Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- **Título**: ArquivoNC – O arquivo web do Jornal Notícias da Covilhã.
- **Área temática**: Memória digital; Arquivos Web; Recuperação e Extração de Informação; Ciência de Dados
- **Candidato**: Rodrigo Silva, Ricardo Campos
- **Email**: rd.silva@ubi.pt, ricardo.campos@ubi.pt
- **Website**: https://arquivonc.ubi.pt

Descrição do Trabalho

Cerca de 80% dos *websites* mudam após um ano. Outros tantos desaparecem totalmente, levando à perda de informação de valor incalculável. O *ArquivoNC* – o arquivo web do jornal Notícias da Covilhã - surge nesse contexto, com o intuito de preservar a história digital da cidade da Covilhã e o legado de um jornal centenário. Com este projeto, recuperamos o acesso a um conjunto de dez anos de dados, perdidos em 2019, aquando do desaparecimento da anterior versão do website (ver Figura 1) e o fim da publicação (em papel) da edição semanal do jornal[1].

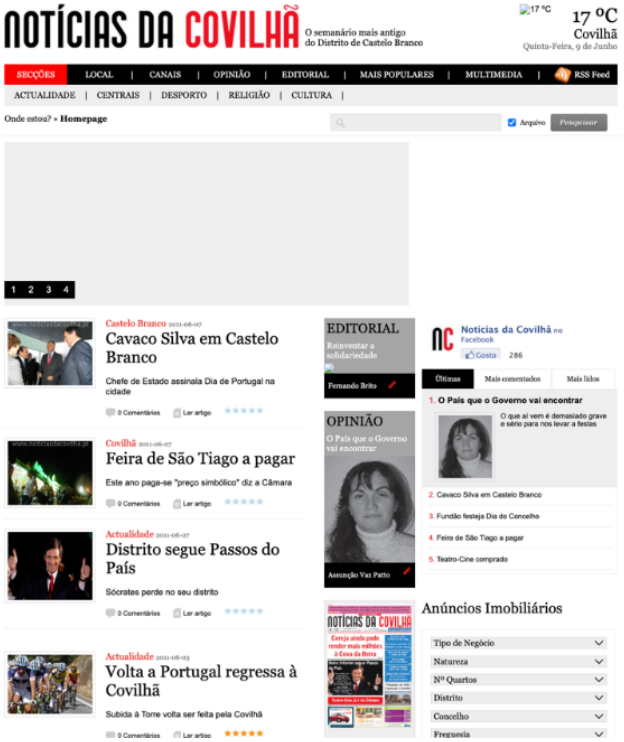

Figura 1 – Edição digital do Jornal Notícias da Covilhã a 09 de junho de 2011.

[1] entretanto retomada a 9 de março de 2023, mas sem a possibilidade de aceder a conteúdos anteriores a 2019.

O projeto agora desenvolvido, disponibiliza o acesso (ver Figura 2) a dez anos de páginas web do jornal Notícias da Covilhã a partir das notícias preservadas pelo Arquivo.pt entre 2009 e 2019, permitindo a jornalistas, radialistas[2], historiadores, estudantes e população em geral, ter acesso a um conjunto de notícias, imagens e capas do jornal publicadas no passado, à data em curso. O seu desenvolvimento é também um importante recurso para o próprio jornal Notícias da Covilhã, que assim recupera o acesso a um conjunto de vasta informação perdida em 2019.

Figura 2 – Website do ArquivoNC (https://arquivonc.ubi.pt). 6 de maio de 2024.

A arquitetura deste projeto é baseada em três módulos: (1) Extração de Informação; (2) Indexação, Pesquisa e Similaridade; (3) Desenvolvimento e Alojamento do Website

Extração de Informação

Para a concretização deste projeto recorremos ao Arquivo.pt. Em concreto, foram consideradas as **2979** versões do website do jornal Notícias da Covilhã preservadas pelo Arquivo.pt no período de tempo compreendido entre 2009 e 2019. Para obter os URLs das 2979 versões, recorremos ao pacote de software Python "PublicNewsArchive" (prémio Arquivo.pt 2022). Para automatizar a extração de informações foram aplicadas técnicas de web scraping que resultaram na obtenção de **2661** notícias, **1327** imagens e **372** capas do jornal. Para complementar a informação extraída e enriquecer o conteúdo das notícias coletadas recorremos ao software YAKE[3], para a extração de palavras-chave relevantes, e ao spaCy[4] para identificar e extrair entidades presentes no texto.

