

Visual Text Analytics: Designing Visualizations for Linguistics and Digital Humanities Research

Instructors: Dr. Christopher Collins - christopher.collins@uoit.ca
Mennatallah El-Assady - menna.el-assady@uni.kn

Description:

This seminar will bring together text analytics research with visual analytics research to investigate the challenges particular to providing visual text analytics tools for linguistics and digital humanities research. The course will cover the basics of text visualization, then move on to analysis tasks common in digital humanities, the state-of-the art visualization techniques for text data, and considerations of usability, trust, and supporting existing work practices in interdisciplinary research.

During the semester, students will meet for a combination of lectures, student presentations of assigned readings, progress updates and peer feedback about course projects. The main output of this seminar will be a well-written and documented implementation of a visual analytics tool to support a specific line of research in linguistics or digital humanities.

Literature:

Jänicke, Stefan and Franzini, Greta and Cheema, Muhammad Faisal and Scheuermann, Gerik. **On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges.** Eurographics Conference on Visualization (EuroVis) - STARs., 2015. https://diglib.eg.org/handle/10.2312/eurovisstar.20151113.083-103

Additional readings will be assigned individually to registered students before each lecture, and students will be expected to augment each reading with a paper of their own choice.

Conditions:

Successful completion of Document Analysis or completion of an online training course (http://bocoup-education.github.io/text-vis-ovc/) before the first class meeting. Students wishing to complete the online tutorial should contact the instructor directly.

It is recommended that students also have completed a course in Information Visualization or Visual Analytics.

Programming skills in Python along with Java and/or Javascript are highly recommended.

Certificate of Achievement:

Successful completion of this seminar will include:

- A 20 minute presentation of prior research and leading of a discussion *on both lecture days*
- Implementation of a visual analytics tool on a specific problem and dataset (topics must be approved by the instructor)
- A 20 minute end-of-term presentation covering the motivations, design decisions, and final implementation of the designed system
- A ~10 page project report including an analysis of background literature and a plan for evaluation

Learning Objectives:

Students will be able to enumerate the design considerations of visual text analytics, and describe the common tasks in linguistics and digital humanities which can be supported by analytics systems. By investigating the literature in this field, as well as hands-on development of an implemented system, students will be able to carry out appropriate research methods for eliciting requirements, creating designs, refining designs through evaluation, and merging text analytics algorithms with visualization algorithms.

Objective:

Students will review common text mining algorithms and visualization toolkits, and how to combine these systems to create new tools for visual text analytics. Students will implement systems which demonstrate their ability to apply appropriate levels of automation to support qualitative research and open-ended analysis.

Workload:

120 hours, of which 4 are spent in class lectures, 4 are spent in class discussions and presentations of readings, 3 are spent in hands-on activities and project brainstorming, and 3 are spent in final presentations. Another 106 hours are expected to be spent on individual work --- reading research papers, preparing presentations, learning new software toolkits (may begin before first class meeting), software development, and report writing.

Course Outline:

Nov 11: Introduction to course project, introduction lecture (1.5h)

- Lecture topics:
    - Course overview
    - What is Digital Humanities?
    - Distant Reading / Franco Moretti
    - Description of course projects, advice on choosing a topic
- Reminder to sign up for readings, project

Nov 16: Submit 0.5 page project proposal by ILIAS

Nov 18: Project pitches + papers discussion lead by students (1.5h)

- Project pitches: Students present their idea for a course project, including the dataset they will use, and the tasks they will support. Project proposals should incorporate any feedback received from the instructors.
- Paper discussion: Each student will lead a presentation and discussion (max 20 minutes each) on one or more papers they have selected. Paper choices must be approved, multiple papers must be thematically linked. The focus of this session is *past works in visual text analytics for the digital humanities*.

