Convergências: estudos em Humanidades Digitais

VOYANT TOOLS E R NA ANÁLISE LITERÁRIA À DISTÂNCIA EM LÍNGUA PORTUGUESA

Voyant tools and R in distance reading literary analysis in Portuguese

Voyant tools y R en el análisis literario a distancia en lengua portuguesa

Diego Giménez[1]

Resumo: O presente artigo objetiva comparar duas ferramentas de leitura distante, utilizando como exemplo corpora da literatura em português. A leitura distante, tal e qual foi definida por Franco Moretti, é um marco hermenêutico e metodológico que permite estudar grandes volumes de texto de maneira quantitativa. Voyant Tools é um site em código aberto que permite a análise textual online e que não requer programação. Quanteda (*Quantitative Analysis of Textual Data*), por sua vez, é um pacote de R para a mineração de dados, também em código aberto. Foi desenvolvido para usuários de R que precisam aplicar processamentos de linguagem natural a textos. Para poder utilizar Quanteda mediante R, são precisos conhecimentos básicos de programação. O artigo, desta maneira, procura comparar as ferramentas citadas, contribuir para a escolha informada de instrumentos de análise textual e fornecer uma base para futuras pesquisas e aplicações na análise literária. Paralelamente, o texto busca problematizar os processos de modelagem, análise e representação, que são construções de conhecimento e, portanto, sujeitas à interpretação.
Palavras-chave*:* Leitura distante. Voyant Tools. Quanteda. R. Literatura.

Abstract: The present article aims to compare two distant reading tools using Portuguese literature corpora as an example. Distant reading, as defined by Franco Moretti, is a hermeneutic and methodological framework that allows for the quantitative study of large volumes of text. Voyant Tools is an open-source web platform that enables online textual analysis without the need for programming. Quanteda (Quantitative Analysis of Textual Data), on the other hand, is an R package for data mining, also open source. It was developed for R users who need to apply natural language processing to texts. To use Quanteda through R, basic programming knowledge is required. The article, therefore, seeks to compare these tools, contribute to informed choices in textual analysis tools, and provide a foundation for future research and applications in literary analysis. Additionally, the text aims to problematize the processes of

[1] Doutor em Estudos Literários. Professor assistente no Departamento de Português da Faculdade de Letras da Universidade de Macau, Macau, (RAEM) China. E-mail: dgimenez@um.edu.mo; Lattes: http://lattes.cnpq.br/5256384887401634; Orcid iD: https://orcid.org/0000-0002-1229-3969.

modeling, analysis, and representation, which are knowledge constructions and, as such, subject to interpretation.
Keywords*:* Distant Reading. Voyant Tools. Quanteda. R. Literature.

Resumen: El presente artículo tiene como objetivo comparar dos herramientas de lectura distante utilizando como ejemplo corpus de la literatura en portugués. La lectura distante, tal como fue definida por Franco Moretti, es un marco hermenéutico y metodológico que permite estudiar grandes volúmenes de texto de manera cuantitativa. Voyant Tools es un sitio web de código abierto que permite el análisis textual en línea y no requiere programación. Quanteda (Quantitative Analysis of Textual Data), por su parte, es un paquete de R para la minería de datos, también de código abierto. Fue desarrollado para usuarios de R que necesitan aplicar procesamiento del lenguaje natural a textos. Para poder utilizar Quanteda mediante R, se requieren conocimientos básicos de programación. De este modo, el artículo busca comparar las herramientas mencionadas, contribuir a la elección informada de herramientas de análisis textual y proporcionar una base para futuras investigaciones y aplicaciones en el análisis literario. Paralelamente, el texto busca problematizar los procesos de modelado, análisis y representación, que son construcciones de conocimiento y, por lo tanto, están sujetos a interpretación.
Palabras clave: Lectura distante. Voyant Tools. Quanteda. R. Literature.

Introdução

No ano 2000, Franco Moretti, em "Conjectures on World Literature", empregou o termo *distant reading* como uma condição de conhecimento:

> *Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, Less is more. If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor. But it's precisely this 'poverty' that makes it possible to handle them, and therefore to know. This is why less is actually more* (Moretti, 2000, p. 57).

No artigo, Moretti argumenta que para estudar a literatura mundial de forma efetiva, é preciso adotar métodos que vão além do *close reading* e que considerem a rede de relações entre literaturas globais. Em 2005, publica *Graphs, Maps, Trees: Abstract Models for a Literary History* (2005), onde desenvolve com maior detalhe o conceito de "leitura distante": "'*Distant reading', I have once called this type of approach; where distance is however not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models*" (Moretti, 2005, p. 1). Para o autor, uma abordagem quantitativa, semelhante à utilizada na história social, poderia

revolucionar os estudos literários ao examinar um corpus maior de obras. Este método permitiria aos investigadores identificarem padrões e tendências dentro da grande massa de fatos, em vez de depender de leituras detalhadas de alguns poucos textos selecionados. O *close reading* é criticado pelo especialista por sua inadequação em lidar com o vasto número de obras literárias que existem. O texto, assim, aponta que mesmo um cânone amplo, como duzentos romances para a Grã-Bretanha do século XIX, representa apenas uma pequena fração do número total de romances publicados durante esse período. Isso torna impraticável entender o campo literário apenas pela leitura atenta, pois esta não abrange o vasto volume de material disponível. A leitura atenta de casos individuais é insuficiente para compreender de forma abrangente um campo tão vasto. O *close reading*, segundo Moretti, resulta ineficaz para compreender sistemas coletivos e padrões mais amplos dentro do campo literário, pois foca na análise detalhada de textos individuais em vez das tendências e estruturas maiores. Uma leitura deste tipo permite aos pesquisadores identificarem padrões e tendências na produção literária que a leitura atenta não pode revelar.

A investigação quantitativa, para Moretti, "*provides a type of data which is ideally independent of interpretations*" (2005, p. 9). Esta independência é simultaneamente uma força e uma limitação. Os dados não fornecem inerentemente as estruturas hermenêuticas necessárias para compreender o significado de padrões, embora os dados quantitativos possam oferecer medições objetivas. Neste ponto, concorda-se parcialmente com Moretti. Enquanto a interpretação dos dados é necessária, também é importante salientar que eles são construções não independentes de interpretação. Tanto a forma de coletar os corpora quanto o algoritmo que os analisa modelam e criam conhecimento. Irá demonstrar-se este argumento ao comparar a forma de selecionar o corpus e de definir as palavras irrelevantes. Do mesmo modo, distintos algoritmos para analisar o *topic modeling* oferecem resultados diferentes.

Moretti gerou diferentes reações com o seu livro, as quais podem ser consultadas na obra *Reading Graphs, Maps, Trees: Responses to Franco Moretti* (2011), editado por Jonathan Goodwin e John Holbo para Parlor Press.

Em 2019, Ted Underwood publicou *Distant Horizons*, onde tenta demonstrar como os arquivos digitais e as ferramentas estatísticas, em vez de reduzir palavras a números, podem aprofundar a compreensão de questões que sempre foram centrais para a investigação humanística. Sem negar, como Moretti, a utilidade de abordagens tradicionais como a leitura

atenta, os estudos narrativos ou os estudos de gênero, Underwood argumenta que também é preciso ler os grandes arcos da mudança literária que têm permanecido ocultos pela sua enorme escala. Utilizando tanto a leitura atenta quanto a distante para rastrear a diferenciação de gêneros, transformação de papéis de gênero e de juízo estético, Underwood demonstra como as ferramentas e os métodos digitais podem chamar a atenção para um campo mais amplo da história literária.

Em Portugal, em 2020, foi publicado na revista *Matlit*, o artigo "Leitura distante em português: resumo do Primeiro Encontro", escrito por vários autores a partir de um encontro que aconteceu em Oslo em outubro de 2019. Para os autores, a leitura distante é uma área interdisciplinar que combina Estudos Literários, Linguística Computacional e Informática Aplicada para analisar grandes coleções de textos. Focada inicialmente em textos literários, a leitura distante expandiu-se para outras fontes, contribuindo para a digitalização e o tratamento virtual de obras literárias, além do desenvolvimento de ferramentas e métodos de extração de informação. Contrapondo-se à leitura atenta, que foca em aspectos qualitativos internos de uma obra, a leitura distante prioriza a perspectiva quantitativa, analisando grandes volumes de obras para testar caracterizações históricas e literárias. A relevância da leitura distante reside na combinação de história literária interpretativa com análise quantitativa, abrindo novas perspectivas para teoria e história literárias e estudos comparativos entre diferentes períodos, línguas e culturas. Para os autores, desde sua proposição por Franco Moretti, a área tem sido desenvolvida por diversos pesquisadores como Mathew Jockers, Geoffrey Rockwell, Stéfan Sinclair, Andrew Piper e Katherine Bode, que, embora críticos em relação a Moretti, contribuíram significativamente para o debate. Com raízes históricas profundas, a leitura distante não é um conceito totalmente novo, mas sua popularização trouxe novas dimensões para os Estudos Literários. Para um entendimento mais profundo, os autores recomendam o trabalho de Ted Underwood, *A Genealogy of Distant Reading* (2017). No contexto português, destacam o volume publicado por Cabral em 2014.

Não é o foco deste artigo traçar a genealogia da leitura distante, mas apenas contextualizar as ferramentas que irão ser comparadas. Objetiva-se, assim, inserir a descrição das ferramentas dentro do debate teórico descrito. As ferramentas de análise, como mencionado no título e no resumo, são o *Voyant Tools*, disponível online, e o Quanteda, que usa a linguagem R de programação. O primeiro não requer conhecimentos de programação por parte do utilizador; o segundo, sim.