Indexação, Pesquisa e Similaridade

Para a indexação e pesquisa dos dados extraídos recorremos à base de dados NoSQL Redis[5]. Em concreto, foram construídos três índices de dados para dar resposta aos diferentes tipos de pesquisa: notícias, imagens e capas de jornal. Por exemplo, cada elemento da notícia (título,

[2] possibilidade da criação de diferentes rúbricas à volta do tema “faz hoje 10 anos que aconteceu...”
[3] http://yake.inesctec.pt/
[4] https://spacy.io
[5] https://redis.io/

conteúdo, data, etc.) é indexado num índice invertido e representado pela frequência dos seus termos utilizando a medida TF.IDF (*Term Frequency - Inverse Document Frequency*). Essa abordagem tradicional permite aos utilizadores pesquisarem por notícias que contenham um termo específico (ou conjunto de termos) num determinado período de tempo. Além disso, cada notícia é também representada no espaço vetorial a partir de um vetor de palavras de 512 dimensões, gerado pelo modelo de linguagem natural BERTimbau[6]. A representação semântica de cada notícia, permite a aplicação de um algoritmo de pesquisa de similaridade (*Approximate Nearest Neighbors*) e está na base de uma das principais funcionalidades deste projeto: **a recomendação de notícias do passado**. Os outros dois índices oferecem ao utilizador a pesquisa num universo de 1327 imagens (incluindo a **recomendação de imagens similares**), bem como a pesquisa de elementos textuais no conjunto das 372 capas indexadas.

Desenvolvimento e Alojamento do website

Para o desenvolvimento do website recorreu-se à framework Flask[7] e à ligação com a base de dados NoSQL Redis para a obtenção dos elementos dinâmicos. A virtualização desta arquitetura é garantida a partir de um sistema Docker[8].

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. Em Portugal, a preservação de conteúdos da web fica a cargo do Arquivo.pt. O objetivo deste projeto passa por utilizar os recursos do Arquivo.pt para preservar a memória digital da Covilhã e o legado do jornal **Notícias da Covilhã** ao tornar o seu conteúdo histórico facilmente acessível a investigadores e ao público em geral.

A disponibilização desses conteúdos através de um website dedicado ao arquivo web do jornal, visa contribuir para a preservação do património local e complementar a informação disponibilizada na atual versão do website, recuperando o acesso a um conjunto de notícias, imagens e capas do jornal perdidas em 2019 com o desaparecimento da anterior versão do website e o fim da publicação (em papel) da edição semanal do jornal (retomada a 9 de março de 2023).

[6] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[7] https://flask.palletsprojects.com/en/3.0.x/
[8] https://www.docker.com/

Resultados Atingidos

Do desenvolvimento deste projeto resultam dois importantes contributos: (1) a compilação de um dataset de dez anos de notícias; e (2) o desenvolvimento e a publicação online de um website.

Dataset

O dataset criado é composto por **2661** notícias, **1327** imagens e **372** capas do jornal Notícias da Covilhã recolhidas a partir de **2979** versões preservadas pelo Arquivo.pt entre 2009 e 2019. A Tabela 1 apresenta as estatísticas estratificadas por ano.

Tabela 1 – Estatísticas do dataset por ano

ANO	VERSÕES	NOTÍCIAS	CAPAS
2009	3	54	2
2010	245	438	39
2011	363	461	49
2012	159	333	26
2013	124	248	17
2014	200	155	26
2015	366	186	50
2016	366	210	53
2017	369	40	43
2018	392	83	48
2019	392	453	20

Website

O website encontra-se disponível a partir do seguinte endereço: https://arquivonc.ubi.pt. A sua organização segue uma estrutura similar à anterior versão do website do jornal do Notícias da Covilhã (com as notícias categorizadas por “Secções”, “Local”, “Opinião” e “Editorial”). procurando manter-se desta forma uma ligação entre os dois websites. Em termos gráficos optou-se por adotar um grafismo mais atual, mantendo-se apenas a cor base do jornal (vermelho). Além das categorias acima referidas, foram também incluídos novos elementos na barra de navegação (ver Figura 3), nomeadamente, a possibilidade de aceder às notícias por “Anos”, às “Capas” e aos “Autores” das notícias. Nessa mesma barra é possível encontrar um item de acesso rápido à “Pesquisa” e aos detalhes do projeto “Sobre”.

