NARRATIVAS AUTOMATIZADAS: um experimento de produção de conteúdo jornalístico através de software[1]

AUTOMATED NARRATIVES: an experiment to produce journalistic content through software

Márcio Carneiro dos Santos[2]

Resumo: Descreve-se o experimento de construção de um software capaz de gerar leads e títulos jornalísticos de forma automatizada a partir de informações obtidas na internet. A possibilidade teórica já prevista por Lage no final do século passado baseia-se na estrutura simples e relativamente rígida desse tipo de construção narrativa, o que facilita a representação ou tradução da sua sintaxe em termos de instruções que os computadores possam executar. Discute-se também as relações entre sociedade, técnica e tecnologia, fazendo um breve histórico sobre a introdução das soluções digitais nas redações jornalísticas e seus impactos. O desenvolvimento foi feito com a linguagem de programação Python e a biblioteca NLTK- *Natural Language Toolkit* – e usou os resultados do Campeonato Brasileiro de Futebol de 2013 publicados em portal da internet como fonte de dados.

Palavras-chave: Narrativas Automatizadas. Jornalismo online. *Python*. Inteligência artificial. NLTK.

Abstract: This paper describes the experiment of building a software able to generate leads and newspaper titles in an automated manner from information obtained from the Internet. The theoretical possibility Lage already provided by the end of the last century is based on simple and relatively rigid structure of this kind of narrative construction, which facilitates the representation and translation of its syntax in terms of instructions that computers can perform. We also discuss the relationship between society, technics and technology, making a brief history of the introduction of digital solutions in newspaper newsrooms and their impacts. The development was done with the Python programming language and library-NLTK Natural Language Toolkit - and used the results of the Brazilian Football Championship of 2013 published in the internet portal as a data source.

Keywords: Automated narratives. Online journalism. Python. Artificial Intelligence. NTLT.

:::

[1]Trabalho apresentado no I Seminário de Pesquisa em Jornalismo Investigativo, realizado na Universidade Anhembi-Morumbi, cidade de São Paulo, entre 24 e 26 de julho de 2014.

[2]Professor Assistente do Departamento de Comunicação Social da UFMA na área de Jornalismo em Redes Digitais. Jornalista, Mestre em Comunicação pela UAM-SP e Doutor pelo programa de Tecnologias da Inteligência e Design Digital (TIDD) da PUC-SP. Coordenador do Laboratório de Convergência de Mídias – LABCOM (www.labcomufma.com) . Email: mcszen@gmail.com.

1 Introdução – o medo e o fascínio das máquinas

Apesar de simplificadora, a versão dualista das relações entre homens e tecnologia ainda hoje é utilizada. Problematizando a neutralidade da técnica no mundo, que existiria apenas como instrumento impessoal, moldado pelas mãos de seu utilizador, fica difícil não pensar nas complexas imbricações entre os aspectos culturais, econômicos, sociais e tecnológicos, em dado momento histórico e lugar. Tal situação nos leva a um amplo espectro de possibilidades com perspectivas naturalísticas, humanísticas, críticas e tecnicistas, incluindo ai as posições mais extremadas dos que temem ou defendem as soluções tecnológicas.

Sejam prometéicos ou fáusticos (RÜDIGER, 2007), apocalípticos ou integrados (ECO, 2006), ciberiluministas ou neoluditas[3], muito esforço tem sido dedicado por áreas como a da Filosofia da Tecnologia e afins para discutir a questão, que se inicia com o conceito de técnica.

É importante ressaltar as diferenças entre técnica e tecnologia. Enquanto a primeira já fazia parte das discussões dos filósofos gregos, a última efetivamente começa a constituir-se, ainda que de forma embrionária, no Renascimento, a partir da junção da ciência aplicada e do objetivo, naquele momento, cada vez mais claro, de dominar a natureza a partir da razão.

Para entender a diferença é preciso voltar cerca de cinco séculos antes do início da era cristã. A *tekhnè* dos gregos segundo Lemos (2002) estava intimamente ligada às ações práticas cobrindo uma ampla faixa de atividades que iam dos ofícios mais simples baseados em trabalhos manuais até às artes e à medicina. Era *tekhnè*, portanto, tudo aquilo produzido pela ação do homem num contraponto ao que era gerado pela natureza.

Essa primeira dicotomia na Grécia de Platão e Aristóteles trazia um julgamento de valor bem definido: o fazer da natureza era superior porque permitia a possibilidade de gerar a si mesmo, de atravessar a fronteira entre a ausência e a presença, de forma independente. A

[3]A ideia do ciberiluminismo está relacionada à visão extremamente positiva e às vezes até ingênua sobre a relação entre tecnologia e seres humanos, representada normalmente por suas características inovadoras, gerando transformações capazes de criar um mundo melhor e mais justo. Já o ludismo vem de Ned Ludd, supostamente, operário que liderou um movimento que pregava a destruição das máquinas nas tecelagens inglesas porque essas reduziam os postos de trabalho. Alguns autores citam Ludd como um personagem criado pelo movimento operário da época para facilitar a propagação da campanha contra a automatização do processo fabril no início da Revolução Industrial.

herança divina e, por isso mais pura, fazia da *phusis*, o princípio da geração das coisas naturais, superior à *tekhnè*, sempre algo inferior, sem a capacidade da *auto-poièses*, ou seja, da autoreprodução.

Se as origens da técnica repousam na antiguidade, o conceito de tecnologia veio bem depois. Ensina-nos Lemos (2002) que a tecnologia é a técnica moderna, muito distante do imaginário da antiguidade e liberta dos seus laços com o divino. Pelo contrário, é a técnica que, baseada na razão e no desenvolvimento científico, na física newtoniana, na matemática cartesiana e no empirismo, transforma a natureza em "objeto de livre conquista" (LEMOS, 2002, p. 45).

Para Rüdiger (2007, p. 175) "a técnica é, em essência, uma mediação do processo de formação da vida humana em condições sociais determinadas". Já a tecnologia é

> o conhecimento operacional que designamos pelo termo técnica enquanto se articula com a forma de saber que chamamos ciência, através da mediação da máquina e, potencialmente, em todas as áreas passíveis de automatização, conforme define o tempo que a criou, a Modernidade (RÜDIGER, 2007, p. 186).

Se para Heidegger, a técnica é um modo de existência do homem no mundo, a partir da modernidade, esse existir tomará um rumo direto de agressão à natureza, agora sujeita ao conhecimento humano e à ideia de um progresso linear, constante e que não podia ser parado. Para muitos, como Sennett (2009), abre-se aqui a caixa de Pandora, a deusa da invenção enviada por Zeus à terra e que para os gregos representava também a cultura das coisas produzidas pelo homem, através das quais este poderia causar danos a si mesmo.

