Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

SOBRE BIG DATA E NEOPOSITIVISMO DIGITAL NA PESQUISA EM HISTÓRIA

Tiago Luís Gil[1,2]

RESUMO

O texto discute o avanço das iniciativas digitais na pesquisa em história, dando particular relevo para o incremento no número de grandes repositórios e conjuntos arquivísticos, tomados aqui como parte de uma tendência de *big data* que precisa ser discutida na disciplina. São apresentados alguns casos de projetos com forte apelo empirísta (como o *Transkribus* e o *Time Machine*) e como estas iniciativas andam *pari passu* com discussões sobre uma suposta irrelevância da teoria. O artigo aponta ideias para o debate, indicando a capacitação digital como uma necessidade na formação de historiadores, diante de um mar de incertezas e algorítimos.

PALAVRAS-CHAVE

big data - neo-positivismo - história digital - ferramentas digitais - fontes históricas digitais.

1 Universidade de Brasília. Brasília, Brasil.

2 Tiago Gil é professor associado de História da América no Departamento de História da Universidade de Brasília e doutor em História Social pela Universidade Federal do Rio de Janeiro (2009). É autor de "Infiéis Transgressores: elites e contrabandistas nas fronteiras do Rio Grande e do Rio Pardo" (Arquivo Nacional, 2007) e "Coisas do caminho: crédito, confiança e informação na economia do comércio de gado entre Viamão e Sorocaba" (Editora da UnB, 2020). Email: tiagoluisgil@gmail.com

CC BY

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

ABOUT BIG DATA AND DIGITAL NEO-POSITIVISM IN HISTORY RESEARCH

ABSTRACT

The text discusses the rise of digital initiatives in historical research, particularly emphasizing the increase in the number of large repositories and archival collections. This proliferation is seen as part of the "big data" trend that needs to be debated in the discipline. Some examples of projects with a strong empirical approach are presented (such as *Transkribus* and *Time Machine*), and how these initiatives are associated with discussions about the supposed irrelevance of theory. The article proposes ideas for debate, indicating digital training as a necessity in preparing historians in the face of a sea of uncertainties and algorithms.

KEYWORDS

big data - Neopositivism - digital history - digital tools - digital historical sources.

Recebido em: 11/02/2024 – Aprovado em: 29/02/2024

CC BY

Nos últimos anos, temos assistido a um contínuo crescimento do número de iniciativas de "informatização" da pesquisa em história, com a utilização de recursos computacionais nas mais diversas temáticas e abordagens. A expansão da internet fez com que um número incomensurável de documentos digitais fosse criado todos os dias, nascidos digitais ou digitalizados. A avalanche de dados é tamanha que diferentes iniciativas foram criadas para dar conta dessa imensidão.[3]

Essas preocupações foram lentamente chegando na pauta da pesquisa em história, tanto no tratamento dos dados *online* quanto no uso de fontes digitalizadas, cada vez mais frequentes em todos os cantos do mundo.[4] A agenda das *big techs* agora é também uma realidade para os historiadores, e o apelo aos chamados *big data* tem sido uma constante, especialmente na Europa e nos Estados Unidos. Nesta discussão, devemos embarcar nos *big data* e abraçá-los, como defendem, por exemplo, os historiadores Guldi e Armitage?[5] Ou devemos nos precaver e avaliar criticamente essas ferramentas, como quer a matemática e ativista Cathy O'Neil?[6] Temos, em nossa formação de historiadores, ferramentas para isso? Esta contribuição visa trazer alguns elementos para esse debate, que começa a ganhar fôlego no Brasil.

Era uma vez o futuro

Eram os idos de 2012 quando Frédéric Kaplan iniciava o *Venice Time Machine*, um projeto milionário e internacional que previa a digitalização de muitos quilômetros de documentos do Arquivo de Estado de Veneza[7] (entidade parceira do projeto), para posterior criação de um "*multidimensional model of Venice and its evolution covering a period of more*

3 Kitchin, 2014; Graham; Milligan; Weingart, 2016.

4 Graham; Milligan; Weingart, 2016.

5 Guldi; Armitage, 2014.

6 O'Neil, 2016.

7 Uso aqui "Arquivo" com a inicial maiúscula para uma completa diferenciação de "arquivo" de computador. A natureza do texto poderia permitir essa dúvida.