Figura 3 – Barra de navegação

A página inicial encontra-se dividida em quatro secções distintas: (1) notícias principais; (2) antigamente era assim; (3) recordar o passado; (4) especial 25 de abril.

Na secção **Notícias Principais** (ver Figura *4*) o utilizador pode navegar por um conjunto de notícias do dia em curso, publicadas há X anos atrás. Esta secção encontra-se implementada para que diferentes notícias sejam apresentadas ao utilizador a cada novo dia. Através desta funcionalidade, garantimos não só o dinamismo do website, mas também uma ligação mais efetiva entre o passado e o presente.

Figura 4 – Página Inicial. Secção “Notícias Principais”

Na secção **Antigamente era Assim** (ver Figura 5) o utilizador pode navegar por um conjunto de capas da semana em curso, publicadas há X anos atrás. Cada capa presente nesta secção é um objeto clicável permitindo aos utilizadores efetuarem ações “zoom-in” e “zoom-out”.

Figura 5 – Página Inicial. Secção “Antigamente era assim”

Na secção **Recordar o Passado** (ver Figura 6) o utilizador pode efetuar pesquisas de texto em notícias, imagens e capas.

Figura 6 – Página Inicial. Secção “Recordar o Passado”.

A pesquisa por **texto** apresenta as notícias que incluem os termos pesquisados no período especificado, sendo possível ordenar os resultados de três maneiras distintas: relevância ou data (ascendente, descendente). No contexto das notícias, as datas assumem um papel central, sendo consideradas o elemento mais importante. A Figura 7 ilustra a pesquisa de texto para a os termos *Universidade da Beira Interior*.

Figura 7 – Página de resultados para a pesquisa de texto *Universidade da Beira Interior*.

Ao clicar numa notícia em específico (ver Figura 8), o utilizador tem acesso ao conteúdo completo da notícia. A palavra-chave mais relevante é enfatizada a vermelho. As restantes top-10 palavras mais relevantes (de acordo com o algoritmo YAKE!) são formatadas a negrito. No final de cada notícia é possível ter acesso ao autor, à categoria e subcategoria da notícia, ao conjunto de palavras-chave determinadas pelo YAKE!, e às entidades identificadas no texto através do spaCy. O utilizador tem ainda a possibilidade de partilhar a notícia em diversas redes sociais, através dos botões de partilha disponibilizados para esse efeito. A incorporação desta funcionalidade visa a divulgação deste projeto a um público mais vasto. Adicionalmente são também apresentadas ao utilizador as top-2 notícias mais similares.

