openMINTED
Open Mining Infrastructure for Text & Data

Facebook.com/openminted

Twitter.com/openminted_eu
#TDMstory

A TEXT AND DATA MINER TELLS HIS STORY

I HELP SCIENTISTS DO SCIENCE

Daniel Duma, PhD candidate at Alan Turing Institute and University of Edinburgh

WHAT ARE YOU WORKING ON?
I'm trying to help scientists do science by creating software that will plug into your existing word processor or text editor. It will recommend papers that you should be aware of, you should read or that you would want to cite.

In other words, the software will effectively say to the scientist: "You don't know of this author, but they wrote a paper in which they said something very similar to what you are saying in your paper, so maybe you should read that paper because maybe it's relevant to your work". This is what I'm trying to build.

WHAT IS THE PROCESS BEHIND THE SOFTWARE?
The software works in two main stages: indexing and retrieval. At indexing time, it accesses the text of scientific articles to create a numeric representation of these in terms of the words they contain. At retrieval time, it extracts a relevant query from the text of a draft paper at the position the author is looking for recommendations, and computes how relevant this query is to every paper it has seen. It can then recommend a number of papers at any particular location in a document by taking the ones that have the highest relevance scores with the extracted query.

IS YOUR PROJECT OPEN SOURCE?
I want this to be a tool that people can use and contribute to. The code is open source (*www.github.com/danieldmm/minerva*) and it only uses open access publications. This is important because it's an absolute nightmare to deal with publishers in terms of accessing their copyrighted material, even simply to index it.

HOW CRITICAL IS IT FOR YOU TO HAVE FREE AND OPEN ACCESS TO DATA?
My software operates in a similar way to a search engine like Google. It needs to access documents in order to index them. If it can't index them, then nobody can find those documents. If the scientific content I am trying to recommend is locked behind a paywall and I am not able to index it, it won't be discoverable.

HOW DO YOU SEE THE FUTURE FOR TDM?
The applications for TDM are infinite. You can do anything. You can eventually make robot scientists or get algorithms to help you in your research at a human level! I think this is the future to come. It's not far away and I'm excited to work on it. I encourage others to join and work on this as well.

Website: www.danielduma.com

Related paper: Duma et al., Rhetorical Classification of Anchor Text for Citation Recommendation, D-Lib magazine, October 2016

European Commission

www.openminted.eu

OpenMinTeD is funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 654021.