|  |
| --- |
| **Python**  **Production of programs that analyze Korean PDF and extract keywords**  **Final Report** |

|  |  |
| --- | --- |
|  | Date : 2023.12.24  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 non-critical words

- 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 -** ranking based on frequency to output the top 15

- 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. Implementation**

**(1)** Extract text from file

- In

Pdf file

-out

Text(Words / variable)

- Explanation

After receiving pdf, import the txt file on a page-by-page basis from pdf and save it in text(variable).

- applied learning

For loop, package, file inout, function

- screen shot

텍스트, 폰트, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

**(2)** Remove non-critical words

- Explanation

Other are excluded and only alphabetic or numeric words are added to the word list.

Gets the list of words that are not important.  
Put the rest in the list called yes\_words, except for the list of words mixed with numbers and non-important words.

- applied learning

for loop, list comprehension

- screen shot

스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 멀티미디어 소프트웨어, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**(3)** ranking based on frequency to output the top 15

- In

Yes\_words

-out

1~15 등 : (단어) n회

- Explanation

Find the frequency of words from the yes\_words list and add them to the dictionary.

After sorting by frequency, it outputs up to the top 15 according to the format.

- applied learning

Dictionary, sorted, for loop

- screen shot

텍스트, 스크린샷, 소프트웨어, 운영 체제이(가) 표시된 사진

자동 생성된 설명

**(4)** save ranking to txt file

- In

Sorted\_word\_rank

-out

순위 빈도수 단어

N n n

N n n

“단어랭킹이 word\_ranking.txt 파일로 저장되었습니다.”

- Explanation

The content of saving the results to a txt file for future use. The file to be saved is specified, and the sorted ranking list is taken and each content is sequentially stored in the txt file through the for loop.

- applied learning

File in out, dictionary, print, for loop

- screen shot

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

**4. Test Result**

**(1)**

- Explanation

Read the letter as txt from the input pdf, remove insignificant words, determine the frequency, and output to the top 15 depending on the format.

- Test Result screen shot

**텍스트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명**

**#** Example of pdf used for testing

**텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명**

**#**Examples of saved text files

**스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명**

**5. 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

Write a code that removes words and investigations that you don't use directly without importing morpheme analyzers.

- reason

I was thinking of importing morpheme analyzers. However, since the analyzer to be imported works using the java environment, we found that it is possible only if a specific environment is established. There was a problem that it was not easily and conveniently available on any computer. Therefore, we didn't import the content, but we directly configured it to be simple and maintained.

In that case, the removal of frequently used but insignificant words can also be maintained from time to time.

**6. Lessons Learned & Feedback**

파이썬으로는 키오스크를 만들어서 프로젝트를 진행해야겠다고 생각하고 기능 등을 구상하고 있었는데 중간고사 이후 파일 관련 내용이 꼭 들어가야하는 단서조항이 생겨서 급선회 하면서 시행착오가 많이 있었습니다.

제가 생각하고 오 기발한데 했던 모든것들이 파이썬 라이브러리로 기본적으로 지원하는게 많아서 좌절감을 많이 느꼈습니다. 이번 프로젝트만 하더라도 원래 형태소 분석기를 통하여 형태소만 남기고 추출을 해주는 라이브러리가 있었는데 자바 환경도 깔려있어야 구동된다는 홈페이지 설명을 읽고 이참에 라이브러리 의존 없이 내가 만들어 보자! 생각을 하고 해당 형태소 분석기 내용은 제거한 후 수작업으로 조사나 불용어를 제거하는 내용을 작성하는 쪽으로 방향을 돌려 진행하였습니다. 다만 언어 분야 특성상 너무 방대한 조사나 불용어가 존재하여 그때그때 랭킹을 확인하고 불용어 같은 경우를 no\_word 리스트에 추가하면서 랭킹을 재 출력하는 번거로움이 붙을 수 있는 내용이지만 직접 구축하려고 시도했다는 점을 높게 사주시면 감사하겠습니다.

다만 라이브러리 의존 없이 진행하려고 하니 계속 여러군대에서 막히다 보니 결과적으로 대화형 인공지능에 의존을 하게 되는 아이러니함이 조금 생겼는데 그래도 제 실력으로 만드려고 최대한 노력하였고 인공지능이 만들어준 부분도 제가 읽고 수정하고 ppt를 확인하면서 복습하고 코드에 주석을 달며 여러 번을 고심한 끝에 만들면서 공부했습니다.

한학기동안 열정적인 모습으로 강의해 주셔서 진심으로 감동받았습니다. 추후 일반선택 학점이 남거나 제가 희망하고 있는 지능실감융합전공 복수전공의 전공필수 학점이 인공지능학부의 교과목과 극소수 일부 겹치는 것으로 알고 있는데 교수님 수업과 겹친다면 꼭 다시 뵙고 싶습니다! 감사합니다.