Extraindo Conteúdo de Sites Jornalísticos - Desenvolvimento da Ferramenta Lnews para Pesquisadores da Comunicação.[1]

Márcio Carneiro dos SANTOS[2]
Universidade Federal do Maranhão , São Luís, MA

RESUMO

Descreve-se a iniciativa de pesquisa aplicada que desenvolveu uma ferramenta de extração de conteúdo, capaz de coletar material em sites jornalísticos de forma automatizada, a fim de permitir aos pesquisadores da Comunicação uma estratégia de coleta de dados mais efetiva, diante do atual contexto de excesso de informação presente no ambiente digital contemporâneo. Desenvolvida em Python e guiada pela proposta epistemológica da Design Science, a solução LNEWS hoje encontra-se em fase de utilização por um grupo de testadores (*beta testers*) que foram selecionados a partir de chamada pública. Esse trabalho está inserido numa iniciativa maior denominada de Métodos Digitais, focada na atualização do conjunto de técnicas e métodos de pesquisa utilizados pela área da Comunicação, a partir de abordagens inter e multidisciplinares.

PALAVRAS-CHAVE: métodos digitais; design science; extração de dados.

1.0 INTRODUÇÃO

A condição contemporânea do ecossistema comunicacional, baseada em redes digitais, massivo crescimento de emissores e explosão informacional, caracterizada por volume, velocidade e variedade, tem sido devidamente registrada nos últimos vinte anos a partir dos trabalhos de Castells (1999), Chwe (2000), Feenberg (2002), Lemos (2002), Santaella (2003), Vilches (2003) entre outros. Suas características de interconexão (NEWMAN, 2010; EASLEY, KLEINBERG, 2010), excesso (GLEICK, 2013) e complexidade (SIMON, 1962; MORIN, 2005), aliadas a um conjunto próprio de especificidades que constituem uma verdadeira ontologia dos entes binários, também já mapeada por Manovich (2001) e Santos (2016), indicam um quadro consolidado de percepção de profundas mudanças que, entretanto, parece ainda não ter impactado de

[1] Trabalho apresentado no GP Conteúdos Digitais e Convergências Tecnológicas no XIX Encontro dos Grupos de Pesquisas em Comunicação, evento componente do 42º Congresso Brasileiro de Ciências da Comunicação.

[2] Doutor em Tecnologias da Inteligência e Design Digital pela PUC-SP. Professor permanente dos Mestrados em Design e Comunicação Profissional da UFMA. Coordenador do LABCOM/DCS/UFMA. E-mail: mcszen@gmail.com .

forma proporcional os métodos e técnicas utilizados pelos pesquisadores da Comunicação para analisar objetos do ciberespaço.

A iniciativa denominada Métodos Digitais, que tem em Rogers (2013) um de seus principais interlocutores, explicita a necessidade, gerada pelas especificidades dos entes de origem binária, de estuda-los a partir métodos e técnicas originados em sua lógica interna.

> Por exemplo, varredura e extração de dados, inteligência coletiva e classificações baseadas em redes sociais, ainda que de diferentes gêneros e espécies, são todas técnicas baseadas na internet para coleta e organização de dados. *Page Rank* e algoritmos similares são meios de ordenação e classificação. Nuvens de palavras e outras formas comuns de visualização explicitam relevância e ressonância. Como poderíamos aprender com eles e outros métodos *online* para reaplica-los? O propósito não seria tanto contribuir para o refinamento e construção de um motor de buscas melhor, uma tarefa que deve ser deixada para a Ciência da Computação e áreas afins. Ao invés disso o propósito seria utiliza-los e entender como eles tratam *hiperlinks*, *hits*, *likes*, *tags*, *datestamps* e outros objetos nativamente digitais. Pensando nesses mecanismos e nos objetos com os quais eles conseguem lidar, os métodos digitais, como uma prática de pesquisa, contribuem para o desenvolvimento de uma metodologia do próprio meio (ROGERS, 2013, E-book).[3]

Santos (2016, p. 32) detalha a proposta:

> Transpondo tal raciocínio ao trabalho de pesquisa, Rogers também separa os métodos eminentemente digitais dos que ele denomina de virtuais, ou seja, que têm sua origem em outros campos e têm sido adaptados para a internet e as redes sociais. A netnografia ou etnografia virtual, os questionários aplicados via e-mail, as entrevistas mediadas pelo computador e pelas redes são algumas das formas adaptadas, diferentes, por exemplo, da mineração e raspagem de dados (data mining e scraping), do acesso direto às APIs das plataformas de mídia social, da utilização de métricas com o Page Rank ou de ferramentas como Open Refine para, respectivamente, coletar, classificar e organizar dados.

Os métodos digitais, dessa forma, não pretendem substituir os atuais, mas complementá-los em situações de pesquisa onde se está lidando com objetos cuja ontologia própria exige um tipo de abordagem mais efetiva.

