Prémio Arquivo.pt

Identificação

- Título: narrArquivo – narrativas a partir da web
- Área temática: Recuperação e Extração de Informação; Criação de narrativas ao longo do tempo
- Candidato: Ricardo Campos, Jorge Duque, Alípio Jorge, Gaël Dias, Célia Nunes
- Email: ricardo.campos@ipt.pt
- Link: http://narrarquivo.inesctec.pt/

Descrição do Trabalho

Nos últimos anos, a quantidade de informação gerada, consumida e armazenada cresceu a uma velocidade vertiginosa dificultando a tarefa dos que procuram informação e a partir dela extrair conhecimento em tempo útil. No domínio da web, os arquivos funcionam como um importante recurso de preservação que permite aceder à informação do passado. Extrair conhecimento a partir destas plataformas, é, no entanto, um processo pouco atrativo para um considerável número de utilizadores habituados a lidar com a versão mais recente da web. Embora o processo de pesquisa e exploração dos resultados tenha melhorado significativamente nos últimos anos, com o desenvolvimento e a implementação de novos algoritmos no domínio da recuperação de informação e visualização dos dados [1], o problema de construir e apresentar estruturas narrativas consistentes no domínio dos Arquivos web encontra-se por resolver [2]. Nestas plataformas, o conteúdo textual é, ainda, e fruto da sua antiguidade, a principal forma de apresentar a informação. A Figura 1 ilustra um destes textos. A notícia, preservada pelo Arquivo.pt na data de 07-01-2017, oferece ao utilizador uma narrativa meramente textual sobre a morte de Mário Soares, antigo Presidente da República.

Figura 1: Notícia preservada pelo Arquivo.pt acerca da morte de Mário Soares ocorrida a 07-01-2017. Origem: https://arquivo.pt/wayback/20170107192028/https://www.dinheirovivo.pt/outras/morreu-mario-soares/

Os anos mais recentes, mostram, no entanto, uma clara tendência, particularmente nas gerações mais novas, para o consumo de informação a partir de diferentes formatos. Impulsionados por este novo paradigma, diferentes *stakeholders*, têm feito um esforço na tentativa de adaptar os seus conteúdos aos hábitos de consumo de um público cada vez mais digital. Meios de comunicação social, como é o caso da BBC no Reino Unido ou do Jornal Público em Portugal, têm estado na linha da frente na apresentação de informação a partir de ferramentas que extraem elementos narrativos e os apresentam aos leitores a partir de formatos mais apelativos, mantendo ao mesmo tempo o essencial da história. As figuras 1 e 2 retratam dois exemplos deste novo paradigma centrado no digital e na sumarização da informação a partir

de estruturas gráficas capazes de enfatizar e resumir os atores principais da história, a sua interação e trajetória ao longo do tempo e do espaço.