Nov 22: Lecture in Document Analysis (1.5h)

- Lecture Topics:
    - Text visualization techniques and challenges overview
    - Challenges specific to visualization analytics in digital humanities
    - Review of NLP algorithms and toolkits
    - Review of visualization toolkits
    - Review of information visualization approaches based on Visualization, Analysis and Design by Tamara Munzner (AK Peters, 2014)

Dec 2: Interim project updates, present prior research related to your project, practical activities (1.5h)

- Project updates: Students will show work-in-progress, design sketches, prototypes. Students are invited to share unexpected challenges, new ideas, and receive feedback and help from peers and instructors.
- Paper discussion: Each student will present one or more papers relevant to *their specific project*. Papers may be selected to describe the domain problem, or they may describe related approaches from the visualization and visual analytics literature. Total presentation and discussion time 20 minutes per student.
- Practical Activities: We will run a design brainstorming and critique activity in class.

Dec 6: Lecture in Document Analysis Course (1.5h)

Lecture topics:

- Open challenges in visual text analytics with a focus on digital humanities;
- Potential guest lecture presenting the humanities point of view;
- In depth case studies from the research of the instructors: *VisArgue* and *Metatation*;
- Critical visualization

Dec 9: Papers discussion lead by students (1.5h)

- Paper discussion: Each student will lead a presentation and discussion (max 20 minutes each) on one or more papers which they have selected. Paper choices must be approved, multiple papers must be thematically linked. The foci of this session are *case studies in interdisciplinary collaboration in visual analytics, what can visual analytics learn from the humanities, and evaluation methods*.
- Students will give a brief project update (2-3 minutes)

Jan 20: Final project presentations and hand in report

- Students will present their final project using a demo, slides, or video (20 min). Presentations and reports should show all the features, describe the rationale behind design decisions, given an overview of the implementation, discuss an evaluation plan, and future work ideas.

Online:

- Course management through ILIAS: https://ilias.uni-konstanz.de/ilias/goto_ilias_uni_crs_602538.html
  - Hand in project proposals
  - Sign up for paper presentations
  - Submit final report
  - Schedule 1 on 1 meetings as needed
- Code management through GitLab: https://git.uni-konstanz.de/
  (Please grant us developer access to your project)

Suggested Papers for Discussion November 18 (you may choose others with approval):
Topic: past works in visual text analytics for the digital humanities

- Choose one or more papers from the past 5 years of the *IEEE InfoVis*, *VAST*, *EuroVis*, *Digital Humanities*, or *Joint Conference on Digital Libraries* conferences or the journals *Transactions on Visualization and Computer Graphics*, *Computer Graphics Forum*, or *Digital Humanities Quarterly* related to text visualization, especially in the humanities.
- Projects listed on http://textvis.lnu.se or mentioned in Jänicke et al.

Suggested Papers for Discussion December 9 (you may choose others with approval):
Topic: case studies in interdisciplinary collaboration in visual analytics, what can visual analytics learn from the humanities, and evaluation methods.

- Choose one or more interdisciplinary papers from the 2016 VIS4DH workshop (http://vis4dh.com), past TextVis workshops with a focus on digital humanities (http://textvis.org), or a venue of your choice (with approval) and discuss the contributions to both visualization and digital humanities.
- Review several visualization for digital humanities papers with attention to the evaluation methods
- Describe, demonstrate, and critique visualization system deployed for digital humanities as found online (you can find some through http://tapor.ca or Google).

Evaluation:
- Quality of in class presentations 30%
- Discussion participation 10%
- Project proposal 10%
- Final project 50%

Course Project - Visual Text Analytics

The Project Challenge

Your course project is to create a visual text analytics tool which addresses a need in the digital humanities. Your project should start with identifying your project challenge, selecting a dataset, identifying a *task* (what do you want to show and why), and sketching ideas. You can meet with the instructors to discuss your sketches and plans, and present them to the class during the project update. Finally, you should implement your idea with a working back end (data processing) and front end (visual interface), write a report about your work, and give a demo and presentation.

The overarching challenges in visual text analytics which we invite you to address are:

- Unifying Close and Distant reading
- “Prosthetic Reading” vs. “Slow Analysis”
- Supporting Serendipitous and Guided Discovery
- Linking Outside Resources into the Analysis
- Aesthetic Experiences
- Annotation of Resources

You can find examples of projects which address these challenges in the slides from the November 11 lecture.