Voyant Tools

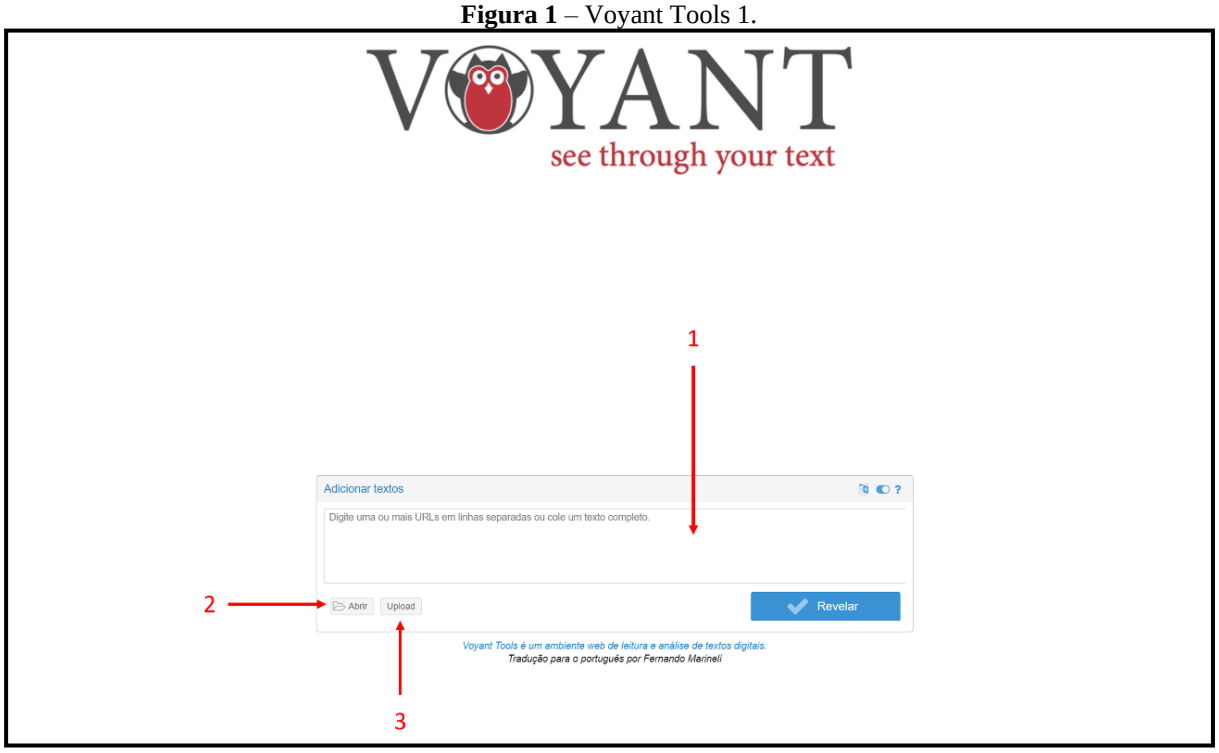
Voyant Tools[2] é uma ferramenta em código aberto que permite a análise textual acessível de forma online. Não são precisos conhecimentos de programação para efetuar as análises textuais. A plataforma pode ser utilizada em vários idiomas e conta com funcionalidades que permitem a análise de diferentes tamanhos e grupos de texto. A análise é feita em *back-end*. A privacidade é mantida por meio da coleta de dados para melhorar as ferramentas e para fins de pesquisa, com os dados dos utilizadores armazenados e analisados de forma anônima, segundo consta na informação da plataforma. O *back-end* refere-se ao mecanismo responsável pelo processamento e pela análise dos dados. Neste caso, envolvem-se as análises computacionais de texto realizadas pelo *Voyant Tools*. O *front-end*, por outro lado, seria a interface com a qual os utilizadores interagem sem terem presente o código.

Corpus

O primeiro passo antes de utilizar o programa é ter um corpus que possa ser analisado pela ferramenta. Existem três formas de carregá-lo: colar o texto completo ou uma URL ao texto na janela "Adicionar textos" (número 1 da imagem a seguir); usar modelos de exemplos do *Voyant* (2), nomeadamente, a obra completa de Shakespeare, a obra completa de Jane Austen ou o *Frankenstein* de Mary Shelley; carregar ficheiros de txt., doc., ou pdf. com OCR na aba "upload" (3).

[2] Disponível em: https://voyant-tools.org/ Acesso em: 27 jul. 2025.

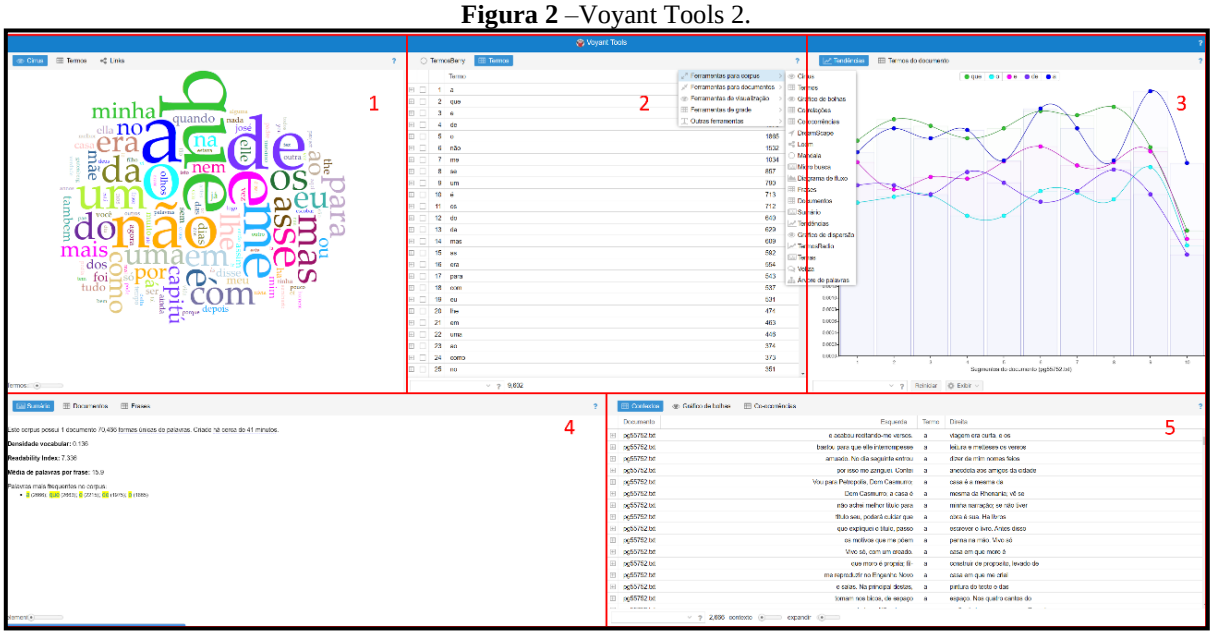
Figura 1 – Voyant Tools 1.



Fonte: https://voyant-tools.org/.

A preparação do corpus para a mineração de dados é muito importante, pois os resultados da análise variam conforme o estado do corpus. O tipo de informação que contém o ficheiro antes da análise pode afetar o resultado do cômputo. Uma vez carregado o corpus na plataforma, o programa faz a análise do texto e oferece, num primeiro momento, resultados por meio da visualização de cinco ferramentas, das múltiplas opções que estão divididas em 5 grupos: “Ferramentas para corpus”; “Ferramentas para documentos”; “Ferramentas de visualização”; “Ferramentas de grade”; e “Outras ferramentas”. Algumas destas partilham grupo.

Figura 2 –Voyant Tools 2.



Fonte: https://voyant-tools.org/.

A plataforma, como consta na imagem (*Voyant Tools 2*), está dividida em cinco visualizações que aparecem por defeito quando se carrega o corpus e que podem ser substituídas a qualquer momento pelo utilizador. A primeira das abas (1) oferece uma nuvem de palavras, ou *cirrus*, na qual constam as palavras mais frequentes, sendo as de maior tamanho as mais citadas. Na segunda (2), vê-se a lista dos termos mais citados no corpus. Na terceira (3), consta um gráfico de tendência, que representa as frequências dos termos em documentos de um corpus ou em segmentos de um documento. Cada série é colorida de acordo com a palavra que representa, e no topo do gráfico uma legenda refere quais palavras estão associadas a quais cores. Na quarta (4), há um sumário que apresenta informações gerais sobre o corpus, como total de palavras, densidade vocabular, média de palavras por frase, índice de legibilidade, palavras mais frequentes, etc. Na quinta janela (5), constam palavras-chave em contexto, em que são exibidas cada ocorrência de uma palavra-chave com um trecho de texto circundante. O utilizador pode modificar cada uma das abas com visualizações mediante as opções que cada janela apresenta. É possível, nas opções de cada visualização tanto exportar a imagem como substituir a ferramenta, definir suas opções e obter sua informação detalhada.

