혁신인재 펠로우쉽 장학 프로그램

★★★

대학원생 취업진로 지원 플랫폼 연구

발표일: 2022.09.16.

팀원: **김재경**, 맹준영, 오관석, 이지은, 허다운

CONTENTS

01
개요

과제 요약

주제 : 웹크롤링 기반의 대학원생 취업 및 진로 지원 플랫폼

- 웹크롤링 기반 대학원생 대상의 취업/진로 지원 DB 및 시스템 구축으로 본교의 경쟁력 제고
- 인공지능 기반 취업/진로 추천 시스템 개발을 통한 대학원생 개인 맞춤형 취업/진로 플랫폼 구축
- 학교 차원에서의 대학원생 전공과 연관된 관심 진로, 기업 및 취업 분야 데이터 확보를 통한 대학원 혁신 도모

연구배경

- **학부생 취업/진로와 대학원생 취업/진로의 차이점이 존재**
  - 학부생 대비 취업/진로 관련 커뮤니티 및 채용 공고 사이트에 대학원생의 취업/진로 정보가 부족함
  - 대학원생의 연구 분야의 세분화로 인해 전문성을 요구하는 맞춤형 취업/진로 가이드가 필요함
- **시대 흐름에 따라 다양하게 변화하는 기업들의 관심 사업 및 연구 동향**
  - 기업들이 다양한 신사업을 발굴, 연구하고자 함
  - 코로나 등으로 인해 빠르게 변화하는 시장에 알맞게 다각도로 변화하는 사업의 흐름 속에서 대학원생들의 취업/진로 설정에 있어 진행 연구와 관련이 높은 진출 가능한 기업에 대한 정보가 필요함
- **대학원생 취업/진로의 중요도에 비해 관련 연구 및 분석을 위한 데이터의 부족**
  - 취업 상담센터 또는 관련 기관에서의 개인 정보가 포함된 취업/진로 데이터 제공에 있어서 시간 및 비용 문제에 따른 라벨 데이터 수집 및 구축의 어려움
  - 자동 크롤링 및 정제된 데이터에 대한 분석 가능한 모델 구현에 대한 기술적인 한계

연구의 목적 및 내용 (1/3)

- **웹크롤링과 딥러닝 기반의 취업 시장 동향 분석 및 기업 추천 데이터베이스 정제 및 구축**

  - MK500 기준 상위 100개 기업을 기준으로 국내 대표 기업 선정 및 목록 구축

  - 선정된 기업과 '채용', '연구', '사업' 등의 키워드로 포털 사이트(네이버)로부터 관련 기사와 텍스트 및 이미지 형식의 채용 공고문을 크롤링한 후, 정제 과정을 거쳐 직무/전공 특화 단어사전 대규모 데이터베이스 구축

  - 이미지 형태의 채용 공고문의 경우, 광학 문자 인식(OCR) 알고리즘을 통해 텍스트로 변환

  - 단어사전 데이터베이스를 기반으로 수집한 기사들의 문장별 핵심 단어에 대한 레이블링 작업 진행 및 자동화 시스템 구축을 위한 딥러닝 기반 기사 문장별 유의미한 단어 추출 모델 개발

연구의 목적 및 내용 (2/3)

- **딥러닝 기반 언어 데이터의 수치 데이터로의 변환 및 기업 추천 시스템 개발**
  - 사전 학습된 NLP 딥러닝 모델(Sentence-BERT)를 활용하여 구축한 직무/전공 단어사전 데이터베이스 내의 모든 핵심 단어를 추천 시스템에 활용될 정량적 수치로의 변환
  - 입력 받은 사용자의 관심 직무/전공 혹은 진행 연구 관련 핵심 단어를 Sentence-BERT를 활용하여 임베딩
  - 임베딩한 입력 정보와 구축한 핵심 단어 간의 기계학습 기반 유사도 분석을 통해 상위 항목을 기준으로 추천 기업 목록 생성

연구의 목적 및 내용 (3/3)

- **구글 스칼라를 이용한 관심 분야 및 연구 키워드 정보 추출 자동화**
  - 학생들의 이메일 계정 정보만을 사용한 관심 분야 및 연구 키워드 정보 추출 자동화
  - 학생들의 이메일 주소를 구글 스칼라에 검색함으로써 저자로 참여한 논문리스트 조회 가능. 검색된 논문들의 키워드 추출을 통해 학생들의 관심 분야 및 연구 키워드 추출 자동화
  - 설문조사 응답과 같은 학생들로부터 얻는 직접적인 데이터가 없이도 충분히 관련 데이터 구축 가능
- **논문 추천 시스템 데이터를 활용한 연구 키워드 추출**
  - 논문 추천 시스템을 통해 구축된 학생들의 관심 분야 데이터 활용
  - 학생들이 직접 기재한 키워드를 사용한다는 점에서 구글 스칼라 데이터를 보완 가능

기대효과

- 고려대학교 대학원생을 대상으로 한 최초의 취업/진로 **지원 통합형 솔루션** 제공으로 학생들의 재학 만족도 증진 및 본교의 **취업 시장 경쟁력 제고**
- 웹크롤링 및 인공지능 기반의 데이터 정제 과정을 거친 **데이터 구축 자동화**를 통한 기존 데이터 구축의 문제점 해결
- 대규모 데이터의 확장 및 대중화 가능
- 대학원 취업/진로 추천 서비스를 진행하는 과정에서 수집한 데이터를 추가로 활용하여 **솔루션 모델의 성능 향상**
- 학생들의 관심 분야 및 연구 키워드에 정보를 얻기 위한 기존의 방법론은 설문 및 피드백에 의존하였지만, 학생 이메일과 구글 스칼라를 활용한 새로운 데이터 구축 방법을 사용하여 이러한 **정보들을 자동적으로 추출** 가능

02
데이터 구축

02 데이터 구축 - 웹크롤링(네이버 기사)

아이디어

- **목적**
  - 기업의 최신 연구 동향 및 인재상 정보를 수집하여 기업 관련 데이터베이스 구축
  - 기업명과 수집 키워드를 함께 검색하여 최근 2년(2020년 1월 1일~2022년 6월 30일) 동안의 네이버 기사를 크롤링
  - 크롤링을 통해 기업명, 수집 키워드, 언론사 URL, 기사 날짜, 기사 제목, 기사 내용을 수집하여 데이터프레임 형태로 구축

- **기업명**
  - MK500 지수 기준의 2021년도 Top-100 리스트에 포함되는 기업

- **수집 키워드**
  - 신사업 투자유치, 논문, 연구 채용, 석박사, 핵심인재
    → 목적에 적절한 총 5개의 키워드 선정

크롤링 과정 (1/4)