Available Datasets

You may work with a dataset below, or you are *encouraged to find an alternative dataset online* - there is too much interesting data available for us to scratch the surface with this list!

Note while the focus is on humanities data (e.g. history, literature), we are open to social sciences data as well (e.g. linguistics).

Bulk Datasets from the British Library: https://data.bl.uk/
Datasets from NLTK (see http://www.nltk.org/book/ch02.html)
VisArgue data (available from instructors)
Project Gutenberg Top 100 Books (available from instructors or from https://www.gutenberg.org)
Public data extracted from the Hathi Trust (see https://goo.gl/ANOYuz)
Baldwin Collection of Historical Children’s Literature (available from instructors)
US Supreme Court decisions (available from instructors)
NRC Emotion Lexicon http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
WikiLeaks data
Enron emails dataset
Film Dialogue Data https://github.com/matthewfdaniels/scripts/

VisPubData http://www.vispubdata.org/site/vispubdata/
Shakespeare Corpus http://lexically.net/wordsmith/support/shakespeare.html
Westbury Lab Wikipedia Collection
http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html
Documenting the American South http://docsouth.unc.edu/docsouthdata/

Describing your User and Task

When selecting a dataset, think about *who* may be interested in this data and *what* they may be interested in. For example, a political scientist may be interested in patterns of argumentation, a linguist may be interested in richness of vocabulary, a literary historian may be interested in the locales in literature over time. Whatever you choose to do, you should propose a specific target stakeholder group and task. You can use online research about the domain to justify your characterization of users and tasks. An ideal method for this step would be to first identify datasets and potential users, and connect with those potential users to interview them to better understand their needs. Due to the shortened nature of the project timeline, we will be asking you to propose the task based on your best judgement and online research. One way to justify a proposed task is to look at tasks which people currently carry out manually. These are often appropriate for visualization to assist in the analytic process (though, not always, as we will discuss in class). Your instructors can help you refine your specific target group and task, and potentially help connect you with a representative of your target community.

Project Report

Project reports and presentations are due January 20. Both presentations and reports should show all the features, describe the rationale behind design decisions, given an overview of the implementation, discuss an evaluation plan, and future work ideas. Reports are expected to be ~10 pages and should include clear images of your prototype. Source code should be submitted through GitLab: https://git.uni-konstanz.de/.

You are free to choose any implementation technologies you find appropriate for your project. Suggestions include using *D3.js* for web-based visualization and Java/Python tools such as *NLTK* for text processing. For a more complete text processing pipeline, you may use the *VisArgue* framework as a platform for your work. *VisArgue* provides standard NLP preprocessing steps in an integrated Java/JavaScript pipeline and works with D3. **If you want to use *VisArgue* contact Menna about attending a tutorial.** Another coding library that may be useful is *TidyText*, which is available for R. See examples here: http://juliasilge.com/blog/Life-Changing-Magic/

You may also make stand alone applications (e.g. Processing/Java) if desired. We will not be assessing code quality, however, we do expect an overview of your implementation, with special attention to what tools and technologies you used, any innovative algorithms you developed, and any approaches to challenges such as scalability. Web-based projects are encouraged as it will allow instructors to try your tools. Otherwise, you may have to meet the

instructors outside of class time to present a detailed demonstration. Alternatively, you can submit a project video demo as part of your report.

Grading of the project will assess:

- Is the project goal (problem, task, and user) well stated in the report?
- Does the design address the stated problem, task, and user?
- Are the design decisions well-justified? Are alternative designs which were considered and rejected?
- Is the evaluation plan technically sound and practical?
- Does the visualization follow good design principles for visual analytics generally?
- Is the use of NLP sufficient and appropriate?
- Quality and clarity of writing.

Milestones

Project pitch due on ILIAS **November 16**
Project pitch presented to class **November 18**
Project update to class **December 2**
Final project report (ILIAS), code (GitLab) and presentations (in class) **January 20**