O projeto Media Cloud Brasil:

Uma análise do tratamento de informações em ambientes de *big data*.

Renato Rocha Souza
FGV - Escola de Matemática Aplicada

Flávio Codeço Coelho
FGV - Escola de Matemática Aplicada

Alexandre Gonçalves
Center for Civic Media - Massachussets Institute of Technology

Evandro Dalben Lopes
FGV - Escola de Matemática Aplicada

Elisa Mussumeci Bianor dos Santos
FGV - Escola de Matemática Aplicada

Victor da Fonseca Machado Junqueira
FGV - Escola de Matemática Aplicada

Introdução

Diversos teóricos procuram abarcar, em suas análises, o fenômeno da concretização de previsões sobre uma "sociedade da informação", ou "do conhecimento", em que a maior força motriz para geração de bens comuns está baseada na informação e nos diversos sistemas especialistas e mediáticos que a manipulam ou dela dependem (GIDDENS, 1991; LEVY, 1993 e 1999; CASTELLS, 1999; TAKAHASHI, 2000; MATTELART, 2002). Os sistemas de informação e de comunicação permeiam e viabilizam virtualmente a grande maioria das atividades humanas, e não mais podemos conceber a sociedade sem sua acentuada imbricação com as tecnologias de informação que nela surgem e a modificam. Acompanhando o desenvolvimento dessas tecnologias, os repositórios de informações que são produzidos durante o desempenho das inúmeras atividades humanas vêm migrando para o ambiente *on-line*, de forma que os registros das produções intelectuais humanas estejam cada vez mais em formatos digitais, acessíveis através de redes e sistemas de computadores.
Em documento recente (IDC, 2012) o IDC mensura e projeta um crescimento da ordem de 30.000% nas taxas mundiais de criação de informação digital entre os anos de 2005 e 2010, aumentando as preocupações sobre sua governança. Para suprir a necessidade de registrar e recuperar as informações, criadas continuamente em ritmos

vertiginosos, e atender a demanda por essas informações, são necessárias metodologias e técnicas inovadoras para manusear grandes acervos de documentos em diversas mídias.
Nesse contexto, faz-se necessário o desenvolvimento de metodologias – e tecnologias associadas – para se enfrentarem os muitos desafios que surgem quando lidamos com massivas quantidades de dados textuais, como nas bibliotecas e arquivos digitais, ou na *World Wide Web*, notadamente quando estes precisam ser regularmente organizados e pesquisados, visando à recuperação em tempo hábil de informações relevantes para algum objetivo específico.
Na academia, os impactos na pesquisa são evidentes, mais notadamente em áreas como a Ciência da Informação e a Ciência da Computação que, além de experimentarem uma grande convergência nas últimas décadas, também têm buscado subsídios em áreas como a Linguística Computacional, Filosofia e Matemática Aplicada, além de suas subáreas como a Visualização de Informação e Bancos de Dados. A pesquisa tem adotado a abordagem de *data driven research*, em que se fazem inúmeros testes de hipóteses em grandes massas de dados como preâmbulos exploratórios e, na medida em que se encontram caminhos promissores derivados dos dados, busca-se aprofundar quantitativa e qualitativamente através de mudanças de focos.
Neste capítulo, apresentamos a pesquisa realizada no escopo do Projeto Media Cloud - Brasil, uma parceria entre a Escola de Matemática Aplicada da Fundação Getulio Vargas, o Massachussets Institute of Technology e a Universidade de Harvard.

O Projeto Media Cloud

O projeto Media Cloud foi lançado em Março de 2009 através de uma parceria entre o MIT Center for Civic Media[1] e o Berkman Center for Internet and Society[2], da Harvard Law School. Constitui-se de uma plataforma para estudar ecossistemas de mídia, ou seja, as relações entre as instituições e os profissionais criadores de mídia - impressa e digital - e os cidadãos. Através do monitoramento programático de milhões de notícias, publicadas online ou transmitidas em canais de televisão, o sistema permite aos pesquisadores monitorar a disseminação de notícias, conceitos e memes, além de permitir a descoberta das redes de atores que pautam a mídia, através da genealogia das notícias. Também permite que se façam análises geográficas - através da identificação da cobertura midiática nas diversas regiões, e políticas, monitorando a abordagem específica dos diversos atores, politicamente identificados com as principais correntes partidárias, segundo os diversos temas de interesse, identificando-se vieses. A plataforma de software é livre, e foi projetada como um substrato para diversos projetos correlatos, centrados em processos de comunicação.

