TDM for Climate Change science

Commercial use (Y/N): No

Related project (*if any – please specify funder and project name*): Funded by European Commission. Ocean Food web Patrol – Climate Effects: Reducing Targeted Uncertainties with an Interactive Network (OCEAN-CERTAIN). Work programme topic: F7- ENV.2013.6.1-1 Climate Related Ocean Processes and Combined Impacts of Multiple Stressors on the Marine Environment.

Author: Pinar Öztürk and Erwin Marsi
Organization: Norwegian University of Science and Technology (NTNU)
Date: 19 December 2014

Use case description – *Describe the problem you are trying to address, the research area it affects, and the TDM application used. Please state all participants, whether and how the results were subsequently used.*

We are interested in *knowledge discovery* related to the science of Climate Change. Our goal is to design and develop a discovery support system to be used by researchers in various disciplines in the Climate Change domain. This is a cross-disciplinary domain where scientists in marine chemistry, marine biologists, geologists, environmental sciences, etc. are trying to better understand the impacts of climate change on the marine food web and the related process of CO2 sequestration through the biological pump. This is a complex problem which attracted a lot of researchers in various disciplines, leading to a huge number of publications. The sheer volume of publications makes it impossible for any researcher to find and read all publications relevant to his/her own research. Differences in the terminologies of various disciplines exacerbate this situation. Consequently, as argued by Swanson (one of the pioneers of text mining on scientific literature), there is a lot of latent public knowledge out there waiting to be discovered. Computational support for the process of scientific discovery seems an obvious direction to go.

We are particularly interested in detecting events and the causes, possibly causal chains, leading to these events. Hence, entity and event detection and relation extraction are important parts of our work. For this, we need access to scientific publications and the rights to perform text mining on them.

Content sources – *List the targeted content sources and types of content to used.*

Our content type of interest is scientific text on any topic related to climate science and marine science in the broadest sense. This involves both free text and semi-structured text such as tables and references. Our primary source is journal articles, but other potential sources include text books, scientific reports, encyclopaedic texts, popular scientific magazines and informative websites. Of particular interest is content from journals with a high impact factor by major publishers such as Elsevier, Springer, Taylor & Francis or Nature Publishing Group.

Wherever available (which is almost universally so for modern publications), access to the XML source of the documents is highly desirable, as extracting text from PDF or HMTL formats is problematic.

Access through a web API adhering to common standards is preferable. In addition, similar access to search engines/interfaces for retrieving content of interest is essential.

Targeted users – *Describe end users, their number and expertise.*

Typical end users are scientists concerned with understanding and explaining the impacts of climate change in marine environments. This includes chemists, biologist, geologists, physicists, oceanographers, geo-chemists, meteorologists, etc. The exact number of potential users is difficult to estimate, but must be at least a couple of thousand.

Impact – *Describe all possible impacts of the use case. Specify any cross-border, societal and economic effects (if possible state any monetary benefits and market advantages) .*

- increase in the speed of scientific discovery
- decrease in the cost of scientific discoveries
- the two factors above will therefore contribute to a faster and better understanding of the effects of climate change
- this in turn will facilitate predictions about the future, as well as
- inform the policy makers responsible for adaptation and mitigation measures

Constraints – *Describe any legal, technical, economic, societal, organizational, cultural, multi-lingual or other limitations and you have encountered.*

Even though our university pays publishers for access to content such as journal articles and other research publications, this is considered to be for human consumption only. From a legal point of view, the licences that our library has with various publishers do not allow bulk downloading of publications.

Text mining usually involves a number of processing steps to clean up, linguistically analyse and annotate the text material. The resulting text corpus would ideally be made available to other researchers in order to guarantee reproducibility of results as well as to allow future research on top of existing results. Unfortunately current licenses prevent such proper scientific behaviour.

Since we do not have access to the document source text (usually in XML format), we have to extract text from PDF or HTML files. This is a non-trivial task that can not be automated in a general way to produce error-free output. In practice, it is a labour intensive and time consuming activity, which takes up a substantial amount of resources that could otherwise be spent on true research. Considering that many researchers or research groups will do the same thing again and again (since we are legally forbidden to share the extracted and processed text), the current TDM policies cost the research communities/organisations, and ultimately the whole society, huge sums of money.

Comments – *Other comments and any recommendations you may have.*