Sete maneiras de usar computadores para entender textos[a]

Por que, afinal, humanistas usam computadores para trabalhar com textos?

Autor: Ted Underwood[b]

Apresentação[c]: O artigo do Prof. Ted Underwood, Ph.D., apresenta conceitos introdutórios para aqueles que querem começar a utilizar ferramentas computacionais em pesquisas na área de Humanidades: de uma breve recapitulação sobre a nomenclatura do campo "Humanidades Digitais" até exemplos de modelagem, o artigo exemplifica usos do *AntConc*, nuvens de palavras, *Google Ngram*, maneiras de representar textos e distingui-los utilizando ferramentas e maneiras diferentes de visualização. O texto original apresenta tom instrucional mais leve, de leitura fluida, o que procuramos manter nesta versão em Língua Portuguesa, algo não comumente visto na tradição acadêmica brasileira, mas que serve muito bem aos propósitos do artigo.

Por que, afinal, humanistas usam computadores para trabalhar com textos?

Parte do foco da nomenclatura "Humanidades Digitais" é designar a tecnologia da informação como algo que pertence às humanidades – e não uma invasora de outro campo.

E é verdade, a interpretação humanística sempre teve uma dimensão tecnológica: nós organizávamos a produção escrita em *commonplace books*[d] (fichamentos) antes mesmo da busca por palavras-chave.

Mas colocar as novas oportunidades de pesquisa como um movimento específico das humanidades chamadas Humanidades Digitais tem o problema de obscurecer o cenário maior. Métodos computacionais estão transformando tanto as ciências sociais e naturais quanto as humanidades, e eles fazem isso através, parcialmente, da criação de novos diálogos entre disciplinas.

Uma das maiores mudanças que máquinas trazem para análise de textos é a mediação de novas conexões para a ciência social. Modelos estatísticos que ajudam sociólogos a compreender estratificação e mudança social não haviam contribuído tanto no passado por que seria difícil conectar modelos quantitativos às evidências ricas, porém amiúde, fornecidas pela escrita. Mas essa barreira está se dissolvendo.

Conforme novos métodos tornam mais fácil a representação de textos não estruturados em modelos estatísticos, várias perguntas fascinantes surgem para cientistas sociais e humanistas. [O'Connor et. al. 2011]

Resumidamente, análise computacional de textos não é uma nova tecnologia específica

---

[a] N.T. Tradução do artigo: UNDERWOOD, Ted. **Seven ways humanists are using computers to understand text.** 2015. The Stone and the Shell. Disponível em: <https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/>. Acesso em: 15 de outubro de 2018.

[b] N.T. Ted Underwood, Ph.D., é Professor da Universidade de Illinois, lecionando Ciências de Informação e Inglês.

[c] N.T. Apresentação escrita pela tradutora.

