Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports

Brandsen, A.

Digging in Documents

Digging in Documents

Using Text Mining to Access the Hidden Knowledge in Dutch Archaeological Excavation Reports

PROEFSCHRIFT

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op dinsdag 15 februari 2021

klokke 11.15 uur

door

Alex Brandsen

Promotoren:	Dr. K. Lambers & Prof.dr. J.C.A. Kolen
Copromotor:	Dr. S. Verberne
Promotiecommissie:	Prof.dr. A.P.J. van den Bosch, Prof.dr. A.N. Brysbaert, Dr. M.G.J. van Erp, Prof.dr. D.R. Fontijn, Prof.dr.ir. W. Kraaij, Prof.dr. J.D. Richards & Prof.dr. H.G.D.G. de Weerdt

Printed by Gildeprint

Lay-out & cover design by Alex Brandsen
AGNES 'spade in book' logo designed by Mike Smethurst
Other vector elements used on cover provided by Freepik.com

All figures in this book are by the author

Acknowledgements

First of all, I would like to thank my grandmother Sietske Haker-Ples, who already knew I was going to be a doctor when I was five years old, and encouraged my curiosity for antiquities by taking me to museums. Also my parents, Silvia Pels and Piet Brandsen, for their never-ending support in my academic endeavours. And to Kayleigh Hines, for putting up with me being busy, stressed and/or absent minded during this research, for supporting me wholeheartedly, and even leaving her home and moving to the Netherlands with me. I would also like to thank Inge van Stokkom, for sending me the job ad for this PhD. Without her, I would have never known about this opportunity and would not be here.

Special thanks go out to my supervision team and promoters: Karsten Lambers, Suzan Verberne, Milco Wansleeben, David Fontijn and Jan Kolen. Each of you helped me tremendously with this research, and your guidance was invaluable. I could not have wished for a better support team.

I am grateful to Martin Koole and Femke Lippok for their excellent collaborations on the research described in Chapter 4 and 8 respectively. And to the entire focus group, for testing and providing feedback: Mette Langbroek, Femke Lippok, Arjan Louwen, Stijn van As, Rik Feiken and Ronald Visser. I would also like to thank Mike Smethurst for designing the AGNES logo.

I would like to thank all of my colleagues at the Faculty of Archaeology and the Data Science Research Programme, specifically Wouter Verschoof-van der Vaart, Anne Dirkson, Hugo de Vos, Daniela Gawehns and Gineke Wiggers, for exchanging ideas, helping me solve problems and inspiring me, as well as the great office banter.

Finally, I would like to thank the various institutes that supported this research: the Faculty of Archaeology and the Data Science Research Programme for funding this PhD position, the Leiden Institute for Advanced Computer Science for access to a web server and their computing cluster, the Leiden ALICE computer cluster for providing the computing power needed for Chapter 7, and the Leiden University Centre for Digital Humanities for funding part of the research described in Chapter 3.

Abstract

The archaeology domain produces large amounts of texts, too much to effectively read or manually search through for research. To alleviate this problem, we created a search system (called AGNES), which combines full text search with entity and geographical search. We first created a manually labelled data set to train a Named Entity Recognition model, which is used to extract entities from text. We also did a user requirement study, and usability evaluation on the system, to make sure it is suitable for archaeological research. In a case study on Early Medieval cremations, we show that using AGNES leads to a knowledge increase when compared to the knowledge of experts, gathered using previously available search engines. This shows that this kind of intelligent search system can help with literature research, find more relevant data, and lead to a better understanding of the past.

Samenvatting (Dutch abstract)

Archeologen produceren grote hoeveelheden teksten, te veel om effectief te kunnen lezen of handmatig te doorzoeken voor onderzoek. Om dit probleem op te lossen hebben we een zoeksysteem ontwikkeld (AGNES), dat zoeken in de volledige tekst van de docmenten combineert met zoeken op entiteiten en zoeken op een kaart. We hebben eerst een handmatig gelabelde dataset gemaakt om een *Named Entity Recognition* model te trainen, dat gebruikt wordt om entiteiten uit tekst te extraheren. We hebben ook een studie gedaan naar de gebruikerseisen en een evaluatie van de *usability* van het systeem, om er zeker van te zijn dat het geschikt is voor archeologisch onderzoek. In een case studie over Vroeg-Middeleeuwse crematies, laten we zien dat het gebruik van AGNES leidt tot een toename van kennis in vergelijking met de kennis van experts, verzameld met behulp van eerder beschikbare zoekmachines. Dit toont aan dat dit soort intelligente zoeksystemen kunnen helpen bij literatuuronderzoek, meer relevante gegevens kunnen vinden, en uiteindelijk kunnen leiden tot een beter beeld van het verleden.

Contents

List of Figures

List of Tables