NLPCC2022 Shared Task5

Multi-label Classification, NER, Content Extraction for Scientific Literature

Training Datasets

CNPIEC KEXIN LTD

April 6th, 2022

Table of Contents

I. TASK OVERVIEW.....1

II. TRACK 1 - Multi-label Classification Model for English Scientific Literature.....2

III. TRACK 2 - Supervised Named Entity Recognition Model for English Domain-specific Texts.....3

IV. TRACK 3 - Detection and Extraction of Inline and Isolated Mathematical Expressions Model for English PDF Articles.....4

I. TASK OVERVIEW

Managing and exploring scientific literature efficiently becomes more and more important for people such as researchers due to the growing large number of publications. Owning to the development of artificial intelligence (AI) and natural language processing (NLP) technology, literature management and exploration can be more efficiently by using AI-based literature classification, literature search and recommendation as well as PDF content extraction functions. Aiming at solving the related challenging fundamental problems, we set three tracks in this task (with three datasets provided):

Track 1: Multi-label Classification Model for English Scientific Literature: develop a multi-label classification model for scientific research literature based on the given metadata (title and abstract) of scientific research literature and corresponding hierarchical labels from a domain-specific topic taxonomy.

Track 2: Supervised Named Entity Recognition Model for English Domain-specific Texts: develop a named entity recognition model for domain-specific texts based on state-of-the-art NLP and deep learning technique with the labeled domain-specific sentences corresponding to seven entity types.

Track 3: Detection and Extraction of Inline and Isolated Mathematical Expressions Model for English PDF Articles: design and develop a model that can automatically recognize and extract the inline and isolated mathematical expressions from the provided English PDF articles without using any commercial third-party APIs.

Organizer: CNPIEC KEXIN LTD and Data Intelligence

Contact: He ZHANG (rd_kexin@cnpiec.com.cn)

II. TRACK 1 - Multi-label Classification Model for English Scientific Literature

Please go to the specified link to obtain training datasets:

Datasets for NLPCC2022.SharedTask5.Track1: https://www.scidb.cn/s/q6NnMr

III. TRACK 2 - Supervised Named Entity Recognition Model for English Domain-specific Texts

Please go to the specified link to obtain training datasets:

Datasets for NLPCC2022.SharedTask5.Track2: https://www.scidb.cn/s/RnURVr

IV. TRACK 3 - Detection and Extraction of Inline and Isolated Mathematical Expressions Model for English PDF Articles

Please go to the specified link to obtain training datasets:

Datasets for NLPCC2022.SharedTask5.Track3: https://www.scidb.cn/en/s/e6fQVb