Prémios Arquivo.pt 2018

Memória Descritiva de trabalho

Identificação

- Título: Conta-me Histórias
- Área temática: Criação de Narrativas Temporais
- Candidato: Arian Pasquali, Vítor Mangaravite, Ricardo Campos, Alípio Jorge, Adam Jatowt
- Email: ricardo.campos@ipt.pt; vitordouzi@gmail.com; arianpasquali@gmail.com; amjorge@fc.up.pt; adam@dl.kuis.kyoto-u.ac.jp

Originalidade e caráter inovador

Ao longo dos últimos anos assistimos a uma tendência crescente de publicação e disponibilização de conteúdos escritos no formato online na forma de artigos noticiosos, comentários ou *posts*, criando novos desafios àqueles que pretendem entender o enredo (*storyline*) de uma notícia. Por outro lado, a manipulação dos factos (*fake news*) e o isolamento intelectual dos utilizadores por via de sítios da Web que utilizam algoritmos que inferem que informação o utilizador pretende ver (*filter bubbles*) levam também a novos desafios quanto à forma como os leitores consomem notícias e formam a sua opinião. Colocam assim em causa, em última instância, a existência de uma democracia plural que facilmente pode ser substituída por uma ditadura digital. O aumento exponencial do volume de dados (*big data*), a diversidade de notícias e um cada vez maior número de fontes (incluindo os media sociais) tornam, no entanto, praticamente impossível ao comum dos cidadãos o acesso, a gestão, e a memorização da informação ao longo do tempo sem recurso a ferramentas auxiliares. Embora o entendimento automático da linguagem natural tenha melhorado significativamente nos últimos anos com o desenvolvimento e a implementação de novos algoritmos no domínio da extração de informação, mineração de texto e a avaliação da credibilidade das notícias[1], o problema de construir estruturas narrativas consistentes encontra-se ainda por resolver [3]. Neste trabalho, pretendemos dar um contributo significativo para o entendimento e acompanhamento de um qualquer tópico (e.g., "*dilma rousseff*") ao longo do tempo. Com vista à concretização deste objetivo propomos o *Conta-me Historias*[2], uma aplicação online que, assente no Arquivo.pt[3], oferece aos cidadãos a possibilidade de entenderem rapidamente quais os principais atores de

[1] https://gate.d5.mpi-inf.mpg.de/credeye/

[2] bit.ly/DemoArquivopt

[3] http://arquivo.pt/

uma estória, as suas relações, motivações, e trajetórias no tempo, sem necessidade de procederem à leitura integral das fontes de dados, contribuindo desta forma para um acesso livre e democrático à informação (assente em fatos e tendencialmente livre de filtros ao fazer uso de diversas fontes jornalísticas). Para garantir a pluralidade e diversidade das fontes de informação, fazemos uso de 24 fontes de notícias eletrónicas (jornais e portais nacionais) que oferecem ao utilizador da aplicação, um espaço de construção de narrativa temporal assente nas suas fontes de informação preferidas. Uma tal ferramenta possibilita ao utilizador comparar a forma mais sensacionalista ou factual como determinados meios noticiosos acompanham um determinado tópico. Oferece ao jornalista um ambiente privilegiado para a pesquisa de eventos passados. Ao historiador, um resgatar da memória relacionada com acontecimentos históricos e ao cidadão, o acesso livre, democrático e plural a um manancial enorme de informação. A comunicação entre o utilizador e o sistema é feita através de uma interface gráfica, onde o utilizador especifica a sua consulta. Uma vez introduzida essa informação, o sistema recorre à API do Arquivo para, de entre as 24 fontes selecionadas, obter todos os títulos das páginas web com referência ao tópico introduzido pelo utilizador. A deteção automática dos melhores títulos noticiosos é feita com recurso ao YAKE![4] [1, 2] um extrator de palavras relevantes desenvolvido pela nossa equipa e que recentemente foi premiado com o Best Short Paper Award na quadragésima edição da European Conference on Information Retrieval (ECIR'18). Adicionalmente fazemos uso do SentiLex-PT01 [4], um léxico de sentimentos para o português desenvolvido por uma equipa de investigadores nacional e que usamos neste projeto para a análise de sentimentos de títulos selecionados como relevantes pelo YAKE!.

