LitViz: Visualizing Literary Data by Means of text2voronoi

Tolga Uslu, Alexander Mehler, Dirk Meyer

Abstract

We present a LitViz, a webbased tool for visualizing literary data which utilizes the text2voronoi algorithm Mehler et al. (2016b) to map natural language texts onto voronoi diagrams. These diagrams can be used, for example, to visually differentiate between (groups of) authors. Text2voronoi utilizes the paradigm of text visualization to reconstruct text classification (e.g., authorship attribution) as a task of image classification. This means that, in contrast to conventional approaches to text classifiction, we do not directly use linguistic features, but explore visual features derived from the texts' visualizations to perform operations on texts. We illustrate LitViz by means of 18 authors, each of whom is represented by 5 literary works.

Keywords: Distant reading, text visualization, text imaging, text2voronoi

1. Introduction

In this paper we present a new tool, called LitViz, for the visual depiction of literary works. To this end, we utilize the text2voronoi algorithm (see Mehler et al. (2016b)) which maps natural language texts to image representations. The idea is to generate images of texts which can be used instead of these texts' symbolic information to characterize them, for example, in terms of authorship, topic or genre. Text2voronoi is in line with the paradigm of text visualization to reconstruct text classification (e.g., authorship attribution) as a task of image classification. In contrast to conventional approaches to text classification, we therefore do not directly use linguistic features, but explore visual features derived from the texts' visualizations in order to identify, for example, their authors. We exemplify LitViz by means of 18 authors each of whom is represent by 5 literary works. LitViz allows for interacting with the visualizations of these works in two modes: two- and three-dimensionally (see Figure 1 and 2).

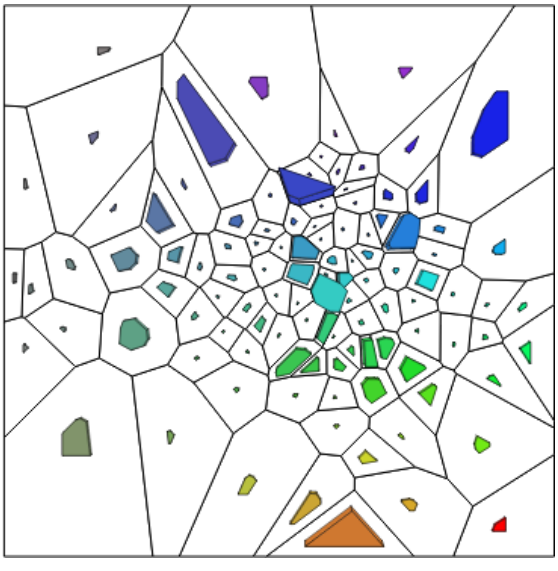

Figure 1: Visual depiction of E.T.A. Hoffmann's *Das steinerne Herz*

2. Related Work

The idea of visualizing literature was inspired by Martin Wattenberg's *The Shape of Song*[1] (Wattenberg, 2001; Wattenberg, 2002). Wattenberg explores identical or otherwise repetitive passages of a composition to visually depict them. This is done by means of semicircles, which combine repeated and repetitive positions in such a way that the micro- and macro-structure of a composition becomes visible. Our idea is to transpose this idea to the visualization of literary data.

Kucher and Kerren (2015) give an overview of state-of-the-art techniques of text visualization and present a website that allows for differentiating between these techniques.

Cao and Cui (2016) provide a systematic review of many advanced visualization techniques and discuss the fundamental notion of information visualization.

Mehler et al. (2016a) present a web tool called Wikidition which allows for automatically generating large-scale editions of text corpora. This is done by using multiple text mining tools for automatically linking lexical, sentential and textual data. The output is stored and visualized using a MediaWiki. Thus, any Wikidition is extensible by its readers based on the wiki principle.

Rockwell and Sinclair (2016) present a detailed web tool, called Voyant tools, for visualizing texts. Unlike Voyant, our focus is on non-standard techniques of visualizing textual data that go beyond histograms, scatterplots, line charts and related tools.

Generally speaking, text visualization supports distant reading as introduced and exemplified by Moretti (2013), Rule et al. (2015) and Michel et al. (2011). These approaches show how visualizations that support distant reading may look like to get overviews of documents by just looking at the final visualizations. LitViz is a tool following this tradition: it utilizes text2voronoi to extend the set of techniques mapping textual data. In this way, it combines Wattenberg's approach with distant reading techniques from the point of view of text visualization.

