



Universiteit
Leiden
The Netherlands

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

Citation

Brandsen, A. (2022, February 15). *Digging in documents: using text mining to access the hidden knowledge in Dutch archaeological excavation reports*. Retrieved from <https://hdl.handle.net/1887/3274287>

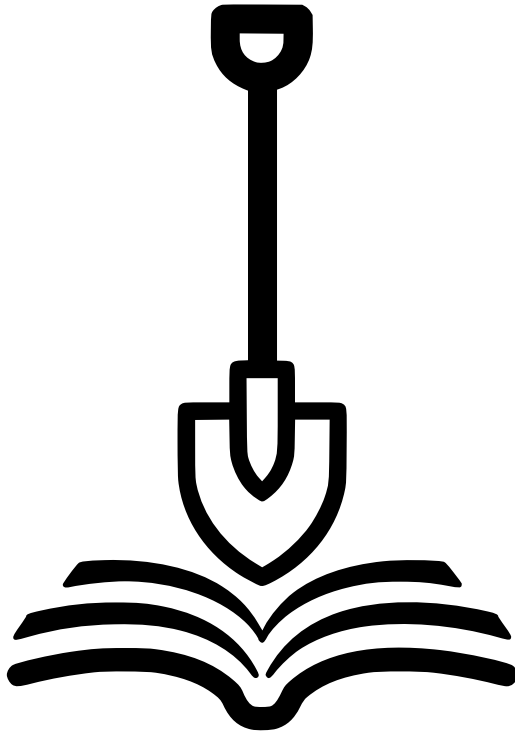
Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3274287>

Note: To cite this publication please use the final published version (if applicable).

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 Weerd

© 2022 Alex Brandsen

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 zoekstelsel ontwikkeld (AGNES), dat zoeken in de volledige tekst van de documenten 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 stelsel, 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

Acknowledgements **I**

Abstract **II**

 Samenvatting (Dutch abstract) II

Contents **III**

List of Figures **IX**

List of Tables **XI**

1 Introduction **1**

 1.1 Research motivation 2

 1.2 Research questions 3

 1.3 Research Methodology 4

 1.3.1 Agile Development Principles 4

 1.3.2 Machine Learning Assessment with Labelled Data 6

 1.3.3 System Evaluation by Focus Group 6

 1.3.4 System Evaluation by Case Study 6

 1.4 Contributions 7

 1.4.1 Software and Data 7

 1.4.2 Scientific Contributions 8

 1.5 Dissertation outline 9

2	Background	11
2.1	Digital Archaeology and Digital Humanities	12
2.1.1	Big Data in Archaeology	14
2.2	Data	16
2.2.1	Malta Convention	16
2.2.2	Grey Literature	17
2.2.3	Repositories	19
2.2.4	The Importance of Archaeological Reports	19
2.2.5	The FAIR Principles	20
2.2.6	Document Properties	21
2.3	Introduction of NLP and Information Retrieval Concepts	22
2.3.1	Information Retrieval	22
2.3.2	Text Mining and Machine Learning	24
2.3.3	Named Entity Recognition	25
2.3.4	Evaluation Metrics	27
2.4	Previous Research	29
2.5	Resources	32
2.5.1	The DANS Corpus	32
2.5.2	Computing Power	32
2.5.3	Ontologies	33
2.5.4	Gold Standard	34
3	Data Set	35
3.1	Introduction	36
3.2	Related Work	37
3.3	Data set Collection	38
3.4	Annotation Setup	39
3.4.1	Annotation Guidelines	39
3.4.2	Entity Types	40
3.4.3	Annotation Process	40
3.5	Annotated Corpus Statistics and Results	40
3.5.1	Inter Annotator Agreement	41
3.5.2	New NER Results	42
3.6	Conclusions	45
3.7	Acknowledgements	45