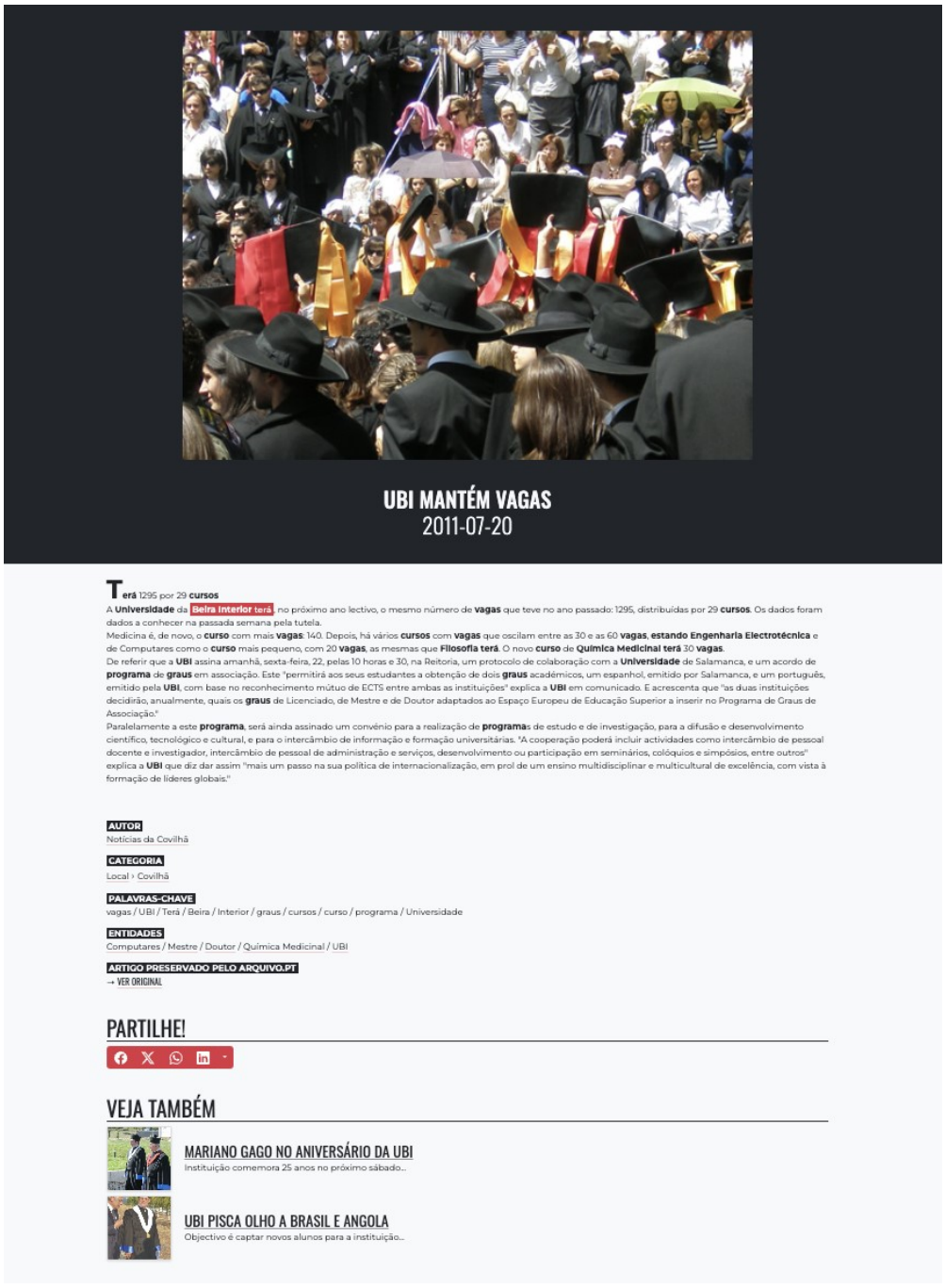

Figura 8 – Notícia com o título "UBI mantém vagas" devolvida para a pesquisa de texto *Universidade da Beira Interior*.

Por fim, é também disponibilizado ao utilizador a possibilidade de aceder ao artigo originalmente preservado pelo Arquivo.pt, e a partir do qual foram extraídas todas as informações (ver Figura 9).

Figura 9 – Artigo original da notícia com o título “UBI mantém vagas” preservado pelo Arquivo.pt

A pesquisa por **imagens** apresenta as imagens associadas aos termos pesquisados. A Figura 10 ilustra este processo para a os termos “*Serra da Estrela*”.

Figura 10 – Página de resultados de imagens para a pesquisa *Serra da Estrela*.

Ao clicar numa imagem em específico, o utilizador tem acesso à notícia associada à imagem, bem como a um conjunto de imagens relacionadas (ver Figura 11), umas das principais inovações (a par da recomendação de notícias) da nossa proposta.

Figura 11 – Imagem que ilustra a notícia “Primavera branca agrada na Serra” e as respetivas imagens relacionadas, devolvida para a pesquisa de imagens *Serra da Estrela*

Finalmente, a pesquisa por **capas** apresenta as capas do jornal que incluem os termos pesquisados no período de tempo especificado. A Figura 12 ilustra a página de resultados para o termo “*Portagens*”, um tópico atualmente em discussão por via de uma recente proposta que via a abolição das chamadas SCUTs em determinadas zonas do país.

Figura 12 – Página de resultados de capas para a pesquisa *Portagens*.

Ao clicar numa capa o utilizador tem a possibilidade não só de efetuar um processo de "*zoom-in*" como também de partilhar a capa através das redes sociais (em dispositivos móveis) ou guardá-la no PC. A Figura 13 ilustra a visualização da capa de 17 de março de 2011 em dois dispositivos diferentes, pc e telemóvel. Nas duas figuras é possível observar a notícia "Portagens atiram 17 mil para o desemprego"

Figura 13 – Visualização da capa a 17 de março de 2011 num PC (esquerda) e num dispositivo móvel (direita), devolvida para a pesquisa de capas *Portagens*

Na secção **Especial 25 de Abril** (ver A Figura 14) o utilizador pode visualizar notícias relacionadas com a comemoração dos 50 anos do Dia da Liberdade na Covilhã.