A proposta do LNEWS vai nessa direção, não como uma ferramenta específica para uma pesquisa e sim como uma meta-ferramenta de utilização ampla para uma categoria de problemas como propõe a linhagem epistemológica da Design Science.

Design Science (DS) é a "ciência que procura consolidar conhecimento sobre o projeto e desenvolvimento de soluções para melhorar sistemas existentes,

[3] Tradução do autor.

resolver problemas e criar novos artefatos" (DRESCH; LACERDA; ANTUNES Jr., 2015, p. 59).

"O artefato, criado pelo homem, representa um intermediador entre um conjunto do conhecimento estabelecido em determinada área e as condições específicas que envolvem o problema que o artefato deverá resolver. (SANTOS, 2016, p.11).

A DS classifica os artefatos em um continuo que vai de um extremo mais abstrato a outro mais tangível e inclui as categorias de constructos, modelos, métodos e instanciações.

Dai a importância também do conceito de classe de problemas.

> O termo classe de problemas que temos utilizado também faz parte dos conceitos importantes da DS. Conjuntos de problemas práticos ou teóricos que tem já estabelecido um conjunto de soluções ou artefatos a eles ligados constituem-se numa classe de problemas. Como exemplo da Comunicação e das Ciências Sociais, poderíamos citar a necessidade geral de coletar dados em repositórios na internet, que poderíamos nomear como coleta de dados digitais. Seja para a produção de uma matéria jornalística, para um plano de gestão ou para a definição de uma política pública sobre determinado tema, com os processos de digitalização e o crescimento do uso de bases de dados, a necessidade de conseguir tais informações, acessando seus repositórios disponíveis na rede, tais como portais de transparência, por exemplo, caracteriza uma classe de problemas onde operam artefatos como os métodos de scraping (raspagem) e extração automatizada, bem como as instanciações disponíveis exemplificadas pelos algoritmos em determinada linguagem de programação, que operam para resolver tais problemas (SANTOS, 2016, p.13).

Deste modo a proposta do LNEWS se insere na categoria de problemas representada pela necessidade enfrentada pelos pesquisadores de extrair conteúdo a partir de sites jornalísticos de forma automatizada, garantindo uma amostragem mais significativa e ganho de tempo, evitando a coleta manual bem mais lenta e cansativa. O artefato, utilizando a terminologia da DS, permite também ao pesquisador da Comunicação a visualização da estrutura interna das editorias/categorias que os veículos utilizam ao classificar as notícias para publicação. Tal informação, sem o uso de recursos computacionais, é de apreensão bem mais difícil. Por fim LNEWS é orientado pela simplicidade justamente por estar focado num conjunto de utilizadores mais distantes das práticas de programação e técnicas de raspagem de dados.

2.0 LNEWS

No seu atual estágio a ferramenta LNEWS permite ao pesquisador:

a) Selecionar um site ou portal jornalístico, com 3 opções pré-selecionadas UOL, ESTADÃO e GLOBO.COM e uma quarta opção onde o usuário pode entrar

com um endereço específico para, por exemplo, explorar algum veículo da sua cidade ou estado, com relevância específica para sua pesquisa;

b) A partir da escolha feita no item anterior o software coleta uma lista de links disponíveis para acesso no momento da consulta onde é possível ver a estrutura de organização das editorias/categorias que são utilizadas pelos emissores na publicação do seu conteúdo;

c) Com a lista o usuário pode escolher qualquer uma das matérias recuperadas e vê-la na íntegra com alguns dos seus meta-dados, quando acessíveis, tais como seus autores. Ainda nesse momento é criada uma lista das palavras mais frequentes e feito um sumário (resumo) do texto. Essa última função ainda apresenta pouca eficiência no caso da língua portuguesa.

d) Após essa etapa é oferecido ao pesquisador salvar sua coleta criando dois arquivos, um com a lista de links para as matérias e o segundo com os textos das mesmas, podendo realizar essa última ação adicionando filtros por editoria ou até pela seleção de matérias que apresentem uma palavra específica. Por exemplo, do total de matérias coletadas é possível salvar apenas as da editoria Política em que a palavra "poder" esteja presente.