Os grandes conflitos mundiais da primeira metade do século XX, o nazismo e o pesadelo da guerra fria e da ameaça nuclear materializaram os piores sonhos dos gregos num mundo que, em tese, deveria ser mais evoluído justamente pela existência da tecnologia.

Nos últimos três séculos a Filosofia da Ciência ocupou muitos pensadores, mas só no século XX, a partir de eventos como a bomba atômica em Hiroshima e Nagasaki e posteriormente as preocupações com as mudanças climáticas, a poluição gerada pelo desenvolvimento industrial[4] e a manipulação genética com a possibilidade, mesmo que

[4]Em janeiro de 2013 a poluição em Pequim chegou a 25 vezes do valor máximo aceitável para o ser humano, gerando inclusive um protesto que constituía-se na venda de latinhas de ar na cidade. Edição do Jornal Nacional – TV Globo – 29/01/2013.

teórica, da clonagem de seres humanos que esse cenário começou a mudar e a produção sobre uma Filosofia da Tecnologia passou a tomar corpo.

A intensidade e a velocidade das mudanças econômicas e sociais nas últimas décadas, de alguma forma, ligadas ao desenvolvimento tecnológico, deram a esse campo um interesse com crescimento exponencial e uma diversidade de correntes e enfoques.

As possibilidades vão do determinismo tecnológico e sua versão radical da tecnologia autônoma de Ellul (1968), que de forma geral coloca os homens à mercê da tecnologia, até versões opostas, como as que pregam a construção social da tecnologia, definida não por parâmetros fora do controle humano mas, pelo contrário, a partir da interação de vários grupos de interesse que definem as linhas gerais do seu desenvolvimento.

Nomes como Heidegger, Arendt e Marcuse representam uma visão crítica do problema, com escritos nem sempre de fácil leitura. Segundo Dusek (2006) há variações para todos. Linguistas anglo-americanos, neo-marxistas, fenomenologistas europeus, existencialistas, hermeneutas, representantes do pragmatismo americano e filósofos pós-modernos como Deleuze, Virilio e mais recentemente Bruno Latour, focalizaram seus olhares sobre a relação entre o homem e a tecnologia, transformando uma temática pouco valorizada em algo com uma centralidade quase inevitável.

Em 1976 foi fundada a Sociedade para a Filosofia e a Tecnologia (SPT), segundo sua própria página pública na internet (THE SOCIETY FOR PHILOSOPHY AND TECHNOLOGY, 2014), uma organização internacional independente que estimula, dá suporte e intermedia discussões filosóficas relevantes sobre tecnologia.

As possibilidades de pensar as relações entre sociedade e tecnologia deram origem a novos campos como o que hoje conhecemos como *Science and Technology Studies* (STS). Nele pensadores como Castells (1999) e Andrew Feenberg (2002) têm se dedicado a formular um cenário compatível com os desafios de estudar uma relação obviamente multifacetada e complexa.

Em sua crítica a visões simplistas sobre o papel da tecnologia no mundo de hoje, Feenberg nos propõe inicialmente uma espécie de mapeamento das posições normalmente apresentadas e a partir delas tenta incorporar questões como democracia, poder e liberdade, como fatores também importantes a considerar nas discussões dos STS.

Na cartografia de Feenberg (2002) das sociedades modernas a tecnologia ocupa um lugar de destaque entre as fontes de poder que se articulam no meio social. Para ele, as decisões políticas que definem muitos dos aspectos da nossa vida cotidiana são direcionadas pela influência dos controladores dos sistemas técnicos, sejam eles das grandes corporações, militares ou de associações profissionais de grupos como físicos, engenheiros e mais recentemente, poderíamos sugerir também, desenvolvedores de software.

Ao fazer tal constatação o autor se remete ao pensamento de Marx que já no século XIX criticava a ideia de que a economia pudesse ser apenas regida por fatores extrapolíticos, através de leis com a da oferta e da procura. Do mesmo modo, imaginar o papel da tecnologia sem avaliar as diversas relações que ela estabelece com a sociedade pode implicar numa visão reduzida do problema.

Num caminho semelhante à crítica marxista de uma economia regulada por uma ordem natural e inexorável, Feenberg (2010) relativiza a racionalidade da tecnologia a partir da constatação de que sua gênese e desenvolvimento acontecem no mundo dos homens e, por isso, também são influenciadas por ele.

> Criação técnica envolve interação entre razão e experiência. Conhecimento da natureza é necessário para fazer um equipamento que funcione. Este é o elemento da atividade técnica que consideramos como racional. Mas o equipamento deve funcionar num mundo social e as lições da experiência nesse mundo influenciam o design (FEENBERG, 2010, p. 17)[5].

Se no campo da Filosofia é amplo o debate, o cinema ao longo de décadas tem traduzido esse imaginário de medo e fascínio em diversos filmes onde as soluções tecnológicas são representadas por robôs, autômatos, máquinas e até sofisticados programas de computador. Naves controladas por entidades automatizadas que se rebelam contra os humanos, como o computador HALL 9000 em "2001 – Odisseia no Espaço" de Kubrick (1968); que decretam sua extinção como em "Exterminador do Futuro" de James Cameron (1984) ou ainda que os escravizam, num mundo digitalmente criado, a "Matrix", para

[5]"Technical creation involves interaction between reason and experience. Knowledge of nature is required to make a working device. This is the element of technical activity we think of as rational. But the device must function in a social world, and the lessons of experience in that world influence design." – Tradução nossa.

utilização da humanidade como simples fonte de energia, dos irmãos Wachowski (1999), são apenas alguns dos inúmeros exemplos que poderíamos citar.

Na série de TV *Star Trek: The next generation* , que também ganhou os cinemas (*Star Trek - First Contact*, Jonathan Frakes, 1996), uma das piores ameaças alienígenas já enfrentadas foi a dos *Borgs*, raça de seres híbridos, biológicos e maquínicos, que rapidamente assumiam o controle das áreas que invadiam, a partir da conversão dos seres que encontravam em sua própria espécie, através da inserção de implantes que faziam as vítimas completamente integradas ao comando central, agindo como uma colônia de insetos, num exército cada vez maior.

Por outro lado, o fascínio pelas máquinas é muito anterior à quase inevitável dependência contemporânea que estabelecemos com celulares, *smartphones*, *tablets* e tantos outros *gadgets* tecnológicos dos quais não queremos mais nos separar.

