Prémio Arquivo.pt

Descrição Sumária do Trabalho

Identificação

- **Título:** Arquivo Público
- **Área temática:** Recuperação e Extração de Informação; Ciência de Dados; Jornalismo de Dados.
- **Candidato:** Diogo Correia; Ricardo Campos
- **Email:** aluno81470@ipt.pt; ricardo.campos@ipt.pt
- **Website:** http://arquivopublico.ipt.pt

Descrição do Trabalho

Este projeto consiste no desenvolvimento de um sistema de recuperação de informação e análise de dados das notícias publicadas pelo jornal Público no domínio http://www.publico.pt/ ao longo do período de tempo compreendido entre 2010 e 2021. Para a realização deste projeto recorremos à API TextSearch do Arquivo.pt para obter as várias *homepages* do jornal Público ao longo do período de tempo considerado. Posteriormente fazemos uso de técnicas de web scrapping para através dessas páginas principais, automatizar o processo de extração de informação, nomeadamente o título, a descrição, a data, o link e o/a autor/a de cada notícia.

A alteração do visual gráfico na página principal do jornal público ocorrida nos anos de 2012 (ver Figura 1) e 2017 (ver Figura 2) obrigou a uma adaptação do processo de webscraping por parte da nossa equipa em cada um desses anos.

Figura 1 – Página do jornal Público em 2012.

Figura 2 – Página do jornal Público em 2017.

No total foram recolhidas 39.788 notícias. A Tabela 1 lista o número total de notícias coletadas em cada ano considerado.

Tabela 1: Total de Notícias Coletadas no período compreendido entre 2010 e 2021

Ano	*Total de Notícias*
2010	4024
2011	6344
2012	2364
2013	1296
2014	3702
2015	6580
2016	8595
2017	8007
2018	4454
2019	18862
2020	2328
2021	689

A escalabilidade desta solução encontra-se garantida para os próximos anos ficando apenas sujeita a adaptações decorrentes de novas alterações na interface gráfica que venham a ser realizadas por parte do Jornal Público. A solução desenvolvida encontra-se publicamente disponível a partir do website arquivopublico.ipt.pt e encontra-se dividida em duas partes distintas: (1) um motor de busca temporal, através do qual os utilizadores poderão fazer pesquisas acerca de qualquer tópico que tenha sido objeto de cobertura noticiosa por parte do jornal Público; (2) uma análise dos dados obtidos, com enfoque no conjunto de palavras-relevantes aí detetadas, localidades, organizações e pessoas mencionadas nos artigos noticiosos.

Para o desenvolvimento do motor de busca recorremos ao Elastic Search[1] uma base de dados NoSQL que permite indexar dados não estruturados. Em particular, usamos o Elastic Search

[1] https://www.elastic.co/pt/

para (1) indexar as informações textuais coletadas no período de tempo considerado; (2) devolver os resultados da pesquisa através de um algoritmo de relevância aí implementado.

Adicionalmente, fazemos uso dos dados obtidos para efetuar uma análise pormenorizada das notícias coletadas. Assim, recorremos ao YAKE[2], um extrator de palavras relevantes desenvolvido pela nossa equipa de investigação, para obter o conjunto de palavras relevantes ao longo dos anos. Adicionalmente fazemos uso do Spacy[3] para obter informações acerca das pessoas, localidades e organizações citadas no conjunto de todas as notícias. A extração das localidades, não sendo um processo totalmente fidedigno (dado tratar-se um processo automático), permite, ainda assim, ter uma noção da cobertura geográfica das notícias indexadas. Nesse sentido, fazemos uso de um processo de geocoding, para através das localidades procedermos à extração das coordenadas geográficas e ao seu correspondente mapeamento.

A singularidade dos tempos vividos nos últimos 2 anos, decorrentes de uma pandemia, impulsionou também a nossa curiosidade na tentativa de entender a evolução e o número de notícias produzidas ao longo dos anos de 2020 a 2021 na página principal do Jornal Público relacionadas com a pandemia. De forma similar ao descrito anteriormente, optámos por obter informações acerca das palavras mais relevantes, bem como das localidades, pessoas e organizações mencionadas nos artigos no primeiro ano do Covid.

Objetivos

Os principais objetivos deste projeto passam por disponibilizar um sistema de pesquisa de informação que permita aos utilizadores efetuar pesquisas sobre qualquer assunto que tenha sido objeto de cobertura noticiosa por parte do Jornal Público no domínio publico.pt, entre os anos 2010 e 2021. Paralelamente, oferecemos aos utilizadores uma análise de dados mais profunda (incluindo sobre o Covid-19) realizada a partir dos dados obtidos.