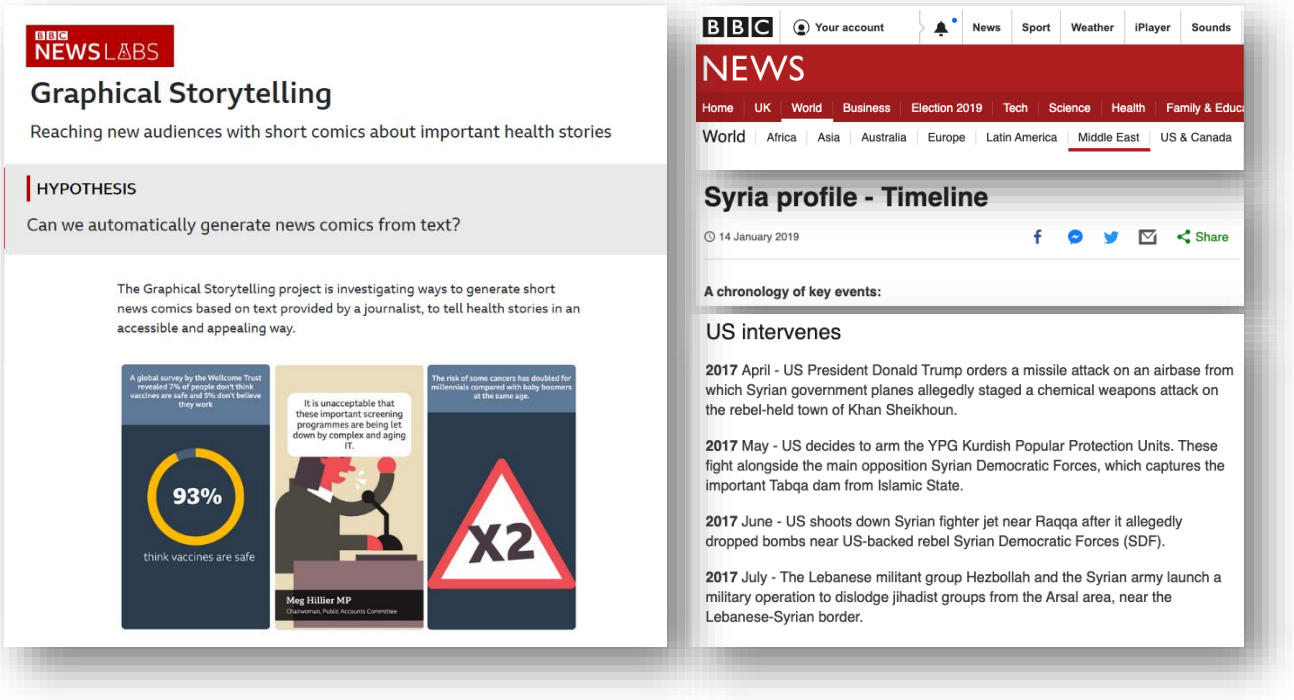

Figura 2: Graphical Storytelling (BBC News Labs) e Syria Timeline (BBC website)

Neste contexto, a representação de textos a partir de timelines (linhas do tempo) surge como alternativa à apresentação de dados feita unicamente a partir de estruturas textuais, oferecendo aos utilizadores a possibilidade de se familiarizarem com um determinado evento num curto espaço de tempo. Apesar da crescente importância das timelines no âmbito da sumarização de dados (ver Figura 3) a partir de vários documentos, pouco se sabe sobre o seu uso e aplicação no contexto de documentos individuais. Por outro lado, o imenso volume de dados presente nos arquivos web, torna proibitiva a construção manual e a disponibilização deste tipo de interface.