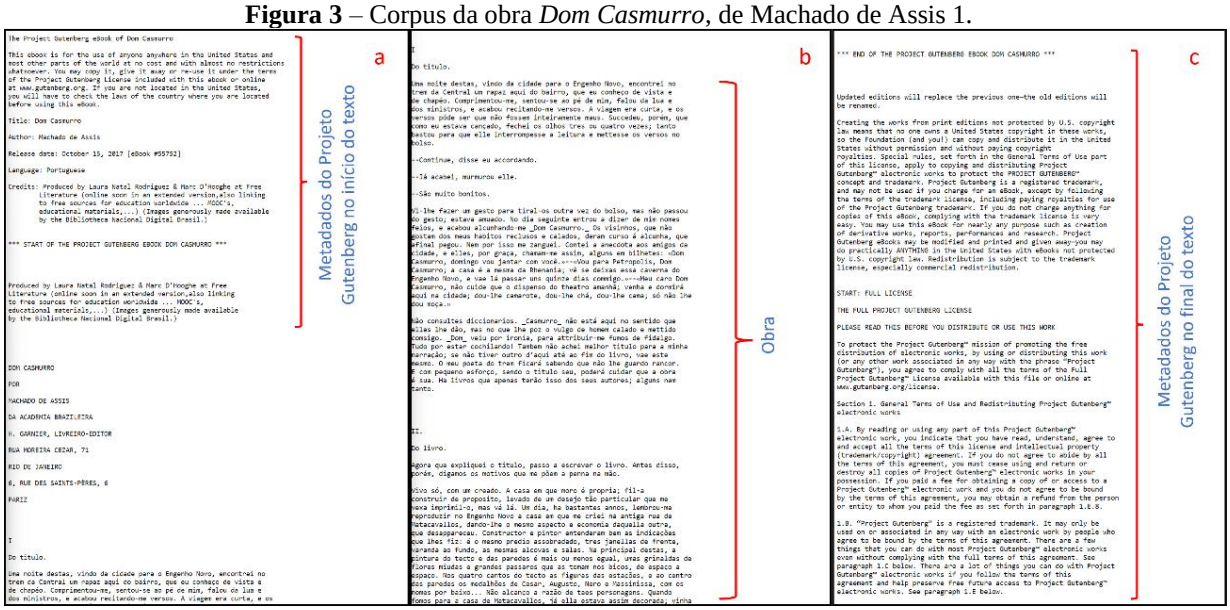
O corpus carregado no exemplo da imagem "*Voyant Tools 2*" corresponde à obra *Dom Casmurro*, de Machado de Assis. O utilizador terá de definir e escolher as ferramentas que irão ser utilizadas na análise. Com o programa, podem ser analisados corpora compostos por uma

ou por várias obras, do mesmo ou de diferentes autores. Um dos primeiros passos, em qualquer caso, depois de carregar o corpus, é ajustar os termos que contemplarão os cômputos e as representações. Por exemplo, podem ser retiradas da representação palavras comuns e consideradas irrelevantes (*stopwords*) de cada língua, como, por exemplo, artigos, preposições, etc. Também podem ser editadas as listas de palavras indesejadas, acrescentando aqueles elementos que não se queira que sejam contemplados no estudo.

Análise do corpus

Vê-se, neste sentido, o exemplo a partir de corpora extraídos do Projeto Gutenberg[3], uma biblioteca digital que oferece acesso gratuito a uma vasta coleção de obras[4]. Pode-se carregar o corpus quer colando a URL ou o texto completo (1, na imagem "*Voyant Tools 1*") quer carregando um ficheiro txt., (2, na imagem "*Voyant Tools 1*"). Na seguinte imagem, mostra-se como aparece a opção *Plain Text UTF-8* da obra *Dom Casmurro* (2024), de Machado de Assis, no Projeto Gutenberg[5].

Figura 3 – Corpus da obra *Dom Casmurro*, de Machado de Assis 1.



Fonte: Projeto Gutenberg.

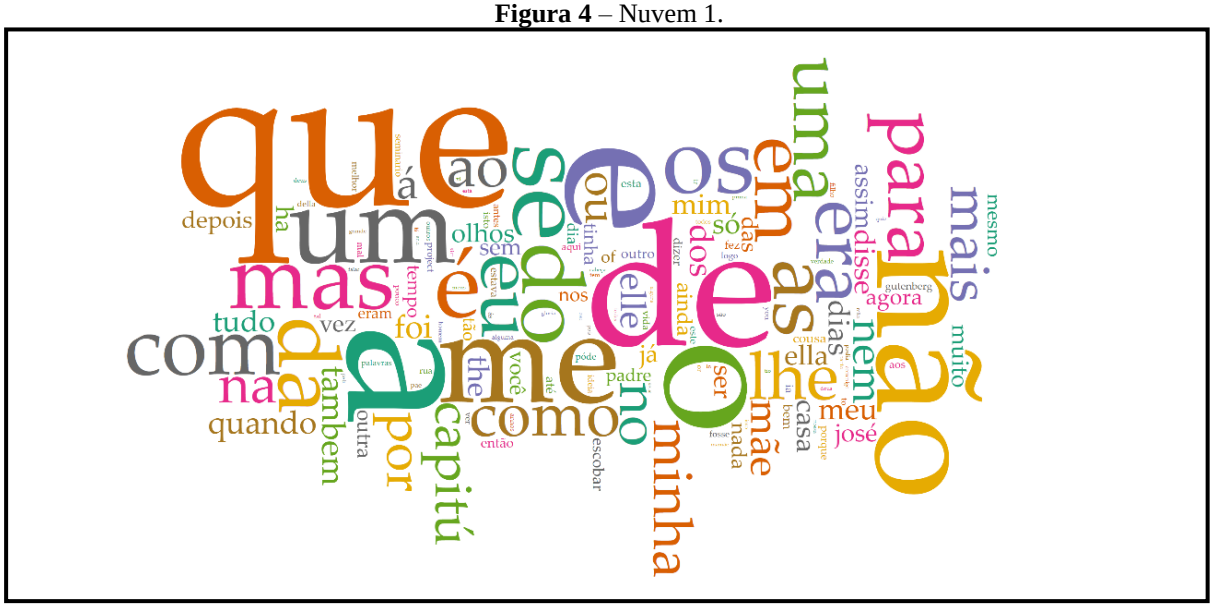
[3] Repositórios de livros eletrônicos gratuitos fundado em 1971 por Michael S. Hart. Disponível em: https://www.gutenberg.org/ Acesso em: 27 jul. 2025.

[4] O maior repositório online de obras literárias dos Países Lusófonos é a Biblioteca Digital de Literatura de Países Lusófonos. Trata-se de uma gerida pelo Núcleo de Pesquisa em Informática, Literatura e Linguística da Universidade Federal de Santa Catarina (NUPILL-UFSC). Disponível em: https://www.literaturabrasileira.ufsc.br/ Acesso em: 27 jul. 2025.

[5] Disponível em: https://www.gutenberg.org/cache/epub/55752/pg55752.txt Acesso em: 27 jul. 2025.

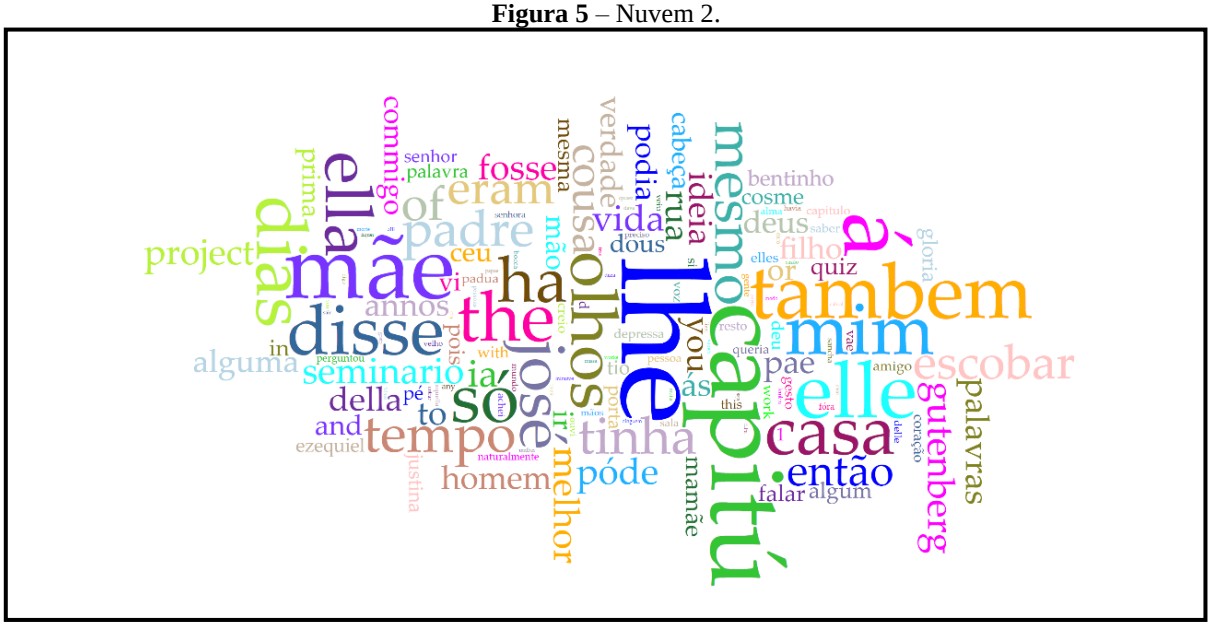
Ao colar diretamente o endereço URL, não é possível tirar os metadados relacionados com o Projeto Gutenberg (“a” e “c”). Portanto, estes dados irão ser analisados pelo programa. Outra opção é copiar e colar apenas o texto da obra (b) ou guardar o texto em um ficheiro txt. e fazer o carregamento uma vez limpos os metadados. Esse passo é mais importante do que parece, e demonstra como a preparação do corpus, dos dados, condiciona a representação destes. Quando Moretti disse que a análise quantitativa oferece dados objetivos, está obviando os processos de modelagem, que criam sentido. Basta, assim, comparar apenas as visualizações das nuvens de palavras mais citadas segundo três formas de trabalhar o corpus:

Figura 4 – Nuvem 1.



