

Embargoed until: 18 April 2019: 7am EST / 12pm UK time / 1 PM CET

The PeerJ logo consists of the word "PeerJ" in a white, sans-serif font, positioned on a solid blue square background.

PRESS RELEASE

Why language technology can't handle Game of Thrones (yet)

Researchers from the Vrije Universiteit Amsterdam and the Dutch Royal Academy's Humanities Cluster evaluated four state-of-the-art tools for recognising names in text, to assess and improve their performance on popular fiction. They find solutions to boost the tools' capability to recognise names in one novel from an accuracy of 7% to 90%.

Natural language processing (NLP) tools are commonly used in many day-to-day applications such as Siri and Google, but the effectiveness of these technologies is not thoroughly understood. Researchers from the Vrije Universiteit Amsterdam and the Dutch Royal Academy's Humanities Cluster have performed a thorough evaluation of four different name recognition tools on popular 40 novels, including *A Game of Thrones*. Their analyses, published in *PeerJ Computer Science*, highlight types of names and texts that are particularly challenging for these tools to identify as well as solutions for mitigating this. In addition, they extracted social networks from the novels to explore differences in story structure. These insights can help make such technologies more robust against genre differences, and can help for example make this technology more useful to journalists wanting to analyse large datasets such as the *Panama Papers*.

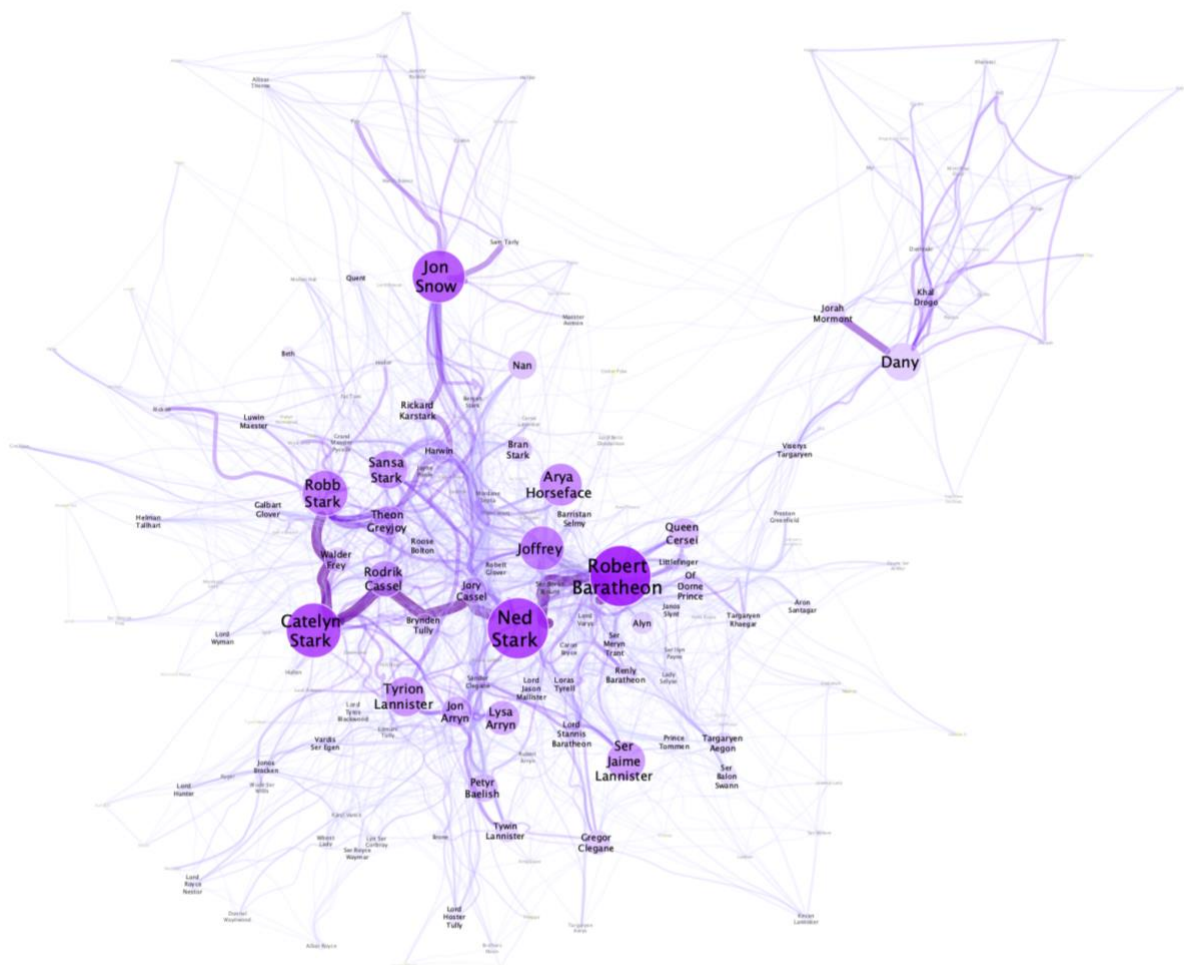
Many NLP tools are based on machine learning; that is, a computer program is trained to identify patterns in text based on previously fed examples. To recognise names in text, it is for example fed many newspaper articles in which humans have meticulously marked the names. The program is then tasked to 'learn' what a name looks like based on context (such as, it being preceded by *Mr*) or the shape of the word (such as that names generally start with a capital letter in English). Now, the problem when applying such a system trained on newspapers to novels, is that authors of novels have much more freedom in their narrative than journalists who need to stick to facts. Fiction authors can make up their own names, such as *Tywin* or *R'hllor*, or use descriptive character names straight from the dictionary such as *Grey Worm*. These names

