FutureTDM
Explore . Analyse . Improve

TOOLS

IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

Facts

Project No: 665940

Program: H2020 | CSA | GARRI-3-2014

Duration: 09/2015 - 08/2017

ContentMine
Zika Tutorial

ContentMine

This tutorial introduces researchers and others interested in the Zika virus to basic and advanced text and data mining methods with the ContentMine toolchain.

Why?

Pandemics are a very important and critical field of research, where fast and open access to the scientific knowledge is a matter of life and deatg. In crisis situations, where every hour matters, openly available software and publications lowers the barrier for collaboration and contribution. This is especially crucial for poorer countries, who are often the ones affected by diseases. This tutorial will help researchers in the field of the Zika virus to get an overview of the research and to dive deeper into it. This can also be seen as a general introduction into pandemics, because it is easy to adapt to other similar cases.

How?

The Open Source software from ContentMine will be used to apply exploratory text and data mining techniques to the Open Access content of Europe PMC. After downloading all relevant publications from the EPMC API, the data is processed so it is usable for text mining, annotation and further analysis. The ContentMine software extracts facts from the publications, automatically adds metadata from established authorities, especially through Wikidata . Everything worked out can be found at https://github.com/ContentMine/FutureTDM

Who?

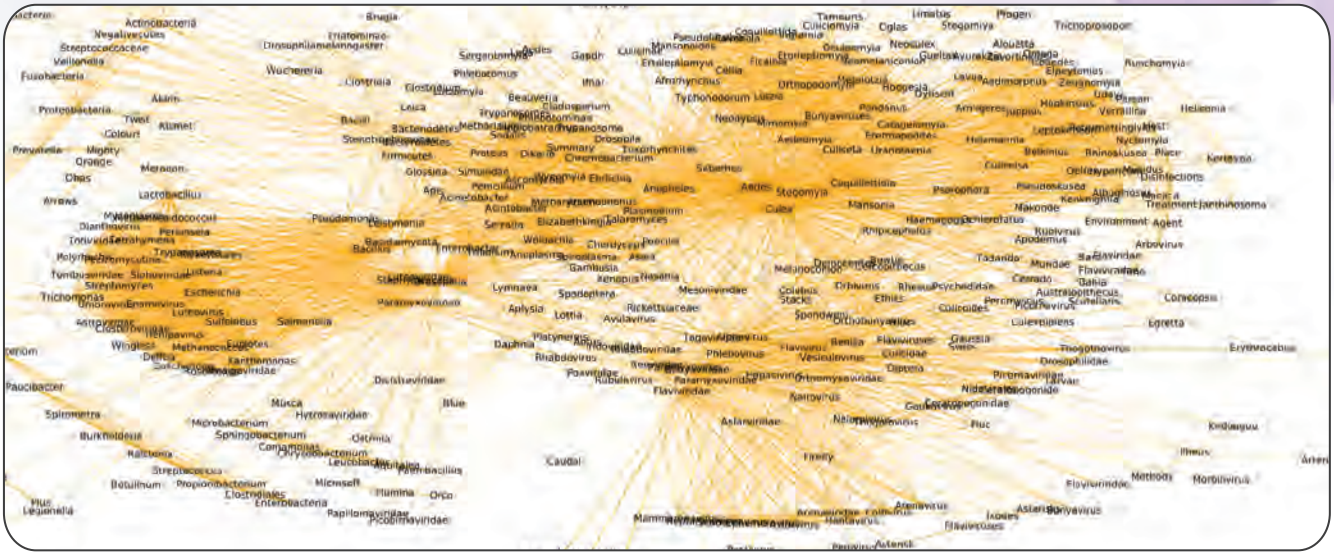
This tutorial has been developed by the ContentMine team as a partner of the FutureTDM project. The project is supported by the Shuttleworth Foundation through the fellowship of Peter Murray-Rust.

ContentMine Ltd (http://contentmine.org/) is a UK non-profit organisation.

Setup

You can use the ContentMine software through our virtual machine image or install the software parts separately. Additionally a Jupyter installation is needed.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.



Get the data

The needed data is acquired via a Europe PMC API query. Define the query, and then use getpapers, a tool for downloading the scientific publications from different repositories. We downloaded all the retrieved papers for "zika", "aedes" and "usutu" (more than 1500).

Extract the facts

After normalizing the data, two ami-plugins are used to extract entities from the text. Ami-word for word-frequencies and ami-species for genus, genus spp. and binomial.

Analyse with a Jupyter notebook

The analysis is done in a Jupyter notebook in python.

Look at the metadata:

A good start to understand a research field, is, to have a look at the authors, the publication dates and the journals. This basic information tells about the degree of maturity of it.

Explore words and species:

Next the frequency of words used is explored for a better understanding of the content in the whole literature downloaded. The frequency of species-terms is then a good way to look out for virus-transmitters. For example Wolbachia, Aedes and Flavivirus are on top of the list.

Follow one specific species:

Finally, we follow one fact to find out more about it. Which one, can be chosen by the researcher. We look on other species occurring with our species in the same publication. Then we look on the publication in which our fact is mentioned. At the end, we plot the full graph of co-occurrences between publications and our species-facts.

Follow Up's

Photo: Aedes aegypti feeding in Dar es Salaam, Tanzania by Muhammad Mahdi Karim (GFDL 1.2)

Graphic: Species Graph by Stefan Kasberger (CC by 4.0)

http://contentmine.org/

Tutorial: https://github.com/ContentMine/FutureTDM