Se na antiguidade e na idade média os relatos sobre autômatos eram restritos, o século XVIII é considerado sua época áurea. No trecho de Devaux (1964) é descrita a apresentação de uma dessas peças que ainda hoje podem ser vistas em Paris, a "Tocadora de Xilofone" de Roentgen, uma boneca musicista que se supõe tenha sido inspirada na figura de Maria Antonieta[6].

> Numa sala do Palácio de Versalhes, entre as saias de balão e os vestidos da corte, o exímio automatista Roentgen, apresenta a Luís XVI outra obra-prima. Aquela Tocadora de xilofone, de corpete decotado e vestido de seda bordada, provoca a curiosidade geral; fala-se do corpo da boneca divinamente modelado debaixo do vestido, pasma-se com a precisão e a sua altiva graciosidade. Uma pequena ária muito viva saltita desperta pela saraivada dos martelos de marfim; todo um século se desprende dessa música elegante e seca; e quando a jovem, debaixo dos anéis do seu penteado alto, volta a cabeça para saudar, a semelhança levanta um sussurro em toda a sala... Mais tocante e mais perfeita no mecanismo que o Escrivão ou a Musicienne, dos Jaquet-Droz, ou o Pato, de Vaucanson, e os seus Flautistas, a Tocadora de xilofone evoca fielmente, quanto a nós, a Primeira Idade do Automatismo (DEVAUX, 1964, p. 7).

[6]A história da boneca pode ser conhecida no documentário "*L'Androïde de Marie-Antoinette*", disponível em: http://www.youtube.com/watch?v=pSxWmJLAaEg .

Figura 1 - Tocadora de Xilofone de Roentgen, restaurada em 1864 por Robert Houdain.

Fonte: Lutice Créations ([200-]).

2 Jornalismo e Tecnologia

Desde os primórdios, o jornalismo esteve ligado a algum tipo de tecnologia, sendo o processo de impressão de Gutemberg e seu desenvolvimento um dos principais fatores que alavancou a expansão dessa atividade.

Muito tempo depois, já no final do século XX, a chegada das redes, da internet e dos computadores às redações iniciou um ciclo de profundas mudanças que até hoje está em andamento e que alguns como Soria (2014) descrevem simplesmente como um *tsunami*, traduzindo o impacto devastador que positiva e negativamente a digitalização de grande parte do processo de produção jornalística tem causado.

Machado (2003), ao descrever o início da mudança, nos ensina que duas posições se estabeleceram para compreender o que estava acontecendo. A primeira, que poderíamos chamar de instrumentalista, entendia que computadores eram apenas mais uma ferramenta a disposição dos jornalistas, artefatos adicionais a serem utilizados na execução do seu trabalho, como antes haviam também sido as inovações do telégrafo, da máquina de escrever e do telex, entre outras.

Já na segunda forma de entender a transformação, a chegada do digital representava uma alteração muito mais extensa, capaz de impactar todas as etapas do processo de produção como também as habilidades necessárias para exercer a função de jornalista, os modelos de negócio dessa cadeia produtiva e os próprios papeis desempenhados tradicionalmente por emissores e receptores em relação aos veículos de massa.

> A falta de clareza sobre as consequências para o jornalismo da disseminação do suporte digital dificulta a compreensão plena das particularidades da prática jornalística nas redes, das mudanças no perfil do profissional, na estrutura organizacional das empresas jornalísticas e das funções que o usuário passa a ocupar no sistema de produção de conteúdos (MACHADO, 2003, p. 2).

Bradshaw e Rohumaa (2011), no histórico que traçam sobre o início do jornalismo *online* no ocidente, indicam os britânicos Today de 1986 como o precursor na produção de conteúdo usando tecnologia digital e o *Daily Telgraph* como um dos primeiros jornais impressos a ter sua versão transposta[7] na ainda pouco conhecida internet de 1994. Era o *Eletronic Telegraph*. Em 1993 o primeiro *browser*[8], o *Mosaic*, havia sido lançado. *BBC Online* em 1997 e *Guardian Unlimited* em 1999 são destaques numa lista de iniciativas ligadas à ideia de levar a atividade jornalística para o ciberespaço.

Na época havia uma grande confiança entre as empresas de mídia de que a internet seria apenas mais um veículo, mais um espaço a conquistar, uma nova fronteira, onde vendendo publicidade e utilizando os modelos de negócio tradicionais todos poderiam prosperar. Tal certeza atraiu muitos investimentos e novas empresas "*dot.com*" surgiram rapidamente mas, em 2001, com a crise que ficou conhecida como o estouro da bolha da internet, percebeu-se que não era tão simples assim lucrar com as iniciativas digitais e que, pelo contrário, o que estava começando era uma corrida pela sobrevivência, onde apenas os que se adaptassem ao novo cenário de forma mais eficiente poderiam continuar.

Passaram-se os anos e novas tecnologias foram continuamente sendo incorporadas ao fazer jornalístico. As bases de dados, a integração de múltiplas mídias para contar uma única história, a capacidade de customizar e segmentar o conteúdo em função dos interesses de usuários cada vez mais exigentes e difíceis de atrair. Surgiram os sistemas de CMS (*Content Management Systems*) que permitiram aos jornalistas publicar diretamente seu conteúdo sem a intermediação de um programador ou especialista em HTML, a linguagem que organiza os elementos de qualquer página na web e que os *browser*s utilizam para construir o que os leitores veem em seus computadores.

[7] Mielnickzuk (2001) nos fala das fases de jornalismo digital, chamando a primeira de fase transpositiva, justamente porque o conteúdo do impresso era apenas copiado para a internet sem grandes alterações.

[8] Um *browser* é um software cliente de internet que solicita ao servidor as páginas que o usuário assinala através do endereço.

Chamar os tradicionais consumidores de notícias de leitores também não é mais tão preciso. A digitalização, o barateamento dos equipamentos para produzir imagens e som, a expansão da infraestrutura da internet e a ubiquidade dos dispositivos móveis fez dos cidadãos produtores de conteúdo, dando a eles um espaço crescente no processo de produção jornalística e constituindo o que alguns chamam de *user generated content* (UGC) , conteúdo gerado por usuários e também de jornalismo participativo, termo que traduz uma série de iniciativas com escopo e dimensão diversos[9], indo do jornalismo produzido por ou para pequenas comunidades até grandes iniciativas que, via internet, ganham alcance internacional.