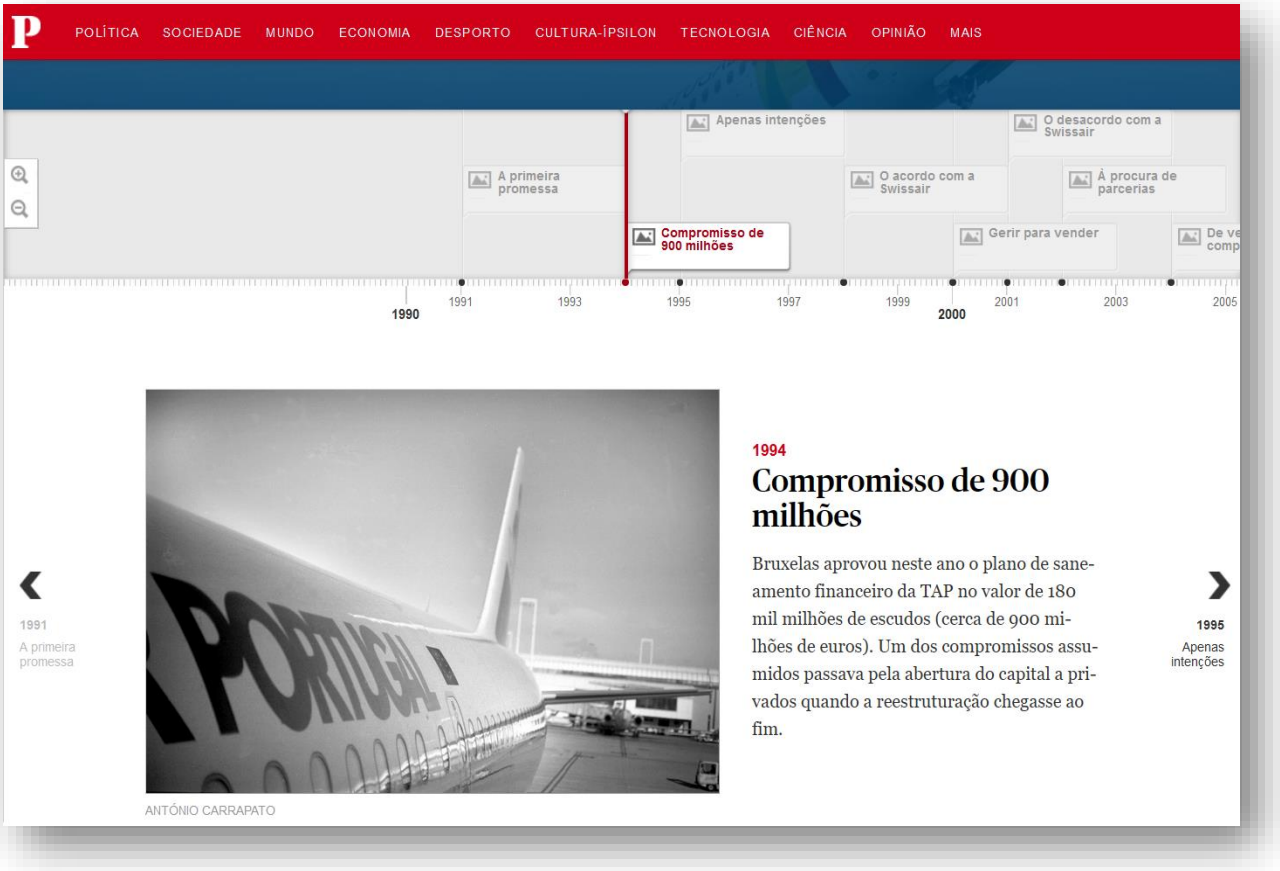

Figura 3: Timeline da privatização da TAP. Jornal Público: https://acervo.publico.pt/economia/interactivo/tap-cronologia-de-uma-privatizacao#0

Neste projeto, apresentamos o *narrArquivo* (http://narrarquivo.inesctec.pt), um website que oferece aos utilizadores a possibilidade de criarem automaticamente narrativas visuais a partir de arquivos da web e da identificação das expressões temporais mais importantes de um texto. Com vista à concretização deste objetivo fazemos uso do (1) *Newspaper 3K*[1], uma biblioteca *python* capaz de extrair conteúdos textuais a partir da grande maioria de *URLs*, do (2) *Heideltime* [4], um algoritmo, estado da arte na identificação de datas, e do (3) *Time-Matters* [3], uma medida de similaridade temporal desenvolvida pela nossa equipa (e anteriormente aplicada no contexto de um conjunto de documentos), para desta feita, detetar, extrair e estimar a importância de

[1] https://newspaper.readthedocs.io

datas no âmbito de um único texto. A partir da página inicial do *narrArquivo* (ver

Figura 4), o utilizador da aplicação é convidado a selecionar uma das "nossas narrativas", um conjunto de textos selecionados a partir de notícias preservadas pelo Arquivo.pt.

