Disease Tagging in Biomedical Literature

Biomedical literature is a hive of valuable information on research topics like diseases, drug/treatment attributes, medical decisions, health effects, population data and epidemiology, and more. With advances in technology, there is a rapid growth in the amount of this literature - making it impossible for researchers and practitioners alone to exhaust all of this valuable information.

With KNIME Software, mining knowledge from text such as disease-related information can be automated. An analytics expert creates a workflow in KNIME Analytics Platform, which contains a model that learns disease names from a set of documents in the biomedical literature. For that, abstracts are automatically extracted from PubMed. These documents (the corpus) are used to train the model, starting with an initial list of disease names (the dictionary). The resulting model is evaluated using documents that were not part of the training. One of the more useful features of these models is that they can extract new information - identifying disease names that were not part of the training set. To check this, we compare the detected disease names with the initial dictionary.

The trained model is then deployed to the KNIME WebPortal via KNIME Server. Here, with the predetermined interaction points (defined by the analytics expert), researchers can interactively inspect the diseases that co-occur in the same documents and explore genetic information associated with these diseases. They can also investigate and filter results as well as visualize them in a network graph (Fig.1).

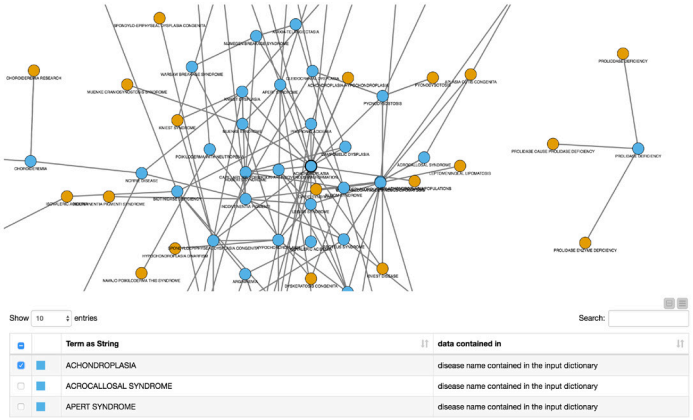

Fig. 1: Interactive view on KNIME WebPortal which users researchers interact with.

Results:

Deploying a KNIME Analytics Workflow on KNIME Server and making it available as an analytical application enables researchers to:
Reduce need to sift manually through biomedical literature

- Automatically gather the latest literature
- Interactively explore and filter results
- Find disease names in the literature that were not part of the original input
- Generate new hypotheses - for example for genetic associations of diseases co-occurring with well-studied diseases

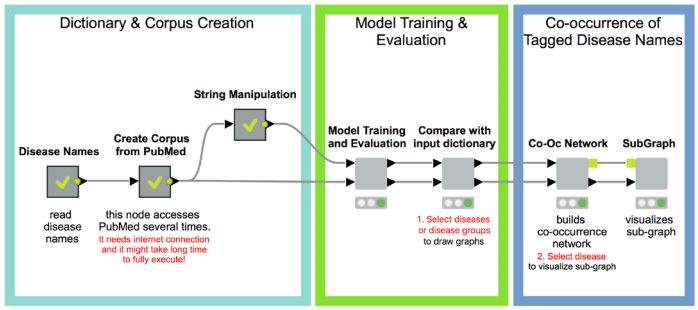

Fig 2. High-level KNIME workflow

The KNIME Textprocessing Extension within KNIME Analytics Platform enables data scientists to build a model that automatically manages large amounts of text data. The StanfordNLP nodes facilitate building and evaluating the model, the Term Co-Occurrence Counter investigates co-occurring diseases, and the Network Mining nodes make it possible to visualize and analyze results. KNIME Server makes the results accessible to researchers and domain experts via the KNIME WebPortal.

Try it out for yourself!

This workflow is available on the KNIME Workflow Hub:
08_Other_Analytics_Types/02_Chemistry_and_Life_Sciences/03_Fun_with_Tags

KNIME AG · Technoparkstrasse 1 · 8005 Zurich, Switzerland · info@knime.com · www.knime.com