[1] http://civic.mit.edu/media-cloud
[2] http://cyber.law.harvard.edu/research/mediacloud#

Como ferramenta computacional, pode-se caracterizar o Media Cloud como uma ferramenta *open source* de análise de conteúdo, que desempenha 5 funções básicas:

- Definição de fontes de mídia de interesse;
- Captura (*crawling*) contínuo de notícias
- Extração de características semânticas dos textos - análise de assunto, análise de conteúdo
- Construção de estatísticas sobre frequências de palavras através de processamento de linguagem natural
- Análise dos resultados

Para estas análises, utilizam-se recursos de visualização de informação, georreferenciamento, dentre outros.
A característica de *Big Data* do projeto pode ser observada nos volumes de dados que são continuamente capturados. Atualmente, são mais de 8 mil fontes de informação, desde grandes veículos da imprensa (como o New York Times) ou blogs de organizações, indústria ou indivíduos, cada uma destas fontes contendo, possivelmente, muitas centenas de *feeds*[3]. As fontes são divididas em categorias, como explicitado a seguir:

- Blogs políticos[4]: os 1000 blogs mais influentes dos Estados Unidos, ranqueados de acordo com o blog *Technorati*[5]. Estes são posteriormente classificados em:
    - Blogs políticos de Esquerda
    - Blogs políticos de Centro
    - Blogs políticos de Direita
- Blogs populares[6]: os 1000 blogs mais importantes, segundo o blog *Bloglines*[7].
- Os 25 maiores veículos de mídia[8], segundo o Google AdPlanner[9]
- Todas os feeds gerados pela Casa Branca[10]

Novas fontes estão sendo constantemente acrescentadas, e diversas análises são disponibilizadas através do sítio online[11], como as que são apresentadas em seções a seguir. A interface principal de consultas ao banco de dados do Media Cloud, aberta ao

[3] http://pt.wikipedia.org/wiki/Feed
[4] http://www.mediacloud.org/dashboard/media/1?media_sets_id=7125
[5] http://technorati.com/
[6] http://www.mediacloud.org/dashboard/media/1?media_sets_id=26
[7] http://www.bloglines.com/
[8] http://www.mediacloud.org/dashboard/media/1?media_sets_id=1
[9] http://www.google.com/adplanner/static/top1000/
[10] http://www.whitehouse.gov/
[11] http://www.mediacloud.org/dashboard/view/1?q1=94946

público em geral, pode ser acessada através de endereço na web e conferida na FIG.1 a seguir:

