**Preprocess**

The aim of preprocess is extracting and editing the information of the xml files of the proteins.

Firstly, Uniprot database downloaded and the information of the subsections in the General annotation (Comments) was extracted.

Secondly, Pubmed references in the text were removed.

Finally, PubMed id’s and abstracts were parsed and saved.

**How to run Preprocess**

Step by step operation:

1. Download the Uniprot xml database: <https://www.uniprot.org/help/downloads>, unzip and place the your folder
2. Install dependencies(given below)
3. Edit the configuration file (Necessary adjustments (for example, giving the file location)
4. Run module main function i.e., python **preprocess\_main.py**

**Dependencies**

1. Python 3.7.3
2. Biopython 1.77 or greater