Abstract. Heritage institutions are exploring new ways to open up their digital collections. In this context, the KB, national library of the Netherlands, has built a data-driven demonstration website based on historical newspapers. This website centers around a currently relevant topic due to the Covid-19 crisis: pandemics. A Toolbox with Notebooks and a sample data set is provided to support students and starting researchers. This paper describes the data selection process, the functionality of the website and corresponding Toolbox, as well as the initial reception.

1 Introduction

Most heritage institutions provide a straightforward way of searching through their digital collections. Users specify a search query and a list with snippets of results is returned. Such an interface works fine when users know what they are looking for, but is not well suited for exploration of a specific topic. An alternative to searching is browsing, but due to how digital collections are often displayed, browsing is not enticing for users [8,12].

Several studies suggest other ways of displaying digital collections, such as a topic-based search interface, in which results are clustered based on a certain topic [8]. Furthermore, institutions can decide to display their textual search results through visualisations [4]. This can, among others, be used to create summaries of the textual data. A benefit of visualisations is the ease with which users can extract information from them [4,9].

Since heritage institutions started digitizing their collections, the use of these collections has changed. A lot of research shifted from mainly manually researching collections to using computer science techniques [10]. In our experience, it often occurs that students or starting researchers are interested in working with digital collections, but lack the skills needed for analysing large amounts of textual data.

Recently, various heritage institutions started experimenting with new ways of exposing their data and providing tools for researchers querying their collections. This is often done by offering ‘Workbenches’ that contain Notebooks: open-source web applications consisting of code and documentation [2,3,11].
The KB, national library of the Netherlands, has closely followed these developments, and decided to set up an experimental data-driven demonstration website. The aim of this website is to examine new ways of displaying the historical newspaper collection of the KB. We created a topic-based website (in Dutch) (http://delpher_demo.kbresearch.nl/) [5], with a relevant topic for the year 2020 due to the Covid-19 crisis: pandemics. The website provides four pandemic related categories, from which users can choose to start their exploratory journey through related historical news articles. The results are summarised by using a timeline and various other visualisations. We deliberately used rather basic visualisations, to be able to provide entry-level demonstration code. A Toolbox with example data and Notebooks is provided for those who are interested in performing these analyses themselves (https://github.com/KBNLresearch/delpher_demo) [6].

In this paper we describe the data collection process and the functionality of the website and Toolbox, after which we conclude with the initial reception of the website.

2 Data collection

The data used for the website is collected from Delpher, a digital heritage archive that provides access to historical books, periodicals and newspapers [7]. For this project, we used the historical newspaper collection.

We started with the selection of pandemic related words by analysing a collection of 295,612 European news articles about Covid-19. These articles were retrieved through the Aylien Coronavirus News dataset [1]. Out of the 50 most commonly used words from these articles, we extracted pandemic related words. This led to the following set: corona, pandemic, outbreak, infection, spread, virus, disease and quarantine. The word ‘corona’ was excluded because there were less than ten articles about this disease in the Delpher news archive. Instead, we added the words ‘flu’ and ‘influenza’ to the set. During the development of the website (November 2020), vaccines and immunity were a hot topic in the Netherlands. Therefore the words ‘vaccin’ and ‘immunity’ were included. Finally, we decided to add ‘Spanish flu’, since this was a noteworthy pandemic in history. The end result was a set of twelve keywords, which we translated to Dutch.

We chose four of these keywords to use as main categories around which the site was built: ‘pandemic’, ‘outbreak’, ‘immunity’ and ‘Spanish flu’. We collected all articles from Delpher in which at least one of these four words was present. Then, we prepared the data for further use on the website, and enriched it by adding metadata about which categories and keywords belong to which article. The remaining keywords where used as sub selections for more in-depth analyses on the website.
3 Tool Description

The homepage presents an introduction and four buttons, each containing a category. The buttons navigate to a page dedicated to this category. Each category page shows a timeline. This timeline shows all the years in which at least one news article from this category was found. Furthermore, the page displays some descriptive analyses. The number of total articles is shown and the content of all articles is summarized in a word cloud. This word cloud shows the 20 most common words from these articles based on their frequency. We only altered the results by removing stop words. The font size and frequency of a word are correlated, which means that a word appears bigger when the frequency is higher (see figure 2).

