Universidade do Minho
Licenciatura em Engenharia Informática

Grupo 11

Março 2022

Processamento de Linguagens
Trabalho Prático 1: CSV

Duarte Moreira (a93321) Lucas Carvalho (a93176) Ricardo Gama (a93237)

Índice

1 Introdução

Na realização do trabalho prático tínhamos como principais objetivos o reforço da nossa capacidade de escrever Expressões Regulares para descrição de padrões em streams de texto, bem como, desenvolver, a partir de expressões regulares, Processadores de Linguagens Regulares que encontrem e transformem textos.

Após a escolha do enunciado, **CSV**, o foco principal era fazer um conversor de um ficheiro CSV - *Comma Separated Values* - para o formato **JSON** - *JavaScript Object Notation*. Para isso, foi necessário especificar os padrões que queríamos encontrar no texto-fonte, através de expressões regulares e desenvolver um filtro de texto para fazer o reconhecimento dos padrões identificados e proceder à transformação pretendida.

2 Desenho e Implementação da Solução

Para a realização desta tarefa o grupo decidiu seguir uma estratégia de *divide and conquer*, ou seja, começamos por dividir o problema em dois. Primeiro, o tratamento do cabeçalho e posteriormente do corpo do ficheiro recebido como input.

O processamento da primeira linha do ficheiro, cabeçalho, consistiu no desenvolvimento de uma expressão regular que capturasse os campos pretendidos de acordo com os exemplos fornecidos no enunciado. O maior desafio nesta parte foi ter em atenção que as colunas que se tratassem de listas poderiam ter variantes tais como: ser uma lista com tamanho fixo ou variável, e para além disso poderiam ou não ter um função de agregação para aplicar à mesma (**avg**, **sum**, **min**, **max**).

A expressão regular utilizada nesta fase foi a segiunte:

- `([^,{]+)(?:\{(\d+),(\d+)\}|\{(\d+)\})?(?:::(\w+))?`

Ao usar esta expressão regular e a função *findall* somos capazes de obter um tuplo para cada coluna do ficheiro. Este irá conter informação do nome da coluna, se for uma lista, o número mínimo e máximo de elementos ou então o número exato e ainda se é seguida de uma função de agregação ou não - (nome,min,max,size,funcao).

Já o segundo problema, consistiu num trabalho mais simples, no entanto mais trabalhoso. Esta segunda parte resume-se a interpretar as linhas lidas do ficheiro de input e com a ajuda dos tuplos calculados anteriormente, ir criando passo a passo, ou neste caso *string* a *string* uma lista de *strings* de forma a que estas se encontrem no formato pretendido - **JSON**. Para isso, recorre-se a dois ciclos aninhados, um para percorrer as linhas e outro as colunas. No meio deste processo, para o tornar menos complexo, são usadas algumas funções auxiliares criadas por nós e mais uma vez é utilizada outra expressão regular, desta vez muito mais simples (`\d+`) para identificar os elementos de uma lista. Uma vez construída a lista de strings é realizado um *join* de forma a juntar todas as strings separando-as com um `\n` e por fim é criado um ficheiro onde é escrito o resultado obtido.

3 Exemplos de Utilização

Nesta secção do relatório, apresentamos os resultados da aplicação do nosso trabalho a diferentes datasets criados pelo grupo.

- **Listas com tamanho definido**

```
Numero,Nome,Curso,Notas{5},,,,,
3162,Candido Faisca,Teatro,12,13,14,15,16
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,18
9234,Ricardo Gama,Informatica,13,14,10,11,17
3493,Duarte Moreira,Informatica,17,16,15,12,11
239,Lucas Carvalho,Informatica,18,10,12,14,11
9235,Daniela Silva,Biologia,10,12,15,13,12
87344,Ana Filipa,Gestao,10,15,16,19,14
```

```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": [12, 13, 14, 15, 16]
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
    "Curso": "Desporto",
    "Notas": [17, 12, 20, 11, 12]
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
    "Curso": "Ciencia Politica",
    "Notas": [18, 19, 19, 20, 18]
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
    "Curso": "Informatica",
    "Notas": [13, 14, 10, 11, 17]
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
    "Curso": "Informatica",
    "Notas": [17, 16, 15, 12, 11]
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
    "Curso": "Informatica",
    "Notas": [18, 10, 12, 14, 11]
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
```

```
    "Curso": "Biologia",
    "Notas": [10, 12, 15, 13, 12]
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": [10, 15, 16, 19, 14]
  }
]
```

- **Listas com um intervalo de tamanhos**

```
Numero,Nome,Curso,Notas{3,5},,,,,
3162,Candido Faisca,Teatro,12,13,14,,
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,
9234,Ricardo Gama,Informatica,13,14,10,11,
3493,Duarte Moreira,Informatica,17,16,15,12,11
239,Lucas Carvalho,Informatica,18,10,12,,
9235,Daniela Silva,Biologia,10,12,15,12,
87344,Ana Filipa,Gestao,10,15,16,,
```

```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": [12, 13, 14]
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
    "Curso": "Desporto",
    "Notas": [17, 12, 20, 11, 12]
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
    "Curso": "Ciencia Politica",
    "Notas": [18, 19, 19, 20]
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
    "Curso": "Informatica",
    "Notas": [13, 14, 10, 11]
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
    "Curso": "Informatica",
    "Notas": [17, 16, 15, 12, 11]
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
```

```
    "Curso": "Informatica",
    "Notas": [18, 10, 12]
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
    "Curso": "Biologia",
    "Notas": [10, 12, 15, 12]
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": [10, 15, 16]
  }
]
```

