Informática para Ciências e Engenharias-B 2017/18

Trabalho Prático Nº 2 — 2017/18

1 Objectivo do Trabalho

Até 1982, a insulina disponível para tratamento de diabetes era obtida por purificação de tecidos animais. Hoje, a insulina humana é produzida em microorganismos geneticamente modificados como *Escherichia coli* e *Saccharomyces cerevisiae*. Investigadores de um laboratório estão a tentar optimizar as estirpes e meio de cultura para maximizar a produção de insulina humana recombinante. Infelizmente, esses investigadores não tiveram informática na sua formação e estão com dificuldade em processar os dados. Eles têm ficheiros de texto com informação acerca das experiências que fizeram. Em cada experiência foi usado um reactor químico para cultivar um lote de microorganismos e medir a concentração de insulina ao longo do tempo. Os ficheiros identificam cada lote, a estirpe, o meio de cultura, a temperatura, e concentração de insulina ao longo do tempo. Agora os investigadores precisam de organizar tudo numa base de dados, calcular estatísticas e gerar gráficos mas não sabem como.

O objectivo deste trabalho é criar um programa capaz de interpretar um ficheiro de texto onde os investigadores, que não sabem programar, possam facilmente especificar que operações querem fazer com os seus dados. Executando as instruções nesse ficheiro, o programa deverá organizar os dados numa base de dados e usá-la para obter as estatísticas e gráficos que os investigadores especificarem.

2 Descrição do Problema

Pretende-se um programa cuja função principal se chame `processar` e receba o nome do ficheiro de texto com as instruções de processamento, fornecido pelos investigadores. O programa deve ler e executar os comandos descritos nesse ficheiro, que deve conter os seguintes comandos:

- `BASE_DADOS ficheiro`

  Este comando deve preceder qualquer comando que interaja com a base de dados e especifica o nome do ficheiro de bases de dados que vai ser usado. Esse ficheiro pode já existir ou pode ser criado pelo sistema de gestão de bases de dados. Note que a base de dados pode mudar, havendo mais que um comando deste tipo. Se o programa encontrar outro comando destes deve fechar a ligação à base de dados anterior e criar uma ligação nova à base de dados corrente.

- `REPORT ficheiro`

  Este comando deve preceder qualquer comando que crie relatórios e especifica o nome do ficheiro de relatório que irá ser preenchido. Quando o programa encontrar este comando deve criar um ficheiro com o nome indicado, ou limpar todo o conteúdo desse ficheiro se ele existir. Se o programa encontrar um comando adicional de `REPORT` deve fechar o acesso ao ficheiro anteriormente especificado e criar o novo ficheiro.

- `CRIAR_TABELAS`

  Ao encontrar este comando o programa deve criar duas tabelas na base de dados cujo nome foi especificado no comando `BASE_DADOS`:

- Tabela `Lotes` com campos para o identificador do lote (uma string única para cada lote), a estirpe (uma string), o código do meio de cultura (uma string) e a temperatura de incubação em graus Celsius (um número inteiro).
- Tabela `Amostras` com campos para o identificador de cada amostra (uma string única para cada amostra), o identificador do lote ao qual essa amostra pertence (uma string), o tempo, em minutos, desde o início da operação até à recolha dessa amostra (um número inteiro) e a concentração medida de insulina em g/L (um número real).

- `CARREGAR ficheiro`

  Ler um ficheiro com o nome indicado e carregar a informação para as tabelas da base de dados `nomeBD`. O ficheiro correspondente a cada relatório tem o seguinte formato:

  Código do lote na primeira linha
  Código da estirpe na segunda linha
  Código do meio na terceira linha
  Temperatura, em graus Celsius, na quarta linha
  Restantes linhas com o identificador de cada amostra, o tempo da amostra (em minutos) e a concentração medida de insulina (em g/L), separados pelo carácter `;`.

  Exemplo de um ficheiro a carregar:

```
B-00004
EC-010
CM-202
36
S-00010;44;0.90
S-00011;73;1.35
S-00012;83;1.57
S-00013;112;2.04
S-00014;136;2.69
S-00015;158;3.40
S-00016;173;3.43
S-00017;198;4.01
S-00018;222;4.76
...
S-00023;334;7.34
S-00024;344;7.56
```

  O identificador do lote, os identificadores das amostras, o tempo e concentração de insulina devem ser inseridos na tabela `Amostras`. A restante informação, bem como o identificador do lote, deve ser inserida na tabela `Lotes`.

- `ESTIRPES minT;maxT;meio`

  Escrever no ficheiro seleccionado para o relatório os códigos das estirpes que tenham sido cultivadas a temperaturas entre `minT` e `maxT` no meio especificado em `meio`. Estes três parâmetros estão separados por `;` e qualquer um pode ser substituido pelo carácter `*` indicando que deve ser ignorado (ou seja, que se aceita qualquer valor).