- 네이버 기사의 검색결과(언론사 URL) 크롤링
  - 네이버 기사의 URL 패턴 분석

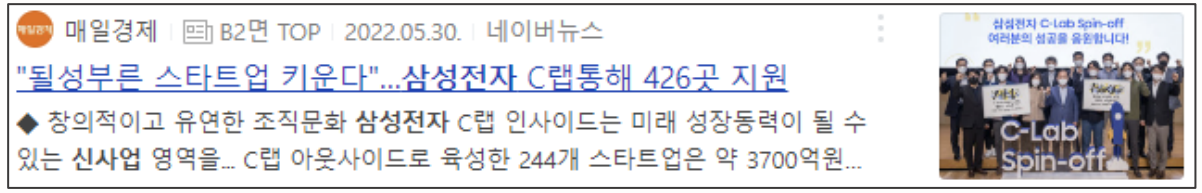
https://search.naver.com/search.naver?&where=news&query=**기업명 및 수집 키워드**&ds=**기간(시작일)**&de=**기간(종료일)**&nso=so:r,p:from**기간(시작일)**to**기간(시작일)**,a:all&start=**기사 순서(1, 11, 21, ...)**

https://search.naver.com/search.naver?&where=news&query=삼성전자 신사업 투자유치&ds=2020.01.01&de=2022.06.30&nso=so:r,p:from20200101to20220630,a:all&**start=1**

https://search.naver.com/search.naver?&where=news&query=삼성전자 신사업 투자유치&ds=2020.01.01&de=2022.06.30&nso=so:r,p:from20200101to20220630,a:all&**start=11**

크롤링 과정 (2/4)

- 네이버 기사의 검색결과(언론사 URL) 크롤링
  - 네이버 기사의 html 패턴 분석




```
<li class="bx" id="sp_nws1">
  <div class="news_wrap api_ani_send">  flex
    <div class="news_area"> == $0
      <div class="news_info">…</div>
      <a href="http://news.mk.co.kr/newsRead.php?no=474702&year=2022"
       class="news_tit" target="_blank" onclick="return goOtherCR(this, 'a
       =nws*e.tit&r=1&i=880000BC_0000000000000000004970727&g=009.0004970727
       &u='+urlencode(this.href));" title=""될성부른 스타트업 키운다"…삼성전
       자 C랩통해 426곳 지원">…</a>
      <div class="news_dsc">…</div>
    </div>
    <a href="http://news.mk.co.kr/newsRead.php?no=474702&year=2022"
     class="dsc_thumb " target="_blank" onclick="return goOtherCR(this, 'a
     =nws*e.img&r=1&i=880000BC_0000000000000000004970727&g=009.0004970727&u
     ='+urlencode(this.href));">…</a>
     ::after
  </div>
</li>
```

③ div.news_area > a

① body

```
<body class="wrap-new api_animation tabsch tabsch_news">
  <div id="nxtt_div" style="display:none;position:absolute;border-width:0;z-index:11000">
  </div>
  <div id="u_skip">…</div>
  <div id="wrap">
    <div id="header_wrap" role="heading" class="type_animation">…</div>
    <script type="text/javascript"> var nx_location_rcode = "09410111" ; </script>
    <div id="container" role="main">
      <div id="content" class="pack_group">
        <h1 class="blind">삼성전자 신사업 투자유치 뉴스검색 결과</h1>
        <div id="main_pack" class="main_pack">
          <script type="text/javascript"> var nx_cr_area_info = [{"n": "tab", "r": 1}];
          </script>
          <script type="text/javascript">…</script>
          <form id="news_form" name="news_form" action="?">…</form>
          <script type="text/javascript">…</script>
          <div id="snb">…</div>
          <script>…</script>
          <section class="sc_new sp_nnews _prs_nws">
            <div class="api_subject_bx">
              <div class="news_pick_area">…</div>
              <div class="group_news">
                <ul class="list_news"> == $0
                  <li class="bx" id="sp_nws1">…</li>
                  <li class="bx" id="sp_nws2">…</li>
                  <li class="bx" id="sp_nws5">…</li>
                  <li class="bx" id="sp_nws8">…</li>
                  <li class="bx" id="sp_nws9">…</li>
                  <li class="bx" id="sp_nws10">…</li>
                  <li class="bx" id="sp_nws11">…</li>
                  <li class="bx" id="sp_nws12">…</li>
                  <li class="bx" id="sp_nws13">…</li>
                  <li class="bx" id="sp_nws14">…</li>
                </ul>
```

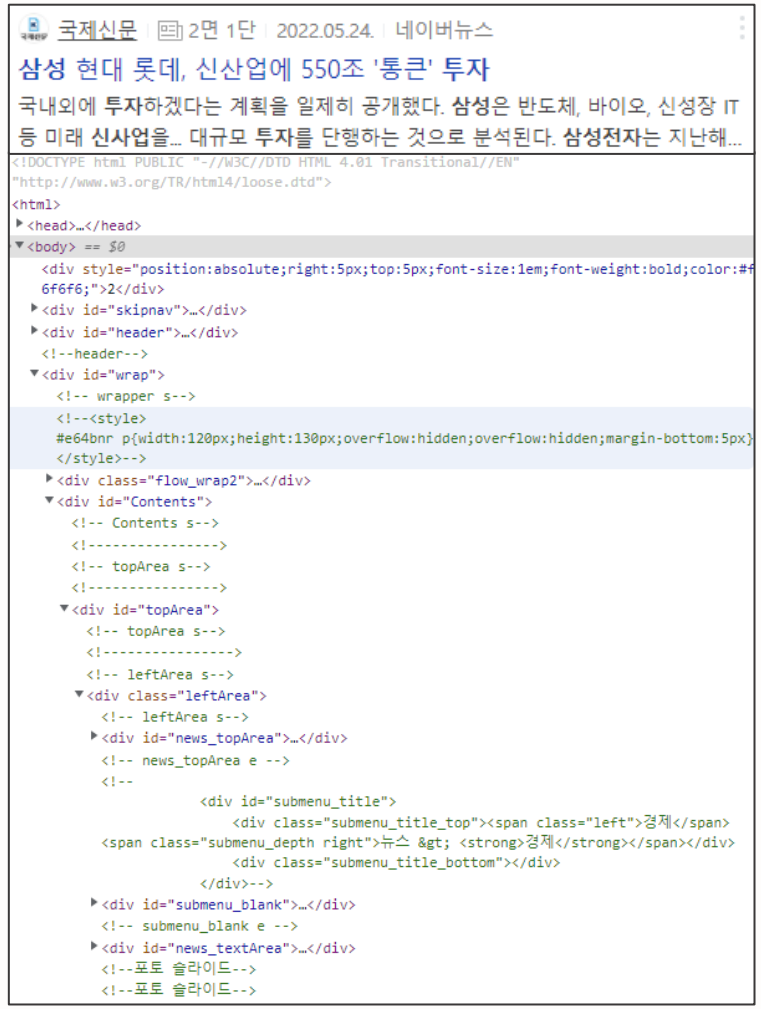
② ul.list_news > li

크롤링 과정 (3/4)