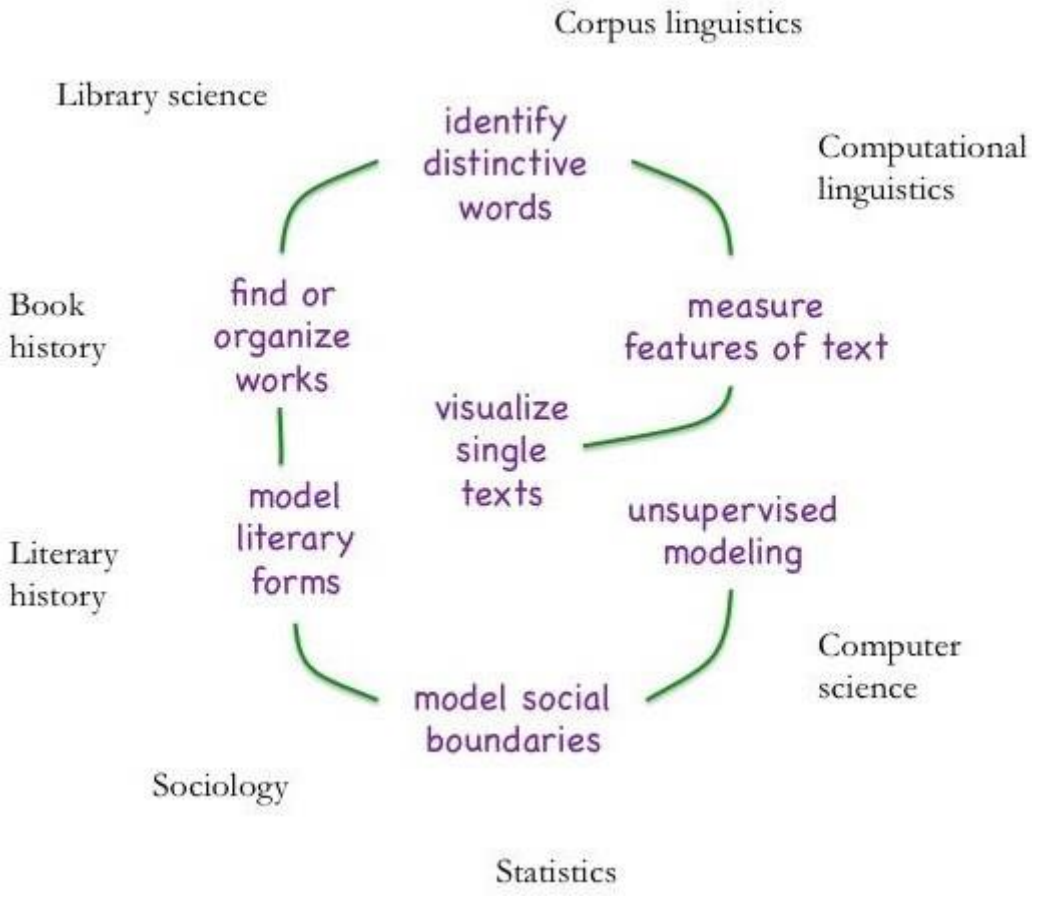
[d] N.T. *Commonplace books* são cadernos de notas, fichamentos e anotações bibliográficas que podem ser organizados de diferentes maneiras.

ou um subcampo das Humanidades Digitais, é uma conversa completamente receptiva em um espaço interdisciplinar. Humanistas, muitas vezes, abordam essa conversa esperando encontrar ferramentas que automatizem tarefas já conhecidas, o que é um bom começo. Mencionarei ferramentas que você pode usar para criar concordância ou uma nuvem de palavras. Não há problemas em parar aí. Formas mais complexas de análise textual começam a parecer uma Ciência Social, e humanistas não estão sob a obrigação de chegar nesse ponto.

Eu devo alertá-los, porém, de que ferramentas digitais abrem as portas para "coisas mais pesadas". A chamada análise de texto ou leitura distante é, de fato, uma conversa interdisciplinar sobre métodos, e se você se afundar nela, talvez descubra que quer tentar diversos itens ainda não categorizados como ferramentas.

O que podemos fazer?

A imagem abaixo é um mapa de algumas coisas que você pode fazer com o texto (inspirado, embora diferente, do mapa de "Humanidades Digitais" de Alan Liu). A ideia é dar uma noção geral de como diferentes atividades estão relacionadas a diferentes tradições disciplinares. O fluxograma vai do centro às extremidades, em espiral, mas essa é apenas uma maneira de organizar a discussão e não necessariamente sugere uma sequência no fluxo de trabalho.



1. Visualizar textos separadamente

Análise textual às vezes é representada como parte de uma "nova moda" nas humanidades. Geralmente, essa noção é incorreta. A maioria dos métodos descritos nessa postagem objetivam revelar padrões que ficam escondidos de leitores individuais – o que tem muito pouco de "moda". Porém há algumas formas de análise categorizadas como leituras superficiais, pois elas visualizam padrões textuais abertos à inspeção direta. Por exemplo, os adorados *cartoons* de Randall Munroe que mostram *plots* de filmes populares ao apontar quais personagens ficam juntos em determinados pontos da narrativa.

