Still Tentative (esp. wrt schedule & student presentations!) Course Description

Hauptseminar Digital humanities from a computational linguistics perspective.

Instructor: Prof. Dr. John Nerbonne j.nerbonne@rug.nl Winter Semester 2024

Mon. & Wed. 8:15-9:45 normally via Zoom, Uni-Tübingen , soemtimes in room 081 at Keplerstr. 2!

Zoom mtg: https://zoom.us/j/96335665372?pwd=S0dYVTl5THZvaFY3S0FBdUFRalVxUT09
Meeting-ID 963 3566 5372
Kenncode (Password): 480210

Note: The course will be held mostly online, but 2-3 times in Tübingen. Hopefully incl. Nov. 6 or 27

Content of course. Digital humanities (DH) has increased enormously in popularity over the last 15 years, benefitting from huge increases in available data and computing power. Baptiste et al. (2011) and Moretti (2013) were widely noticed, but the work had been going on for a long time (Robinson 2003, Busa 1980). More and more digitally sophisticated work is being done in history and literature departments, joining their humanities' cousins such as linguistics, archeology and musicology, where the digital turn was effected earlier. An important focus is text analysis, where computational linguistics (CL) plays a natural role.

This seminar will focus on text analysis as practiced in DH, focusing on topics that CL is poised to contribute to, including stylometry, authorship attribution (Nerbonne 2007), and authorship profiling; sentiment analysis (Napier & Shamir 2018), incl. the detection of hate speech and the like; topic modeling (Jockers 2013) and work on the history of stories. Given time and interest we may examine other topics, e.g., the role of social networks in literary analysis, spelling variations in philology (see van Dalen-Oskam & van Zundert 2007),

The intention is to survey topics popular in DH, and to require only limited technical ability, making the course accessible to non-CL students. Students wishing to earn nine points will have to conduct a project involving programming, however.

Requirements. There are three levels of participation worth 3, 6 or 9 ECTS points. All students will be expected to *participate in the seminars*, both via discussion of lectures but also during the discussion of *obligatory exercises*. *A roughly 15-20-min. presentation(s) of a research paper* will also be expected. The choice of paper must be made by Nov. 10th, an outline of the talk two weeks before presentation (but by Dec 11 for those presenting in the first week of Jan.), and a rough draft of approx. 12-15 intended slides must also be turned in a week before presentation (and by Dec. 18th for those presenting in the first week of Jan.). These requirements hold for all participants, and will result in 3 ECTS points.

Signup for presentations: https://docs.google.com/document/d/1f3QO-C2lPy9xk8R9UVekykKPYXNhuN6P43kwkyYiMoY/edit?usp=sharing

Those wishing to earn 6 ECTS must complete all of the requirements for 3 points and must additionally write a paper, preferably on the same topic as the presentation, of 2K-3K words. I will say more on the desired format later in the course. Alternative topics may be proposed but must be approved.

Those wishing to earn 9 ECTS points must complete all of the requirements for 3 points and must additionally conduct a project in DH involving some implementation. The topic of the project must be approved ahead of time. The work must be reported on in a paper of 3,5-4K words.

Potential presentation topics are printed in red below, and note that each line is a presentation topic. Thus Dec. 20 lists two topics. Guidelines for presentations are in the general section of the course website (on Moodle) as well as an example presentation from another course. Most topics involve a single paper, but the two papers by Brett & Blei on topic modeling are an exception. Please notify the instructor at least two weeks ahead of time if you wish to present. As usual, those who claim a topic first have the first rights to it. Sign up via the Google Doc or email your interest!

If you have a text-related topic in DH in mind that you'd would like to present, please feel free to propose it to me. I promise to consider it critically. Last year, the detection of abusive language was a popular topic – for praiseworthy reasons, so I've included some more papers on detecting cyberbullying and detecting fake news, too. But it's also quite difficult (Waseem & Hovy 2016, Chandrasekran et al. 2017, Nguyen et al. 2020), so please consider this before taking it on.