CC BY

than 1000 years".[8] O projeto previa a instalação de grandes *scanners* no Arquivo e sua operação por engenheiros (caso do próprio Kaplan) para posterior reconhecimento digital automático dos manuscritos. Mediante elaborados recursos de programação, seria possível a identificação de nomes de pessoas e lugares, que acarretaria, quase como uma obviedade, a descoberta de redes de relacionamentos e a (também óbvia) criação de extensos gráficos de redes sociais, mostrando as (certamente simples) conexões entre as pessoas e seus espaços de atuação:

> By combining this mass of information, it is possible to reconstruct large segments of the city's past: complete biographies, political dynamics, or even the appearance of buildings and entire neighborhoods. The information extracted from the primary and secondary sources are organized in a semantic graph of linked data and unfolded in space and time in an historical geographical information system[9]

Era um conto de fadas digital. A moderna engenharia da computação agora procurava fazer aquilo que inúmeros historiadores até então não haviam conseguido: varrer uma série de documentos extremamente longa de modo exaustivo. Não se tratava apenas de digitalizar, mas de reconhecer um sem-número de caligrafias diferentes produzidas em uma diversidade de idiomas (lembrando que Veneza era uma potência comercial que mantinha embaixadores em terras longínquas), inclusive o latim e o vêneto.

Assim, não se tratava apenas de reconhecer os textos e seu teor, mas obter informações que permitissem o reconhecimento de personagens e suas vidas, posteriormente montadas e contadas (também para o grande público) com o uso de modernos artifícios computacionais. O conto de fadas se fazia ainda mais fantástico, pois o projeto, em seus *websites*, utilizaram com esmero recursos gráficos como vídeos, mapas coloridos e gráficos animados, além de uma fala impressionante (e impressionista) em uma *TED Talk*. O futuro enfim havia alcançado os territórios do passado, e

8 Kaplan, 2015.

9 Kaplan, 2015.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

os historiadores estavam prestes a se tornar espectadores privilegiados da capacidade computacional que os engenheiros eram capazes de aportar.

O conto de fadas, contudo, teve um fim. Era setembro de 2019 e o Arquivo de Estado de Veneza lançava uma nota à imprensa informando a suspensão dos acordos de cooperação então vigentes com a École Polytechnique Fédérale (EPFL), de Lausanne, onde Kaplan atuava, e a Universidade Ca'Foscari de Veneza, também parceira. O motivo era a falta de transparência referente às atividades realizadas pela equipe de Lausanne sobre os procedimentos adotados e os resultados obtidos. O Arquivo de Veneza reclamava de uma indesejada hierarquia entre as instituições, a exclusão de seus técnicos dos trabalhos, a falta de discussão sobre os procedimentos de digitalização e catalogação e, finalmente, a transparência sobre as decisões tomadas e a análise dos primeiros resultados. Segundo o então diretor do Arquivo, Gianni Doria, a decisão de encerrar os trabalhos teria sido mútua, após uma série de tentativas de acordos que especificasse uma nova política de interação. A EPFL, contudo, dizia ter sido uma decisão unilateral do Arquivo e informava que buscaria novos diálogos, o que não se concretizou.[10]

A decisão do Arquivo de Veneza não destacava apenas a discussão técnica e política sobre as tratativas. Ao final do documento, uma importante observação era feita, no sentido de discutir os termos da parceria "*nella convinzione che non sia sufficiente digitalizzare i documenti, anche attraverso strumenti e algoritmi complessi, per comprendere la storia di Venezia*".[11] Era uma nota com perfume e brio locais. Mas era também uma nota de resistência a um projeto que tentou colonizar seus documentos com armas ainda desconhecidas e com um poder de fogo não completamente explicado.

Não se tratava apenas de uma questão contratual ou de falta de comunicação. O que estava em jogo era a noção do que seria a história na cabeça das diferentes equipes. A noção de história do Arquivo era uma

10 Castelvecchi, 2019, p. 607.

11 Archivio di Stato di Venezia, 2019. "*Com a convicção que não é suficiente digitalizar os documentos, mesmo com o uso de instrumentos e algoritmos complexos, para compreender a história de Veneza*".

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

visão tradicional, mas muito conectada à leitura erudita e lenta das fontes. A noção de Kaplan, aparentemente mais moderna, dispensava o erudito para substituí-lo por um algoritmo. Ambas eram fundamentalmente empiristas, mas a da EPFL não se oferecia à crítica (das fontes) e partia da premissa de que os dados se organizariam por si mesmos, independentemente de uma teoria de fundo; ou, melhor, os dados se exibiriam de uma forma aceitável para todas as leituras de mundo possíveis, tanto de *experts* quanto do grande público.[12]

O projeto *Venice Time Machine* tomou, nos anos seguintes, um caminho diferente, que merece análise detida, a ser efetuada em outro momento. No entanto, convém destacar suas implicações políticas, bastante complexas, distantes da aparência de uma solução meramente técnica para problemas de historiadores. O projeto agora se volta para criar focos de *Time Machine* em várias cidades da Europa, sem jamais perder de vista a ideia de que o uso de algoritmos pode "amplificar" as fontes disponíveis.[13]

O canto da sereia

É famosa a passagem da *Odisseia* que fala sobre o momento em que Ulisses passara com seu navio junto à ilha ao redor da qual abundavam sereias capazes de atrair, com seu belíssimo canto, as embarcações para perto das rochas, naufragando-as.[14] Para desfrutar do canto sem correr riscos, Ulisses tapou o ouvido de seus marinheiros e se fez amarrar ao mastro do navio.