- **Função de Agregação: AVG**

```
Numero,Nome,Curso,Notas{3,5}::avg,,,,,
3162,Candido Faisca,Teatro,12,13,14,,
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,
9234,Ricardo Gama,Informatica,13,14,10,11,
3493,Duarte Moreira,Informatica,17,16,15,12,11
239,Lucas Carvalho,Informatica,18,10,12,,
9235,Daniela Silva,Biologia,10,12,15,,
87344,Ana Filipa,Gestao,10,15,16,19,
```

```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": 13.0
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
    "Curso": "Desporto",
    "Notas": 14.4
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
    "Curso": "Ciencia Politica",
    "Notas": 19.0
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
    "Curso": "Informatica",
    "Notas": 12.0
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
```

```
    "Curso": "Informatica",
    "Notas": 14.2
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
    "Curso": "Informatica",
    "Notas": 13.33
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
    "Curso": "Biologia",
    "Notas": 12.33
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": 15.0
  }
]
```

- **Função de Agregação: SUM**

```
Numero,Nome,Curso,Notas{3,5}::sum,,,,,
3162,Candido Faisca,Teatro,12,13,14,,
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,
9234,Ricardo Gama,Informatica,13,14,10,11,
3493,Duarte Moreira,Informatica,17,16,15,12,11
239,Lucas Carvalho,Informatica,18,10,12,,
9235,Daniela Silva,Biologia,10,12,15,13,
87344,Ana Filipa,Gestao,10,15,16,19,14
```

```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": 39
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
    "Curso": "Desporto",
    "Notas": 72
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
    "Curso": "Ciencia Politica",
    "Notas": 76
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
```

```
    "Curso": "Informatica",
    "Notas": 48
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
    "Curso": "Informatica",
    "Notas": 71
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
    "Curso": "Informatica",
    "Notas": 40
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
    "Curso": "Biologia",
    "Notas": 50
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": 74
  }
]
```


- **Função de Agregação: MIN**

```
Numero,Nome,Curso,Notas{2,5}::min,,,,,
3162,Candido Faisca,Teatro,12,13,14,,
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,
9234,Ricardo Gama,Informatica,13,14,10,11,
3493,Duarte Moreira,Informatica,17,16,15,12,11
239,Lucas Carvalho,Informatica,18,10,,,
9235,Daniela Silva,Biologia,10,12,15,,
87344,Ana Filipa,Gestao,10,15,16,19,14
```


```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": 12
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
    "Curso": "Desporto",
    "Notas": 11
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
```

```
    "Curso": "Ciencia Politica",
    "Notas": 18
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
    "Curso": "Informatica",
    "Notas": 10
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
    "Curso": "Informatica",
    "Notas": 11
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
    "Curso": "Informatica",
    "Notas": 10
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
    "Curso": "Biologia",
    "Notas": 10
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": 10
  }
]
```


- **Função de Agregação: MAX**

```
Numero,Nome,Curso,Notas{3,6}::max,,,,,,
3162,Candido Faisca,Teatro,12,13,14,,,
7777,Cristiano Ronaldo,Desporto,17,12,20,11,12,13
264,Marcelo Sousa,Ciencia Politica,18,19,19,20,,
9234,Ricardo Gama,Informatica,13,14,10,11,,
3493,Duarte Moreira,Informatica,17,16,15,12,11,10
239,Lucas Carvalho,Informatica,18,10,12,,,
9235,Daniela Silva,Biologia,10,12,15,13,12
87344,Ana Filipa,Gestao,10,15,16,19,14,
```


```
[
  {
    "Numero": "3162",
    "Nome": "Candido Faisca",
    "Curso": "Teatro",
    "Notas": 14
  },
  {
    "Numero": "7777",
    "Nome": "Cristiano Ronaldo",
```

```
    "Curso": "Desporto",
    "Notas": 20
  },
  {
    "Numero": "264",
    "Nome": "Marcelo Sousa",
    "Curso": "Ciencia Politica",
    "Notas": 20
  },
  {
    "Numero": "9234",
    "Nome": "Ricardo Gama",
    "Curso": "Informatica",
    "Notas": 14
  },
  {
    "Numero": "3493",
    "Nome": "Duarte Moreira",
    "Curso": "Informatica",
    "Notas": 17
  },
  {
    "Numero": "239",
    "Nome": "Lucas Carvalho",
    "Curso": "Informatica",
    "Notas": 18
  },
  {
    "Numero": "9235",
    "Nome": "Daniela Silva",
    "Curso": "Biologia",
    "Notas": 15
  },
  {
    "Numero": "87344",
    "Nome": "Ana Filipa",
    "Curso": "Gestao",
    "Notas": 19
  }
]
```

4 Conclusões

Após a realização do trabalho prático número 1, consideramos que foi bastante benéfico para o aprofundamento da matéria lecionada na unidade curricular de **Processamento de Linguagens**, com destaque para as **Expressões Regulares** e do módulo **re** do **Python**, que foram essenciais para gerar os filtros de texto necessários para a realização do trabalho.

Acreditamos que conseguimos atingir os objetivos principais do trabalho. Conseguimos gerar um conversor de um ficheiro **CSV** para o formato **JSON**, tendo em conta todas as condicionantes adicionais que este trabalho nos propunha: listas com tamanho definido, com um intervalo de tamanhos e funções de agregação. Para isto, foi importante a captura inicial dos valores do cabeçalho num *array*, onde cada posição continha um tuplo com informações relativas a cada coluna, sendo assim possível verificar se esta continha uma função de agregação, se tinha um valor máximo e mínimo para o intervalo ou se se tratava apenas de um tamanho definido.