Project Proposal

Data Engineering & Analytics on Medical Tech-Publications

Project Description:

There is a free ONG to collect the medical technical publications named "***pubmed***", based upon the data available we will do necessary transformation activities to identify the maximum number of publishers on the same technical aspects as drugs, diseases, etc.

Data Brief:

Data is available on the website in zip folders, each zip folder has an XML file with details of multiple technical papers details as author, abstract, grant, published date, country, Publication Company, title, etc.

Anticipated Difficulties:

- Downloading the top 10 zip folders from the website using a web scraping technique.
- Automating script to unzip the folder and collect the required data from the nested xml file to a dataset.

Timeline:

- Time required for Data collection one week.
- Data quality check and modelling one week.
- Analysing and visualizing the data for two weeks.