O impacto da tecnologia no jornalismo também obrigou a revisão de alguns conceitos clássicos como o da pirâmide invertida e do lead. A necessidade da atualização constante e a pressão do tempo criaram novas formas narrativas onde a notícia é construída em camadas, a partir das unidades de informação que vão se tornando disponíveis, sendo conectadas pelos *hiperlinks* e cuja estrutura pode ir de materiais praticamente brutos, sem qualquer edição, a pacotes completos do jornalismo tradicional incluindo análises, desdobramentos e contextualização. O conceito de resolução semântica de Fidalgo (2003) descreve o processo, fazendo uma analogia com as imagens digitais que, a partir do aumento do número de *pixels*[10] que as formam, permitem gradualmente melhor visualização e compreensão. Na redação digital as partículas de informação chegam em fluxo contínuo e com elas construímos nossas histórias, iniciando as vezes apenas com poucas palavras na área de "últimas notícias" e, quando merecem, chegando às grandes reportagens contadas no ambiente digital como a premiada Snow Fall (BRANCH, ([200-])) do *New York Times*.

Mesmo com a necessidade de aprendizado em novas habilidades e devendo agora ser capaz de trafegar por tecnologias, equipamentos e conceitos como o de SEO[11] e SMO, a figura do jornalista, reconfigurada pela tecnologia ainda permanece essencial como elemento

[9]Ver em Knight e Cook (2013) a distinção entre os dois conceitos.
[10]O conceito de pixel parte da ideia de que as imagens digitais são formadas por matrizes de pontos que definem a resolução da tela e traduz a menor unidade constituinte da representação das imagens quando são gerenciadas por computadores.
[11]Sigla para *Search Engine Optimization*, conjunto de técnicas para melhorar o posicionamento da página de internet em mecanismo de busca como Google. SMO (*Social Media Optimization*) seria o equivalente para as redes sociais.

do processo de produção; mas em 2010, os primeiros exemplos de algo mais radical começaram a aparecer.

3 Narrativas Automatizadas - *Narrative Science e Automated Insigths*

Morozov (2012) utilizando um sugestivo título, "Um robô roubou o meu Pulitzer !" relata os primeiros movimentos de empresas de inteligência artificial, entre elas a *Narrative Science*[12], no negócio de gerar notícias. O produto da empresa: conteúdo jornalístico automatizado vendido como serviço para portais de notícias, principalmente da área de esportes e finanças, onde uma boa parte da informação utilizada advém de números e relações entre grandezas mensuráveis como a cotação do dólar ou o resultado de uma partida de futebol.

Figura 2 - Print de matéria sobre jornalismo automático

HOME / FUTURE TENSE : WHAT'S TO COME?

Article from future tense
ASU | NEW AMERICA | SLATE

A Robot Stole My Pulitzer!

How automated journalism and loss of reading privacy may hurt civil discourse.

By Evgeny Morozov | Posted Monday, March 19, 2012, at 7:11 AM ET

Can technology be autonomous? Does it lead a life of its own and operate independently of human guidance? From the French theologian Jacques Ellul to the Unabomber, this used to be widely accepted. Today, however, most historians and sociologists of technology dismiss it as naive and inaccurate.

Yet the world of modern finance is increasingly dependent on automated trading, with sophisticated computer algorithms finding and exploiting pricing irregularities that are invisible to ordinary traders.

Meanwhile, *Forbes*—one of financial journalism's most venerable institutions—now employs a company called Narrative Science to automatically generate online articles about what to expect from upcoming corporate earnings statements. Just feed it some statistics and, within seconds, the clever software produces highly readable stories. Or, as *Forbes* puts it, "*Narrative Science*, through its proprietary artificial intelligence platform, transforms data into stories and insights."

Automated journalism like that produced by Narrative Science could perhaps save media jobs, but it can also hurt civil discourse.
William Gottlieb/Library of Congress

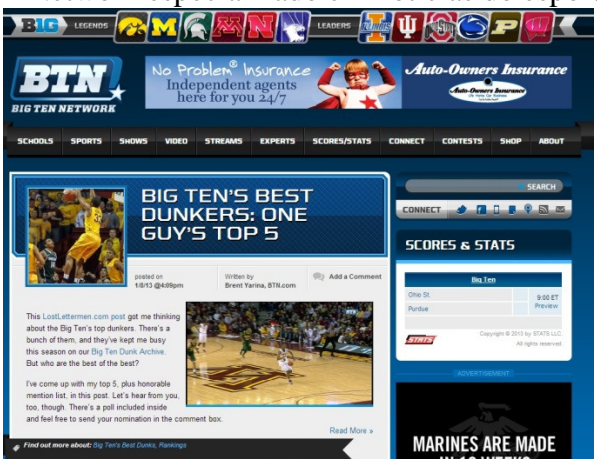
Fonte: Morozov (2012).

Narrative Science (NS) nasceu de um projeto de pesquisa chamado "*Stats Monkey*" desenvolvido por alunos e professores de ciência da computação e jornalismo da *Northwestern University* a partir do InfoLab e que basicamente escrevia resumos sobre resultados de jogos do baseball americano. Em 2010 a empresa mudou de nome e logo em

[12]http://narrativescience.com/

seguida patenteou uma plataforma de autoria baseada em inteligência artificial chamada Quill.

Figura 3 - Portal Big Ten Network especializado em notícias do esporte e cliente da NS



Fonte: *Big Ten Network* (2014).

Automated Insights (AI) é outra companhia que já fornece conteúdo jornalístico automatizado para diversos clientes. Nascida com o nome de *StatSheet*, em 2008, a empresa recebeu financiamento de uma entidade de apoio à inovação no estado da Carolina do Norte nos EUA e iniciou um percurso de desenvolvimento que em 2014 contabilizou, segundo seu site oficial (AUTOMATED INSIGHTS, 2013), mais de 300 milhões de textos escritos automaticamente, entre relatórios empresariais e notícias jornalísticas.

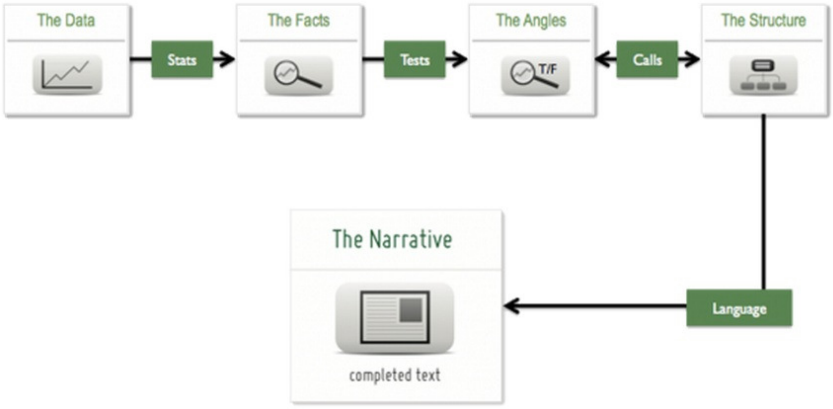
Figura 4 - Exemplos de conteúdo publicado por AI em plataformas móveis

Fonte: *Automated Insights* (2013).

