

Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- Título: Arquivo 25 de Abril
- Área temática: História de Portugal e Imprensa nacional online
- Candidato: Miguel Garcia Tavares Henriques
- Email: miguel.garcia.th@gmail.com

Descrição do Trabalho

Este projecto disponibiliza um arquivo online de artigos jornalísticos sobre várias personalidades, eventos e movimentos que tiveram relevância no antes e durante o 25 de Abril de 1974. Para cada um destes elementos foram recolhidos artigos de vários órgãos de comunicação social de referência (com publicação online) tendo em conta este contexto histórico. Este arquivo está disponível num web site (sugestão: ver num computador): <https://arquivo25abril.com/>

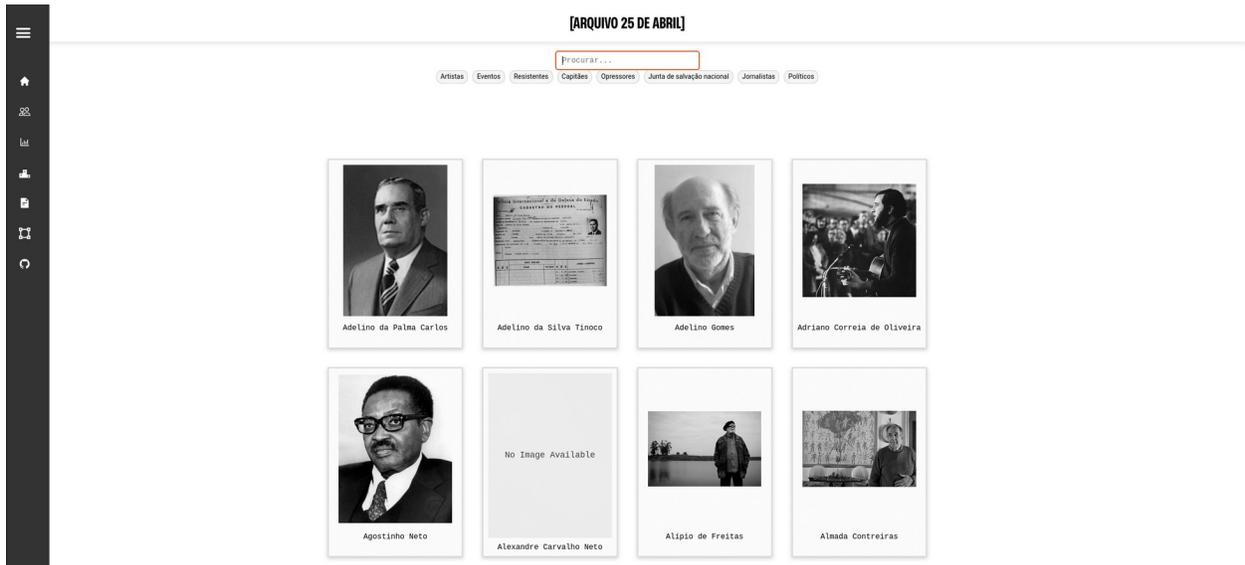
O site apresenta na página inicial um formulário de pesquisa aberto (este formulário permite fazer *full-text search* aos artigos recolhidos). Na barra lateral existem vários menus:



[ARQUIVO 25 DE ABRIL]

Entidades: Esta página apresenta todas as entidades de pesquisa. As entidades são carregadas com paginação automática (*infinite scroll*) para dar uma experiência melhor ao utilizador. Além disso é possível fazer pesquisa por entidades (um formulário permite fazer *full-text search* às entidades). É possível ver por categoria. Quando se passa o rato por cima de uma entidade aparece uma *tooltip* com a biografia de a entidade. Quando se clica numa entidade são listados todos os artigos associados à entidade ordenados por relevância (descrescente). Na página de listagem de artigos, os artigos são carregados com paginação automática, e quando se passa o rato por cima do título do artigo aparece uma *tooltip* com a *preview* (imagem) do artigo original ([linkToScreenshot](#)). Quando clicamos no artigo, somos redirecionados para o artigo original no Arquivo.pt.