Schedule

	Week of …	Monday	Wed.
1	Oct. 16 Intro	Dept. orientation mtg. No class.	Intro., orientation Ted Talk, Michel & Lieberman (homework)
2	Oct. 23	Michel et al. (2011) Google n-grams exercises (Ex.1)	Student interests.
3	Oct. 30	Annotated Text (TEI)	All Saint's Day, no class
4	Nov. 6	Discussion Ex. 1 Culturomics	Discussion paper presentation, choices
	11:00	Parlamint Exercise (Ex. 2)	
5	Nov 13	Stylometry & Text Classification Nerbonne (2007)	Intro *Stylo* Eder et al. (2016)
6	Nov. 20	Intro *Stylo*, Part II *Stylo* Exercise, *The Federalist* (Ex.3)	Clustering & MDS – groups!
7	Nov. 27	Bayes & Naïve Bayes	Discussion Ex. 2 Parlamint
8	Dec. 4	Burrows Delta	Deltas, Evert et al. (2017)
9	Dec. 11	Ex. 3 Discuss, Stud. presentations	Eder '15 "Rolling Stylomtry" Van Dalen/V Zundert '07 Walewein
	11 am		
10	Dec. 18	"Imposters" Koppel/Winter 2014 Wilhelmus, Kestemont et al. 2017	Typical wds., Klaußner et al '15 Rhyme-Features, Kestemont '12
	Dec 25/Jan 1	Holidays	Holidays
11	Jan. 8.	Author profiling Estival et al. 2007 Plank & Hovy 2015	Hate speech, Waseem & Hovy '16, Cyberbullying, Van Hee et al., '18
12	Jan 15	Sentiment Analysis Lit, Taboada et al '06	Emotions in Parliament, Rheault et al. Shakespeare Chars, Nalisnick & Baird '13,
13	Jan 22	Topic Modeling, Jockers '12, Ex. 4 Brett / Blei 2012	Goldstone & Underwood '12 & '14 History ideas, Mimno '12
14	Jan 29	Jurafsky '14 Games & Gender, Erdur	Search older texts, Jurish et al. '14 Restoring old texts, Assael et al. '22,
15	Feb. 5	Information Theory in DH	Rao et al. 2009, Liberman 2009 Need to engage!

Options for reserve:

1. Lg. history? Bizzoni et al. '20 / Semantics Tamahsebi et al.?
2. Forgotten books? https://forgotten-books.netlify.app/index.html
3. Music lyrics, rhyme, scanning (application): Elena Gonzalez Blanco, Madrid

Literature

'*' next to an entry in the list below means either that I haven't read it yet (it's too new), or that it's more difficult than most. Caution is advised.
Note, too, that I have purged the papers on dialectology from an older version of this list, because I may give a course on dialectometry in Tübingen next semester. Those topics won't be covered here.

General introductions

Students often want to refer to a general textbook to orient themselves in a filed that's being introduced. Here are some candidates

*Drucker, J. (2021). *The digital humanities coursebook: an introduction to digital methods for research and scholarship*. Routledge.

Terras, M., Nyhan, J., & Vanhoutte, E. (Eds.). (2016). *Defining digital humanities: a reader*. Routledge.

Note (Nov. 9) Adding Argamon et al. (2003) for profiling, highlighting Chandrasekran et al. 2017, Verhoeven et al. 2016 for profiling, Jebaseelie & Kiukbakraran (2012) for sentiment analysis for marketing, Desai, Mitali, & Mayuri A. Mehta (2016) on methods for sentiment analysis, Nguyen et al. 2020 based on expressions of interest)

Specific papers

Argamon S., M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

*Assael, Yannis, et al. (2022) "Restoring and attributing ancient texts using deep neural networks." *Nature* 603.7900: 280-283. *https://www.nature.com/articles/s41586-022-04448-z.pdf*

Barbrook, Adrian C., et al. (1998) "The phylogeny of the Canterbury Tales." *Nature* 394.6696: 839-839.

Bizzoni, Yuri, et al. (2020) "Linguistic variation and change in 250 years of English scientific writing: a data-driven approach." *Frontiers in Artificial Intelligence*. Art. 73.

Blei, David M. (2012) "Topic modeling and digital humanities." *Journal of Digital Humanities* 2.1: 8-11.

Brett, Megan R. (2012) "Topic modeling: A basic introduction." *Journal of digital humanities* 2.1: 2-1.

Burrows, John (2017) Interview. https://www.youtube.com/watch?v=0QpJFAjdKz8

Busa, Roberto (1980) "The annals of humanities computing: The Index Thomisticus." *Computers and the Humanities.* 83-90.

Chandrasekharan, Eshwar, et al. (2017) "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech." *Proceedings of the ACM on Human-Computer Interaction* 1 CSCW. 1-22.

van Dalen-Oskam, Karina & Joris Van Zundert. (2007) "Delta for middle Dutch—author and copyist distinction in *Walewein*." *Literary and Linguistic Computing* 22.3: 345-362.

