# Identificação

- Título: **Politiquices**
- Área temática: investigação jornalística; processamento de linguagem natural; web semântica; análise de redes/grafos;
- Candidato: David Soares Batista
- Email: dsbatista@gmail.com

# Descrição do Trabalho

O Politiquices (www.politiquices.pt) é uma ferramenta que dá a possibilidade de explorar e analisar as relações de **apoio** e **oposição,** entre personalidades políticas, expressas em títulos de notícias preservadas no Arquivo.pt, como exemplificado na Figura 1.

Além de identificar as relações de oposição/apoio, também associa as personalidades envolvidas ao seu identificador na Wikidata[1]. Deste modo, a informação associada à personalidade (e.g.: filiação política, cargos políticos, etc.) fica a fazer parte implícita da relação e pode ser usada na análise das relações extraídas.

Este processo resulta numa rede de relações de apoio e oposição entre personalidades, e implicitamente dos seus partidos políticos, cobrindo um período de cerca de 24 anos. É possível fazer perguntas a esta rede, para qualquer período de tempo, como:

- Que personalidades do BE se opuseram a Jerónimo de Sousa?
- Que acusações fez Passos Coelho a José Sócrates?
- Quem de dentro do PS se opôs/apoiou a José Sócrates?
- Que personalidades do PSD foram apoiadas por Assunção Cristas?
- Que personalidades afiliadas ao BE se opuseram a personalidades do PCP?
- Que personalidades do PS apoiaram personalidades do PSD?
- Que políticos do BE estiveram envolvidos em conflitos internos ao próprio partido?


Figura 1: Exemplos de títulos de notícias com relações explícitas de oposição e apoio.

[1] https://www.wikidata.org/wiki/Q182367 - exemplo da página na Wikidata de José Sócrates

Obtendo como resposta uma lista notícias preservadas no Arquivo.pt que evidenciam as relações de apoio e oposição entre personalidades e/ou partidos, podendo-se depois consultá-las ou descarregar a lista de notícias num formato estruturado.

Também é possível explorar as relações através de um grafo interactivo, podendo aplicar diferentes filtros de forma a seleccionar as personalidades e as relações visíveis no grafo. Os nós no grafo representam personalidades e os arcos notícias que dão suporte às relações entre personalidades. Através dos filtros pode-se definir:

- o tipo de relações a visualizar
- o intervalo de tempo associado às notícias
- o número de notícias necessárias para que exista uma ligação entre dois nós

Podendo depois navegar-se no grafo, clicando nos arcos para aceder às notícias e nos nós para ver o perfil individual de uma personalidade, como representado na Figura 2.