Arce (2009) já havia feito uma avaliação da possibilidade de automatização dos discursos incluindo ai as ideias de Lage (1997) sobre o tema, ambos, entretanto, em caráter teórico e não experimental.

Coppin (2010, p. 24) esclarece que, no campo da inteligência artificial, uma das principais questões está relacionada à representação da realidade que será utilizada pelo programa de computador, já que "para um computador poder solucionar um problema relacionado ao mundo real, ele primeiro precisa de um meio para representar o mundo real internamente. Ao lidar com aquela representação interna, o computador torna-se capaz de solucionar problemas".

Figura 5 - Processo de transformação de dados brutos em narrativas utilizado pela NS



Fonte: *Narrative Science* (2010).

Na questão específica do conteúdo jornalístico, as empresas citadas começaram a produzir *leads* basicamente por ser uma forma que apresenta uma estrutura interna bastante definida e por isso traduzível de modo mais fácil para uma sequencia de instruções a ser realizadas por uma máquina.

4 Modelagem de experimento com resultados do futebol

Para construir nosso experimento de narrativa automatizada nos propomos a desenvolver um código de programação capaz de escrever pequenos textos sobre os resultados do campeonato brasileiro de futebol de 2013. Utilizamos a linguagem de programação Python[13] por considera-la de mais fácil aprendizagem para não especialistas em programação como jornalistas e profissionais da comunicação[14].

A linguagem Python permite a utilização de diversos módulos de programação já desenvolvidos previamente e com finalidades específicas, facilitando a construção das soluções a partir da combinação de funções cujo código já existe. A biblioteca NLTK[15] – *Natural Language Toolkit* – que utilizamos nesse projeto é um desses exemplos e incorpora um grande número de recursos para o processamento de textos.

A modelagem do problema foi feita a partir da seguinte sequencia: obter resultados dos jogos e informações complementares tais como local da partida e número da rodada; registrar essas informações em alguma estrutura simples de arquivo que pudesse posteriormente ser consultada para a construção do material; traduzir as próprias regras do torneio em termos de variáveis e relações para que a sintaxe do regulamento pudesse orientar a concatenação dos elementos do texto; gerar as frases a partir dos resultados das operações realizadas com os dados coletados nas partidas.

No primeiro protótipo, a opção de pedir ao usuário que entrasse com todas as informações relativas aos jogos, tais como local, times e resultado, mostrou-se ineficiente

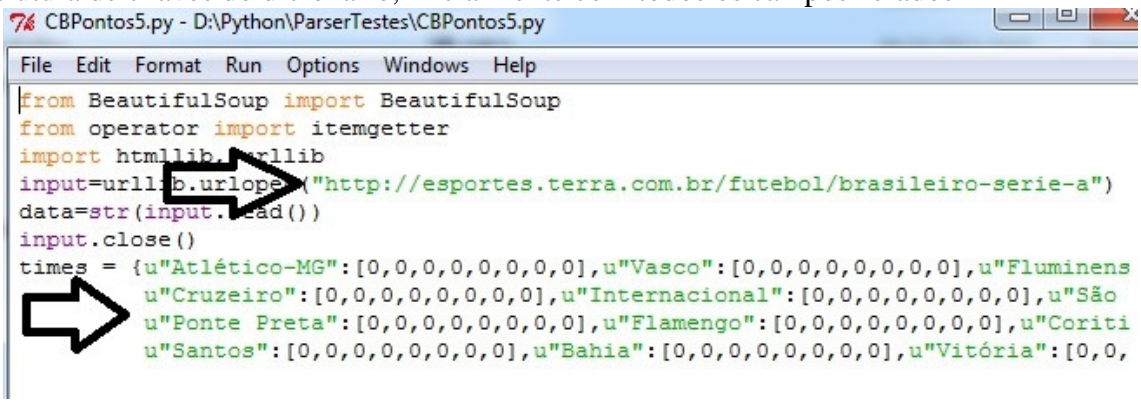
[13]<www.python.org>.
[14]Projetos envolvendo programação e jornalismo têm sido desenvolvidos, com exemplos na área do Jornalismo Investigativo, no intuito de extrair e processar dados em grandes quantidades e utilizar essas informações para a construção de infográficos e narrativas no jornalismo digital. <http://gijn.org/>.
[15]<www.nltk.org>.

devido a grande chance de erros de digitação e inserção de termos em formato inesperado que faziam o software travar com muita facilidade.

Assim partimos para uma solução que, a partir de um endereço específico na internet onde esses dados fossem disponibilizados, conseguia ler todas as informações iniciais de forma automática e mais rápida. Para os testes selecionamos a página do portal Terra dedicada à área de esportes que publicava a cada rodada os resultados e a tabela atualizada do campeonato (PORTAL TERRA, 2014)[16]. A tabela foi usada como instrumento de validação dos cálculos do software já que ela também totalizava as métricas que as regras do torneio geravam tais como número de jogos, pontos ganhos, gols feitos, gols sofridos, saldo de gols e índice de aproveitamento.

Definida a estratégia de coleta dos dados fizemos a parte do código que salvava esses elementos associando-os a cada time numa estrutura que na linguagem Python é conhecida por dicionário, onde a cada elemento, chamado de chave, são associados valores diversos, cada um representando alguma informação gerada a partir dos resultados dos jogos.

Figura 6 – Parte do código que mostra o endereço de extração dos dados e os times na estrutura de chaves do dicionário, inicialmente com todos os campos zerados



Fonte: Elaborado pelo autor.

