Automatic Acronym Extraction

Team extracTLA

Pavan, Guru Subramanian, Krishnakanth

Problem statement

- While reading technical journals, we find ourselves getting stumbled across acronyms
- And let’s be honest many a time we don’t know what the full form is and it is even frustrating sometimes when there is no full form adjacent to it

Proposed Solution

- To make a program which takes an input document and gets all the acronyms and their full forms, if given in the document

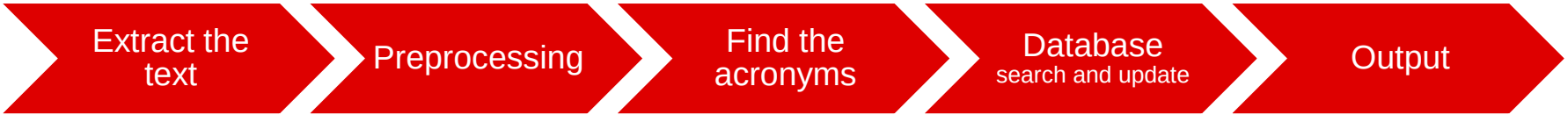
- To maintain a database of acronyms and their full forms

- To interface the database with the program to search for full forms and updating the database with new acronyms-full forms

Functionalities of the Python program

- The program takes input as a website URL (or) PDF URL (or) path to a local PDF file and extracts the text from them
- The program comes bundled with a database of all the acronyms it has encountered since the time the program was written
- The acronyms are stored in a Unix database, to enable interfacing with other languages/scripts
- Preprocess the text and identify acronyms and their full-forms, if expanded at least once around the acronym in the text
- Search the database to find missing full-forms
- Output all the acronyms whose full forms have been found
- Saves a new copy of the document with full forms appended after every acronym (Yes! It defeats the purpose of using acronyms, but is convenient for newbies)

Flow/Algorithm



- URL / PDF URL / local PDF path
- **urllib** to download PDF from URL
- **PyPDF** to read text from PDF
- **BeautifulSoup** to get rendered text from webpage

- Regular expression to remove punctuation and escape characters

- Iterate the words in the text
- Identify acronyms
- Search for full forms in 20 words around the particular acronym

- **dbm** to manage the Unix database
- Search DB for un-expanded acronyms
- Update DB with found acronym-full forms

- Print the acronym list and found full forms
- Save an updated doc with full forms at all occurrences

References

- Download PDF from URL using urllib
- Read text from PDF file using PyPDF
- Text from Webpage using BeautifulSoup
- Remove punctuation using regex

Thank You!