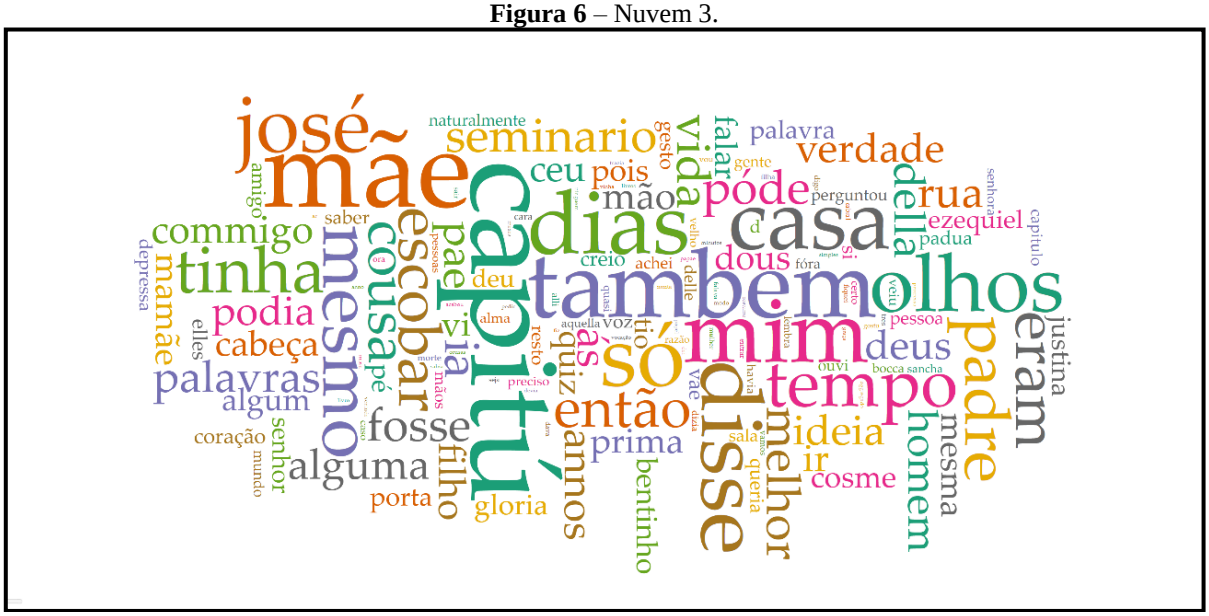
Fonte: https://voyant-tools.org/.

Figura 5 – Nuvem 2.



Fonte: https://voyant-tools.org/.

Figura 6 – Nuvem 3.



Fonte: https://voyant-tools.org/.

As três representações de palavras mediante nuvens correspondem ao mesmo corpus, que foi tratado de forma diferente. Na primeira imagem (Nuvem 1), representa-se o corpus tal e como consta no endereço *Plain Text UTF-8* da obra de Machado no Projeto Gutenberg. Portanto, além da obra, constam também metadados relacionados ao projeto. Na segunda imagem (Nuvem 2), foram retiradas as palavras comuns do português (artigos, preposições e alguns pronomes). Chama a atenção na segunda imagem, por um lado, que há palavras em inglês e, por outro, que há palavras com uma ortografia antiga. As palavras em inglês (*the*,

Gutenberg, project…) correspondem aos metadados mencionados, que são alheios à obra de Machado. As palavras com outra ortografia podem corresponder a obras anteriores aos diferentes acordos ortográficos. No caso da obra de Machado, distinguem-se consoantes geminadas (delle, della…) e algumas preposições com acentos agudos que, por não estarem no dicionário atualizado de palavras comuns, constam na análise e na representação. Na terceira imagem (Nuvem 3), foram editadas as listas de palavras não desejadas na análise, nomeadamente, as que estão em inglês relacionadas com os metadados e as palavras comuns em uma ortografia anterior ao último acordo ortográfico.

Este exemplo pode parecer uma questão menor, mas demonstra claramente como a preparação do corpus e o algoritmo condicionam o resultado da análise e da representação. Por esse motivo, argumenta-se contra Moretti que os dados não são objetivos, como pretende o investigador. Os dados e as representações são construções que supõem escolhas de interpretação. Assim, acredita-se que é absolutamente imprescindível que investigações que utilizem este tipo de ferramentas descrevam minuciosamente a forma como prepararam o corpus e os programas e/ou algoritmos que utilizaram nas análises. O Voyant Tools permite exportar as análises em diferentes formatos para serem utilizadas e citadas, como em HTML para usar numa página web. Além disso, pode-se exportar as referências da visualização segundo as normas de referência bibliográfica MLA, Chicago, APA, ou *Bib Tex*. Também é possível criar um caderno de notas (*Spyral Notebook*), no qual se visualiza parte do código e descrições que podem ser guardados em perfis de usuário no *GitHub*.

Resumo

O Voyant Tools, assim, resulta uma ferramenta completa com muitas opções de análise e visualização que permitem estudar textos sem necessidade de ter conhecimentos de programação. Crê-se que é importante que um utilizador que pretenda publicar os resultados de análise com o programa descreva bem tanto a preparação do corpus quanto as ferramentas e o modo como foram preparadas para poder partilhar como modelou e representou os dados, especificando, desta maneira, os processos de interpretação que acompanham a cada etapa. Por exemplo, a partir da obra que foi estudada: (a) criou-se o corpus a partir da obra *Dom Casmurro*,

de Machado de Assis, no repositório do Projeto Gutenberg, na opção *Plain Text UTF-8*[6]; (b) colou-se a ligação na plataforma Voyant; (c) realizou-se a primeira análise; (d) retirou-se as palavras indesejadas, por defeito, do português; (e) editou-se a lista com os elementos indesejados, incluindo os termos em inglês provenientes dos metadados e das palavras em ortografia antiga; (f) descarregou-se as análises e representações de dados; (g) se serão utilizadas em publicações, pode-se exportar o código e as referências bibliográficas da análise ou criar um caderno de notas a ser partilhado em uma conta *GitHub*.

Os pontos negativos do programa podem ser enumerados da seguinte forma: (1) a plataforma depende da conexão à internet. Se a ligação falhar, não há forma de realizar as análises; (2) por ser em *back-end*, o utilizador não tem controle total sobre o processo de análise. O acesso ao algoritmo é parcial; (3) em caso de utilização de um corpus que não consta na internet mediante ligação URL, e seja preciso carregá-lo a partir de ficheiros pessoais, não fica claro o destino desse corpus dentro da plataforma. Esse elemento é relevante segundo o nível de privacidade que se pretenda com os dados que estão sendo empregados.

Quanteda

R é uma linguagem de programação criada por Ross Ihaka e Robert Gentleman. O R é utilizado por mineradores de dados para análise e desenvolvimento de software estatístico. O ambiente de software R oficial é um *software* livre de código aberto, disponível sob a Licença Pública Geral GNU, e faz parte do pacote GNU[7]. Ele é escrito principalmente em C, Fortran e R (parcialmente auto-hospedado). Entre os seus vários pacotes especializados de R, destaca-se o Quanteda, utilizado para a manipulação e análise de dados textuais[8]. Ele foi criado por Kenneth Benoit, Kohei Watanabe e outros colaboradores, em código aberto, para os utilizadores de R que necessitam aplicar processamento de linguagem natural a textos.

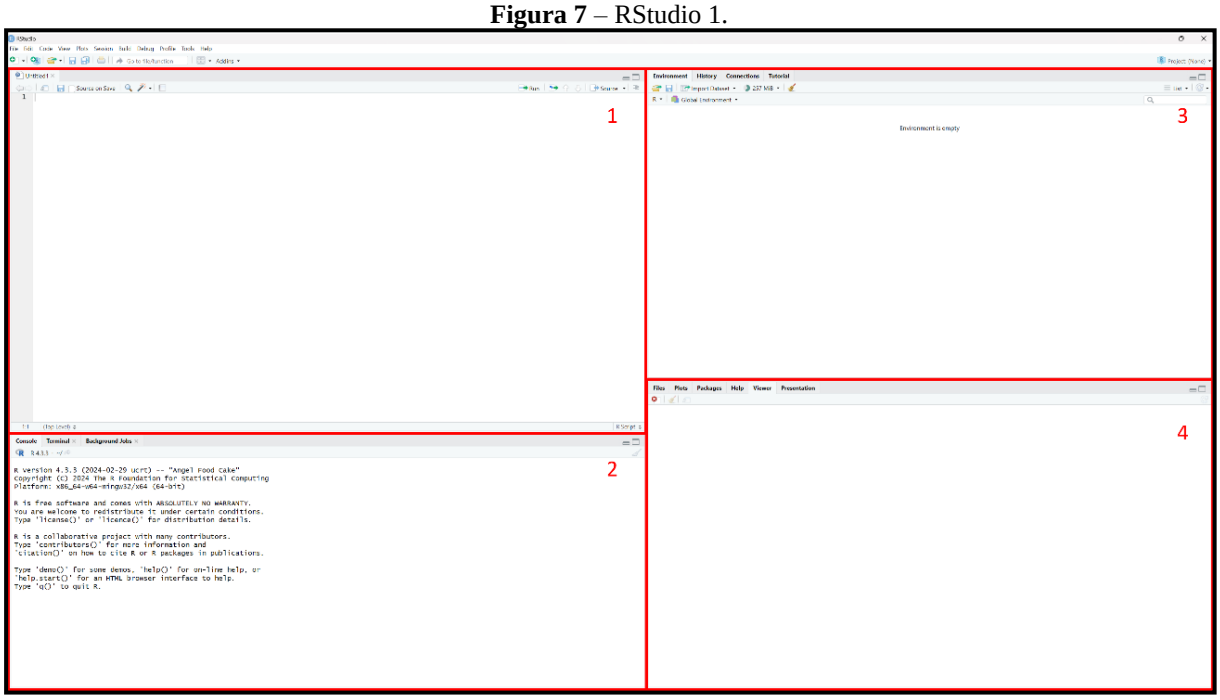
[6] https://www.gutenberg.org/cache/epub/55752/pg55752.txt (acesso em: 27 jul. 2025).
[7] O GNU é um projeto de software livre iniciado por Richard Stallman em 1983 e cujo propósito foi o desenvolvimento de um sistema operativo completo e de código aberto, semelhante ao Unix, mas composto exclusivamente por software livre. O ambiente oficial de software R integra o pacote GNU e está disponível sob a Licença Pública Geral GNU, permitindo a sua utilização, modificação e distribuição de forma livre.
[8] Pacotes e bibliotecas no contexto do R são conjuntos de ferramentas que os utilizadores podem aplicar para expandir as funcionalidades da linguagem. Um pacote consiste em código, dados e documentação agregados com o propósito de simplificar a execução de tarefas específicas. As bibliotecas, por sua vez, são coleções de pacotes que podem ser carregadas conforme necessário para oferecer funcionalidades adicionais.

