Project Proposal

Pavan Bhagwan Choudhari | Priya Garg

PubMed is a free search engine that comprises citations for biomedical literature from MEDLINE, life science journals, and online books. The search engine uses MeSH Headings to help researchers and experts find the resource papers for a target topic. However, these documents are manually annotated. The plan is to build better annotation methods that could reduce manual burden.

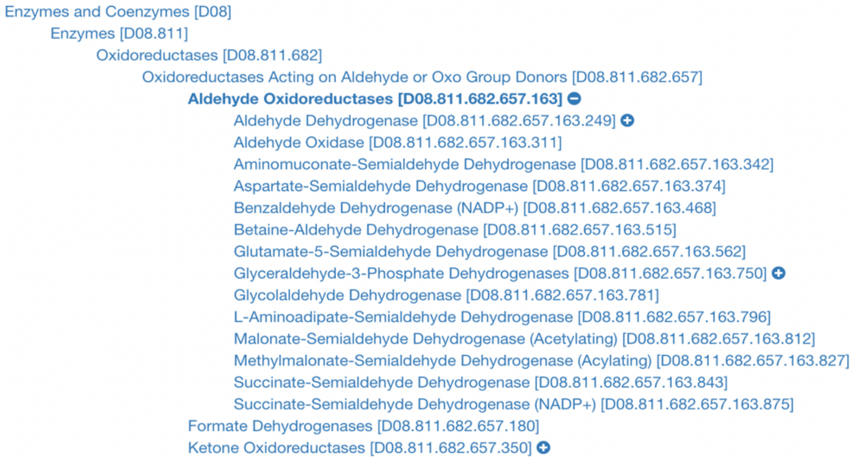
Our dataset is available at this link. If a research article is on aldehyde oxidoreductases, then figure 1, shows the tree structure for that topic, which can be traced back to its root domain, Chemicals and Drugs here as shown in figure 2. The focus of this project to classify the 16 root categories. As shown in figure 3, data contains descriptor name, which is reverse mapped to these categories. Also, the original data size is around 34 million rows (79 GB), from which we will sample ~50,000 records. We plan to use python, pandas, nltk, spacey, sklearn, pytorch, transformers, gradio, seaborn.

The plan is to test a wide array of models. Most of the models are provided by libraries. We plan to start with N-gram features that can be built using CountVectorizer, and training models for different orders of n-grams. This is not a competition of how much data can be fit in memory (sparse vectors are too big), but a comparison of using 1-gram, 2-gram and 3-gram as features. An alternative is to build a TFIDF model with TfidfVectorizer and train a classifier model. We also might work with Logistic Regression, SVM and Tree based models. The final goal is to get better domain specific features by using transformers as sequence classifiers. A particularly interesting pretrained BERT model is microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract that can be fine-tuned for this sequence classification task. For BERT, we'll have to add a few custom final layers for classification, and a custom DataLoader. F1-score will be leveraged to compare the models.

Since there is a lot of data to sample from, visualizations like bar charts will ensure a good sample as the training set. Also, word clouds would give us insights on medical terms in research articles.

The preliminary source of reference are the articles with similar work (internet research), and documentations for Python libraries. Our plan is to start the project on 11/19/2022. We expect one week for the preliminary work (data exploration and baseline models), and another week for the implementation of complicated models.

Both of us want to get knowledge and experience by working on each model type, so we will split the work evenly. For instance, if one builds BERT the other can experiment on Roberta. The same applies for other model categories.



- Enzymes and Coenzymes [D08]
  - Enzymes [D08.811]
    - Oxidoreductases [D08.811.682]
      - Oxidoreductases Acting on Aldehyde or Oxo Group Donors [D08.811.682.657]
        - **Aldehyde Oxidoreductases [D08.811.682.657.163]** ⊖
          - Aldehyde Dehydrogenase [D08.811.682.657.163.249] ⊕
          - Aldehyde Oxidase [D08.811.682.657.163.311]
          - Aminomuconate-Semialdehyde Dehydrogenase [D08.811.682.657.163.342]
          - Aspartate-Semialdehyde Dehydrogenase [D08.811.682.657.163.374]
          - Benzaldehyde Dehydrogenase (NADP+) [D08.811.682.657.163.468]
          - Betaine-Aldehyde Dehydrogenase [D08.811.682.657.163.515]
          - Glutamate-5-Semialdehyde Dehydrogenase [D08.811.682.657.163.562]
          - Glyceraldehyde-3-Phosphate Dehydrogenases [D08.811.682.657.163.750] ⊕
          - Glycolaldehyde Dehydrogenase [D08.811.682.657.163.781]
          - L-Aminoadipate-Semialdehyde Dehydrogenase [D08.811.682.657.163.796]
          - Malonate-Semialdehyde Dehydrogenase (Acetylating) [D08.811.682.657.163.812]
          - Methylmalonate-Semialdehyde Dehydrogenase (Acylating) [D08.811.682.657.163.827]
          - Succinate-Semialdehyde Dehydrogenase [D08.811.682.657.163.843]
          - Succinate-Semialdehyde Dehydrogenase (NADP+) [D08.811.682.657.163.875]
        - Formate Dehydrogenases [D08.811.682.657.180]
        - Ketone Oxidoreductases [D08.811.682.657.350] ⊕