4	Text Classification	47
4.1	Introduction	48
4.2	Related work	50
4.2.1	Text mining in the Archaeological Domain	50
4.2.2	Multi-label Text Classification	51
4.3	Data	52
4.3.1	Source Data	53
4.3.2	ABR Ontology	53
4.3.3	Definition of Categories	54
4.3.4	Obtaining the document labels from the data	54
4.3.5	Exploration of the Extracted Labels	55
4.3.6	Pre-processing the metadata	56
4.4	Methods	58
4.4.1	Document Pre-processing	58
4.4.2	Document Filtering	59
4.4.3	Balancing the Training Set	59
4.4.4	Construction of a Manually Labelled Reference Set	61
4.4.5	Classification Methods	62
4.4.6	Selection round	63
4.5	Results	64
4.5.1	Selection Round	64
4.5.2	Pre-processing Optimisation	66
4.5.3	Best Methods per Category	67
4.6	Conclusion	69
4.6.1	Future Work	73
5	User Requirement Solicitation	75
5.1	Introduction	76
5.2	Prior work	78
5.3	Introducing AGNES	79
5.3.1	Named Entity Recognition	80
5.3.2	Indexing & front end	82
5.4	User study	83
5.4.1	Definition of target audience	83
5.4.2	Focus Group	85
5.4.3	Prototype for discussion	86
5.4.4	Workshops	86
5.4.5	Results	87
5.5	Future Work	88

5.6	Conclusions	90
6	Usability Evaluation	91
6.1	Introduction	92
6.2	Background	94
6.2.1	Access to archaeological data	94
6.2.2	Feedback on existing systems from our user group	95
6.2.3	Related work in usability studies	95
6.3	AGNES	96
6.4	User Study Setup	99
6.4.1	Workshops in the Archaeological Grey-literature Named Entity Search (AGNES) project	99
6.4.2	Compilation of the focus group	99
6.4.3	Design and procedure	100
6.5	Analysis and Results	101
6.5.1	Information Needs	102
6.5.2	Query Strategies and Effectiveness	103
6.5.3	Evaluation and User Satisfaction	104
6.6	Discussion	107
6.7	Conclusions	109
6.7.1	Future Work	111
7	Using BERT for Named Entity Recognition	113
7.1	Introduction	114
7.2	Related Work	118
7.2.1	Knowledge-driven and Data-driven NER	118
7.2.2	NER for Document Retrieval	119
7.2.3	IR and NER in Archaeology	119
7.2.4	Language- and Domain-specific BERT Models	120
7.3	Data	121
7.3.1	Pre-processing	122
7.4	Methods	123
7.4.1	Baselines	123
7.4.2	Fine-tuning BERT for Dutch Archaeology and NER	123
7.4.3	Ensemble Methods	124
7.4.4	Entity-driven Document Search	125
7.5	Results	126
7.5.1	Model Stability and Quality	126
7.5.2	Ensembles	128

7.5.3	Analysis of the Retrieval Collection	130
7.6	Discussion	132
7.6.1	Error Analysis	132
7.6.2	Tokenisation Issues	136
7.7	Conclusion	137
8	Case Study	139
8.1	Introduction	140
8.2	Methods	143
8.2.1	AGNES	144
8.2.2	Search Process for our Case Study	145
8.2.3	Evaluation: Comparison to Existing Knowledge	146
8.3	Results	146
8.3.1	Information Needs and Queries	146
8.3.2	Retrieved Documents	148
8.3.3	Comparison	149
8.4	Discussion	150
8.4.1	Archaeological Significance	150
8.4.2	Potential of AGNES for Archaeological Research	153
8.4.3	Future Work	154
8.5	Conclusions	155
9	Discussion	157
9.1	Development-led Archaeology and the Role of AGNES	158
9.2	Catching the By-Catch	159
9.3	Synthesising Research	161
9.4	MEAN & FAIR Data	162
9.5	Taming Big Data	163
9.6	The Problem with Complexity	164
9.7	Evaluation Metrics	165
9.8	Conclusion	166
9.8.1	Answers to Research Questions	166
9.8.2	Answer to Problem Statement	169
9.9	Future Research	170
9.9.1	EXALT	170
9.9.2	Long Term Ideas	172
9.9.3	Recommendations	173
	Bibliography	177