A Figura 1 mostra a interface gráfica do *Conta-me Histórias*. De uma forma simples os utilizadores podem interagir e testar a aplicação em 2 modos de teste distintos. No primeiro, selecionando uma de entre 5 pesquisas definidas pela equipa de investigação (ver "*Examples*"). No segundo, digitando as suas próprias pesquisas. Em "*advanced options*" o utilizador pode especificar um intervalo de tempo (e.g., últimos 5 anos, últimos 10 anos, etc) bem como escolher o fornecedor de conteúdos ("all" significa que o utilizador pretende obter informações tendo por base todos os jornais ou portais nacionais disponíveis na lista). A criação das narrativas temporais está naturalmente dependente dos resultados devolvidos pela API do Arquivo, pelo que a especificação de um período temporal de 10 anos não garante à partida a obtenção de resultados compreendidos entre [2008; 2018][5].

[4] https://boiling-castle-88317.herokuapp.com/

[5] A atual versão da API retorna resultados maioritariamente concentrados no período compreendido entre 2011 a 2016.

Figura 1: Interface de Pesquisa

Em qualquer dos dois casos, os resultados podem ser explorados fazendo uso de duas dimensões: (1) Narrativa; e (2) Termos Relacionados. Na narrativa temporal o utilizador terá acesso a uma memória descritiva do tópico que poderá ser explorada com recurso a uma timeline horizontal ou vertical. A Figura 2 mostra as potencialidades da aplicação, ao explorar o tópico “Dilma Rousseff" ao longo dos últimos 10 anos. A determinação dos períodos de tempo mais relevantes neste espaço temporal é feita tendo por base os picos de publicação dos artigos publicados entre 2011 e 2016. Frases marcadas a vermelho indicam um sentimento negativo, enquanto que frases marcadas a verde indicam um sentimento positivo. Frases sem qualquer tipo de marcação indicam um sentimento neutro. Uma observação atenta dos resultados permite ter acesso a momentos marcantes da ex-presidente Dilma Rousseff no período em questão. Nos trechos selecionados como relevantes é possível identificar outros atores relacionados com a personagem principal, nomeadamente “Cavaco Silva”, “Lula da Silva”, “José Serra”, “José Sócrates”, “Luis Amado” e “Barack Obama” que à data dos acontecimentos exerciam um cargo de chefia. Adicionalmente é possível também identificar algumas entidades geográficas, em particular “Brasil”, “Portugal”, “Rio de Janeiro” e “Médio Oriente”.

