Learning about Text and Data Mining, the Future of Open Science

Martine Oudenhoven, Nancy Pontika

The volume of digital data is doubling every two years (EMC, 2014). In the world of science, the cumulative total of articles published since 1665 is estimated to be more than 50 million (Jinha, 2010). There is a wealth of knowledge hidden in this massive volume of articles, but reading and analyzing all of them manually is not humanly possible. Text and data mining (TDM) can provide a solution. It can read and analyze millions of texts quickly and reveal patterns and trends that can lead to new discoveries in various fields, for example in research analytics, medicine, agriculture and social sciences.

However, researchers encounter challenges while trying to mine. Not all text and data miners possess the technical skills that are needed for the existing tools. There are also legal barriers, as the current copyright law, the right to read articles does not include the right to mine them. And thirdly there are interoperability barriers. Even if the data are openly accessible, they are often only available on publishers' websites that support only their own technology for accessing this information (Knoth and Pontika, 2016). As a result, combining services on one dataset is almost impossible.

The European project OpenMinTeD provides solutions, by working on a new infrastructure for text and data mining. The project will:

- provide an extensive collection of TDM tools and services, which can be used across disciplines and communities.
- give access to big amounts of mineable open science + content, both full-text and data.
- establish interoperability standards and build a standard layer. In this way, miners can combine different TDM services to produce their data.
- Provide training and support for researchers, content providers and service providers (developers of TDM tools and services), in the form of workshops as well as online resources and courses.
- Encourage developers to further develop existing open tools and services.

Our goal is to make the world of open science mineable, therefore we work with content providers of open text and data, such as CORE and OpenAIRE. In order ensure that the platform and services meet user requirements, OpenMinTeD works together with different user communities, including research analytics, life sciences, agriculture & biodiversity and social sciences. OpenMinTeD also collaborates with another project, FutureTDM, that focuses on the technical and legal barriers for text and data mining on a policy level.

OpenMinTeD is a project funded by the European Commission and started in June 2015. It is halfway through it 3-year contract. We are already working on a sustainability plan and we will work closely with the OpenAIRE project and CORE to continue the infrastructural work on text and data mining in the long-term. All tools and services that are built as part of OpenMinTeD are open access and open source. Our mission is to make the immense amount of text and data that currently exists discoverable, and we believe TDM is the way to do this.