Instalação

Para utilizar o R, devem-se instalar os programas R, RStudio e Rtools (no caso de usuários de Windows). R é a linguagem de programação base. No site oficial do CRAN[9] pode ser descarregado o programa. Uma vez instalado o R, a seguir pode-se descarregar o RStudio, uma plataforma de desenvolvimento integrado (IDE) para R, disponível no site oficial do RStudio[10]. A versão gratuita (RStudio Desktop) é suficiente para a maioria dos utilizadores.

No caso de uso do sistema operativo *Windows*, será necessário instalar o *RTools*[11] para compilar pacotes que requerem compilação de código. O *RTools* pode ser obtido no site oficial do *RTools*[12]. Ao abrir o *RStudio* pela primeira vez, este deve detectar automaticamente a instalação do R. Uma vez feitas as instalações e configurações, o *RStudio* pode ser utilizado para escrever e executar código em R. Os utilizadores podem criar *scripts* (código), carregar pacotes, importar dados e realizar análises estatísticas e visualizações de dados.

Figura 7 – RStudio 1.



Fonte: RStudio (Diego Giménez).

[9] CRAN (*Comprehensive R Archive Network*) é uma rede de servidores que armazenam e distribuem software relacionado com R. Disponível em: https://cran.r-project.org/ Acesso em: 27 jul. 2025.
[10] Disponível em: https://www.rstudio.com/ Acesso em: 27 jul. 2025.
[11] O RTools é um conjunto de ferramentas necessário para compilar pacotes R em sistemas operativos Windows. O RTools é utilizado para garantir que o código pode ser compilado e executado corretamente.
[12] Disponível em: https://cran.r-project.org/bin/windows/Rtools/ Acesso em: 27 jul. 2025.

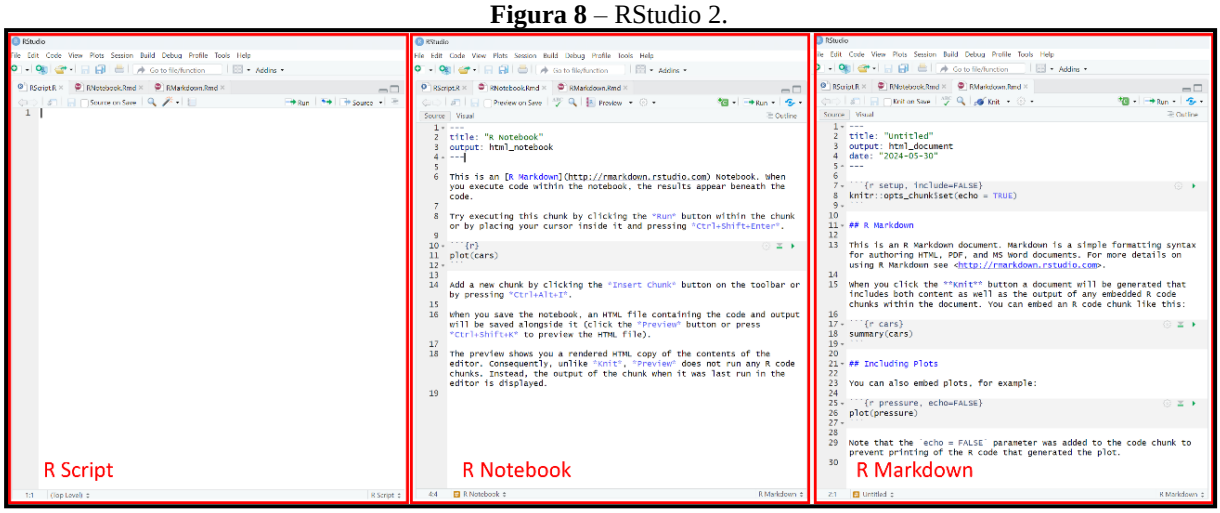
A plataforma do RStudio está composta por quatro janelas (imagem RStudio 1). A janela número 1 corresponde ao editor de *script*, onde se escreve e edita o código R. Podem ser criados *scripts*, abrir *scripts* existentes e executar linhas individuais ou todo o *script*. É onde se escreve o código. A número 2 é a console, onde o código é executado e os resultados são exibidos. Após executar um comando no editor de *script*, o resultado aparecerá no console. No número 3, constam ambiente, histórico, conexões e tutorial. O ambiente mostra os objetos (como variáveis, funções, etc.) que estão atualmente carregados na memória do R. Isso inclui os dados que carregou ou criou durante a sessão. O histórico mostra os comandos que executou durante a sessão atual do RStudio. As conexões permitem conectar-se a bancos de dados externos ou outros recursos de computação. No número 4, pode-se ver os ficheiros e as matrizes que a análise vai realizando. Na quarta janela, percebem-se o visualizador de ficheiros, os *plots* e pacotes. O visualizador permite examinar os detalhes dos objetos, como visualizar os dados em uma estrutura de dados, como um *dataframe*; ficheiros mostram os arquivos no diretório de trabalho atual e permite navegar entre eles; *plots*, quando são criados gráficos no R, eles aparecem neste guia; "pacotes" exibe informações sobre os pacotes instalados no seu ambiente R. Isso inclui os pacotes que carregou durante a sessão atual e os que estão disponíveis para instalação.

Utilização do RStudio

No RStudio, são utilizados diferentes tipos de ficheiros para diversas finalidades. O R Script, o R Notebook e o R Markdown podem ser interpretados como fases da investigação e da escrita em que primeiro se testa o código e a análise (R Script); a seguir se descrevem os passos junto com o código (R Notebook); e finalmente se prepara um texto para a publicação (R Markdown). Assim, o R Script, por exemplo, onde o código R é escrito e armazenado, é um ficheiro de texto simples. A extensão de ficheiro é ".R". O R Notebook, mediante a execução de blocos de código e a visualização imediata dos resultados, combina código R com texto explicativo. A extensão de ficheiro é ".Rmd". O R Markdown é utilizado para criar documentos que integram código R, texto explicativo e resultados de análises, como gráficos e tabelas. Através do R Markdown, podem ser gerados documentos em diversos formatos, incluindo HTML, PDF e Word. Este tipo de ficheiro é particularmente útil para a criação de relatórios,

apresentações e documentos técnicos que requerem a inclusão de código e resultados. A extensão de ficheiro também é ".Rmd".

Figura 8 – RStudio 2.



Fonte: RStudio (Diego Giménez).

Uma vez que se tem acesso ao RStudio e se conhecem os tipos de ficheiros, é preciso instalar dentro do RStudio tanto o pacote Quanteda quanto os restantes pacotes que serão precisos para efetuar as análises. Lembra-se que os pacotes são compostos por código, dados e documentação que foram reunidos para facilitar a realização de tarefas específicas. As bibliotecas são coleções de pacotes que podem ser carregadas conforme necessário para fornecer funcionalidades adicionais.

Existem duas maneiras de carregar os pacotes: mediante a opção "instalar pacotes", dentro de ferramentas (*tools*), no menu superior. "Instalar pacotes" leva ao diretório do CRAN; através de um *script* (código) no editor de *scripts*. A dificuldade inicial ao utilizar o R e o RStudio, para quem não tem conhecimentos de programação ou da linguagem em questão, é saber qual é precisamente o código. No caso dos estudos de literatura em português, pode-se consultar modelos de análise com código pronto e, assim, adaptar os códigos a diferentes projetos. Neste sentido, um modelo de utilização de R em textos literários pode ser consultado em Giménez e Gomide (2022)[13].