Esses *cartoons* revelam pouco que não sabemos. Eles são divertidos, em parte, por que as narrativas abordadas são familiares: nós podemos redescobrir um material já conhecido de uma maneira gráfica que torna fácil ampliar ou reduzir o escopo que damos a detalhes. Gráficos de redes que conectam personagens são legais por um motivo similar. Ainda é discutível o que (se algo) eles revelam, e é importante notar que redes de relacionamentos ficcionais funcionam diferentemente das da realidade [Elson, et al., 2010]. Mas as pessoas as acham interessantes.
Em algum nível, não existe nada que não pudesse ser descoberto por um leitor atento, mas a crítica as considera útil. Se você quer fazer a concordância para alguma obra (ou para uma biblioteca inteira), *AntConc* é uma ferramenta útil. Estratégias de visualização são um tópico que merecem uma discussão separada.

2. Escolher características para representar textos

Um pesquisador lidando com a análise computacional de um texto precisa responder duas questões. Primeiro: como irá representar os textos? Segundo: o que irá fazer com essas representações quando as conseguir? Muito do que vem a seguir aborda a segunda questão, por que há uma infinidade de respostas para a primeira – e elas não necessariamente limitam as da segunda.

Na prática, textos são comumente representados por contagem de várias palavras que eles possuem (tratados como "sacos de palavras"). Como essa representação é radicalmente diferente de uma leitura sequencial comum, as pessoas tendem a ficar surpresas com sua funcionalidade. Porém, o objetivo da análise computacional não é reproduzir modelos de entendimento que leitores já alcançaram. Se estamos tentando revelar padrões em larga escala que não estariam evidentes para leitores comuns, não é necessário retraçar os padrões sintáticos que organizam a compreensão de um trecho específico. Na realidade, diversas questões macroscópicas são registradas no nível da escolha de palavras: autoria, tema, gênero, público alvo, etc. A própria popularidade do *Google Ngram* mostra que achamos a frequência de palavras interessante.

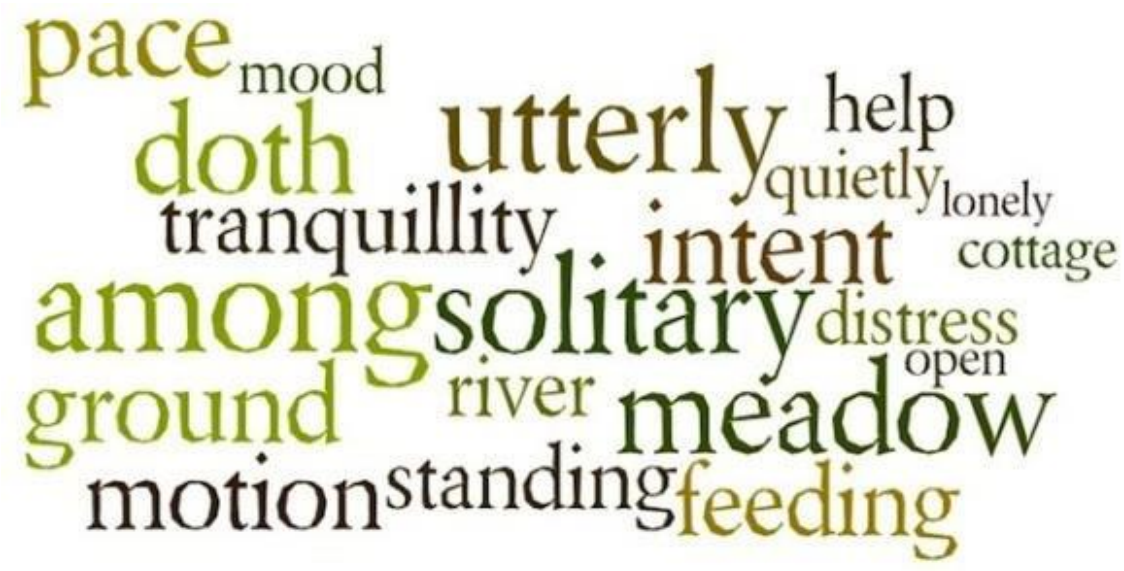
Há diversas outras maneiras de representar textos. Você pode contar sentenças de duas palavras, ou até medir o espaço em branco. Informações qualitativas que não podem ser mesuradas são representadas como uma variável categórica. Também é possível considerar a sintaxe, se necessário. Linguistas computacionais estão ficando muito bons em analisar sentenças e muitos de seus insights estão compilados em projetos