[ARQUIVO 25 DE ABRIL]

ADELINO GOMES

*il | Carlos Gil
otas no 25 de
o mundo fora /
BLICO*

Público

alista que registou momentos de outros conflitos pelo mundo, é rtigo do jornal Público. No dia da lancia para a revista Flama, foi saiu imediatamente para as ruas contencimentos com sua máquina taria Judite, recorda a sua partida ra preparação, tendo até mesmo estados. O filho mais novo, Daniel rita do pai, mostrando a multidão eber os passageiros do primeiro a esperança e da revolução. Gil ular, mas também os momentos feridos perto da sede da PIDE. O destaca o papel de Gil como uma capacidade de registrar e narrar

*"Estes homens estiveram
frente a frente, tinham o
dedo no gatilho e não
dispararam" / Entrevista /
P*

Público

O livro "Os Rapazes dos Tanques", de Alfredo Cunha e Adelino Gomes, lançado em 2014, conta a história dos militares que defenderam o regime no 25 de Abril de 1974, no Terreiro do Paço. A obra revela, pela primeira vez com ambição, a perspectiva dos derrotados da revolução, focando-se nas memórias e experiências de oficiais, furiéis, cabos e soldados que estiveram frente a frente com os revolucionários. Um dos principais focos é a descoberta, após 39 anos de busca, da identidade do cabo apontador que recusou disparar contra a coluna de Salgueiro Maia. A pesquisa envolveu extensa investigação em arquivos militares, cartas e entrevistas, confrontando-se com dificuldades como a desconfiança inicial dos militares e a falta de registos oficiais da derrota. O encontro com o cabo

*Tertúlia literária assinala
25 de abril no Porto -
Babel*



Manuel Magalhães e Miguel Marinho. Desde novembro de 2011, o Porto de Encontro recebeu mais de duas dezenas de autores, incluindo Manuel António Pina, Gonçalo M. Tavares, Mário de Carvalho, Miguel Miranda e Dulce Maria Cardoso. Esta tertúlia literária, dedicada ao 25 de Abril, promete ser um evento rico em debate e reflexão sobre a importância desta data na história de Portugal. A escolha dos livros

*Livro sobre Salguei
lançado no dia ei
faria 70 anos - Obs*

Observador

No dia em que Salgueiro Maia completaria o 40º aniversário do 25 de Abril, a Anc a terceira edição do livro "Capitão de Abril", novos materiais. O prefácio, escrito por ' antigo membro do Movimento das Forças A Salgueiro Maia como um herói da histó Lourenço recorda como Salgueiro Maia escritos que dariam origem ao livro, expre de partilhar as suas experiências. Salguei em 1992, sem ver o livro publicado, ma amigos concretizaram o seu desejo em edição, lançada a 1 de julho, inclui três nov António de Sousa Duarte (biógrafo de Armando Fernandes e João de Melo (extrai de Paixões"). Estes juntam-se aos depoir de Matos Gomes, Francisco Sousa Tavares Cruzeiro e Vasco Lourenço, presentes na Duas entrevistas também são incluídas p

Top: Esta página permite ver o top (entre 1 a 100) de artigos. Permite ver o top ordenado por relevância ou por entidades associadas (esta listagem tem paginação automática). Quando se clica no artigo listado podemos ver as entidades associadas do lado direito (ao clicar na entidade vamos para listagem de todos os artigos dessa entidade), e do lado esquerdo temos um botão para o artigo original no Arquivo.pt.