- 언론사별 기사 내용 크롤링
  - 언론사의 URL 호출 및 기사 내용 추출

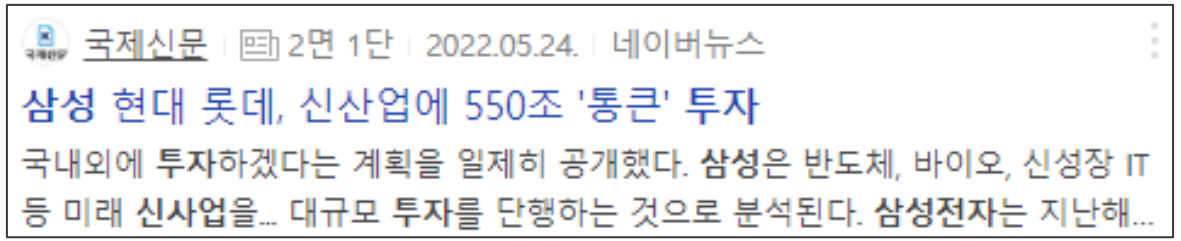
언론사별로
각기 다른 html 구조를
가진다는 문제점!!

문장 단위의 정제 과정을
거쳐 모든 텍스트를 크롤링



크롤링 과정 (4/4)

- 언론사별 기사 내용 크롤링
  - 언론사의 URL 호출 및 기사 내용 추출



"될성부른 스타트업 키운다"...**삼성전자** C랩통해 426곳 지원. **삼성전자**가 4월 29일 경기 수원 삼성디지털시티에서 박학규 사장(앞줄 왼쪽 넷째) 등 경영진과 창업자들이 참석한 가운데 `상반기 C랩 스핀오프 론칭데이`를 열었다. **삼성전자**는 국내 스타트업 생태계 활성화와 창업 지원을 위해 C랩(Creative Lab) 프로그램을 운영 중이다. ... C랩 아웃사이드는 스타트업이 성공적으로 성장할 수 있도록 기술 지원부터 **투자유치**까지 전폭적으로 지원하고, **삼성전자**와의 사업 협력 방안 모색 기회도 제공한다. ... **삼성전자**는 CSR 비전 '함께가요 미래로! Enabling People(가능성을 만들어가는 사람들)' 아래 C랩 아웃사이드, 삼성미래기술육성사업, 스마트공장, 협력회사 상생펀드 등 상생 활동과 청소년 교육 사회공헌 활동을 펼치고 있다.

삼성그룹 현대자동차그룹 롯데그룹은 윤석열 정부 출범에 맞춰 반도체, 바이오, 모빌리티 등 **신산업**에 향후 5년간 총 550조 원을 국내외에 투자하겠다는 계획을 일제히 공개했다. 조 바이든 미국 대통령이 윤석열 대통령과 지난 20일 경기도 평택시 **삼성전자** 반도체공장을 방문해 이재용 **삼성전자** 부회장의 안내를 받으며 공장을 시찰하고 있다. 삼성은 반도체, 바이오, 신성장 IT 등 미래 **신사업**을 중심으로 향후 5년간 450조 원(관계사 합산)을 투자할 계획이라고 14일 밝혔다. 삼성은 "사업의 성공이 연관 산업 발전과 국민소득 증대로 이어져 국가 경제 발전을 이끌어가는 '선순환 구조'를 구축할 것으로 기대한다"며 "삼성의 파운드리 사업이 세계 1위로 성장하면 **삼성전자**보다 큰 기업이 국내에 추가로 생기는 것과 비슷한 경제적 효과가 있다"고 설명했다. ...

02 데이터 구축 - 웹크롤링(네이버 기사)

크롤링 결과

- 100개의 기업 및 5개의 수집 키워드를 통해 기업 관련 데이터베이스 구축
  - 기업명, 수집 키워드, 언론사 URL, 기사 제목, 기사 날짜, 기사 내용을 데이터프레임 형태로 수집

기업명	수집 키워드	언론사 URL	기사 제목	기사 날짜	기사 내용
company	search	url	title	date	output
카카오뱅크	신사업 투자유치	http://www.investchosun.com/?p=326	[Invest]대기업 신사업 확대·유동성 과잉에...글로벌 PEF, 앞다	2021.05.21.	대기업 신사업 확대·유동성 과잉에...글로벌 PEF, 앞다
카카오뱅크	신사업 투자유치	http://theviewers.co.kr/View.aspx?No=	'통신 끌고 신사업 밀고' SK텔레콤·KT, 1분기 동반 어닝서프라	2021.05.11.	'통신 끌고 신사업 밀고' SK텔레콤·KT, 1분기 동
카카오뱅크	신사업 투자유치	http://www.startuptoday.co.kr/news/ar	통신 3사, 신사업 성장세로 '깜짝실적'...14분기만에 합산이익	2021.05.11.	통신 3사, 신사업 성장세로 '깜짝실적'...14분기만에 합
카카오뱅크	신사업 투자유치	http://www.seoulwire.com/news/article	자본 확충나선 카뱅·케뱅 "건전성·신사업 잡는다"	2020.11.02.	자본 확충나선 카뱅·케뱅 "건전성·신사업 잡는다" <
카카오뱅크	신사업 투자유치	http://www.wsobi.com/news/articleVie	카카오 1분기 영업이익 882억원 달성...커머스·콘텐츠 '신사업'	2020.05.07.	카카오 1분기 영업이익 882억원 달성...커머스·콘텐츠
카카오뱅크	논문	http://news.mk.co.kr/newsRead.php?nc	네이버·카카오 뭇지 않네...K스타트업도 '세계적 논문' 속속 입	2022.02.14.	네이버·카카오 뭇지 않네...K스타트업도 `세계적 논문`
카카오뱅크	논문	http://www.aitimes.com/news/articleVi	논문-수업 모두 잡은 IEEE의 젊은과학자상 수상자...KAIST 서창	2021.07.26.	논문-수업 모두 잡은 IEEE의 젊은과학자상 수상자...KA
삼성전자	연구 채용	http://www.newspim.com/news/view/2	삼성전자, 반도체 초격차 위해 여섯 번째 연구시설 신축 검토	2022.06.29.	삼성전자, 반도체 초격차 위해 여섯 번째 연구시설 신
삼성전자	연구 채용	https://www.chosun.com/economy/mc	대만 TSMC의 박사 연봉 얼마길래... 삼성전자 직원들 깜짝 [왕	2022.06.21.	삼성전자 직원들 깜짝 [왕개미연구소] - 조선일보.
삼성전자	연구 채용	https://www.news1.kr/articles/?470574(	RE100 가입 저울질하는 삼성, '안전환경' 연구소 개편·인재 확	2022.06.09.	RE100 가입 저울질하는 삼성, '안전환경' 연구소 개편·
삼성전자	연구 채용	https://misaeng.chosun.com/site/data/	2026년까지 34만명 채용...대기업 취업문 열린다	2022.06.02.	2026년까지 34만명 채용...대기업 취업문 열린다 - 1등