[1] http://turbulence.org/Works/song/gallery/gallery.html

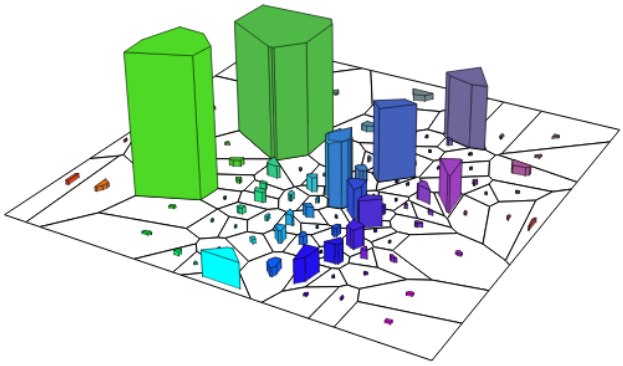

Figure 2: 3D visualization of Franz Kafka's *Der Kübelreiter*.

3. Model

Our goal is to generate images from literary works in a way that text classifiers can be fed by the features of these iconic representations in order to perform classification experiments, for which usually linguistic features are explored. This is the task of the text2voronoi algorithm, which calculates image representations of texts in four steps Mehler et al. (2016b): In the first step, the input text is analyzed by means of TextImager Hemati et al. (2016) to extract linguistic features in the usual way, that is, features, spanning a vector space of linguistic data. In the second step, the resulting vector space is used to compute embeddings for each of the extracted linguistic features. Embeddings are produced by means of word2vec (Mikolov et al., 2013). In the third step, a voronoi tessellation of the embedded features is computed. As a result, each lexical feature is mapped onto a separate voronoi cell whose neighborhood reflects the feature's syntagmatic and paradigmatic associations with other features of the same space. The topology of the voronoi cells spans a voronoi diagram that visually represents the input text. Each of these cells is characterized by its filling level, transparency and height (third dimension) thereby reflecting its co-occurrence statistics within the input text, while the position and size of a cell is determined by the embedding of the corresponding feature – for the mathematical details of this algorithm see Mehler et al. (2016b). Finally, the text2voronoi algorithm extracts visual features from the voronoi diagrams to feed classifiers performing classifications of the input texts.

LitViz utilizes the first three steps of this algorithm. Unlike the classical text2voronoi procedure, it does not address the final step of classification. Rather, it gives access to voronoi diagrams of input texts via a two-dimensional graphical interface, which can be transformed into a three-dimensional one by means of user interaction. These two- and three-dimensional text representations can be used by the user of LitViz to interact with the underlying input texts in order to highlight single voronoi cells, to change her or his reading perspective or to visually compare voronoi diagrams of different texts. In this way, LitViz paves the way to a kind of a *comparative distant reading* by making accessible the visual depictions of different texts in an interactive manner.

4. The LitViz Tool

We have selected 18 authors of German literature each of whom is represent by 5 literary works. The works are taken from the Project Gutenberg (`https://www.gutenberg.org/`) and visualized by means of the text2voronoi algorithm. Any of these examples is made accessible by the front page of LitViz (see Figure 3). When hovering over a voronoi cell of the voronoi diagram of a sample work, information about the underlying linguistic feature represented by this cell is displayed. According to Mehler et al. (2016b), we call these images *VoTes*: Voronoi diagram of a Text. LitViz presents VoTes via a graphical user interface for two- and three-dimensional interactive graphics. In this way, we go beyond Wattenberg's 2D depictions of musical pieces.

The second page (tab) of LitViz gives access to the comparison tool. Here the user first selects the number of VoTes to be compared. Then the user selects a subset of works of the authors to be compared. In the example in Figure 5, we compare four VoTes of two authors: two VoTes of two works of Heinrich Heine (top) and two VoTes of Heinrich Mann (bottom). It is easy to see that these VoTes fall into two classes, depending on the underlying authorship. Heinrich Mann's two VoTes are organized around a center that is composed of many small cells, while there is a small subgroup of peripheral cells that are large. In contrast to this, the two VoTes of Heinrich Heine do not display such a center and are more evenly distributed in terms of their size. It is a main task of LitViz to allow for such comparisons. In this way, that is, by interacting with the texts' image representations and by using the mouse-over technique, the user can study single features and how they are related to other features of the same representational space.

Last but not least, LitViz provides a so-called custom tab. Here, the user can upload and visualize its own texts. It is then possible to set filter options using an option tool (see Figure 4) in order to further restrict the visualization.