Figura 14 – Página Inicial. Secção Especial 25 de Abril

No desenvolvimento deste projeto, foram tidas em consideração preocupações ao nível da responsividade do website de forma a garantir o acesso a partir de diferentes dispositivos. A Figura 15 ilustra a visualização de diferentes secções da página web a partir de um dispositivo móvel.

Figura 15 – Visualização de diferentes secções a partir de um dispositivo móvel

Originalidade e caráter inovador

Este projeto é, de acordo com o nosso melhor conhecimento, o primeiro projeto a fazer uso dos dados do Arquivo.pt para a reconstrução/recriação de um website, a partir das versões preservadas por esta infraestrutura ao longo de um período alargado de 10 anos. Neste contexto, recuperamos o acesso a um vasto conjunto de informação inacessível ao público em geral desde 2019. Com a disponibilização online do website, contribuímos para a preservação da memória digital da Covilhã e do jornal Notícias da Covilhã. O projeto, destaca-se também pela integração de métodos de recuperação de informação e de algoritmos de similaridade de texto que permitem a recomendação de notícias e imagens similares. Através de uma interface gráfica moderna, os utilizadores podem assim pesquisar por notícias, imagens e capas do jornal do passado. O facto de as notícias do passado serem diariamente atualizadas na página principal, a par da possibilidade de os utilizadores poderem partilhar notícias através das redes sociais, são dois dos fatores que acreditamos, contribuirão para a utilização desta plataforma.

Impacto social (aplicação e utilidade social)

O projeto possibilita aos cidadãos da região da Beira Interior reavivar memórias da vida local, recuperando parte da história recente da Covilhã e das localidades que fazem parte da Beira Interior, em particular das regiões cobertas na Secção “Local”, i.e., Fundão, Castelo Branco, Guarda e Belmonte. Através desta plataforma é possível, por exemplo, escrutinar a efetiva concretização das propostas políticas apresentadas nos últimos 10 anos, consultar os artigos escritos por determinados autores (é o caso de um artigo escrito pelo ex-Presidente da República Mário Soares, aquando de uma homenagem que lhe foi feita na Covilhã), pesquisar por assuntos atuais (e.g., scuts, portagens) ou por personalidades públicas com ligação à cidade (e.g., António Guterres ou José Sócrates).

Impacto científico (aplicação e utilidade científica)

Este projeto facilita a investigação a jornalistas, radialistas, estudantes, entre outros. O projeto chama também a atenção para a necessidade de garantir a preservação de conteúdos dos jornais locais em colaboração, por exemplo, com o meio académico.

Relevância da utilização do Arquivo.pt

Para a realização deste trabalho fizemos uso dos dados preservados pelo Arquivo.pt, infraestrutura sem a qual não teria sido possível concretizar este projeto. No total foram coletadas **2661** notícias, **1327** imagens e **372** capas do jornal Notícias da Covilhã recolhidas a partir de **2979** versões preservadas pelo Arquivo.pt ao longo de 10 anos (2009 – 2019). A inexistência de dados anteriores a 2009 e posteriores a 2019, impossibilitam uma recolha mais alargada e são um exemplo claro da importância deste tipo de infraestruturas.

Comentários adicionais

Este projeto foi desenvolvido por Rodrigo Silva (aluno da Universidade da Beira Interior) sob a orientação de Ricardo Campos (Professor da Universidade da Beira Interior e afiliado ao INESCTEC e Ci2@IPT) no âmbito da unidade curricular de projeto de 1º ciclo em Engenharia Informática da Universidade da Beira Interior (UBI).

O projeto foi financiado por fundos nacionais através da agência de financiamento portuguesa, FCT - Fundação para a Ciência e a Tecnologia no contexto do projeto StorySense (DOI 10.54499/2022.144 09312.PTDC). Contou com a colaboração de Sérgio Nunes (Área de Sistemas e Desenvolvimento, UBI) e Paulo Crispim (Centro de Informática e Sistemas (CIS), IPTomar).

Recursos complementares

Arquivo.pt, https://arquivo.pt

PublicNewsArchive, https://github.com/diogocorreia01/PublicNewsArchive

YAKE!, https://yake.inesctec.pt

The Past Web – Exploring Web Archives,
https://link.springer.com/book/10.1007/978-3-030-63291-5
Dainel Gomes, Elena Demidova, Jane Wintes, Thomas Risse

Searching images in a web archive,
https://sobre.arquivo.pt/wp-content/uploads/SearchingImagesWebArchiveDSAA-202final.pdf
André Mourão, Daniel Gomes