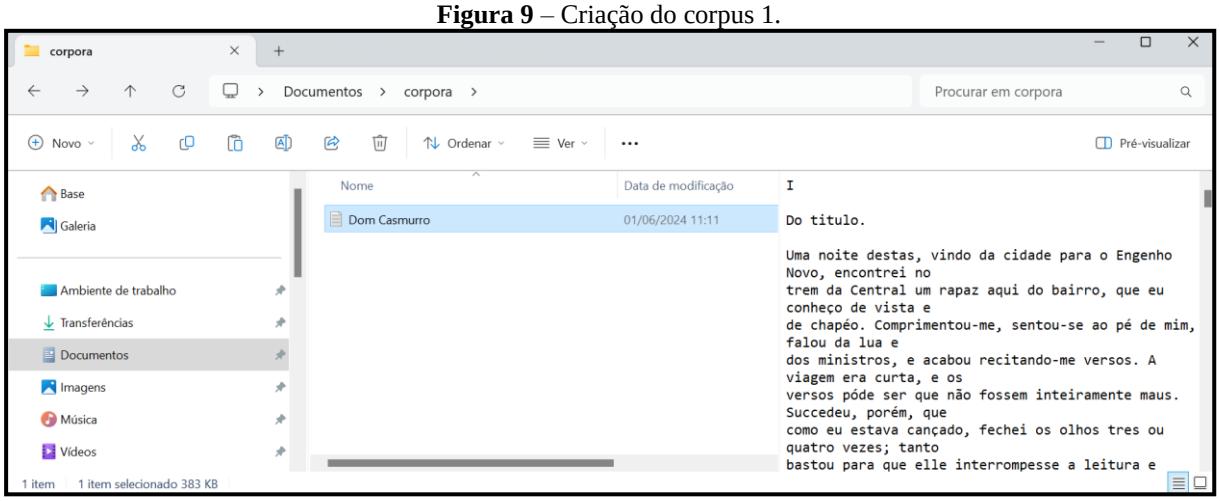
[13] Outros modelos de análise literária com R podem ser conferidos em Giménez (2024).

Análise com o RStudio

A utilização do RStudio e as primeiras análises passam, desse modo, por diferentes etapas: descarregar o *software*; preparar o ambiente descarregando os pacotes; e analisar os dados. Para esta última etapa, é preciso, em primeiro lugar possuir corpora que serão o objeto da análise. O corpus pode estar composto por um documento (*Dom Casmurro* de Machado de Assis, no exemplo deste texto) ou por um conjunto de documentos (mais de um livro de Machado de Assis). É importante detalhar a forma em que se prepara o corpus, isto é, se contém metadados relevantes ou não, da mesma maneira como foi descrito com o Voyant Tools.

Para este artigo, foi criado um ficheiro txt. que contém a obra *Dom Casmurro* sem metadados irrelevantes. Em português, pode-se descarregar as obras desde repositórios gratuitos como Projeto Gutenberg ou a Biblioteca Digital de Literatura de Países Lusófonos (UFSC), como já foi mencionado. Uma vez criado o ficheiro com a obra, criou-se na pasta "documentos" do computador, uma pasta com o nome "corpora", na qual se colocou o ficheiro com a obra de Machado.

Figura 9 – Criação do corpus 1.

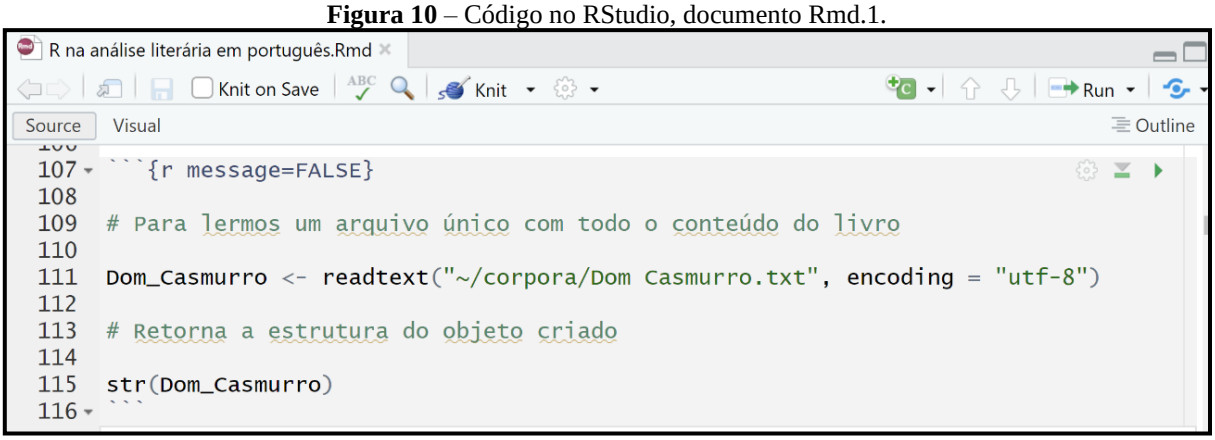


Fonte: RStudio (Diego Giménez).

Após descarregados os programas e criado o ficheiro com o texto a analisar, procede-se a dar indicações ao programa, mediante código, para que leia o ficheiro e o transforme em um corpus nos parâmetros de R. Uma vez instalado o R, o *Rtools* e o *RStudio* e tendo o corpus em uma pasta, basta carregar no *RStudio* o ficheiro Rmd, disponível nos repositórios mencionados (nota de rodapé 8), e seguir as indicações, adaptando-as ao corpus que se quiser analisar e especificando o diretório onde consta dito corpus. A seguir, descreve-se os passos mais

importantes para a execução das análises. Uma descrição completa do ficheiro excede os limites do presente texto. Um dos primeiros passos, assim, é carregar o corpus em R:

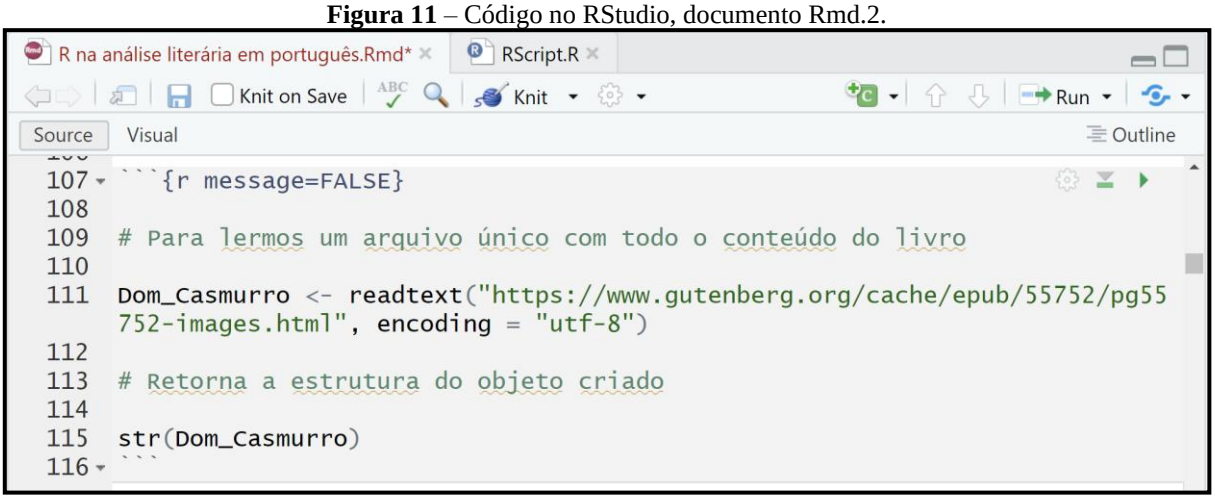
Figura 10 – Código no RStudio, documento Rmd.1.



Fonte: RStudio (Diego Giménez).

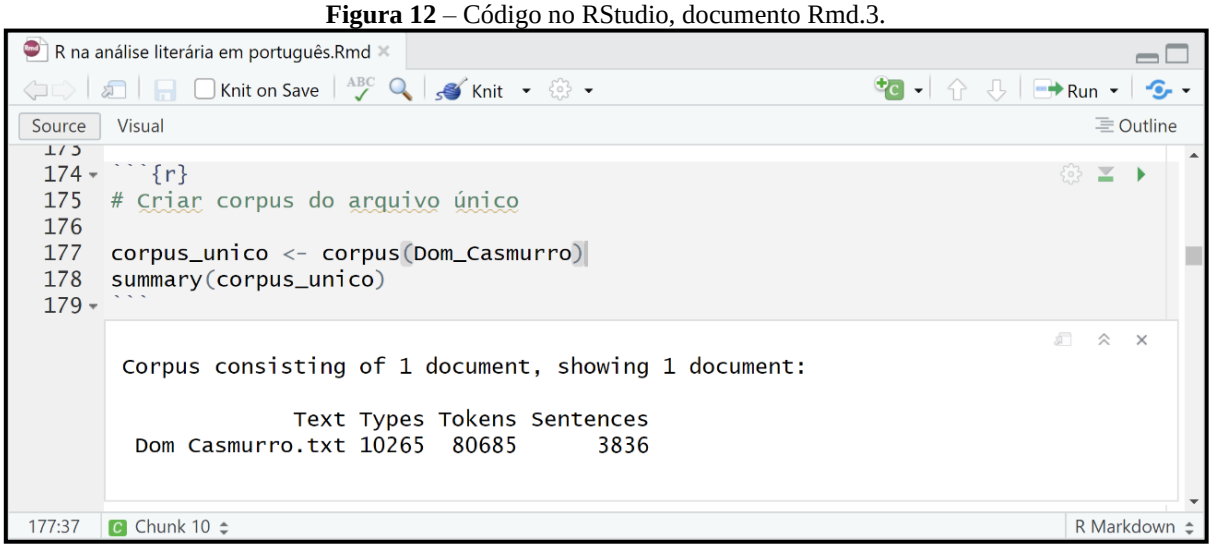
O código em R acima lê um arquivo de texto que contém o livro *Dom Casmurro* com codificação UTF-8 e depois exibe a estrutura do objeto resultante. Com o código, indica-se ao programa que deve ler o ficheiro txt. "Dom Casmurro" que está na pasta "corpora" dentro de "documentos" (imagem "Criação de corpus 1"). Outra forma de ler o corpus é indicando o endereço do texto em uma página on-line. Neste caso, sim é preciso ter conexão à rede:

Figura 11 – Código no RStudio, documento Rmd.2.



