Word Clouds with Spatial Stable Word Positions across Multiple Text Witnesses

Dähne, Janis

janis.daehne@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Pöckelmann, Marcus

marcus.poeckelmann@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Ritter, Jörg

joerg.ritter@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Introduction

Word clouds are a common and widespread visualization technique [1][2], with prominent layout algorithms like Wordle [3], and can be seen as a distant-reading tool. Word clouds usually represent the supposedly most meaningful words of a text, according to some criteria, e.g., word length, part-of-speech tags, frequency (occurrences), etc. To be able to apply word clouds to larger text witnesses, the word clouds must be restricted to the top X words. Especially if the word clouds have to be dynamically regenerated when the criteria change. The number X should be specified by the user as a parameter.

Another criterion, especially for interactive applications, is to define a selection window of a desired length that restricts the words displayed to those within the selection window. The window can be moved over the text to restrict the scope of the word cloud to the desired segments. The word cloud changes when the window is moved, as different words need to be displayed. This raises a problem when the same words occur in successive word clouds. To track the relevance of words throughout the text, each word should be in the same place in all successive clouds while moving the window, i.e., the clouds should be spatial stable. The few approaches [4][5][6] known in literature to generate stable word clouds are discussed in full detail by Herold et at. [9], which also introduces a first algorithm to keep word positions spatial stable in one cloud.

This paper describes how the approach in [9] can be generalized to multiple word clouds, allowing it to be used for comparing multiple text witnesses in scholarly editions. Our generalization ensures that the same word is always placed at the same position across all clouds even if the selection window is moved. We call this property synchronous. We also describe how to improve the necessary calculations that allow us to use multiple clouds interactively.

We have integrated our approach into LERA [7][8], a collation tool for scholarly editions [1]. LERA can be used to discover and inspect changes among text witnesses and we think our stable word clouds can facilitate this process.

Previous state

A simple solution to guarantee spatial stable word positions would be to reserve a separate position for each unique word. This is not practical as the available drawing space is limited. However, it is possible to identify groups of words that never occur together at the same time and thus can share the same position.

Identifying such groups is achieved in [9] by building a word-segment occurrence matrix, merging consecutive segment vectors in order to represent the selection window and finally identifying orthogonal word vectors that indicate that the corresponding words never occur together within the selection window. This former approach was originally integrated in LERA [9], see Fig. 1, but has some issues. It does not take into account the actual word dimensions, thus cannot guarantee an overlap-free layout. Another limitation is, that it can only handle a single text witness. For multiple text witnesses the words might appear at a different positions in each cloud, breaking spatial stability across the word clouds. Finally, the runtimes are too slow for interactive applications, especially for multiple text witnesses.

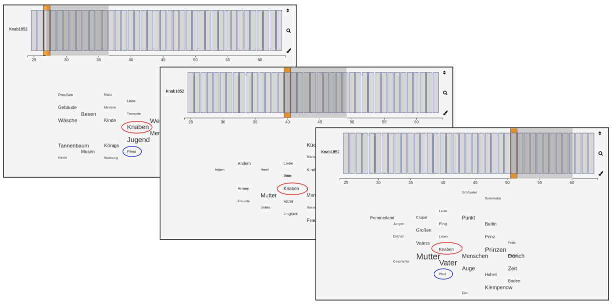


Figure 1. Spatial stable word cloud for the single text witness “Aus der Knabenzeit” by Karl Gutzkows (edition of 1852) for three different selection windows (beginning, middle and end of the text). The words stay (e.g. *Knaben*, *red*) and reappear (e.g. *Pferd*, *blue*) at fixed positions.

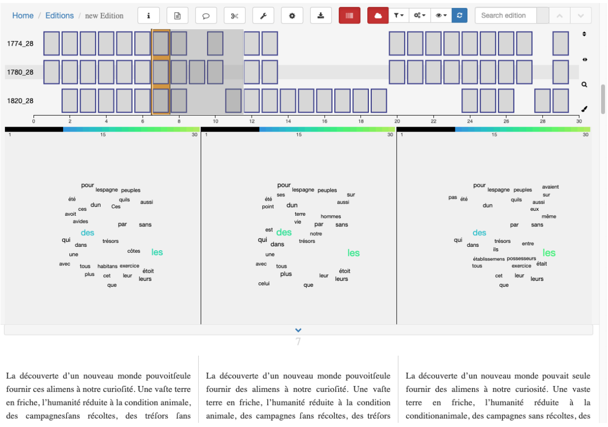
Improving the spatial stable clouds

A major improvement in runtime was achieved by utilizing bitsets/bit-vectors for the word vectors in the occurrence matrix. This is possible since only the occurrence of a word in each segment has to be stored. The orthogonal check for two bitset word vectors is very efficient as it uses bitwise operations and we further optimized it, by returning a negative result as fast as we encounter two non-orthogonal bytes. The groups can now be understood as the bitwise *OR* of the vectors of all words within a group.

We also improved the runtime of creating the groups described in [9] by removing some checks and instead reordering the word vector positions in the occurrence matrix. In our implementation, we used word frequencies as a criterion for word relevance, as this is most useful for our applications and it allows for another optimization. We precalculate a matrix whose columns represent the cumulative frequencies of a word, so that we can calculate the frequency of a word for any window size in one operation.

