Mestrado em Engenharia Informática
Ciência de Dados
Ano Letivo 2023/2024

Mini Projeto I

Trabalho Individual
Data Acquisition (Files, Packages, APIs)

© Ricardo Campos
ricardo.campos@ubi.pt

O trabalho prático é obrigatório para a obtenção de aprovação na unidade curricular. A não entrega durante o prazo previsto implica a automática reprovação dos alunos.

Objetivo: Familiarização com o processo de aquisição de dados através de ficheiros, packages e APIs.

Entrega: Os trabalhos (em formato notebook – devidamente documentados) devem ser inseridos na plataforma de e-learning (moodle) até 18/03/2024, 23h59.

Realização do trabalho: Os trabalhos devem ser realizados individualmente.

Tarefa 1: Familiarização com a obtenção de dados a partir de ficheiros *pdf*.

1. Reúna um conjunto aproximado de 100 ficheiros em formato *pdf* relacionados com uma temática à sua escolha (e.g., artigos científicos; documentos do parlamento europeu; patentes; programas eleitorais; etc). Proceda à extração do texto de cada ficheiro com recurso a bibliotecas Python.
2. Guarde os conteúdos num ficheiro JSON adotando uma estrutura de dados apropriada com vista a guardar todos os dados relevantes obtidos. Por exemplo, no caso de um programa eleitoral, seria adequado guardar o nome do partido político, o líder do partido à data da eleição, a designação da eleição, a data da eleição, o texto, assim como outros elementos relevantes extraídos a partir da aplicação de ferramentas de NLP ao texto, nomeadamente, palavras-relevantes, entidades (NER – *named entity recognition*), datas e outras que achar adequadas. Seja criativo.
3. Carregue o ficheiro JSON (anteriormente criado) para o seu ambiente de programação.

4. Imprima o conteúdo do ficheiro JSON, restrito aos 5 primeiros registos.
5. Crie uma nuvem de palavras a partir dos textos coletados. Seja criativo. Por exemplo, crie diferentes *word clouds* se tiver mais do que um período de tempo. Para ver alguns exemplos de como criar uma *wordcloud* clique no seguinte link: https://github.com/amueller/word_cloud/blob/master/examples/simple.py

Tarefa 2: Familiarização com a obtenção de dados a partir de packages Python

1. Recorra ao package do *wikipedia* [https://pypi.org/project/wikipedia/] para criar um *dataset* de 2000 imagens relacionadas com duas temáticas distintas à sua escolha (e.g., covid e desporto).

Tarefa 3: Familiarização com a obtenção de dados a partir de APIs

1. Recorra à API "Text Search" ou à API "Image Search" do Arquivo.pt para reunir um conjunto elevado de textos ou imagens que deverá guardar no seu computador.
2. Guarde os textos ou imagens na pasta */data/aaaa.mm.dd*, onde *aaaa.mm.dd* é um valor que deverá ser dinamicamente obtido a partir da data de execução do código (e.g., 2020.05.26, no caso de o código ser executado no dia 26/05/2020). Proceda também à gravação das informações correspondentes (dos textos ou das imagens) num ficheiro JSON dentro da mesma pasta.
3. Carregue o ficheiro JSON em memória e percorra os conteúdos de 5 dos registos.
4. Reúna dois colegas e elabore uma proposta de candidatura ao Prémio Arquivo.pt 2024. Detalhe e explore, junto com os seus colegas, uma descrição sumária da ideia tendo em conta o seu impacto social e científico, a relevância da utilização do Arquivo.pt, originalidade e a exequibilidade da concretização do projeto. Para consultar os premiados das edições anteriores clique no seguinte link: https://sobre.arquivo.pt/pt/colabore/premios-arquivo-pt/