5. Conclusion

We introduced a novel web tool, called LitViz, for visually depicting natural language texts based on the text2voronoi algorithm. LitViz enables the comparison of the visualizations of different texts. This allows, for example, for comparing the styles of the underlying authors *visually*. In this way, we extend the existing tool palette of distant reading. LitViz can be accessed via `http://alba.hucompute.org/text2voronoi`.

6. Bibliographical References

Cao, N. and Cui, W. (2016). *Introduction to Text Visualization*. Atlantis Briefs in Artificial Intelligence. Atlantis Press.

Hemati, W., Uslu, T., and Mehler, A. (2016). TextImager: a distributed uima-based system for NLP. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 59–63.

Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community in-

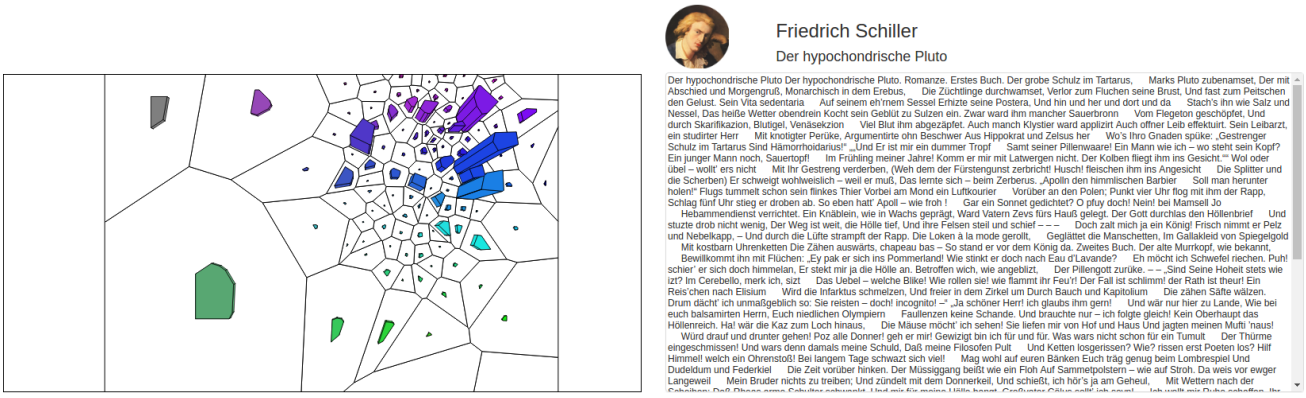

Figure 3: Front page of LitViz.

