Exploring text mining techniques to structure a digitized catalogue

Karen Goes [1] , Sara Veldhoen [2] & Steven Claeyssens [2]

[1] VU Amsterdam; [2] KB, National Library of the Netherlands

This research aims to obtain structured data from digitized catalogues, the national bibliography. The national bibliography lists the books, journal titles and maps published in the Netherlands since 1846. It contains details about every publication, such as the author, the publisher, the number of pages and the retail price. The national bibliography reflects the changing culture and interests of a nation over time, which makes it an invaluable resource for the history of the Netherlands. Between 1846 and 1994 the national bibliography was published in instalments. The Koninklijke Bibliotheek (KB) [1], the National Library of the Netherlands, has digitized these catalogue books and published them online as PDFs, JPGs and plain text (OCR) on Delpher [2] and DBNL [3], the Digital Library of Dutch Literature.

However, this data is unstructured, uncorrected and cannot be processed computationally. To structure and retrieve information from the catalogues different text mining techniques are used. Two main issues arise when working with the uncorrected OCR output.

The first issue concerns the layout of the OCR output: the majority of the catalogues has two columns per page, however, the OCR processing does not always pick up on this. Therefore, an analysis of the OCR is performed to determine the quality of each catalogue and whether or not it can be automatically processed. From a total of 31 catalogues a selection is made based on the OCR analysis for further processing. This resulted in only 9 usable catalogues for this research.

Secondly, the boundaries of bibliographical entries in the catalogues are clearly marked: the first word is printed in boldface and the continuation lines are indented. These properties are lost in the OCR, so we cannot rely on them when processing the plain text. The lines in the OCR output are grouped together to form a single line per bibliographical entry in the catalogue. This is done automatically utilizing the alphabetic nature of the catalogues. Each beginning of a new letter in the catalogues is marked manually in the OCR output to give the automatic process a guideline as to which lines might be the start of a new bibliographical entry.

Each bibliographical entry contains information such as the author, title, publisher and retail price. In the next stage of this research, this information is extracted from the bibliographical entries using different text mining techniques namely, regular expressions, a probabilistic context-free grammar, and named entity recognition.

---

[1] www.kb.nl
[2] www.delpher.nl
[3] www.dbnl.org

The first technique relies on splitting the bibliographical entries on specific separators and regular expressions. The specific separators are characters such as a comma, semicolon or dash, that consistently divide different pieces of metadata within the entry. This is used for metadata such as the author and title, since their length and format will be different across the different entries. Regular expressions are used for metadata that has a consistent format such as the retail price.

Another technique that is used to extract the metadata from the bibliographical entries is a probabilistic context-free grammar. This technique is only used for two of the nine catalogues, since it is very inefficient and time-consuming to create. This technique also uses the separating characters, as well as probabilities attached to them. An advantage of this is that OCR error can be taken into account, for example by giving a comma a 95% chance of being a comma and a 5% chance of being a full stop.

Named entity recognition is used as a third technique. This technique is only used to extract publishers from the entries, since this piece of metadata is expected to form a clear entity within the entry. The Dutch Spacy model [4] is used to perform the named entity recognition.

The extracted information is improved using external knowledge such as a list of Dutch surnames for authors and Dutch and Belgian city names. Named entity recognition is also used as external knowledge to fill in 'gaps' in the extracted data where the afore mentioned techniques were unable to extract a city or publisher from a bibliographical entry.

To evaluate the extraction process, both an automatic and manual method are applied. For the three most recent catalogues, the KB has provided data that can be used as reference data to evaluate the extracted information. For the remaining catalogues one hundred gold bibliographical entries are manually annotated to create a gold standard.

The evaluation results show that the rule-based system with regular expressions is the best technique to extract metadata from the bibliographical entries. This technique has the lowest average percentage of 'no match' matches across all volumes and metadata, which indicates that it extracts the correct metadata from the entries. Additionally, the use of external knowledge improves the performance for this technique, whereas for the others it worsens them.

To account for OCR errors fuzzy matching is applied, which relies on the Levenshtein distance measure, which deems a value correct if it corresponds with the gold value for at least 80%. The fuzzy matching proofs to be a valuable matching type to catch OCR errors that would otherwise have caused a 'no match'.

The extracted data is transformed to a standard format for bibliographic metadata, which makes it structured and searchable data. The data can also be linked and matched with for instance the national catalogue of publications that are kept in Dutch libraries to find out which part of Dutch publishing activity is lost or which part is already available in a digital form.

[4] https://spacy.io/models/nl