De Smedt, Tom, et al. (2018) "Multilingual cross-domain perspectives on online hate speech." *arXiv preprint arXiv:1809.03944*.

Eder, M. (2016). “Rolling stylometry" *Digital Scholarship in the Humanities* 31(3): 457-69. See the blog on running the stylo analysis: https://computationalstylistics.github.io/blog/rolling_stylometry/

Eder, M. et al.. (2016) "Stylometry with R: a package for computational text analysis." *The R Journal* 8.1.

Erdur, N. (2022). Gender in Genshin Impact: A Corpus-Assisted Discourse Analysis. *Language Education and Technology,* 2(1). https://www.researchgate.net/publication/371220355_Gender_in_Genshin_Impact_A_Corpus-Assisted_Discourse_Analysis.

Estill, Laura, and Luis Meneses (2018) "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare’s Plays." *Digital Studies/Le champ numérique* 8.1. https://www.digitalstudies.org/articles/10.16995/dscn.295/

Estival, Dominique, et al. (2007) "Author profiling for English emails." *Proc. 10th Conf. Pacific Association Computational Linguistics*. 263-272.

Evert, Stefan, et al. (2017) "Understanding and explaining Delta measures for authorship attribution." *Digital Scholarship in the Humanities* 32.suppl_2: ii4-ii16.

Feng, Song et al. (2012). Syntactic stylometry for deception detection. In *Proc. 50th Meeting Association for Computational Linguistics (Volume 2)*. 171-175.

Desai, Mitali, & Mayuri A. Mehta (2016). "Techniques for sentiment analysis of Twitter data: A comprehensive survey." *Int. Conf. on Computing, Communication & Automation* (ICCCA). IEEE.

Goldstone, Andrew and Ted Underwood (2012) "What can topic models of PMLA teach us about the history of literary scholarship." *Journal of Digital Humanities* 2.1: 39-48.

Goldstone, Andrew and Ted Underwood (2014) "The quiet transformations of literary studies: What thirteen thousand scholars could tell us." *New Literary History* 45.3 (2014): 359-384.

Gottscharek et al. (2011) “Toward information retrieval for historical document collections” *Int. J. Doc. Analysis & Recog.* 14, 159-171

Van Hee, Cynthia, et al. (2018) Automatic detection of cyberbullying in social media text. *PloS one* 13.10: e0203794.

Jebaseeli, A. Nisha, and E. Kirubakaran (2012) A survey on sentiment analysis of (product) reviews. *International Journal of Computer Applications* 47.11

Jockers, Matthew L. (2013) *Macroanalysis: Digital methods and literary history*. U. Illinois Press. Jockers, Jockers, Matthew L. (2012) The LDA Buffet: A Topic Modeling Fable. https://www.matthewjockers.net/macroanalysisbook/lda/

Jurafsky, Dan, et al. (2014) "Narrative framing of consumer sentiment in online restaurant reviews." *First Monday 19*(4). https://doi.org/10.5210/fm.v19i4.4944.

(SLP) Jurafsky, Daniel, and James H. Martin. 3rd ed. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. In prep. but avail. at https://web.stanford.edu/~jurafsky/slp3/

Jurish, Bryan et al. (2014). Querying the *Deutsches Textarchiv*. In: *MindTheGap@ iConference*. 25-30. kaskade.dwds.de/~moocow/software/DTA-CAB/doc/html/DTA.CAB.WebServiceHowto.html

Kestemont, M. (2012). Stylometry for medieval authorship studies: an application to rhyme words. *Digital Philology: A Journal of Medieval Cultures*, *1*(1), 42-72.

Kestemont, Mike, et al. (2017) *Van wie is het Wilhelmus?: de auteur van het Nederlandse volkslied met de computer onderzocht*. Amsterdam University Press.

Kestemont, Mike, et al. (2017) "Did a poet with donkey ears write the oldest anthem in the world?" *Digital humanities 2017. Book of abstracts (Montreal, August 9, 2017)*.

Kestemont, Mike, and Folgert Karsdorp. "Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity." *Proceedings http://ceur-ws. org ISSN* 1613 (2020): 0073.

Kestemont, Mike, et al. "Forgotten books: The application of unseen species models to the survival of culture." *Science* 375.6582 (2022): 765-769.

Klaussner, Carmen et al. (2015). Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities*, *30*(suppl_1), i114-i129.