Resultados Atingidos

Através deste projeto foi possível coletar, analisar e disponibilizar através de um sistema de pesquisa, acesso a 39.788 notícias entre os anos de 2010 e 2021, grande parte dos quais (cerca de 19.000) relativas ao ano 2019. Os anos anteriores a 2010 (com exceção do ano de 2004 e 2005) revelaram-se ter poucas notícias coletadas e preservas no Arquivo.pt, razão pela qual não procedemos à sua coleta e respetiva análise. O projeto em questão encontra-se disponível a partir do link arquivopublico.ipt.pt. No topo do website encontra-se disponível o sistema de busca que permite ao utilizador pesquisar por um tópico do seu interesse. A Figura 3 exemplifica esse processo através da *query* "*Cristiano Ronaldo*". O website encontra-se parametrizado para que a pesquisa e os resultados a devolver cubram o espectro de notícias compreendidas entre 2010 e 2021. Vale a pena no entanto notar que as opções avançadas dão ao utilizador a possibilidade

[2] http://yake.inesctec.pt
[3] https://spacy.io/

de restringir a pesquisa ao período de tempo compreendido. É importante também referir que o website desenvolvido é responsivo podendo ser acedido via pc local ou dispositivo móvel.

Figura 3 – Exemplo de query Vladimir Putin.

Os resultados obtidos podem ser visualizados a partir de uma lista de resultados em modo similar ao que acontece com o Google. A Figura 4 mostra 5 de um total de 95 resultados.

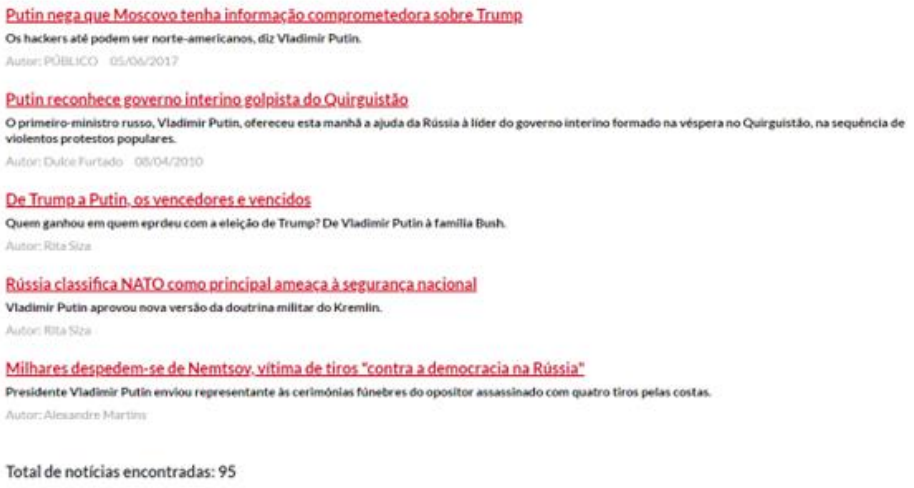

Figura 4 – Resultados obtidos para a query Putin.

Ao clicar num dos títulos, o utilizador tem acesso à página web através do Arquivo.pt (ver Figura 5).

Figura 5 – Página web preservada pelo Arquivo.pt.

O utilizador tem ainda disponível algumas sugestões relevantes para pesquisa (ver Figura 6).

Figura 6 – Sugestões de Pesquisa.

Adicionalmente, facultamos aos utilizadores do Arquivo Público uma análise de dados obtida a partir dos dados coletados. A

Figura 7 ilustra o conjunto das cinco organizações localidades e pessoas mais referidas nos artigos do jornal Público (no domínio publico.pt) no ano de 2015.

Ao clicar nas imagens os utilizadores são automaticamente redirecionados para o conjunto de artigos que mencionam a respetiva entidade no ano em causa.

