Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- Título: Imaginarium
- Área temática: Informática, processamento e reconhecimento de imagem e *machine learning*
- Candidato: Diogo António Montez de Sousa
- Email: montezdesousa@gmail.com

Descrição do Trabalho

O *Imaginarium* (https://imaginarium.pages.dev) é um *website* que permite a pesquisa de imagens idênticas na base de dados do *Arquivo.pt* partindo de uma imagem, em vez de palavras. O design do site, vídeo e o próprio nome são inspirados num artefacto mágico do universo *Harry Potter* chamado pensieve, que permite as personagens guardar e rever memórias. Este item estabelece uma ponte para a premissa de viajar no tempo subjacente ao concurso e ao *Arquivo.pt* em geral.

Figura 0. - Pensieve

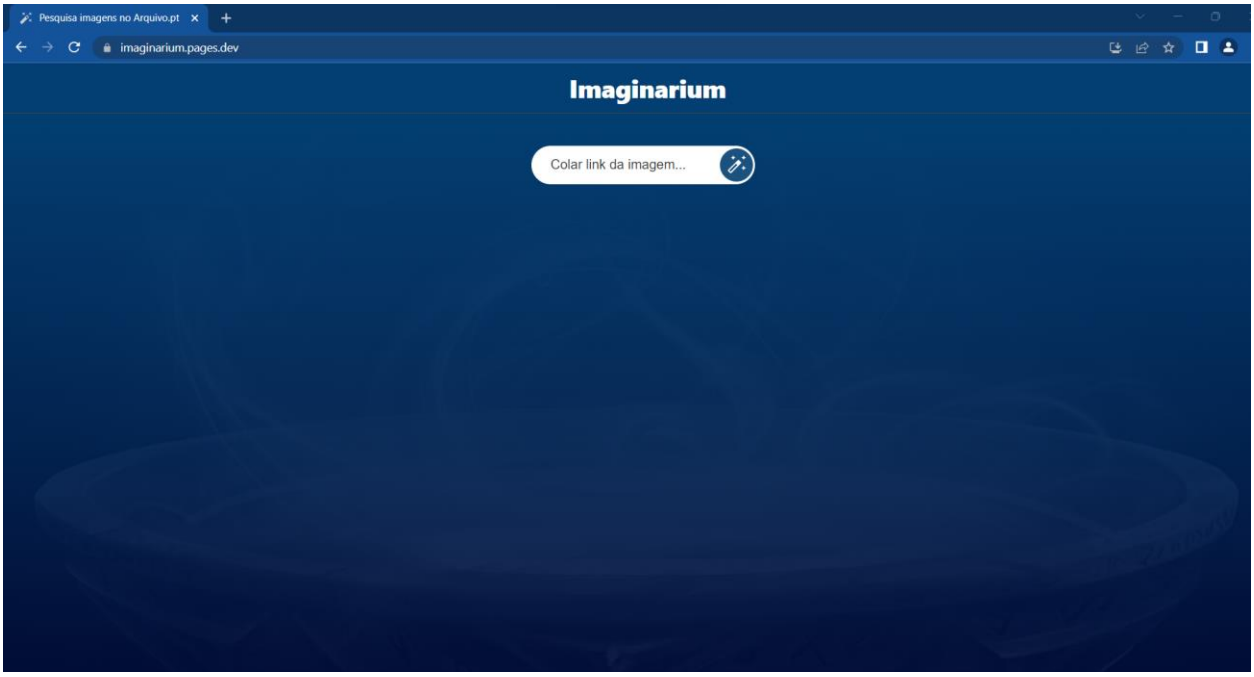

Figura 1. - Primeira página *Imaginarium*

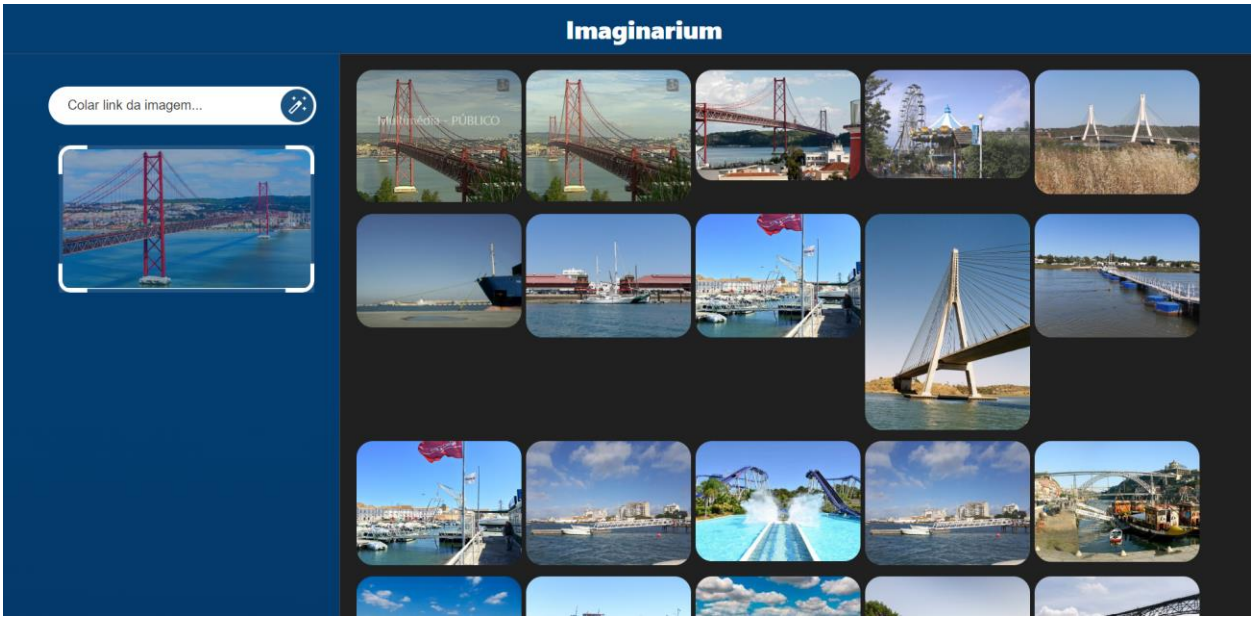

Figura 2. - Pesquisa *Imaginarium* - Ponte 25 de Abril

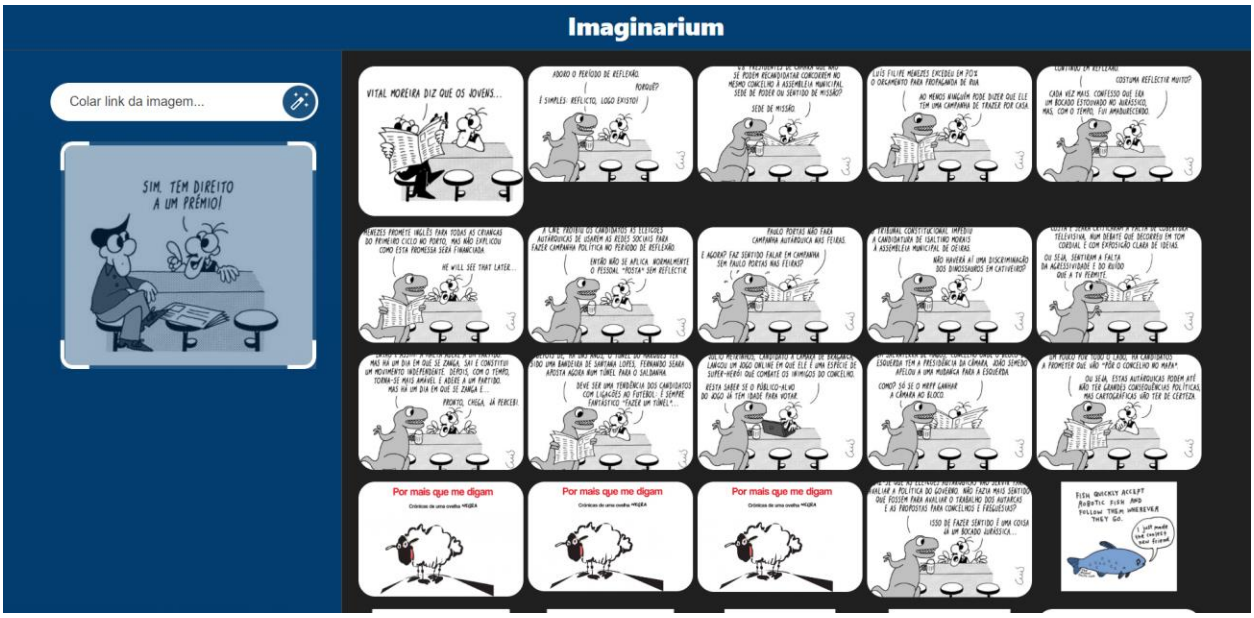

Figura 3. - Pesquisa *Imaginarium* - Bartoon Público

Esta funcionalidade foi desenvolvida com o objetivo de ser integrada no próprio motor de busca do *Arquivo.pt*. A referência mais popular deste tipo de pesquisa (*reverse image search*) é o *Google Lens*, que nos permite encontrar imagens idênticas à que fornecemos. Na Figura 4. podemos ver um exemplo pesquisando o logotipo do Arquivo.pt no *Google Lens*.