acessíveis como o *Natural Language Toolkit*. Certamente haverá questões de pesquisa – envolvendo, por exemplo, o conceito de caractere – que irão requerer análise sintática. Elas são, porém, questões para quem está começando na área.

3. Identificar vocabulário distintivo

Pode ser bem fácil ter insights úteis no nível da dicção. São essas as alegações que acadêmicos da literatura há muito tempo fazem: *Norton Anthology of English Literature* prova que William Wordsworth emblematiza a alienação romântica usando palavras como "solitário, por si só, sozinho" [Greenblatt et. al., 16].
Pesquisadores aprenderam a tomar cuidado com essas afirmações. Creio que Wordsworth escreva a palavra "sozinho", mas ele realmente a usa mais do que outros escritores? "Sozinho" é uma palavra comum. Como distinguimos insights reais de reflexões pouco significativas?

A linguística de corpus desenvolveu várias maneiras de identificar se alguma palavra de fato é mais presente em uma amostra em comparação com outras. Uma das mais usadas é *Dunning's log-likelihood: Ben Schmidt has explained why it works*, acessível online pelo Voyant ou para download pelo *AntConc*. Então, se você tem uma amostra de um autor (como Wordsworth) e um corpus de referência para contrastá-la (como uma coletânea de outros poemas), é bem simples de identificar termos que tipificam a obra de Wordsworth em relação a outras obras. (Existem maneiras de mesurar o alto uso de uma palavra; Ada Kilgarriff recomenda um teste Mann-Whitney). Realmente há boas evidências que mostram que "solitário" é uma das palavras que distinguem o trabalho de Wordsworth.

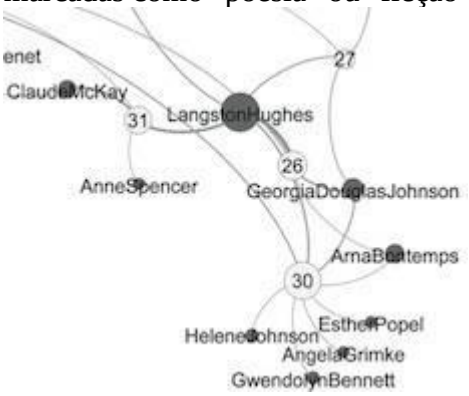


Palavras que são consistentemente mais comuns em trabalhos de William Wordsworth do que em outros poetas de 1780 a 1850. Usei os gráficos do Wordle, mas as palavras foram selecionadas por um teste de Mann-Whitney, que mede a sobrerrepresentação em relação a um contexto - não pelo método (sem contexto) do Wordle.

Também é fácil de organizar os resultados em uma nuvem de palavras – se você quiser. Pessoas tiram sarro das nuvens, com alguma razão; elas são chamativas mas não oferecem muita informação. Eu as uso em postagens de blog, pela extravagância, mas não faria isso em um artigo.