Fonte: RStudio (Diego Giménez).

Figura 12 – Código no RStudio, documento Rmd.3.



Fonte: RStudio (Diego Giménez).

O código cria um corpus a partir do objeto que contém o texto do livro "Dom Casmurro" e exibe um resumo estatístico desse corpus, fornecendo uma visão geral das características do texto contido nele. O objeto "corpus" é o formato necessário para que o Quanteda possa processar e gerar informações sobre o(s) texto(s). Para isso, basta aplicar a função "corpus". Automaticamente, o texto é segmentado em *tokens* e frases. *Tokens* correspondem a todas as ocorrências (incluindo repetições) de palavras, bem como a outros itens como pontuação, números e símbolos. Ao investigar o corpus com a função *summary*, obtém-se a contagem de frases, *tokens* e *types* (o número de *tokens* distintos em um corpus). Em seguida, deve-se *tokenizar* o corpus:

Figura 13 – Código no RStudio, documento Rmd.4.

```
```{r}
# tokenizar nosso corpora
toks_unico <- tokens(corpus_unico)

# remover pontuação (corpus limpo com regex)
toks_nopunct_unico <- tokens(corpus_unico, remove_punct = TRUE)

```
```

Fonte: RStudio (Diego Giménez).

O código *tokeniza* o corpus "corpus_unico", ou seja, divide o texto em unidades menores chamadas *tokens* e armazena o resultado em "toks_unico". *Tokeniza* novamente o corpus "corpus_unico", mas desta vez, removendo todos os caracteres de pontuação, e armazena o resultado em "toks_nopunct_unico". Essas operações são comuns em análises de

texto, onde a *tokenização* facilita a manipulação e análise do texto, e a remoção de pontuação pode ser útil para certas análises onde a pontuação não é relevante. E ainda se pode retirar palavras comuns ou irrelevantes para a análise:

Figura 14 – Código no RStudio, documento Rmd.5.

```
```{r}
# eliminar stopwords do corpus feito com um único arquivo
toks_nostop <- tokens_select(toks_unico, pattern = c("é", "porqu", "ha", "ond",
"tudo", "toda", "porque", "onde", "todo", "tão", "ter", "grand", "elle", "sobr",
stopwords("pt")), selection = "remove")

# eliminar tokens específicios do corpus feito com um arquivo, após eliminação
das pontuações
toks_selected_unico <- tokens_select(toks_nopunct_unico, pattern = c("é",
"porqu", "ha", "ond", "tudo", "toda", "porque", "onde", "todo", "tão", "ter",
"grand", "elle", "sobr", stopwords("pt")), selection = "remove")
```
```

Fonte: RStudio (Diego Giménez).

Desta forma, removeram-se as *stopwords* e algumas palavras específicas do conjunto de *tokens* do corpus original, resultando em "toks_nostop". Removeu-se também as mesmas *stopwords* e palavras específicas do conjunto de *tokens* do corpus que teve a pontuação removida, resultando em "toks_selected_unico". Esses procedimentos também são comuns em análises de texto para limpar os dados e focar nas palavras que são mais relevantes para a análise subsequente. Posteriormente, é preciso criar uma matriz de documentos e termos (DFM):

Figura 15 – Código no RStudio, documento Rmd.6.

```
```{r}
# aqui podemos ver as 20 palavras mais frequentes quando removemos números,
símbolos e pontuação

dfm_selected_unico <- dfm(toks_selected_unico)
print("remoção de tokens selecionados no corpus previamente limpo com regex e sem
stopwords")
topfeatures(dfm_selected_unico, 20)

```
```

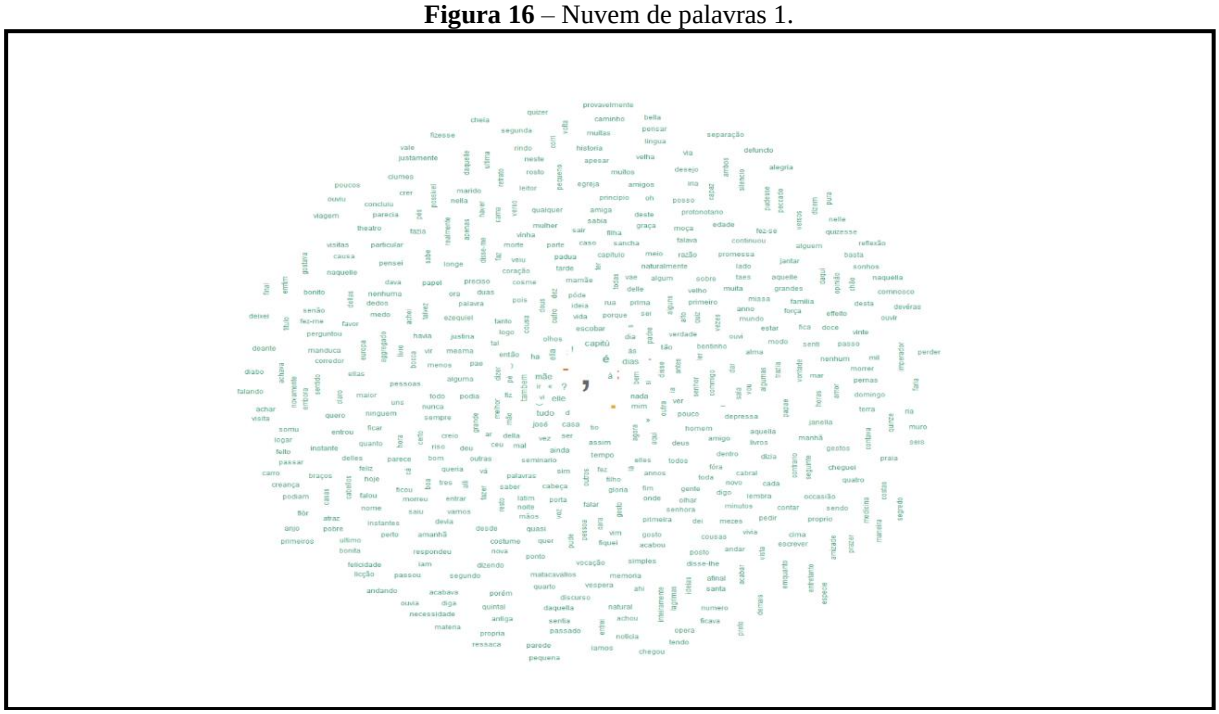
Fonte: RStudio (Diego Giménez).

Assim, mediante o código acima descrito, cria-se uma matriz de documentos e termos (DFM) a partir dos *tokens* do corpus que foram limpos (sem pontuação e sem *stopwords*). A análise exibe as 20 palavras mais frequentes na DFM resultante. Esses passos são habituais em análises de texto para identificar as palavras mais frequentes em um corpus após a limpeza dos

dados, o que é capaz de fornecer *insights* sobre o conteúdo predominante do texto, e são matrizes que podem ser utilizadas em outras análises.

