Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- Título: Desarquivo
- Áreas temáticas: Tecnologia, Investigação Jornalística, Verificação de Factos
- Candidato: Miguel Sozinho Ramalho
- Email: mr.miguel.ramalho@gmail.com

Descrição do Trabalho

O Desarquivo é um projeto que procura democratizar e complementar o **jornalismo de investigação** e a **verificação de factos**.

Tem por base a análise de milhões de notícias pertencentes aos principais jornais generalistas portugueses, escritas nos últimos 20 anos, e que se encontram preservadas no Arquivo.pt.

Permite explorar o grafo das ligações entre pessoas, organizações e locais que estavam escondidas no conteúdo jornalístico português, através de uma interface visual web simples e robusta.

Objetivos

O principal objetivo do Desarquivo é explicitar as ligações, até agora implícitas, presentes no histórico de notícias portuguesas, que acabam por ser um espelho da realidade do país, e facilitar a sua exploração para objetivos como a investigação jornalística ou verificação de factos. Apesar de serem estes os casos de uso primários, não há razão para que não seja também uma ferramenta de aprendizagem e exploração do contexto português, sua história, seus intervenientes, no fundo, de como as peças do puzzle que é Portugal encaixam.

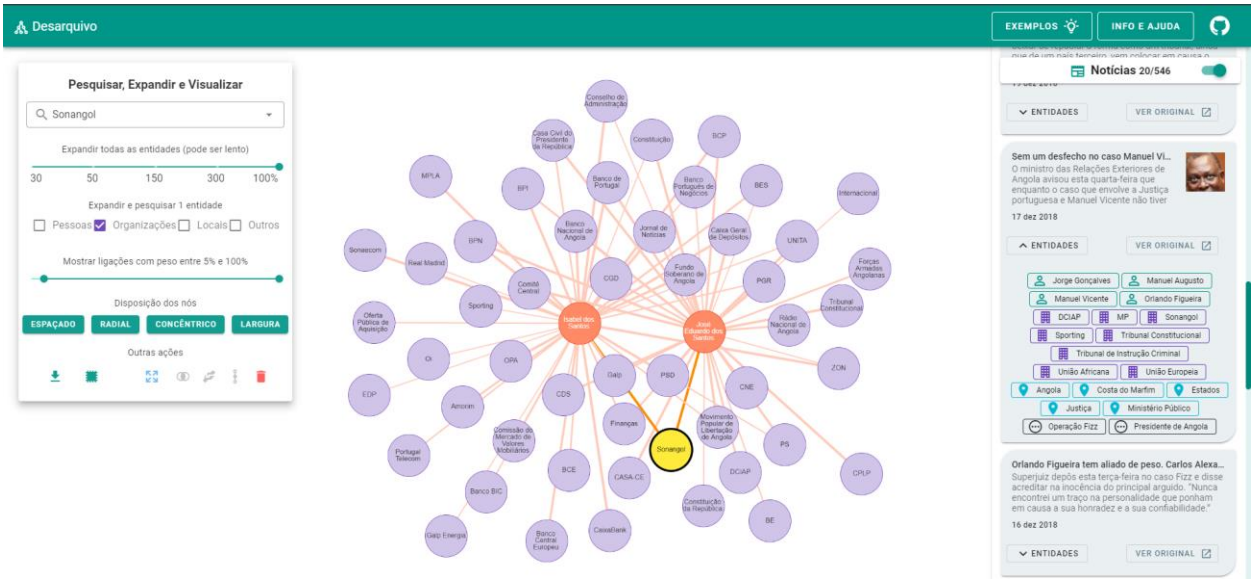
Portanto, idealiza-se que as ligações entre as pessoas, organizações (empresas, partidos políticos, universidades, sociedades, …), locais e outras entidades que são mencionadas em conteúdo jornalístico presente no Arquivo.pt sejam pesquisáveis de uma forma que facilita a investigação e exploração das relações entre elas.

Resultados Atingidos

Os objetivos deste trabalho desdobram-se sobre várias tarefas que culminaram na concretização dos objetivos acima propostos, como tal destacam-se os seguintes resultados, por ordem sequencial:

- Processo de extração
    - Análise de notícias arquivadas a partir da API do Arquivo.pt;
    - Extração do conteúdo relevante das páginas de notícias (ignorando links, excertos, anúncios, entre outro conteúdo poluidor);
    - Identificação das entidades (pessoas, organizações, locais e outros) mencionadas em cada notícia;
    - Extração das relações entre entidades, com base na presença mútua nas mesmas notícias;
    - Organização dos dados numa base de dados não relacional.
- Processo de limpeza
    - Identificação e processamento de entidades duplicadas/sinónimas. Por exemplo: “Marcelo Rebelo de Sousa”, “Professor Marcelo” e “Presidente Marcelo” dizem respeito à mesma pessoa; “UE” e “União Europeia” dizem respeito à mesma organização (este processo não é à prova de exceções, mas obteve resultados suficientemente bons para melhorar a qualidade dos dados finais);
    - Criação de um grafo de ligações e entidades únicas numa base de dados para grafos.
- Interface visual para exploração do grafo onde é possível
    - Pesquisar entidades pelo seu nome;
    - Expandir as relações de uma entidade;
    - Explorar das relações entre diferentes entidades (diretas e indiretas);
    - Gerar várias disposições visuais do grafo;
    - Interagir com o grafo;
    - Visualizar as notícias associadas a cada entidade;
    - Visualizar as notícias associadas a cada ligação entre duas entidades;
    - Exportar a visualização do grafo em formato imagem;

- Partir de exemplos base e explorá-los ou alterá-los.