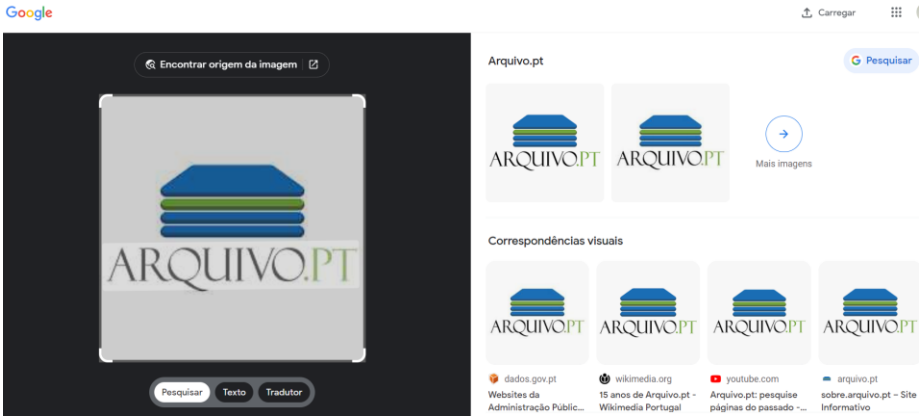

Figura 4. - Pesquisa *Google Lens* - logotipo *Arquivo.pt*

A concretização do projeto passou por 3 fases:

1) Pré-processamento

Recolha de dados de imagens na base de dados do *Arquivo.pt* através da API oficial ImageSearch-API https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1.1-(beta).

Os dados recolhidos de cada imagem foram os seguintes:

1. URL de acesso à imagem
2. Título da página do arquivo que contém a imagem
3. URL para a página do arquivo que contém a imagem

Por simplicidade, a recolha de dados focou-se apenas nalguns domínios em específico tais como o *publico.pt* e jornais centenários como *jn.pt*, *dnoticias.pt*, entre outros. Na sua totalidade foram recolhidas 43,250 referências para imagens.

É importante referir que se assim se entender a base de dados pode ser carregada com mais imagens, sendo potencialmente extensível a toda a base do arquivo.

2) Implementação

1) Extração das características de cada imagem

O primeiro passo foi redimensionar as imagens para 299 x 299 pixels, de modo que todas sejam comparáveis. De seguida criou-se uma matriz de convulsão para cada imagem através de uma 50-layer *convolutional neural network* (ResNet50) pré-treinada com mais de um milhão de imagens (*ImageNet*). A convulsão pode ser equiparada à impressão digital de cada imagem na base de dados e no nosso caso significou a compressão de 299x299x3=268,203 elementos (# de pixels vezes 3 canais de cor - RGB) num vetor de 2,048 elementos. Ver Figura 5.

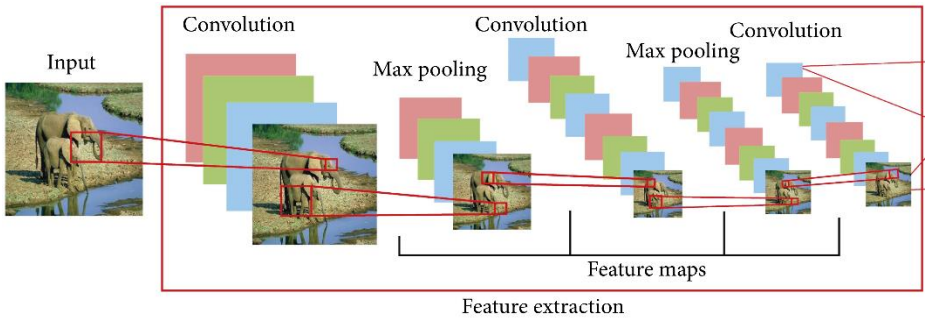

Figura 5. Representação CNN

2) Criação do modelo que aplica o algoritmo KNN (K-Nearest Neighbours)

Nesta fase, criamos um modelo que aplica o KNN para as convoluções geradas no ponto anterior. Isto leva alguns minutos a treinar dependendo da quantidade de imagens, mas só é necessário fazer uma vez e de seguida estaremos sempre a reutilizar o modelo. O KNN é um algoritmo de aprendizagem supervisionada que é usado para classificar os *k* vizinhos mais próximos de determinado item (por simplicidade no projeto definiu-se k=200), como ilustrado na Figura 6. Imagine-se que o círculo vermelho representa uma imagem pesquisada por um utilizador. Os círculos mais próximos do vermelho correspondem às imagens mais semelhantes à que

pesquisamos. Logicamente quanto mais nos afastamos do círculo vermelho, menor será a semelhança das imagens que se vão encontrando.