선정된 기업으로부터 핵심 키워드 기반 기사 크롤링 결과

데이터 정제

- **딥러닝 기반 논문 초록을 통한 연구 키워드 추출**
  - 번역한 논문 초록에서 명사만 추출 후, 한글에 특화된 Sentence-BERT 모델을 활용한 키워드 추출 진행
  - N-gram 형태의 텍스트를 사용하여 키워드 후보군 설정
    - N-gram의 n 값은 2~3으로 설정. 즉, 2~3개의 단어 조합으로 구성됨
  - MMR(Maximum Limit Relegance)을 사용한 **전체 문장**과 후보 키워드 중 유의미한 키워드 추출

```
[ ]  mmr(doc_embedding, candidate_embeddings, candidates, top_n=5, diversity=0.1)

     ['인천 연수구 송도동', '인천 연수구 바이오', '올해 촬영 인천', '연수구 송도동 바이오', '연수구 송도동']

[ ]  # 상대적으로 높은 diversity값은 다양한 키워드 5개
     mmr(doc_embedding, candidate_embeddings, candidates, top_n=5, diversity=0.8)

     ['인천 연수구 송도동', '바이오 뉴스 기사', '대학생 취업 준비', '현장 탐방 제약', '영업 이익 기록']
```

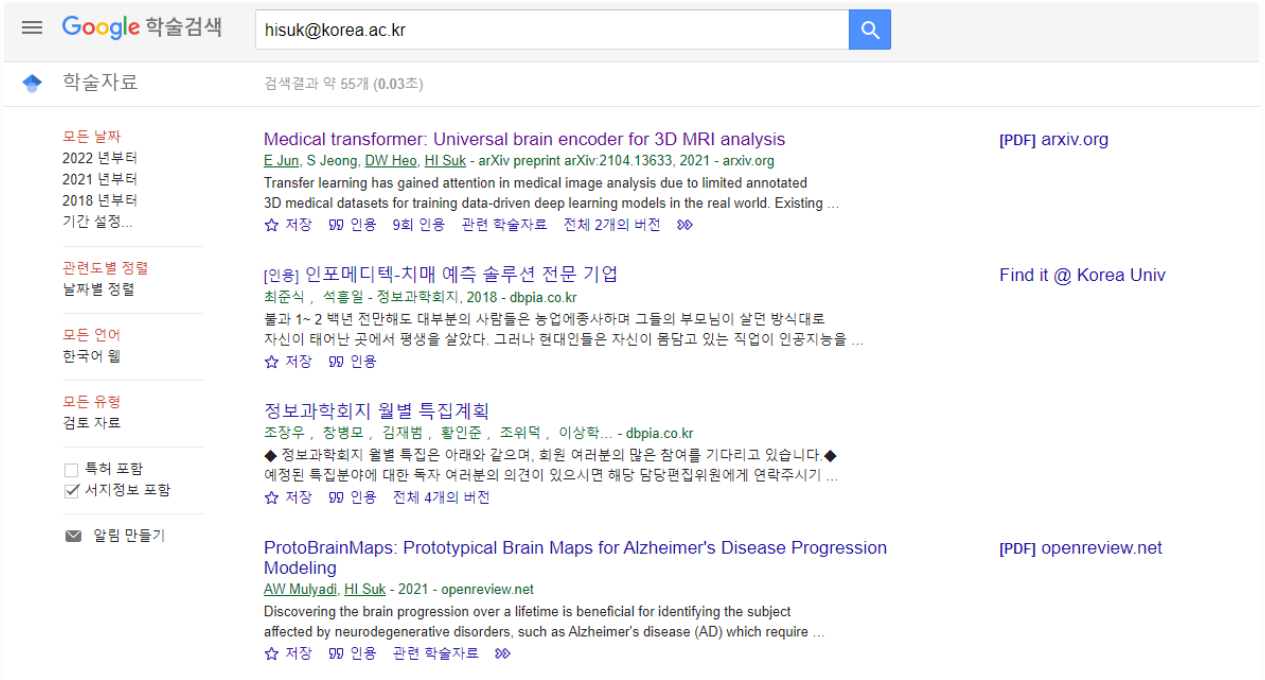
딥러닝 기반 논문 초록을 통한 연구 키워드 추출 결과 예시

02 데이터 구축 - 웹크롤링(구글 스칼라)

아이디어

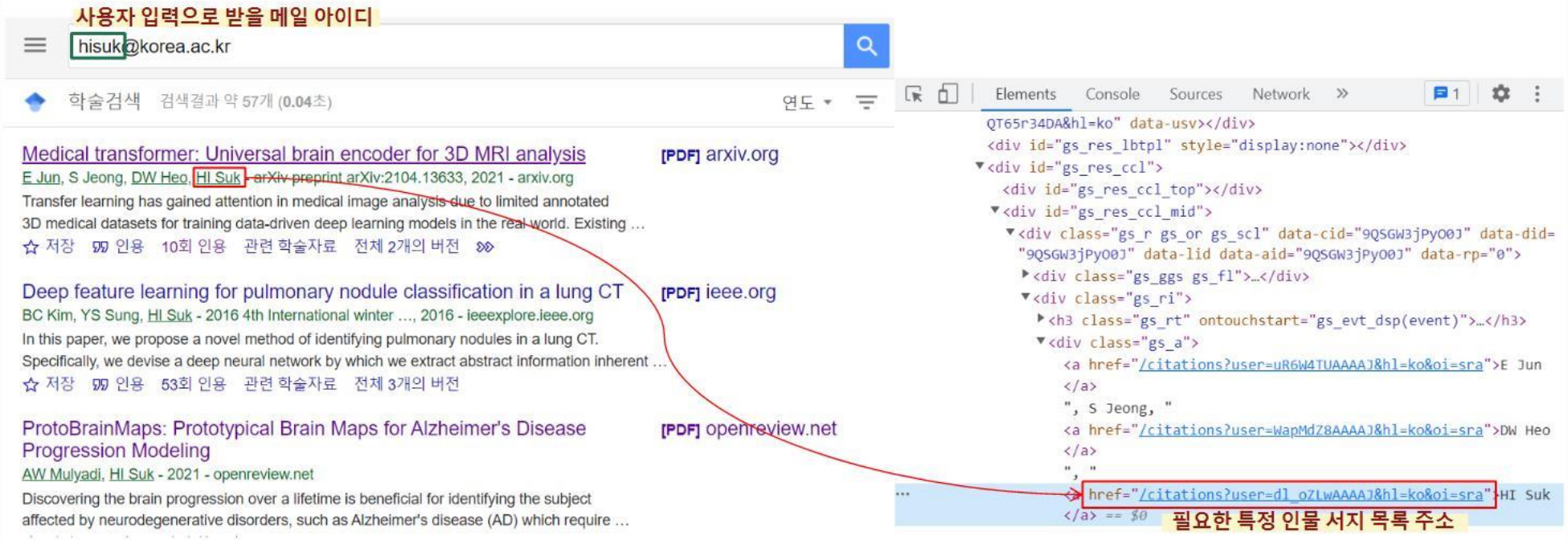
- 학생들의 설문조사 응답 없이도 **관심 분야 및 연구 키워드에 대한 정보를 자동으로 얻는 방법에 대한 필요성**이 크지만 학교에서 공식적으로 접근 가능한 논문 데이터는 졸업 시점에 국한된다는 한계를 가짐

