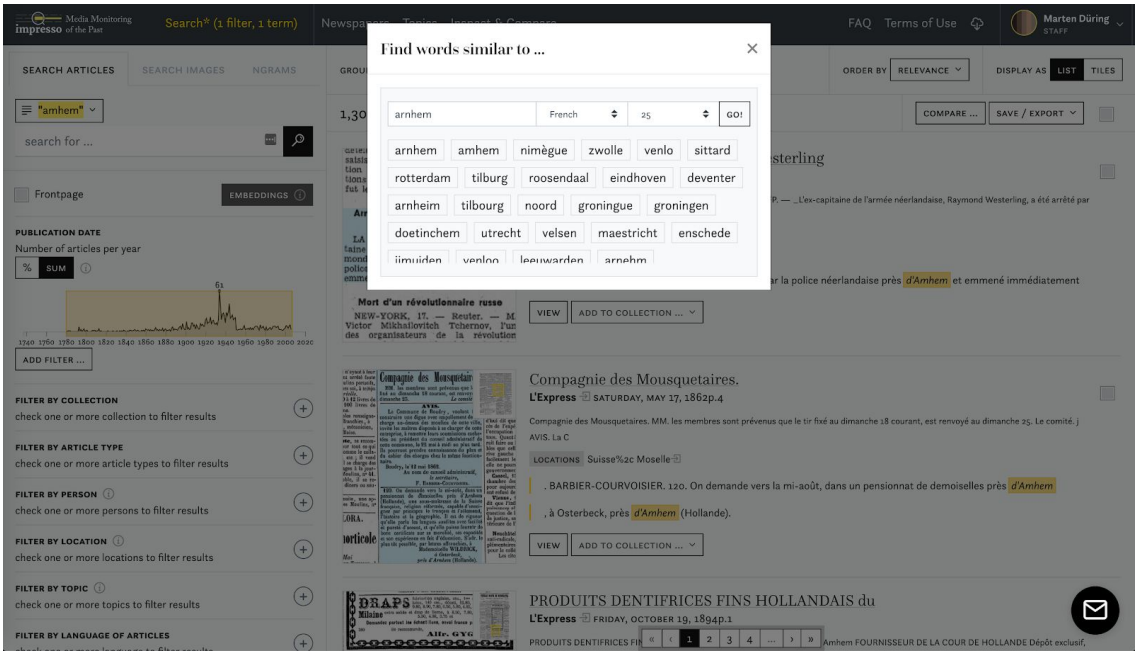
Deep-diving in NLP enhanced digitised newspapers: A hands-on session with the *impresso* interface.

The growing digitisation and online publication of millions of pages of newspapers does not equate with an easier access for research purposes. The interdisciplinary project *impresso*[1] developed a methodologically-reflected technological framework to enable new ways of engaging with multilingual digital content of historical newspapers and new approaches to address historical questions. More precisely, the project has three main foci: the application of text mining techniques to transform unstructured, large-scale and noisy and textual content into semantically indexed, structured, and linked data; the co-design and implementation of an innovative visualization interface to enable the seamless exploration of complex and vast amounts of historical data; and the effective use of these new tools and methods, whose strengths and weaknesses are best evaluated when challenged by scholars working on historical research questions. The making of the *impresso* interface is interdisciplinary and collaborative: text mining techniques and their accomodation in the interface are iteratively refined according to historians' needs and feedback, and regularly questioned from a (digital) history methodological and epistemological point of view.

[1] https://impresso-project.ch. The impresso project is supported by the Swiss National Science Foundation under grant CR- SII5_173719.



Screenshot from the impresso interface showing the word embeddings component.

The *impresso* interface[2] was designed with the overall ambition to facilitate content discovery based on text mining and the critical assessment of the underlying corpus following best practices in historical research. During the demonstration we will use a sample query to illustrate *impresso's* iterative exploration workflows. This includes, for example, word embeddings to retrieve frequent OCR mistakes, synonyms, word neighbors or spelling variations; search facets based on pre-existing metadata (e.g. title, country of publication) and newly extracted semantic annotations (e.g. topic models, named entities); the creation of collections with up to 10.000 articles and their visual comparison based on metadata; the (visual) exploration of text reuse clusters, topics and n-grams. The interface was designed to foster iterative query-building, meaning that a given query can be altered within a component by either broadening or restricting its scope. In practice, this translates in the capacity of a user to e.g. expand her search result lists by complementing the search terms with historical spelling variations, filter by inter-related topics, exclude irrelevant entities, and adjust time filters based on n-gram frequencies.

Our contribution will begin with a brief introduction of the project and will be followed by a demonstration of *impresso's* exploratory workflows using an example query. Participants are

[2] https://impresso-project.ch/app

encouraged to participate in a hands-on test of the interface (temporary login credentials will be provided). In the third and final part we will discuss questions which may have arisen during the demonstration or hands-on sessions.