4. Buscar e organizar

Esse subtítulo é um resumo do enorme número de maneiras que podemos usar a tecnologia da informação para organizar coleções de material escrito ou orientar nossos olhos no espaço discursivo. Humanistas já fazem isso o tempo todo, é claro: confiamos muito na pesquisa na web, bem como na pesquisa de palavras-chave em catálogos de bibliotecas e bancos de dados de texto completo.
Mas nosso conjunto atual de estratégias pode não ser suficiente para todas as coisas que queremos encontrar. Isso é óbvio para historiadores, que trabalham extensivamente com material não publicado. Porém, o mesmo é verdade para livros impressos: obras de poesia ou ficção publicadas antes de 1960, por exemplo, muitas vezes não são marcadas como "poesia" ou "ficção".



Detalhe da Fig. 7 em *So and Long*, "Network Analysis and the Sociology of Modernism".

Mesmo se acreditássemos que a tarefa de encontrar coisas estivesse resolvida, ainda precisaríamos de maneiras de mapear ou organizar coleções. Um interessante tópico de pesquisa nos últimos anos envolveu o mapeamento das conexões sociais concretas que organizam a produção literária. Natalie Houston mapeou conexões entre poetas vitorianos e editoras; Hoyt Long e Richard Jean So mostraram como os escritores são listados por publicação nos mesmos periódicos [Houston 2014; So and Long 2013].

Existem, é claro, centenas de outras maneiras pelas quais humanistas podem querer organizar material. Mapas costumam ser usados para visualizar referências a locais de publicação. Outra abordagem óbvia é agrupar os trabalhos por algum tipo de similaridade textual.
Não há ferramentas criadas especificamente para dar suporte a grande parte desse trabalho. Existem ferramentas para construir visualizações, mas muitas vezes a maior parte do problema é encontrar ou construir os metadados de que você precisa.

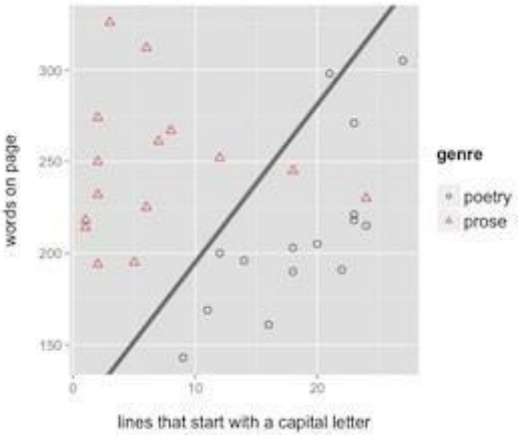
5. Modelar formas e gêneros literários

Um modelo é uma representação simplificada de algo e, em princípio, pode ser construído com palavras, madeira ou qualquer material de sua preferência. Na prática, nas ciências sociais, os modelos estatísticos são frequentemente equações que descrevem a probabilidade da associação entre variáveis. Muitas vezes, a "variável de

resposta" é o que se está tentando entender (forma literária, comportamento de voto, etc.), e as "variáveis preditoras" são as que você suspeita que podem ajudar a explicar o fenômeno.

Esta não é a única maneira de abordar a análise de textos. Historicamente, os humanistas tenderam a começar escolhendo algum aspecto do texto para medir primeiro, e depois elaboravam algum argumento em relação ao significado do que mediam. Eu fiz isso e, de fato, pode funcionar. Porém os cientistas sociais preferem o caminho contrário: identificar um conceito que se está tentando entender e depois tentar modelá-lo. Pode-se dizer que sua abordagem parece excessivamente sistemática.

Construir modelos pode ajudar humanistas de várias maneiras. Classicamente, os cientistas sociais modelam conceitos para entendê-los melhor. Se você está tentando entender a diferença entre dois gêneros ou formas, criar um modelo pode ajudar a identificar o que os diferencia.
Estudiosos também podem enquadrar modelos de gêneros totalmente novos: Andrew Piper faz em um ensaio recente sobre o "romance de conversação".