- **학생들의 이메일 주소를 구글 스칼라에 검색**함으로써 저자로 참여한 논문리스트 조회 가능. 검색된 논문들의 키워드 추출을 통해 학생들의 관심 분야 및 연구 키워드 추출 자동화



크롤링 과정 (1/7)

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 이메일 입력을 통한 서지 목록 URL 추출
    - 특정 사용자의 서지 목록 URL 추출 내용 및 방법 예시



추출하고자 하는 특정 인물 서지 목록 주소 추출 방법 예시

크롤링 과정 (2/7)

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 이메일 입력을 통한 서지 목록 URL 추출
    - 관련 코드(1/2)

```
## Google 학술 검색 중 특정 인물 논문 list를 보여주는 URL 찾기 위한 코드
if student:
  # 학생인 경우 URL 경로
  tmp_url = 'https://scholar.google.com/scholar?hl=ko&as_sdt=0%2C5&q=' + subject_id + '%40' + school_mail + '&btnG='
else:
  # 교수인 경우 URL 경로
  tmp_url = 'https://scholar.google.com/scholar?start=50&q='+subject_id+'%40'+school_mail+'&hl=ko&as_sdt=0,5'

# 크롤링 진행하는 컴퓨터의 고유 header
headers = {'User-Agent' : '........'}
tmp_req = requests.get(tmp_url, headers=headers)
tmp_sbj_soup = BeautifulSoup(tmp_req.content, 'html.parser')
tmp_mail_list = tmp_sbj_soup.select('div.gs_a > a')
```

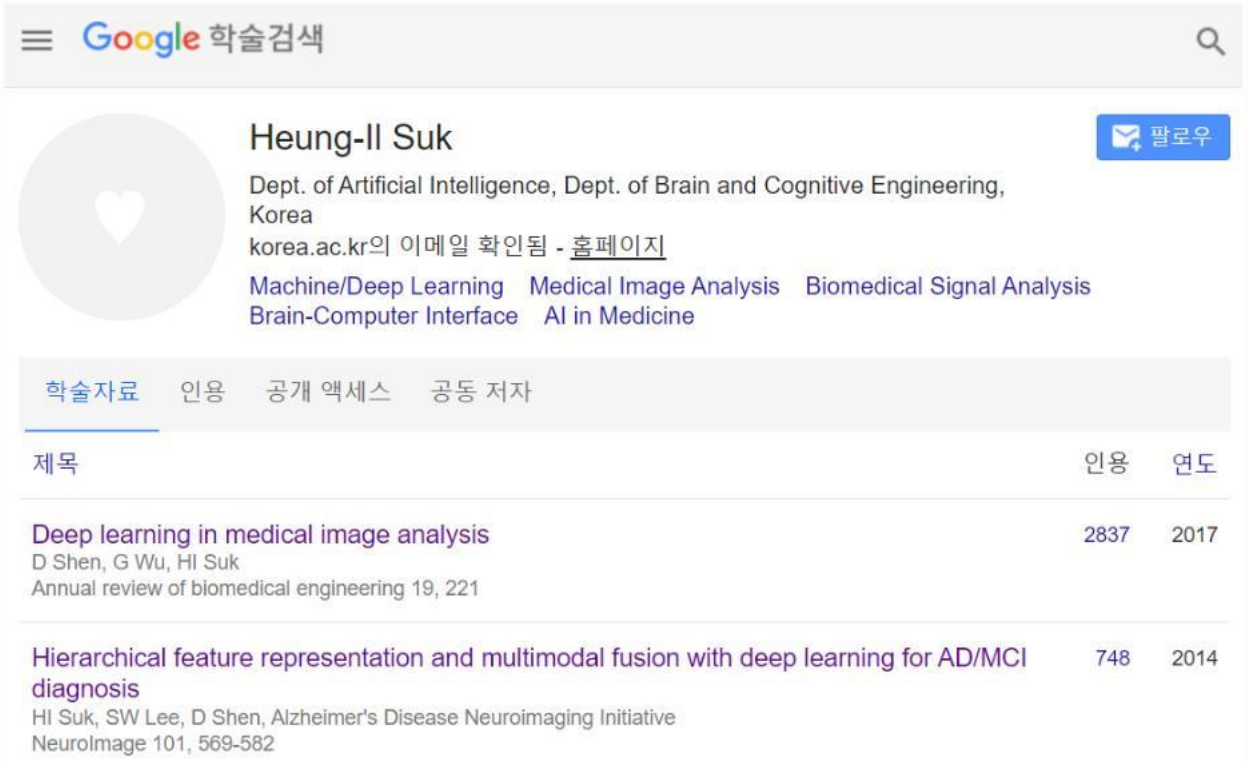
크롤링 과정 (3/7)