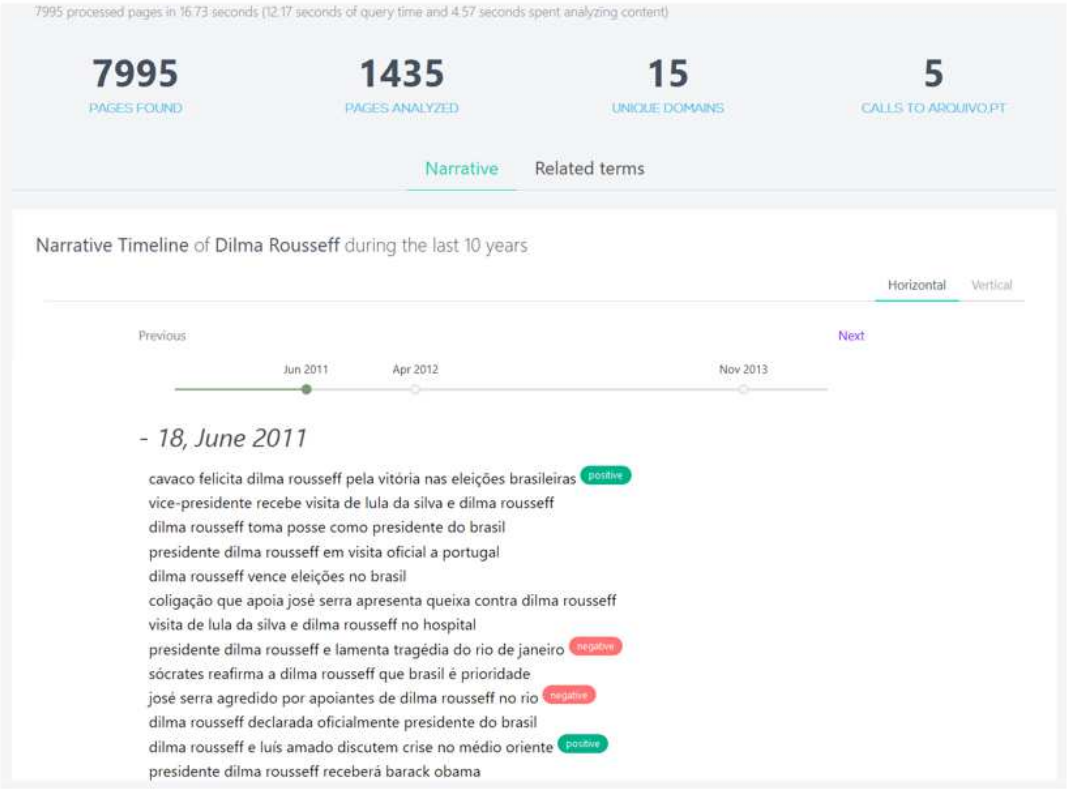

Figura 2: Narrativa Horizontal (pesquisa: "Dilma Rousseff" nos últimos 10 anos)

Em adição à narrativa temporal, disponibilizamos também as palavras que aparecem próximas ao tópico de pesquisa. Na Figura 3 é possível observar os termos relevantes associados à pesquisa Dilma Rousseff.

Figura 3: Termos relacionados com a pesquisa: "Dilma Rousseff" nos últimos 10 anos

É importante notar que os tempos de resposta, com variações entre os 10 e os 40 segundos, são maioritariamente inflacionados pela utilização da API do Arquivo. No topo da Figura 2 é possível observar que dos 16.73s necessários à obtenção dos resultados, apenas 4.57s dizem respeito à análise dos conteúdos. Uma solução *in-house*, que evite o recurso à API do Arquivo.PT, pode assim melhorar consideravelmente os tempos de retorno. Na lista abaixo, sumariamos os principais contributos deste trabalho:

1. Neste projeto, propomos uma aplicação que, ao fazer uso do acervo do Arquivo.pt, oferece ao utilizador um acesso privilegiado a um conjunto de conteúdos passíveis de serem explorados ao longo de um determinado espaço temporal;
2. Disponibilizamos uma solução que assente no YAKE! (uma abordagem não supervisionada) é facilmente adaptável a diferentes *providers* de conteúdos, línguas e contextos, por não fazer uso de coleções externas, ferramentas linguísticas ou qualquer tipo de treino de dados;
3. Finalmente, disponibilizamos ao utilizador da nossa aplicação, um ambiente *online* (responsivo, i.e., adaptável a PCs e *smartphones*), para que a nossa abordagem possa ser testada e avaliada em tempo real pelo utilizador, que assim terá a possibilidade de testar diferentes pesquisas/cenários.

Impacto social (aplicação e utilidade social)

Numa sociedade claramente marcada pela pós-verdade e pelas *fake news*, ferramentas como o Arquivo.pt são um contributo fundamental na preservação da memória coletiva. O *Conta-me Histórias*, surge, neste contexto, com o objetivo de oferecer ao cidadão uma forma adicional de exploração dos dados do Arquivo. Acreditamos que a construção de narrativas temporais acerca de um determinado tópico, são um contributo fundamental para uma sociedade civil mais bem informada e um auxiliar precioso para que estudantes, jornalistas, políticos, investigadores e cidadãos, possam, de uma forma geral, ter acesso a uma resenha histórica (factual e

transparente) acerca de um determinado tópico. A pesquisa “passos coelho impostos”[6] é apresentada neste contexto, como um exemplo bastante claro das potencialidades da aplicação. Uma leitura atenta dos resultados, permite observar as alterações no posicionamento do antigo primeiro ministro, no que à temática dos impostos diz respeito. Em junho de 2011 (ver Figura 4), em plena campanha eleitoral, contra o então governo de José Sócrates Passos Coelho “acusa” os socialistas de quererem aumentar os impostos.

