Scientific Publication Mining

Scientific research continues to expand both human understanding of our world and solve societal problems through technical progress. One way that this progress is documented is through scientific publications. However, there are now millions of publications available for researchers from all science and technology domains. Consequently, it is nearly impossible for humans to thoroughly research across these millions of publications. The goal of this challenge is to develop and apply machine learning and statistical techniques to mine these publications and identify key characteristics and patterns that can be used by human researchers to develop useful knowledge and further enhance scientific discovery.

The dataset available for this task consist of scientific publication records. The metadata for each publication include title, abstract, author list and the publication venue and date. For a portion of the publications the full-text of the paper will also be available. The participants are welcome to use external data in their approaches as long as that data is publicly accessible. All participants will be asked to document all external data sources, and detail how the data was used.

Challenge Questions

1. Identify the individual or group of individuals who appear to be the expert in a particular field or sub-field.
    a. Experts are people with high level of knowledge in a certain area. Recognizing experts can be beneficial to students familiarizing themselves with a new area or to scientists looking for collaborators. The goal of this task is to employ different methods, for example modelling or graph-based algorithms, and apply them on the dataset to discover people with high level of expertise. The response to this task should include example output, such as the model or graph developed with highlighted important nodes or a list of names, and a description of tools and methods used to produce the output.
2. Identify topics that have been researched across all publications.
    a. Given a collection of documents, the goal of this task is to extract topics that recur in the collection so that a person not familiar with the collection can quickly explore its contents. The aim is to assist human understanding, so a good solution should identify topics in a way that makes sense to a person. This task could explore for example graph or text clustering methods. The solution should also include a description of methods used for the task.
3. Visualize the geographic distribution of the topics in the publications.
    a. Researchers are associated with different institutions across the globe. Following up on the previous task, the goal of this task is to visualize the identified topics

with respect to their geographical distribution. Are there certain locations which focus on specific topics? The solution should again contain a description of how was the output produced.

4. Identify how topics have shifted over time.
    a. The goal of this task is understanding popularity evolution of topics over time, or in other words how the knowledge base is changing over time with the influx of new topics, growth or decay of older topics. Understanding the popularity of topics is important because it helps in identifying trending topics. Same as in case of the previous tasks, the solution should include example output and a description of methods used to produce the output.
5. Given a research proposal, determine whether the proposed work has been accomplished previously.
    a. Choosing which proposals to fund is a complicated task, because the evaluators needs to be aware of the research area and whether the proposed research is novel. The goal of this task is to identify whether there are any publications which have previously tackled the proposed research.