[ARQUIVO 25 DE ABRIL]

Relevância Entidades Top 5 de artigos

arquivo.pt

OBSERVADOR

Como o 25 de Abril transformou a sede da PIDE num destino de turismo revolucionário /premium

16 Junho 2019 271

Como o 25 de Abril transformou a sede da PIDE num destino de turismo revolucionário – Observador

Observador

P

Política · PSD · PCP · PS · CDS-PP · BE · PAN

INVESTIGAÇÃO - TRANSIÇÃO PARA A DEMOCRACIA NO MNE (II)

Os diplomatas do Estado Novo são salazaristas, socialistas e até comunistas

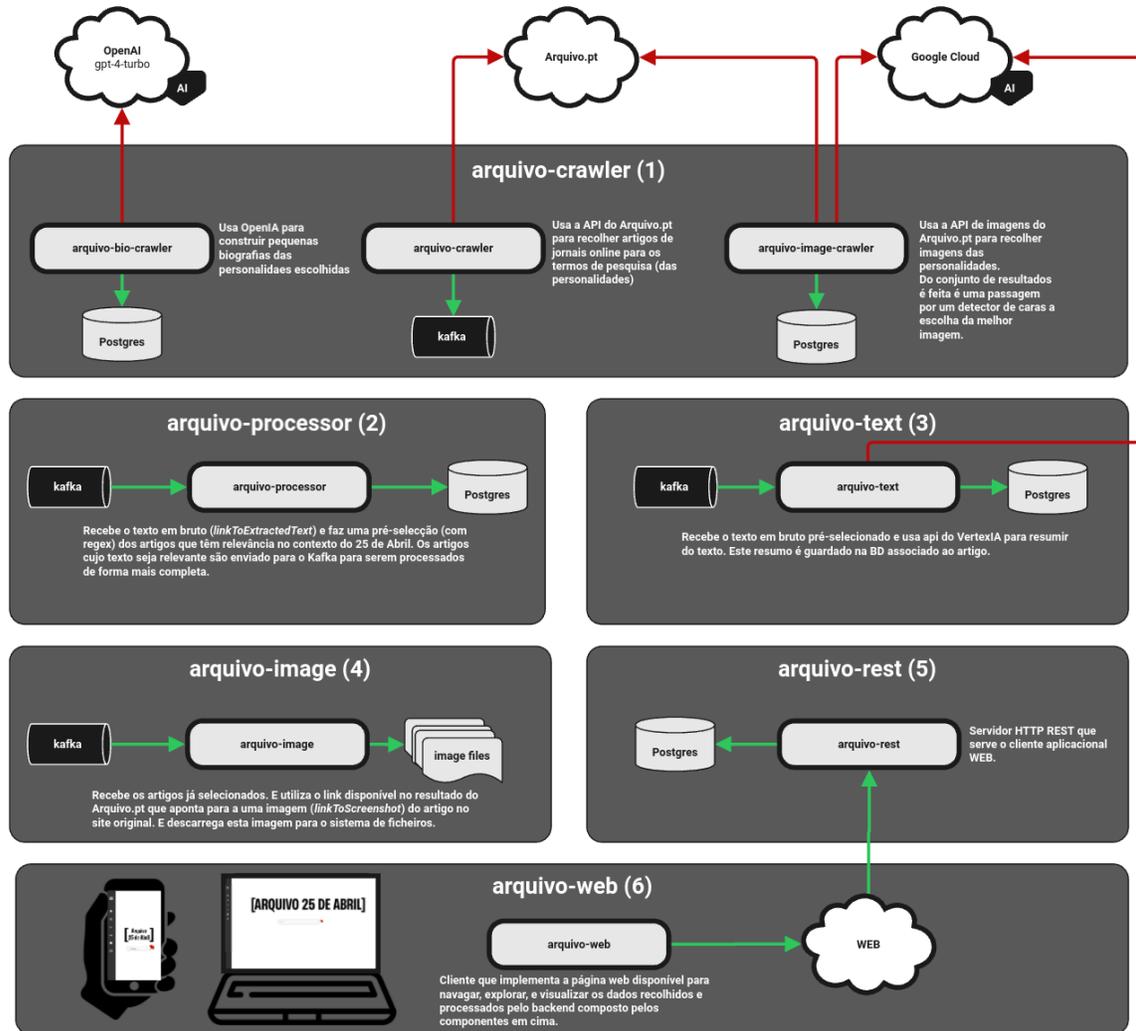
No Estado Novo havia diplomatas comunistas? "Claro que sim!", "o ministério não era monolítico". Nos 45 anos do Verão