Um modelo estatístico imaginário simples que distingue páginas de poesia de páginas de prosa

Em outros casos, o motivo da modelagem não será realmente descrever ou explicar um conceito, mas simplesmente reconhecê-lo em escalas maiores. Descobri que precisava criar modelos preditivos simplesmente para encontrar a ficção, a poesia e o drama em uma coleção de 850.000 volumes.

A preferência entre modelar para explicar e modelar para prever tem sido discutida extensamente em outras disciplinas [Shmueli, 2010]. Porém, os modelos estatísticos ainda não foram amplamente utilizados em pesquisas históricas e humanistas podem encontrar formas de usá-los que não sejam comuns em outras disciplinas. Por exemplo, uma vez que modelamos um fenômeno, podemos questionar sobre a estabilidade diacrônica do seu padrão. (Um modelo treinado para reconhecer gêneros de uma década específica faz previsões igualmente boas sobre a próxima década?)

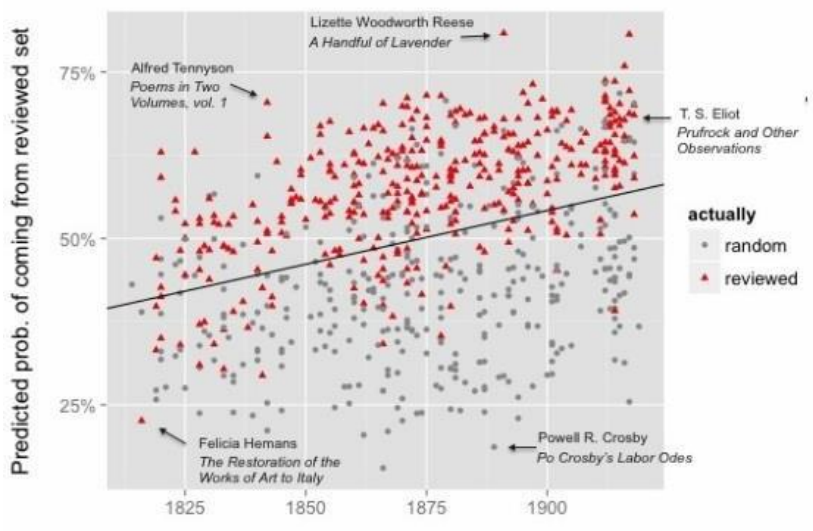
Existem muitos pacotes de software que podem ajudar a inferir os modelos dos seus dados. Mas avaliar a validade e a adequação de um modelo é mais complicado. É importante entender completamente os métodos que estamos usando, e isso provavelmente exigirá um pouco de leitura. Pode-se começar compreendendo as

suposições implícitas em modelos lineares simples e trabalhar com modelos mais complexos produzidos por algoritmos de aprendizado de máquina [Sculley e Pasanek, 2008].

Particularmente, porém, é importante aprender sobre o problema do "*overfitting*". Parte da razão pela qual os modelos estatísticos estão se tornando mais úteis nas humanidades é que novos métodos tornam possível o uso de centenas ou milhares de variáveis, para representar texto não estruturado (esses pacotes de palavras tendem a conter muitas variáveis). No entanto, um grande número de variáveis aumenta o risco de haver uma sobrecarga de dados, e você precisa saber como evitar isso.

6. Modelar barreiras sociais

Não há razão para que os modelos estatísticos de texto fiquem restritos a questões de gênero e de forma. Textos também estão envolvidos em todos os tipos de transações sociais, e esses contextos são frequentemente identificáveis na escrita.
Por exemplo, Jordan Sellers e eu estudamos recentemente a história da distinção literária treinando modelos para distinguir entre poesia revisada em periódicos de elite e uma seleção aleatória de volumes extraídos de uma biblioteca digital. Há muitas coisas que podemos aprender fazendo isso, mas o resultado mais relevante é que os padrões implícitos que distinguem o discurso poético da elite se mostram relativamente estáveis ao longo do século.



