DARIAH-DE

Die quantitative Analyse großer Datenbestände in den Geisteswissenschaften: eine Kommentierte Bibliographie (R 5.2.2)

Version 19.12.2014
Cluster 5
Verantwortlicher Partner Universität Würzburg

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:

Stefan Pernes, UWÜ
Steffen Pielström, UWÜ

Revisionsverlauf:

Datum	Autor	Kommentare
19.11.2014	Pernes / Pielström	Entwurf
19.12.2014	Pernes / Pielström	Ergänzungen aus dem Konsortium

Inhaltsverzeichnis

Die quantitative Analyse großer Datenbestände in den Geisteswissenschaften: eine Kommentierte Bibliographie

Stefan Pernes[1] & Steffen Pielström[2]

Julius-Maximilians-Universität Würzburg

Die folgende, im Rahmen der Aktivitäten des Cluster 5 (*Big Data in den Geisteswissenschaften*) entwickelte Bibliografie dient dazu, den aktuellen Forschungsstand zu Verfahren der quantitativen Textanalyse in den *Digital Humanities* festzuhalten und sichtbar zu machen. Die Zusammenstellung basiert auf der laufend aktualisierten Zotero[3] Bibliographie von DARIAH-DE, *Doing Digital Humanities*[4]. Diese Bibliographie, die 2012 von DARIAH-DE Mitarbeitern begonnen wurde und mittlerweile durch eine *Community* von zur Zeit 108 aktiven Wissenschaftlern beobachtet und ergänzt wird, bietet eine gute Übersicht über das Spektrum an Publikationen, die momentan in der wissenschaftlichen Praxis der *Digital Humanities* wahrgenommen werden. Die Inhalte sind entsprechend der *TaDiRAH*-Taxonomie (*Taxonomy of Digital Research Activities in the Humanities*[5]) nach Untersuchungsobjekten, Aktivitäten und technischen Verfahren angeordnet, was die Auswahl für ein bestimmtes Forschungsvorhaben relevanter Titel vereinfacht. Die vorliegende Zusammenstellung greift hieraus jene Publikationen auf, die für die quantitative Analyse großer Datenbestände relevant sind. Hierbei werden wir zum sowohl auf theoretische Reflexionen, als auch auf etablierte und neue Verfahren zur Datenverarbeitung, -anreicherung und -analyse eingehen. Im Zentrum des vorliegenden Reports steht dabei der Überblick über die Forschungsliteratur. Er bildet damit die Grundlage für den künftigen Report 5.2.3 der eine inhaltliche Beschreibung des Forschungsstandes und der kritischen Auseinandersetzung mit dem für und wieder bestimmter Verfahren gewidmet sein wird.

---

[1] stefan.pernes@uni-wuerzburg.de
[2] pielstroem@biozentrum.uni-wuerzburg.de
[3] https://www.zotero.org/
[4] https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography/
[5] https://github.com/dhtaxonomy/TaDiRAH