Artigos associados:

- Artigo de António Costa
- Artigo de António Oliveira Salazar
- Artigo de Costa Cordeiro
- Artigo de António Guterres
- Artigo de José Saramago
- Artigo de José Cardoso Pires
- Artigo de João de Deus
- Artigo de Sá Carneiro
- Artigo de Mário Soares

Sobre: Nesta página é feita apresentada a motivação, algumas considerações pessoais, agradecimentos e referências.

Arquitectura: Nesta página é apresentada (figura) e descrita a arquitectura do projecto. Cada componente é explicado com algum detalhe. E são explicadas algumas decisões.



(1) arquivo-crawler: Este componente é responsável pela recolha de dados de várias fontes externas. O projecto assenta em dois conceitos principais: entidades de pesquisa (pessoas, eventos, locais) e artigos noticiosos. Para cada entidade de pesquisa, recolhemos uma biografia, uma fotografia e artigos noticiosos relacionados. Isto é feito através de três crawlers especializados.

arquivo-bio-crawler: Cada entidade de pesquisa tem uma biografia associada. Esta biografia permite ao utilizador do site saber quem é a entidade de pesquisa. Este crawler utiliza a API da OpenAI para obter uma biografia tendo em conta o contexto Histórico do Estado Novo e do 25 de Abril com as entidades de pesquisa.

Porquê IA? Numa versão anterior do projecto, utilizámos o Wikipedia (DBPedia) para obter a

biografia, no entanto, os resultados não eram completos e a informação não estava abreviada. Muitas vezes os resultados tinham metacaracteres e não eram legíveis ou relevantes neste contexto. No entanto, a OpenIA, mesmo utilizando o gpt-4-turbo não é uma solução perfeita. Isto é, não conseguimos obter uma biografia completa e relevante para todas as entidades. Estas LLMs ainda têm algumas limitações e não conseguem obter informação relevante para todas as entidades.

arquivo-crawler: Este componente recolhe os dados do Arquivo.pt através da [API](#). Para cada uma das entidades de pesquisa e para cada site é criado um URL para fazer a recolha de artigos, este vai retornar todos os resultados do Arquivo.pt entre a data inicial que o Arquivo.pt permite e o dia corrente. O próprio crawler itera sobre as várias páginas. Este componente não faz nenhum tipo de processamento, apenas faz o crawling e envia os dados para o arquivo-processor.

arquivo-image-crawler: Este componente recolhe fotografias para as entidades de pesquisa. A fonte de dados é o Arquivo.pt através da API de imagens. Para garantir que as imagens são relevantes, utilizámos vários critérios de selecção para garantir que das imagens devolvidas pelo Arquivo.pt escolhamos a mais correcta para a entidade de pesquisa em questão. Os dois critérios principais foram encontrar referências do nome da entidade nos campos da imagem: `pageTitle`, `imgAlt`, e `imgCaption`. Depois utilizámos um detector de caras da `google-api`. Estes foram os critérios para escolher a melhor imagem.

(2) arquivo-processor: Este componente é o orquestrador principal do sistema. É responsável por: - Receber dados dos crawlers através do Kafka - Fazer uma primeira análise de relevância dos artigos usando palavras-chave contextuais - Distribuir o trabalho para os componentes especializados (arquivo-text e arquivo-image) - Gerir o estado do processamento de cada artigo O scoring inicial é feito através de expressões regulares que procuram termos relacionados com o Estado Novo e o 25 de Abril, bem como referências às entidades de pesquisa. Apenas os artigos que ultrapassem um threshold mínimo são enviados para processamento adicional.