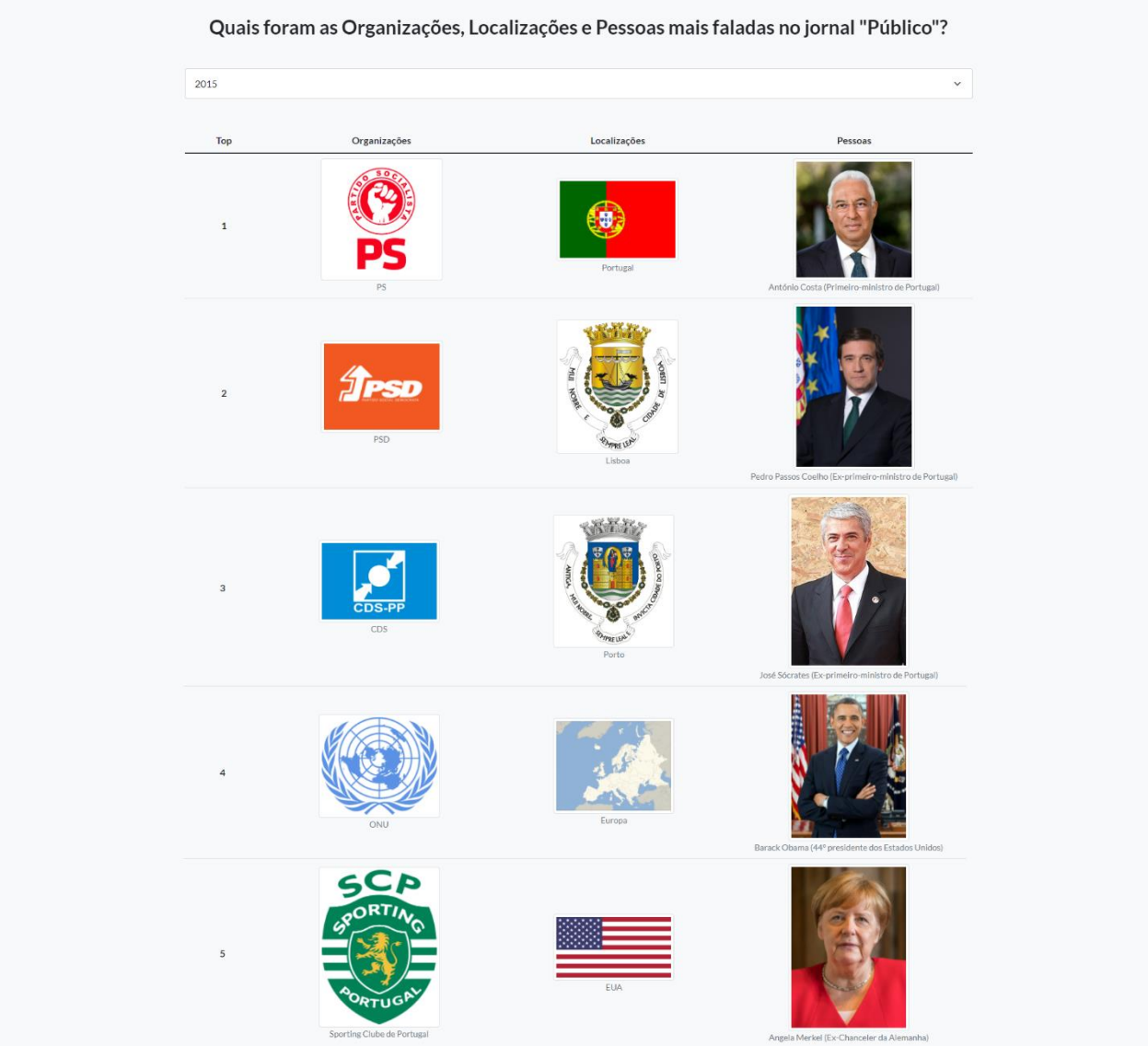

Figura 7 – Top-2 de entidades em 2015.

Com base nas localidades detetadas procedemos a um processo de geocoding que nos permite mapear as localidades. A Figura 8 ilustra esse mapeamento para o ano de 2015. É importante observar que o processo aqui realizado é todo ele automático, da identificação das localidades no texto à extração das coordenadas, pelo que não é possível garantir a sua total efetividade, razão pela qual os dados aqui apresentados devem ser lidos com precaução.

Figura 8 – Top-2 de entidades em 2015.

Os utilizadores têm ainda a possibilidade de observar as keywords mais mencionadas ao longo de cada ano em notícias do jornal Público através de uma Wordcloud novamente com dados obtidos a partir do Arquivo.pt (ver Figura 9):

Figura 9 – Nuvem de palavras no ano de 2015.

Os utilizadores do Arquivo Público podem ainda encontrar uma análise relativa ao impacto que a Covid-19 teve nas notícias do jornal Público, ficando a conhecer a evolução do volume de notícias relacionadas com a pandemia, as organizações, localidades e pessoas mais mencionadas em notícias Covid-19, um mapa interativo com as localidades mencionadas em notícias Covid-19 e uma Wordcloud com as palavras-relevantes mais mencionadas nas notícias. A Figura 10 ilustra uma dessas análises.

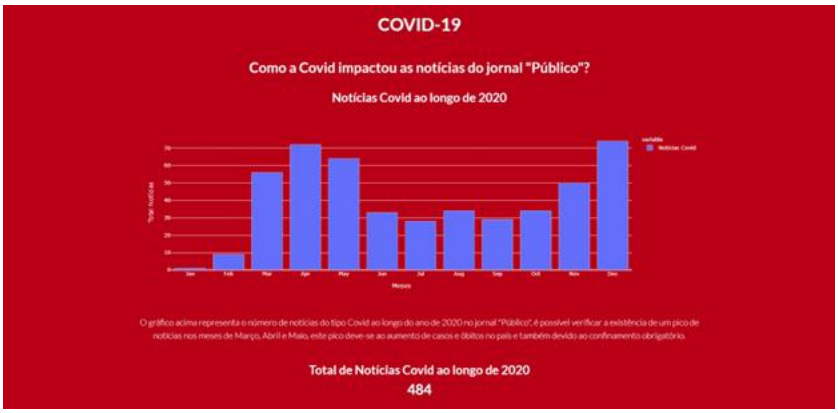

Figura 10 – Notícias acerca do Covid-19 na primeira página do Jornal Público.

