97.33%)
```

3-Tarefa

Dados dois *datasets*, contendo a correlação entre as várias submissões e a ordem cronológica de entrada no sistema, pretende-se obter os grupos com maior semelhança entre si, para identificar possíveis cópias.

Note que a percentagem obtida para cada grupo é a média de correlação entre todos os nós do grupo.

É requisito técnico a utilização de grafos, para a análise do problema.

Os *datasets* devem ser lidos desde o formato cvs (*comma separate values*).

Deve apresentar como output, o resultado do agrupamento, diretamente na consola.

Faça uma análise ao algoritmo, sobre a robustez (aplica-se em todos os casos) e eficiência (o processamentos dos dados é rápido ou existem otimizações a realizar...).

Finalmente, analise os dados, discuta os resultados e conclua sobre o projeto.

3-Requisitos

(a) Utilize os princípios e técnicas estudadas em AED.

(b) Siga uma metodologia de desenvolvimento baseado em testes (*TDD - Test Driven Development*).

4-Relatório

O TP1 incluiu o desenvolvimento de um relatório, que deve conter:

(a) Uma descrição do problema e a abordagem do grupo para a sua resolução.

(b) O desenho de um diagrama de classes, com a análise realizada.

(c) O pseudo-código do algoritmo desenvolvido.

(d) Resultados.

(e) Discussão.

(f) Conclusão.

5-Entrega

O trabalho prático deve ser entregue em formato digital, incluindo o código do projeto e o relatório, com indicação dos nomes dos elementos do grupo (máximo 2 elementos), para o email *tiago.candeias@ismat.pt*, até ao dia 26-06-2023.

6-Avaliação

Esta parte do trabalho vale 50% da nota da componente prática da unidade curricular. Os critérios de avaliação são: clareza do texto, simplicidade do algoritmo, clareza das explicações, qualidade do código e comentários no código.