Figura 4: Página Inicial do narrArquivo

Ao utilizador é oferecida também a possibilidade de especificar um endereço web[2] ou mesmo introduzir um texto (conforme teremos oportunidade de observar mais à frente). Para exemplificarmos o processo de construção de narrativa tomamos como base o endereço web[3] da notícia do website *dinheiro vivo* que dá conta da morte do antigo Presidente da República Mário Soares. A Figura 5 ilustra o processo de criação da narrativa visual e chama a atenção para as diferenças existentes entre esta página e a página web coletada em 2017 pelo Arquivo.pt (ver Figura 1). O "texto anotado", componente do *narrArquivo*, mostra uma escala de cores onde cada data é etiquetada de acordo com a sua importância no texto (com os scores a variar entre 0 e 1). Por defeito apenas as datas determinadas como mais relevantes pelo nosso algoritmo são mostradas ao utilizador, um processo particularmente útil em textos longos onde a existência de datas menos relevantes é propensa de acontecer. Observe-se também que expressões temporais relativas, como é o caso de "este sábado" ou mesmo "92 anos" são automaticamente etiquetadas pelo sistema. Adicionalmente o algoritmo salienta no texto as palavras relevantes que contribuem para o cálculo da importância das datas. Como regra base, uma data será tanto ou mais relevante quanto mais próxima estiver de palavras-chave importantes. Para determinar estas palavras, fazemos uso do *YAKE*[4] [5] um extrator de palavras relevantes desenvolvido pela nossa equipa.

Figura 5: Texto anotado

[2] a partir de notícias preservadas pelo Arquivo.pt ou através de notícias disponíveis na web atual
[3] https://arquivo.pt/wayback/20170107192028/https://www.dinheirovivo.pt/outras/morreu-mario-soares/
[4] http://yake.inesctec.pt

Em adição ao texto anotado, o utilizador tem a possibilidade de recorrer ao componente *storyline* (ver Figura 6) para ter uma visão geral da narrativa da estória e navegar entre os diferentes períodos temporais automaticamente ancorados na linha do tempo. Textos que contenham datas relevantes serão mostrados ao utilizador junto com um título extraído automaticamente com recurso ao *YAKE*. Esse mesmo título (na figura abaixo, "*Comunidade Económica Europeia*") servirá de base para obter a imagem associada. Com vista a esse objetivo recorremos à *API* de imagens[5] do Arquivo.pt recentemente lançada por esta infraestrutura.

Figura 6: Storyline

[5] https://github.com/arquivo/pwa-technologies/wiki/ImageSearch-API-v1.1-(beta)

Um outro aspeto importante, prende-se com a capacidade do sistema em associar datas relativas (e.g., "*este sábado*") a datas explícitas (e.g., "*07-01-2017*") não necessariamente referidas no texto. A Figura 7 ilustra um desses exemplos. No texto é possível observar que a expressão temporal "*este sábado*" é automaticamente associada à data de publicação do documento[6] ("*07-01-2017*").

Figura 7: Storyline. Normalização de datas relativas.

Os restantes dois componentes da narrativa visual oferecem ao utilizador uma visão agrupada das datas (componente agrupamento temporal) e uma nuvem de palavras (ver Figura 8) obtida com recurso às palavras-chave mais importantes automaticamente determinadas pelo *YAKE* a partir do texto base.