Não há dúvida de que o *Transkribus*, uma ferramenta de reconhecimento automático de manuscritos, seja comparável a um canto belíssimo. Criada a partir de um projeto da Universidade de Innsbruck, ela é capaz de identificar texto em um documento manuscrito, reconhecer as linhas

12 Teoria, aqui, aparece não como uma teoria específica, como um modelo explicativo em particular e tampouco como um campo disciplinar, mas como um componente fundamental do processo de produção de conhecimento.

13 Time Machine, 2024.

14 Homero, 1997

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

e, finalmente, transcrever o texto, identificando uma "zona" da imagem (mediante coordenadas cartesianas dos *pixels* da imagem) com um certo trecho de texto, por HTR (*Handwriting text recognition*, ou reconhecimento de texto manuscrito).

O projeto começou com uma experiência prévia de Günter Mühlberger na digitalização e reconhecimento de caracteres com tecnologia OCR (*Optical Character Recognition*, ou reconhecimento óptico de caracteres) em jornais alemães dos anos 1990.[15] A partir daí, a equipe ao redor de Mühlberger foi rumando para novos desafios, e, no início dos anos 2000, já estava trabalhando em projetos com reconhecimento de manuscritos, ainda que com diversos revezes.[16]

Talvez seja importante explicar a diferença entre as duas tecnologias. Enquanto os sistemas OCR se valem da semelhança das letras com base em uma padronização muito restrita, como é a dos impressos, para a qual são utilizados, a HTR não pode contar com a padronização do formato das letras por conta da multiplicidade de formas da escrita humana, mesmo se tratando de uma mesma caligrafia. Para tanto, as soluções de reconhecimento de manuscrito devem ser bem mais elaboradas, o que implica a criação de modelos de manuscritos, baseados em treinamento de inteligência artificial.[17]

Em 2013, as iniciativas de Mühlberger tomaram a forma de um projeto que vislumbrava efetivamente o reconhecimento de textos manuscritos. Esse projeto *tranScriptorium*, foi a base do *Transkribus*, que posteriormente foi desenvolvido em parceria com a Universidade de Valência. A ferramenta foi mantida dentro do projeto original até 2016, quando foi criada a cooperativa READ, que, desde então, é responsável pela manutenção e pelo desenvolvimento de novas funcionalidades.

O *Transkribus* é certamente uma boa ferramenta, belo como o canto das sereias. O problema sempre é o risco de naufrágio. A tecnologia HTR não pode ser entendida como uma ferramenta neutra e, muito menos,

15 Stauder, 2023.

16 Stauder, 2023.

17 Kahle *et al.*, 2017, p. 19-24.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

descontextualizada de um mundo onde o grande volume de dados é sedutor. Por si só, o *Transkribus* não é uma ferramenta que necessariamente alimenta uma obsessão empirista, mas ele certamente contribui para isso.

Tal como no *Venice Time Machine*, a maioria dos textos que apresentam o projeto e seus *outputs* é fortemente tecnicista.[18] Outra parte expressiva é composta pela descrição dos benefícios e pelo elenco do potencial público consumidor, em que historiadores e grande público estão lado a lado, como se a leitura das fontes antigas fosse a mesma ou muito próxima. Não há uma discussão sobre a visão de mundo que orienta a leitura ou sobre a ambição de ter tanto material para análise. Ter a possibilidade de buscar dados em milhões de documentos antigos parece uma necessidade inquestionável.

A voracidade por documentos nem sempre é sinônimo de um empirismo devorador. Para contar a história de pessoas simples, sobre as quais poucos documentos foram gerados, precisamos ler e descartar milhares de potenciais documentos que poderiam — talvez — mencionar aquela pessoa. Consultamos muitos, utilizamos poucos. Isso ocorre porque as fontes sobre as pessoas do passado não foram geradas de modo homogêneo. Elas são completamente desiguais e refletem, no agregado do que foi produzido e do que restou, escolhas associadas com os poderes de cada tempo. Pessoas comuns são pouco descritas nas fontes e é razoável contar a história delas. Para tanto, uma saída comum é observar todos os documentos possíveis.