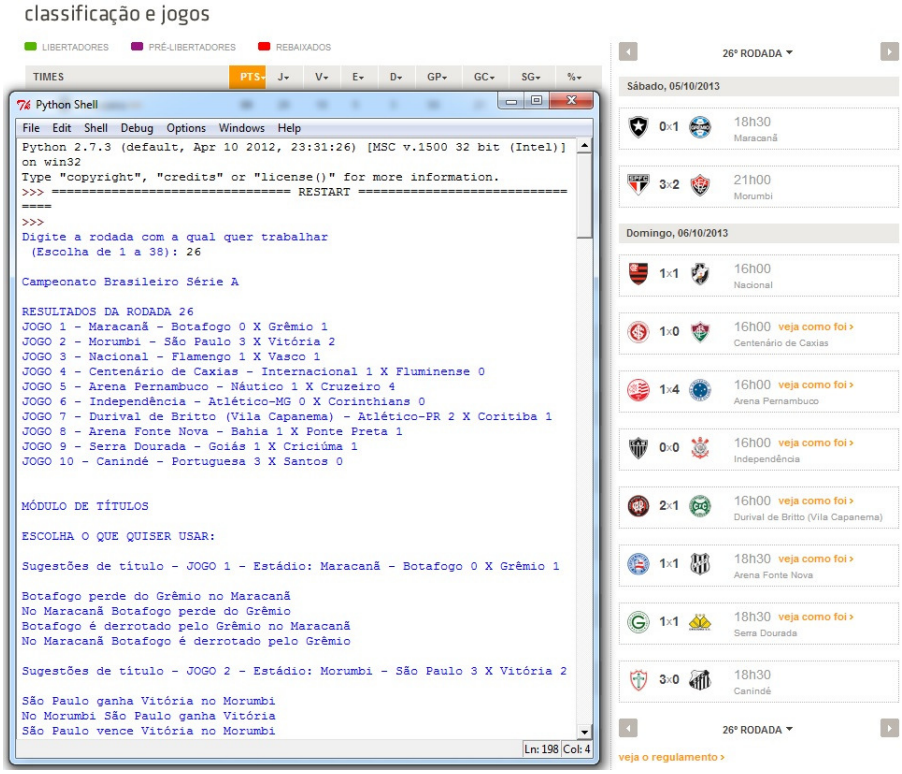
Ao iniciar o código o usuário é demandado apenas a escolher o número da rodada que deseja explorar. A partir dai o software coleta os resultados de todas as rodadas até chegar à selecionada e assim vai registrando os resultados e acumulando-os na estrutura do dicionário.

[16]Atualmente o endereço da tabela é <http://esportes.terra.com.br/futebol/brasileiro-serie-a/tabela>.

É interessante observar que o que é retirado do portal da internet são apenas os resultados dos jogos. Com eles o software aplica as regras do torneio para calcular os outros valores associados ao time. Por exemplo, ao coletar o resultado de determinada partida, o software compara o número de gols dos dois times envolvidos, se um deles é maior do que o outro, o de valor maior ganhou a partida e, por isso, no registro referente a pontos ganhos são acrescidas três unidades. O perdedor não soma nada no registro e no caso de saldo de gols iguais, uma unidade é acrescentada a cada um dos times indicando os pontos por um empate.

A parte do texto também está associada a essa lógica. Foi criada uma lista de verbos a ser usada de acordo com o contexto do resultado. Assim "vence", "ganha", "bate" e outras construções semelhantes são escolhidas de forma aleatória pelo software. Quando a diferença de gols entre o vencedor ou o perdedor é maior que dois, indicando uma forte superioridade no placar, uma outra lista de verbos é acionada com opções como "arrasa" , "liquida" e até "humilha", pensando num texto mais sensacionalista. Assim o título é gerado concatenando as informações básicas que já temos, no caso, o nome dos times e o placar, como em "Flamengo vence Botafogo no Engenhão por 3 a 2". Considerando que o software capturou da internet os times envolvidos e o resultado, bem como o lugar onde aconteceu, basta comparar os gols do placar, definir o vencedor e dai escolher um dos verbos da lista disponível concatenando todos esses elementos numa estrutura simples.

Figura 7 - Tela que compara a página do portal com os resultados e a tela gerada pelo programa onde podem ser vistos primeiro os dados registrados e depois as sugestões de título baseadas nos resultados



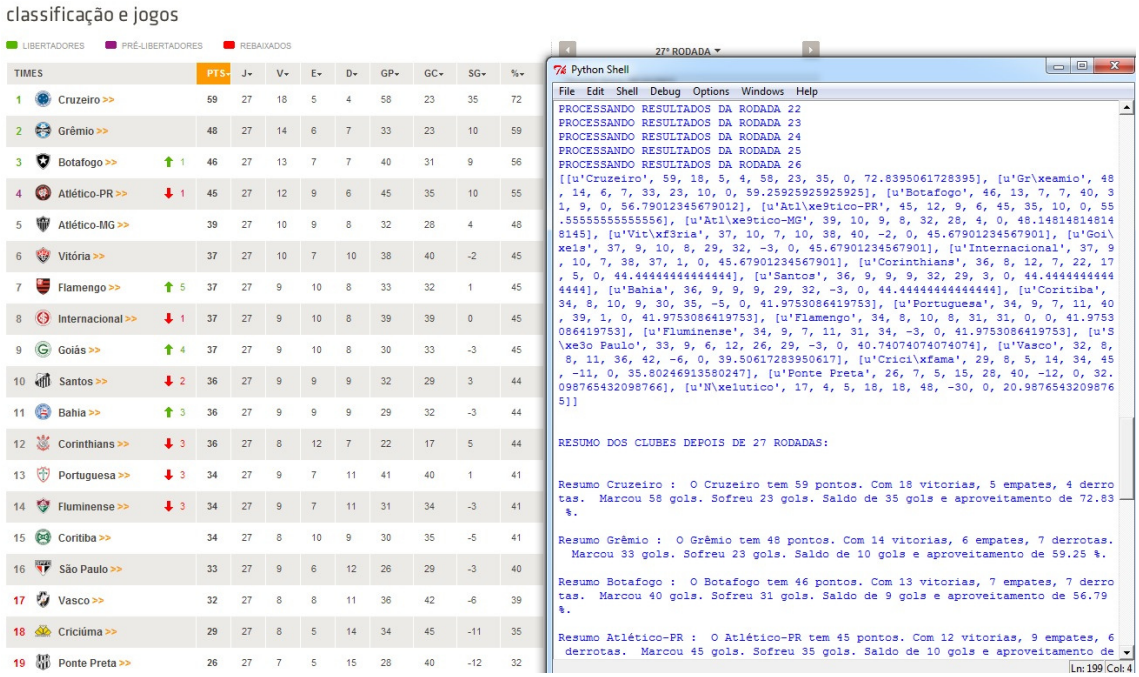
Fonte: Elaborado pelo autor.

À medida que o software registra os jogos das rodadas ele vai atualizando todos os parâmetros adicionais já listados, que são representações definidas pelas próprias regras do torneio, incluindo na estrutura de dicionário que foi criada um conjunto de dados que será utilizado para inferir várias outras informações como a própria posição do time na tabela, o número de pontos que cada um tem e seu aproveitamento, calculado dividindo o total de pontos conseguidos pelo total de pontos disputados.

Como exemplo, se um time disputou 10 jogos, ou seja, 30 pontos, valor que teria se vencesse todas as partidas e de fato tem apenas 3, uma vitória e um empate ou três empates, seu rendimento seria de apenas 10%, ou seja, bastante baixo.

