# A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias

Evandro Cunha[1]
Gabriel Magno[2]
Virgilio Almeida[3]

## Introdução

Dentro do contexto da linguística de corpus, área que se ocupa da coleta e da análise de grandes conjuntos de dados linguísticos reais, tem-se mostrado necessário o desenvolvimento de ferramentas capazes de auxiliar na coleta e na organização de material. Frequentemente, conteúdo disponível em meio digital (por exemplo, em páginas da Web) é importante para que o pesquisador realize determinadas investigações; contudo, observa-se que, em certas ocasiões, tais investigações não são levadas a cabo pela ausência de ferramental adequado.

Em particular, um tipo de conteúdo muito relevante para pesquisadores em humanidades digitais – inclusive linguistas – são os comentários publicados por leitores de notícias em portais online. A análise desses comentários permite o estudo de questões linguísticas nos mais variados domínios, como o lexical, o morfossintático e o pragmático, além de uma série de outras questões em diversas áreas do conhecimento, entre elas a sociologia, as ciências políticas e a comunicação social.

Neste artigo, são apresentados dois recursos que visam auxiliar os pesquisadores interessados em trabalhar com comentários de leitores em portais online: (a) um coletor de comentários de portais de notícias; e (b) um corpus composto por comentários de leitores do portal UOL. Ambas os recursos receberam o nome de Xereta – uma brincadeira com o significado coloquial do termo, que, de acordo com o dicionário Michaelis[4], significa "que ou aquele que se intromete em assuntos alheios de forma inconveniente; bisbilhoteiro, intrometido".

Na próxima seção, são apresentadas as principais características dos textos postados por leitores como comentários em portais de notícias. Nas seções seguintes, são discutidos os resultados

1 Doutorando em Linguística na Universiteit Leiden (Holanda) e em Ciência da Computação na Universidade Federal de Minas Gerais (UFMG). E-mail: evandrocunha@dcc.ufmg.br

2 Doutorando em Ciência da Computação na Universidade Federal de Minas Gerais (UFMG). E-mail: magno@dcc.ufmg.br

3 Professor Titular do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais (UFMG) e Faculty Associate do Berkman Klein Center for Internet and Society, Harvard University (EUA).
E-mail: virgilio@dcc.ufmg.br

4 http://michaelis.uol.com.br

dos esforços para a elaboração do coletor e do corpus Xereta e apresentam-se, de maneira geral, os principais desafios e limitações conceituais e computacionais enfrentados durante sua elaboração.

## Comentários em portais de notícias

Portais de notícias são atualmente responsáveis por grande parte do volume de tráfego na Internet brasileira. De acordo com dados da Alexa Internet, cinco entre os dez sites mais acessados do país em junho de 2017 podem ser considerados dessa categoria (ALEXA, 2017). A partir da década de 2000, com o advento da Web 2.0 (CUNHA, 2012, p.8), tais portais passaram a incorporar a participação do leitor em suas notícias: na mesma página em que uma certa notícia é publicada, um leitor é capaz de expor sua opinião e torná-la disponível para os demais interessados na notícia em questão – criando, assim, não apenas conteúdo, mas sobretudo conversações e comunidades (COMM, 2009 *apud* LEAL & ROSSINI, 2012).

A Figura 1 ilustra alguns desses comentários publicados por leitores em uma notícia do portal UOL[5]. É interessante notar que uma notícia mais popular pode alcançar um número considerável de comentários, às vezes ultrapassando centenas deles. Alguns portais permitem que um comentário seja postado em resposta ao comentário de outro usuário e até mesmo que o leitor avalie positivamente ou negativamente um comentário publicado.


Figura 1. Comentários publicados por leitores em uma notícia do portal UOL. Os nomes e as fotos dos usuários foram tornados anônimos por questões de privacidade.

5 http://www.uol.com.br

Diversos estudos apresentam ferramentas que lidam com comentários de portais de notícias. Por exemplo, Hsu et al. (2009) propõem uma abordagem para ranquear comentários com base na qualidade – o que pode ser útil, entre outras coisas, para filtrar spams –, enquanto Potthast & Becker (2010) apresentam uma ferramenta para auxiliar a sumarização e a visualização de opiniões expressas na forma de comentários na Internet. Outros estudos utilizam comentários na Web como objeto de investigação: como exemplo, pode-se citar o trabalho de Potthast (2009), que analisa a natureza descritiva de comentários em objetos textuais (incluindo em notícias), e o trabalho de Reyes et al. (2010), que avalia aspectos humorísticos em comentários. Há também uma série de estudos que focam na análise de comentários em contextos específicos, como os contextos político (LEAL & ROSSINI, 2012) e legislativo (ROSSINI & MAIA, 2014; MAIA et al., 2015).