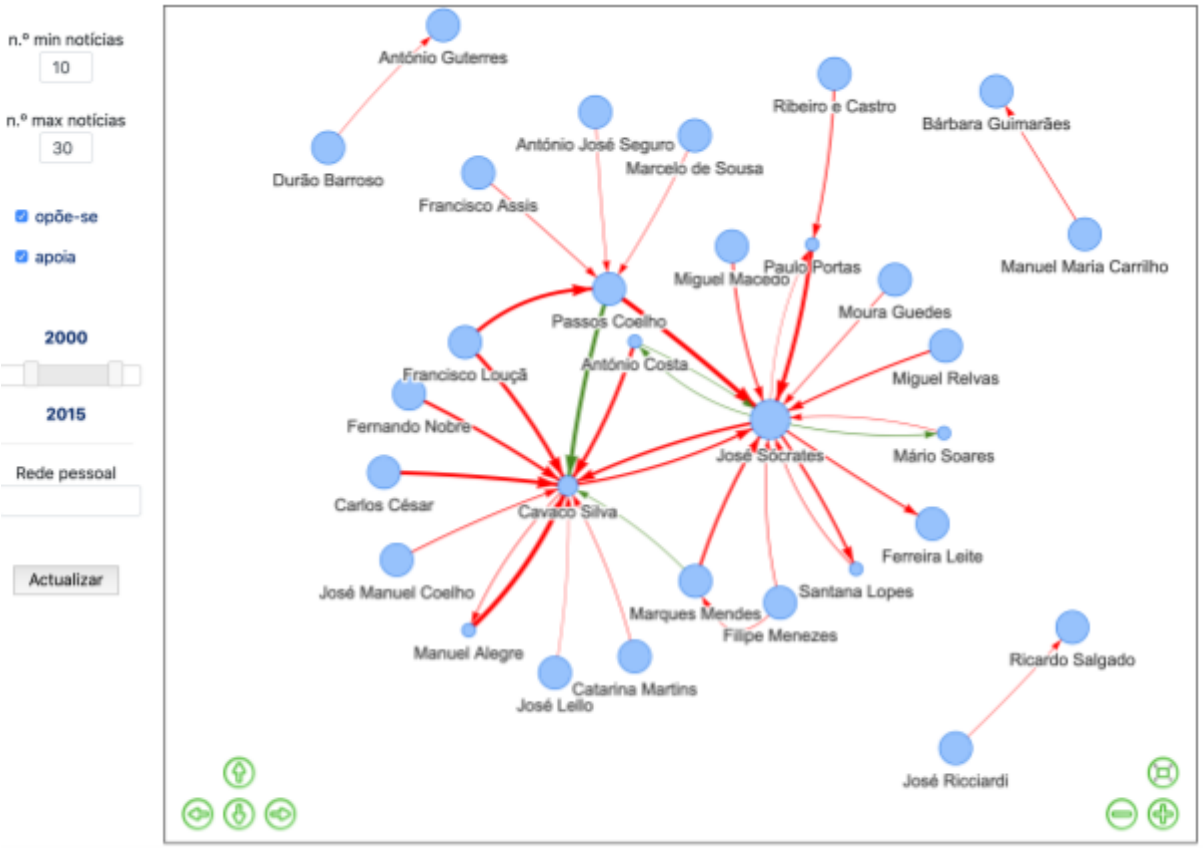


Figura 2: Exploração da rede relações através de um grafo interactivo.

## Objetivos

O Politiquices tem como objectivo principal facilitar a exploração de relações de apoio ou oposição entre personalidades políticas expressas em notícias portuguesas, podendo ser usado como ferramenta de apoio ao jornalismo de investigação.

Utilizando o Arquivo.pt como fonte de notícias expressando relações políticas e a Wikidata como a base de conhecimento que descreve de forma estruturada as personalidades e partidos políticos, o objectivo foi o de construir um base dados sob a forma de grafo que permitisse investigar as relações de apoio e oposição na política.

O segundo objetivo foi disponibilizar acesso à base de dados permitindo explorá-la de diferentes formas: fazendo interrogações, navegando num grafo, ou analisando as relações de cada personalidade individualmente.

## Resultados Atingidos

O principal resultado deste trabalho é a base de dados contendo relações entre personalidades políticas, interligada com notícias preservadas no Arquivo.pt e com referências das personalidades envolvidas à Wikidata. É representada sob a forma de uma rede ou grafo e publicada de forma livre no formato aberto *Resource Description Framework* (RDF)[2], permitindo a sua integração noutras aplicações. Na sua versão mais recente, contém 592 personalidades políticas em cerca de 7.200 notícias distribuídas por um período de 24 anos. É possível descarregar a base de dados[3] e fazer uso dos seus dados localmente ou noutras aplicações.

A interface desenvolvida, é também um resultado deste trabalho, que permite explorar e interagir com os dados de diferentes formas, permitindo selecionar as relações, personalidades, períodos de tempo, e exportar os resultados para um formato estruturado.

Um terceiro resultado, são os vários componentes de *software* disponibilizados livremente[4] que permitem gerar a base de dados de relações:

- Um modelo de *machine learning* para detectar relações de suporte/oposição nas notícias recolhidas do Arquivo.pt bem como os dados anotados para treino
- Detecção de personalidades em títulos de notícias e ligação com a Wikidata
- Comunicação com o *endpoint* da Wikidata de forma a recolher informações de personalidades políticas relevantes e partidos políticos
- Recolha e conversão de notícias do Arquivo.pt para um formato de texto limpo que permitiu ser processado por algoritmos de *machine learning*

[2] https://www.w3.org/RDF/
[3] https://www.politiquices.pt/about
[4] https://github.com/davidsbatista/politiquices

## Originalidade e caráter inovador

No contexto de análise de notícias de política em Português, e tanto quanto é do meu conhecimento, julgo este ser um dos primeiros trabalhos a analisar interações de apoio e oposição entre personalidades políticas expressas nas notícias, recorrendo a técnicas de processamento de linguagem natural e análise de redes.

