Keyword Extraction

Extraction of Java Keywords from PDF

Konijeti Koteswara Nikhil,
Amrita University,
Amritpuri
Kerala,690525.

Overview:

Given a PDF containing Java notes, my task was to extract Java Keywords from the PDF, I have used Natural Language Processing and Text Mining to complete this task successfully.

Goals:

1. My goal was to extract the keywords(like Inheritance, encapsulation, multithreading) from the document and mentioning the keywords in order of their weightages in an excel sheet.

Specifications:

I have used some of the Natural Language Processing methods and packages like word_tokenize,textract,stem etc and some other python methods which included split,lower,remove and Collections package.

ExcelWriter

This was the method which was the last and the crucial step which converts the Data frame into an Excel sheet.

Milestones:

I. Extracting text from the given PDF

Using textract package to extract text from the given PDF was kind of tough as it contained several methods(tesseract) and had to use them when needed. And word_tokenize method which converts text into individual tokens.I have used them successfully which resulted the desired output.

II. Identification of keywords and converting my Data Frame into an Excel Sheet

There were many keywords extracted but had to find the keywords which are relating to Java and finally adding the columns(keywords and no.of times the keyword has occurred) to the Data frame and converting the Data frame into an Excel sheet using ExcelWriter.

Conclusion:

These were some of the milestones which I had to reach and the hurdles which I had to cross.Last but not the least I had fun doing this project.