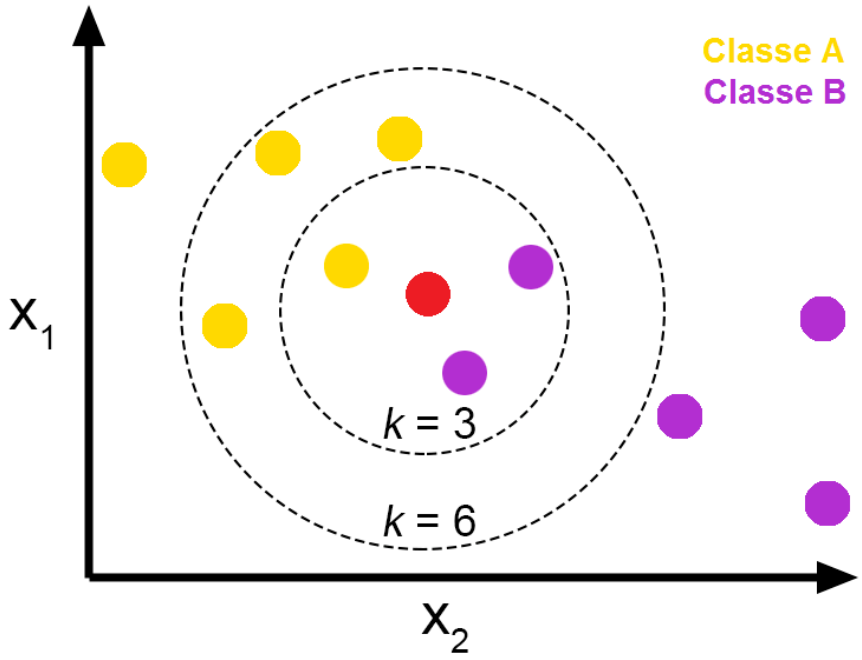

Figura 6. Representação algoritmo KNN

3) Execução

Ao fornecer o URL de qualquer imagem ao *Imaginarium*, este irá transformar a imagem referenciada pelo URL numa convolução, e enviá-la para o modelo KNN (a imagem corresponde ao círculo vermelho na Figura 1). O modelo KNN e alguns processos auxiliares retornam os URLs das 200 imagens com as convoluções mais semelhantes à da pesquisa, ordenadas por uma pontuação de similaridade.

Objetivos

O objetivo deste trabalho é o desenvolvimento de uma ferramenta que facilite e fomente o acesso, pesquisa e consequente utilização do *Arquivo.pt*, em particular recorrendo ao reconhecimento de imagens. Esta utilização poderá ser tanto de caráter científico como lúdico.

Resultados Atingidos

O principal resultado foi a criação do website *Imaginarium* (https://imaginarium.pages.dev) totalmente funcional, no qual é possível colar ou arrastar o endereço de qualquer imagem disponível online para a barra de pesquisa e obter uma lista de duzentas imagens ordenadas por

grau de similaridade. Estas imagens integram notícias publicadas num subconjunto dos *websites* gravados no arquivo, no entanto existe a possibilidade de estender o motor de busca a outras bases de dados. Ao passar o cursor por cada uma das imagens devolvidas pelo *Imaginarium* é possível ler o título da notícia à qual a imagem corresponde e, ao clicar na mesma, aceder à própria notícia no *Arquivo.pt*.

Originalidade e caráter inovador

Um dos fatores que demonstra a originalidade do trabalho é o facto da funcionalidade desenvolvida não existir no Wayback Machine, o motor de busca do Internet Archive que é um dos mais populares arquivos da internet mundial. Por outro lado, a relevância da pesquisa por imagens é evidenciada pelo valor que o próprio *Google* lhe atribui fazendo parte do motor de busca principal, bem como uma aplicação web independente – *Google Lens*.

O projeto *Imaginarium* também se destaca pela utilização de métodos de processamento e reconhecimento de imagem avançados e *machine learning* para facilitar a descoberta e consequente utilização da informação disponibilizada pelo Arquivo.pt.

Uma das particularidades deste projeto é que ao mesmo tempo que proporciona uma inovação metodológica de interesse e utilidade científica para investigadores de diversas áreas, acreditamos que também atrairá a atenção de qualquer cidadão apresentando assim uma utilidade social relevante. Adicionalmente, proporciona uma funcionalidade que pode ser estendida e adaptada ao motor de busca do *Arquivo.pt*, o que torna o projeto potencialmente duradouro no tempo e que a sua utilização não se extingue no âmbito deste concurso.