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 이메일 입력을 통한 서지 목록 URL 추출
    - 관련 코드(2/2)

```
# 검색된 이름 갯수 기반, 가장 많이 이름이 나온 인물이 우리가 찾고자 하는 인물일 것이라 가정
if len(tmp_mail_list) == 0:
  subject_google_scholar = tmp_sbj_soup.find_all('a', href=True)[6].get('href')
else:
  name_info = []
  for i in range(len(tmp_mail_list)):
    name_info.append(str(tmp_mail_list[i]).split('sra">')[-1].split('</a')[0])
  unique_name = np.unique(name_info)
  appeared_name = {}
  for ii in range(len(unique_name)):
    tmp_max_numb = len(np.where(np.array(name_info) == unique_name[ii])[0])
    if ii == 0:
      max_name = unique_name[ii]
      max_numb = tmp_max_numb
    else:
        if tmp_max_numb > max_numb:
          max_name = unique_name[ii]
          max_numb = tmp_max_numb
  max_name_list_nm = np.where(np.array(name_info) == max_name)[0]
  subject_google_scholar = tmp_mail_list[max_name_list_nm[0]].get('href')

# 특정 인물의 서지 목록을 확인할 수 있는 URL
surl = f'https://scholar.google.com{subject_google_scholar}'
```

크롤링 과정 (4/7)

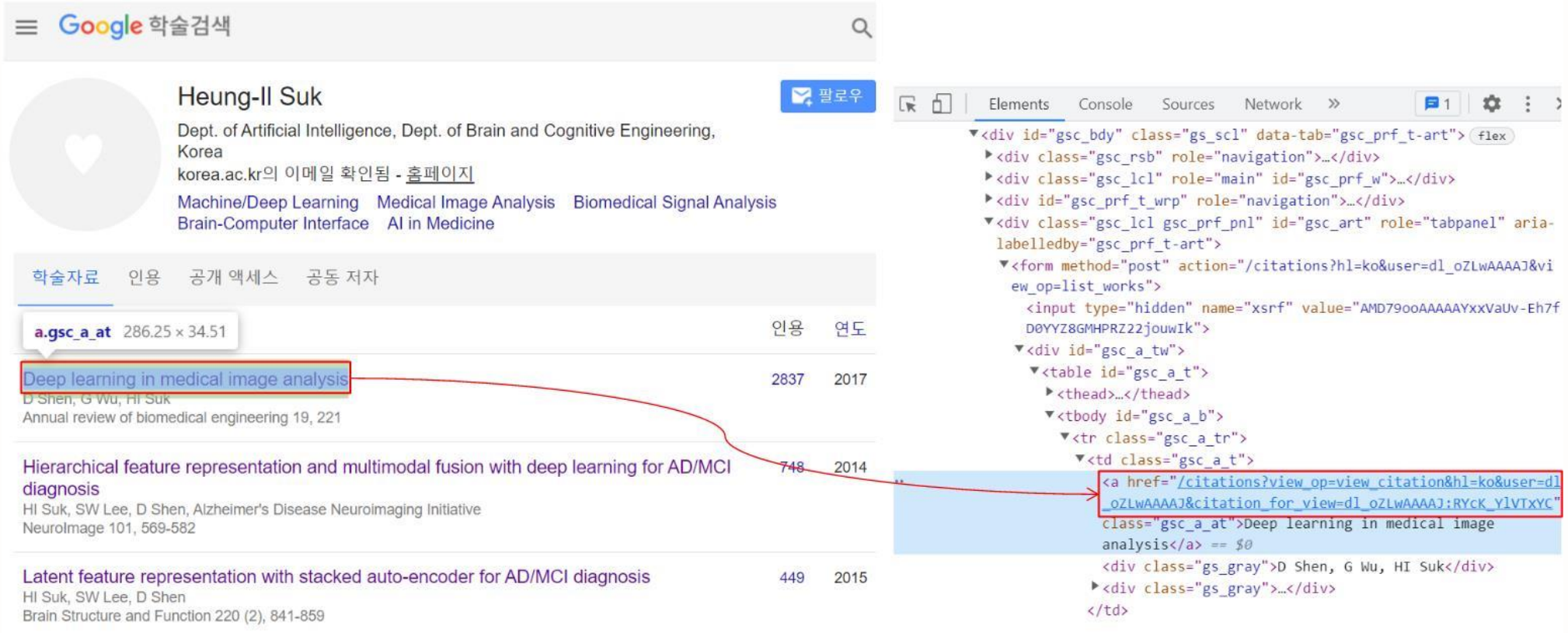
- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 이메일 입력을 통한 서지 목록 URL 추출
    - 추출한 URL 통해 확인 가능한 페이지



특정 인물의 서지 목록 주소를 통해 확인 가능한 페이지 예시

크롤링 과정 (5/7)

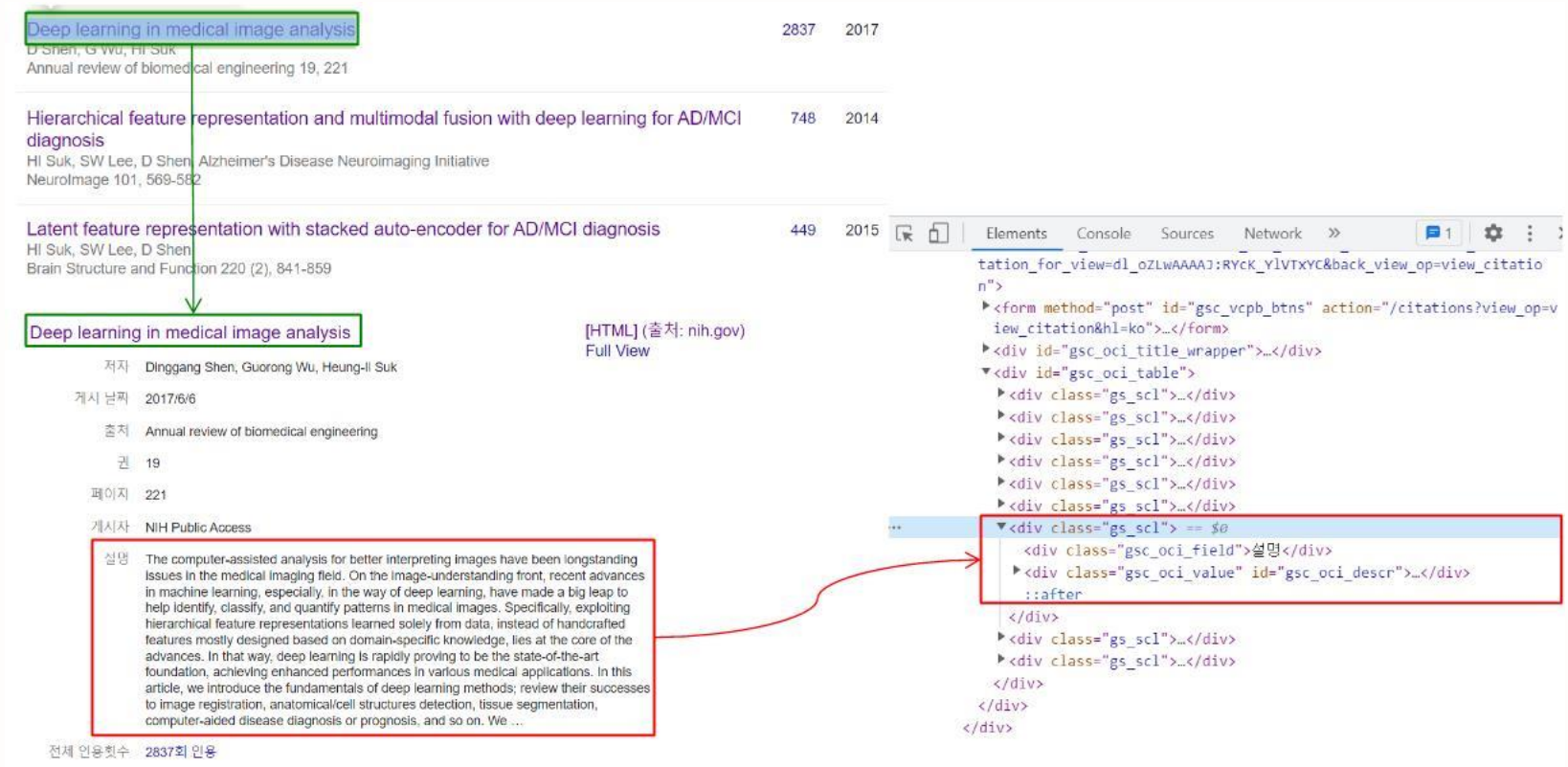
- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 논문 초록 크롤링
    - 개인의 서지 목록 페이지로부터 각 논문의 초록이 담긴 페이지 URL 주소 추출