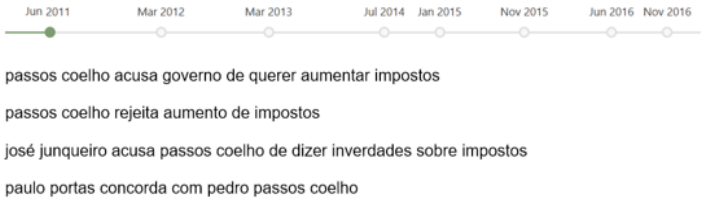

Figura 4: Narrativa horizontal de “Passos Coelho Impostos” em junho de 2011.

A vitória nas eleições em 2012 acabaria por levar Passos Coelho a primeiro-ministro do XIX Governo Constitucional (21 de junho de 2011 - 30 de outubro de 2015) e a um ajustamento do seu discurso no que ao aumento dos impostos diz respeito (ver Figura 5 e 6).

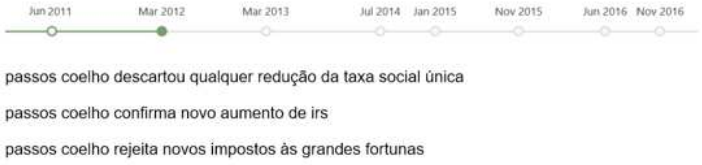

Figura 5: Narrativa horizontal de “Passos Coelho Impostos” em março de 2012.

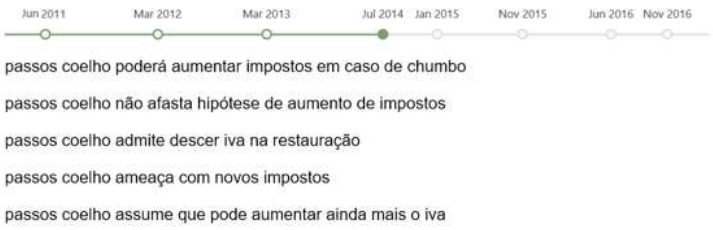

Figura 6: Narrativa horizontal de “Passos Coelho Impostos” em julho de 2014.

A inversão no discurso, dá-se apenas em meados de 2015 (ver Figura 7), com o aproximar de nova campanha eleitoral que levará Passos Coelho à formação do XX Governo Constitucional (30 de Outubro de 2015 - 26 de Novembro de 2015).

[6] Nota: a escolha de Passos Coelho é meramente casual e totalmente desprovida de motivações ou ideologias políticas, sendo considerada neste contexto, dada a distância dos acontecimentos, permitir uma análise no tempo.

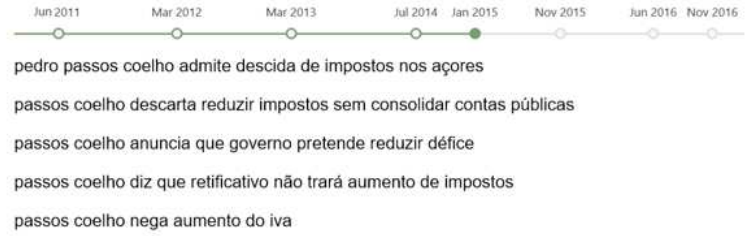

Figura 7: Narrativa horizontal de "Passos Coelho Impostos" em janeiro de 2015.

Em novembro de 2015 com a saída do Governo, Passos Coelho acaba por convocar eleições no PSD e vê José Eduardo Martins ser inicialmente apontado como seu sucessor (ver Figura 8).

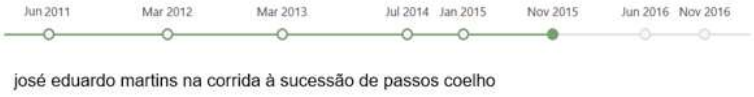

Figura 8: Narrativa horizontal de "Passos Coelho Impostos" em novembro de 2015.

Mas é Rui Rio, como mais tarde se viria a confirmar, que marca terreno na sucessão ao líder do partido. No regresso à bancada parlamentar do PSD e já depois de ganhar as eleições internas, Passos Coelho volta a contestar as propostas do governo que no seu entendimento implicam inesperados aumentos, exigindo também explicações sobre o caso Banif que marca o início da legislatura de António Costa (ver Figura 9).