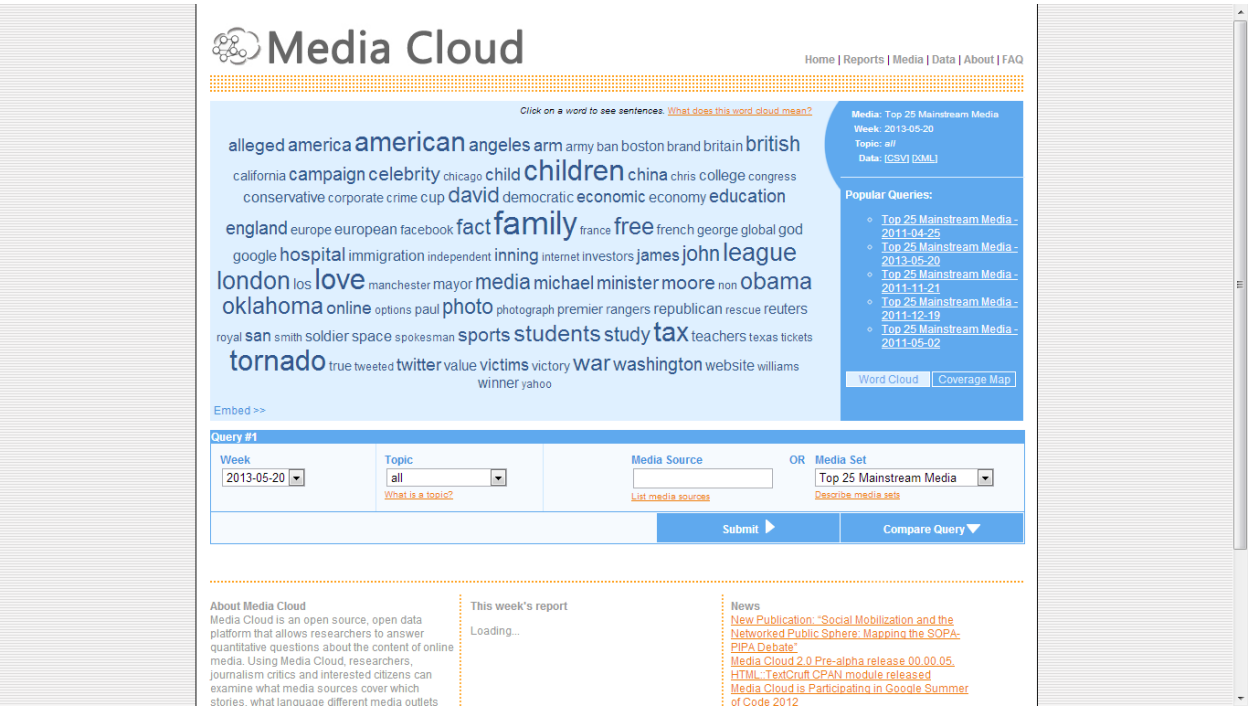

Figura 1: Interface de consulta do banco de dados do Media Cloud (EUA). Fonte: http://www.mediacloud.org/dashboard/view/1?q1=94946

O Projeto Media Cloud Brasil

O Projeto "Media Cloud Brasil – Panóptico de Fluxos Textuais na Sociedade Brasileira" foi desenhado tendo como inspiração o Projeto Media Cloud do MIT, e como objetivo:

> "...a estruturação de um processo contínuo de captação de uma miríade de dados de cunho textual (jurídicos, legislativos, midiáticos, acadêmicos, etc.) visando à realização de análises de cunho acadêmico, semântico, estatístico e orgânico, permitindo a construção de modelos panópticos para percepção e inferência sobre a conjuntura brasileira e realização de análises preditivas.".

No tocante à tecnologia, propunham-se os seguintes resultados:

- Indexação de texto completo de grandes coleções de documentos, a serem armazenados em servidores da FGV, constituindo material para análises e pesquisas;
- Disponibilização de aplicação *web* para realização de consultas nas coleções disponíveis;

- Disponibilização de análises e Incorporação das funcionalidades da ferramenta PyPLN de engenharia textual, oferecendo capacidades analíticas avançadas baseadas em processamento de linguagem natural;
- Disponibilizar API para automatização do gerenciamento de coleções, consultas às mesmas e tarefas analíticas customizadas.

Em relação aos produtos de pesquisa, que abrangem áreas como Ciência da Informação, Linguística, Sociologia, Direito e Política, podemos citar:

- Análises de pautas de mídia e seus atores;
- Análise de impacto de notícias;
- Análises de rede de atores;
- Estudos sobre a genealogia de conceitos;
- Análise de evolução cultural através de gradientes terminológicos;
- Análise lexicográficas históricas;
- Análise semântica do discurso;
- Criação semiautomática de taxonomias e ontologias através da extração de conhecimento de textos;
- Classificação semiautomática de textos;
- Predição de conjunturas através de textos;
- Análise de impacto de redes sociais na mídia tradicional.

Cronologia

Em Outubro de 2012, a partir de contato com o Centro de Tecnologia e Sociedade da Escola de Direito da Fundação Getulio Vargas (FGV), foi estabelecida uma parceria interna entre as Escolas de Matemática Aplicada e a Escola de Direito, para buscar replicar e ampliar o escopo do projeto Media Cloud no estudo do panorama da mídia brasileira. Com o apoio financeiro da presidência da FGV, estruturou-se ambiente tecnológico para o empreendimento. A partir do lançamento do projeto Media Cloud BR, no início de 2013, os pesquisadores brasileiros Flávio Codeço Coelho e Renato Rocha Souza realizaram visitas às sedes do *Center for Civic Media* , no *Massachusets Institute of Technology*, e do *Berkman Center for Internet and Society*, na *Harvard University*, para estabelecer as bases da parceria. internacional. Nesta ocasião foram realizadas exposições dos propósitos dos projetos, alinhamento das características comuns e comparações entre as diferentes tecnologias adotadas, assim como entre as conjunturas econômica e política dos dois países. Estabeleceu-se parceria que prevê o intercâmbio de dados, de tecnologias e de pesquisadores nos anos vindouros, no sentido de análise conjunta dos resultados e aprimoramento mútuo do processo. Em Julho de 2013 foi estabelecida a plataforma tecnológica para captura de *feeds* nacionais para realização do mapeamento do panorama da mídia brasileira, assim