do not behave like 'normal' names, thus NLP systems have difficulty recognising them in a text.

The experiments performed by Niels Dekker (Trifork B.V.), Tobias Kuhn (Vrije Universiteit Amsterdam) and Marieke van Erp (KNAW Humanities Cluster) also highlight the flexibility of language and how names are contextualised in stories. It is for example possible to refer to *Daenerys Targaryen* as *Daenerys* and *she*, but she is also known as *Dany*, *Daenerys Stormborn*, *Mother of Dragons*, *Khaleesi*, *the Unburnt* and *Mhysa*. The social network created for *A Game of Thrones*, illustrates for example that *Dany* is used by her friends, and her full name *Daenerys* only by her enemies (in her absence).

The research described in this publication shows that more attention should be paid to the performance of NLP tools and that there is still work to do before 'text' can be fully understood by computers.

Images:



Network visualisation showing that Dany/Daenerys is not close to other main characters in 'A Game of Thrones'. **Image credit:** N. M. Dekker, CC BY-SA 4.0



mauricio-santos-503880-unsplash.jpg: Photo by mauRÍCIO santos on Unsplash (public domain)

Full Media Pack including image:

https://drive.google.com/drive/folders/18_dei3tcQD-fuwaqcuskw0D1c2huivUA?usp=sharing

###

EMBARGOED until 18 April 2019: 7 am EST / 12 midday UK local time

Link to the Published Version of the article (quote this link in your story – the link will ONLY work after the embargo lifts): <http://peerj.com/articles/cs-189>
your readers will be able to freely access this article at this URL.

Citation to the article: Dekker N, Kuhn T, van Erp M. 2019. Evaluating named entity recognition tools for extracting social networks from novels. PeerJ Computer Science 5:e189 <https://doi.org/10.7717/peerj-cs.189>

###

About:

PeerJ is an Open Access publisher of seven peer-reviewed journals and a preprint server. PeerJ's mission is to help the world efficiently publish its knowledge. All works published by PeerJ are Open Access and published using a Creative Commons license (CC-BY 4.0). PeerJ is based in San Diego, CA and the UK and can be accessed at peerj.com

PeerJ Computer Science is the peer-reviewed, open access journal covering all [subject areas](#) in computer science, with the backing of a [prestigious advisory board](#) and more than [300 academic editors](#).

PeerJ has an Editorial Board of over 1,900 respected academics, including 5 Nobel Laureates. PeerJ was the recipient of the 2013 ALPSP Award for Publishing Innovation. PeerJ Media Resources (including logos) can be found at: peerj.com/about/press

###

Media Contacts:

Thijs van der Veen, communications advisor KNAW Humanities Cluster, thijs.van.der.veen@bb.huc.knaw.nl + 31 (6) 46 11 03 99 (English, Dutch; European time zones)

For the authors:

Niels Dekker, niels.dekker@trifork.nl +31 6 44 03 14 37 (English, Dutch; European time zones)

Dr. Marieke van Erp, marieke.van.erp@dh.huc.knaw.nl +31 (0) 20 4628 627 / +31 (6) 499 099 59 (English, Dutch; European time zones)

For PeerJ: email: press@peerj.com , <https://peerj.com/about/press/>

Note: If you would like to join the PeerJ Press Release list, please register at: <http://bit.ly/PressList>