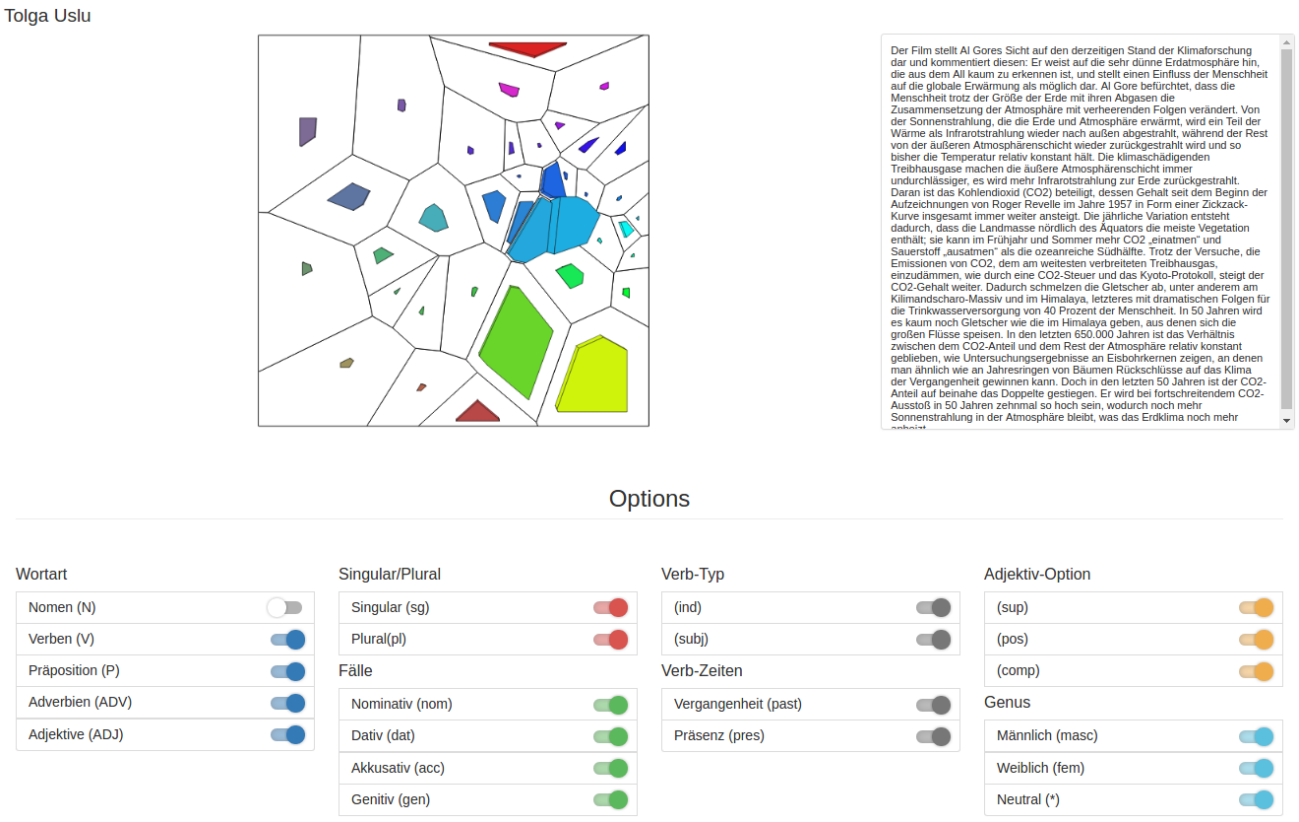

Figure 4: Custom VoTe with filter options.

sights. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pages 117–121. IEEE.

Mehler, A., Gleim, R., vor der Brück, T., Hemati, W., Uslu, T., and Eger, S. (2016a). Wikidition: Automatic lexiconization and linkification of text corpora. *Information Technology*, pages 70–79.

Mehler, A., Uslu, T., and Hemati, W. (2016b). Text2Voronoi: An image-driven approach to differential diagnosis. In *Proceedings of the 5th Workshop on Vision and Language (VL'16) hosted by the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin*.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Rockwell, G. and Sinclair, S. (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.

Rule, A., Cointet, J.-P., and Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.

Wattenberg, M. (2001). The shape of song. *Website http://www. turbulence. org/Works/song/mono. html*.

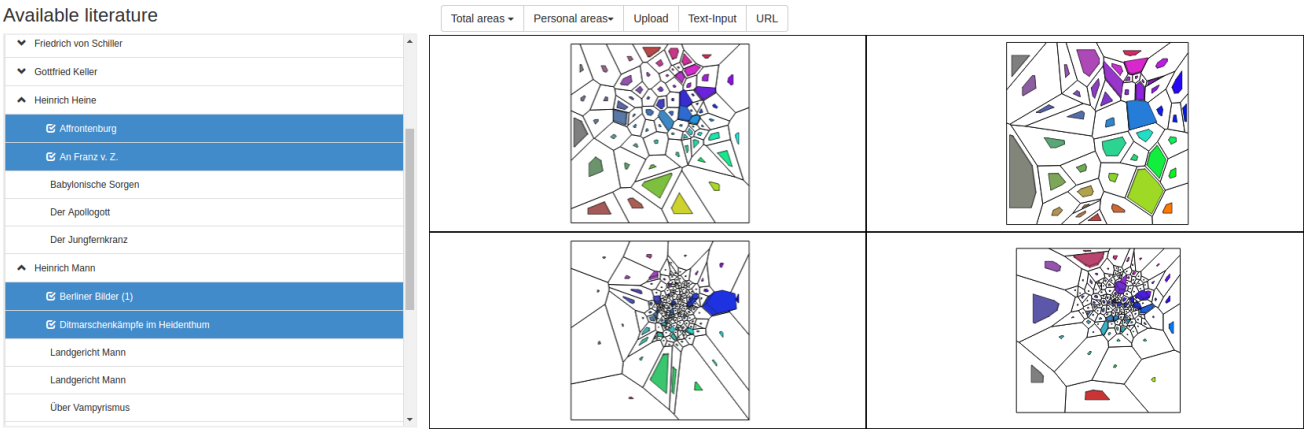

Figure 5: Comparison tool: Heinrich Heine (top) in comparison to Heinrich Mann (bottom).

Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110–116. IEEE.