Originalidade e caráter inovador

Este projeto, é o primeiro a oferecer aos utilizadores (jornalistas incluídos) a possibilidade de pesquisarem o acervo de notícias do jornal Público, a partir das notícias arquivadas ao longo dos anos pelo Arquivo.pt. O projeto académico aqui apresentado não deve ser encarado como uma substituição da funcionalidade de pesquisa oferecida pelo próprio Jornal Público. Contrariamente a essa funcionalidade, o Arquivo Público é um projeto dedicado única e exclusivamente à pesquisa de informação na homepage do publico.pt a partir de dados coletados pelo Arquivo.pt. A realização deste projeto permitiu-nos desenvolver uma arquitetura que pode agora ser adaptada a outros projetos, permitindo dessa forma que jornais de menor projeção e com poucos recursos financeiros possam oferecer aos seus utilizadores um sistema de pesquisa a partir dos dados coletados pelo Arquivo.pt

Com base nos dados coletados e de acordo com o nosso melhor conhecimento, somos também o primeiro projeto a fazer uma análise mais detalhada das notícias publicadas ao longo dos anos nas dimensões temporais, geográficas e na deteção automática de entidades (uma funcionalidade que não se encontra disponível no jornal Público).

Impacto social (aplicação e utilidade social)

A disponibilização do projeto a partir de um link público permite a qualquer utilizador efetuar pesquisas sobre determinados tópicos que tenham sido objeto de notícia no passado. A existência desta plataforma revela-se assim um importante contributo na tentativa de tornar a informação acessível aos utilizadores, nomeadamente a utilizadores interessados em pesquisar informação de destaque (leia-se publicada na homepage). Adicionalmente, a análise levada a cabo através dos dados coletados ao longo dos anos permite aos utilizadores ter uma noção das principais entidades focadas em notícias do jornal público. Fruto da pandemia vivida optámos por replicar essa análise a partir das notícias com referência ao Covid19. As figuras publicadas no website mostram a evolução da cobertura noticiosa ao longo do período de tempo considerado.

Impacto científico (aplicação e utilidade científica)

No âmbito deste projeto foram aplicadas um conjunto de ferramentas de cariz científico, nomeadamente na construção do motor de busca e na análise de dados efetuada. O projeto aqui apresentado exemplifica a forma como este tipo de ferramentas pode ser usado no contexto da ciência dos dados e mais concretamente na área do jornalismo de dados. O caso dos dados do Covid-19 é, na nossa opinião, um interessante exemplo do que estes dados têm para oferecer para a comunidade em geral e para os jornalistas em particular. Os dados aqui apresentados, permite ainda que os utilizadores da plataforma possam observar a evolução ao longo dos anos, que as notícias do jornal Público têm sofrido, nomeadamente as nível das organizações, localidades e pessoas aí mencionadas.

Relevância da utilização do Arquivo.pt

Para a realização deste trabalho recorremos à API TextSearch do Arquivo.pt para dessa forma obter os dados no período de tempo considerado. A disponibilização e o acesso a estes dados revelaram-se um passo fundamental, sem o qual não teria sido possível concretizar este projeto. Neste projeto, recorremos, para efeitos demonstrativos à utilização dos dados do jornal Público. A arquitetura desenhada permite, no entanto, que novos projetos similares a este possam vir a ser concretizados no futuro, fazendo novamente uso dos dados coletados pelo Arquivo.pt.

Comentários adicionais

Este projeto foi desenvolvido por Diogo Correia e Ricardo Campos e resulta da continuação do trabalho efetuado durante o ano letivo 2021/2022 na UC de Projeto Integrado do Tesp em Informática lecionado no Instituto Politécnico de Tomar (Escola Superior de Tecnologia de Abrantes). A versão inicial deste projeto (realizada no âmbito da referida UC) contou com a colaboração dos alunos Guilherme Oliveira e Diogo Graça e versou uma análise dos dados coletados ao longo dos anos pelo Arquivo.pt do Jornal Mirante (histórico jornal de Santarém).

Agradecimentos: Paulo Crispim (Gab Informática do IPT); Vasco Campos (INESC TEC).

Recursos complementares

- Arquivo Público: http://arquivopublico.ipt.pt
- Vídeo de participação: https://youtu.be/ciMEdMu2s5U