Python Programming and Practice

Development daily news keyword summary program

Progress Report: 2

Date: 2023.12.10

Name: 지연우

ID: 224491

1. Introduction

1) Background

It's time consuming and very cumbersome to check out all the tons of news pouring

out every day. But we can't completely stop paying attention to current events, so

we need a more efficient way to get information. To solve this problem, a program

that can easily check current events information in short sentences or keywords is

needed.

2) Project goal

Aim to create a program that summarizes articles uploaded every day in short

sentences or words.

3) Differences from existing programs

The article summary function previously provided by Naver News only summarizes

each article. This, of course, helps save time, but it is difficult to easily find out

various current events information. We analyze all articles to show the user a few

key sentences or keywords, making it easy to get current events information.

2. Functional Requirement

1) Collect articles

- function to collect articles uploaded daily

(1) Collect information on popular articles by media company

- Crawl the titles, body text of the top five popular articles of each media company

2) Summarize the articles

- Summarize each article in short sentences and words.

(1) Create article summary

- Summarize the article in one sentence.

(2) Extract keywords

- Extract one keyword of articles.

3) Show keywords

- Show keywords to user.

(1) Sort keywords in order of frequency

- Show the keywords that appear in more articles first.

4) Show summary of articles by keyword

- When a user selects a keyword, it shows a summary of the article in which that keyword appears.

3. Progress

1) Function Implementation

(1) Collect articles

- Input: Link to Naver News Ranking page
- Output: TSV file containing information of the articles
- Description: Crawl the titles, body text and url of the top five popular articles from each media company.
- Applied concepts: loops, functions, string manipulation, exception handling, file write,
 __name__ == "__main__".
- Code Screenshot

```
# 넥이버 뉴스의 현장뉴스 펜이지에서 연론사발 조회수 상위 5개 기사의 정보를 수집하는 함수
# 각 기시에 대형 crawl_article 함수를 사용하여 기사 정보를 열어오고, tsv파일에 저장
2usages **Yearwood》=1

def crawl_rankings(file_name):
    rankings_link = "https://news.naver.com/main/ranking/popularDay.naver"
    soup = get_soup(rankings_link)
    rankingnes_boxes = soup.select(".rankingnews_box") # 연론사별 형칭 5위까지의 기사 정보가 담긴 백스를 선택

with open(file_name, 'w', encoding="utf8") as fp:
    fp.wmite("연론사물\t]\AMB\t]\AMB\t]\AMB\t]\AMB\t]\AMB\t]\AMB\t]

for box in rankingnews_boxes:
    company_name = box.select_one(".rankingnews_name").text
# 영어 기사는 제외하도록 명은 신문은 수집하지 않음
    if company_name in ("allowidge", "allowidge"):
    continue
    article = box.select("li")

for i, article in enumerate(articles):
    try:
    # 원크는 https://n.news.naver.com/article/연론사코드/기사변호?ntype=RANKGIN 형태
    article_data = crawl_article(article_link)
    fp.write(company_name + "\t" + \t".", "join(article_data) + "\n")
# 원제기준에 해당하는 기사가 없어 당칭 5위까지 기사가 존재하지 않을 때
    except:
    print(company_name, i+1, "번째 기사 수집 실패")

if __name__ == "__main__":
    file = f"./crawledData{str(datetime.today().date()).replace(_odc '-', __newc '')}.tsv"
    crawl_rankings(file)
```

(2) Summarize the articles

- Input: Crawled articles data
- Output: Summaries of the articles and a dictionary with a keyword extracted from a summary as a key and a list of indexes of the sentence in which the keyword appeared as a value
- Description: Summarizes articles using artificial intelligence models and calculates the frequency of keyword appearance by extracting nouns to be used as keywords from summary sentences
- Applied concepts: loops, functions, list comprehension, file read, file write, dictionary
- Code Screenshot

