Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- Título: **Olh'ó Regional**
- Área temática: Análise do jornalismo local em Portugal
- Candidato: Miguel Ramalho, João Carvalho
- Email: mr.miguel.ramalho@gmail.com, carrapa.joao.c@gmail.com

Descrição do Trabalho

O *Olh'ó Regional* é uma plataforma web de disponibilização e análise de dados sobre jornais e notícias regionais em Portugal nos últimos 30 anos.

Objetivos

1. Implementar e disponibilizar publicamente um processo automatizado de recolha e extração de notícias em larga escala, a partir do Arquivo.pt;
2. Compilar, enriquecer e partilhar uma base de dados de jornais e notícias regionais em Portugal dos últimos 30 anos;
3. Facilitar um interface de pesquisa e análise sobre a base de dados construída;
4. Apresentar casos de estudo que incidam sobre os dados recolhidos e que sirvam de inspiração à reutilização e maior adoção do projeto por cidadãos e académicos;
5. Permitir a preservação de notícias regionais no Arquivo.pt, que ainda não se encontrem lá.

Resultados Atingidos

O primeiro resultado a destacar é a validação da nossa hipótese: sim, é possível, através de ferramentas como o Arquivo.pt, criar um repositório de notícias regionais de Portugal com dimensão e variedade para que sobre ele sejam feitas análises que ajudem a compreender a realidade do jornalismo local em Portugal.

Destacamos ainda a materialização dos nossos objetivos no portal web *Olh'ó Regional*. Lá, vivem a nossa metodologia, um dashboard de análise sobre os dados recolhidos e enriquecidos, casos de estudo que exemplificam o tipo de análises que pode ser feita, e mais

funcionalidades como enviar uma notícia ainda online e não arquivada para o Arquivo.pt com um clique, procurar os jornais existentes e as suas notícias, aceder aos nossos dados enriquecidos, e fazer contribuições ou sugestões para o futuro do projeto.

Originalidade e caráter inovador

Destacamos:

- Integração de diferentes fontes de dados: Entidade Reguladora para a Comunicação Social (ERC), o Arquivo.pt, a Wayback Machine e a Memória da Imprensa Portuguesa;
- Um processo de download de notícias em larga escala que funciona para centenas de websites, precisando apenas do seu URL;
- Combinação de diferentes métodos de enriquecimento de dados, obtendo notícias sobre as quais produzimos metadados geográficos, temporais, semânticos, relacionais, de género e de autoria;
- Um sistema de exploração de todos esses dados, desenhado para ser acessível e útil tanto para especialistas como cidadãos comuns;
- Exploração exemplificativa dos nossos dados e plataforma através de temas como: desertos de notícias, desigualdade de género em citações, ou mesmo gastronomia local Portuguesa;
- Uma ferramenta e ambiente que encoraja cidadãos e académicos a explorar o jornalismo local, quer à escala nacional quer ao nível dos jornais e municípios individuais.

Impacto social (aplicação e utilidade social)

Numa perspetiva imediata, este trabalho pode ser usado como uma plataforma centralizada de acesso a notícias de jornais regionais que já não se encontram disponíveis. Aproximando cidadãos dos seus meios de notícias regionais, de histórias do seu próprio passado que já não estão online, e de memória dos próprios jornais que muitas vezes têm uma visibilidade ou capacidade de auto-preservação noticiosa reduzida.

Numa perspetiva futura, este projeto abre portas a novas ideias, iniciativas e aplicações que integrem e reflitam sobre os meios regionais de notícias no panorama de notícias nacional.

Impacto científico (aplicação e utilidade científica)

Uma das linhas orientadoras do *Olh'ó Regional* foi garantir que tanto a metodologia como o resultado final trariam utilidade à comunidade científica. Isso manifesta-se na partilha e documentação no nosso repositório de código com toda a metodologia de extração e análise

de dados, que envolveu técnicas avançadas de computação e machine learning. Também se manifesta na disponibilização dos dados extraídos e trabalhados a todos os que os queiram utilizar, quer através de uma transferência integral quer através da exploração na plataforma desenvolvida. Por fim, o design do dashboard de exploração de dados também foi criado com o intuito de permitir rapidamente testar hipóteses sobre notícias regionais, funcionando como ponto de partida para investigações mais profundas. Por fim, esperamos que a nossa disponibilidade em ajudar qualquer académico ou centro de investigação a tirar partido do código ou dados venha também a impactar o seu trabalho e produção científica sobre o tema.

Relevância da utilização do Arquivo.pt

O Arquivo.pt foi fundamental para a recolha de toda a informação que já não se encontra disponível online. Possibilitou que os dados usados no projeto cubram os últimos 30 anos, e websites que já nem existem, desbloqueando análises temporais e descobrindo conteúdos que de outra forma não podiam ser consultados ou sequer confirmados contra uma fonte de verdade. A sua API é bastante generosa e permitiu que milhões de páginas fossem extraídas a custo zero.