1. '*Big Data*' in den Geisteswissenschaften

Die große Mehrheit der digital verfügbaren geisteswissenschaftlichen Forschungdaten liegt in Form von Texten vor. Theoretische Reflexionen zur computergestützten Textanalyse setzen sich mit der Frage auseinander, wie ein technisch vermittelter Zugang zu sprachlicher Bedeutung und eine damit einhergehende, für die klassischen Textwissenschaften ungewohnte, Organisation von Arbeitsschritten und Material als Ergänzung zu, wenn nicht sogar als direkte Weiterentwicklung typisch qualitativer hermeneutischer Interpretationsverfahren eingesetzt werden kann. Eine Beobachtung, die als kleinster gemeinsamer Nenner der folgenden Ansätze gelten darf, ist das Konzept einer Suche nach möglichst invarianten Prinzipien und Mustern. Es handelt sich dabei um eine Suchbewegung, die unabhängig von historischen Epochen und Kulturräumen stattfindet und alle möglichen Arten von Texten und Objekten umfasst (Bod, 2013). Angewandt auf aktuelle Datenbestände, wie zum Beispiel auf das Google Books Korpus, bestehend aus 15 Millionen Volltexten, gelangt diese Suchbewegung aufgrund ihrer Größenordnung zu einer neuen Qualität und wird seither auch als *Culturomics* bezeichnet (Aiden & Michel, 2011). In einem enger abgesteckten Rahmen ermöglicht die neue Größenordnung der Datenbestände in den Geisteswissenschaften ein gezieltes, empirisches Nachprüfen theoretischer Erkenntnisse. So können anhand quantifizierbarer Befunde beispielsweise die Entwicklung literarischer Strömungen nachempfunden werden - hinsichtlich ihrer Verbreitung in bestimmten Regionen und demografischer Gruppen, aber auch hinsichtlich ihrer Formierung ausgehend von einzelnen Autoren und Texten (Jockers, 2013). Die computergestützte Modellierung wird hier zu einer experimentellen und iterativen Praxis, die es ermöglicht, bestehende Annahmen und Generalisierungen jenseits von unüberprüfbaren Eindrücken und Anekdoten neu zu beleuchten (Bode, 2012). Die Kernkompetenz des Computers liegt hierbei im Überblicken von Textmengen und in der Erkennung von Mustern, die sich der menschlichen Wahrnehmung entziehen, sie lässt sich unter Umständen jedoch auch ausweiten, über rigide Berechnungen hinaus, hin zu Formen der spielerischen Auseinandersetzung, zu einer Einbindung von Intuition und Subjektivität (Ramsay, 2011). Ausgehend von diesen paradigmatischen Veränderungen im Bereich der Textinterpretation, wurde bereits angeregt, die Philologie anstelle der Philosophie als Bezugs- und Metawissenschaft für die Geisteswissenschaften einzusetzen (McGann, 2014).

1.1 Landmark studies

Ein wichtiger Schritt, um dieses neue Forschungsfeld einer breiteren Öffentlichkeit zugänglich zu machen, wurde im Rahmen des Google Books Projektes geleistet (Aiden & Michel, 2011, 2014). Zu dem Ansatz, Schlüsselworte im diachronen Verlauf zu verfolgen und einander gegenüberzustellen, finden sich jedoch einige Vorläufer, darunter frühere Arbeiten aus der Korpuslinguistik (Brunet, 1989). Innerhalb der zeitgenössischen *Digital Humanities* hat sich das Verfahren des *Topic Modeling* - ein statistisches Modell zur Entdeckung abstrakter Themenkomplexe in Dokumentensammlungen - bewährt, was sich auch daran ausmachen lässt, dass das Verfahren in einigen umfangreicheren Arbeiten zum Einsatz kommt (Underwood & Goldstone, 2012; Heuser & Le-Khac, 2012; Schmidt, 2013; Evans, 2014). Dabei scheint der probabilistische Charakter der Ergebnisse zu einer Verknüpfung quantitativer und qualitativer Herangehensweisen beizutragen. Neben solchen weitgehend offenen Zugängen zum Material, werden weiterhin auch Korpora im Sinn einer kontrollierten Grundgesamtheit erstellt, um darin verankerte Aussagen treffen zu können - so zum Beispiel ein Korpus bestehend aus mehreren Millionen von Metaphern (Pasanek & Sculley, 2008), oder ein Korpus, das sämtliche Passagen einer vom Erzähler unterbrochenen Figurenrede in den Werken von Charles Dickens versammelt (Mahlberg & Smith, 2012). Darüber hinaus sind auch experimentelle Arbeiten zu nennen, wie zum Beispiel *Mandala Browser*, eine grafische Oberfläche, die eine spielerisch-intuitive Exploration von Korpora ermöglicht (Brown *et al.*, 2011)

1.2 Probleme und Grenzen quantitativer Verfahren

Gerade im Kontext der Geisteswissenschaften ist es naheliegend, die rigide Strukturierung, die von computergestützten Verfahren benötigt wird, kritisch zu hinterfragen. Gefordert werden kann beispielsweise, dass zukünftige Werkzeuge der Textanalyse ein holistischeres Bild vom menschlichen Lesevorgang unterstützen und dahingehend Anknüpfungspunkte bereitstellen sollen (Sinclair, 2003). Kritisch beleuchtet wurden auch Fragen nach der statistischen Grundgesamtheit, *Recall*-Problemen und einer Tendenz, das zu analysieren, was leicht zu finden ist (Ball, 1994). Dementsprechend wurde für die Anwendung statistischer Methoden in den Geisteswissenschaften - nicht zuletzt wegen des inhärenten subjektiven Gehalts der Fragestellungen - eine rigorose Qualitätskontrolle gefordert (Sculley & Pasanek,

2008). Im Hinblick auf die Wartung von Datenbeständen wurde auch von den Auswirkungen und Lösungsansätzen zu einer Inkonsistenz in föderierten Datenbanken berichtet (Upchurch, 2012). Darüber hinaus wird neben den technischen Fragen zu *'Big Data'* auch die Notwendigkeit gesehen, die sozioökonomischen und politischen Konstellationen, anhand derer solche großen Datenbestände erzeugt werden, zu beleuchten (Vis, 2013).