Liberman, Mark et al. (2009) "Conditional entropy and the Indus Script" Discussion of Yadav et al. (2010) languagelog.ldc.upenn.edu/nll/?p=1374

Masías, Víctor Hugo, et al. (2015) "Shakespeare, social media and social networks." *IEEE Technology and Society Magazine* 34.4: 17-30.

Michel, Jean-Baptiste & Erez Lieberman Aiden (2011) "Culturomics" (TedTalk) https://auth.ted.com/users/new?context=ted.www%2Fmain-nav&referer=https%3A%2F%2Fwww.ted.com%2Fusers%2Fauth%2Fted_oauth2

Michel, Jean-Baptiste, et al. (2011) "Quantitative analysis of culture using millions of digitized books." *Science* 331.6014: 176-182.

Miller, John E., et al. (2020) "Using lexical language models to detect borrowings in monolingual wordlists." *PLoS ONE* 15.12: e0242709.

Mimno, David. (2012) "Computational historiography: Data mining in a century of classics journals." *Journal on Computing and Cultural Heritage (JOCCH)* 5.1: 1-19.

Moretti, Franco. (2013) *Distant reading*. Verso Books.

Nalisnick, Eric T., and Henry S. Baird. (2013) "Character-to-character sentiment analysis in Shakespeare's plays." *Proc. 51st ACL (Volume 2: Short Papers)*.

Napier, Kathleen, and Lior Shamir. "Quantitative sentiment analysis of lyrics in popular music." *Journal of Popular Music Studies* 30.4 (2018): 161-176.

Nerbonne, John (2007) "The exact analysis of text." *Foreword to 3rd ed.,* Frederick Mosteller and David Wallace *Inference and Disputed Authorship: The Federalist Papers.* CSLI: Stanford. xi-xx.

Newman, David J., and Sharon Block (2006). "Probabilistic topic decomposition of an eighteenth-century American newspaper." *Journal of the American Society for Information Science and Technology* 57.6: 753-767. Avail. at Citeseer.

Nguyen, Dong, et al. (2020) "How we do things with words: Analyzing text as social and cultural data." *Frontiers in Artificial Intelligence*. 1-14.

Oshikawa, Ray et al. (2020) "A survey on natural language processing for fake news detection." *Proc. 12th Language Resources & Evaluation Conference.* 6086-6093.

Plank, B., & Hovy, D. (2015). Personality traits on twitter—or—how to get 1,500 personality tests a week. In *Proc. 6th workshop comp. appr. subjectivity, sentiment and social media analysis* 92-98.

Qi, Qianqian et al. (2022) "A Comparison of Latent Semantic Analysis and Correspondence Analysis for Text Mining." *arXiv preprint arXiv:2108.06197*.

Rao, Rajesh PN, et al. (2009) "Entropic evidence for linguistic structure in the Indus script." *Science* 324.5931 1165-1165.

Robinson, Peter (2003) "The History, Discoveries, and Aims of the Canterbury Tales Project." *The*

Chaucer Review 38.2: 126-139. https://www.jstor.org/stable/pdf/25094241.pdf

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, *11*(12), e0168843.

Sprugnoli, Rachele, et al. (2016) "Towards sentiment analysis for historical texts." *Digital Scholarship in the Humanities* 31.4: 762-772.

Štajner, Sanja, & Seren Yenikent. (2020) "A survey of automatic personality detection from texts." *Proc. 28th international COLING*. 6284-6295.

Taboada, Maite,et al. (2006) "Sentiment classification techniques for tracking literary reputation." *LREC workshop: towards computational models of literary analysis*. Avail. via CiteSeer.

Tehrani, Jamshid J. (2013) "The phylogeny of little red riding hood." *PloS one* 8.11: e78871.

Thaisen, Jacob. (2013) "Gamelyn's Place among the early exemplars for Chaucer's Canterbury Tales." *Neophilologus* 97.2: 395-415.

Verhoeven et al. (2016) "Twisty: a multilingual twitter stylometry corpus for gender and personality profiling." *Proc 10th LREC*

Waseem, Z., & Hovy, D. (2016). "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In *Proceedings of the NAACL student research workshop*. 88-93.

Yadav, N., Joglekar, H., Rao, R. P., Vahia, M. N., Adhikari, R., & Mahadevan, I. (2010). Statistical analysis of the Indus script using n-grams. *PloS one*, 5(3). https://doi.org/10.1371/journal.pone.0009506

Including Language Log discussion, see Liberman (2009) languagelog.ldc.upenn.edu/nll/?p=1374