Ricerca e analisi del testo (”text analysis”)

Caroline Muller

RICERCA E ANALISI DEL TESTO («TEXT ANALYSIS»)

Caroline Muller

L'analisi di grandi volumi di dati non è nuova: storici e storiche hanno approfittato, sin dalla fine degli anni Cinquanta, delle possibilità offerte dai computer. Dopo un periodo di crisi dovuto alle critiche metodologiche rivolte alle statistiche seriali, l'ingresso nell'era digitale ha rivitalizzato queste modalità d'analisi, ad esempio riguardo all'analisi del testo ("text mining"). L'inserimento in rete di grandi volumi di testi (atti parlamentari, giornali, corrispondenza privata, politica o pontificia…) consente di porre in maniera differente alcune questioni storiche. La lettura «distante» (Moretti) aiuta a visualizzare i processi, le reti, a evidenziare le informazioni altrimenti invisibili senza il supporto dell'algoritmo. Il progetto Numapresse di Julien Schuh studia, a partire da un database di migliaia di giornali digitalizzati, ad esempio, lo spazio che occupano i rimandi, gli echi e i riferimenti nella stampa francese, costruendo così una storia della viralità mediatica. La lettura remota e l'analisi dei testi contribuiscono alla nascita di nuovi approcci ai processi storici, rendendo possibile formulare ipotesi o rileggere, sfumare e completare quelli nati dalla lettura «ravvicinata» dei documenti.

Perché un tale lavoro sia possibile, è necessario tenere conto del "trattamento" dei documenti, cioè del modo in cui passiamo dalla struttura iniziale dell'informazione (il giornale, le sue rubriche, poi il suo repertorio fotografico) e un formato di testo interrogabile da un robot. Per fare questo, i documenti devono essere "ocerizzati"[1], cioè trasformati in testo da un software di riconoscimento automatico. Questo è uno dei grandi progetti legati alla ricerca testuale. Affinché il corpus testuale possa essere "scavato" o incrociato con altri, deve esse-

[1] Il neologismo fa riferimento al trattamento attraverso un software di riconoscimento ottico dei caratteri (in inglese Optical Character Recognition, OCR) [NdT].

re stato oggetto di diversi procedimenti: l'"ocerizzazione" e la codifica, il tutto in una logica di interoperabilità, così da poter essere impiegato dalla maggior parte dei sistemi e strumenti. Anche il corpus testuale deve essere rintracciabile, il che implica lo sviluppo di adeguati strumenti di ricerca. Fermo restando il rispetto di questi differenti vincoli, compete poi ai ricercatori cogliere queste opportunità, questione che mette alla luce le nuove esigenze nel campo della formazione. Progetti su larga scala come la Venice Time Machine (EPFL) prevedono la digitalizzazione, il trattamento di dati, la ricerca e la visualizzazione di migliaia di documenti testuali contenuti negli archivi di Venezia, che alla fine consentiranno la realizzazione di una «macchina per tornare indietro nel tempo».

L'analisi testuale non riguarda solamente i corpora digitalizzati, ma include anche nuove fonti testuali per scrivere la storia, come lo status sui social network, i tweet e gli archivi di pagine web. Questi documenti dischiudono molte piste di ricerca; si può menzionare lo studio di come i social network digitali influenzino le celebrazioni (Clavert 2018), la storia delle telecomunicazioni e di Internet (Schafer 2018), la storia della costruzione di comunità "virtuali" (Milligan 2017). L'ingresso nell'era dei *big data* corrisponde quindi all'inizio dell'era dell'utilizzo da parte degli storici delle fonti native digitali interrogabili per la ricerca testuale e la loro successiva visualizzazione.

Bibliografia

F. Clavert, *Face au passé: la Grande Guerre sur Twitter*, «Le Temps des médias» 31 (2018), 2, pp. 173-186, URL: < https://www.cairn.info/revue-le-temps-des-medias-2018-2-page-173.htm > [consultato il 27 maggio 2019].

S. Graham, I. Milligan, S. Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London, World Scientific Publishing Company, 2015.

C. Lemercier, C. Zalc, *Quantitative Methods in the Humanities: An Introduction*, Charlottesville-London, University of Virginia Press, 2019.

I. Milligan, *Welcome to the web: The online community of GeoCities during the early years of the World Wide Web*, «UWSpace» (2017), URL: < http://hdl.handle.net/10012/11859 > [consultato il 28 maggio 2019].

F. Moretti, *Distant Reading*, London, Verso Books, 2013.

V. Schafer, *En construction: la fabrique française d'internet et du web dans les années 1990*, Bry-sur-Marne, INA, 2018.

Sitografia

Numapresse, URL : < https://numapresse.hypotheses.org/1 > [consultato il 27 maggio 2019].