Instituto Politécnico de Tomar

UNIDADE DEPARTAMENTAL DE TECNOLOGIAS DE INFORMAÇÃO E COMUNICAÇÃO

Mestrado em Engenharia Informática – Internet das Coisas

Análise de Grande Volume de Dados

Ano Letivo 2023/2024

Projeto I
Trabalho de Grupo
Aquisição, Armazenamento e Recuperação de Dados em Larga Escala

© Ricardo Campos
ricardo.campos@ipt.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 em larga escala (APIs e webscraping), armazenamento e recuperação (redis ou elasticsearch)

Entrega: Os trabalhos (em formato notebook – devidamente documentados) devem ser inseridos na plataforma de e-learning (moodle) até 22/04/2024, 18h00.

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 Web Scraping

Extraia informação de uma página web estática à sua escolha (e.g., extrair informações dos destaques listados na página web do https://ticketline.sapo.pt/

Tarefa 4: Familiarização com a obtenção de dados a partir de APIs

1. Obtenha um conjunto elevado de textos (a guardar no seu computador em formato json) com recurso a uma API. Deverá proceder à recolha dos dados (e.g., de hora a hora) durante o espaço de 5 dias consecutivos (pode usar uma máquina virtual gratuita - https://www.pythonanywhere.com/ - em alternativa ao seu computador pessoal). Registe o desenrolar do processo de obtenção de dados num ficheiro de *logs* (e.g., via biblioteca *logging*).
2. Carregue o ficheiro JSON em memória e percorra os conteúdos de 5 dos registos.

Tarefa 5: Familiarização com o armazenamento e recuperação de dados em larga escala (redis ou elasticsearch)

1. Proceda à indexação dos textos (coletados na tarefa 4) no redis ou elasticsearch.
2. Exemplifique o processo de recuperação de informação com recurso a um conjunto de queries.