XML Validation, XPath Retrieval, and XML Manipulation with DOM

XML Validation:

- Validated the initial XML file „**islamqa_10thousand_500.xml**“ using multiple techniques:
  - Employed **xmllint** tool for validating with both external and internal DTD.
  - Used **xmllint** for XML Schema validation.
  - Utilized **Jing tool** for RelaxNG validation.
- Ensured data integrity and adherence to defined rules through validation.

XML Manipulation with DOM:

- Employed DOM parser in Python to manipulate the XML file ‚**islamqa_10thousand_500.xml**‘:
  - Sorted elements based on answer length in increasing order.
  - Added the "answer_length" node under the 'answer' element.
  - Added the "References" node to include Hadith books and "Categories" node to provide insights on the topic of the question and answer nodes within the 'entry' element.
  - After XML manipulation, the new XML file ‚**islamqa_10thousand_500_DOM.xml**‘ was re-validated using RelaxNG to ensure compliance with the defined schema.

Validating the final xml file:

- To check the validity of the final version of xml file with embeddings ‚**islamqa_final_1500.xml‘**, **xmllint** tool was used for XML Schema validation.

XPath Retrieval and Dataset Statistics on final xml file ‚islamqa_final_1500.xml‘ :

- Utilized XPath to access the XML file and retrieve essential data outputs. For example, XML titles, questions, answers containing a specific word for targeted searches.
- Gathered dataset statistics and identified patterns using XPath queries, for example count of reference books/theologians, count of a specific word mentioned in the XML file or gather nodes mentioning the intersection of selected words