  Exemplo de comandos de resumo:

```
ESTIRPES 30,36;*
ESTIRPES *;*;*
ESTIRPES *;*;CM-206
```

Neste exemplo, o primeiro comando de resumo pede os códigos das estirpes cultivadas a temperaturas entre os 30 e 36 graus Celsius, qualquer que seja o meio de cultura. O segundo comando pede a listagem de todos as estirpes na base de dados. O terceiro pede os códigos das estirpes cultivadas no meio de cultura `CM-206`. A listagem das estirpes não deve conter repetições. Ver em 2.1 mais detalhes sobre o formato da informação a escrever no ficheiro de relatório.

- `GRAFICO ficheiro;estirpe`

  Criar um gráfico com todas as amostras obtidas para a estirpe indicada em **`estirpe`** e gravá-lo no ficheiro especificado em **`ficheiro`**. O gráfico deve apresentar no eixo das abcissas (x) o tempo em que cada amostra foi retirada, em horas, e no eixo das ordenadas (y) a concentração de insulina em grama por litro. Ou seja, querem ver a produtividade dessa estirpe. Deve também mostrar a recta de regressão linear para os dados apresentados, assumindo que a concentração inicial de insulina é zero. Seja esta recta:

  $y = \beta x$

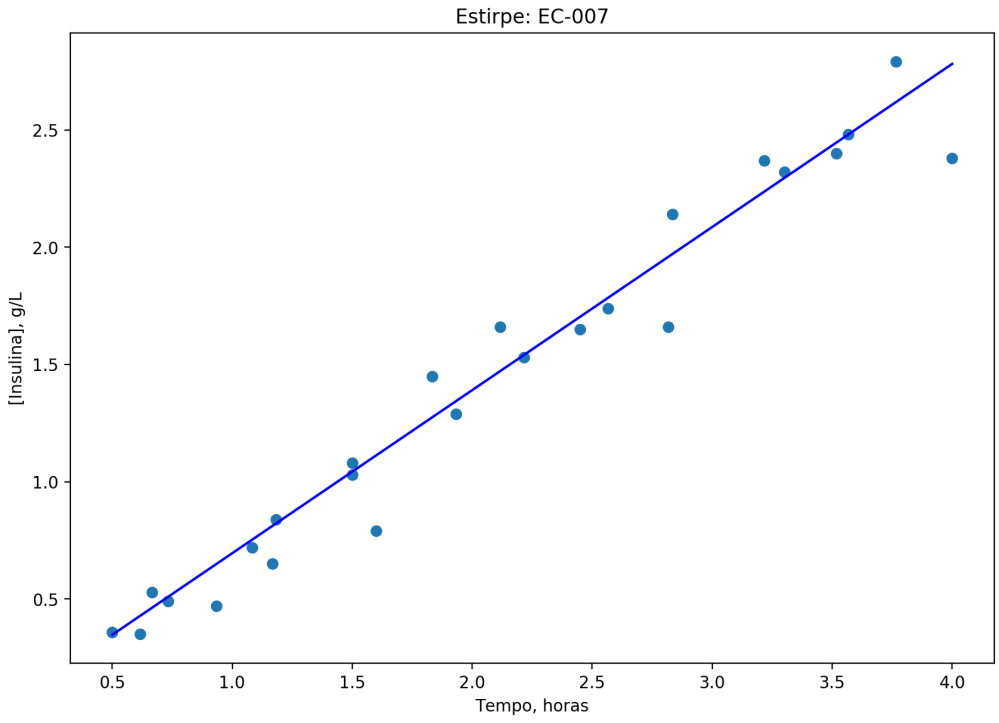
  o parâmetro β pode ser calculado pela seguintes expressão:

  $$\beta = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

  Onde x_i e y_i são os valores de tempo e tempo (em horas) e de concentração de insulina (grama por Litro de solução). A linha da regressão linear deve ser traçada desde o menor valor de tempo ao maior valor de tempo (em horas) nesse conjunto de pontos.

  Este gráfico ilustra o resultado esperado para a estirpe com o código `EC-007` usando os ficheiros fornecidos.

Minutos	[Insulina]
37	0.35
56	0.47
70	0.65
96	0.79
30	0.36
40	0.53
65	0.72
90	1.08
110	1.45
127	1.66
147	1.65
170	2.14
193	2.37
211	2.4
226	2.79
44	0.49
71	0.84
90	1.03
116	1.29
133	1.53
154	1.74
169	1.66
198	2.32
214	2.48
240	2.38



2.1 Ficheiros de relatório

Cada relatório, num ficheiro especificado no comando `REPORT`, tem várias linhas por cada comando do tipo `ESTIRPES` que ocorre no ficheiro dos comandos. Para cada comando `ESTIRPES` deve haver uma linha indicando quantas estirpes satisfazem as condições e quais as condições impostas. De seguida deve haver uma linha por cada estirpe que satisfaz as condições impostas, indicando respectivo código. Note que pode haver mais do que um comando `ESTIRPES` e que os resultados de todos os comandos `ESTIRPES` devem constar no ficheiro indicado no comando `REPORT`. Se houver um segundo comando `REPORT`, os comandos `ESTIRPES` que vierem a seguir devem escrever no novo ficheiro de relatório.