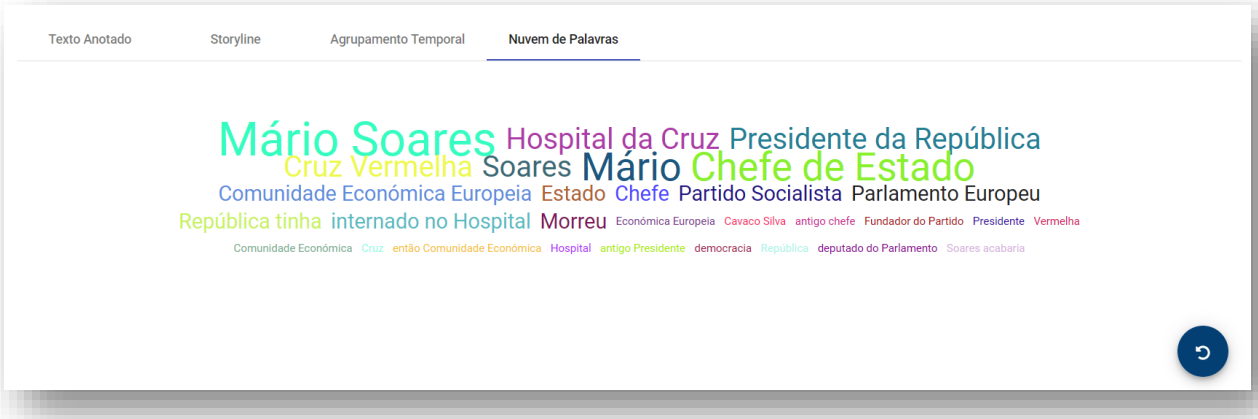

Figura 8: Nuvem de palavras gerada a partir de notícia preservada pelo Arquivo.pt

[6] inferida a partir do URL ou explicitamente definida pelo utilizador a partir das opções avançadas da pesquisa.

Em alternativa à especificação de um *URL* ou às "*nossas narrativas*", o utilizador poderá também proceder à introdução (*copy*/*paste*) de um texto, sempre e quando este inclua, naturalmente, alguma forma de evidência temporal. A Figura 9 ilustra essa possibilidade e chama a atenção para o facto de o sistema poder receber textos ou endereços web noutros idiomas que não o português e ainda assim continuar a fazer uso do acervo de imagens do Arquivo.pt para o processo de construção da narrativa visual.

