Cecilia Graiff, Richard Prußas, Fabian Strobel:

DraCor

01.02.2022

1. Problem and Research Question

- ‚DraCor' short for ‚drama corpus' → collection of mostly european plays

- accessible through API or web platform dracor.org

- data: raw text data, metadata, network data

- most existing research specific to certain kind of drama, often SNA

- RQ: How can we represent the history of plays using stylometry?

  - How does that representation relate to findings of traditional literary studies?
  - How does the choice of textual properties to compare influence the results?
  - What are similarities/differences in italian vs. german drama?

# 2. Methods

1. Loading the corpora, exploratory data analysis

2. Vectorization with tf-idf (evt. lemmatization, removal of stop words, …)

3. Dimensionality reduction with PCA

4. Choice of cluster number with elbow plot or based on prior existing analyses

5. K-means clustering

6. Export of data (probably as .csv file)

7. Import of data in R-Shiny and display of visualizations in an interactive Web-App

3. Expected Results
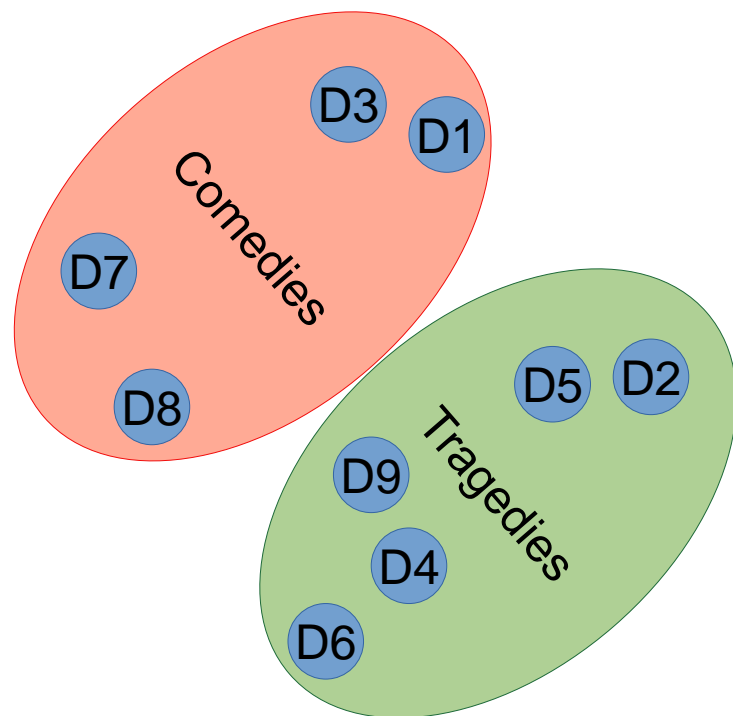
We hope for the results to represent categories from comparative literature analysis.
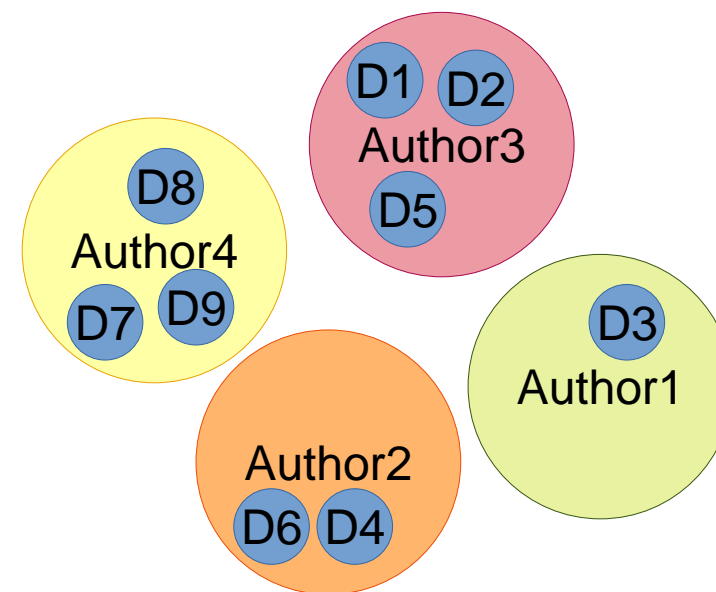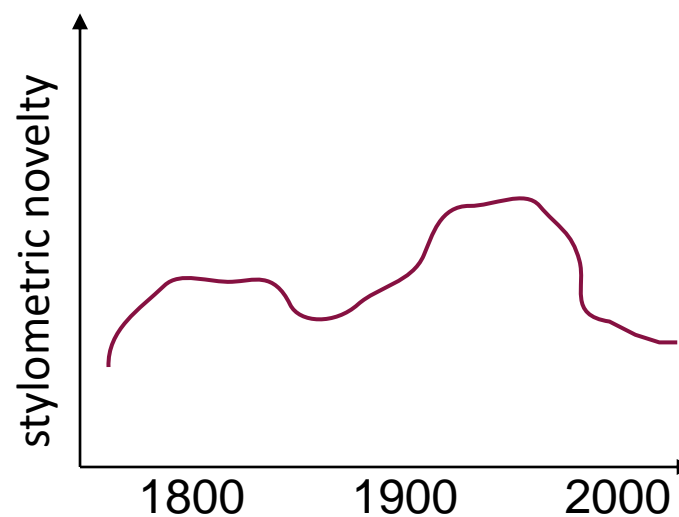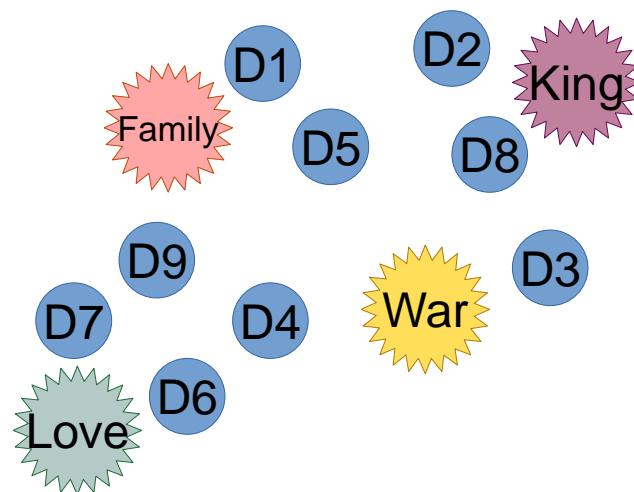
<u>Bibliography</u>

● **Chang, K. and DeDeo, S. (2020): Divergence and the Complexity of Difference in Text and Culture. In: Journal of Cultural Analytics 4.11 (2020): 1-36. doi: 10.22148/001c.17585.**

● **Fischer, F., et al. (2019): Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, doi:10.5281/zenodo.4284002.**

● **Jurczyk, T. (2021): Clustering with Scikit-Learn in Python. Programming Historian, (10).**

● **Pfister, M. (2001): Das Drama.**

● **Scherer, S. (2013): Einführung in die Dramen-Analyse.**