Impacto social (aplicação e utilidade social)

Para além da utilização de caráter científico, arquivos digitais como o Arquivo.pt deverão ser acessíveis ao público na era digital em que vivemos (2). O *Imaginarium* disponibiliza uma forma simples de pesquisa de artigos de jornais sobre temas de interesse social, sem implicar que se recorra à escrita de palavras-chave, algo que por vezes dificulta o processo de pesquisa. Ou seja, será possível fazer uma pesquisa no Arquivo.pt mesmo desconhecendo os nomes e/ou os temas específicos a escrever na barra de pesquisa. Por exemplo, qualquer cidadão poderá pesquisar uma imagem de uma figura pública de relevância política do passado e da qual desconheça o nome e obterá um conjunto de notícias relacionadas com a mesma.

Impacto científico (aplicação e utilidade científica)

Este recurso irá criar variadas possibilidades para a investigação histórica e temporal de diversos tópicos como heranças culturais (por exemplo pinturas de artistas históricos portugueses), acontecimentos de caráter internacional (por exemplo o desastre das Torres Gémeas), ou acontecimentos de interesse nacional (conquista do Euro 2016 pela seleção portuguesa de futebol) entre outros (1). Será ainda possível a contextualização de imagens de interesse histórico no panorama social e a investigação da sua evolução ao longo dos anos através da análise das notícias a que as imagens correspondem (1).

Este projeto proporciona uma ferramenta inovadora que aprimora a forma como a valiosa base de dados disponibilizada pelo Arquivo.pt pode ser acedida, explorada, analisada e mapeada.

O valor dos arquivos digitais e do desenvolvimento de ferramentas que facilitem a sua utilização, particularmente arquivos digitais de edições de jornais, têm vindo a ser reconhecidos pela comunidade científica (3). Outros projetos a nível Europeu como o *NewsEye* (projeto fundado pela *European Union's Horizon 2020 research and innovation programme* que terminou em Dezembro de 2022) já se dedicam a investigar possibilidades de utilização de arquivos digitais, o que demonstra a relevância científica, o potencial e caráter atual do projeto *Imaginarium* (2).

Relevância da utilização do Arquivo.pt

Os jornais consistem numa coletânea única de eventos culturais, políticos e sociais. Particularmente jornais mais antigos e abrangentes apresentam uma riqueza única de informação relevante para a sociedade. Apenas no Arquivo.pt seria possível aceder a uma extensa coletânea digital desses mesmos jornais (2).

Para além disso, o *Arquivo.pt* disponibiliza ferramentas indispensáveis que possibilitam a utilização das imagens na base de dados tais como a *ImageSearch-API* do *Arquivo.pt* (https://github.com/arquivo/pwa-technologies/wiki/APIs,https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1.1-(beta)) e o tutorial *Como explorar o Arquivo.pt usando Python* (3)

Comentários adicionais

Não se aplica.

Recursos complementares

Referências

(1) Viola, L. (2022). Review: Newspaper Navigator. Reviews in Digital Humanities, III(6). https://doi.org/10.21428/3e88f64f.e86411ad, Revisão do projeto Newspaper Navigator que permite pesquisa por imagens em jornais históricos

(2) Gomes, D. Web archives as research infrastructure for digital societies: the case study of Arquivo. pt. Archeion, 2022(123)., https://sobre.arquivo.pt/wp-content/uploads/GomesArquivoPTCaseStudy2022.pdf, Case-study Arquivo.pt

(3) NewsEye - A Digital Investigator for Historical Newspapers, https://www.newseye.eu/, Website do projeto NewsEye

(4) Tutorial: como explorar o Arquivo.pt usando Python, https://sobre.arquivo.pt/pt/tutorial-para-humanidades/, Tutorial: como explorar o Arquivo.pt usando Python

(5) Case studies - NewsEye, https://www.newseye.eu/case-studies/, Repositório de case-studies projeto NewsEye

(6) Leon Yin - Reverse image search engines using out of the box machine learning libraries, https://www.youtube.com/watch?v=-5BAepEE9I8&ab_channel=PyData, Apresentação sobre motores de busca através de imagens

(7) The Anatomy of a Web Archive Image Search Engine - Technical Report, https://sobre.arquivo.pt/wp-content/uploads/The_Anatomy_of_a_Web_Archive_Image_Search_Engine_tech_report.pdf, Artigo técnico sobre a arquitetura do motor de busca de imagens do Arquivo.pt