Prémio Arquivo.pt

Descrição Sumária do Trabalho (documento-modelo)

Instruções gerais:

- Apresente de forma concisa os contributos do trabalho de acordo com os critérios de avaliação.
- Responda diretamente às questões colocadas.
- Use linguagem clara e sucinta.
- Pode incluir imagens e respetivas legendas.
- O tamanho máximo do ficheiro é de 10 MB.
- O limite de palavras é de 2 000.
- Submeta a versão final num dos seguintes formatos: DOCX ou PDF.

Identificação

- Título: **Portugal, a que custo?**
- Área temática: **Memória económica · Jornalismo de dados · Humanidades Digitais**
- Candidato: **Bernardo Morais Chagas**
- Email: **bernardo.chagas@tecnico.ulisboa.pt**

Descrição do Trabalho

Descreva sucintamente o trabalho realizado.

O projeto “Portugal, a que custo?” é uma plataforma de memória económica que usa 4.338 artigos jornalísticos preservados pelo Arquivo.pt para responder, com provas primárias verificáveis, à questão mais pessoais da economia portuguesa como, “Quando comecei a trabalhar, era mais fácil ou difícil de viver em Portugal do que hoje?”

O projeto extrai automaticamente dados económicos como os preços de habitação, combustíveis, salários e inflação, analisando os artigos da imprensa portuguesa arquivados entre 1996 e 2026, usando a TextSearch API do Arquivo.pt e extração por modelos de linguagem (LLM/Groq). Com esses dados, calcula um Índice de Dificuldade Económica (IDE) inédito para cada ano, numa escala 0–100.

O resultado é um site interativo com gráfico do IDE 1996–2026, timeline de notícias por ano, calculadora de poder de compra e um questionário didático. Cada valor numérico mostra o título exato do artigo que o sustenta, com link permanente para a página arquivada no Arquivo.pt.

Endereço do projeto: https://web.tecnico.ulisboa.pt/~bernardo.chagas/Portugal_a_que_custo/

Objetivos

Quais os principais objetivos do trabalho?

1. Resposta verificada: Permitir que qualquer cidadão compare a dificuldade económica de diferentes décadas com base em fontes primárias de imprensa, e não em opiniões ou perceções.

2. Demonstração metodológica: Provar que a TextSearch API do Arquivo.pt pode ser usada como fonte primária de extração de dados quantitativos históricos, com pipeline reproduzível e documentado.

3. Dataset académico em acesso aberto: Produzir e disponibilizar uma série histórica de preços (1996–2026) extraídos de fontes jornalísticas arquivadas, com valor independente para investigadores.

Resultados Atingidos

Quais os principais resultados do trabalho?

Pipeline de dados (Fases 1–3):

- 4.338 snippets recolhidos via TextSearch API (29 consultas × 6 domínios de imprensa; 503 ficheiros JSON)
- 1.425 registos extraídos com relevância ≥4/5 — taxa de extração de 33,5%
- Score de confiança médio: 74,9% (mediana 78,5%; intervalo 43,3%–99,4%); calculado por consistência inter-artigos, volume e relevância
- Cobertura temporal: dados reais do Arquivo.pt para 1997–2021; fallback INE/DGEG/Pordata para os restantes anos (assinalado no site)

Índice de Dificuldade Económica (IDE):

- Implementação de índice que permite a comparação direta entre anos
- Índice composto 0–100, pesos: Habitação 35%, Salário 30%, Combustíveis 20%, Inflação 15%

Funcionalidades do site:

- Hero interativo: seletor de ano que calcula o IDE
- Gráfico IDE 1996–2026 com anotações nos momentos históricos
- Timeline: 4 cards por ano (habitação, combustíveis, salário, inflação) com título do artigo sempre visível e link para o Arquivo.pt

- Calculadora de Poder de Compra: narrativa em prosa, cada dado com fonte e link
- Comparador de Anos e Quiz didático com explicação baseada em manchetes arquivadas

Originalidade e caráter inovador

Quais os contributos inovadores que demonstram a originalidade do trabalho?

IDE — métrica inédita: Ao contrário dos índices do INE, o IDE é calculado exclusivamente sobre o discurso jornalístico arquivado, medindo não o que aconteceu, mas o que a imprensa reportou. A divergência entre os dois é ela própria informativa.

Traceabilidade primária total: Cada valor numérico do site tem o título do artigo que o sustenta, o jornal, a data e o link permanente para o Arquivo.pt. Não existe nenhum dado sem fonte verificável.

Pipeline reproduzível em Humanidades Digitais: O fluxo TextSearch API → extração LLM → agregação → frontend estático está inteiramente documentado e publicado, podendo ser replicado para outras temáticas ou arquivos web.

Dicionário histórico de linguagem económica: As consultas foram adaptadas ao vocabulário de cada época: "escudos/contos" antes de 2002, "carestia de vida" antes de 2010, "fogos" antes de 2005. O pipeline converte automaticamente preços em escudos para euros (1€ = 200,482$), criando uma série histórica contínua.

Impacto social (aplicação e utilidade social)

Descreva como o seu trabalho poderá ser útil para os cidadãos?