(3) arquivo-text: Este componente utiliza IA para processar e analisar o conteúdo dos artigos. O processamento é feito em duas fases: 1. Geração de um resumo do artigo usando o VertexAI (PaLM2) 2. Análise de relevância do resumo em relação ao contexto histórico e às entidades associadas

Porquê IA? Na primeira versão deste projeto (2024), um dos problemas com que nos deparámos foi a dificuldade em limpar o texto "raw" disponível em `LinkToExtractedText`. Este texto é muito grande e contém muitos metacaracteres e tem todo o texto já sem tags de HTML presente na página. Por exemplo, em sites de notícias existem muitas colunas com notícias mais pequenas, cabeçalhos, rodapés, etc. Isto trazia dois problemas: (1) Por vezes era difícil perceber se o texto era realmente relevante no contexto por haver muito "lixo" à volta; (2) a forma como apresentávamos o texto do artigo de forma a mostrar ao utilizador do que se tratava o artigo, não era legível. Desta forma, decidimos utilizar um modelo de linguagem para fazer o resumo do texto. Ao fazer um resumo semântico de um texto "com

lixo" conseguimos ter um texto que é focado no que realmente é relevante, e depois fazer uma análise do contexto do projecto sobre esse texto reduzido e focado.

(4) arquivo-image: Este componente é responsável por fazer o download e processamento das imagens dos artigos. Recebe mensagens Kafka do arquivo-processor com os metadados do artigo e o URL da imagem (campo linkToScreenshot). As imagens são processadas, redimensionadas e optimizadas antes de serem armazenadas. Esta abordagem assíncrona permite-nos: - Reduzir a carga no Arquivo.pt - Optimizar o espaço em disco - Processar apenas imagens de artigos relevantes

(5) arquivo-rest: Este componente é responsável por fornecer uma API REST para disponibilizar os dados processados. Esta API é utilizada pela aplicação web.

(6) arquivo-web: Este componente é a aplicação web que suporta esta website. É uma aplicação React que utiliza o framework Vite.

Github: É um link para o repositório github do projecto. Todo o código está acessível. É possível correr o arquivo localmente com docker. Existe já um ficheiro com o esquema da base de dados e todos os dados das entidades.

Objetivos

O principal objectivo deste projecto é promover a memória do que foi a divulgação do maior acontecimento do século XX em Portugal, enquanto decorrem as celebrações dos 50 anos do 25 de Abril. Além disso, este projecto também pretende destacar a relevância da imprensa online como um arquivo histórico (através de um olhar jornalístico). A imprensa online é um meio de informação muito acessível, em particular, para os mais novos. Por isso, é de relevar a importância da imprensa como a fonte de informação verdadeira e isenta. E claro, este projecto também serve de homenagem aos que lutaram por um Portugal livre e sem o qual iniciativas como o Prémio Arquivo.pt não existiriam.

Resultados Atingidos

O principal resultado foi a criação de um arquivo temático com uma visão jornalística da história do 25 de Abril. O número considerável de artigos que foram escritos ao longo destes anos na imprensa portuguesa sobre o 25 de Abril (de uma forma geral), contribuíram para que este resultado seja (1) mais completo pois tem uma cobertura maior e (2) mais rico pois é possível cruzar diferentes fontes para a mesma história.

Originalidade e carácter inovador

Desde que se iniciaram as celebrações dos 51 anos do 25 de Abril de 1974 são muitas as iniciativas de preservação da memória histórica. Os jornais, as câmaras municipais e cinematecas têm feito pedidos de recolha de fotografias e vídeos da época. Surgem vários documentários, filmes e livros sobre o tema. No entanto, este projecto, tanto quanto sei, é o primeiro a reunir de uma forma organizada e ordenada material que estava disperso em vários repositórios (os jornais online) para essa preservação histórica.

Impacto social (aplicação e utilidade social)

Qualquer pessoa que se interesse pela História Portuguesa do Séc. XX encontrará neste projecto uma fonte de informação previamente curada pelos órgãos de comunicação social.

O contexto actual da política mundial tornam cada vez mais necessário conhecer bem a História recente. Através deste projecto é possível conhecer através de várias personalidades e temas várias histórias do fim da ditadura de Salazar e do início da democracia com o 25 de Abril.