como a estruturação de uma plataforma específica - PyPLN - para processamento de linguagem natural e análise dos *corpora* textuais a serem gerados. Em Agosto deste mesmo ano, a Escola de Matemática Aplicada da FGV recebeu o pesquisador Alexandre Gonçalves,[12] do MIT, que, junto aos bolsistas de iniciação científica Elisa Mussumeci e Victor Fonseca, iniciaram a classificação manual dos milhares de sítios web nacionais identificados como relevantes para análise. Cada conjunto de feeds virá a constituir uma coleção em banco de dados para processamento pela plataforma PyPLN

A plataforma analítica PyPLN

Em Março de 2012 os pesquisadores da Escola de Matemática Aplicada iniciaram o desenvolvimento de uma plataforma completa de processamento de linguagem natural denominada "PyPLN"[13], em código aberto, e disponível para a comunidade. Esta plataforma, 100% aberta e gratuita, foi desenvolvida em uma linguagem de programação não proprietária - Python - orientada a objeto e adotada em cursos de computação das principais universidades do mundo, o que garante sua acessibilidade e benefício para a comunidade. A ferramenta conta atualmente com módulos para execução das seguintes tarefas:

- Conversão de formatos (PDF, HTML, RTF);
- Estatísticas textuais básicas (contagem de palavras, contagem de sentenças, repertório);
- Construção de histogramas de frequência palavras;
- Marcação morfossintática dos textos.

E, no momento, estão sendo implementadas as seguintes funcionalidades:

- Extração de informação para indexação automática;
- Análise de sentimentos;
- População semiautomática de ontologias;
- Agrupamento (*cluster*) de documentos;
- Extração de entidades nomeadas.

A ideia é que a arquitetura seja modular, permitindo que novas funcionalidades sejam agregadas por indivíduos que desejem contribuir. O código pode ser examinado e clonado do site do projeto no *GitHub*[14], além de se poderem acessar às funcionalidades

[12] http://civic.mit.edu/users/alexandre-goncalves
[13] http://pypln.org/
[14] https://github.com/NAMD/pypln.api

através dos servidores de demonstração[15] da FGV, cuja tela de acesso aparece nas FIG.2 e FIG.3. a seguir:

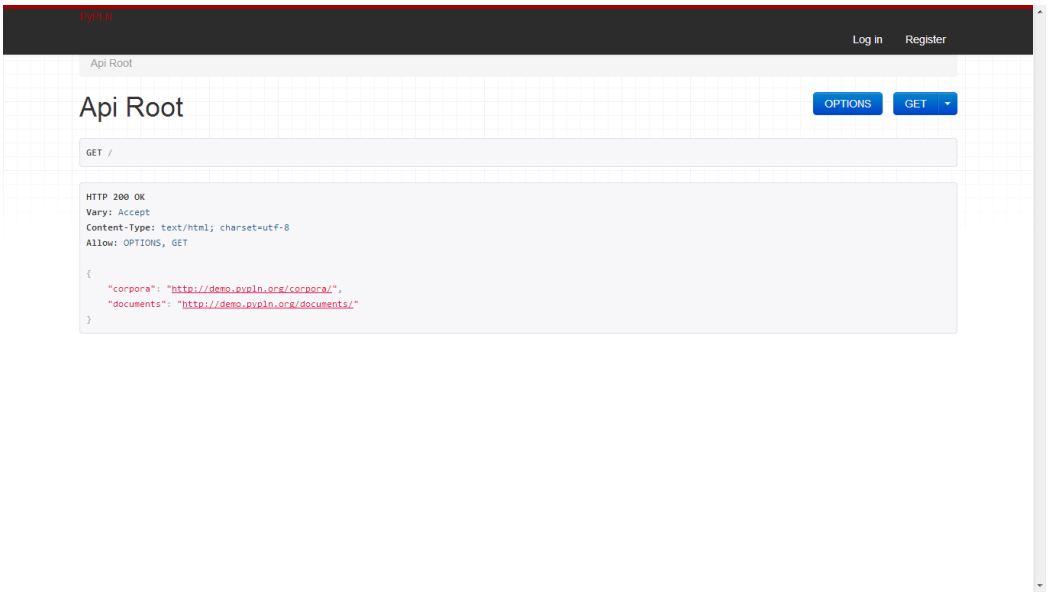

Figura 2: Interface do PyPLN para programadores.

Figura 3: Interface do PyPLN para usuários.

O acesso pode ser feito através de uma interface web ou através de chamadas à própria API, embutidas em programas de usuários.

[15] http://demo.pypln.org/ e http://fgv.pypln.org/

No escopo deste projeto, foi estabelecida parceria com o LEEEL[16] (Laboratório de Estudos Empíricos e Experimentais da Linguagem) da Escola de Letras da UFMG, departamento de Linguística, para produção científica e parceria no desenvolvimento de aplicativos e software para a comunidade de Processamento de Linguagem Natural. No escopo desta parceria, foi comprada uma licença do marcador morfossintático PALAVRAS[17], para tratamento do português, que já foi incorporado à plataforma PyPLN mencionada.

Alguns resultados da pesquisa

Embora o projeto seja bastante recente, e, termos de Brasil, a contraparte americana já possui um longo histórico de realizações. Nas subseções a seguir, são apresentados alguns dos projetos derivados da plataforma Media Cloud.

Traçando a Origem de Notícias na Mídiasfera

Um dos estudos de maior impacto realizado a partir do acervo do Mediacloud foi o acompanhamento da mobilização social em torno do projeto de lei americano SOPA-PIPA (BENKLER et al. 2013)[18].
O projeto de lei SOPA (Stop Online Piracy Act)-PIPA(Protect Intelectual property Act) foi introduzido no congresso Americano, em setembro de 2010 como um projeto bi-partidário para dar direitos ao Departamento de Justiça Americano a fechar unilateralmente web-sites envolvidos com a pirataria de filmes e música assim como o comércio de bens contrabandeados. Este projeto gerou intensa rejeição popular que acabou desencadeando a sua retirada da pauta de votações do congresso 17 meses mais tarde.
BENKLER et al., valeram-se de um conjunto de 9757 artigos publicados entre setembro de 2010 e Janeiro de 2012. Estes artigos foram extraídos do banco de dados do Mediacloud, selecionados através de busca pela sigla do projeto de lei (SOPA, PIPA) e complementados com outros artigos linkados a partir dos comentários dos artigos previamente selecionados.
Esta análise permitiu a caracterização do processo de amplificação pela mídia de um debate originalmente iniciado por uma comunidade muito restrita de técnicos e ativistas.

[16] http://www.letras.ufmg.br/CMS/index.asp?pasta=leel&path=20101229104322.asp&title=Equipe
[17] http://beta.visl.sdu.dk/constraint_grammar.html
[18] http://cyber.law.harvard.edu/research/mediacloud/2013/mapping_sopa_pipa/