A preocupação com a inclusão de pessoas menos aparentes nas fontes históricas não é, contudo, a tônica do *Transkribus*. Não há qualquer discussão sobre a seletividade da memória e se, ao fim e ao cabo, esse incremento empírico não vai acabar, finalmente, aumentando a desigualdade já existente na narrativa histórica, dando ainda mais importância a quem já é muito conhecido. A tônica do projeto está centrada numa metáfora: *unlock the past*, desbloquear, destravar ou abrir o passado. Esse é o mote do projeto, apresentado na página principal e na publicidade da ferramenta, que sugere que o passado estaria trancado. Sendo uma

18 Stauder, 2023; Kahle *et al.*, 2017 .

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

ferramenta de reconhecimento de caracteres, fica a impressão de que basta ler todos os documentos para liberá-lo. O passado está preso na paleografia.

A noção de *unlock* aparece em outros contextos na página do projeto. Ao apresentar o conjunto de ferramentas associadas ao *Transkribus*, a metáfora reaparece, com uma ligeira variação: *unlock history*. São as *Features to unlock history*, descritas na sequência: *AI Text Recognition; Custom AI Training; Field & Table Recognition; Powerful Text Editor; Publishing & Search Tool*. A observação histórica se resume a abrir documentos e fazer buscas por palavras-chave em fontes com texto já reconhecido. A ideia da teoria como uma importante ferramenta de decifração não aparece aqui. A maré dos dados nos guiará.

A noção de *unlock history* aparece reforçada nas palavras do criador do *Transkribus*, Günter Mühlberger, em uma breve entrevista com sua equipe em 2023. Segundo ele, "*there are still so many interesting documents out there waiting to be discovered: Exploring them with HTR will be a big boost to historical research*".[19] Os documentos estão apenas esperando sua descoberta, e o encontro com o documento será um grande incremento na pesquisa em história. Não é preciso insistir que se trata de uma concepção de história arraigada em certo empirismo do século XIX. É o uso da inteligência artificial com uma imaginação romântica.

O problema não é a metáfora. Ginzburg e Prosperi usaram a mesma metáfora de "destrancar" ou arrombar (*forzare*) em um livro publicado em 1975, quando procuravam decifrar as condições de produção e as perspectivas teológicas de uma obra do século XVI. Contudo, nesse mo-

19 Stauder, 2023

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

mento, a "chave-mestra" para arrombar o segredo era um conceito teórico e não um amontoado de manuscritos.[20]

Sobre o neopositivismo digital

As histórias do *Venice Time Machine* e do *Transkribus* não são casos isolados. São projetos próximos do campo dos historiadores e, talvez, por isso nos pareçam mais interessantes. Nossa vida digital cotidiana está marcada por "soluções" digitais que organizam milhões de dados e nos oferecem respostas tidas como "neutras" e aceitáveis, sem qualquer comprometimento político ou teórico. Ferramentas como os buscadores são um exemplo perfeito disso. Todos os dias utilizamos buscadores para encontrar informação, sem ter a menor ideia sobre as decisões que aquelas ferramentas tomam (automaticamente, mas criadas por um cérebro humano) para hierarquizar as respostas, excluir ou incluir itens e muitas outras variáveis.[21] De Certeau nos lembrava, em 1972, de que a seleção dos dados é o primeiro momento em que a teoria opera.[22] O que fazer agora, quando um algoritmo seleciona os materiais com os quais lidamos não apenas na pesquisa, mas também em nossas vidas?

> Par sa puissance et son efficacité, par son ambiguïté, Google neutralise le sens critique qui nous permettrait de garder à l'esprit que lorsque nous y cherchons une information, le moteur de recherche mondial avance une représentation particulière de la réalité pour nous répondre, et non la réalité elle-même.[23]

Esse apelo à técnica como resposta primordial aos problemas dos historiadores não é privilégio de engenheiros que criam motores de busca, transcritores automáticos ou "modelos multidimensionais" da evolução de um dado lugar ao longo do tempo. Ele já penetrou profunda-

20 Ginzburg; Prosperi, 1975.

21 Mounier, 2018.

22 De Certeau, 1978 [1972].

23 Mounier, 2018.

mente na nossa disciplina, não somente pela amplitude (cada vez maior) das ferramentas digitais em nosso cotidiano, mas também pela vertigem na observação dos acervos *online* cada vez mais avassaladores, sejam aqueles nascidos digitais, sejam os agora digitalizados. Encontramos uma grande euforia de muitos historiadores diante do *big data* e das soluções técnicas.[24] Há quem defenda que o uso de grandes volumes de dados, aliado a novas formas de visualização, pode ser disruptivo e dar início a novas epistemologias.[25]