No âmbito de projectos associados ao Arquivo.pt existem trabalhos realizados com algumas semelhanças, no sentido em que também exploram o texto das notícias contidas no Arquivo.pt, e.g.: "*Conta-me Histórias*", "*Desarquivo*", "*Arquivo de Opinião*".

No entanto, o Politiquices destaca-se em três aspectos:

- foco num domínio concreto dentro dos vários temas existentes nas notícias preservadas no Arquivo.pt
- integração dos resultados com uma base de conhecimento aberta
- disponibilização dos resultados num formato aberto e *standard*

Os projetos referidos acima retornam resultados do Arquivo.pt de uma forma densa e cobrindo todo o espectro de notícias, sendo que o utilizador posteriormente tem que filtrar os resultados para encontrar o que procura. O Politiquices permite questionar diretamente o Arquivo.pt, dentro de um domínio concreto. Tanto quanto é do meu conhecimento isto é único no âmbito de projectos relacionados com o Arquivo.pt.

A interligação dos dados analisados com a Wikidata permite lidar com o problema de ambiguidade de nomes. Por exemplo, o nome "Jardim" poderá ser uma referência a Vera Jardim, Alberto João Jardim, ou Jardim Gonçalves. Ao analisar o título e quando necessário o corpo da notícia e recorrendo à Wikidata para desambiguação o nome "Jardim" presente num título é sempre desambiguado para uma personalidade única. Além de facilitar o processo de desambiguação, permite também enriquecer as relações com mais dados associados à personalidade política, e.g.: filiação política, cargos políticos, legislaturas em que participou, relações familiares, etc.

A base de dados poderá ser descarregada e consultada localmente[5]. A interligação com os dados da Wikidata permite fazer interrogações mais específicas do que aquelas disponibilizadas na interface, por exemplo:

- Existem personalidades relacionadas familiarmente e envolvidas numa relação de oposição/suporte? (i.e., através as relações familiares associadas a cada personalidade presentes na Wikidata)

- Que personalidades estão envolvidas em conflitos existentes dentro de um mesmo Governo? (i.e., através da lista das legislaturas para cada personalidade, informação também disponível na Wikidata)

[5] https://www.politiquices.pt/about

## Impacto social (aplicação e utilidade social)

O Politiquices poderá ser utilizado como uma ferramenta de suporte à investigação jornalística. Um jornalista pode rapidamente reunir uma colecção de notícias contendo interacções de apoio ou oposição envolvendo personalidades e partidos políticos.

Poderá também ser usado para examinar ou detectar padrões. Explorando a linha de tempo de relações de uma personalidade é possível procurar por padrões onde personalidade emerge como alvo ou origem de relações de apoio e oposição, como exemplificado na Figura 3.

Um caso de uso interessante será de comparar as relações de apoio ou oposição antes de uma personalidade política tomar posse de determinado cargo público com as relações depois de a personalidade ter assumido o cargo, ou ver que relações de apoio subitamente emergiram.


Figura 3. Exemplo de uma linha de tempo das relações de uma personalidade.

A existência de uma ferramenta que mantém e disponibiliza o acesso a um repositório de interacções entre personalidades políticas permite, não só fazer o seu acompanhamento, como também analisar em retrospectiva essas mesmo interacções dentro de determinados contextos.

## Impacto científico (aplicação e utilidade científica)

Para a comunidade científica há dois resultados importantes: os dados anotados e o grafo de relações.

Os dados anotados permitem que novos algoritmos sejam treinados para detectar relações e ligar as personalidades com a Wikidata, e servem de incentivo ao desenvolvimento de melhores algoritmos de processamento de linguagem natural no contexto de notícias de âmbito político em Português.

O grafo de relações permite a cientistas de áreas como: sociologia, ciência política e computação realizar diversos estudos, por exemplo:

- Encontrar comunidades de apoio e oposição em função do tempo e verificar se há mudanças entre alianças e oposições

- Enriquecer as relações, categorizando-as num ou mais tópicos com base numa análise detalhada do texto da notícia

- Estudar os triângulos políticos: se as personalidades políticas X e Y sempre acusam ou defendem uma terceira personalidade Z, qual será a relação típica entre X e Y ?