Captura de ecrã da interface final produzida com um exemplo das principais organizações e empresas com que Isabel dos Santos e o seu pai se relacionam. Do lado esquerdo, o painel de interação e pesquisa. No centro, o grafo interativo. Do lado direito, notícias relativas à "Sonangol" por ser a organização selecionada no momento da captura.

Originalidade e caráter inovador

- Utilização, pela primeira vez, de conteúdo de notícias, arquivadas no Arquivo.pt, para inferir entidades que nelas participam e como estas se relacionam, criando um grafo que espelha a realidade portuguesa;
- Criação de uma forma inovadora de explorar esta rede de ligações que é, ao mesmo tempo, acessível a qualquer cidadão sem formação específica;
- Abertura do precedente deste tipo de ferramenta para outros contextos e países.

Impacto social

No que toca à investigação jornalística, o Desarquivo procura ter um impacto direto na forma com esta atividade é e pode vir a ser conduzida; procura aumentar a produtividade e, eventualmente, potenciar a descoberta de relações e pormenores que não são tão fáceis de identificar da forma tradicional (motores de busca por texto). Este impacto vai de encontro ao que foi conseguido em projetos como os de análise automática de fugas de informação que têm vindo a ser do

conhecimento público (Panama papers, Luanda Leaks, entre outras). Naturalmente, este impacto direto resulta num impacto indireto na vida dos cidadãos comuns por haver mais uma forma de procurar transparência e promover a justiça.

No que toca à verificação de factos, o Desarquivo serve como ferramenta de combate à desinformação e *fake-news* por representar uma nova forma de contrastar aquilo que é afirmado com aquilo que está registado nos últimos 20 anos de notícias em Portugal.

De forma mais geral, o desarquivo serve também como um motor de pesquisa que procura responder de forma diferente a questões que os cidadãos tenham e mesmo ajudar à compreensão das estruturas políticas, académicas, desportivas, culturais, etc. em Portugal. Os exemplos disponibilizados na plataforma realçam isso mesmo.

Por fim, o Desarquivo propõe-se como ferramenta de valorização do trabalho jornalístico que, desta forma, não fica apenas arquivado para consulta histórica. Cada notícia e cada artigo adiciona mais valor aos anteriores e, juntos, convergem numa rede que representa a visão que o país tem sobre si e sobre o exterior, ao longo do tempo.

Impacto científico

Em primeiro lugar, este projeto pode servir como fonte de inspiração para estudos que venham a aplicar a mesma estratégia a novos contextos.

Em segundo lugar, vem incentivar o desenvolvimento de melhores algoritmos nas áreas de processamento de linguagem natural, extração automática de notícias, reconhecimento de entidades, exploração de grafos, deteção de comunidades, entre outros para aplicação direta no contexto de notícias escritas em português.

Uma outra linha de investigação que daqui pode derivar é a automatização da deteção de *fake-news* com base em grafos de relações e notícias. Dado que as notícias de fontes jornalísticas bem estabelecidas e curadas representa, regra geral, o ponto de partida de qualquer trabalho de verificação de factos, este tipo de trabalhos tem agora um novo formato para explorar, que conjuga precisamente o conteúdo jornalístico de confiança com uma estrutura relacional que reforça a informação verdadeira e descarta a pouco fundamentada.

Por fim, todo o processo de recolha, processamento e análise de notícias encontra-se documentado de forma replicável, e ficará aberta ao público no fim do presente concurso. Tal representa um ponto de partida para investigadores que se queiram debruçar sobre alguns dos desafios que o projeto acarta e melhorar ou adaptar o trabalho já feito, podendo mesmo haver uma contribuição comunitária nas versões seguintes do projeto.

Para além do grafo de relações entre entidades, foi também gerado um grafo de relações diretas entre entidades e notícias. Ambos serão disponibilizados publicamente como fontes para trabalhos de investigação científica. De realçar que a interface desenvolvida visa o uso pelo