A programação tem se tornado uma solução bastante disponível como ferramenta de pesquisa, e estamos assistindo ao surgimento de diversos cursos não simplesmente de Humanidades Digitais, mas de História Digital propriamente dita. A crítica não pode se bastar ao fato de as ditas "humanidades" serem de difícil classificação, de saber até que ponto essas disciplinas compartilham, de fato, epistemologias e práticas. Ela precisa discutir os limites da técnica. [26]

Um exemplo dessa abordagem foi uma obra de 2014, *The History Manifesto*, que teve grande repercussão e gerou um extenso debate. A maior parte do debate realizado se restringiu à discussão sobre a *longue durée* apresentada por Guldi e Armitage, seus autores. Segundo eles, novas pesquisas em história com maior abrangência cronológica teriam um efeito benéfico após anos de pesquisas de tempo curto, que seriam responsáveis, entre outros problemas, por um afastamento da história de um potencial grande público e pela perda de relevância dos historiadores como figuras importantes na formação da opinião pública. Mas um aspecto foi menos debatido: o uso de *big data*, defendido pelos autores, a partir de ferramentas digitais. A defesa do *big data* aparece, agora, como uma solução para velhos problemas.[27]

Guldi e Armitage não estão sozinhos. Temos uma profusão de revistas especializadas que vêm surgindo sobre o tema das humanidades

24 Kitchin, 2016; Maud Ehrmann *et al.*, 2021; Jemielniak, 2020.

25 Moretti, 1999, 2005; Kitchin, 2006, p. 20-29.

26 Mounier, 2018.

27 Guldi; Armitage, 2014.

digitais e da história digital. Não são publicações acríticas, pelo contrário, mas o espaço reservado para a discussão teórica é muito reduzido em comparação com a potencial difusão de "receitas de bolo" digitais. Em 2007, surge o *Digital Humanities Quarterly*. Em 2011, aparece o *Journal of Digital Humanities*. Em 2015, era a vez do *Journal of Open Humanities Data*. Em 2017, aparecia uma publicação em língua italiana, o *Umanistica Digitale*, criado pela também recente *Associazione per l'Informatica Umanistica e la Cultura Digitale* (2011). Em 2019, o *International Journal of Digital Humanities*. Em 2020 surgiam duas novas publicações: a francesa *Humanités numériques* e a italiana *Magazén: international journal for digital and public humanities*. Estou considerando aqui apenas publicações com foco integral no tema, sem considerar os vários *dossiers* que foram lançados.

Desde 2010, há um surto de publicações técnicas voltadas para a pesquisa em humanidades. Um dos primeiros exemplos é *Macroanalysis: digital methods and literary history*, de Matthew Jockers (2011).[28] Em 2014, Folgert Karsdorp apresentou uma edição virtual (em um "notebook virtual") de um curso da linguagem de programação python, com *Python Programming for the Humanities*, que gerou a edição posterior de *Humanities data analysis: case studies with Python* (2021), em colaboração com Mike Kestemont e Allen Riddell.[29] Em 2018, Brian Kokensparger lançava *Guide to Programming for the Digital Humanities*, também focado na linguagem python.[30] Em 2020, Jemielniak lançava seu *Thick big data: doing digital social sciences*, demonstrando que a voga das humanidades digitais andava mesmo na direção dos chamados *big data*.[31] Também há uma grande difusão de cursos de verão, mestrados e doutorados erguidos sobre essa nova temática febril.

A tônica de todos esses cursos e publicações é fundamentalmente centrada na técnica como ferramenta geral para grandes volumes de

28 Jockers, 2013.

29 Karsdorp; Kestemont; Riddell, 2021.

30 Kokensparger, 2018.

31 Jemielniak, 2020.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

dados. A discussão teórica, além de ter pouco apelo, é geralmente tratada como um problema individual, o que não seria de todo ruim, se o debate existisse. O problema, como dito, é que todo esse movimento ocorre em paralelo a outro grave processo: a negação da teoria. Essa é a denúncia que faz Pierre Mounier ao analisar um texto como o de Chris Anderson, *The End of Theory*, no qual o autor explicitamente avalia que a grande quantidade de dados cada vez mais acessível tornará, em breve, o método científico obsoleto.[32] Não se trata de uma afirmação desconectada da realidade e do ambiente em que surgiram todas aquelas obras que fizemos referência — ou mesmo aqueles cursos técnicos. O projeto *Time Machine* não tinha uma proposta estruturalmente distante dessa perspectiva.