2. Textklassifikation

Ein bereits weit etabliertes Betätigungsfeld der *Digital Humanities*, bei dem Datenbestände quantitativ analysiert werden, ist die Klassifikation von Texten, insbesondere die Autorenschaftsattribution. Zu dem Problem der Textklassifikation sind zahlreiche Publikationen veröffentlicht worden, in denen methodische Verfahren anhand von Fallbeispielen vorgestellt und beschrieben werden. Diese Studien basieren überwiegend auf für Autoren charakteristischen Häufigkeiten bestimmter Wörter. Um diese zu erfassen, werden aus den zu untersuchenden Texten Wortfrequenzlisten generiert, die sich z.B. mittels Diskriminanzanalyse z.B. (Mosteller & Wallace, 1963; Craig, 1999) oder *Principal Component Analysis* vergleichen lassen (z.B. Burrows, 1989; 2005; Binongo, 2003). Ein anderer, weit verbreiteter Ansatz besteht darin, ein Distanzmaß für zwei solche Wortfrequenzlisten zu berechnen um die Unterschiedlichkeit der Texte zu quantifizieren und sie darauf basierend mit einem *Clustering* Algorithmus zu gruppieren (Dunning, 1993; Hoover, 2004, 2008; Labbé & Labbé, 2006; Garcia & Martin, 2007; Popescu & Dinu, 2009; Eder, 2010; Rybicki & Eder, 2011; Black, 2012; Kestemeont *et al.*, 2012; Marsden *et al.*, 2013). Ähnliche, auf einfachen *Features*, wie Wortfrequenzen, basierende Verfahren finden im Übrigen nicht nur bei der Identifikation Literarischer Stile Anwendung, es gibt auch Versuche, sie z.B. auf den Stil von Malern anzuwenden (Shamir, 2012). Verfahren aus dem Bereich des *Text Mining* und des Maschinellen Lernens - d.h. der Anwendung statistischer Verfahren zur automatischen Verarbeitung und Analyse großer Textmengen - finden nicht nur in der Autorenschaftattribution (Koppel *et al.*, 2008; Jockers & Witten, 2010; Eder & Rybicki, 2013) Anwendung, sondern werden auch jenseits dieses Problems für andere Textklassifikationsaufgaben eingesetzt (Sebastiani & Ricerche, 2002; Horton *et al.*, 2009). Solche Verfahren ermöglichen z.B. die Unterscheidung von übersetzten und nicht übersetzten Texten (Baroni & Bernardini, 2006), die Klassifikation nach Gender, Alter, ethnischem Hintergrund oder politischer Gesinnung des Autors (Argamon *et al.*, 2009; Bamman *et al.*, 2012; Dahlhof, 2012), oder auch die Abschätzung des Zeitraumes, in dem ein Text verfasst

wurde (Tilahun *et al.*, 2013). Ein weiteres Anwendungsgebiet ist die Klassifikation von Abschnitten innerhalb eines Textes, z.B. zur Identifikation direkter Rede oder der Gedanken von Romanfiguren (Brunner, 2013). Eine Übersicht zur Klassifikation literarischer Texte aus korpuslinguistischer Sicht liefert Biber (2011).