A questão "era mais difícil antes?" é universal e frequentemente política. O projeto "Portugal, a que custo?" transforma um debate de perceções num exercício de literacia económica baseado em fontes primárias de imprensa preservada.

A Calculadora de Poder de Compra personaliza os dados: quem entrou no mercado de trabalho em 2003 vê, em prosa, o que o seu primeiro salário mínimo representava em rendas e litros de gasolina, comparado com 2026, com a fonte de cada número.

O Quiz didático "Conheces o teu Portugal económico?" revela as respostas corretas com o título real do artigo arquivado, tornando o Arquivo.pt tangível para públicos não académicos e incentivando a sua exploração direta.

Impacto científico (aplicação e utilidade científica)

Descreva como o seu trabalho poderá ser útil para investigadores?

Humanidades Digitais e Web Archiving: Demonstra que corpora de arquivo web podem ser fonte primária de extração de dados quantitativos históricos abordagem ainda pouco explorada para o arquivo da web portuguesa.

Dataset em acesso aberto: Os 1.425 registos extraídos de imprensa portuguesa (1997–2021), com score de confiança e ligação ao artigo original, estão disponíveis no GitHub e podem ser referenciados em trabalhos de investigação sobre economia portuguesa ou história da imprensa.

Jornalismo computacional: A metodologia de extração com LLM sobre snippets de arquivo web, com score de confiança composto (consistência 40% + volume 35% + relevância 25%), constitui uma contribuição metodológica para o campo.

Relevância da utilização do Arquivo.pt

Exponha a necessidade de recorrer ao Arquivo.pt para obter recursos únicos para o trabalho realizado. Apresente exemplos de conteúdos preservados que utilizou.

O projeto "Portugal, a que custo?" é metodologicamente dependente do Arquivo.pt de forma insubstituível. O INE mede o que aconteceu; o Arquivo.pt preserva o que a imprensa reportou. São coisas diferentes, e essa diferença é o objeto do projeto. Todos os valores económicos foram extraídos de artigos jornalísticos preservados pelo Arquivo.pt usando a TextSearch API.

Exemplos de conteúdos preservados utilizados:

- Reportagens do Diário de Notícias sobre o salário mínimo nacional: categoria com maior cobertura (708 registos, média 30,8 fontes/ano, presentes em 23 dos 29 anos cobertos).
- Notícias sobre preços de combustíveis que confirmaram o pico histórico de 2005 (gasolina 1,044 €/L), tornando 2005 o ano mais difícil do IDE.
- Artigos sobre "carestia de vida" (termo da época), capturados por consultas historicamente adaptadas ao vocabulário jornalístico de cada período.
- Artigos do Público (1996–2021) sobre preços de habitação em Lisboa, com valores em escudos/contos antes de 2002, convertidos automaticamente para euros.
- Reportagens sobre a transição escudo → euro (2001–2003), integrando a primeira série histórica contínua de preços em euros para o período escudo.

A TextSearch API foi usada com 29 consultas, 4 categorias económicas, 6 domínios de imprensa e intervalos temporais adaptados à evolução da linguagem jornalística. O rate limit de 250 requests em 180 segundos foi escrupulosamente respeitado.

Comentários adicionais

Inclua informação adicional que considere relevante para demonstrar o valor do trabalho.

Transparência do score de confiança: O site exibe, para cada ano, o indicador de fiabilidade (0–100%). Os anos 1996 e 2022 a 2026, sem registos extraídos do Arquivo.pt, estão assinalados como baseados em dados de fallback INE/DGEG/Pordata.

Custo zero e metodologia reproduzível: Todo o pipeline usa recursos gratuitos (TextSearch API, Groq free tier, React + deploy estático). Os scripts estão publicados no GitHub, permitindo a qualquer investigador replicar ou adaptar o processo.

Recursos complementares

Liste referências disponíveis online que complementem o trabalho. Por exemplo, podem ser ligações para documentos, websites, imagens ou aplicações.

Apenas serão considerados recursos disponíveis em acesso aberto através da Internet. Utilize o seguinte formato para listar as referências:

- Custo Portugal, https://web.tecnico.ulisboa.pt/~bernardo.chagas/Portugal_a_que_custo/ : plataforma interativa com IDE 1996–2026, Timeline, Calculadora e Quiz
- Repositório GitHub, https://github.com/BernardoMChagas/Custo-Portugal-Arquivo.pt : código-fonte, scripts e dataset em acesso aberto (CC BY 4.0)
- Dataset final_for_frontend.json, https://github.com/BernardoMChagas/Custo-Portugal-Arquivo.pt/tree/main/data : 1.425 registos de preços extraídos de imprensa portuguesa arquivada (1997–2021)
- Arquivo.pt TextSearch API, https://arquivo.pt/textsearch: API utilizada para recolha dos 4.338 snippets
- Pordata, https://www.pordata.pt : fonte dos dados de fallback (salário mínimo)
- INE, https://www.ine.pt : fonte dos dados de fallback (inflação média anual e preço habitação)
- DGEG, https://www.dgeg.gov.pt : fonte dos dados de fallback (preços históricos de combustíveis)