A proposta de Anderson não é uma posição isolada. Há muitos engenheiros, matemáticos e estatísticos que apostam na capacidade das ferramentas estatísticas, operando mediante correlações e gerando conhecimento novo. A partir de certos padrões, seria possível obter insights "*born from the data*"[33] em uma nova era na qual "*the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory*".[34]

Mounier centra seu foco em outro caso, o dos *culturomics*, como o Google Ngram Viewer[35] e o muito recentemente lançado *Gallicagram*[36]. Nesses programas, milhões de livros são organizados e tornados pesquisáveis a partir de unidades básicas de texto, os *ngram*, que permitem identificar a frequência do uso de certos termos ao longo de certos períodos de tempo em *corpora* textuais imensos. Mounier discute o próprio exemplo dado por Michel *et al.*[37], o caso do termo "Chagall" em milhões de livros de língua alemã, dando ênfase para a queda desse termo no

32 Anderson, 2008.

33 Kitchin, 2014.

34 Kitchin, 2014.

35 Michel *et al.*, 2011, 2024.

36 Azoulay; Courson, 2021; Courson *et al.*, 2023.

37 Michel *et al.*, 2011.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

período do nazismo, quando a referência a artistas judeus, como Chagall, era restrita. Mounier destaca que essa associação só foi possível graças não ao grande volume de dados do Google Ngram Viewer, mas ao conhecimento prévio feito com base em trabalhos não necessariamente digitais ou quantitativos. Como faríamos para interpretar achados estatísticos de processos sobre os quais não temos a menor ideia? Como interpretar, por exemplo, uma possível queda nos registros de compra e venda em cartórios de Veneza (para retomar o caso anterior)? Como uma queda nos negócios ou nos registros de negócios, apenas para dar duas possibilidades bem simples?

Ao fim e ao cabo, a análise de *big data* não permite qualquer crítica documental, pois não temos a menor ideia de como ela foi efetivamente produzida, já que ela se baseia na relação dos dados brutos entre si. Entender a existência de um documento histórico, sua criação e sua preservação ao longo do tempo implica elucidar as diversas seleções realizadas socialmente ao longo do tempo, fazendo com que algumas coisas sejam lembradas e outras tantas esquecidas. Os Arquivos, assim como os livros nas bibliotecas (para retomar os casos do *Google Ngram Viewer* e do *Venice Time Machine*), não são amostras representativas do mundo ou do passado, mas construções sociais geradas por uma sociedade em transformação. Como disse Mounier, "*Ce n'est pas l'absence d'exhaustivité des données qui compte, c'est le fait que leur constitution est le résultat d'une intention humaine*".[38]

A ameaça do *big data* é bastante visível se pensamos, como foi feito aqui, como uma ameaça ao método científico e à noção de teoria, mais especificamente. Por trás disso tudo, existe uma noção bastante difusa entre os historiadores, que é a separação entre teoria e técnica. Essa noção fora a base da experiência do *Venice Time Machine*, mas é muito mais difundida. Ela está presente em todos aqueles manuais de programação para historiadores e cursos a que fizemos referência. Ela é finalmente aceita, pois muitos historiadores terceirizam suas atividades de levantamento documental e, mais recentemente, de processamento de

38 Mounier, 2018.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

dados com o uso de ferramentas digitais. Adotar um *software* que faça o trabalho parece algo absolutamente condizente com a pesquisa tal como ela sempre foi feita. Nisso, não há nenhum problema. Os problemas se colocam quando não sabemos ao certo o que o software faz.

Caminhando para uma conclusão (com um jogo)

Uma das técnicas mais usadas no meio das Humanidades Digitais, que aparece com força em diversos livros de história digital, é o chamado *Topic Modeling* (ou extração de temas). Trata-se de uma técnica que promete identificar palavras-chave para um texto a partir do texto em si, como que tentando adivinhar o assunto do escrito sem que seja necessária a leitura por parte de um ser humano. É uma promessa interessante, que poderá ajudar os pesquisadores a escolher suas leituras a partir de uma filtragem prévia por palavras-chave. A ideia de usar palavras-chave é antiga, mas ela sempre dependia de uma ação humana prévia, fosse por parte da autoria ou atribuídas por profissionais de bibliotecas. A promessa agora é automatizar essa tarefa, humanamente impossível no atual contexto de produção frenética e publicação acelerada.[39]

