Digging in documents – using text mining to unlock the hidden knowledge in Dutch archaeological reports

Brandsen, A.

Colofon

TMA 59, 2018
30ste jaargang
Prijs los nummer: €12,-

Het *Tijdschrift voor Mediterrane Archeologie* is een onafhankelijk tijdschrift dat aandacht besteedt aan actueel archeologisch onderzoek in de mediterrane wereld, in het bijzonder verricht vanuit Nederland en België. Het overnemen van artikelen is toegestaan mits met bronvermelding. Bijdragen van lezers kunnen al dan niet verkort door de redactie worden geplaatst.

TMA verschijnt twee keer per jaar. Opgave kan schriftelijk of via onze website. Een abonnement kost €20,-. Studenten betalen €15,- (onder vermelding van studentnummer).
Het abonnement loopt van 1 januari tot en met 31 december en wordt automatisch verlengd, tenzij een maand van tevoren schriftelijk is opgezegd.

Adres:
Tijdschrift voor Mediterrane Archeologie
Poststraat 6
9712 ER Groningen

Bankgegevens:
Stichting ter Ondersteuning Oudheidkundig Onderzoek
IBAN: NL14INGB0005859344
BIC: INGBNL2A

KvK: 41014777

TMA online:
- tijdschrift@mediterrane-archeologie.nl
- mediterrane-archeologie.nl
- rug.academia.edu/TMATijdschriftvoorMediterraneArcheologie
- facebook.com/mediterranearcheologie

TMA komt tot stand in samenwerking met Barkhuis Publishing, Eelde

ISSN 0922-3312
81999/SOOO

Inhoudsopgave

Artikelen

Recensies

→

Introducties op lopend onderzoek

Digging in documents – using text mining to unlock the hidden knowledge in Dutch archaeological reports

Promotieonderzoek (Universiteit Leiden, Leiden Centre of Data Science), Alex Brandsen

Mede ten gevolge van de ratificatie van het Verdrag van Malta produceren archeologen in Nederland rond de 60.000 rapporten per jaar. Deze 'grijze literatuur' wordt gepubliceerd buiten de traditionele drukkerijen om en is over het algemeen moeilijk te vinden. De informatie in deze rapporten is echter van groot belang en kan veel bijdragen aan wetenschappelijk onderzoek, mits archeologen de benodigde informatie kunnen vinden in deze *big data*.

Vroeger was het moeilijk om deze rapporten te verkrijgen, maar tegenwoordig stellen zowel DANS als de Rijksdienst voor het Cultureel Erfgoed een groot aantal van deze documenten online beschikbaar. Het probleem is dat via deze systemen alleen te zoeken is op metadata, die bijvoorbeeld beschrijven dat een rapport de Middeleeuwen behandelt. Er wordt echter niet vermeld dat er ook enkele artefacten uit de Bronstijd zijn gevonden, terwijl deze objecten belangrijk zouden kunnen zijn voor een onderzoek over de Bronstijd. Daarom is het nodig om alle tekst goed doorzoekbaar te maken.

Dit kan worden gerealiseerd met een *full-text search*, zoals in Google, maar ook hier kunnen zich problemen voordoen. Bij de zoekterm "Middeleeuwen" vindt een *full-text search* bijvoorbeeld niet "Middeleeuwse" en zeker niet "1000 na Christus". Deze synonymie is een veelvoorkomend fenomeen in rapporten. Ook het omgekeerde probleem komt voor, namelijk wanneer één woord verschillende betekenissen heeft. Om al deze complicaties het hoofd te bieden, moet een zoeksysteem taal 'begrijpen' en specifiek archeologische concepten kunnen herkennen.

In mijn project wordt *text mining* toegepast om automatisch relevante archeologische concepten te herkennen in tekst. Hiervoor gebruik ik *machine learning* (of machinaal leren), een vorm van kunstmatige intelligentie die op basis van voorbeelden uit handmatig geannoteerde teksten nieuwe woorden automatisch kan classificeren. Hier is in het verleden wel mee geëxperimenteerd,[1] maar er is helaas geen bruikbaar systeem uit voortgekomen.

Het doel van dit project is om een webapplicatie te bouwen: AGNES (*Archaeological Grey literature Named Entity Search*), die archeologen de mogelijkheid biedt om op een slimme en efficiënte manier door documenten te zoeken. Alhoewel de focus op dit moment op de Nederlandse archeologie ligt, is dit systeem ook relevant voor de mediterrane archeologie, aangezien bovengenoemde problemen in elk archeologisch onderzoeksgebied voorkomen. Een doel voor de toekomst is dan ook om dit systeem uit te breiden naar andere gebieden en talen.

Na zijn BA en MSc in Archeologie heeft *Alex Brandsen* (a.brandsen@arch.leidenuniv.nl) een paar jaar in de commerciële web-developmentsector gewerkt. Hij is nu aangesteld als promovendus bij de faculteit Archeologie en het Leiden Centre of Data Science aan de Universiteit Leiden.

Eindnoot

1 Paijmans, H. & Brandsen, A. 2010, "Searching in archaeological texts: Problems and solutions using an artificial intelligence approach", *PalArch's Journal of Vertebrate Palaeontology*, vol. 7, no. 2, pp. 1-6; Richards, J., Tudhope, D. & Vlachidis, A. 2015, "Text mining in archaeology: Extracting information from archaeological reports" in *Mathematics and archaeology*, (red.) J.A. Barceló & I. Bogdanovic, pp. 240-254.

Figuur 1. Het logo van de webapplicatie AGNES.