I-Analyzer: a flexible interface for full-text search, filtering and visualization

Janssen, Berit
b.d.janssen@uu.nl
Utrecht University, the Netherlands

Stiphout, Mees
m.vanstiphout@uu.nl
Utrecht University, the Netherlands

van der Plas, Luka
l.p.vanderplas@uu.nl
Utrecht University, the Netherlands

Distant reading in a corpus of text resources is a recurring need in digital humanities research. It allows for quick extraction of key words and phrases from a text, see patterns in texts over time, or find a relevant subset of documents for close reading. With the availability of newspaper corpora with sufficient OCR quality and high-quality metadata, Utrecht University's Research Software Lab started development on the newspaper corpus tool *I-analyzer* in 2016. From the start, the goal of I-analyzer has been to provide researchers and students with the stepping stones that distant reading and other text mining can supply, without them having to be proficient coders. Built to be compatible with different data formats (currently xml, csv, and html) and configurable to collect and index text and most
types of metadata, I-Analyzer has since been extended with many digital humanities corpora - newspapers and otherwise - as well as more visualization and search functionality.

Other text mining applications have also been built for this goal, such as *Delpher* (Der Weduwen 2015) or *Voyant Tools* (Sinclair / Rockwell 2012), but they either do not provide the option to work with multiple corpora, or they do not offer full-text search, or filtering and visualization through metadata. The *Gale Digital Scholar* tool does offer all these functionalities to some extent, but, unlike I-Analyzer, it is a commercial product that researchers (or their affiliated institutions) need to pay to use. I-Analyzer is open-source, and several of the corpora hosted on the platform are as well, which means that anyone can use and even adapt I-Analyzer and the open-source corpora hosted there.

New corpora can be added to I-Analyzer relatively quickly, by defining the format of the source data. Data are indexed using *Elasticsearch*, which enables fast full-text search and filtering. The user is presented with an interface resembling a search engine: using a search bar and a side bar of metadata filters, they can inspect documents matching their search for close reading or download a subset of the corpus to process further offline. Currently, visualization options include document- and term frequency of the search term, as well as a representation of the most frequent n-grams including the search term. Moreover, for selected text corpora, diachronic word models were trained using Word2Vec (Mikolov et al. 2013) to inspect which words co-occur most frequently with a given search term over time, comparable to capabilities of *ShiCo* (Martinez-Ortiz et al. 2016) or *Hansard-shiny*.

I-Analyzer has been used in various research projects, ranging from the impact of translations on reader experience through book reviews (Kotze et al. 2021), the occurrence of concepts and names on Jewish funerary inscriptions (Saar 2021) to the conceptual history of words surrounding democracy in parliamentary speeches (Ihalainen et al. 2022). Currently, we are working on extending the infrastructure such that it is going to be possible for corpora to be added by researchers themselves, without the need for one of our developers to directly oversee this process. To this end, we invite feedback on the interface, and look forward to further contribute to the field of Digital Humanities with our continued development and support of I-Analyzer.

Bibliography

Ihalainen, Pasi / Janssen, Berit / Marjanen, Jani / Vaara, Ville (2022). "Building and Testing a
Comparative Interface on Northwest European Historical Parliamentary Debates: Relative
Term Frequency Analysis of British Representative Democracy." In *CEUR Workshop
Proceedings* (Vol. 3133). CEUR-WS.

Kotze, Haidee / Janssen, Berit / Koolen, Corina / van der Plas, Luka / Van Egdom, Gys-Walt (2021). "Norms, affect and evaluation in the reception of literary translations in multilingual online reading communities: Deriving cognitive-evaluative templates from big data." *Translation, Cognition
& Behavior* , 4(2), 147-186.

Martinez-Ortiz, Carlos / Kenter, Tom / Wevers, Melvin / Huijnen, Pim / Verheul, Jaap / Van Eijnatten, Joris / ... & Van den Bosch, Antal (2016). "Design and implementation of ShiCo: Visualising shifting concepts over time." *HistoInformatics* , 1632, 11-19.

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg S. / Dean, Jeff (2013). "Distributed
representations of words and phrases and their compositionality." *Advances in neural
information processing systems* , 26.

Saar, Ortal-Paz (2021). The PEACE Portal-Revisiting the Sea of Stone. DH Benelux 2021, 1-21.

Sinclair, S., & Rockwell, G. (2012). "Teaching computer-assisted text analysis: Approaches to
learning new methodologies." *Digital Humanities Pedagogy: Practices, Principles, and
Politics* , 241-63.

der Weduwen, Arthur. T. (2015). "Towards a complete bibliography of seventeenth-century Dutch
newspapers: Delpher and its applications." *Tijdschrift voor Tijdschriftstudies* .