O leitor que nos acompanhou até aqui é agora convidado a um jogo. Pensar em palavras-chave para o texto que acabou de ler. O artigo publicado tem aquelas palavras atribuídas pelo autor, mas quem lê é convidado a reavaliar essa escolha, sempre muito subjetiva. Feita essa tarefa, vejamos como dois algoritmos resolveram o mesmo problema. O primeiro é aquele disponível no *website* "nocodefunctions.com"[40], criado pelo professor de economia da *Emlyon Business School*, Clément Levallois, que tem algumas ferramentas digitais consideradas excelentes por alguns estudos.[41] O resultado teve vários níveis de resposta, mas os primeiros dois, mais confiáveis, apontaram as seguintes expressões:

39 Graham; Milligan; Weingart, 2016.

40 Levallois, 2024.

41 Ribeiro *et al.*, 2016; Levallois, 2013.

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

tópico 0: ideia, menor ideia, venice time machine, análise, google, documental

tópico 1: projeto, documentos, ferramenta, transkribus, fontes

Talvez o leitor esteja frustrado ou procurando alguma razão para certas palavras que estão nessas listas de palavras-chave. A descrição sobre o funcionamento do algoritmo pode ser encontrada direto no *website* de processamento. Antes de condenar (ou dar alguma razão) para os resultados do algoritmo de Levallois, vejamos outro resultado, produzido por um algoritmo criado pelo autor deste artigo com o uso da linguagem de programação python. Trata-se de um código que despreza tudo o que não for substantivo e busca as palavras mais relacionadas entre si — as que mais aparecem estando na mesma frase —, o que não significa que sejam as mais frequentes. O resultado foi o conjunto de palavras formado por "projeto", "ferramenta", "historiador", "dado", "documento" e "história".

Na opinião do leitor, esses resultados seriam totalmente desprezíveis? Erraram feio? Tem alguma proximidade com o texto lido? A dificuldade não está aqui em julgar a eficiência da ferramenta. A questão é que o resultado terá sempre alguma parcela de verossimilhança e outra dose de absurdo, sendo sua classificação como "bom" ou "ruim" uma decisão para a subjetividade de cada um. Para quem acabou de ler o texto, o julgamento é fácil. Mas e para quem não leu? Pois é para estes últimos que as análises de *big data* foram construídas, e essa tendência não parece dar sinais de arrefecimento. Estamos preparados para isso? Temos conhecimentos para fazer a crítica desses instrumentos que se tornam cada dia mais presentes? Estamos discutindo tudo isso nas universidades e centros de pesquisa? Ou seria o analfabetismo digital a saída mais aconselhável?

Bibliografia

ANDERSON, Chris. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, [*s. l.*], 2008. Disponível em: https://www.wired.com/2008/06/pb-theory/. Acesso em: 14 mar. 2024.

AZOULAY, Benjamin; COURSON, Benoît De. Gallicagram: un outil de lexicométrie pour la recherche. *SocArXiv*, [*s. l.*], 8 dez. 2021. doi: https://doi.org/10.31235/

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

osf.io/84bf3.

CASTELVECCHI, Davide. Venice 'Time Machine' Project Suspended amid Data Row. *Nature*, [*s. l.*], v. 574, n. 7780, 31 out. 2019, p. 607. doi: https://doi.org/10.1038/d41586-019-03240-w.

COURSON, Benoît De *et al.* Gallicagram : les archives de presse sous les rotatives de la statistique textuelle. *Corpus*, [*s. l.*], n. 24, 15 jan. 2023. doi: https://doi.org/10.4000/corpus.7944.

DE CERTEAU, Michel. A operação histórica. *In*: LE GOFF, Jacques; NORA, Pierre (org.). *História:* novos problemas. São Paulo: Livraria Francisco Alves Editora, 1978.

EHRMANN, Maud *et al*. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406 [cs]*, 23 set. 2021. Disponível em: http://arxiv.org/abs/2109.11406. Acesso em: 14 mar. 2024.

GINZBURG, Carlo; PROSPERI, Adriano. *Giochi di pazienza*: un seminario sul Beneficio di Cristo. Vol. 258. Torino: Einaudi, 1975.

"Google Books Ngram Viewer", 2024. https://books.google.com/ngrams/.

GRAHAM, Shawn; MILLIGAN, Ian; WEINGART, Scott. *Exploring Big Historical Data*: The Historian's Macroscope. [*S. l.*]: World Scientific Publishing Company, 2016.

GRAHAM, Shawn; MILLIGAN, Ian; WEINGART, Scott. *Exploring Big Historical Data*: The Historian's Macroscope. [*S. l.*]: World Scientific Publishing Company, 2016.

GULDI, Jo; ARMITAGE, David. *The history manifesto*. Cambridge, United Kingdom: Cambridge University Press, 2014.

HOMERO. *Odisséia Em Versos*. São Paulo: Ediouro, 1997.