Figure 1: aldehyde oxidoreductases tree view

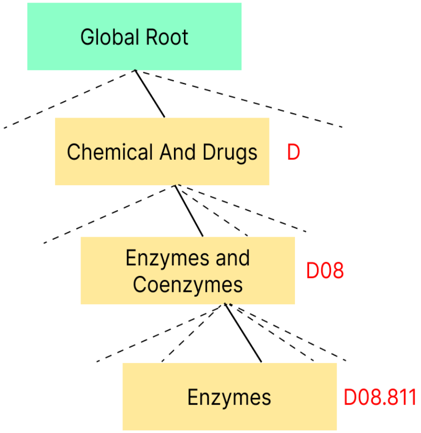
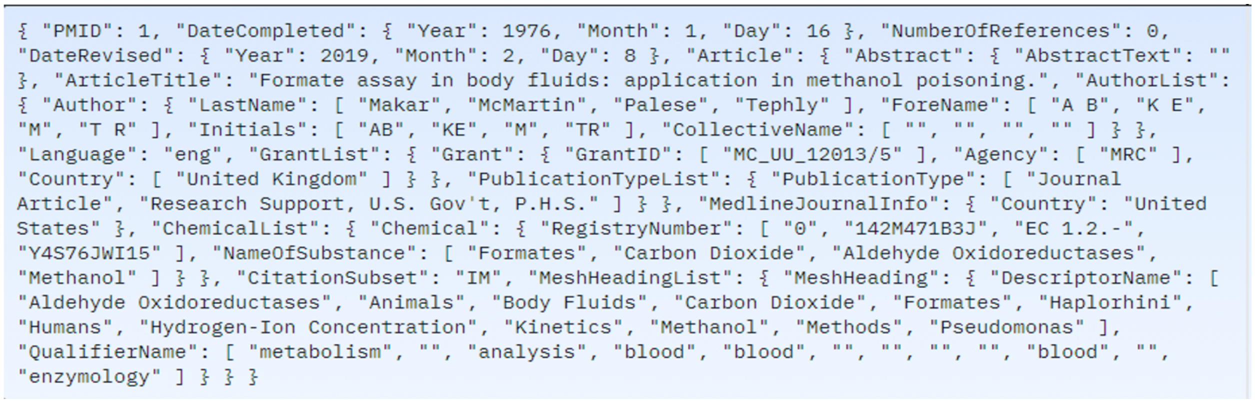


Figure 2: Tree Structure for Category 'D'



```
{ "PMID": 1, "DateCompleted": { "Year": 1976, "Month": 1, "Day": 16 }, "NumberOfReferences": 0,
"DateRevised": { "Year": 2019, "Month": 2, "Day": 8 }, "Article": { "Abstract": { "AbstractText": ""
}, "ArticleTitle": "Formate assay in body fluids: application in methanol poisoning.", "AuthorList":
{ "Author": { "LastName": [ "Makar", "McMartin", "Palese", "Tephly" ], "ForeName": [ "A B", "K E",
"M", "T R" ], "Initials": [ "AB", "KE", "M", "TR" ], "CollectiveName": [ "", "", "", "" ] } },
"Language": "eng", "GrantList": { "Grant": { "GrantID": [ "MC_UU_12013/5" ], "Agency": [ "MRC" ],
"Country": [ "United Kingdom" ] } }, "PublicationTypeList": { "PublicationType": [ "Journal
Article", "Research Support, U.S. Gov't, P.H.S." ] } }, "MedlineJournalInfo": { "Country": "United
States" }, "ChemicalList": { "Chemical": { "RegistryNumber": [ "0", "142M471B3J", "EC 1.2.-",
"Y4S76JWI15" ], "NameOfSubstance": [ "Formates", "Carbon Dioxide", "Aldehyde Oxidoreductases",
"Methanol" ] } }, "CitationSubset": "IM", "MeshHeadingList": { "MeshHeading": { "DescriptorName": [
"Aldehyde Oxidoreductases", "Animals", "Body Fluids", "Carbon Dioxide", "Formates", "Haplorhini",
"Humans", "Hydrogen-Ion Concentration", "Kinetics", "Methanol", "Methods", "Pseudomonas" ],
"QualifierName": [ "metabolism", "", "analysis", "blood", "blood", "", "", "", "", "blood", "",
"enzymology" ] } } }
```

Figure 3: Single record