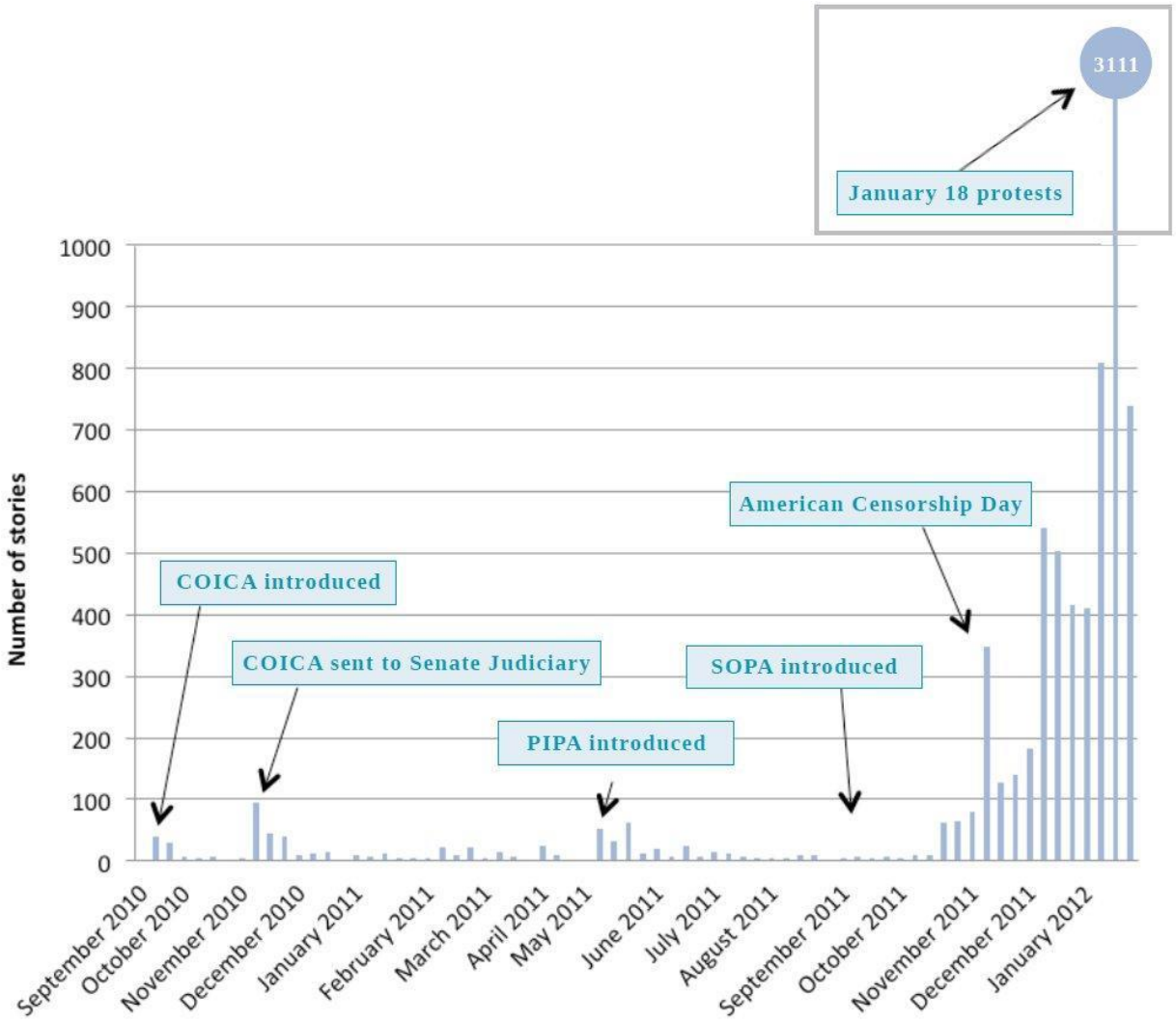

Figura 4: Evolução do número de artigos publicados durantes o período entre a apresentação do projeto de lei, e a sua remoção. Fonte: Benkler et al., 2013.

Mapeando o Globo

O Projeto Mapeando o Globo (Mapping the Globe) busca construir mapas e visualizações interativas que buscam auxiliar o entendimento sobre quais assuntos são cobertos pelo jornal Boston Globe, como pode ser visto na FIG.5. A importância do mapeamento de aonde a mídia foca sua atenção, tanto em termos de quantidade quanto de qualidade, é a possibilidade de entender as regiões que são cobertas e os tipos de evento para os quais se direciona atenção. Estas inormações podem ser cobertas com bancos de dados sobre população, renda e crime.

