Processamento de Linguagens
Trabalho Prático 1
Universidade do Minho
Licenciatura em Engenharia Informática

Grupo 22

David Pereira Alves

Rui Miguel Borges Braga

Tiago Lucas Alves

27 de março de 2022

Conteúdo

1 Introdução

Neste relatório é apresentado o desenvolvimento do trabalho prático número 1 da Unidade Curricular de Processamento de Linguagens. O tema escolhido pelo grupo foi o processador de registos de exames médicos desportivos no qual se pretendia trabalhar com um dataset gerado no âmbito do registo de exames médicos desportivos, criar indicadores estatísticos através do dataset e representar esses indicadores através de código html.

Neste relatório está presente a estratégia utilizada e a forma como foram calculados os indicadores para cada funcionalidade pedida.

2 Estratégia Utilizada

De modo a processar o dataset "emd.csv", a estratégia utilizada foi processar cada linha do dataset com o seguinte filtro:

```
   (?P<Id>\d{7}[A-Za-z0-9]{17}),
(?P<Index>\d\d?),
(?P<Data>\d{4}-\d{2}-\d{2}),
(?P<Primeironome>[A-Z][a-z]+),
(?P<Ultimonome>[A-Z][a-z]+),
(?P<Idade>\d{1,2}),
(?P<Genero>(M|F)),
(?P<Morada>[A-Z][a-z]+),
(?P<Modalidade>[A-Z][A-Za-zçãé]+),
(?P<Clube>[A-Z][A-Za-zã]+),
(?P<Email>[A-Za-z.]+[A-Za-z]+@[A-Za-z.ã]+[A-Za-z]),
(?P<Federado>true|false),
(?P<Resultado>true|false)
```

Desta forma, através dos identificadores do Python(?P), é possível criar para cada paramêtro(id, index, data,etc) uma entrada para um dicionário, sendo, desta forma, feito o parsing do ficheiro para uma lista de dicionários em que cada posição da lista corresponde aos dados de um exame médico desportivo. De seguida, sobre esse dicionário são calculados os indicadores estatísticos e, posteriormente, a escrita do código html em ficheiros para apresentação dos dados. Relativamente à apresentação dos indicadores, estes estão presentes no ficheiro index.html e, em cada um deles, existe a opção "mais infomação aqui"que permite vizualizar toda a informação que foi utilizada para calcular os dados estatísticos.

3 Indicadores Estatísticos

3.1 Datas extermas dos registos no dataset

No cálculo da maior e menor data de registo do exame médico reuniu-se o nome e a data de registo de todos os atletas do dataset numa lista e ordenou-se essa informação pela data de realização do exame médico desportivo, sendo a primeira posição da lista o exame médico desportivo mais antigo e a ultima posição o exame médico mais recente. No campo "mais informação aqui"encontra-se esta lista ordenada por data.

3.2 Distribuição por género em cada ano e no total

No cálculo da distribuição por género em cada ano e no total é criado um dicionário que tem como chave os anos em que se realizaram os exames médicos e como valor possui outro dicionário com chaves "Masculino"e "Feminino"que tem listas com os nomes dos atletas. Seguidamente, para calcular quantos atletas efetuaram registos num determinado ano basta apenas calcular o tamanho das listas dos atletas ao ano correspondente. No campo "mais informação aqui"encontra-se estas listas ordenada por ano e género.

3.3 Distribuição por modalidade em cada ano e no total

No cálculo da distribuição por modalidade em cada ano e no total é criado um dicionário em que cada chave é associada a um ano de realização de um exame médico desportivo e, para cada ano, é criado um dicionário no qual as chaves são as modalidades encontradas, sendo associada cada modalidade uma lista com os nomes dos atletas e os respetivos clubes. De seguida é calculado, para cada ano e modalidade, o número de atletas correspondentes no dicionário de cada modalidade através do comprimento da lista associada a cada modalidade. No campo "mais informação aqui"encontra-se para cada ano e modalidade os atletas e o respetivo clube ao qual pertencem.

