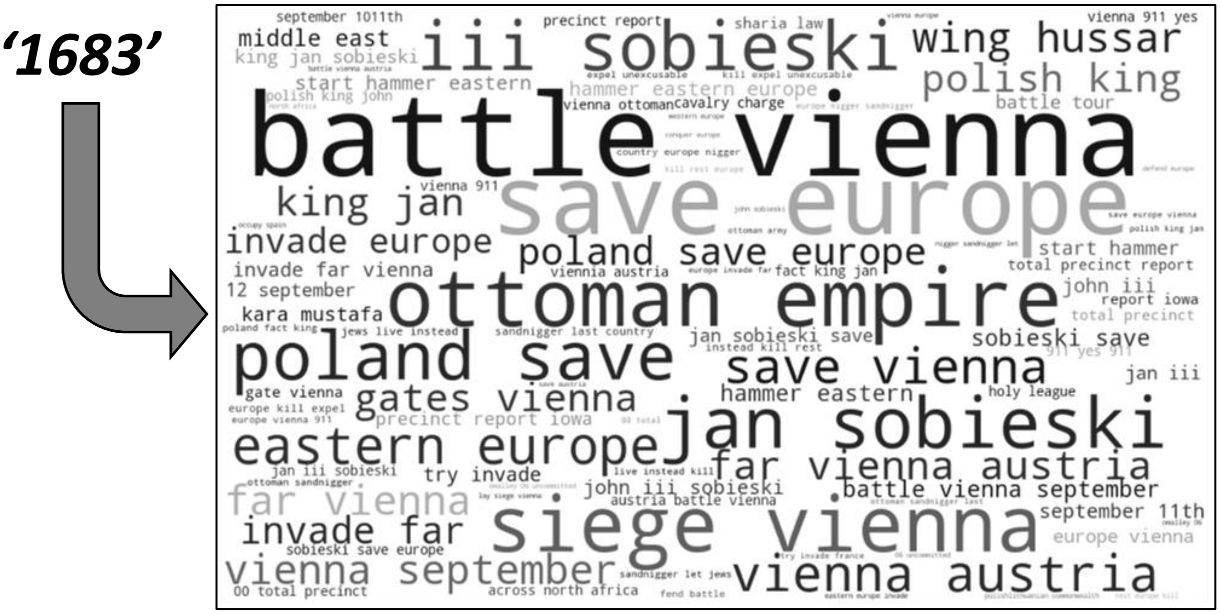
Beyond keywords: An NLP method for identifying corpus-specific historical terms in cultural datasets.

This contribution presents a technique for identifying and extracting information in natural language textual datasets specifically related to history and the past without relying on specific domain knowledge or a list of predetermined keywords. Whilst a common and straightforward method for retrieving information in computational and digital humanities, keyword searches are often limited to the researcher's preexisting knowledge of how cultural phenomena are terminologically and lexically represented in the material at hand. This problem is especially salient for historians and other scholars seeking to identify or reconstruct the historiographical consciousness of a specific community without imparting preconceived assumptions about the terms and features that make up such historical understanding in the source material. Instead, this proposed approach leverages natural language processing techniques to offer a novel and efficient way to retrieve text data relating to history and the past, and consequently provides a method for constructing a list of *corpus-specific* keywords queryable for further exploration.

For the purpose of testing this method, the entire dataset from the political section named "/pol/" (short for 'Politically Incorrect') of the online discussion forum 4chan.org was downloaded between 2013 and 2022. Comprising almost 100 GB of data and around 300 million user comments, this dataset serves as a relevant sample of a community discussing a variety of topics wherein references to history and the past are expected to be present, albeit not being the main focal point of communal interaction. In order to extract relevant data on /pol/-users' historical understanding, the Named Entity Recognition component of the SpaCy library was first used as an initial filter to recognize and extract comments containing references to *DATE* entities (i.e. any expressions that represent specific dates or periods of time mentioned in the text based on a model pre-trained on large amounts of annotated data). While this drastically reduced the size of the data (from 300 million to approximately 28 million comments), many false positives remained that were irrelevant to historical. To solve this, a series of regular expression patterns were created to identity comments mentioning specific temporal signifiers, i.e. references to past years, decades and centuries in both numerical and written form (as well as abbreviations such as 'the 50s'), yielding a filtered dataset of about 6.6 million comments, each annotated with a unique temporal signifier. Next, by grouping comments into discrete temporal-specific documents, Term Frequency-Inverse Document Frequency (TF-IDF) – a techniques for statistically evaluating a term's importance by considering its frequency both within a document and the entire corpus – was then used to identify the top-most significant terms relating to each temporal group, i.e. a collection of corpus-specific historical terms based on their relative association to select time periods (see the figure below for an illustrative example of this).

Finally, to further refine the list of historical terms, the neural network-based embedding technique of word2vec was used to expand the list of relevant keywords. Once a set of historical terms had been extracted based on their association to specific temporal signifiers using TF-IDF analysis, word2vec was used to find relevant historical terms *not* dependent on the use of temporal signifiers, but instead through finding words semantically similar to the TF-IDF terms based on their distributional proximity in the vector space. Using this method on the /pol/-dataset, a total of 1.302 unique historical terms were identified which, when used as queryable keywords on the initial dataset, produced a dataset of 13.144.987 comments (about 4.3% of the total comments) that effectively represent a subset of historical discourse within the larger community attuned to their specific cultural and terminological practices, without relying on any preexisting domain knowledge or predetermined list of keywords.



Caption: Example of how a TF-IDF analysis on comments grouped according to the temporal unit '1683' enables the retrieval of multiple corpus-specific historical terms useful for constructing a list of queryable keywords. This list can be further expanded to more accurately represent the semantic space of historical discourse in a given dataset, by representing them as vectors in a neural network and find semantically similar words based on distributional patterns.