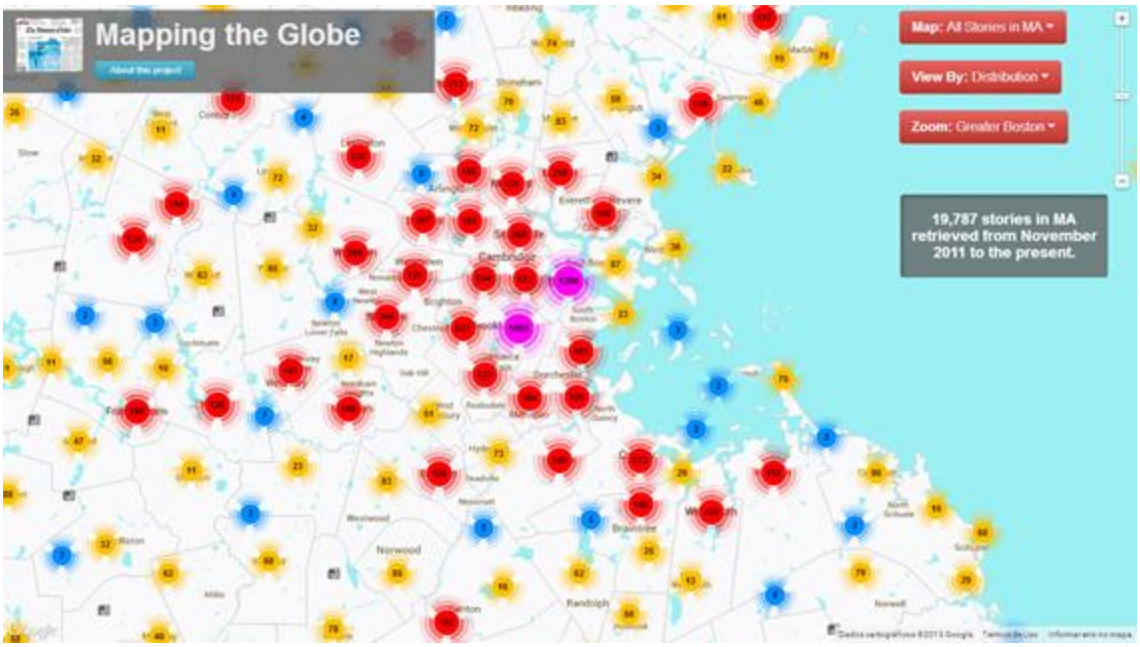

Figura 5: Interface do Projeto Mapping the Globe. Fonte: http://globe.mediameter.org/

No Projeto Media Cloud BR, uma iniciativa semelhante, denominada “Justice Cloud”. Nesta vertente, buscamos mapear os processos jurídicos, tais como detalhados no sítio do Trubunal de Justiça do Rio de Janeiro (TJRJ). Este mapeamento permitirá identificar o perfil de litigância no estado, e tecer inferências correlacionais, a partir das análises semânticas dos tipos de processos. A interface do demonstrador online - ainda em construção - pode ser vista na FIG 6.

Figura 6: Interface do Projeto Justice Cloud. Fonte: http://mbjcloud.appspot.com/

Considerações e trabalhos futuros

O projeto Media Cloud, e sua plataforma analítica PyPLN, prometem contribuir sobremaneira para o estado da arte das técnicas de análise de assunto e de conteúdo, tanto pelas tecnologias inovadoras para tratamento de grandes massas de dados textuais, quanto pelas confluências interdisciplinares, que agregam, em termos de teoria e metodologia, os campos da Ciência da Informação, Ciência da Computação e Linguística; e em termos de contexto e escopo, o arcabouço das ciências sociais, como Direito, Comunicação, Política e Sociologia. O material gerado constituirá um farto material para as pesquisas nestas áreas, em décadas a porvir.

Referências

BENKLER, Yochai, Hal Roberts, Robert Faris, Alicia Solow-Niederman, and Bruce Etling. "Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate." Berkman Center Research Publication 2013-16 (2013).

CASTELLS, M. A Sociedade em Rede. São Paulo: Paz e Terra, 1999. 617p.

GIDDENS, A. As Conseqüências da Modernidade. São Paulo: Ed. Unesp, 1991.

IDC. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. 2012. Disponível em: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
Acesso em: Set. 2013.

LÉVY, Pierre. As Tecnologias da Inteligência: o futuro do pensamento na era da informática. São Paulo: Editora 34, 1993. 203p.

LÉVY, Pierre. Cibercultura. São Paulo: Editora 34, 1999. 260p.

MATTELART, Armand. História da sociedade da informação. São Paulo: Loyola, 2002.

TAKAHASHI, Tadao (org). Sociedade da informação no Brasil: Livro Verde. Brasília: Ministério da Ciência e Tecnologia, 2000.