3.4 Distribuição por idade e género

No cálculo da distribuição por idade e género, inicializa-se uma lista com 4 listas vazias, sendo que estas irão corresponder às combinações possíveis de género e das duas faixas etárias(`< 35 anos e >= 35`) e faz-se uma travessia da lista de exames médicos desportivos, retirando o nome do atleta, idade e género, de forma a preencher estas listas com os seus respetivos atletas e a sua informação. Nas listas, a informação dos atletas é adicionada sobre a forma de tuplos (nome, idade), sendo que, depois, as listas são ordenadas por ordem crescente de idade como critério principal e por ordem alfabética de nome como segundo critério. No fim, percorremos as listas já ordenadas, de forma a preencher as tabelas presentes no .html.

3.5 Distribuição por morada

No cálculo da distribuição por morada, inicializa-se um dicionário em que as chaves vão ser as moradas e os valores serão listas de tuplos com informação dos atletas. Assim sendo, fez-se uma travessia da lista de atletas, retirando a cada um deles o seu nome, morada e modalidade respetivas, de modo a preencher este dicionário. De seguida, ordenámos cada uma das listas do dicionário por ordem alfabética dos nomes dos atletas e transformámos o dicionário numa lista que está ordenada por ordem alfabética das moradas. Seguidamente, criámos outro dicionário com o número de atletas por cada morada. No final, apresentámos no html este último dicionário com a lista de moradas e o número de atletas respetivo, em que cada morada tem uma referência para lista de atletas presentes na mesma.

3.6 Distribuição por estatuto de federado em cada ano

No cálculo da distribuição por estatuto de federado em cada ano, utilizámos dois dicionários para além da lista com os atletas. Um dos dicionários diria respeito aos atletas federados por ano enquanto o outro continha os não federados em cada ano. Cada um destes dicionários teria elementos chave-valor onde a chave é o ano em que os exames foram realizados e o valor é uma lista com a informação dos atletas que realizaram exames nesse ano. De modo a obter a informação sobre o número de atletas federados ou não federados, basta saber o tamanho da lista para cada ano. Assim podemos escrever no ficheiro .html os dados da distribuição dos atletas federados por ano. No fim percorremos todos os anos de cada um dos dicionários para preencher o .html correspondente a esta query, onde guardamos em tabelas os dados dos atletas que nos permitiram obter a informação para este indicador.

3.7 Percentagem de aptos e não aptos por ano

No cálculo da percentagem de aptos e não aptos por ano, recorreu-se a dois dicionários para além da lista com os atletas. Um dos dicionários diria respeito aos atletas aptos por ano enquanto o outro continha os não aptos em cada ano. Cada um destes dicionários teria elementos chave-valor onde a chave é o ano em que os exames foram realizados e o valor é uma lista com a informação dos atletas que realizaram exames nesse ano. De modo a obter a informação sobre o número de atletas aptos ou não aptos, basta calcular o tamanho da lista para cada ano. Depois podemos escrever no ficheiro .html os dados da distribuição dos atletas aptos por ano. No fim percorremos todos os anos de cada um dos dicionários para preencher o .html correspondente a esta query, onde guardamos em tabelas os dados dos atletas que nos permitiram obter a informação para este indicador.

4 Conclusão e Trabalho Futuro

Neste primeiro trabalho prático da UC de Processamento de Linguagens, o objetivo era criar um processador de registos de exames médicos que , utilizando um dataset com este tipo de informação, fosse capaz de criar um ficheiro .html que contivesse indicadores estatisticos sobre os dados. Este projeto incidiu principalmente no processamento dos dados, tendo o grupo só utilizado uma expressão regular para captar toda a informação de um exame médico desportivo.

No fim da realização do trabalho, o grupo sente que realizou um trabalho positivo onde conseguiu desenvolver mais competências como desenvolvimento em Python, modulo re, processamento de texto, html e no processamento de dados.

No futuro, podem ainda ser implementados mais indicadores e melhorar os ficheiros .html gerados para conterem a informação de forma mais organizada e com gráficos ajudem à leitura dos indicadores.