Appendices	205
A Category frequencies	207
B Filter list	209
C Category frequencies test set	211
D Curriculum Vitae	213
Glossary	217

List of Figures

3.1	CRF F1 score for each entity type per 1/10th chunk of data added to the training set.	43
3.2	Confusion matrix showing percentages for each combination of predicted and annotated entity type.	44
4.1	The number of documents and available metadata values.	56
4.2	An overview of the frequencies of the eight time period categories. X axis labels as per table 4.2a.	60
4.3	An overview of the frequencies of the eleven site type categories. X axis labels as per table 4.2b.	60
4.4	An overview of the frequencies of the eight categories for time period classification, as captured within our reference set.	61
4.5	An overview of the frequencies of the eleven categories for site type classification, as captured within our reference set.	61
4.6	Plot of the frequency of time period labels and the associated F1 score for that label. A trend line has been added to illustrate the correlation (Pearson's $r = 0.56$).	71
4.7	Plot of the frequency of subject labels and the associated F1 score for that label. A trend line has been added to illustrate the correlation (Pearson's $r = 0.28$).	71
5.1	AGNES Logo	80
5.2	AGNES Workflow	84

5.3	2D representation of clustered word embeddings.	89
6.1	Screenshot of AGNES version 0.3. Pictured here is a query for ‘artefact:axe AND (period:neolithic OR period:mesolithic) AND fulltext:burnt’, with the results on a map and in a list underneath (with snippets). On the left we can see the facets, used to filter results on period, type of document, and subject.	98
6.2	Line plot showing the number of new issues raised for each user . .	105
6.3	Word cloud of all feedback given, both positive and negative (translated from Dutch to English, ‘ahn’ is the height model of the Netherlands)	106
6.4	Line plot showing for each user, how much time they spent formulating one element of a query, for each new query they attempted. The black line is the average over all the users.	108
7.1	Query interface showing query for “Artefact: urn AND Context: cremation AND startdate < -2000 AND enddate > -800 AND fulltext: upside down”. Interface and query translated to English for the readers’ convenience.	126
7.2	Distribution of F1 scores over ten runs with different seeds, for each of the 5 folds (50 runs per model). The zero scores for multiBERT are runs where the model failed to learn.	127
7.3	Graph showing for each year in each detected time period, how often it occurs in our data set, labelled by ArcheoBERTje. For clarity, years before 10,000 BCE are not included. Major time periods are denoted with dashed lines.	131
7.4	Confusion matrix between true labels and ArcheoBERTje predictions.	133
8.1	Screenshot of the AGNES query interface (translated from Dutch)	144
8.2	Map of The Netherlands showing known sites (red circles) and previously unknown sites found with AGNES (blue squares). Yellow diamonds indicate known Early Medieval sites (with or without cremations) as recorded in the Archis system. Province names marked in black.	151

List of Tables

2.1	Illustrating the true/false positive/negative categories.	28
3.1	Descriptions and examples for each entity type. Examples are translated from Dutch.	39
3.2	Annotated corpus statistics.	41
3.3	Number of annotations per entity type in the data set	41
3.4	Inter-annotator agreement measures on 100 sentence test document. Calculated by doing pairwise comparisons between all combinations of annotators and averaging the results.	42
3.5	F1 scores for entity types and overall micro F1 compared between the previous and new data set. Species wasn't included in old data set, so we only present the score for the new data set.	42
4.1	Examples of noise introduced by (1) OCR mistakes, (2) PDF to text conversion and (3) manual metadata entry in free text fields (locations in time period field). Errors are underlined.	49
4.2	Overview of the included labels, full names and the number of sub-categories for each main category in time periods and site types. Category names are translated from Dutch.	55
4.3	Examples showing the conversion of free text metadata entries to structured label codes.	58