The layout of the clouds is also improved by ensuring that no word overlaps can occur, allowing for an even more compact layout. This is achieved by measuring all words with their required font sizes beforehand in the browser. An accurate bounding box for a group is than determined by the largest bounding box of its members. To get an upper bound on the required space for the words we declared a maximum font size for the words. This would change the semantics of the cloud, as the font size correlates with the frequency (or some other measure). We compensated for that by introducing a color scale that is used for frequencies above some threshold. This is in contrast to other word clouds that normally uses colors randomly to better distinct words but do not carry any additional information.

Finally, we generalized the spatial stable word clouds for multiple text witnesses by making them synchronous., i.e., words occurring in more than one text witness are always placed at the same position across all word clouds, even if the selection window is moved, see Fig. 2. This makes it easier to visually track a word across word clouds. We achieved this by taking all word vectors of all texts into account during the process of creating groups. However, we still need to process information for every text individually, e.g., top X words, in order for every word cloud to be used on its own. All steps are optimized in a way that all interactions with the clouds, like moving the selection window or changing the size of the selection window can be done interactively.



La découverte d’un nouveau monde pouvoitſeule fournir ces alimens à notre curioſité. Une vaſte terre en friche, l’humanité réduite à la condition animale, des campagnesſans récoltes, des tréſors ſans

La découverte d’un nouveau monde pouvoitſeule fournir des alimens à notre curioſité. Une vaſte terre en friche, l’humanité réduite à la condition animale, des campagnes ſans récoltes, des tréſors

La découverte d’un nouveau monde pouvait seule fournir des alimens à notre curiosité. Une vaste terre en friche, l’humanité réduite à la conditionanimale, des campagnes sans récoltes, des

Figure 2. Spatial stable word clouds can now share word positions across multiple text witnesses, in this case for three editions of “Histoire des deux Indes”. The prominent words are “les” and “des” (by frequency) for an interactively adjustable selection window of 5 segments, starting at row seven in the text alignment, only the top 30 words in this window by frequency are used. Note that all words, that occur in multiple clouds, are at the same positions.

Conclusion

The stable word clouds presented in this paper allow to effectively track changes in the relevance of words throughout a text, even across multiple text witnesses. In combination with a collation tool for providing the data basis and a selection tool, the synchronous spatial stable clouds provide a novel distant-reading approach to analyze and visualize text changes for scholarly editions. The improvements presented here make the approach so efficient that it can be used in interactive applications. An interactive demo version of the entire system can be found at: https://lera.uzi.uni-halle.de.

Notes

1. Homepage of LERA: https://lera.uzi.uni-halle.de

Bibliography

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., Granitzer, M. (2008): *On the beauty and usability of tag clouds*. In: 2008 12th International Conference Information Visualisation. pp. 17–25. IEEE. https://doi.org/10.1109/IV.2008.89

Lohmann, S., Ziegler, J., Tetzlaff, L. (2009): *Comparison of tag cloud layouts: Task-related performance and visual performance and visual exploration*. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 392–404. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-03655-2_43

Feinberg, J. (2010): *Wordle*. In: Steele, J., Iliinsky, N. (eds.) Beautiful Visualization Looking at Data Through the Eyes of Experts. O’Reilly Media, Newton

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H. (2010): *Context preserving dynamic word cloud visualization*. In: 2010 IEEE Pacific Visualization Symposium (PacificVis), pp. 121–128. IEEE. https://doi.org/10.1109/MCG.2010.102

Castella, Q., Sutton, C. (2014): *Word storms: multiples of word clouds for visual comparison of documents*. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 665–676. ACM. https://doi.org/10.1145/2566486.2567977

Silva e Silva, L.G., Assunção, R.M. (2018): **Cowords: a probabilistic model for multiple word clouds**. J. Appl. Stat. 45(15), 2697–2717. https://doi.org/10.1080/02664763.2018.1435633

Bremer, T.,Molitor, P., Pöckelmann, M., Ritter, J., Schütz S. (2015): *Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen - Das Fallbeispiel der Histoire philosphique des deux Indes von Guillaume Thomas Raynal*. In: Editio, R. v. Nutt-Kofoth, B. Plachta and W. Woesler (eds), Volume 29, Issue 1, pp. 29–51. https://doi.org/10.1515/editio-2015-004

Pöckelmann, M., Medek, A., Ritter, J., Molitor, P. (2022): *LERA — an interactive platform for synoptical representations of multiple text witnesses*. In: Digital Scholarship in the Humanities, Volume 38, Issue 1, pp. 330–346. https://doi.org/10.1093/llc/fqac021

Herold, E., Pöckelmann, M., Berg, C., Ritter, R., Hall, M.M. (2019): *Stable Word-Clouds for Visualising Text-Changes Over Time*. International Conference on Theory and Practice of Digital Libraries, TPDL2019, Oslo. Springer, Cham. pp. 224-237. https://doi.org/10.1007/978-3-030-30760-8_20