개인의 서지 목록 페이지로부터 각 논문의 초록이 담긴 페이지 URL 주소 추출 관련 예시

크롤링 과정 (6/7)

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 논문 초록 크롤링
    - 추출한 URL을 통해 해당 논문 페이지에서 초록 추출



추출한 URL을 통해 확인 가능한 페이지 및 초록 추출 관련 예시

크롤링 과정 (7/7)

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 논문 초록 크롤링
    - 관련 코드

```
## 특정 인물의 서지 목록에서 논문 리스트 및 초록 추출을 위한 코드
get_sbj_url = requests.get(surl)
sbj_soup = BeautifulSoup(get_sbj_url.content, 'html.parser')
sbj_article_list = sbj_soup.select('td.gsc_a_t > a')

# 크롤링 데이터 저장을 위한 csv 폴더 및 파일 설정
notebooks_base_dir = '.'
dt = datetime.today().strftime('%Y%m%d_%H%M%S')
sv_fname = os.path.join(notebooks_base_dir, subject_id+'_'+str(dt)+'.csv')
f = open(sv_fname, 'w', encoding='utf-8', newline='')
wt = csv.writer(f)
wt.writerow(['email', 'name', 'title', 'abstract'])
```

```
for i in range(len(sbj_article_list)):
  print(sbj_article_list[i])
  web_name = sbj_article_list[i].get('href')

  # 초록 추출 관련
  web_page_full = f'https://scholar.google.com{web_name}'
  get_paper_url = requests.get(web_page_full)
  paper_soup = BeautifulSoup(get_paper_url.content, 'html.parser')

  abs_content = paper_soup.select('div.gsh_small')
  paper_title = str(sbj_article_list[i]).split('>')[-2].split('<')[0]
  print(web_page_full)
  print(list(abs_content))

  # [이메일, 이름, 논문 제목, 초록] 순으로 CSV 파일에 추출한 정보 저장
  out = [subject_id+'@'+school_mail, max_name, paper_title, abs_content[0].text]
  wt.writerow(out)

f.close()
```

크롤링 결과

- **이메일 기반 논문 초록 크롤링**
  - 특정 인물(사용자)의 논문 초록 크롤링
    - 크롤링 결과

	A	B	C	D
1	email	name	title	abstract
2	hisuk@korea.ac	hisuk	Deep learning in	The computer-assisted analysis for better interpreting images have been longstanding issues in the medical imaging field. On the image-understanding front, recent advances in machine learning, especia
3	hisuk@korea.ac	hisuk	Hierarchical feat	For the last decade, it has been shown that neuroimaging can be a potential tool for the diagnosis of Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI), and also fusion c
4	hisuk@korea.ac	hisuk	Latent feature re	Recently, there have been great interests for computer-aided diagnosis of Alzheimer's disease (AD) and its prodromal stage, mild cognitive impairment (MCI). Unlike the previous methods that considere
5	hisuk@korea.ac	hisuk	Deep learning ba	Combining multi-modality brain data for disease diagnosis commonly leads to improved performance. A challenge in using multi-modality data is that the data are commonly incomplete; namely, some m
6	hisuk@korea.ac	hisuk	Deep learning-ba	In recent years, there has been a great interest in computer-aided diagnosis of Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI). Unlike the previous methods that cons
7	hisuk@korea.ac	hisuk	State-space moc	Studies on resting-state functional Magnetic Resonance Imaging (rs-fMRI) have shown that different brain regions still actively interact with each other while a subject is at rest, and such functional interac
8	hisuk@korea.ac	hisuk	Hand gesture re	In this paper, we propose a new method for recognizing hand gestures in a continuous video stream using a dynamic Bayesian network or DBN model. The proposed method of DBN-based inference is p
9	hisuk@korea.ac	hisuk	A novel Bayesia	As there has been a paradigm shift in the learning load from a human subject to a computer, machine learning has been considered as a useful tool for Brain-Computer Interfaces (BCIs). In this paper, we
10	hisuk@korea.ac	hisuk	Deep ensemble	Recent studies on brain imaging analysis witnessed the core roles of machine learning techniques in computer-assisted intervention for brain disease diagnosis. Of various machine-learning techniques, s
11	hisuk@korea.ac	hisuk	A novel relationa	In this paper, we focus on joint regression and classification for Alzheimer's disease diagnosis and propose a new feature selection method by embedding the relational information inherent in the observa
12	hisuk@korea.ac	hisuk	A novel matrix-s	Recent studies on AD/MCI diagnosis have shown that the tasks of identifying brain disease and predicting clinical scores are highly related to each other. Furthermore, it has been shown that feature sele
13	hisuk@korea.ac	hisuk	Subspace regula	The high feature-dimension and low sample-size problem is one of the major challenges in the study of computer-aided Alzheimer's disease (AD) diagnosis. To circumvent this problem, feature selection a
14	hisuk@korea.ac	hisuk	Person authentic	In this paper, we propose a new biometric system based on the neurophysiological features of face-specific visual self representation in a human brain, which can be measured by ElectroEncephaloGraph
15	hisuk@korea.ac	hisuk	Deep sparse mu	Recently, neuroimaging-based Alzheimer's disease (AD) or mild cognitive impairment (MCI) diagnosis has attracted researchers in the field, due to the increasing prevalence of the diseases. Unfortunate
16	hisuk@korea.ac	hisuk	Commanding a b	In this work, we propose a novel brain-controlled wheelchair, one of the major applications of brain-machine interfaces (BMIs), that allows an individual with mobility impairments to perform daily living acti
17	hisuk@korea.ac	hisuk	Non-homogenec	Neuronal power attenuation or enhancement in specific frequency bands over the sensorimotor cortex, called Event-Related Desynchronization (ERD) or Event-Related Synchronization (ERS), respective
18	hisuk@korea.ac	hisuk	Canonical featur	Fusing information from different imaging modalities is crucial for more accurate identification of the brain state because imaging data of different modalities can provide complementary perspectives on th
19	hisuk@korea.ac	hisuk	Subject and clas	EEG-based discrimination among motor imagery states has been widely studied for brain-computer interfaces (BCIs) due to the great potential for real-life applications. However, in terms of designing a r
20	hisuk@korea.ac	hisuk	Multimodal mani	As the early stage of Alzheimer's disease (AD), mild cognitive impairment (MCI) has high chance to convert to AD. Effective prediction of such conversion from MCI to AD is of great importance for early
21	hisuk@korea.ac	hisuk	Matrix-similarity	Recent studies on Alzheimer's Disease (AD) or its prodromal stage, Mild Cognitive Impairment (MCI), diagnosis presented that the tasks of identifying brain disease status and predicting clinical scores ba