4.4	Overview of the scores for each method. Abbreviations refer to the following: TF-IDF (Sklearn, linear SVM with TF-IDF weights), D2V (Sklearn, linear SVM with Doc2Vec vectors), ONT (Sklearn, linear SVM classification based on ontology extracted entities) and SCY (Sklearn, linear SVM classification based on spaCy retrieved entities).	64
4.5	Overview of the top ten F1 scores for time period classification. PP = numerical values referring to pre-processing steps as described in Section 4.4.1, Aug = number of augments of the training set. . .	65
4.6	Overview of the top ten F1 scores for site types classification. PP = numerical values referring to pre-processing steps as described in Section 4.4.1, Aug = number of augments of the training set. . .	65
4.7	Overview of the best methods per individual category for time period classification and the overall average of these best methods. Column names yield the meaning as provided in the previous section.	68
4.8	Overview of the best methods per individual category for site type classification and the overall average of these best methods. Column names yield the meaning as provided in the previous section.	68
4.9	An overview of the F1 scores for all main and sub-categories for time period classification. Main categories are denoted in bold. . .	70
4.10	An overview of the F1 scores for the main and sub-categories for site type classification. Sub-categories not present within the reference test set are not included. Again, main categories are denoted in bold.	70
5.1	Synonymy and Polysemy examples	80
5.2	Precision, recall and F1-scores for the 3 targeted entities, on a scale of 0 to 1.	82
5.3	Overview of participants in focus group per category	86
5.4	Features and average scores (0-3) across focus group (n = 9), sorted by average score, descending.	88
6.1	Overview of participants in usability evaluation per category . . .	100
6.2	Three examples of user generated tasks and their associated queries and query reformulations (translated from Dutch).	102
6.3	Feedback split into positive and negative, with for each word how often it occurs in that context. Words only mentioned once are not included.	107

7.1	Descriptions and examples for each entity type. Examples are translated from Dutch. Adapted from (Brandesen <i>et al.</i> , 2020, p. 4574).	122
7.2	Micro average precision, recall and F1 score at token level (B and I labels), over 10 runs with different seeds, for each of the 5 folds (50 runs total). Standard deviation of F1 over the 10 runs is added in brackets for the Bidirectional Encoder Representations from Transformers (BERT) models. Standard deviation of precision and recall lies between 0.006 and 0.020. The ‘Fails’ column indicates the number of times the model failed to learn ($F1 = 0$).	127
7.3	The 10 most frequent error combinations between the 3 models for which at least one model has the correct prediction. Errors are marked in red.	129
7.4	Micro F1 score, precision and recall for the six ensemble methods, for one run over five folds. ArcheoBERTje results averaged over 50 runs and the optimised production model are added for comparison. The ArcheoBERTje predictions used as features for CRF are from the production model. The baseline features are the word- and context-based features used for CRF in prior work.	129
7.5	Overview of entities detected in the entire corpus, showing total and unique counts, plus the top 5 for each entity (translated from Dutch where relevant).	130
7.6	ArcheoBERTje precision, recall and F1 score for each label.	134
8.1	All nine queries used to retrieve results, in the order in which they were issued. An English translation is given for Dutch terms. Asterisks (*) are wildcards.	147
8.2	Overview of relevant, irrelevant and possibly relevant results. Relevant results are divided into previously known and unknown sites.	147
8.3	Overview of the different categories of irrelevant results. Percentages are rounded to whole numbers.	149
A.1	An overview of the frequencies for all site type categories. Main categories are denoted in bold.	208
A.2	An overview of the frequencies for all time period categories. Main categories are denoted in bold.	208
B.1	An overview of different types of lists and included terms.	209

C.1 An overview of the frequencies for all time period categories captured by the reference test set. Main categories are denoted in bold. 211

C.2 An overview of the F1 scores for the main and sub-categories for site type classification as captured by the reference test set. Sub-categories not present within the reference test set are not included. Again, main categories are denoted in bold. 212