



Exploring text mining techniques to structure a digitised catalogue



Karen Goes¹, Sara Veldhoen² & Steven Claeyssens²

kwmgoes@gmail.com, {sara.veldhoen; steven.claeyssens}@kb.nl

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

Introduction

This research is conducted as part of an internship at the National Library of The Netherlands. The National Library collects all books that are published in the Netherlands, this has actively been done since

The Brinkman Catalogue of Books is part of the collection and the subject of this research. The catalogue lists all the books published in the Netherlands within a specific time frame. It is available in monthly and yearly volumes, as well as volumes covering multiple years. These volumes are part of the national bibliography, and are currently only available as unstructured data, on paper and partially digitised.

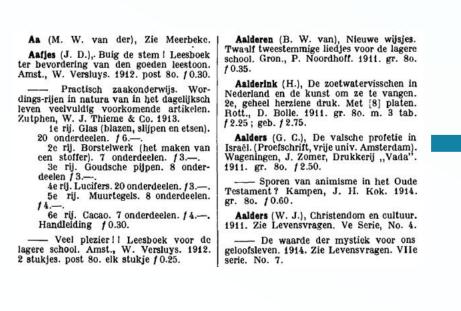
This research uses the digitised catalogue volumes to extract the data from them and turn them into structured data, that is machinereadable and searchable. To accomplish this the use of multiple text mining techniques is explored to determine the best method for this

Using the structured data the National Library can make a comparison with their own collection to identify books that are missing from their collection, especially from before 1974.



Materials

- 31 Brinkman catalogue volumes, originally published between 1833 and 1980. Two material types are available per catalogue volume: PDF scan
- OCR output



elken zin. Breda, van Gutick en Hermans. 1836. kl. 8°.

— Geschied- en aardrijkskundige beschrijving van het koningstijk der Nederlanden en het Groot-hertogdom Luxemburg. Gorinch. J. Noorduyn en Zn. 1841. gr 8° f 4,80.

— Beknopte geschied- en aardrijkskundige beschrijving van het koningrijk der Nederl. en het Groot-hertogdom Luxemburg; benevens een kort overzigt van Nederl. bezittingen buiten Europa. Gorinch. J. Noorduyn en Zn. 1844. 3° verm. dr. 1848. kl. 8°. f 0,17½.

- China en zijne bewoners geschetst voor jonge lieden. Amst. G. J. A. Beijerinck

afgegaan van een beknopt overzigt van de vestiging en uitbreiding der magt van Nederl. aldaar, Amst. J. F. Schleüer. 1845—49. rov. 8°. met pl. en kaarten. 1°—19° afl. Elke afl. f 0.50.

/lanes (J. D.),. Buig de stem I Leesboek le rij. Glas (blazon, slijpen en etsen). 2e rij. Borstelwerk (het maken van een stoffer). 7 onderdeelen. f 3.-. 6e rij. Cacao. 7 onderdeelen. f 4.-.

Aalderen (B. W. van), Nieuwe w(jsjes Aa (M. W. van der), Zie Meerbeke. Aalderink (H.), De zoetwatervisschen in Nederland en de kunst om ze te vangen. 1911. gr. 80. f 2.50. deelen f 3.-. Sporen van animisme in het Oude 4e rij. Lucifers. 20 onderdeelen. 13.-. Testament ? Kampen, J. H. Kok. 1914.

Gulick en Hermans. 1833. 12°. met kaartje. • f 2,60. Herinneringen uit het gebied der geschiedenis betrekkelijk de • Nederlanden. 4mst. Nieuwe herinneringen uit het gebied der gesc• hieden• is be• trekkelijk • de Nederlanden. .Amst. J. C. van I-esteren. 1837. gr. 8°. f 3,60. Lees- en vertaalboekje voor de hoogste klassen der.Fransche seholen, met eene woordenlijst, waarin de vertaling der daarin voorkomende woorden opgegeven wordt. ,dmst. Schalekamp , van de Grampel en fakker. 1836. kl. 8°. . • f 0,60. Zamenspraken in de Nederd. en Fransche talen, met woordelijke overzetting van elken zin. Breda, van Gulick en Germans. 1836. kI. 8°. . f 0,25. Geschied- en aardrijkskundige beschrijving van het koningrijk der • Nederlanden er het Groot-hertogdom Luxemburg. Gorinch. J. Noorduyn en Zn. 1841. gr 8° f 4.80. Beknopte geschied- en aardrijkskundige beschrijving van het koningrijk der Nederl. en het Groot-hertogdom Luxemburg; benevens een kort overzigt van Nederl. bezittinge buiten Europa. Gorinch. J. Noorduyn en Zn. 1844. 3e verin. dr. 1848. kl.8°. f0,17I China en zijne bewoners geschetst voor jonge lieden. Amst. G. J. A. Beijerinck. 1845. kl. 8". met houtsneepl. f 1,80.