## Descrição das ferramentas (coletor e corpus Xereta)

Nesta seção, são descritos o coletor e o corpus de comentários Xereta. Por não se tratar de uma publicação de escopo computacional, não são tratados aqui detalhes de implementação do coletor, mas apenas informações gerais sobre seu funcionamento.

### *Coletor Xereta*

O objetivo do coletor Xereta é, dada uma lista de páginas de notícias, retornar um arquivo contendo todos os comentários publicados nessas páginas. Uma de suas principais características é seu funcionamento simples, permitindo sua utilização por usuários com conhecimento apenas básico de computação.

Inicialmente, é importante destacar que o coletor Xereta foi desenvolvido em código aberto e livre para uso, modificação e distribuição. Isso significa que ele pode ser baixado[6] e utilizado gratuitamente, sem necessidade de nenhum tipo de registro, e que, caso o usuário sinta necessidade, pode alterá-lo livremente para melhor atender aos seus objetivos – situação em que, naturalmente, é necessário que o usuário possua conhecimentos de programação. O coletor Xereta possui a licença GNU GPL (General Public License), que garante as liberdades de executar o programa para qualquer propósito, de estudar como ele funciona e adaptá-lo às suas necessidades, de redistribuir cópias e, finalmente, de aperfeiçoá-lo e liberar seus aperfeiçoamentos de modo que toda a comunidade se beneficie deles.

O coletor Xereta está disponível em duas formas diferentes: para download e para uso online. O arquivo disponível para download é o mesmo utilizado na versão online. A versão para download é destinada sobretudo àqueles que desejam conhecer e aprimorar o código, enquanto a versão online é mais indicada para uso geral, sobretudo por usuários mais inexperientes.

6 A partir dos endereços http://xereta.herokuapp.com ou http://www.dcc.ufmg.br/~evandrocunha/xereta

A versão online do coletor Xereta está disponível na página xereta.herokuapp.com (caso em algum momento essa página esteja fora do ar, recomenda-se procurar por instruções no endereço www.dcc.ufmg.br/~evandrocunha). Para realizar a coleta a partir do site, basta inserir na caixa disponível a lista de URLs de onde os comentários devem ser coletados. O arquivo "comentarios.csv", que será baixado automaticamente, poderá ser aberto em um editor de planilhas (como o LibreOffice Calc ou o Microsoft Excel). Na versão do coletor vigente no momento desta publicação, o arquivo contendo os comentários possui as seguintes oito colunas:

1) número de identificação do comentário, fornecido pelo portal;
2) número de identificação do comentário ao qual o comentário em questão responde, caso o comentário em questão seja uma resposta a outro comentário;
3) título da notícia onde o comentário foi publicado;
4) nome ou apelido fornecido pelo usuário que comentou;
5) texto do comentário em si;
6) número de avaliações positivas do comentário, caso o portal permita que os comentários sejam avaliados positivamente;
7) número de avaliações negativas do comentário, caso o portal permita que os comentários sejam avaliados negativamente;
8) URL da notícia.

O principal desafio na elaboração do coletor Xereta está no fato de que as estruturas das páginas de cada portal de notícias (UOL, Folha, G1, Terra etc.) são completamente diferentes umas das outras. Além disso, pode acontecer que as estruturas das páginas de seções diferentes (entretenimento, esportes, política etc.) de um mesmo site sejam também diferentes. Tudo isso faz com que o código utilizado para coletar comentários de um determinado portal não possa ser reaproveitado para coletar a partir de outro portal: basicamente, é necessário gerar um código modificado para cada novo portal a ser coletado, muitas vezes sendo necessário visitar as várias seções do site para identificar em que medida elas se diferenciam entre si.

Ademais, pode-se verificar que as informações disponíveis para coleta também variam: por exemplo, enquanto alguns portais permitem que os comentários sejam avaliados positivamente e negativamente, outros permitem apenas avaliações positivas ("curtidas"), enquanto outros sequer oferecem esse serviço. Essas diferenças interferem na decisão final de quais campos devem ser incluídos no arquivo de saída. Por uma decisão de implementação, optou-se por gerar um arquivo de saída único que compila todos os comentários das URLs fornecidas, independentemente do portal de origem. Assim, foi necessário definir um conjunto de campos de informação que contemplasse de forma consistente comentários de quaisquer portais. No caso das avaliações, na primeira

versão disponível do coletor Xereta, optou-se por fornecer, sempre que possível, a quantidade de comentários positivos e negativos, pois considerou-se ser uma informação relevante em muitos contextos. Entretanto, quando o portal não oferece o serviço de avaliações, o coletor retorna o valor "0" no arquivo de saída. É importante, portanto, que o pesquisador esteja ciente disso quando for utilizar as informações contidas nos campos "avaliações positivas" e "avaliações negativas" de seu arquivo de saída.

