**Projeto Integrado – 09/02/2024**

Na passada aula de Projeto Integrado, colocamos em prática a planificação para este semestre. Conjuntamente decidimos extrair informações do site do Jornal Online Nós Uminho ([www.nos.uminho.pt](http://www.nos.uminho.pt)).



Figura 1 - Jornal online da UMinho

Primeiro, verificamos os números do jornal disponíveis online. Há todos os meses uma nova edição e a atual é a nº 128, de dezembro de 2023. No jornal inserem-se diversas rubricas de artigos. O URL do nº 128 é [www.nos.uminho.pt/History.aspx?id=1149](http://www.nos.uminho.pt/History.aspx?id=1149) e, à medida que o número presente no final do URL decresce, verificamos que temos acesso a edições anteriores. Após vários testes, verificamos também que existem algumas edições em falta e chegamos à conclusão de que a edição mais antiga que pode ser acedida é datada de junho de 2015, correspondendo ao nº 53 (URL: [www.nos.uminho.pt/History.aspx?id=1067](http://www.nos.uminho.pt/History.aspx?id=1067)). Decidimos então extrair as informações desta edição nº 53.

No *Windows Powershell*, demos início ao processo de extração instalando a ferramenta *wget* que, muito resumidamente, é responsável por fazer o download de ficheiros da internet. De seguida, já na linha de comandos, criamos uma pasta com o nome *NosUm* (que seria o destino da informação extraída) e abrimo-la. A partir daqui, utilizando o *wget* e o URL da página da qual pretendemos, extraímos as informações utilizando o seguinte comando:



Figura 2 - Extração com wget

Desta forma, todas as informações extraídas foram guardadas na pasta previamente criada.

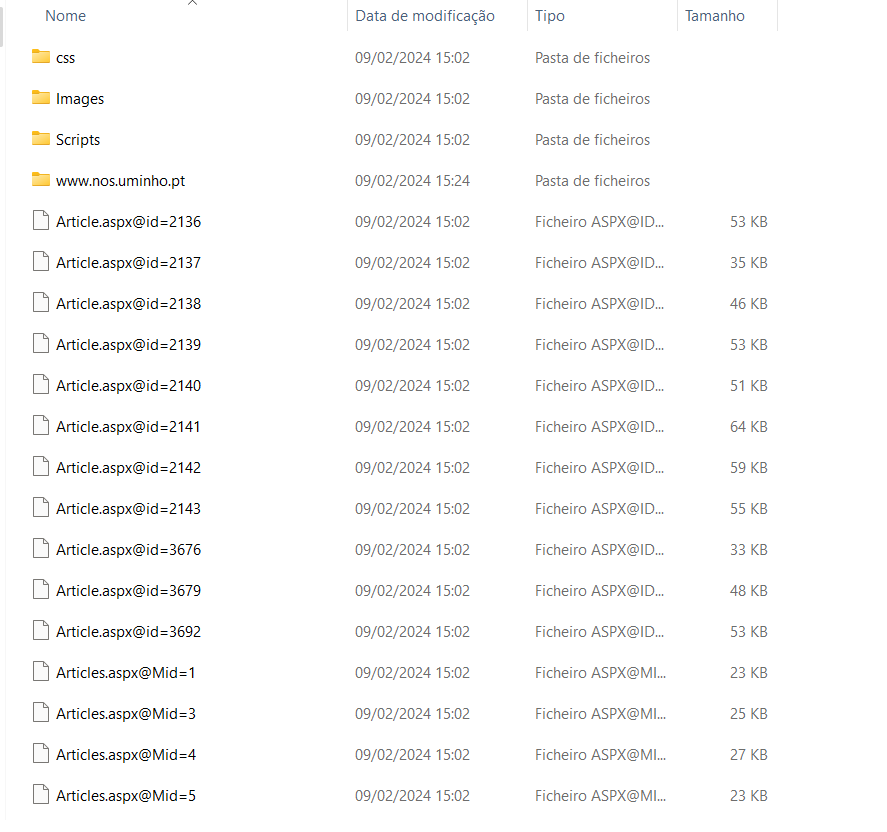


Figura 3 - Conteúdo da pasta NosUm após a extração

Depois desta experiência, testamos fazer a mesma coisa, mas utilizando *Python*. Apenas para experimentação, fizemos a extração desde o artigo cuja terminação é 1148 até 1150.



Figura 4 - Extração com Python

Neste ponto, já com as extrações feitas, era necessário fazer a “limpeza” dos dados e retirar apenas aquilo que nos interessa. Desta forma, novamente no *Windows Powershell*, instalamos a ferramenta *ripgrep*, que serve para procurar padrões nos nossos artigos. Como o nosso interesse principal são as entrevistas, procuramos pelo padrão “UM de nós”, visto que este é o nome da rubrica em que se inserem as entrevistas a pessoas que fazem parte da comunidade da Universidade do Minho.



Figura 5 - Teste ripgrep

Os resultados não corresponderam às nossas expectativas pois, com este padrão, não foi possível extrair apenas as entrevistas. O motivo que descartou este padrão é muito simples: em todas as abas do jornal, aparecem, na barra lateral e na superior, referências a esta rubrica, ou seja, retira-nos mais informação do que aquilo que queremos. Deste modo, deixamos para a aula seguinte a busca por um padrão que nos retire apenas as entrevistas.

Para finalizar a aula, aprendemos a converter os ficheiros html que extraímos, em outros formatos, por exemplo txt ou markdown. Esta conversão é feita de forma muito simples. Testamos com o URL 3677:



Figura 6 - Conversão para Markdown

Desta forma, na pasta *NosUM*, um ficheiro markdown com o nome 3677 a partir do ficheiro html que extraímos da web, foi criado.