Automatic Acronym Extraction

Team extracTLA - *Pavan, Guru Subramanian, Krishnakanth*

Problem Statement

While reading technical journals, we find ourselves getting stumbled across acronyms. To make a program which takes an input document and gets all the acronyms and their full forms and maintains a database of acronyms

Implementation Details

1. Text Extraction

   The program supports three types of inputs for finding the acronyms:

   - Path to a local PDF file
   - URL of a PDF on web (https)
   - URL of webpage (https)

   For web PDF, the PDF file is downloaded first and deleted after text extraction. urllib has been used to download the PDF file

   For extracting text from a local PDF or the downloaded PDF, pdfplumber module has been used for text extraction. pdfplumber has been chosen instead of PyPDF for its failure in some of the test cases.

   For webpages, BeautifulSoup has been used to first get parsed HTML and then ignore the text in tags like meta, head, comments, script, style, documents.

   In all the above inputs the text should be in text format and not as an image.

2. Algorithm

   The extracted text is first preprocessed to replace punctuation with spaces using regex. This will help overcome having conditions for parentheses, full stops and other punctuations near the acronyms and their full forms

   The processed text is then read word by word to find acronyms, words with all caps and an exception of ‘s’ allowed as the last character.

   Whenever such acronyms are found, the first letters of 10 words on either side of the acronym are considered. The case of the letters is ignored in this step. The first found match of the first letters, with the acronyms, is taken as the full forms of that acronym

   Unexpanded acronyms are stored to later search in the DB

3. DBMS

   The python’s dbm module, which can maintain Unix DB’s is used to for the DBMS. It has a key-value dictionary type structure which is the best suited in acronym finding.

Using this Unix DB opens the window to interface the database with other languages/scripts.

The keys and values are stored as bytes, with keys being a fixed number of bytes.

The lib provides a get method to get the value of a particular key. The assert method can be used to assign a value to a particular key

It is assumed that each acronyms has only one expansion possible

4. Output and updating DB

   Previously stored unexpanded acronyms are searched for in the DB.

   All the acronyms whose full forms are found are printed. Other unexpanded acronyms in the text are also printed

   The database is updated with new acronyms-full forms found in the text.

Testing

IEEE paper : https://ieeexplore.ieee.org/document/7112806

Known Bugs

- One full form for one acronym is assumed
- Words like of, it, the, … are not ignored in the full form

Results and Conclusions

- Found some reliable sources for python library usage
- Collaborating on a project and integrating individual codes

References

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