The page also displays a line chart showing the number of found articles per year. This can be switched to a bar chart that shows the number of articles per keyword. The user can scroll through the timeline to discover the various years in which articles were found. By selecting a year, the before mentioned descriptive analyses adapt to the selection. The page is also extended with buttons for keywords that were found in the articles corresponding to this selection. Furthermore, a bar chart with the number of articles per keyword is shown (see figure 1).

When a further selection is made by choosing a keyword, a word tree is displayed. The word tree shows the relationship between keywords that are co-occurring in the articles. The bigger the font size, the more frequent the words occur together (see figure 2). The word tree is set as default, but the user can switch back to the bar chart.
Fig. 2. Examples of a word cloud (left) and a word tree (right) for the category 'pandemic' for the year 1927 and with 'ziekte' (disease) as selected keyword.

The page contains a link to the original scans of the selected articles, to give users the opportunity to explore them on Delpher. Clicking on the words in the word cloud also navigates to Delpher. In that case, the Delpher result is further narrowed down to not only the category, but also to the corresponding word in the word cloud and, if applicable, the earlier selected keyword.

There is an option to download a file with metadata and a 'bag of words' of each article from the current selection. Finally, a link to the Toolbox is provided.

4 Toolbox: example data and Jupyter Notebooks

We provided a Toolbox for students or starting researchers to help them getting started with analysing textual data themselves. The Toolbox is a Github repository containing Jupyter Notebooks and example data. The Notebooks guide users through basic preparation and analysis techniques. Users can also download data sets from the demonstration website and use them for further analysis.

The complete code of the demonstration website was also made available. This code can be used as a starting point for creating other topic-based websites or to make an improved version of our website.

5 Conclusion

The demonstration website was promoted through several social media channels. The feedback we received was positive. Users liked the way they where able to explore topics. They particularly liked the timelines and word clouds, and the fact that the website was showcasing a currently relevant topic. Multiple request where made for more information about how to replicate the visualisations, after which we showed them the Toolbox. Thus, a recommendation for further development would be to give the Toolbox a more prominent place on the website. To determine the actual added value of this website, a more comprehensive evaluation is desirable in the near future.
A Covid-19 Inspired Data-driven Approach to Explore Historical Newspapers

References

1. Aylien: Free news intelligence dataset (2020), https://aylien.com/resources/datasets/coronavirus-dataset
2. BVMC. labs: Glam jupyter notebooks (2020), http://data.cervantesvirtual.com/blog/notebooks/
3. Clariah media suite: Jupyter notebooks (2020), https://mediasuite.clariah.nl/documentation/howtos/jupyter-notebooks
4. Conner, C., Samuel, J., Kretinin, A., Samuel, Y., Nadeau, L.: A picture for the words! textual visualization in big data analytics. arXiv preprint arXiv:2005.07849 (2020). https://doi.org/10.13140/RG.2.2.25351.83360
5. Cuper, M.: Met delpher de diepte in: pandemieën (2020), http://delpher_demo.kb.research.nl/
6. Cuper, M.: Toolbox delpher demo site (2020), https://github.com/KBNLresearch/delpher_demo
7. Delpher: https://www.delpher.nl/
8. Grant, C.E., George, C.P., Kanjial, V., Nirkhiwale, S., Wilson, J.N., Wang, D.Z.: A topic-based search, visualization, and exploration system. Proceedings of the 28th International FLAIRS Conference (2015)
9. Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: Text analytics based on word clouds. In: 2014 47th Hawaii International Conference on System Sciences. pp. 1833–1842 (2014). https://doi.org/10.1109/HICSS.2014.231
10. Janicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In: Borgo, R., Ganovelli, F., Viola, I. (eds.) Eurographics Conference on Visualization (EuroVis) - STARs. The Eurographics Association (2015). https://doi.org/10.2312/eurovisstar.20151113
11. Sherratt, T.: Glam workbench (2020), https://glam-workbench.net/
12. Whitelaw, M.: Generous interfaces for digital cultural collections. Digital Humanities Quarterly 9 (2015)