openMINTED
Open Mining Infrastructure for Text & Data

Facebook.com/openminted

Twitter.com/openminted_eu
#TDMstory

A TEXT AND DATA MINER TELLS HIS STORY

ANALYSING THE RECENT PAST

Federico Nanni about his work as a Post-Doc at the University of Mannheim

HOW HAVE YOU APPLIED TDM DURING YOUR STUDIES?
Web archives preserve an unprecedented abundance of primary sources which allow us to track, examine and understand major recent events and transformations in our society. A topic of interest for both historians and political scientists is the rise of Euroscepticism as a consequence of the recent economic crisis. To support such studies, I've used TDM to build event-centric sub-collections from large archives. These include the core documents related to the event being studied and, more importantly, documents which describe related aspects (e.g., premises and consequences). This is done by identifying relevant concepts and entities from a knowledge base and then detecting mentions of these entities in documents. These mentions are interpreted as an indicator for relevance.

WHAT CHALLENGES DO YOU FACE WHEN USING TDM?
One barrier is getting the data for research. When it comes to web archives, national libraries often don't release data for text mining. You only have access through a search engine and are, for example, not able to scrape the data. As a result, the data cannot be used. An additional issue emerges when the archives are 'open' but do not provide an easy way to access their data, for example via an API. You then have to write your own code to be able to download the data, which is what most researchers end up doing but this requires programming skills. Another issue is not knowing whether the data made available via API is complete. This information should always be easily accessible, not hidden somewhere on the website. And, as a last issue, there's reproducibility. If a political scientist proposes something and the data is incomplete or cannot be reproduced, this has a negative impact on research. People often share the script they used to gather data but this type of code can be messy and not easy to use by others.

HOW COULD THIS BE IMPROVED?
A best-practice for researchers is to clearly document all your code, all the things you've tried, all the parameters you've tuned to get the results so others can reproduce your data, correct you or give feedback on what you're doing.

WHAT DO YOU WANT TO SEE IN FUTURE?
Historians and political scientists should do more text mining. In the future we will have more and more born-digital and digitised sources so it's fundamental for us to learn how to access and examine these sources.

Related paper: Nanni et al., Building Entity Centric Event Collections, Proc. of JCDL, 2017, https://ub-madoc.bib.uni-mannheim.de/42529/1/paper.pdf

European Commission

www.openminted.eu

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