Impacto científico (aplicação e utilidade científica)

Historiadores, jornalistas, ou escritores podem encontrar neste projecto vários factos, histórias e curiosidades que através deste projecto estão organizados pelos actores da história e ordenados pela sua relevância no tema do 25 de Abril. Um investigador pode procurar por

palavras-chave, por personalidade ou até encontrar todos os artigos relacionados com o 25 de Abril em cada uma das fontes noticiosas utilizadas.

Relevância da utilização do Arquivo.pt

O Arquivo.pt preserva toda Web Portuguesa desde 1991. Neste projecto o Arquivo.pt funcionou como um motor de busca de artigos de jornais, visto que é possível fazer pesquisas por um termo, num determinado site e num intervalo temporal. Sem este trabalho do Arquivo.pt seria impossível recolher alguns dos artigos que já não estão acessíveis através dos sites actuais.

Comentários adicionais

No desenvolvimento deste projecto foram encontradas algumas limitações. Em particular na recolha de imagens e biografias para as personalidades.

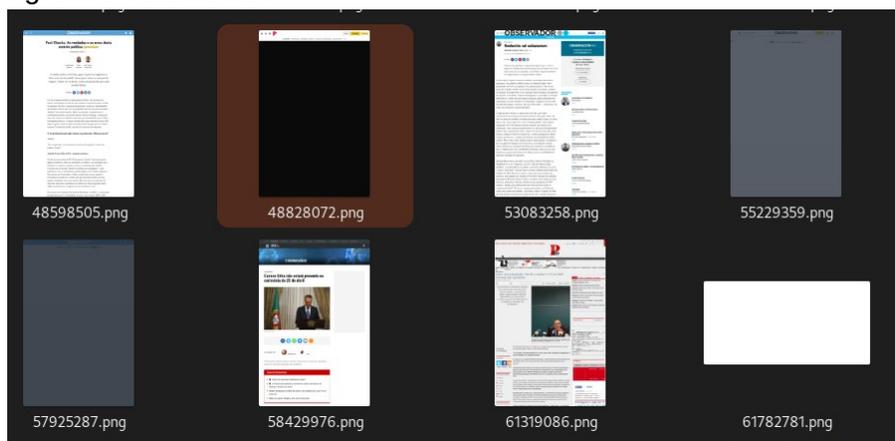
Biografias:

Algumas das personalidades pesquisadas não têm biografia associada. O modelo utilizado pela OpenIA não conseguiu criar (com base nas suas fontes) uma biografia. No entanto, para algumas destas personalidades, mesmo fazendo uma pesquisa na web não se encontram grandes resultados (e.g., wikipedia). É de notar que grande parte destas personalidades são ex-PIDEs, o que demonstra que ao longo destes 50 anos estas pessoas permaneceram na sombra do escrutínio e da História.

Imagens:

1) A API de imagens do Arquivo.pt foi também utilizada para recolher fotografias das personalidades pesquisadas. Para garantir que esta associação era correcta foram utilizados alguns critérios: scoring do texto das imagens e utilização de um detector de caras. No entanto, ainda existiram alguns casos em que a imagem seleccionada automaticamente não representava a personalidade em questão. Para esses casos residuais, as imagens for corrigidas manualmente.

2) As imagens utilizadas para fazer um preview dos artigos nem sempre é útil. O screenshot às vezes está branco/preto ou por vezes não tem informação útil visível (por exemplo, o artigo tem acesso reservado). Esta imagem vem directamente dos resultados do arquivo, por isso não foi possível fazer grande tratamento.



Recursos complementares

- *25 de Abril na Imprensa online, <https://arquivo25abril.com/>, o site onde é disponibilizado este projecto*
- *Github do projecto: <https://github.com/MiguelGarciaTH/25abril>*
- *API Arquivo.pt <https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API> o API do Arquivo.pt para web sites*
- *ImageSearch [https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1.1-\(beta\)](https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1.1-(beta)) a API do Arquivo.pt para imagens*