##### Downloading Pubmed

Downloaded the Pubmed dataset in XML format through bash script

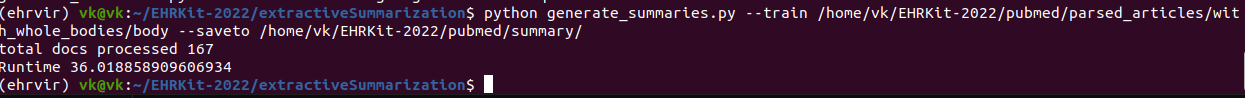
**Parsing Pubmed Articles**

From another directory, we parsed each article to 3 plaintext files: one containing its abstract, one containing its body, and one containing both. For the body, we will choose to parse either the whole body or just its introduction section. We have parsed for both the whole body and just introduction section and saved them as a separate folder. Up to 757123 articles can be parsed using their whole bodies, and 45,655 with just their intros.

**Generating Summaries**

**generate\_summaries.py** generates Lexrank summaries using this summarizers package for a directory of source documents.

* --train: Takes in the directory path containing training documents, which are used for calculating idf scores.
* --saveto: Takes in the directory path for saving the generated summaries (saved with the extension *name*.sum)
* --test (optional): Takes in the directory path containing documents for which to produce summaries. The default is the documents in the train directory.
* --ntrain (optional): First n number of documents to train on. Default is all documents.
* --ntest (optional): First n number of documents to produce summaries of. Default is all documents.
* --threshold (optional): Sets threshold for Lexrank algorithm. Default is 0.03.
* --size (optional): Sets size (number of sentences) of summaries produced. Default is 1.



**rouge\_scores.py** generates rouge score for the summaries produced using this summarizers package.

* --summaries: Takes in the directory path containing summaries in the format name.sum.
* --references: Takes in the directory path for the references of the summaries produced in the format name.tgt. Each tgt file must have same number of newlines as corresponding .sum file.
* --saveto (optional): Takes in the file path for saving the ROUGE score results.