Geschiedkundige beschrijving • van • de sta• ll Breda • en oms• treken. • Gorinchem, J

foorduyn en Zoon. 1845. gr. 8°. f 3,90. Nederlands Oost-Indie, of beschrijving der Nederl.~bezittingen in Oos• t-Indie, voor

afgegaan van een beknopt overzigt van de vestiging en uitbreiding der magt van Nederl. aldaar. dmst. J. F. Schle~er. 1845-49. roy. 8°. met pl. en kaarten. 1e-19e afl. Elke afl. f 0,50.

Rott., D. Bolle. 1911. gr. 80. m. 3 tab. f 2.25; gob. f 2.75. Aalders (G. C.), De valsche profetie in Israel. (Proefschrift, vrije univ. Amsterdam) Wageningen, J. Zomer, Drukkerij ,, Vada". 5e rij. Muurtegels. 8 onderdeelen. gr. 80. f 0.60. Handleiding f 0.30. Veel plezier I I Leesboek voor de lagere school. Amst., W. Versluys. 1912. 2 stukjes. post 80. elk stukje 1 0.25. 1911. Zie Levensvragen. Ve Serie, No. 4. - De waarde der mystiek voor ons geloofsleven. 1914. Zie Levensvragen. Vile serie. No. 7.

Methodology

Deal with non-alphabetical order

Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude; geheel opnieuw bew. door W. N. van der Sluys; geill. door G. van

Alkmaar : Kluitman, 1979. - 158 p : ill. ; 18 cm. - (Kluitman

Alkmaar Kluitman, 1979. - 158 p : ill. ; 18 cm. - (Kluitman jeugdserie ; J 1081) (Pietje Bell serie) ISBN 90-206-1081-3 : f. 3.25 8030302 Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude; geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van Straaten. - 22e dr. -

jeugdserie; J 1081) (Pietje Bell serie) ISBN 90-206-1081-3: f. 3.25 Abkoude, Chr. van - Pietje Bell is weer aan de gang / door Chr. van Abkoude. - 24e dr. - Alkmaar Kluitman, 1979. - 160 p.: ill.; 18 cm. - (Kluitman jeugdserie ; J 1042) (Pietje Bell serie) ISBN 90-206-1042-2 : f. 3.25 8030299

geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van Straaten. - 22e dr. - Alkmaar Kluitman, 1979. - 158 p : ill. ; 18 cm. - (Kluitman jeugdserie ; J 1081) (Pietje Bell serie) ISBN Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude; geheel opnieuw bew. door W. N. van der Sluys ; geill. door G. van Straaten. - 22e dr. - Alkmaar : Kluitman, 1979. - 158 p : ill. ; 18

cm. - (Kluitman jeugdserie ; J 1081) (Pietje Bell serie) ISBN

Abkoude, Chr. van - Pietje Bell in Amerika / door Chr. van Abkoude;

Abkoude, Chr. van - Pietje Bell is weer aan de gang / door Chr. van Abkoude. - 24e dr. - Alkmaar Kluitman, 1979. - 160 p. : ill. ; 18 cm. (Kluitman jeugdserie ; J 1042) (Pietje Bell serie) ISBN 90-206-1042-2 : f. 3.25 8030299

Replace dashes

Aabye, Karen: Amazone, machtige stroom. (Zuid-Holl. U.M.). fl. 6.-; geb. fl. 7.90. - In prijs ver-hoogd: fl. 6.45; geb. fl. 8.50. -- Martine. (Zuid-Holl. U.M.). Geb. fl. 11.50. - In prijs verlaagd:

- Vrouw, ga de zon tegemoet. [Kvinde, gaa mod solen]. Het avontuurlijke leven van Marianne Holler. Vert. [uit bet Deens] van Cath. van Eysden. 's-Gra-venh., Zuid-Holl. U.M. [1965]. 24 x 16. 314 blz. [Cultuurserie]. Geb. fl. 14.90.

Aabye, Karen: Amazone, machtige stroom. (Zuid-Holl. U.M.). fl. 6.-; geb. fl. 7.90. - In prijs ver-hoogd: fl. 6.45; geb. fl. 8.50. Aabye, Karen: Martine. (Zuid-Holl. U.M.). Geb. fl. 11.50. - In prijs

verlaagd: geb. fl. 6.90.

Aabye, Karen: Vrouw, ga de zon tegemoet. [Kvinde, gaa mod solen]. Het avontuurlijke leven van Marianne Holler. Vert. [uit bet Deens] van Cath. van Eysden. 's-Gra-venh., Zuid-Holl. U.M. [1965]. 24 x 16. 314 blz. [Cultuurserie]. Geb. fl. 14.90.