JEMIELNIAK, Dariusz. *Thick big data*: doing digital social sciences. New product. New York: Oxford University Press, 2020.

JOCKERS, Matthew Lee. *Macroanalysis: digital methods and literary history*. Topics in the digital humanities. Urbana: University of Illinois Press, 2013.

KAHLE, Philip *et al.*. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *In*: IAPR INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR), 14., 2017, Kyoto. *Anais* [...]. Kyoto: IEEE, 2017. p. 19-24. https://doi.org/10.1109/ICDAR.2017.307.

KAPLAN, Frédéric. The Venice Time Machine. *In*: ACM SYMPOSIUM ON DOCUMENT ENGINEERING, 2015, Lausanne Switzerland. *Proceedings* [...]. Lausanne Switzerland: ACM, 2015. p. 73. doi: https://doi.org/10.1145/2682571.2797071.

KARSDORP, Folgert; KESTEMONT, Mike; RIDDELL, Allen. *Humanities data analysis*:

Almanack, Guarulhos, n. 36, ep00124, 2024
http://doi.org/10.1590/2236-463336ep00124

case studies with Python. Princeton: Princeton University Press, 2021.

KITCHIN, Rob. Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, [*s. l.*], v. 1, n. 1, 1 abr. 2014. doi: https://doi.org/10.1177/2053951714528481.

KITCHIN, Rob. Positivistic geography and spatial science". *In*: KITCHIN, Rob. *Approaches to human geography*. London: Sage, 2006. p. 20-29. Disponível em: https://scholar.google.com/citations?view_op=view_citation&hl=pt-PT&user=Y_3-GBQAAAAJ&cstart=200&pagesize=100&sortby=pubdate&citation_for_view=Y_3-GBQAAAAJ:hEXC_dOfxuUC. Acesso em: 14 mar. 2024.

KOKENSPARGER, Brian. *Guide to Programming for the Digital Humanities*: Lessons for Introductory Python. New York: Springer, 2018.

LEVALLOIS, Clément. *Nocode Functions*, 2024. Disponível em: https://nocodefunctions.com/. Acesso em: 14 mar. 2024.

LEVALLOIS, Clement. Umigon: sentiment analysis for tweets based on lexicons and heuristics. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, SEMEVAL, 2013, Atlanta. *Proceedings* [...]. Atlanta: [*s. l.*], 2013. p. 414-417. Disponível em: https://scholar.google.com/scholar?cluster=9774134509927520 80&hl=en&oi=scholarr. Acesso em: 14 mar. 2024.

MICHEL, Jean-Baptiste *et al*. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, [*s. l.*], v. 331, n. 6014, p. p. 176-182, 14 jan. 2011. doi: https://doi.org/10.1126/science.1199644.

MINISTERIO DELLA CULTURA. Archivio di Stato di Venezia. *Sospensione dei rapporti con EPFL su Time Machine*, 2019. Disponível em: https://www.archiviodistatovenezia.it/it/eventi/news/sospensione-dei-rapporti-con-epfl-su-time-machine.html. Acesso em: 14 mar. 2024.

MORETTI, Franco. *Atlas of the European Novel, 1800-1900*. Verso, 1999. Disponível em: https://books.google.com.br/books?id=ja2MUXS_YQUC. Acesso em: 14 mar. 2024.

MORETTI, Franco. *Graphs, Maps, Trees*: Abstract Models for a Literary History. Verso, 2005. Disponível em: https://books.google.com.br/books?id=YL2kvMIF8hEC. Acesso em: 14 mar. 2024.

MOUNIER, Pierre. *Les humanités numériques*: Une histoire critique. Paris: Éditions de la Maison des sciences de l'homme, 2018. doi: https://doi.org/10.4000/books.editionsmsh.12006.

O'NEIL, Cathy. *Weapons of math destruction*: how big data increases inequality and threatens democracy. 1st. ed. New York: Crown, 2016.

RIBEIRO, Filipe Nunes *et al*. SentiBench - a benchmark comparison of state-of-

-the-practice sentiment analysis methods. *arXiv*, [*s. l.*], 14 jul. 2016. Disponível em: http://arxiv.org/abs/1512.01818. Acesso em: 14 mar. 2024.

STAUDER, Florian. A Short History of Transkribus with Günter Mühlberger. *READ--COOP*, 22 fev. 2023. Disponível em: https://readcoop.eu/a-short-history-of--transkribus-with-gunter-muhlberger/. Acesso em: 14 mar. 2024.

TIME MACHINE EUROPE. *About Us*, 2024. Disponível em: https://www.timemachine.eu/about-us/. Acesso em: 14 mar. 2024.