

Problem: Classify given set of Pubmed abstracts (biomedical literature abstracts) into four classes:

- a) Abstracts containing Drug adverse events
- b) Abstracts containing Congenital anomalies
- c) Abstracts containing congenital anomalies caused by drug usage.
- d) Others

Dataset: Pubmed (<https://pubmed.ncbi.nlm.nih.gov/>)

Programming language: Python preferred

Background

Pubmed is a repository of all biomedical literature. NCBI, the agency managing Pubmed, provides e-utilities to download abstracts from their database using curl and other python programs (Entrez e-utils). Examples of using these e-utilities are given in this link (<https://www.ncbi.nlm.nih.gov/books/NBK25498/>). The examples in the link provide use of the utilities in Perl, but similar functionality is also present in Python (using requests library).

MeSH (or Medical Subject Headings) is a system devised by NCBI to provide 'topics' to every abstract in the Pubmed database (e.g. <https://www.ncbi.nlm.nih.gov/mesh/68064419>). Both drug side-effects and congenital anomalies have entries in the MeSH database (<https://www.ncbi.nlm.nih.gov/mesh/68064420> and <https://www.ncbi.nlm.nih.gov/mesh/68000013>). The MeSH terms provide an easy way to download abstracts belonging to these topics (the first example in the Entrez utility link above shows how to search for MeSH terms in Pubmed). Note that NCBI uses semi-manual methods to assign MeSH topics to abstracts, so it is not a simple keyword based method.

Individual tasks

Task 1: Write a Python program to download at most 10,000 abstracts (in either XML format or just the plain text abstracts) using Entrez utilities belonging to each of the four classes given at the top of the page.

Task 2 (major task): Write a classifier using any AI/ML technique to classify a given novel set of abstracts into the four classes given above. Perform n-fold cross validation and provide performance metrics for the classification.

Note: The approaches used are more important at this stage than fine-tuning for high performance. You can use open-source software for the task.