Extracted meta data

1. Generate letter

2. Group lines on first

3. Loop to deal with

and incorrect

4. Replace dashes

Integrate

Gazetteers

NER

external knowledge

non-alphabetical

alphabetical orders

sections

| Author | Title | City | Publisher | Year | Pages | Size | ISBN | Price | Product number |
|-------------------|-----------------------------------------------------|------------|----------------------|-----------|---------|------------|------------|----------|----------------|
| Unknown | A Paris | Zutphen | Thieme | 1978 | 25 | 30 cm | 9003285004 | f. 5.20 | 7905126 |
| Unknown | A Paris | Zutphen | Thieme | 1979 | 25 | 30 cm | 9003285004 | f. 5.20 | 7938217 |
| Unknown | Unknown | A + | aardrijkskunde. | Unknown | 15 | 22 cm | 902571112X | f. 3.25 | 7924032 |
| Aa, A. J. van der | Aardrijkskundig woordenboek der Nederlanden | Zaltbommel | Europese Bibliotheek | . Unknown | 136 | 18 cm | N/A | f. 55 | 7928260 |
| Aafjes, Bertus | Kleine Isar, de vierde koning : halleluja, de wonde | Amsterdam | Meulenhoff | 1979 | 217 | 14 x 21 cm | 9029012323 | f. 25 | 7956335 |
| Aafjes, Bertus | Het koningsgraf : honderd en een sonnetten | Bussum | Agathon | 1979 | 119 | 19 cm | 9026956436 | f. 9.90 | 7931389 |
| Aafjes, Bertus | Een ladder tegen een wolk | Amsterdam | Querido | 1979 | 158 | 19 cm | 902149468X | f. 6.20 | 7926190 |
| Aafjes, Bertus | Limburg, dierbaar oord | Amsterdam | Meulenhoff | 1979 | Unknown | Unknown | 9029007915 | f. 26.50 | 7927311 |
| - | | | | | | | | | |

Evaluation

• Manually created evaluation data

For a total of 100 bibliographical entries the meta data is manually copied down to create evaluation data. Entries from random pages in a volume are copied until the 100 entries are reached. This is done for all volumes from 1850 to 1971.

printed catalogues

OCR output of

printed catalogues

Alphabetical

book entries

Structured

PICA+ data

Selection of

volumes

Extract

meta data

Splitting and regular expressions

PCFG

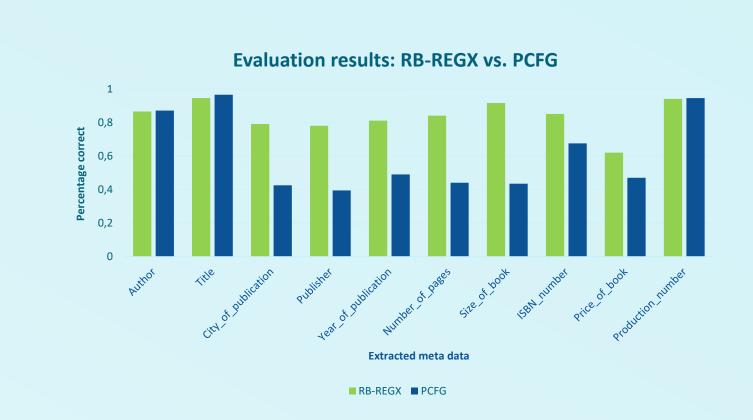
Guidelines are created to ensure that the process of copying down the meta data is done in the same manner across all the volumes and entries.

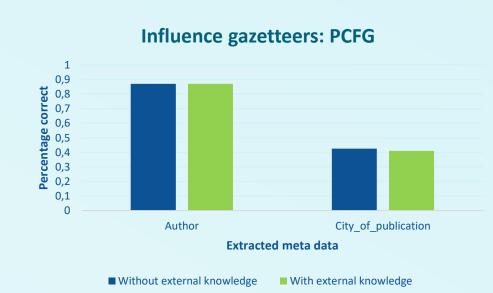
② Digitally available evaluation data

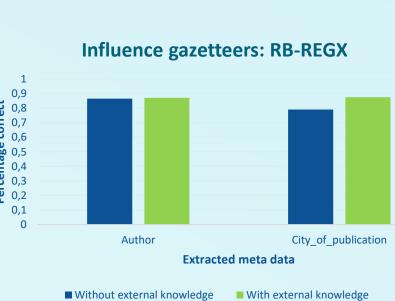
For the volumes from 1975 to 1980, the National Library has provided data that can be used as evaluation data. This data can be linked to the extracted data using the Brinkman production number, which can be found in both the catalogues and the library data.

Again, 100 bibliographical entries are used for the evaluation taken at random. With as selection requirement a Brinkman production number that can be linked to the library data, which is already digitally available.

Evaluation results







Conclusion

The evaluation results show that the RB-REGX system is the best technique as it is able to extract meta data from all the entries and with a high accuracy.

Across all the volumes the lowest extraction percentage is for the ISBN

The RB-REGX technique is more efficient to implement, and it performs better than the PCFG.

Future recommendations

- Redo the OCR process to minimise the loss of volumes and simplify the formation of bibliographical entries.
- Use the output of the current research as annotated data for a machine learning system.
- Process other parts of the catalogue such as the bibliographical entries sorted by topic.