Na Figura 2, é apresentado o diagrama de funcionamento do coletor Xereta. Primeiramente, pode-se observar que o coletor recebe como entrada um único arquivo com a lista de todas as URLs a serem coletadas e, no final, gera um único arquivo de saída contendo os comentários de todas as URLs e seus respectivos campos de informação. Internamente, a primeira tarefa do programa é consumir o arquivo de URLs e armazená-las em uma lista. Em seguida, o programa itera sobre essa lista, consumindo uma URL por vez. Como mencionado anteriormente, o algoritmo para extrair os comentários e suas informações é diferente para cada portal, tornando-se necessária a criação de um módulo específico para cada site. O coletor então verifica o texto da URL em si para identificar o portal de origem e, consequentemente, qual módulo de coleta utilizar. É importante frisar que, apesar da necessidade de desenvolvimento de módulos individuais, o coletor Xereta é flexível o suficiente para permitir a inclusão de novos módulos de coleta referentes a outros portais, bastando o programador implementá-los utilizando o mesmo formato dos demais. Após a extração dos comentários de uma determinada URL, eles são incluídos em uma lista comum onde todos os comentários coletados já estão armazenados. Ao final, quando toda a lista de URLs é percorrida, o programa salva a lista de comentários no arquivo de saída.

## *Corpus Xereta*

O corpus Xereta é um conjunto de comentários coletado e organizado pelos próprios autores do coletor Xereta – utilizando-se, naturalmente, de tal coletor. Assim como o coletor, o corpus Xereta está disponível na página xereta.herokuapp.com (caso em algum momento essa página esteja fora do ar, recomenda-se procurar por instruções no endereço www.dcc.ufmg.br/~evandrocunha). A versão preliminar do corpus Xereta conta com 25.441 comentários postados em notícias publicadas entre janeiro e julho de 2014 nas seguintes seções do portal UOL: "Blogs", "Ciência", "Cotidiano", "Opinião" e "Política". Na segunda versão, em fase final de coleta no momento da publicação deste artigo, serão disponibilizados mais de um milhão de comentários extraídos a partir de notícias publicadas entre 2012 e 2014 em várias seções do site. Para a elaboração das listas contendo URLs relevantes para coleta, foi utilizado o *sitemap* do próprio UOL[7].

7 *Sitemap* é uma lista contendo as URLs das páginas que compõem um site. O *sitemap* do UOL está disponível no endereço http://sitemaps.uol.com.br