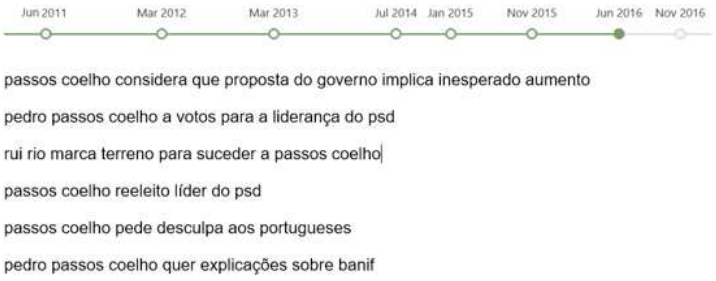

Figura 9: Narrativa horizontal de "Passos Coelho Impostos" em junho de 2016.

A resenha histórica aqui apresentada, é apenas um, dos muitos exemplos da utilidade da nossa aplicação e do seu potencial impacto social. Outros exemplos, de acontecimentos e personagens recentes da nossa história, podem ser experimentados na nossa aplicação.

Impacto científico (aplicação e utilidade científica)

Este trabalho foi realizado por uma equipa de investigação que tem desenvolvido e implementado soluções no âmbito da extração de informação e da construção de narrativas. Depois do desenvolvimento do YAKE!, e da organização do primeiro workshop em Extração Narrativa a

partir de Textos, encontramo-nos presentemente a organizar um special issue[7] no IPM Journal, a que se junta uma candidatura a um projeto do Portugal 2020 (a aguardar decisão). Como resultado da nossa participação neste concurso, pretendemos a breve prazo submeter um artigo científico ao CIKM'18 com a descrição dos resultados obtidos, bem como continuar a explorar e a desenvolver novas funcionalidades na aplicação. Em particular, ponderamos a aplicação de ferramentas NER (named entity recognition) para a deteção formal de atores (pessoas e organizações) e espaços geográficos (cidades, países). Adicionalmente, perspetivamos aplicar o nosso extrator de palavras relevantes (YAKE!) às várias versões de uma mesma página web. Uma tal abordagem, permitirá também aos cientistas sociais de áreas como sociologia, ciência política, ciências da comunicação e história, investigar, por exemplo, os principais tópicos discutidos ao longo do tempo na página principal de um partido, e desta forma observar eventuais diferenças na estratégia adotada por diferentes líderes partidários e governos. Acreditamos que o *Conta-me Histórias* é, neste contexto, um importante contributo para a comunidade científica ao abrir novas oportunidades de investigação e exploração dos dados do Arquivo.pt.

Relevância da utilização do Arquivo.pt

Neste trabalho usámos o Arquivo.pt como fonte de informação da nossa aplicação. Com base na API pudemos ter acesso a um conjunto de notícias publicadas ao longo dos últimos anos por diversos *providers* de conteúdos e desta forma construir uma narrativa temporal para cada pesquisa. Em particular recorremos aos websites de 24 fontes de informação, a saber:

- Público;
- Diário de Notícias;
- DNotícias;
- RTP;
- Correio da Manhã;
- IOL;
- TVI24;
- Notícias SAPO;
- SAPO;
- Expresso;
- Sol;
- Jornal de Negócios;
- A Bola;
- Jornal de Notícias;
- Sic Notícias;
- Lux;
- Jornal i;
- Google Notícias;
- Dinheiro Vivo;

[7] https://www.journals.elsevier.com/information-processing-and-management/call-for-papers/special-issue-on-narrative-extraction-from-texts-text2story

- AEIOU;
- TSF;
- Meios & Publicidade;
- Sábado;
- Jornal Económico;

Recursos complementares

[1] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. ECIR'18
[2] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A. (2018). YAKE! Collection-independent Automatic Keyword Extractor. ECIR'18
[3] Jorge, A., Campos, R., Jatowt, A., Nunes, S. (2018). First International Workshop on *Narrative Extraction from Text* (Text2Story'18). ECIR'18
[4] Silva, M., & Carvalho, P., & Costa, C., & Sarmento, L. (2010). Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis. Technical Report. TR 10-08. University of Lisbon, LASIGE.