UNIVERSIDADE BEIRA INTERIOR

FACULDADE ENGENHARIA
Departamento de Informática

Notícias da Covilhã @ Arquivo Web

Proposta de Projeto. Lic. Eng. Informática

Orientador: Ricardo Campos (ricardo.campos@ubi.pt)

Objetivos

O crescente aumento de informações publicadas na web, sob a forma de textos, imagens, vídeos e áudios, tem sido uma característica marcante da era digital. Curiosamente, nunca como antes, se perderam tantos conteúdos impedindo que as gerações atuais e futuras tenham acesso a um registo histórico da web, tal como hoje a conhecemos. Diversos estudos referem que 80% das páginas da web desaparecem ou mudam passado apenas 1 ano [1]. Nesse contexto, a importância da preservação de conteúdos web torna-se fundamental. Em Portugal, a preservação de conteúdos da web fica a cargo do Arquivo.pt, uma importante iniciativa dedicada a obter, arquivar e disponibilizar o conteúdo da web portuguesa.

O objetivo deste projeto passa por utilizar esses recursos para preservar a história e o legado do jornal "Notícias da Covilhã" ao tornar o seu conteúdo histórico (mais de 3,000 versões preservadas) facilmente acessível ao público e investigadores. A disponibilização desses conteúdos através de um website responsivo e dedicado ao arquivo web do jornal, visa contribuir para a preservação do património local e complementar a informação disponibilizada na atual versão do website.

Além da disponibilização dos conteúdos, o projeto tem como objetivo a criação de um sistema de pesquisa que permita aos utilizadores do website recuperar informações a partir dos dados coletados do Arquivo.pt.

Plano de Trabalho

T1: Definição e planeamento (2 semanas)

T2: Recolha e armazenamento dos dados do Arquivo.pt com recurso ao desenvolvimento de scripts em Python (2 semanas)

T3: Criação de uma imagem Docker com vista à integração dos vários componentes (1 semana)

T4: Desenvolvimento do website com base no layout anterior do jornal (6 semanas)

T5: Implementação do sistema de pesquisa com recurso ao Elastic Search (base de dados noslq) e ao algoritmo de recuperação de informação BM25 (3 semanas)

T6: Testes Finais (2 semanas)

T7: Documentação e Apresentação (2 semanas)

A Tabela 1 apresenta a distribuição das tarefas por cada uma das 15 semanas.

Tabela 1: Cronologia das tarefas (T) por semana (S).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
T1															
T2															
T3															
T4															
T5															
T6															
T7															

Requisitos Técnicos / Académicos

- conhecimentos sólidos em Python (para a recolha de dados) ou disposição para aprender.
- conhecimento em linguagens de programação web (HTML, CSS, JavaScript) ou Flask para o desenvolvimento do website.
- experiência com Elasticsearch (base de dados nosql) ou disposição para aprender.
- familiaridade com a criação de imagens Docker.
- bons conhecimentos de programação, engenharia de software e composição web

Resultados Esperados

- script python de obtenção dos dados e respetivo dataset coletado
- website responsivo (código a disponibilizar no github do aluno) e disponível online
- relatório
- candidatura ao Prémio Arquivo.pt 2024 (lista de vencedores dos prémios anteriores: https://sobre.arquivo.pt/pt/colabore/premios-arquivo-pt/)

Bibliografia

[1] Gomes, D., Demidova, E., Winters, J., and Risse, T. (2021). The Past Web: Exploring Web Archives. Springer.