Informática para Ciências e Engenharias-B 2018/19

Trabalho Prático Nº 2 — 2018/19

1 Objectivo do Trabalho

Um grupo de historiadores está a estudar uma série de experiências antigas de química e precisa de ajuda a organizar os dados. Têm os apontamentos de um grupo de químicos que determinou as fórmulas empíricas de vários compostos orgânicos, alguns deles também caracterizados quanto aos seus atributos químicos (alcanos, ácidos carboxílicos, cetonas, etc.). Têm também listas dos compostos e das duas fórmulas químicas. A diferença entre uma fórmula química e uma fórmula empírica é que enquanto a fórmula química especifica o número de átomos de cada elemento que existe numa molécula do composto, a fórmula empírica apenas nos dá a proporção entre o número de átomos dos vários elementos. Um dos problemas que os historiadores têm de resolver é o de descobrir quais os compostos que podem ter sido identificados nas experiências cujos registos estão a estudar.

O objectivo deste trabalho é criar um programa capaz de interpretar um ficheiro de texto onde os historiadores, 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 historiadores 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 da base 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 garantir que a ligação à base de dados anterior é fechada e que as operações subsequentes sejam feitas na base de dados corrente.

- `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 `Compostos` com campos para o identificador do composto (um número inteiro único para cada composto), o nome do composto (uma string), a fórmula química do composto (uma string) e o ponto de ebulição Celsius (um número fraccionário).
  - Tabela `Atributos` com campos para o identificador do composto (um número inteiro) e o nome do atributo associado a esse composto (uma string). Como um composto pode ter vários atributos e vários compostos podem partilhar atributos, ambos os campos admitem valores repetidos. No entanto, não faz sentido que a tabela `Atributos` tenha

duas vezes o mesmo atributo para o mesmo composto, portanto o par formado pelo identificador do composto e o atributo não pode ser repetido na tabela.

- `CARREGAR ficheiro`

  Ler um ficheiro com o nome indicado e carregar a informação para as tabelas da base de dados `nomeBD`. O ficheiro tem, em cada linha, separados por `;`, os valores do identificador do composto, nome do composto, fórmula do composto, ponto de ebulição e, finalmente, os atributos químicos associados ao composto. Os atributos podem ser um ou mais e, no caso de serem vários, estarão separados por vírgula.

  Exemplo de um ficheiro a carregar:

```
1;Acetic Acid;C2H4O2;117.9;Carboxylic acid
2;Acetone;C3H6O;56.2;Ketone
3;Alanine;C3H7NO2;213;Amine,Carboxylic acid
4;Aniline;C6H7N;184.1;Amine,Aromatic
5;Benzene;C6H6;80.1;Aromatic
```

  O identificador do composto, o seu nome, fórmula e ponto de ebulição deve devem ser inseridos na tabela `Compostos`. Os atributos químicos do composto devem ser inseridos, em conjunto com o identificador do composto, na tabela `Atributos`.

- `REPORT nome_ficheiro`

  Especifica o ficheiro para o onde as listas de compostos correspondentes ao critério de selecção serão gravadas. Todos os comandos do tipo `LISTA` (ver abaixo) que sigam um comando `REPORT` devem acrescentar estas listas ao ficheiro especificado no comando `REPORT`. Se um novo ficheiro for especificado em `REPORT` então os comandos subsequentes do tipo `LISTA` deverão ser gravados nesse ficheiro. Por exemplo, neste caso:

```
REPORT relatorio.txt
LISTA CH;*
LISTA CH3;Alkane
LISTA CH2O;*
LISTA CH2O;Ester
REPORT relatorio2.txt
LISTA C3H7;*
LISTA C4H9;*
```

  os primeiros quatro comandos `LISTA` deverão resultar na escrita dos resultados no ficheiro `relatorio.txt` e os outros dois, a seguir ao segundo comando `REPORT`, deverão resultar na escrita do ficheiro `relatorio2.txt`.

- `LISTA formula_empirica;atributo`

  Escrever no ficheiro seleccionado no comando `REPORT` os nomes e fórmulas químicas dos compostos que tenham o atributo especificado e que sejam compatíveis com a fórmula empírica especificada. A seguir ao comando `LISTA` há um espaço, a fórmula empírica, um carácter de ponto e vírgula e o nome de um atributo. Notem que o nome do atributo pode conter o carácter espaço. Em alternativa, o nome do atributo pode ser substituído pelo asterisco, `*`, indicando qualquer atributo.