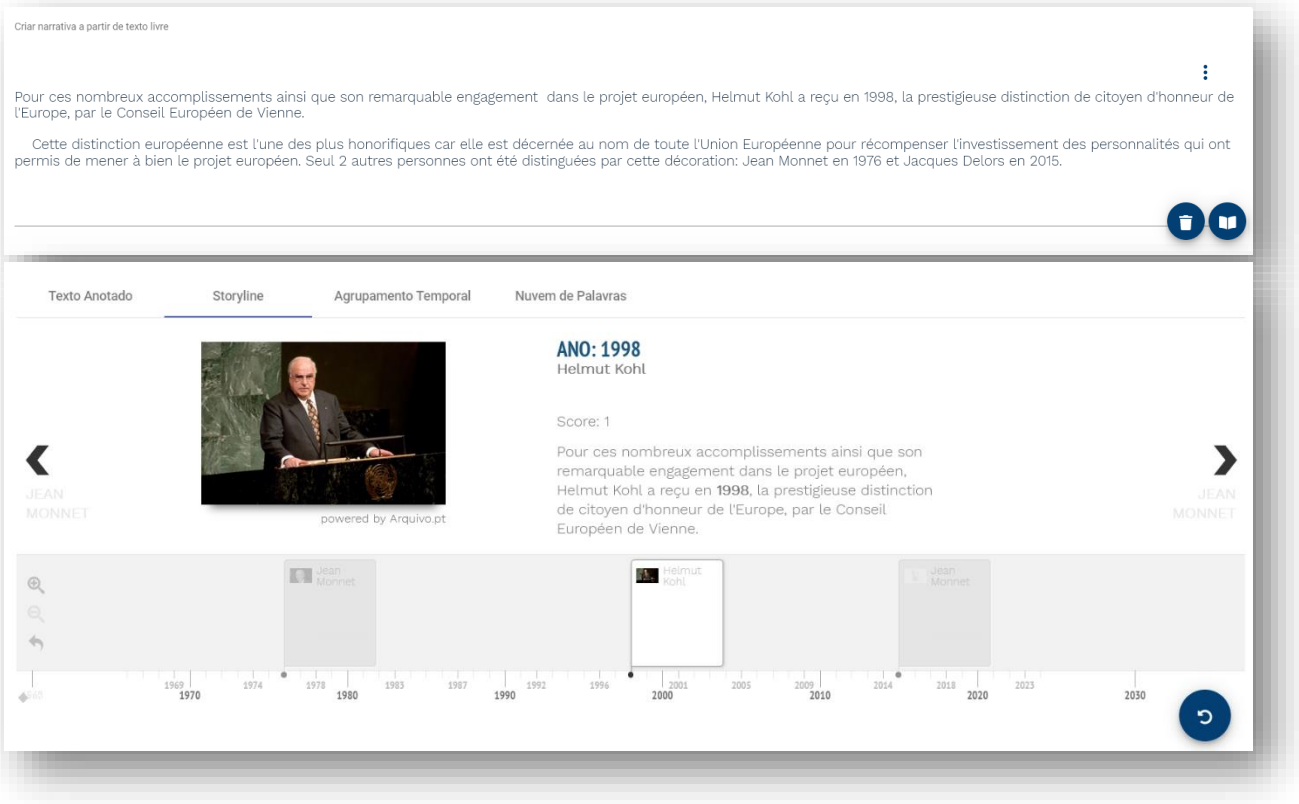

Figura 9: Componente introdução de texto livre na língua francesa.

Objetivos

Neste projeto pretendemos aplicar o algoritmo *Time-Matters* na identificação e estimação da importância de expressões temporais a partir de textos individuais, bem como propor uma alternativa à disponibilização de estruturas meramente textuais, oferecendo ao utilizador a possibilidade de criar narrativas visuais automáticas, com foco temporal e recurso ao uso de textos e imagens preservadas pelo Arquivo.pt.

Resultados Atingidos

Em concreto, (1) disponibilizamos online um website (responsivo, i.e., adaptável a PCs e smartphones) que faz uso do acervo de documentos e imagens do Arquivo.pt para oferecer ao utilizador a possibilidade de criar narrativas visuais; (2) propomos uma solução que, assente no *Time-Matters* e no *YAKE,* é facilmente adaptável a diferentes coleções, idiomas e contextos, por não fazer uso de coleções externas ou qualquer tipo de treino de dados; (3) finalmente,

disponibilizamos online[7] o código do nosso algoritmo, abrindo portas a que outros investigadores e empresas dele possam fazer uso no contexto de futuros projetos.

Originalidade e caráter inovador

Embora o processo de sumarização temporal seja relativamente comum a partir de múltiplos documentos, pouco se sabe sobre a sua aplicação em documentos únicos e a sua utilização na criação de narrativas visuais. Neste projeto, propomos a criação automática de narrativas visuais a partir de documentos coletados pelo Arquivo.pt. A aplicação deste algoritmo no contexto dos arquivos web é, na nossa perspetiva, um importante contributo na tentativa de manter estas infraestruturas apelativas para o utilizador comum e, ao mesmo tempo, tratar de as aproximar das tendências mais atuais no que diz respeito aos novos formatos de consumo de informação. De acordo com o nosso melhor conhecimento, tal, não foi ainda feito no passado. A automatização deste processo gera naturalmente erros, mas também desafios interessantes aos quais apenas o futuro poderá dar resposta, quer na necessidade de melhorar a deteção de eventos e a sua relação com os aspetos temporais, quer na necessidade de melhorar a efetividade do sistema de pesquisa de imagens. Conduzir este processo de forma manual é, no entanto, e dado o elevado volume de dados aí presente, uma tarefa impossível no contexto dos arquivos da web. Com a disponibilização do *narrArquivo* online pretendemos contribuir para a resolução deste problema. A eventual integração direta deste componente no site do Arquivo.pt ou em qualquer outro website é possível, no futuro, a partir da simples disponibilização de um botão através do qual o utilizador poderá criar a respetiva narrativa.