Comentários adicionais

Com o projeto publicado, vamos contactar equipas de investigação que tenham trabalhado o tema do jornalismo local em Portugal e dos seus desafios, não só para partilhar o projeto mas para integrar as suas sugestões e necessidades no tipo de processamento de dados e na forma como os disponibilizamos à comunidade científica.

Tendo trabalhado longas horas sobre os dados fornecidos pela ERC identificámos uma lista de desafios e sugestões à forma como os dados são representados e validados, iremos entrar de novo em contacto para não só partilhar construtivamente a nossa experiência como para divulgar o presente projeto.

Por fim, segue-se uma descrição detalhada da Metodologia de extração e processamento de dados:

1. Identificação de fontes de dados para o trabalho:

   O trabalho baseia-se em dois conjuntos de dados principais, os jornais e as notícias. Identificou-se a ERC como a fonte primária para dados dos jornais. Identificou-se a exposição Memória da Imprensa Portuguesa como uma fonte complementar para o dataset dos jornais. Identificou-se o Arquivo.pt como a principal fonte de dados para a recolha de notícias e validação das características dos jornais Identificaram-se a Wayback Machine e os sites dos próprios jornais que estão ativos, como fontes complementares de dados para a recolha de notícias.

2. Recolha dos dados da ERC sobre os jornais e critérios de exclusão de meios noticiosos:

   Contactou-se a ERC, que prontamente disponibilizou um dataset contendo a lista completa de publicações periódicas registadas em Portugal, ativas e inativas entre 2000 e 2026. A lista contém os seguintes campos por publicação e a metodologia de exclusão de publicações seguiu os seguintes critérios:
   - Todas as publicações de âmbito não regional foram excluídas.
   - Todas as publicações sem suporte online foram excluídas.
   - Todas as publicações de informação especializada foram excluídas.
   Define-se neste ponto os jornais que sobreviveram aos critérios de exclusão como "jornais regionais". A aplicação destes critérios de exclusão resultou numa lista de 663 jornais regionais.

   Em seguida procedeu-se a uma verificação individual de cada um dos registos no Arquivo.pt ou site ativo para validar que a publicação tem ou teve em qualquer espaço temporal, uma presença online verificável. Verificou-se também o conteúdo das notícias e foram excluídos todos os registos que apresentaram um carácter diferente de notícias regionais (como desporto, doutrinário, especializado ou notícias nacionais ou internacionais). A aplicação destes critérios de exclusão resultou numa lista de 462 jornais regionais.

   Incrementou-se também a lista com registos provenientes do projecto Memória da Imprensa Portuguesa, o que permitiu resgatar registos que estavam declarados na lista da ERC como não tendo site online mas que acabaram por ter. Procedeu-se à mesma verificação manual que no passo anterior o que permitiu fechar a lista com 573 jornais regionais.

   Reconhecemos que a esta lista faltarão Órgãos de Comunicação Social com presença online não comunicada à ERC, que só uma revisão manual mais extensa poderá colmatar.

3. Enriquecimento dos dados sobre os jornais regionais:

   Identificou-se manualmente as regiões, distritos e municípios que cada jornal cobre, fazendo uso dos respectivos Estatutos Editoriais e análise das próprias notícias. Esta cobertura é distinta da localização da sede: um jornal sediado em Lisboa pode cobrir a região do Alentejo, por exemplo.

4. Recolha das notícias dos jornais regionais:

   Procedeu-se a realizar uma extração/scrapping de todas as notícias em todos os jornais regionais. Os scrappers começaram por recolher todas as notícias do Arquivo.pt, e para os casos em que não foi possível, recolheram as notícias dos sites ativos e Wayback Machine. O nosso scraper consegue-se adaptar a variações na organização dos

websites mas ainda há vários jornais com notícias online para os quais só um código específico conseguirá extrair as notícias corretamente, a dimensão desse esforço não coube nesta primeira fase de desenvolvimento e enquadra-se nos principais objetivos futuros para a plataforma.

5. Processamento dos dados e categorização:

   Com base no texto disponível no título e corpo das notícias, encontraram-se os temas mais comuns que formaram as categorias usadas no processo de categorização das mesmas. A lista inicial de categorias foi feita manualmente e foi usada uma abordagem de descoberta automática de novas categorias que seleciona grupos significativos de notícias e identifica novas possíveis categorias, as quais foram manualmente verificadas.

6. Apresentação e disponibilização dos dados:

   Criou-se um site que funciona como plataforma de acesso aos conjuntos de dados e que permite fazer exploração e análise sobre os mesmos. Disponibilizou-se também na plataforma links para acesso a todos os conjuntos de dados relevantes.

Recursos complementares

- Plataforma "*Olh'ó Regional*", https://olho-regional.pages.dev/ : Plataforma principal para este projeto, descrita ao longo deste documento, e que convidamos a explorar
- Código do projeto, https://github.com/olho-regional/olho-regional : Código com licença de acesso aberto para reutilização; espaço público de esclarecimento de dúvidas sobre a metodologia e código, onde qualquer um pode pedir novas funcionalidades ou reportar erros