A fórmula empírica de um composto indica a proporção de átomos de cada elemento desse composto. Por exemplo, a fórmula química do benzeno é C_6H_6, porque cada molécula de benzeno tem 6 átomos de carbono e 6 de hidrogénio, mas a sua fórmula empírica é CH porque a proporção de átomos de carbono e de hidrogénio é de 1 para 1 e é esta a proporção que é determinada empiricamente.

A forma de determinar se uma fórmula química corresponde a uma fórmula empírica é verificar se tem os mesmos elementos e se a proporção de átomos entre a fórmula química e a fórmula empírica é a mesma para todos os elementos. No exemplo acima, do benzeno, a fórmula química tem seis átomos de carbono e a empírica tem um, numa proporção de 6/1. No caso do hidrogénio é o mesmo, uma proporção de 6/1, por isso a fórmula empírica corresponde a esta fórmula química. Se a fórmula química fosse a do etano, C_2H_6, não corresponderia porque a proporção de átomos de carbono seria 2/1 e a de hidrogénio de 6/1. Outras, como por exemplo a anilina, C_6H_7N, podia ser logo eliminada porque o azoto só está presente numa das fórmulas.

Encontrando o comando `LISTA`, o programa deve seleccionar todos os compostos na base de dados que contenham o atributo especificado e verificar, para cada um, se é compatível com a fórmula empírica. Deve escrever no ficheiro seleccionado no comando `REPORT` a linha do o comando `LISTA` que está a ser executado e, a seguir a esta e com uma indentação de um tabulador (`'\t'`), uma linha por cada composto que corresponda aos critérios pedidos com o nome do composto e a sua fórmula química. Por exemplo, para os comandos ilustrados à esquerda, o resultado no ficheiro de relatório deve ser este mostrado à direita. Notem a indentação das linhas referentes aos compostos encontrados.

```
LISTA CH;*
LISTA CH3;Alkane
LISTA CH2O;*
LISTA CH2O;Ester
```

```
LISTA CH;*
    Benzene;C6H6
LISTA CH3;Alkane
    Ethane;C2H6
LISTA CH2O;*
    Acetic Acid;C2H4O2
    Methyl Formate;C2H4O2
LISTA CH2O;Ester
    Methyl Formate;C2H4O2
```

Nota: para executar este comando, o vosso programa pode assumir que os compostos nesta base de dados têm elementos com apenas uma letra (e.g. C, N, O, H) e que as fórmulas químicas estão escritas de forma a que cada elemento só apareça uma vez na fórmula. Por exemplo, a alanina será escrita com a fórmula química $C_3H_7NO_2$ e não $C2H_4NH_2COO_H$. Assumir isto simplifica significativamente a implementação.

- `GRAFICO ficheiro;atributo`

  Criar um gráfico com a relação entre o logaritmo do número de átomos de carbono e a temperatura de ebulição de todos os compostos com o atributo especificado. O gráfico deve apresentar no eixo das abcissas (x) o logaritmo do número de átomos de carbono de cada composto e, no eixo das ordenadas (y), o ponto de ebulição de cada composto. Deve também mostrar a recta de regressão linear para os dados apresentados. Seja esta recta:

$$y = \alpha + \beta x$$

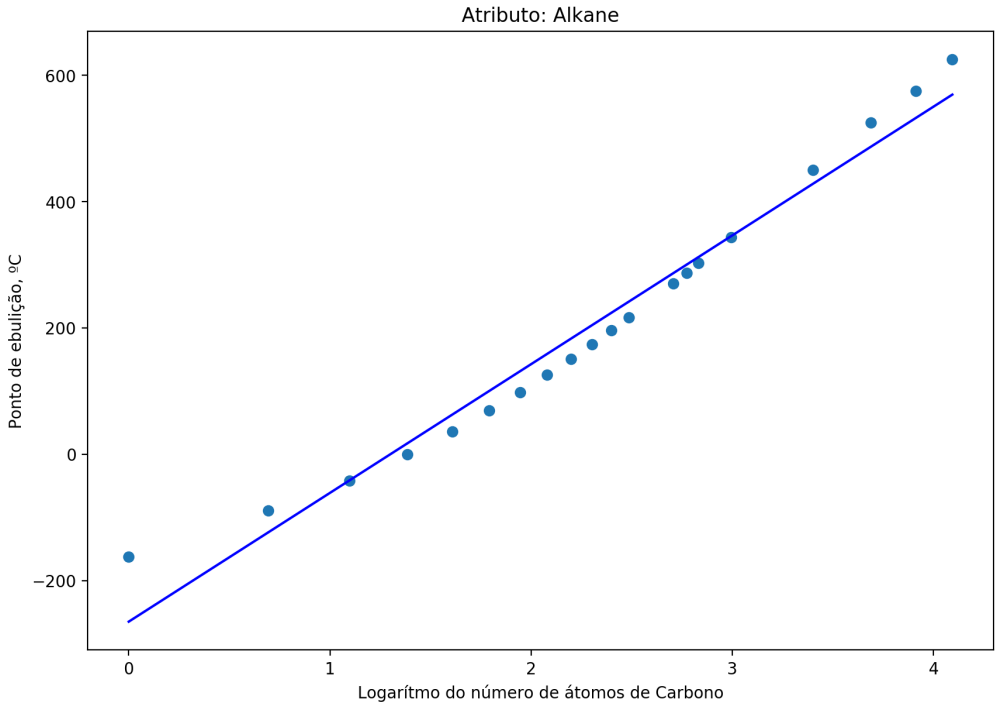
  os parâmetros α e β podem ser calculado pela seguintes expressões:

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \alpha = \frac{\sum y_i}{N} - \beta \frac{\sum x_i}{N}$$

Onde x_i e y_i são os valores do logaritmo do número de átomos de carbono e do ponto de ebulição dos N compostos considerados e $\bar{x}$ e $\bar{y}$ as respectivas médias. A linha da regressão linear deve ser traçada desde o menor valor do logaritmo do número de átomos de carbono e o maior valor do logaritmo do número de átomos de carbono nesse conjunto.

Este gráfico ilustra o resultado esperado para o atributo `Alkane`.

Átomos C	Ebulição
4	0.0
10	174.0
12	216.0
2	-89.0
17	303.0
7	98.0
60	625.0
16	287.0
6	69.0
20	343.0
1	-162.0
9	151.0
8	126.0
50	575.0
15	270.0
5	36.0
3	-42.0
40	525.0
30	450.0
11	196.0



3 Dados do Trabalho

O arquivo `trabalho2.zip` tem os ficheiros `compostos.txt` e `compostos2.txt`, bem como o ficheiro de comandos `orders.txt` 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 seu endereço oficial da FCT. Esse ficheiro .zip tem de conter pelo menos um ficheiro chamado `tp2.py` que implemente a função `processar` pedida no enunciado. Esta função deve estar implementada **exactamente como especificado, com esse nome e recebendo como argumento o nome do ficheiro de comandos**. O ficheiro zip pode conter outros módulos caso queiram organizar o código em vários módulos mas não é necessário incluir outros ficheiros que não os módulos que vocês criarem. Não incluam os ficheiros de dados, nem o sqlite3.exe, imagens ou qualquer outro material desnecessário. Ponham no zip apenas o código que implementaram.

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 colega.

O prazo para a entrega deste trabalho termina no final do dia **7 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.
- Reutilização de funções em diferentes contextos, sempre que adequado.
- 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 dos ficheiros de 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 `Compostos` e `Atributos`, 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 um atributo que existe na base de dados;
  - nenhum comando `CARREGAR` irá aparecer antes do comando `CRIAR_TABELAS`.
  - os comandos `REPORT` e `BASE_DADOS` aparecem pelo menos uma vez antes de ser necessário criar relatórios ou aceder à base de dados.
  - todas as fórmulas químicas contém elementos com apenas uma letra (e.g. C, N, O, H)
  - em todas as fórmulas químicas cada elemento aparece apenas uma única vez, seguido do número de átomos total desse elemento. Por exemplo, o ácido acético não será representado por CH_3COOH mas sim por $C_2H_4O_2$.

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

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.