Esses números permitem ao software escrever textos com mais informações.

Figura 8 - Tela do software que demonstra a estrutura do dicionário atualizada e um pequeno texto de resumo da situação do time no campeonato a partir dos elementos registrados



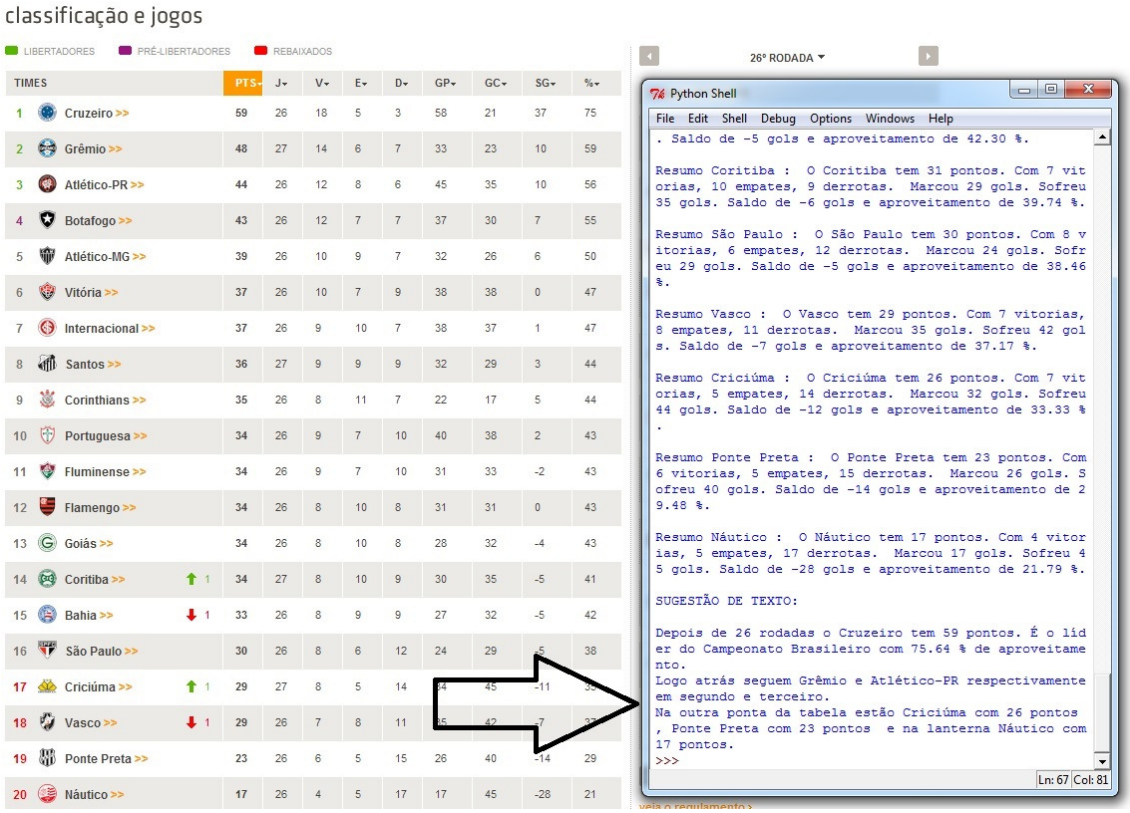
Fonte: Elaborado pelo autor.

Num nível com um pouco mais de complexidade é escrito então um lead com informações gerais sobre a situação do campeonato naquela rodada. Basicamente a partir do resumo que é inferido com a situação individual de cada time, o software faz o texto indicando os líderes com seus números e os lanternas do torneio, aspectos que normalmente são enfatizados em noticiário desse tipo. É importante ressaltar que a tabela do site é aqui utilizada apenas para validar os resultados já que o software retira da internet somente os jogos, seus resultados e o local onde aconteceram. O resto ele mesmo calcula, baseado nas regras do torneio que estão representadas para ele em termos de instruções de programação.

A construção desse conteúdo, apesar de um pouco mais complicada, também parte da ideia de concatenar unidades menores de informação a partir de listas de palavras e expressões comuns nesse tipo de texto.

A título de ilustração seria como definir uma estrutura prévia onde alguns elementos, no caso o nome dos times e suas métricas, podem ser imaginados como lacunas a ser preenchidas por quem estiver naquelas posições em uma determinada rodada. A ideia de arquivos dinâmicos, ou seja, que se alteram ao longo do tempo pode ser utilizada aqui.

Figura 9 - Tela do software com o que seria o lead construído a partir das informações lidas sobre o campeonato em determinada rodada



Fonte: Elaborado pelo autor.

5 Conclusões

Apesar de ter sido conduzido apenas em caráter exploratório, o experimento indica a possibilidade real e não apenas teórica de produzir alguns tipos de estruturas jornalísticas de forma automatizada.

Ficou claro que conteúdos baseados em informações numéricas e relações que podem ser reduzidas mais facilmente a expressões matemáticas, baseadas numa sintaxe mais restrita, como a que pode ser extraída das regras de um torneio esportivo, por exemplo, são mais fáceis de reproduzir.

Da mesma forma que fizemos com os resultados do Campeonato Brasileiro, não seria difícil construir algo semelhante para gerar pequenos textos informando as variações do câmbio ou de ações em bolsas de valores, a previsão do tempo para cidades ou regiões e

outros conteúdos que, se observarmos, são construídos com uma estrutura que se repete com pequenas variações.

A capacidade de coletar e processar informações com grande quantidade e variedade parece indicar o potencial de uso desse tipo de solução, principalmente no jornalismo online e nos grandes portais da internet que precisam atualizar seus conteúdos com mais agilidade.

Formas mais complexas de programas já estão em uso comercial, como nos exemplos da *Narrative Science* e da *Automated Insights*. Na área das plataformas de mídias sociais também já existem várias soluções capazes de extrair dados da massa de conteúdo que geramos diariamente através do que postamos e dai inferir que tópicos têm mais potencial de replicação, ou de tornarem-se um viral, nos termos do marketing, o que também indicaria que uma nova forma de medir o valor de noticiabilidade de determinado tema já estaria disponível, dessa vez não a partir da avaliação do jornalista, mas da métrica extraída das informações sobre o interesse dos usuários naquele assunto[17].