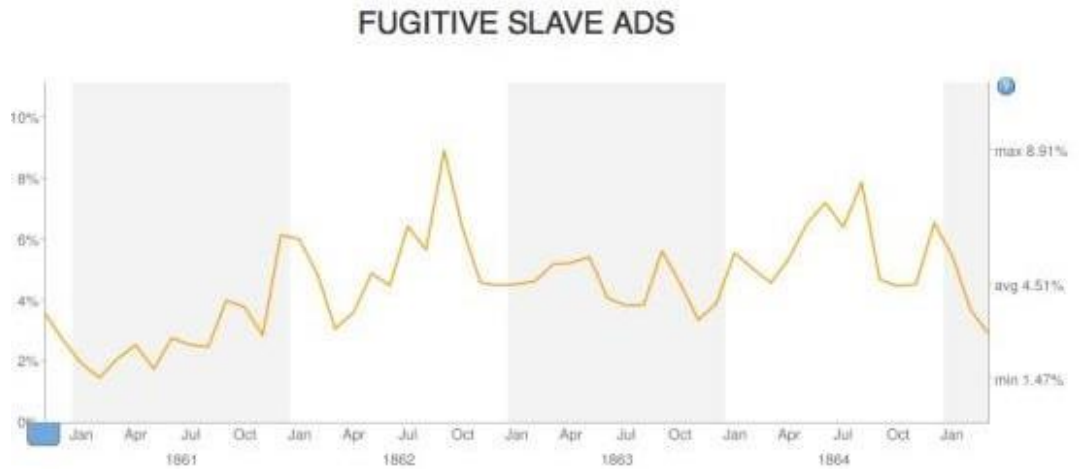
Questões semelhantes poderiam ser formuladas sobre a história política ou legal.

7. Modelagem não supervisionada

Os modelos que discutimos até agora são supervisionados no sentido de que eles têm um objetivo explícito. Digamos que você já saiba quais romances foram revisados em periódicos proeminentes e quais não foram, você está desenvolvendo um modelo para descobrir se existem padrões nos próprios textos que podem nos ajudar a explicar esse limite social ou a traçar sua história.
Mas os avanços no aprendizado de máquina também tornaram possível treinar modelos não supervisionados. Supomos que você use uma coleção de textos desconhecidos e peça a um algoritmo para organizar a coleção encontrando agrupamentos ou padrões

de algum tipo especificado de forma imprecisa. Você não necessariamente sabe quais padrões surgirão.

Se isso soa epistemologicamente arriscado, você não está errado. Como o círculo hermenêutico não nos permite obter algo do nada, a modelagem não supervisionada envolve, inevitavelmente, muitas suposições (explícitas). Pode, no entanto, ser extremamente útil como uma heurística exploratória e, às vezes, como base para argumentos. Uma família de algoritmos não supervisionados chamada "modelagem de tópico" tem atraído muita atenção nos últimos anos, tanto de cientistas sociais quanto de humanistas. Robert K. Nelson usou a modelagem de tópicos, por exemplo, para identificar padrões de publicação em um jornal da era da Guerra Civil de Richmond.



Estou colocando modelos não supervisionados no final desta lista porque eles podem ser atrativos demais. A modelagem de tópicos é perfeitamente projetada para workshops e demonstrações, já que você não precisa começar com uma pergunta de pesquisa específica. Um grupo de pessoas com interesses diferentes pode simplesmente derramar uma coleção de textos no computador, reunir-se e ver quais padrões emergem.

Geralmente, padrões interessantes surgem: a modelagem de tópicos pode ser uma ferramenta poderosa para a descoberta. Mas seria um erro considerar este fluxo de trabalho como paradigmático para análise de texto. Normalmente, os pesquisadores iniciam com perguntas de pesquisa específicas e, por essa razão, suspeito que muitas vezes preferimos modelos supervisionados.