3. Verarbeitung, Annotation, *Enrichment*

Den meisten der zuvor beschriebenen Textklassifikationsprobleme ist gemein, daß sie sich durch die Analyse der Häufigkeit von Wörtern, also einfachen, bereits in den Rohdaten verfügbaren *Features*, lösen lassen. Andere Forschungsfragen können aber durchaus auf Vorverarbeitungsschritte angewiesen sein, insbesondere auf Annotationen, die die Rohdaten mit zusätzlichen oder aus den Rohdaten abgeleiteten Informationen versehen, und im Extremfall sogar der alleinige Gegenstand der darauf folgenden Analyseschritte sein können. Die geisteswissenschaftliche Forschung hat zahlreiche Konventionen zur digitalen Annotation für unterschiedlichste Quellengattungen etabliert (siehe z.B. Walsh, 2012), die Informationen selbst werden jedoch meist von Hand in den Datensatz eingefügt. Bei Forschungsvorhaben deren Datenbasis aber sehr groß und wenig, oder gar nicht annotiert bzw. vorverarbeitet ist, wird die Automatisierung dieser Arbeitsschritte unerlässlich. Automatisierte Annotationsverfahren können hierbei sowohl genutzt werden, sowohl um die ganzen Objekte (Werke, Texte oder Einträge) einer Sammlung mit Metadaten anzureichern und zu organisieren (Herkt, 1991; Nicolas *et al.*, 2010), als auch um die Daten innerhalb dieser Objekte, wie z.B. einzelne Worte oder Textpassagen, für weitere Analyseschritte zu annotieren (Bobenhausen & Gehl, 2009; Unsworth, 2011). Solche Verfahren können Informationen auch aus externen Quellen beziehen und in die Forschungdaten einarbeiten (Zhang & Iria, 2009).

4. *Feature Extraction*

Neben unbearbeiteten Rohdaten und annotierten Metadaten können auch von diesen abgeleitete oder aus diesen erzeugte Parameter einen Teil, oder auch den gesamten Gegenstand der angestrebten Analyse darstellen. Im einfachsten Fall werden *Features*, die in die Analyse eingehen sollen, nach bestimmten Regeln vorselektiert (Yang & Pedersen, 1997; Forman, 2003; Guyon & Elisseeff, 2003; Li *et al.*, 2009; Eder, 2010; Rybicki & Eder, 2011) oder gewichtet (Hoover, 2004; Waltman *et al.*, 2010). Andere Verfahren erzeugen auf der Basis automatisierter analytischer Verfahren aus den vorhandenen Informationen zusätzliche,

abgeleitete *Features*. Beispiele hierfür sind *Topic Modelling* (Blei, 2012), *Opinion Mining* und *Sentiment Analysis* (Pang & Lee, 2008).

5. Lehrbücher

Eine mögliche Auswahl an Lehrbüchern für das vorliegende Forschungsfeld zu erstellen ist nicht trivial und kann keinen Anspruch auf Vollständigkeit erheben. Das ist einerseits begründet in der Interdisziplinarität des Themas, andererseits in der Diversifizierung möglicher Quellen - immer öfter werden didaktisch hochwertig aufbereitete Anweisungen in *Blogs* der Öffentlichkeit zur Verfügung gestellt. Die in den Lehrbüchern behandelten Themen reichen von Einführungen zu Register, Genre und Stil (Biber & Conrad 2009), zu Stilometrie (Holmes & Kardos 2003), Worthäufigkeitsverteilungen (Baayen 2001) und Statistik mit *R* (Baayen 2008), bis hin zum *Mining* großer Datenbestände, das vor allem anhand von *Machine Learning* bewerkstelligt wird (Feinerer 2008; Karatzoglou & Feinerer 2010; Rajaraman & Ullman 2011; Han 2011; Murphy 2012; Miner 2012). Neben dem klassischen Lehrbuch-Format ist (stellvertretend für eine Vielzahl solcher *Tutorials*) auf zwei weitere Einführungen zu *Text Mining* im Allgemeinen (Underwood 2012) und *Topic Modeling* im Speziellen (Riddell 2013) hinzuweisen.

Bibliographie

Aiden, E., & Michel, J.-B. (2013). *Uncharted: big data as a lens on human culture*. New York: Riverhead Books.

Argamon, S., Cooney, C., Horton, R., Olsen, M., Stein, S., & Voyer, R. (2009). Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters, *3*(2).

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht; Boston: Kluwer Academic.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R* (1. publ.). Cambridge: Cambridge University Press.

Ball, C. N. (1994). Automated Text Analysis: Cautionary Tales. *Literary and Linguistic Computing*, *9*(4), 295–302. doi:10.1093/llc/9.4.295

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2012). Gender identity and lexical variation in social media. *Journal of Sociolinguistic*. doi:10.1111/josl.12080

Baroni, M., & Bernardini, S. (2006). A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, *21*(3), 259–274. doi:10.1093/llc/fqi039

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature*, *1*(1), 15–23. doi:10.1075/ssol.1.1.02bib

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK; New York: Cambridge University Press.

Binongo, J. N. G. (2003). Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. *Chance*, *16*(2), 9–17.

