ML4H
Project 2

NLP/Representation learning
19.03.2020

NLP/Representation learning project

- As medical records with text is hard to come by...
- The aim is here to use NLP and representation learning methods to improve the ML model that predicts the given target.
    - Evaluation of the project is not going to be strictly based on the evaluation metrics, but as well the discussion on why and how various transformation on text helps (or not!) to get a better model.
        - Use visualization tools for data exploration!

Predict readmission within 30 days of discharge

- Based on the Diabetes 130-US hospitals for years 1999-2008 UCI data set:
  - Subsampled into 10k samples
  - Standard 60-20-20 splits for train-test-validation
  - Expanded with the descriptions of the ICD codes
- **Problem**: simple binary classification (target variable: '**readmitted'**)
  - ~65.7% of the samples are readmitted
- Feature types:
  - Ordinal: 'num_medications'
  - Categorical: 'race', 'gender'
  - Text: 'diag_1_desc', 'diag_2_desc', 'diag_3_desc'
- Note: many missing values! ('?')
- **Evaluation metrics**: AUROC, F1-score

COVID-19 Open Research Dataset Challenge

- Metadata for papers from these sources are combined: CZI, PMC, BioRxiv/MedRxiv. (total records 29500)
  - CZI 1236 records
  - PMC 27337
  - bioRxiv 566
  - medRxiv 361
- 17K of the paper records have PDFs and the hash of the PDFs are in 'sha'
- For PMC sourced papers, one paper's metadata can be associated with one or more PDFs/shas under that paper - a PDF/sha correponding (sic!) to the main article, and possibly additional PDF/shas corresponding to supporting materials for the article.

COVID-19 Open Research Dataset Challenge

- 13K of the PDFs were processed with fulltext ('has_full_text'=True)
- Various 'keys' are populated with the metadata:
  - 'pmcid': populated for all PMC paper records (27337 non null)
  - 'doi': populated for all BioRxiv/MedRxiv paper records and most of the other records (26357 non null)
  - 'WHO #Covidence': populated for all CZI records and none of the other records (1236 non null)
  - 'pubmed_id': populated for some of the records
  - 'Microsoft Academic Paper ID': populated for some of the records

Deliverables

- Conda env YAML file to reproduce your runtime environment
- A separate Jupyter notebook for each task
- Deadline(s):
  - For the first task a final solution by 09.04.2020@23:59 CEST
  - For the second task a WIP by 09.04.2020@23:59 CEST and your final solution by 16.04.2020@23:59 UTC (right after you have submitted to kaggle)