Python

Production of programs that analyze Korean PDF and extract keywords

progress report **#1**

Date : 2023.11.03

Name : Haegeon Lee

ID : 183014

1. Introduction

1) Background

With the advent of the information age, the era of reading and interpreting a large amount of documents has arrived. If the amount of documents is vast when a particular document is first encountered, fatigue accumulates before reading. Therefore, if the Python program provides keywords that readers should focus on first, the speed increases rapidly and more efficient reading is possible. Therefore, the existence of these programs is necessary.

2) Project goal

It aims to develop a program that analyzes pdf to analyze what key keywords are based on the frequency of words and provides them to users.

3) Differences from existing programs

Existing programs are produced based on English, so there is a big difference in reading Korean pdf and it is impossible to use. Therefore, there is a difference from existing programs because we focus on extracting keywords from Korean pdf file for Koreans.

2. Functional Requirement

1) Function 1 – Extract text from file

- Extract the text file from the pdf file and store it in the list.

2) Function 2

- Remove investigations such as '은','는','이','가' from list tocken. And import Morpheme analyzer to make up a list with only nouns. We will also add the ability to remove words that are frequently used but are not important.

3) Function 3

- Saves the frequency of nouns in dictionary form and outputs the most written value.

4) Function 4

- We plan to save the result value as a txt file so that it can be used in the future.

3. progress

1) Implementation of a feature

(1) Extract text from file

- In

Pdf file

-out

Words (variable / List of words separated by spaces)

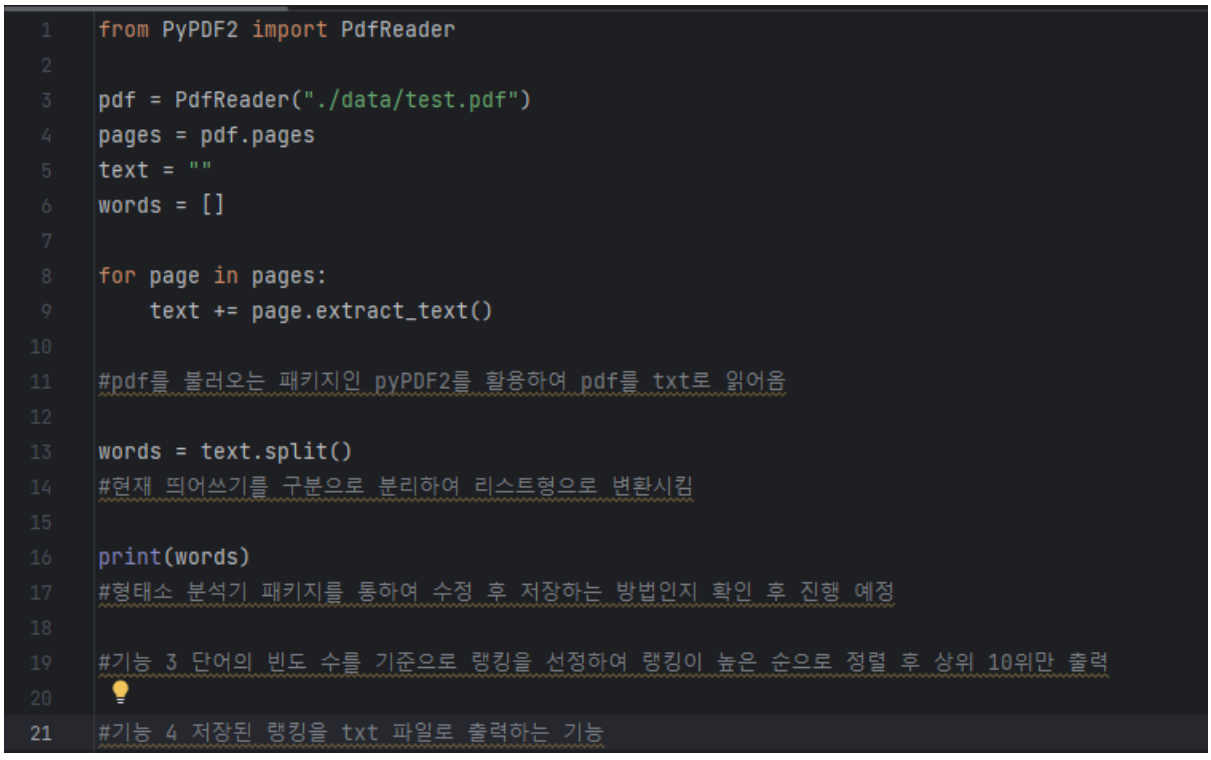
- Explanation

Enter the pdf file as Python and save as a list what is imported into txt as a function inside the package.

- applied learning

For loop, package, file inout

- screen shot



```
from PyPDF2 import PdfReader

pdf = PdfReader("./data/test.pdf")
pages = pdf.pages
text = ""
words = []

for page in pages:
    text += page.extract_text()

#pdf를 불러오는 패키지인 pyPDF2를 활용하여 pdf를 txt로 읽어옴

words = text.split()
#현재 띄어쓰기를 구분으로 분리하여 리스트형으로 변환시킴

print(words)
#형태소 분석기 패키지를 통하여 수정 후 저장하는 방법인지 확인 후 진행 예정

#기능 3 단어의 빈도 수를 기준으로 랭킹을 선정하여 랭킹이 높은 순으로 정렬 후 상위 10위만 출력

#기능 4 저장된 랭킹을 txt 파일로 출력하는 기능
```

2) Test Result

(1) Feature Name Tested

- Explanation

Enter the pdf file as Python and save as a list what is imported into txt as a function inside the package.

- Test Result screen shot

```
C:\Users\haegun2\PycharmProjects\SW1\venv\Scripts\python.exe "C:\Users\haegun2\PycharmProje
['-', '1', '-', '보도참고자료', '배포', '일시', '2023.', '3.', '8.(수)', '담당', '부서공항정책관',

종료 코드 0(으)로 완료된 프로세스
```

```
'담당자사무관', '이양구(044-201-4138)', '주무관', '임태호(044-201-4589)', '보도일시', '3.', '8.(수)',
```

4. Changes to plan

1) Change History Subject

- before

Remove investigations such as '은','는','이','가' from list tocken. And import Morpheme analyzer to make up a list with only nouns.

- after

i will also add the ability to remove words that are frequently used but are not important. And I plan to save the result value as a txt file so that it can be used in the future.

- reason

often used, and if it is idiomatic or repeatedly used for words needed to construct sentences, it is in the upper ranks.

5. Schedule

<table>
<tr><th></th><th>11/3</th><th>11/10</th><th>11/26</th><th>12/17</th><th>12/18</th><th>12/20</th><th>12/23</th></tr>
<tr><td>Write Proposal</td><td>Finished</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Function 1</td><td></td><td colspan="3">75% finished --></td><td></td><td></td><td></td></tr>
<tr><td>Function 2</td><td></td><td></td><td colspan="2">------></td><td></td><td></td><td></td></tr>
<tr><td>Function 3</td><td></td><td></td><td></td><td></td><td colspan="2">------></td><td></td></tr>
<tr><td>Function 4</td><td></td><td></td><td></td><td></td><td></td><td colspan="2">------></td></tr>
</table>