Black, C. (2012, March 14). Clustering with Compression for the Historian. Retrieved November 25, 2012, from http://journalofdigitalhumanities.org/1-1/clustering-with-compression-for-the-historian-by-chad-black/

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. doi:10.1145/2133806.2133826

Bobenhausen, K., & Gehl, G. (2009). Automatisches metrisches Markup deutschsprachiger Gedichte. *Jahrbuch für Computerphilologie*, *7*. Retrieved from http://computerphilologie.tu-darmstadt.de/jg07/bobgehl.html

Bod, R. (2013). *A new history of the humanities: the search for principles and patterns from Antiquity to the present*. Oxford: Oxford Univ. Press.

Bode, K. (2012). *Reading by numbers: recalibrating the literary field*. New York: Anthem Press.

Brown, S. I., Ruecker, S., Antoniuk, J., Farnel, S., Gooding, M., Sinclair, S., … Gabriele, S. (2011). Reading Orlando with the Mandala Browser: A Case Study in Algorithmic Criticism via Experimental Visualization. *Digital Studies / Le Champ Numérique*, *2*(1).

Brunet, E. (1989). L'exploitation des grands corpus: Le bestiaire de la littérature française. *Literary and Linguistic Computing*, *4*(2), 121–134. doi:10.1093/llc/4.2.121

Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing*. doi:10.1093/llc/fqt024

Burrows, J. F. (1989). "An ocean where each kind. . .": Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, *23*(4-5), 309–321. doi:10.1007/BF02176636

Burrows, J. F. (2005). Who Wrote Shamela? Verifying the Authorship of a Parodic Text. *Literary and Linguistic Computing*, (20), 437–450.

Craig, H. (1999). Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, *14*(1), 103 –113. doi:10.1093/llc/14.1.103

Dahllof, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches--A comparative study of classifiability. *Literary and Linguistic Computing*, *27*(2), 139–153. doi:10.1093/llc/fqs010

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74.

Eder, M. (2010). Does Size Matter? Authorship Attribution, Small Samples, Big Problem. In *Digital Humanities 2010: Conference Abstracts* (pp. 132–134). London: King's College London. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-744.html

Eder, M., & Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, *28*(2), 229–236. doi:10.1093/llc/fqs036

Evans, M. S. (2014). A Computational Approach to Qualitative Analysis in Large Textual Datasets. *PLoS ONE*, *9*(2). doi:10.1371/journal.pone.0087908

Feinerer, I. (2008). An Introduction to Text Mining in R. *Rnews*, *8*(2), 19–22.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, *3*, 1289–1305.

Garcia, A. M., & Martin, J. C. (2007). Function Words in Authorship Attribution Studies. *Literary and Linguistic Computing*, *22*(1), 49–66. doi:10.1093/llc/fql048

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

Han, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Burlington, MA: Elsevier.

Herkt, M. (1991). *Anwendungsmöglichkeiten computergestützter Erfassungs- und Auswertungshilfen am Beispiel der Güter- und Einkünfteverzeichnisse des Kollegiatstiftes St. Mauritz in Münster*. Bochum: Brockmeyer.

Heuser, R., & Le-Khac, L. (2012). *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*. Standford CA: Literary Lab, Stanford University.

Holmes, D. I., & Kardos, J. (2003). Who Was the Author? An Introduction to Stylometry. *Chance*, *16*(2), 5–8.

Hoover, D. L. (2004). Delta Prime? *Literary and Linguistic Computing*, *19*(4), 477 –495. doi:10.1093/llc/19.4.477

Hoover, D. L. (2008). Searching for Style in Modern American Poetry. In S. Zyngier (Ed.), *Directions in Empirical Literary Studies: Essays in Honor of Willie van Peer* (pp. 211–227). Amsterdam: John Benjamins. Retrieved from 10.1075/lal.5.18hoo

Horton, R., Morrissey, R., Olsen, M., Roe, G., & Voyer, R. (2009). Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie. *Digital Humanities Quarterly*, *3*(2).

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*(2), 215–223. doi:10.1093/llc/fqq001

Karatzoglou, A., & Feinerer, I. (2010). Kernel-based machine learning for fast text mining in R. *Computational Statistics & Data Analysis*, *54*(2), 290–297. doi:16/j.csda.2009.09.023

Kestemont, M., Luyckx, K., Daelemans, W., & Crombez, T. (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*, *93*(3), 340–356. doi:10.1080/0013838X.2012.668793

Koppel, M., Schler, J., & Argamon, S. (2008). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26.