Os desdobramentos desse tipo de tecnologia no mercado ainda não podem ser avaliados. É fato que muitas redações têm reduzido postos de trabalho e/ou se utilizado de mão de obra contratada de outras formas diferentes da relação formal registrada em carteira. Temporários, estagiários e colaboradores que são pagos como pessoa jurídica são situações comuns. Com o desenvolvimento de soluções mais sofisticadas de produção de conteúdo a partir de software é possível considerar que novas alterações podem estar a caminho.

É importante ressaltar também que mesmo as soluções mais complexas de inteligência artificial ainda estão distantes de replicar as sutilezas e complexidades de um bom texto jornalístico, principalmente numa língua como a portuguesa, que até hoje apresenta dificuldades para outras categorias de software como os de reconhecimento de voz e tradução para conseguirem níveis altos de acerto.

Por outro lado, a precarização do trabalho e a replicação indiscriminada de releases e conteúdos gerados por fontes, justificados de forma simplista pela pressão do tempo e pela necessidade de atualização constante, são um risco para esses profissionais já que, como foi demonstrado, as operações simples e baseadas em estruturas comuns têm muito mais chances de serem replicadas automaticamente.

[17]Como exemplo ver: <http://www.newswhip.com/Brasil>.

O aprofundamento no trabalho de apuração, o jornalismo investigativo, a extração de relações complexas a partir de dados inter-relacionados e a criação de infográficos e formas alternativas de visualização de informações nos parecem bons exemplos de como a atividade humana pode continuar sendo essencial no que se considera um jornalismo de qualidade. A melhoria dos currículos e dos programas de formação na área também terão papel importante nos impactos dessas novas tecnologias.

Se "resistir é inútil"[18] parece ser uma afirmação intimamente ligada às relações entre homens e técnica na história das sociedades, no campo do jornalismo, um texto criativo e bem elaborado poderá nos garantir a convivência pacífica com as soluções automatizadas que tem seu valor em processos repetitivos e de baixo nível de execução.

Muito mais nociva do que a geração de textos jornalísticos via software parece ser a automatização dos jornalistas que deixam de exercer a ação humana e complexa ligada à sua atividade, no exercício das práticas da sua profissão, simplesmente replicando conteúdos ou realizando de forma descuidada parte do seu trabalho. Esse parece ser o grande problema que teremos que enfrentar, sejamos nós céticos, temerosos ou fascinados por tecnologia.

6 Referências

ARCE, Tacyana. O lead automatizado: uma possibilidade de tratamento da informação para o jornalismo impresso diário. **Revista Exacta**, Belo Horizonte, v. 2, n. 3, 2009.

AUTOMATED INSIGHTS. 2013. Disponível em: <www.automatedinsights.com>. Acesso em: 10 jan. 2013.

BIG TEN NETWORK. 2014. Disponível em: <www.btn.com>. Acesso em: 12 abr. 2014.

BRADSHAW, Paul; ROHUMAA, Liisa. **The online journalism handbook**: skills to survive and thrive in the digital age. Essex: Pearson Education, 2011.

BRANCH, John. Snow Fall: the avalanche at Tunnel Creeak. **The New York Times,** New York, (200-]). Disponível em: <http://www.nytimes.com/projects/2012/snow-fall/?forceredirect=yes#/?part=tunnel-creek>. Acesso em: 2 jun. 2014.

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

[18]"Resistance is futile", frase repetida pelos *Borgs* da séria *Star Trek* para suas vítimas (tradução nossa).

COPPIN, Ben. **Inteligência artificial**. Rio de Janeiro: LTC, 2010.

DEVAUX, Pierre. **Autómatos, automatismo e automatização**. Tradução Luis Borges Coelho. Lisboa: Editorial Gleba, 1964. (Coleção Horizonte, n.3).

DUSEK, Val. E-book. **Philosophy of technology**: an introduction. Malden, MA: Blackwell Publishing, 2006.

ECO, Humberto. **Apocalípticos e integrados**. São Paulo: Perspectiva, 2006.

ELLUL, Jacques. **A técnica e o desafio do século**. Rio de Janeiro: Paz e Terra, 1968.

FEENBERG, Andrew. E-book. **Transforming technology**: a critical theory revisited. New York: Oxford University Press, 2002.

_____. E-book. **Between reason and experience**. Essays in technology and modernity. Cambridge, MA: Mit Press, 2010.

FIDALGO,Antonio. Sintaxe e semântica das notícias on-line. Para um jornalismo assente em base de dados. In: FIDALGO, António; SERRA, Paulo (Org.). Informação e Comunicação Online. **Jornalismo Online**. v. 1.Covilhã: Universidade da Beira Interior/Portugal, 2003.

LAGE, Nilson. O lead clássico como base para a automação do discurso informativo. In: CONGRESSO BRASILEIRO DE PESQUISADORES DA COMUNICAÇÃO INTERCOM, 20., 1997, Santos. **Anais**... Santos, SP. 1997.

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

LUTICE CRÉATIONS. Paris, [2000-]. Disponível em: <http://www.automates-boites-musique.com/>. Acesso em: 7 abr. 2014.

KNIGHT, Megan; COOK, Clare. **Social media for journalists**: principles e practice. Londres: Sage, 2013.

MACHADO, Elias**. O ciberespaço como fonte para os jornalistas**. Salvador: Calandra, 2003.

MIELNICZUK, Luciana. **Características e implicações do jornalismo na web**. 2001. Disponível em: <http://200.18.45.42/professores/chmoraes/comunicacao-digital/13-2001_mielniczuk_caracteristicasimplicacoes.pdf>. Acesso em: 8 set. 2010.

MOROZOV, Evgeny. **A robot stoled my Pulitzer!**: future tense. 2012. Disponível em: <http://www.slate.com/articles/technology/future_tense/2012/03/narrative_science_robot_journalists_customized_news_and_the_danger_to_civil_discourse_.html>. Acesso em: 11 abr. 2014.

NARRATIVE SCIENCE. 2010. Disponível em: <www.narrativescience.com>. Acesso em: 10 jan. 2013.

PORTAL TERRA. Esporte. 2014. Disponível em: <http://esportes.terra.com.br/futebol/brasileiro-serie-a>. Acesso em: 31 maio 2014.

RÜDIGER, Francisco. **Introdução às teorias da cibercultura**: tecnocracia, humanismo e crítica no pensamento contemporâneo. 2. ed. Porto Alegre: Sulina, 2007.

SENNETT, R. **O artífice**. Rio de Janeiro: Record, 2009.

SORIA, Carlos. **Convergência de mídias**. 2014. Palestra apresentada ao Seminário sobre Integração Multimídia, São Luís, 2014.

THE SOCIETY FOR PHILOSOPHY AND TECHNOLOGY. 2014. Disponível em: <http://www.spt.org/>. Acesso em: 2 jun. 2014