Uma Aplicação *Web* para Sumarização Multidocumento de Notícias em Português

Rafael Ribaldo, Francisco A. Cabelo, Thiago A. S. Pardo
Instituto de Ciências Matemáticas e de Computação, USP/São Carlos

1. Objetivos

Nas últimas décadas, muitas tecnologias novas têm surgido, trazendo com isso não somente um crescente aumento no volume de informação, mas também a necessidade do tratamento automático da mesma, dado que o seu processamento se torna cada vez mais difícil. Diante disso, a tarefa de sumarização automática (SA) multidocumento, a qual consiste em produzir automaticamente um único sumário a partir de um grupo de textos sobre o mesmo assunto, torna-se uma tarefa de grande importância.

O objetivo deste trabalho foi criar uma ferramenta (aplicação *web*) capaz de lidar com notícias jornalísticas on-line, por meio da adaptação de técnicas clássicas para a sumarização multidocumento.

2. Métodos e Procedimentos

Muitos métodos e técnicas têm sido aplicados em diferentes abordagens da SA. O método de Salton et al. (1997), por exemplo, constrói o sumário a partir da relevância das informações contidas nos documentos. Esse foi o método adotado neste trabalho.

A aplicação *web* criada possibilita a extração e a sumarização de documentos, por meio da busca por notícias que versam sobre um mesmo assunto. Para isso, utilizou-se 1) da *API* do buscador *Google News* para a recuperação dos textos; 2) da ferramenta de remoção de conteúdo irrelevante de páginas *web* *NCleaner* (Evert, 2008) e 3) do sumarizador criado a partir do modelo de Salton et al., chamado *RSumm* (Ribaldo et al., 2012).

3. Resultados e Conclusões

A Figura 1 mostra o aplicativo *web* desenvolvido, seguido da Figura 2, que apresenta o sumário gerado a partir da busca realizada previamente.

Abaixo, temos uma pesquisa pelo termo "CPI Cachoeira", onde o retorno, contendo os 8 documentos mais relevantes, de acordo com o *Google News*, é então sumarizado.

Figura 1 – Botão "Sumarizar"

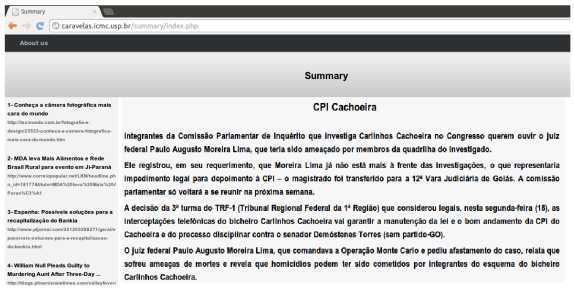

Figura 2 – Sumário

Referências

Evert, Stefan (2008). A lightweight and efficient tool for cleaning Web pages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco, May, 2008.

Ribaldo, R.; Akabane, A.T.; Rino, L.H.M.; Pardo, T.A.S. (2012). Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In the *Proceedings of the 10th International Conference on Computational Processing of Portuguese* (LNAI 7243), pp. 260-271. April 17-20, Coimbra, Portugal.

Salton, G.; Singhal A.; Mitra, M; Buckley C. (1997). *Automatic Text Structuring And Summarization*. Information Processing & Management, Vol. 33, No, 2, pp. 193-207.