

Toepassing van benoemde entiteitsherkenning op die *Woordeboek van Suid-Afrikaanse Plekname*

Outeurs:

Christiaan du Plooy,
E Kotzé,
J du Plessis

Affiliësie:

Departement Rekenaar-
wetenskap en Informatika,
Universiteit van die Vrystaat
Posbus 339, Bloemfontein,
9300

Korresponderende outeur:

DC du Plooy
E-pos:
dcduplooy@gmail.com

Hoe om hierdie artikel aan te haal:

Christiaan du Plooy,
E Kotzé, J du Plessis,
Toepassing van benoemde
entiteitsherkenning op die
*Woordeboek van Suid-
Afrikaanse Plekname*,
*Suid-Afrikaanse Tydskrif
vir Natuurwetenskap en
Tegnologie* 38(1) (2019).
[https://doi.org/10.36303/
SATNT.2019.38.1.765](https://doi.org/10.36303/SATNT.2019.38.1.765)

Kopiereg:

© 2019. Authors.
Licensee: *Die Suid-
Afrikaanse Akademie vir
Wetenskap en Kuns*.
Hierdie werk is onder
die Creative Commons
Attribution License
gelisensieer.

Application of named entity recognition on the *Dictionary of Southern African Place Names*: This research project aims to address the design of an artefact to extract entities and relationships from the *Dictionary of Southern African Place Names*. Such an artefact would be valuable in the field of digital tourism: enriching user interfaces and providing tourists with information about relevant South African locations.

Inligting word al duisende jare aangeteken deur van simbole wat klanke verteenwoordig, gebruik te maak, soortgelyk aan die moderne alfabet. Geleerdes is só toegelaat om oor lang afstande te kommunikeer, gebeurtenisse aan te teken, waarnemings te maak en om geskrewe kunsvorme te skep. Sedert die begin van die rekenaaraera word die meeste inligting egter digitaal opgeneem en in gestruktureerde, semi- en ongestruktureerde dataformate op gesofistikeerde rekenaarsstelsels gestoor. Die soek en ontginning van nuwe inligting in ongestruktureerde dataformate kan uitdagend wees. Moderne rekenaars is egter geoptimeer om data te stoor, te soek en te analiseer, ongeag die formaat waarin die data gestoor word.

Inligtingonttrekkingstegnieke word voortdurend in natuurliketaalprosessering (NLP) gebruik om benoemde entiteite en die verhoudings tussen benoemde entiteite te onttrek uit semi- en ongestruktureerde databronne. Dubbelsinnige woorde, idiomatiese en figuurlike taalgebruik en die konteks waarin die teks gebruik word, maak dit egter moeilik vir rekenaars om algemene voorstellings van die teks op te bou.

Die doel van hierdie studie is om inligtingonttrekkingstegnieke op die *Woordeboek van Suid-Afrikaanse Plekname* (Raper et al 2014) toe te pas om gestruktureerde data uit die ongestruktureerde teks te verkry. Die gestruktureerde data verkry uit die *Woordeboek van Suid-Afrikaanse Plekname* kan van waarde wees op die gebied van digitale toerisme, aangesien dit koppelvlakke van mobiele toepassings of webwerwe kan verryk en aan potensiële toeriste inligting kan verskaf oor die plekke in Suid-Afrika wat hulle besoek of beplan om te besoek.

Inskrywings van individuele plekke in die *Woordeboek van Suid-Afrikaanse Plekname* bestaan uit: die naam van die plek, die provinsie waarin die plek geleë is, die breedte- en lengtegraad van die plek en 'n kort beskrywing van die plek. Die kort beskrywing van die plek bestaan uit: 'n kort aanduiding van die geografiese kategorie van die plek, die afstand en rigting na ander nabygeleë plekke, 'n kort opsomming van die plek en historiese gebeurtenisse, en die taal en oorsprong van die naam van die plek. Die benoemde-entiteitsherkenningstelsel word toegepas op die kort beskrywing van die plek, met die doel om persoonsname, plekname, taalname, betekenis, datums en tipe (plekke) vanuit die beskrywing te onttrek. Die verhoudings tussen benoemde entiteite word dan algoritmies bepaal en visueel voorgestel.

Verskeie klassifiseerders word vergelyk, insluitend steunvektorklassifiseerders, Bernoulli Naïewe Bayes-klassifiseerders en besluitnemingsbome. Die klassifiseerders word geleer om benoemde entiteite te onttrek deur na die kenmerke van benoemde entiteite te kyk in die "Groningen Meaning Bank"-korpus. Die mees gepaste klassifiseerder word dan geïdentifiseer deur die akkuraatheid, presisie, herroeping en f-tellings van die klassifiseerders te bepaal en te vergelyk.

Literatuurverwysings

Raper, P.E., Möller, L.A. & Du Plessis, T., 2014, 'Dictionary of Southern African Place Names' (4th ed), Johannesburg, Jonathan Ball.

Nota: 'n Seleksie van referaatopsommings: Studentesimposium in die Natuurwetenskappe, 25–26 Oktober 2018, SA Akademiegebou, Pretoria, Suid-Afrika. Gasredakteurs: Prof Rudi Pretorius (Departement Geografie, Universiteit van Suid-Afrika); Prof Chris Swanepoel (Departement Besluitkunde, Universiteit van Suid-Afrika); Me Andrea Lombard (Departement Geografie, Universiteit van Suid-Afrika)