Impacto social (aplicação e utilidade social)

Numa era em que o volume de dados é enorme (e com tendência para crescer) a disponibilização deste tipo de ferramentas é um importante contributo para que os utilizadores comuns possam beneficiar do acesso a narrativas geradas automaticamente em cima de coleções que, pelo seu volume e antiguidade, tornariam praticamente impossível uma geração manual. A utilização de elementos gráficos como complemento à informação textual, surge enquadrada na recente tendência de disponibilizar aos leitores diferentes formatos para consumo de informação por forma a que estes possam rapidamente ficar familiarizados com o tópico. Exemplos de casos de uso são as narrativas existentes na página inicial do *narrArquivo*. Num espectro mais alargado, este tipo de aplicações pode também ser útil a todos os que pretendam extrair informação estruturada (nomeadamente temporal) a partir de documentos da web atual (por via da disponibilização de um qualquer *URL* ou texto livre). Artigos financeiros, clínicos, noticiosos[8] ou documentos longos, como é o caso da *wikipedia*[9], são alguns dos principais candidatos a poderem beneficiar deste tipo de sumarização temporal.

[7] https://github.com/liaad/time-matters
[8] https://www.publico.pt/2021/02/18/ciencia/noticia/tres-missoes-chegaram-planeta-vermelho-mes-1951138
[9] https://pt.wikipedia.org/wiki/Caso_BPN

Impacto científico (aplicação e utilidade científica)

O *narrArquivo* é um projeto que resulta de uma linha de investigação denominada *Text2Story*[10] e que decorre no âmbito de um projeto Portugal 2020. Dois dos autores do *narrArquivo*, Ricardo Campos e Alípio Jorge lideram atualmente o projeto em curso. Para o desenvolvimento do *narrArquivo* recorremos a dois algoritmos desenvolvidos pela nossa equipa, o *YAKE* [5] e o *Time-Matters* [3], o último dos quais anteriormente utilizado na identificação de expressões temporais relevantes a partir de múltiplos documentos. Ao invés, neste projeto, focamo-nos em usá-lo no caso particular da identificação de instâncias temporais a partir de textos individuais e na criação de narrativas visuais a partir de textos coletados pelo Arquivo.pt. Paralelamente à disponibilização do website, tornamos público o código-fonte do algoritmo de suporte a esta aplicação, um importante contributo que visa proporcionar aos interessados uma base para futuros desenvolvimentos. O trabalho aqui apresentado levanta uma série de novos desafios que procuraremos vir a responder no futuro, nomeadamente na identificação de relações entre eventos e expressões temporais com vista a melhorar a efetividade do sistema. No âmbito do projeto em curso iniciámos recentemente a execução de uma bolsa de doutoramento que visa trabalhar este problema.

Relevância da utilização do Arquivo.pt

Neste projeto propomos a criação e a visualização de narrativas a partir de textos selecionados do Arquivo.pt. Os exemplos disponibilizados (ver

Figura 4) visam demonstrar a utilidade deste tipo de aplicações em documentos coletados no passado e aproximá-los, tanto quanto possível, dos hábitos de consumo atuais, mantendo, ao

[10] http://text2story.inesctec.pt

mesmo tempo, a atratividade dos arquivos web, sobretudo junto de uma geração habituada a consumir dados a partir de diferentes formatos. Para a criação do componente de *storyline* recorremos à *API* de imagens do Arquivo.pt. Em concreto, cada instância temporal relevante detetada pelo nosso algoritmo, é ilustrada com recurso a um acervo de 1.800 milhões de imagens. O uso deste componente é dos pontos mais importantes deste projeto.

Recursos complementares

[1] Time-Matters: Temporal Unfolding of Texts
[2] Exploding TV Sets & Disappointing Laptops: Archival Content Suggestion by Finding Interesting Content from the Past
[3] Identifying Top Relevant Dates for Implicit Time Sensitive Queries
[4] Multilingual and cross-domain temporal tagging. Language Resources and Evaluation
[5] YAKE! Keyword Extraction from Single Documents using Multiple Local Features