Este é o conteúdo esperado do ficheiro de relatório `report.txt` com os dados e comandos fornecidos juntamente com este enunciado no ficheiro `teste.txt`

```
2 estirpes cultivadas com o meio CM-206 entre 28°C e 30°C
EC-013
EC-007
3 estirpes cultivadas com o meio CM-206 entre *°C e *°C
EC-010
EC-013
EC-007
4 estirpes cultivadas com o meio * entre 29°C e 30°C
EC-011
EC-008
EC-013
EC-007
```

2.2 Opcional: Determinar a melhor combinação

Os investigadores estão também interessados em descobrir a combinação de estirpe, meio e temperatura que maximiza a produtividade. Para resolver este problema, devem considerar todas as combinações testadas destes três parâmetros, reunir todos os dados disponíveis para cada combinação (que podem estar em vários lotes) e calcular, para cada combinação, a taxa de produção de insulina, dada pelo declive (β) da regressão linear da concentração (g/L) em função do tempo (em horas). No final de processar os comandos todos do ficheiro de comandos, o programa deve escrever no último ficheiro de relatório seleccionado uma linha indicando a melhor combinação de estirpe, temperatura e meio de cultura, e a taxa de produção em grama por Litro por hora:

`Melhor:EC-010, 36°C, CM-202 com 1.33g/L/h`

Notem que esta alínea vale apenas um valor e é um pouco mais complexa do que outras partes do trabalho que contam mais para a avaliação. Recomendo que não a resolvam a menos que já tenham feito o restante trabalho todo.

3 Dados do Trabalho

O arquivo `trabalho2.zip` tem os ficheiros `teste.txt` e os ficheiros `N.txt`, com N variando de 1 a 19, que podem ser usados como exemplos para testar o seu programa. Notem, no entanto, que estes testes não são exaustivos. É aconselhável fazerem também os vossos testes.

4 Entrega do Trabalho

O trabalho é entregue em formato .zip para o endereço `praticasice@gmail.com`, por ambos os elementos do grupo, cada um usando o se endereço oficial da FCT. Esse ficheiro .zip tem de conter pelo menos um ficheiro chamado `tp2.zip` que implemente a função `processar` pedida no enunciado. Este ficheiro pode conter outros módulos caso queiram organizar o código em vários módulos mas não é necessário incluir outros ficheiros que não os módulos que os alunos criarem. Não incluam os ficheiros de dados, nem o Sqlite.

Ambos os elementos do grupo têm de enviar exactamente o mesmo ficheiro .zip. Se algum não enviar o ficheiro, assume-se que não fez o trabalho, não prejudicando a nota do cologa.

O prazo para a entrega deste trabalho termina ao meio dia do dia **3 de Junho**, mas é aconselhável entregar o trabalho pelo menos um dia antes para poderem resolver qualquer problema com a entrega. Os alunos são responsáveis pela entrega correcta do trabalho.

5 Critérios de Avaliação do Trabalho

De acordo com o Regulamento de Avaliação de Conhecimentos da FCT/UNL,[1] os estudantes directamente envolvidos numa fraude são liminarmente reprovados na disciplina. Em ICE, considera-se que um aluno que **dá** ou que **recebe** código num trabalho comete fraude. Os alunos que cometerem fraude num trabalho não obterão frequência.

Os trabalhos serão avaliados de acordo com os seguintes critérios.

- Utilização correcta dos elementos básicos da linguagem Python.
- Decomposição adequada do problema em sub-problemas.
- Código legível e documentado.
- Criação correcta das tabelas na base de dados especificada.
- Inserção correcta dos dados dos ficheiros nas tabelas respectivas da base de dados.
- Criação e conteúdo correctos do relatório.
- Cálculo correcto da recta de regressão e criação do gráfico.
- Implementação genérica. O programa deve ser capaz de processar os dados mesmo que varie o nome da base de dados, o nome do ficheiro de comandos ou o nome do ficheiro com os resultados, e mesmo que o ficheiro de comandos especifique várias bases de dados e ficheiros de relatórios. Só deve assumir que:
  - as tabelas da base de dados são `Lotes` e `Amostras`, com os campos indicados no enunciado;
  - os comandos a processar são dos tipos indicados;
  - os ficheiros referidos nos comandos `CARREGAR` existem;
  - qualquer comando `GRAFICO` referirá uma estirpe que existe na base de dados;
  - nenhum comando `CARREGAR` irá aparecer antes do comando `CRIAR_TABELAS`.

[1] Em `http://www.fct.unl.pt/sites/default/files/documentos/estudante/informacao_academica/Reg_Aval.pdf`

– os comandos `REPORT` e `BASE_DADOS` aparecem pelo menos uma vez antes de ser necessário criar relatórios ou aceder à base de dados.

Para a nota do trabalho ser 20, tudo tem de estar certo. Mas a nota do trabalho será um número entre zero e vinte. Portanto, se o programa for decomposto em funções, a incorrecção de uma função não deve impedir a programação das outras. Por exemplo, é preferível fazer um programa que processe só alguns comandos a não fazer programa nenhum.