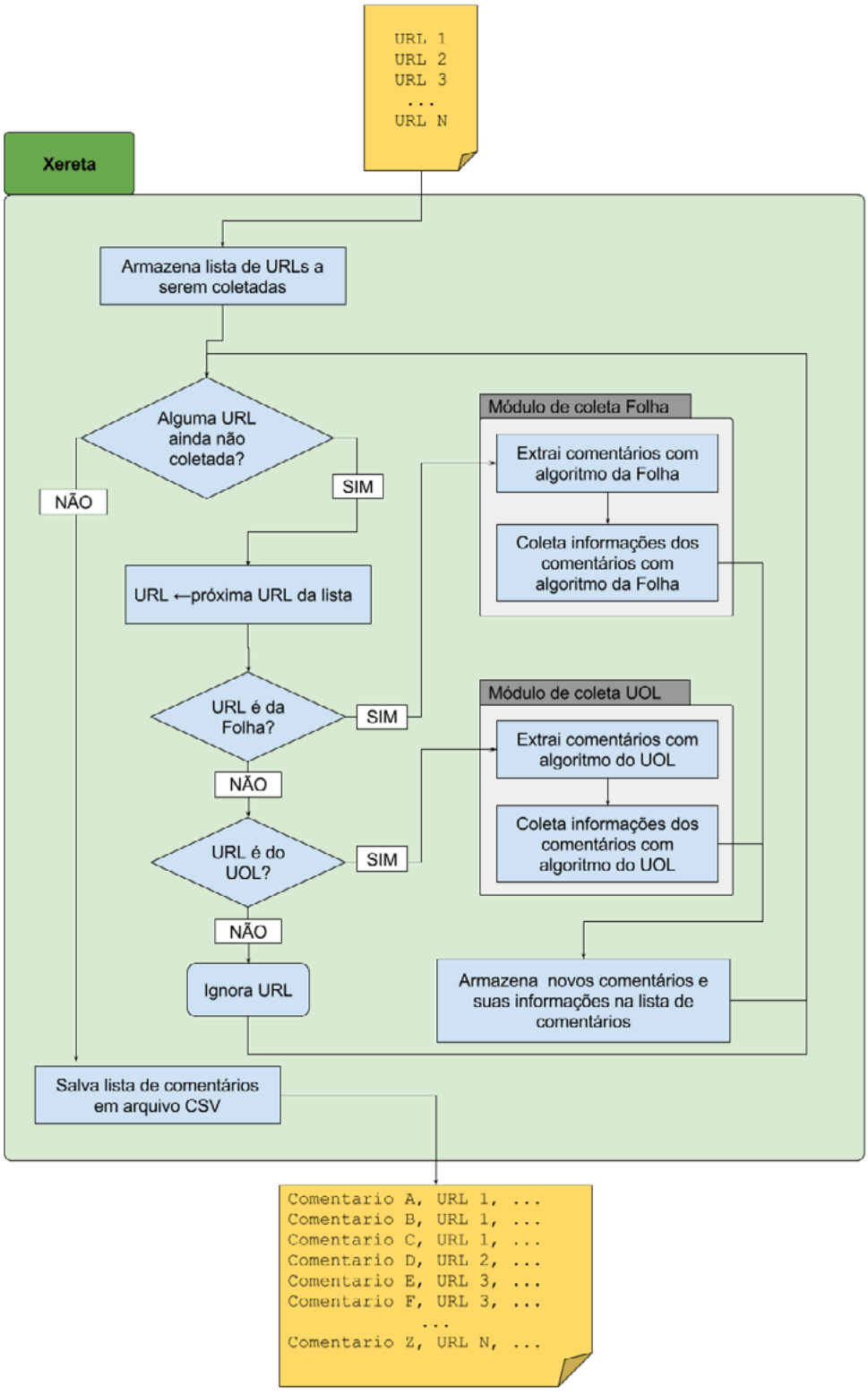


Figura 2. Diagrama de funcionamento do coletor Xereta

## Considerações finais

Neste artigo, são apresentados e descritos dois recursos importantes para a pesquisa nas áreas de humanidades digitais e, em particular, para os estudos em linguística de corpus: um coletor de comentários publicados em notícias de portais da Internet e um corpus composto por uma seleção de comentários do portal UOL. Aqui, mostra-se que o coletor possui funcionamento simples, podendo ser utilizado até mesmo por usuários com conhecimentos limitados de computação. Com relação ao corpus, mostra-se que ele contém comentários de diversas seções do portal UOL, tornando possível, por exemplo, a análise de características textuais presentes em diferentes seções de um portal de notícias online.

Na sequência dos trabalhos, pretende-se aperfeiçoar o coletor para que ele seja capaz de obter conteúdo de ainda mais portais de notícias, tanto brasileiros (G1, Terra etc.) quanto estrangeiros (NYT, Washington Post etc.). Paralelamente, pretende-se disponibilizar uma nova versão do corpus, com cerca de um milhão de comentários.

Como mencionado anteriormente, o coletor Xereta é desenvolvido em código aberto e livre para uso, modificação e distribuição. Os autores deste trabalho acreditam que a criação de uma comunidade de desenvolvedores e usuários interessados em melhorar a ferramenta seria importante para seu pleno desenvolvimento. Por essa razão, os interessados em fazer parte dessa comunidade são encorajados a entrar em contato com os autores.

## Agradecimentos

Agradecemos à Patrícia Rossini, que, enquanto doutoranda em Comunicação Social na UFMG, nos motivou a elaborar o coletor Xereta e nos auxiliou testando versões preliminares em suas próprias pesquisas.

## Referências bibliográficas

ALEXA. *Top Sites in Brazil*. Disponível em http://www.alexa.com/topsites/countries;0/BR Acessado em 08/06/2017.

CUNHA, Evandro L.T.P. *Etiquetagem de micromensagens no Twitter: uma abordagem linguística*. Dissertação de mestrado. Belo Horizonte: Universidade Federal de Minas Gerais, 2012.

HSU, Chiao-Fang; KHABIRI, Elham; CAVERLEE, James. Ranking comments on the social web. In: Computational Science and Engineering, pp. 90-97, 2009.

LEAL, Paulo Roberto Figueira; ROSSINI, Patrícia Gonçalves da Conceição. As campanhas eleitorais no contexto da política personalizada. In: *Comunicação Política e Eleitoral no Brasil: Perspectivas e limitações no dinamismo político*. 2012.

MAIA, Rousiley et al. Sobre a importância de examinar diferentes ambientes online em estudos de deliberação. Opinião Pública, v. 21, n. 2, 2015.

POTTHAST, Martin. Measuring the descriptiveness of web comments. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 724-725, 2009.

POTTHAST, Martin; BECKER, Steffen. Opinion summarization of web comments. Advances in Information Retrieval, pp. 668-669, 2010.

REYES, Antonio et al. Evaluating Humour Features on Web Comments. In: LREC. 2010.

ROSSINI, Patrícia GC; MAIA, Rousiley CM. Is Political Participation Online Effective? Handbook of Research on Advanced ICT Integration for Governance and Policy Modeling, 2014.