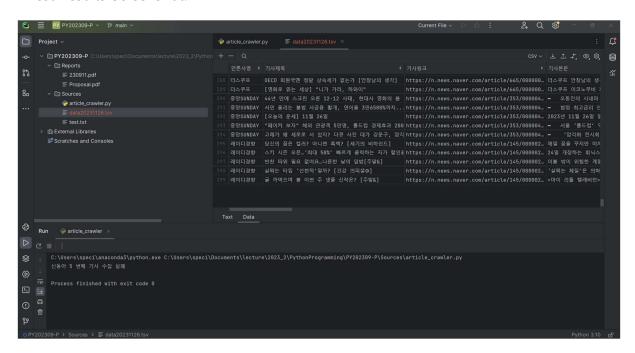
```
🥏 summarizer.py 🗵
       from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
      from konlpy.tag import Hannanum
      # 기사 10개씩 <u>나누어서 요약하도록</u> 설정
           model = AutoModelForSeq2SeqLM.from_pretrained('eenzeenee/t5-base-korean-summarization')
           tokenizer = AutoTokenizer.from_pretrained('eenzeenee/t5-base-korean-summarization')
               inputs = tokenizer(inputs, max_length=512, truncation=True, return_tensors="pt", padding=True)
output = model.generate(**inputs, num_beams=3, do_sample=True, min_length=10, max_length=100)
decoded_output = tokenizer.batch_decode(output, skip_special_tokens=True)
                result += decoded_output
      def get_keywords(summaries):
           keyword_dic = {}
           stop_words = ["등", "씨", "것"]
               words = set([word for word in hannanum.nouns(summary) if word not in stop_words])
                for word in words:
                     if word in keyword_dic:
                         keyword_dic[word].append(i)
                         keyword_dic[word] = [i]
```

2) Test Results

(1) Collect articles

- Description: Successfully crawled information from 399 articles across 80 media outlets on Naver News Ranking page. For 'Shindonga' (a specific media outlet), only articles up to the 4th rank were available, resulting in a failure message when attempting to collect the 5th article.

- Test Results Screenshot:



(2) Summarize the articles

- Description: The summary of the article's body works well, but in the case of keywords, contrary to expectations, many insignificant words appear. There is a need to improve the logic related to keywords.
- Test Results Screenshot:

"거의 다 먹었는데"...이가장 식당 급식 밥통서 나온 회색 수건: 강원도 한 스키장 식당의 단체급식용 밥통에서 사용하면 수건이 나와 이용격이 눈살을 찌푸리는 일이 발생했다.
''(채터리' 르빛나다.')응물 80%' 트럼프 돌아오면 발어질 일: 트럼프 전 미국 대통령이 미국 대선에서 재집권에 성공하면 보복 관세를 확대하고 친환경 에너지 지원책을 대거 페기하겠다고 선언했다.
[단독] 밀었던 대북병원의 배신..'가짜 아킬레스건 수술이 유통된 것으로 확인됐으며 전문
아침밥 정겨주고 오후 3시 참퇴.'울산을 3배' 기적 일어났다.' 이모주상사의 사내 합계특수출산을은 2021년 1.97명으로 뛰어 10년 만에 일본 3대 종합상사의 출산을이 세 배 넘게 뛴 현상을 일본 사회도 가'
"16 주식 미리 사물을 걸"...주가 목동하자 후회 쏟아졌다 [연계소문]: 16연터티인먼트가 그룹 불락핑크와 원전체 활동에 대한 제계약을 체결됐다고 밝혀자 업계 안밖에서 후회의 목소리가 나왔다.
"남성 돌과 동거, 키 크고 예됐다"...35년전 조지아주 피살 한인여성 신원 확인: SBS 그것이 알고 싶다가 35년 동안 미제로 남아있던 김정은 씨 피설사건의 전모를 파헤졌다. 한적한 마을의 쓰레기 수거함에서 '발거아내와 소대성' 성관계 영상 SNS 올린 군 간부 집유..왜: 의외부지법 형사3단독은 SNS에 성관계 영상을 올린 혐의로 기소된 전방지역 직업군인 14 씨에게 정역 1년에 집행유에 2년을 선고됐다.
수백 명 먹었는데...스키장서 밥 푸다 따라 나온 수건 논란: 강원도 내 한 스키장의 모식당에서 식사를 한 고격이 급식용 바르에서 수건을 발견해 논란이 되자 해당 업체는 같은 사례가 발생하지 않도록 조치하이준석 "신당, 1월 안에 매듭..이낙연, 같이 그리는 울탁리 돼야": 이 전 국민의림 대표는 용 대통령의 변화를 유원한 변수라고 말하여 신당 창당과 관련해 낮어도 1월 안에 매듭지을 수 있을 것이라고 밝혔다. 수능 수확 1등급 96.5%가 이과성...울해도 '문과침공' 여전: 2024학년도 대략수학능력시험에서 수학 1등급 학생 중 96.5 가 미적분을 선택한 이과 확성인 것으로 나타났으며 교육별 유물리가 두드리지면서 목대균 왕이다.였던 55세 여성이 '오징어게임 살사판' 상금 60억 따란 비결은: 댓물리스 오징어게임 더 결단지 우속자인 혈단은 서비이별 소에 출면한 뒤 두려움 없이 자신이 돼 삶을 해져 나가겠다고 세계 시작 수명생 홍인 거짓 광고, 업계 1~4위 학원 다 걸렸다. 입시업계 최대 학원인 메가스터디와 시대인재 등이 대학수학능력시험 수점생을 끌어들이기 위해 부당 광고를 한 혐의로 과정금 제재를 맞았다. 이들은 소국민의됨 영입인재 1호 "이 당이나 저 당이나 모두 퇴보 정치" 쓰소리: 국민의원이 인재 1호로 영입한 한 소이청소년과의원 원장은 육아 방법에 대해 조연하고 있다. 유류보 캡처 하 소아청소년과의원 원장이 내원 존간 이 당이나 저 당이나 모두 되고 경제 보석이 나왔다.
[단독] "SPC 불때 닷 매를 70% 급리" 단킨 점포 두 곳 손대소: 에스피씨그를 산하 비알코리아 제일 단킨 점주들은 본사의 위법평위로 매를이 70 이상 급략해 수억의 손해를 입었다고 주장하며 소송을 제기적으로 참고 96.5%가 이과성"... '문과 점공' 한국 서제진다: 2024학년도 수하에서 수확 명역 1등급의 이과상 독자지 전상이 심화됐다는 입시업계 본석이 나왔다.
[단독] "SPC 불때 닷 매를 70% 급리" 단킨 점포 두 곳 손대소: 에스피씨그를 산하 비알코리아 제일 단킨 전주들은 본사의 위법평위로 매를이 70 이상 급략해 수억의 손해를 입었다고 주장하여 소송을 제기적으로 참고하는 집안에 가장 기관을 기존한 부분의 상대 2명을 기존한 보다는 전체 모산에 가장 이 의적성 인기에 용치한 이라면 19점을 다면 기원이 경제을 보다는 전체 문화 이라면 의원인 기원 의원인 기원 용치와 이라면 의원인 기원 의원인 기원으로 가장 수입에 무료 기원으로 보다는 전체 가장 1점을 되었다고 보는 1점을 기관하는 구원이 가장 1점을 기관하는 기원이 의원에 가장 1점을 기원하는 기관 기원으로 1점을 기원으로 1점을 보다는 1점을 기원으로 1점을 기원으로 1

```
---- 키워드 목록 상위 10개 -----
현상: [3, 9, 14]
수: [8, 11, 13]
1등급: [9, 14, 17]
수학: [9, 14, 17]
이: [9, 14, 17]
이: [9, 11, 19]
대학수학능력시험: [9, 11, 17]
2024학년: [9, 14, 17]
수능: [11, 14, 17]
발생: [0, 7]
```

4. Changes in Comparison to the Plan

1) Removal of Comment Count-related Feature

- Previously: Comment count was collected in Function 1, and when showing article summaries in Function 4, they were displayed in descending order of comment count.
- Now: Comment count is no longer collected in Function 1. In Function 4, article summaries are displayed randomly.
- Reason: Comment count information was not present in the crawled HTML documents. Although considering using Selenium to collect comment counts was an option, it was decided to exclude this feature as it was not deemed crucial, and the time required for article collection became excessively long.

5. Schedule

- Indicating Progress

업무	11/3	11/17	12/1	12/15	12/22
Write a proposal	Complete				

Function 1	Complete		
Function 2	In progress		
Function 3		Not Started	
Function 4		Not Started	
Final Report			Not Started