Em suma, há muitas coisas novas que os humanistas podem fazer com o texto, desde novas versões de coisas que sempre fizemos, até modelar experimentos que nos aprofundam no terreno metodológico das ciências sociais. Alguns desses projetos podem ser realizados com uma ferramenta simples, mas os mais ambiciosos exigem um pouco de familiaridade com um ambiente de análise de dados como o *Rstudio*, ou até mesmo uma linguagem de programação como *Python*, e, mais importante, com as suposições que permeiam a ciência social quantitativa. Por essa razão, não espero que esses métodos se tornem universalmente difundidos nas humanidades tão logo. Em princípio, tudo o que está acima é acessível para alunos de graduação com um ou dois semestres de preparação - mas não é o tipo de preparação que cursos de Inglês ou História garantam.

Geralmente deixo as postagens do blog intocadas para documentar o que acontecia na época. Mas as coisas estão mudando rapidamente, e é muito trabalhoso revisar completamente uma postagem de pesquisa como essa a cada poucos anos, então neste caso eu posso continuar alterando e adicionando informações com o passar do tempo. Vou sinalizar minhas edições com a data entre colchetes.

Como citar esse recurso: UNDERWOOD, T. Sete maneiras de usar computadores para entender textos. Tradução de Mariana L. Souza. The Stone and the Shell, 04 jun. 2015. Título original: Seven ways humanists are using computers to understand text. Disponível em: https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/. Acesso em: 14 jan. 2019.

Bibliografia Selecionada

ELSON, David K.; DAMES, Nicholas; MCKEOWN, Kathleen R.. Extracting Social Networks from Literary Fiction. **Proceedings Of The 48th Annual Meeting Of The Association For Computational Linguistics**, Uppsala, Suécia, p.138-147, 2010.

GREENBLATT, Stephen. **Norton Anthology of English Literature.** 8. ed. New York: Ww Norton, 2006.

HOUSTON. Toward a Computational Analysis of Victorian Poetics. **Victorian Studies**, [s.l.], v. 56, n. 3, p.498-510, 2014. Indiana University Press. http://dx.doi.org/10.2979/victorianstudies.56.3.498.

NOWVISKIE, Bethany Paige. **Speculative Computing:** Instruments for Interpretative Scholarship. Charlottesville: University Of Virginia, 2004.

O'CONNOR, Brendan; BAMMAN, David; SMITH, Noah A. Computational Text Analysis for Social Science: Model Assumptions and Complexity. **Nips Workshop On Computational Social Science,** p.1-10, dez. 2011.

PIPER, Andrew. Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel. **New Literary History**, [s.l.], v. 46, n. 1, p.63-98, 2015. Johns Hopkins University Press. http://dx.doi.org/10.1353/nlh.2015.0008.

SCULLEY, D.; PASANEK, Bradley M. Meaning and mining: the impact of implicit assumptions in data mining for the humanities. **Digital Scholarship In Humanities**, v. 23, n. 4, p.409-424, 12 set. 2008.

SHMUELI, Galit. To Explain or to Predict? **Statistical Science**, [s.l.], v. 25, n. 3, p.289-310, ago. 2010. Institute of Mathematical Statistics. http://dx.doi.org/10.1214/10-sts330.

SO, R. J.; LONG, H.. Network Analysis and the Sociology of Modernism. **Boundary 2**, [s.l.], v. 40, n. 2, p.147-182, 1 jun. 2013. Duke University Press.

http://dx.doi.org/10.1215/01903659-2151839.

STALLYBRASS, Peter. Against Thinking. **Modern Language Association**, v. 122, n. 5, p.1580-1587, jan. 2007.

WILLIAMS, Jeffrey J. **The New Modesty in Literary Criticism.** Disponível em: <https://www.chronicle.com/article/The-New-Modesty-in-Literary/150993>. Acesso em: 05 jan. 2015.