Pretendo em breve submeter um artigo científico descrevendo os resultados deste trabalho, e uma descrição dos dados anotados usados para treinar os algoritmos, e dos primeiros resultados alcançados. Paralelamente conto continuar a trabalhar e melhorar os algoritmos que fazem a ligação das personalidades com a Wikidata e a detecção de relações.

Julgo que os resultados do Politiquices representam um complemento importante nos recursos já existentes para o processamento de linguagem natural em Português, mas também um ponto de partida no desafio de fazer análise de sentimento entre personalidades políticas expressas em notícias.

## Relevância da utilização do Arquivo.pt

Grande parte das interações entre personalidades políticas são descritas nas notícias publicadas em diferentes meios de comunicação. O acesso a uma colecção de dados com um espectro temporal alargado de notícias é fundamental para se poder fazer uma análise deste tipo de interacções.

O Arquivo.pt permitiu ao Politiquices ter acesso a notícias de relevância política provenientes de diferentes fontes de informação (e.g., jornais *on-line*, *websites* de canais de TV, rádio e portais agregadores de conteúdos) e para um período de tempo de cerca de 24 anos.

Todos os artigos foram recolhidos usando as interfaces de programação (API) de acesso ao arquivo. Nomeadamente a interface de pesquisa[6] para aceder a resultados, e a interface *CDX server*[7] que permitiu obter os limites da dimensão temporal das recolhas feitas para cada domínio.

A API de pesquisa foi usada restringindo os resultados a ocorrências de nomes de personalidades políticas e tendo como alvo 45 domínios selecionados[8] para um período de tempo variando com o domínio de pesquisa, mas nos extremos entre 1996 e 2019.

## Comentários adicionais

Para fazer parte da base de dados do Politiquices uma notícia necessita de preencher os seguintes requisitos:

- mencionar duas personalidades políticas no título
- ambas as personalidades terem uma página na Wikidata
- O título conter uma relação de oposição/apoio entre essas duas personalidades

O processo de selecção de notícias do Arquivo.pt é automatizado e depende de vários componentes de *software* que recorrem a algoritmos de *machine learning* para detectar os requisitos descritos acima, nomeadamente:

- identificar referências a personalidades políticas expressas nos títulos
- associar as personalidades à sua página na Wikidata
- detectar a relação de oposição ou apoio e também a sua direcção (e.g: 'X é apoiado por Y' é diferente de 'X apoia Y')

Naturalmente estes algoritmos nem sempre dão os resultados correctos. Ajustei os parâmetros dos algoritmos de forma a sacrificar o número de resultados em favor de resultados correctos, ou seja, a ter poucas notícias/relações, mas potencialmente corretas, ao invés de um grande número de notícias.

---

[6] https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API
[7] https://github.com/arquivo/pwa-technologies/wiki/URL-search:-CDX-server-API
[8] https://github.com/davidsbatista/politiquices/blob/master/politiquices/nlp/data_sources/arquivo_pt/config_data/domains.txt

É possível que existam alguns erros na classificação das relações. No entanto, investigando é possível melhorar estes algoritmos. Conto lançar futuramente versões atualizadas do grafo à medida que optimizo e melhoro os algoritmos, tendo também em consideração a existência de mais dados no Arquivo.pt.

Todos os componentes de *software* e dados anotados são disponibilizados livremente de forma a poderem ser replicados os resultados. Mas também de poder dar, a potenciais interessados, a oportunidade de contribuir para o projecto ou usar o código para outros projectos que envolvam extrair informação estruturada do texto nas notícias no Arquivo.pt.

De forma a completar o grafo foram também usados artigos de notícias de outras fontes. Nomeadamente, da colecção CHAVE - disponibilizada pela Linguateca - que contém artigos do jornal PÚBLICO publicados em 1995 e 1996, e também alguns artigos recolhidos diretamente a partir do site do publico.pt que não fazem parte do Arquivo.pt.

## Recursos complementares

- *"Wikidata: a free collaborative knowledgebase"*, wikidata.org: uma base de conhecimento aberta, colaborativa e estruturada permitindo a utilização e integração dos seus dados noutras aplicações;

- "*Colecção CHAVE*", https://www.linguateca.pt/CHAVE/: textos completos das edições do jornal PÚBLICO publicados nos anos de 1994 e 1995;