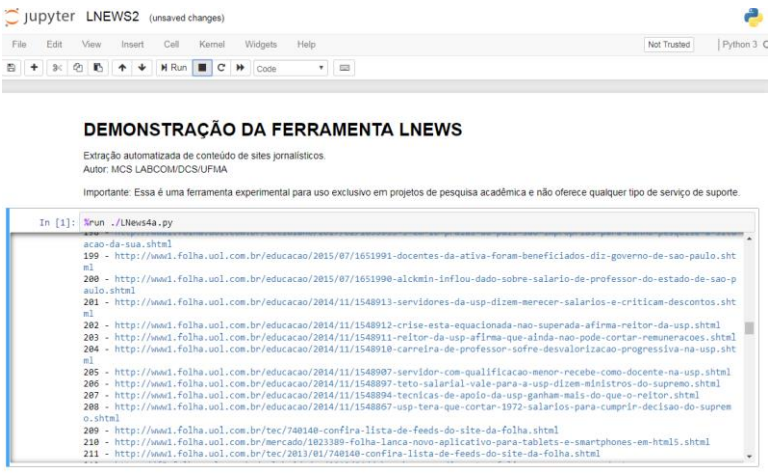

Figura 1 – Tela com lista de links extraídos. Fonte: do autor.

Outro módulo da solução, denominado LNEWS Análise, já trabalha não mais com os sites diretamente e sim com os arquivos de textos extraídos. A partir deles estão sendo desenvolvidas funções para calcular e visualizar métricas tais como a diversidade léxica, que traduz a riqueza na utilização das palavras do texto e as palavras frequentes que podem ser também listadas ou apresentadas num gráfico.

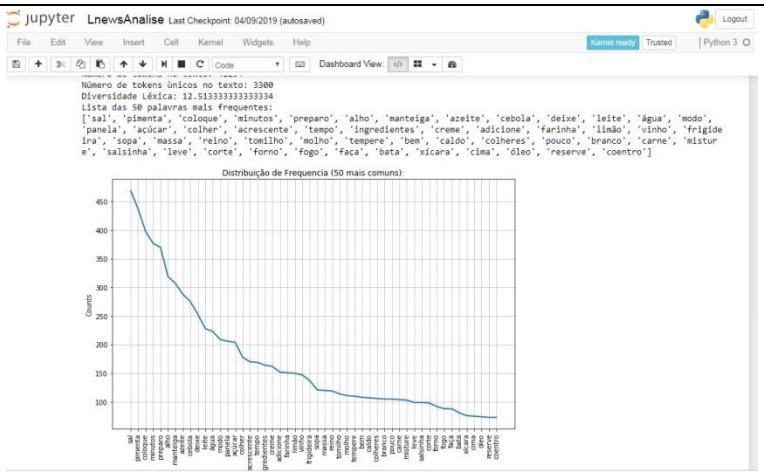

Figura 2 – Tela Lnews Análise com palavras frequentes do texto e gráfico. Fonte : do autor .

Hoje o LNEWS ainda encontra-se em fase de testes (beta) a partir de um grupo de pesquisadores que se inscreveram numa chamada pública para testá-lo em suas próprias pesquisas. Cada um recebeu um link para explorar a solução num ambiente web. O procedimento acertado é que, após cada sessão de utilização, o *beta tester* preencherá um relatório simples indicando se conseguiu realizar as tarefas específicas da sua pesquisa, que dificuldades teve e que sugestões faria para o desenvolvimento do software. Mais informações sobre esse processo e o LNEWS estão disponíveis em https://www.labcomdata.com.br/ .

3.0 REFERÊNCIAS

CASTELLS, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999.

CHWE, Michael S. Communication and coordination in social networks. **Review of Economic Studies**, 67, p. 128-156, 2000.

DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JR, José Antonio Valle. **Design Science Research**: método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015.

EASLEY, David; KLEINBERG, Jon. **Networks, Crowds and Markets**: reasoning about a highly connected world. Nova York: Cambridge University Press, 2010.
FEENBERG, Andrew. **Transforming technology**: a critical theory revisited. New York: Oxford University Press, [E-book], 2002.

GLEICK, James. **A informação**: uma história, uma teoria, uma enxurrada. São Paulo, Companhia das Letras, 2013.

LEMOS, André. **Cibercultura**: tecnologia e vida social na cultura contemporânea. 4. ed. Porto Alegre: Sulina, 2002.

MANOVICH, L. **The language of new media**. Massachusetts: Mit Press. 2001.

MORIN, Edgar. **Introdução ao pensamento complexo**. 4.ed. Porto Alegre: Sulina, 2005.

NEWMAN, M. E. **Networks:** an introduction. Nova York: Oxford University Press, 2010.

ROGERS, Richard. **Digital Methods**. Cambridge: Mit Press. [E-book], 2013.

SANTAELLA, Lúcia. **Culturas e artes do pós-humano**: da cultura das mídias à cibercultura. São Paulo: Paulus, 2003.

SANTOS, Márcio. **Comunicação digital e jornalismo de inserção**: como big data, inteligência artificial, realidade aumentada e internet das coisas estão mudando a produção de conteúdo informativo. São Luís: LABCOM Digital, 2016.

SIMON, Herbert A. **The architecture of complexity**. In: Proceedings of the American Philosophical Society. v. 106, n. 6. dez, 1962.

VILCHES, Lorenzo. **A migração digital**. São Paulo: Loyola, 2003.