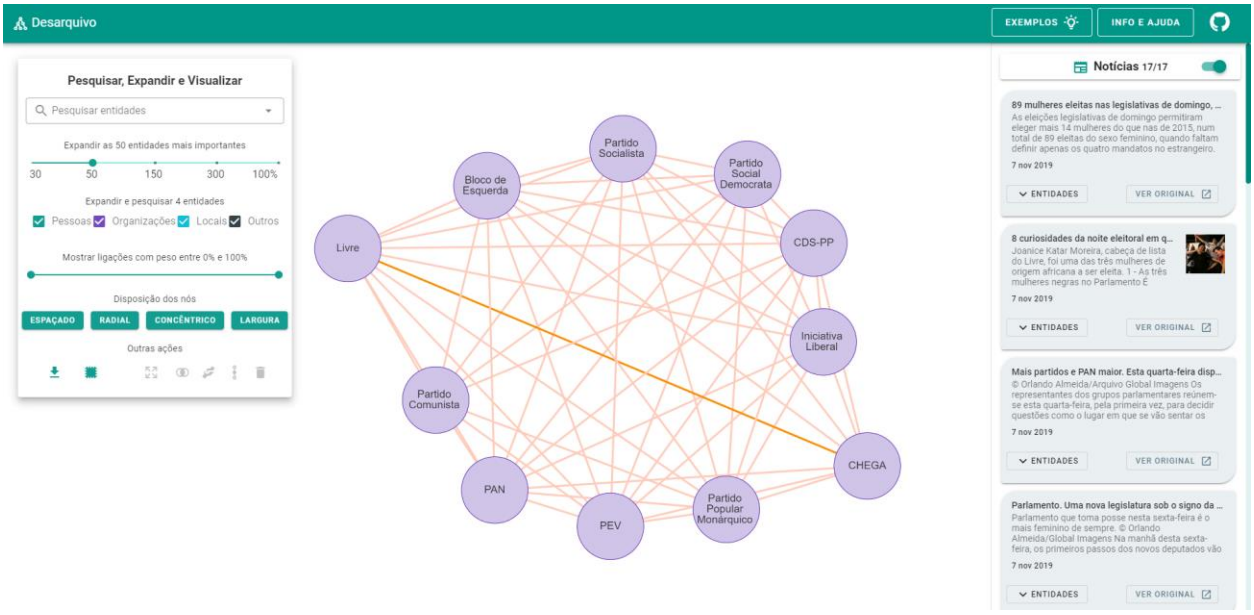
cidadão comum mas que investigadores conseguirão fazer pesquisas muito mais complexas através de linguagens de pesquisa em bases de dados de grafos, como sendo o CypherQL.

Relevância da utilização do Arquivo.pt

Sem o Arquivo.pt dificilmente se conseguiria ter acesso a conteúdo jornalístico tão consistente e uniformizado tanto ao longo do tempo como nos diferentes temas que aborda. É mesmo seguro afirmar que nunca seria possível fazer um trabalho tão completo sem o Arquivo.pt precisamente porque cada fonte de notícias que se quisesse analisar iria requerer um esforço individual grande, face ao acesso holístico que se consegue com as APIs de pesquisa disponibilizadas pelo Arquivo.pt. Destas, foram utilizadas:

- Arquivo.pt API v.0.2 (para pesquisa de notícias)
- URL search: CDX server API (para otimização das pesquisas feitas sobre a primeira, com base nas datas em que os diferentes sites tinham notícias arquivadas)

Ainda esteve em cima da mesa a hipótese de usar a ImageSearch API v1 (beta) para obter imagens a associar a cada entidade na visualização gráfica mas tal não foi feito por se estimar que houvesse um impacto negativo na performance da ferramenta quando estivessem muitos nós a ser visualizados, não estando essa experiência excluída de versões futuras.



Exemplo de conteúdo preservado utilizado: As notícias do lado direito ("89 mulheres nas legislativas de domingo…", "8 curiosidades da noite eleitoral…", "Mais partidos e PAN maior...", "Parlamento. Um nova legislatura sob o signo da …") são todas exemplos de notícias arquivadas de onde foram extraídas entidades e, que neste caso, correspondem à ligação entre o partido Livre e o partido CHEGA.

O nome fala por si, mas o Desarquivo vem mostrar que o valor do Arquivo.pt, mais do que histórico, é intemporal e que, sem o Arquivo.pt, não haveria nada para *desarquivar*.

Comentários adicionais

De forma a tornar o uso do Desarquivo mais fácil, foram criadas duas secções na própria página (canto superior direito) que dão acesso tanto a exemplos como a instruções de utilização. Sobretudo porque se considera que um dos custos da inovação é, precisamente, a adaptação.

Recursos complementares

- Desarquivo, https://msramalho.github.io/desarquivo/ : A página principal do projeto, onde se encontra a interface visual, instruções e exemplos de utilização;
- Código desenvolvido para o desarquivo, https://drive.google.com/file/d/13rRen-uRFHmLXN5gp8M0wJZP1kh6xnDZ/view?usp=sharing, Onde se encontra o código respetivo à recolha, processamento, análise, organização e visualização de dados, bem como instruções para investigadores;
- Repositório *open-source* do código e dados (este link será tornado público no final do concurso e terá o mesmo conteúdo que o link acima mas com potencial de colaboração comunitária), https://github.com/msramalho/desarquivo : Onde se encontrará o código respetivo à recolha, processamento, análise, organização e visualização de dados, bem como instruções para investigadores. Vai permitir que haja utilizadores a contribuir para a ferramenta, quer por sugestões e identificação de erros quer por adição de novas funcionalidades;
- Vídeo de participação, https://youtu.be/tVlOUuRqIVU, O mesmo vídeo que foi submetido para o concurso, mas online.