Labbé, C., & Labbé, D. (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*, *21*(3), 311 –326. doi:10.1093/llc/fqi063

Li, S., Xia, R., Zong, C., & Huang, C.-R. (2009). A framework of feature selection methods for text categorization. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2* (pp. 692–700). Stroudsburg, PA, USA: Association for Computational Linguistics.

Marsden, J., Budden, D., Craig, H., & Moscato, P. (2013). Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon. *PLoS ONE*, *8*(6), e66813. doi:10.1371/journal.pone.0066813

McGann, J. (2014). *A New Republic of Letters. Memory and Scholarship in the Age of Digital Reproduction*. Boston, MA: Harvard Univ. Press.

Mahlberg, M., & Smith, C. (2012). Dickens, the suspended quotation and the corpus. *Language and Literature*, *21*(1), 51–65. doi:10.1177/0963947011432058

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176 –182. doi:10.1126/science.1199644

Miner, G. (2012). Chapter 2 - The Seven Practice Areas of Text Analytics. In *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (pp. 29–41). Boston: Academic Press.

Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, *58*(302), 275–309. doi:10.2307/2283270

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press.

Nicholas, N., Ward, N., & Blinco, K. (2010). Abstract Modelling of Digital Identifiers. *Ariadne*, *62*.

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Retrieved from http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

Pasanek, B., & Sculley, D. (2008). Mining Millions of Metaphors. *Literary and Linguistic Computing*, *23*(3), 345–360. doi:10.1093/llc/fqn010

Popescu, M., & Dinu, L. P. (2009). Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis. In *Proceedings of the International Conference RANLP* (pp. 349–354).

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Ramsay, S. (2011). *Reading machines : toward an algorithmic criticism*. Urbana Ill.: University of Illinois Press.

Riddell, Allan (2013): Text Analysis with Topic Models for the Humanities and Social Sciences. Retrieved November 18, 2014, from http://de.dariah.eu/tatom

Rybicki, J., & Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, *26*(3), 315–321. doi:10.1093/llc/fqr031

Schmidt, B. M. (2013). Words Alone: Dismantling Topic Models in the Humanities. Retrieved April 15, 2013, from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Sculley, D., & Pasanek, B. M. (2008). Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities. *Literary and Linguistic Computing*, *23*(4), 409–424. doi:10.1093/llc/fqn019

Sebastiani, F., & Ricerche, C. N. D. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*, 1–47.

Shamir, L. (2012). Computer Analysis Reveals Similarities between the Artistic Styles of Van Gogh and Pollock. *Leonardo*, *45*(2), 149–154. doi:10.1162/LEON_a_00281

Sinclair, S. (2003). Computer-Assisted Reading: Reconceiving Text Analysis. *Literary and Linguistic Computing*, *18*(2), 175 –184. doi:10.1093/llc/18.2.175

Tilahun, G., Feuerverger, A., & Gervers, M. (2013). Dating medieval English charters. *arXiv:1301.2405*, *6*(4). doi:10.1214/12-AOAS566

Timothy Allen, C. C. (2010). Plundering Philosophers: Identifying Sources of the Encyclopédie. *Journal of the Association for History and Computing*. Retrieved from http://hdl.handle.net/2027/spo.3310410.0013.107

Underwood, T. (2012). Where to start with text mining. Retrieved from http://tedunderwood.wordpress.com/2012/08/14/where-to-start-with-text-mining/

Underwood, T., & Goldstone, A. (2012). What can topic models of PMLA teach us about the history of literary scholarship? Retrieved December 16, 2012, from http://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/

Unsworth, J. (2011). Computational Work with Very Large Text Collections. *Journal of the Text Encoding Initiative*, (Issue 1).

Upchurch, C. (2012). Full-Text Databases and Historical Research: Cautionary Results from a Ten-Year Study. *Journal of Social History*. doi:10.1093/jsh/shs035

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10). doi:10.5210/fm.v18i10.4878

Walsh, J. A. (2012). Comic Book Markup Language: An Introduction and Rationale, *6*(1).

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization (pp. 412–420). Morgan Kaufmann Publishers.

Zhang, Z., & Iria, J. (2009). A novel approach to automatic gazetteer generation using Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources* (pp. 1–9). Stroudsburg, PA, USA: Association for Computational Linguistics.