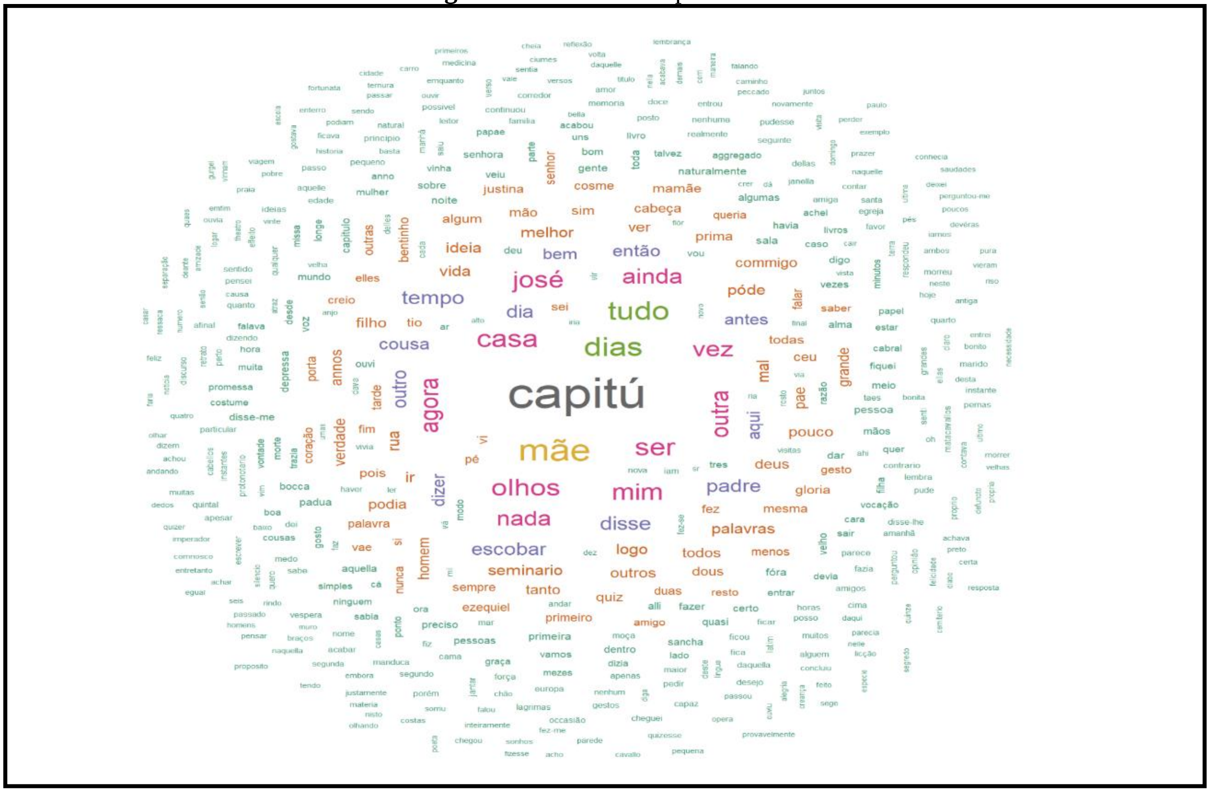
Essa é a descrição dos primeiros passos para poder utilizar o Quanteda para a mineração textual. De forma resumida, pediu-se ao programa que lesse o documento "Dom Casmurro", guardado na pasta "corpora"; posteriormente, indicou-se ao programa para criar um "corpus" a partir da leitura em R do documento; a seguir se *tokenizou* o corpus; na sequência limpou-se o corpus tirando as palavras irrelevantes; e, finalmente, criou-se uma matriz com os termos mais frequentes. A partir desses primeiros passos, pode-se efetuar outros tipos de análise. Neste artigo, também se mostrará a visualização das nuvens de palavras que, dependendo do processo de criação e de limpeza do corpus, também sofre alterações, fato que demonstra a tese deste artigo de que as representações e os resultados das análises são construções:

Figura 16 – Nuvem de palavras 1.



Fonte: RStudio (Diego Giménez).

Figura 17 – Nuvem de palavras 2.



Fonte: RStudio (Diego Giménez).

A primeira imagem foi criada a partir do corpus criado com o ficheiro “Dom Casmurro”, que foi *tokenizado*, e do qual não se retirou as palavras indesejadas, a pontuação e os símbolos. A segunda imagem corresponde ao mesmo corpus, mas com remoção de palavras indesejadas, números, símbolos e sinais de pontuação. Da mesma maneira que aconteceu com o *Voyant Tools*, o Quanteda demonstra como a preparação dos dados e a manipulação deles afetam diretamente a representação.

Resumo

De forma semelhante ao Voyant Tools, o uso do Quanteda apresenta uma metodologia determinada. No caso da obra de *Dom Casmurro*, a análise acompanhou a metodologia a seguir: a) instalação dos programas e pacotes necessários; b) criação do corpus numa pasta nos documentos; c) leitura do corpus pelo RStudio, nos arquivos guardados no computador (*off-line*) ou mediante um endereço a um texto on-line; d) criação e *tokenização* do corpus com o R.

O pacote Quanteda é uma ferramenta potente para a análise textual e a mineração de dados que oferece maior maneabilidade, ao permitir a codificação. O utilizador não apenas programa, senão também tem acesso ao processamento dos dados e um maior controle da modelagem e do destino do corpus. Outro benefício é que o programa pode ser utilizado sem ligação à internet. Entre os pontos menos positivos, pode-se destacar que se trata de uma ferramenta que requer conhecimentos básicos de programação. Quando os pacotes ou a plataforma são atualizados, os códigos utilizados podem ficar obsoletos. É necessária uma verificação do código, caso se pretenda certa estabilidade ou reprodutibilidade. Esse fato deve ser contemplado na hora de redigir um projeto de investigação que use essas ferramentas.

Conclusões

Ambas as ferramentas permitem análises completas de corpus textuais de diferentes tamanhos e composições. A mineração e o processamento de dados textuais, assim como a existência de um corpus previamente selecionado pelo pesquisador, são comuns às duas ferramentas no que diz respeito à metodologia. O utilizador de Voyant Tools não precisa ter conhecimentos de programação, dado que a plataforma permite o processamento online nos servidores do programa. O Quanteda, por outro lado, requer a instalação dos programas em um computador e permite o acesso off-line. O processamento dos dados, assim, é realizado no computador do utilizador. São necessários conhecimentos de programação.

Em segundo lugar, no que concerne à usabilidade, o Voyant Tools é mais fácil de utilizar ao apresentar uma interface mais intuitiva, na qual basta apenas carregar um documento ou indicar o endereço on-line em que se encontra o texto que se pretende analisar. Enquanto isso, o Quanteda requer o uso de código para a utilização, o carregamento e a leitura do corpus.

Finalmente, sobre a capacidade de análise e representação, o Voyant Tools contém várias ferramentas de visualização e análise. Como o processamento dos dados é realizado nos servidores do programa, o utilizador tem menos controle sobre a modelagem dos dados. O Quanteda, por sua vez, oferece um maior controle sobre a modelagem, embora seja preciso programar. Que os processamentos dos dados sejam feitos nos servidores do programa ou no computador do utilizador é um fato que deve ser contemplado no que atinge à privacidade e direitos sobre os dados e o corpus que estão sendo analisados.

Seja qual for o programa escolhido para a pesquisa, acredita-se que a especificação do processo de preparação do corpus e da modelagem dos dados seja imprescindível para garantir

a transparência e a reprodutibilidade das análises com dados textuais. Tanto a preparação do corpus quanto a aplicação de algoritmos de análise têm impacto direto nos resultados e nas suas representações gráficas. As representações produzidas por essas ferramentas são o resultado de decisões analíticas e interpretativas conscientes, desde a seleção do corpus até a definição de parâmetros para exclusão de palavras (*stopwords*), escolha de algoritmos e formas de visualização. Os dados não são neutros e as ferramentas digitais supõem interpretação. A dimensão interpretativa inerente à investigação com ferramentas digitais que acompanha os processos de preparação, modelagem e análise, devem ser explicitados pelos investigadores para garantir a transparência.

No artigo em questão, tanto o Voyant Tools quanto o Quanteda oferecem resultados diferentes nas análises, dependendo da manipulação do corpus. Conhecer as ferramentas utilizadas nas análises, detalhar o modo de coleta e preparação do corpus e descrever o tipo de algoritmo empregado são passos fundamentais para que se possa compartilhar resultados e metodologias passíveis de reprodução, verificação e contestação. Nesse sentido, o programa Quanteda oferece maior flexibilidade ao permitir um acesso mais direto ao processo de análise e representação. Em função dos objetivos da pesquisa, do grau de familiaridade com programação, do nível de controle desejado sobre o processamento dos dados, e das questões éticas e metodológicas envolvidas, deve ser feita a escolha entre Voyant Tools e Quanteda, e não simplesmente com base em critérios técnicos. O presente texto, assim, visou refletir sobre os modos como a tecnologia transforma as práticas críticas, e contribuir para o desenvolvimento de uma cultura de uso consciente, crítico e interpretativo das humanidades digitais.

Referências

ALVES, D. As Humanidades Digitais como uma comunidade de práticas dentro do formalismo acadêmico: dos exemplos internacionais ao caso português. **Ler História**, 69, 2016. Disponível em: https://doi.org/10.4000/lerhistoria.2496. Acesso em: 21 set. 2024.

BENOIT ET AL. Quanteda: An R package for the quantitative analysis of textual data. **Journal of Open Source Software**, 3(30), p. 774, 2018. Disponível em: https://doi.org/10.21105/joss.00774. Acesso em: 21 set. 2024.

CABRAL, M.J. *et al*. **Lire de près, de loin: close vs distant reading.** Paris: Garnier, 2014.

GIMÉNEZ, D & GOMIDE, A. Pesquisa Literária com R: Análise Quantitativa de Dados Textuais, Quanteda tomando como exemplo o Livro do Desassossego. **Estudos do Século**

XX, 22, p. 135-153, 2022. Disponível em: https://doi.org/10.14195/1647-8622_22_7. Acesso em: 27 jul. 2025.

GIMÉNEZ, D. R na análise literária em português. **Rpubs**, 2024. Disponível em: https://hdl.handle.net/10316/116796. Acesso em: 27 jul. 2025.

GOODWIN, J. *et al*. **Reading Graphs, Maps, and Trees: Responses to Franco Moretti**. SC: Parlor Press, 2011.

MORETTI, F. Conjectures on World Literature, em **New Left review**, n. 1, p. 64-68, 2000.

MORETTI, F. **Graphs, Maps, Trees: Abstract Models for Literary History**. London: Verso Books, 2005.

SANTOS, D. et al. Leitura distante em português: resumo do Primeiro Encontro. **MATLIT: Materialidades da Literatura**, v. 8, n. 1, p. 279-298, 2020. Disponível em: https://doi.org/10.14195/2182-8830_8-1_16. Acesso em: 21 set. 2024.

UNDERWOOD, T. A Genealogy of Distant Reading. **DHQ: Digital Humanities Quarterly**, v. 11, n. 2, 2017. Disponível em: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html. Acesso em: 21 set. 2024.

UNDERWOOD, T. **Distant Horizons: Digital Evidence and Literary Change**. Chicago: The University of Chicago Press, 2019.

Recebido em: 22 de setembro de 2024
Aceito em: 2 de agosto de 2025