Question Answering System

Team :

1. Bhavana Talluri (S20170010025)
2. P . Sai Jahnavi Shanvitha (S20170010107)

Description:

An IR system built to retrieve information in the domains of authors and books (literature domain). We currently spanned the author and book infoboxes from wikimedia dumps and temporarily stored them in the form of tuple for each entity of infobox.This data is later converted into xml format through a python script. This xml file is then used to upload the data on Solr in the respective cores for both authors and books.

Now another scipt is written to take query in the form of question from the user and this is preprocessed to remove stop words, wh words, and other less important words(using nlp libraries).The other important words from the query are captured into a list. These words are passed to solr(running in localhost) through client url (cURL). The reponse obtained is first subjected to byte decoding (because of json loads) and then converted back to list .This information is displayed to the user.

Technologies Used:

SAX parser, Solr- 8.3.1(Querying, storing and searching), cURL, nltk library, python

Procedure to run the code:

The jupyter notebook (Wikimedia.ipynb) contains all the functionality of the system and the code has comments explaining the functionality of each snippet.\

- Code for parsing wikimedia dump, conversion to xml file and uploading it to solr need not be run, (unless until we would like to update data) as we have already stored the xml files and uploaded data to Solr.
- The import statements in the beginning needs to be run and also all the cells under the heading " Question Answering system" to see the functionality of the system
- A sample of 5 queries and their answers are already shown in the output
- The nltk part (one of NLP libraries) used takes a minute or two to process and return a refined query.
- The search results are obtained within 2 seconds.

Challenges to be adressed:

- The data used in the project has been obtained after parsing 1011 authors and 1065 authors from the wikimedia dumps.Entire data needs to be parsed to run the code over a larger collection to obtain complete results (We couldnt parse complete collection due to lack of computational resources)
- The synonyms for the attributes present in the infoboxes haven't been accounted for. As a result, the query should contain the words which exactly match the attribute names in the infobox.