논문 초록 크롤링 결과 예시

번역

- **영어 논문 초록 번역**
  - 구글 번역 API 활용

```
[ ] abstract
```

```
['The computer-assisted analysis for better interpreting images have been longstanding issues in the medical imaging field. On the image-understanding front, recent advances in machine learning, especially, in the way of deep learning, have made a big leap to help identify, classify, and quantify patterns in medical images. Specifically, exploiting hierarchical feature representations learned solely from data, instead of handcrafted features mostly designed based on domain-specific knowledge, lies at the core of the advances. In that way, deep learning is rapidly proving to be the state-of-the-art foundation, achieving enhanced performances in various medical applications. In this article, we introduce the fundamentals of deep learning methods; review their successes to image registration, anatomical/cell structures detection, tissue segmentation, computer-aided disease diagnosis or prognosis, and so on. We …',
 "For the last decade, it has been shown that neuroimaging can be a potential tool for the diagnosis of Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI), and also fusion of different modalities can further provide the complementary information to enhance diagnostic accuracy. Here, we focus on the problems of both feature representation and fusion of multimodal information from Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). To our best knowledge, the previous methods in the literature mostly used hand-crafted features such as cortical thickness, gray matter densities from MRI, or voxel intensities from PET, and then combined these multimodal features by simply concatenating into a long vector or transforming into a higher-dimensional kernel space. In this paper, we propose a novel method for a high-level latent and shared feature representation …"]
```

```
[ ] translates = []

    for i in range(len(abstract)):
      translate = translator.translate(abstract[i], dest="ko").text
      translates.append(translate)
```

```
[ ] translates
```

```
['더 나은 해석 이미지를위한 컴퓨터 보조 분석은 의료 이미징 분야에서 오랜 문제가되었습니다.이미지 이해 전선에서, 특히 기계 학습의 발전, 특히 딥 러닝 방식에서 의료 이미지의 패턴을 식별, 분류 및 정량화하는 데 큰 도약을했습니다.구체적으로, 도메인 별 지식을 기반으로 설계된 수제 기능 대신 데이터에서만 배운 계층 적 기능 표현을 악용하는 것은 진보의 핵심에 있습니다.그런 식으로 딥 러닝은 최첨단 재단으로 빠르게 입증되어 다양한 의료 응용 프로그램에서 강화 된 공연을 달성합니다.이 기사에서는 딥 러닝 방법의 기본 사항을 소개합니다.이미지 등록, 해부학/세포 구조 검출, 조직 세분화, 컴퓨터 보조 질병 진단 또는 예후 등의 성공을 검토하십시오.우리 …',
 '지난 10 년 동안, 신경 영상은 알츠하이머 병 (AD)의 진단을위한 잠재적 도구가 될 수 있고, 혈전 단계, 가벼운인지 장애 (MCI) 및 다른 양식의 융합이 추가 정보를 추가로 제공 할 수 있음이 밝혀졌습니다.진단 정확도를 향상시킵니다.여기, 우리는 자기 공명 영상 (MRI) 및 양전자 방출 단층 촬영 (PET)의 기능 표현 및 복합 정보의 융합 문제에 중점을 둡니다.우리가 아는 한, 문헌의 이전 방법은 주로 대뇌 피질 두께, MRI의 회백질 밀도 또는 PET의 복셀 강도와 같은 손으로 만들어진 특징을 사용한 다음 간단히 긴 벡터로 연결하거나 변형하여 이러한 멀티 모달 특징을 결합했습니다.고차원 커널 공간.이 논문에서, 우리는 고급 잠재 및 공유 기능 표현을위한 새로운 방법을 제안합니다…']
```

구글 번역 API를 활용한 논문 초록 번역 결과 예시

한계점 및 해결 방안

- **실제 사용하는 이메일과 논문에 사용한 이메일 주소가 다른 경우 및 개인 구글 스칼라 페이지가 없는 경우에 따른 한계점**
  - 추출 가능한 초록이 존재하지 않음
  - 지도 교수님 연구 논문들과 연구 분야가 동일할 것으로 가정하여 지도 교수님 논문 초록 추출로 대체

- **개인 구글 스칼라 페이지 URL 추출 시 발생하는 한계점**
  - 개인의 메일 주소 입력 후 나오는 페이지에서 확인 가능한 이름의 빈도수를 활용함에 따라 공동 연구자 혹은 지도교수님이 해당 인물로 추출될 가능성이 있음
  - 해당 공동 연구자 혹은 지도교수님의 연구분야 또한 확인하고자 하는 개인의 연구분야와 유사할 것이라 가정하여 추출된 개인 구글 스칼라 페이지로부터 초록을 추출하는 것으로 대체

- **다양한 키워드 추출이 어려운 경우에 따른 한계점**
  - 추출된 키워들 간의 유사성이 높아 다양한 주제를 담지 못하는 한계점을 MMR방법의 diversity 하이퍼파라미터를 통해 해결함

아이디어

- **논문 추천 시스템을 운영하면서 축적된 학생들의 관심사 데이터를 활용**
  - 구글 스칼라를 통한 학생들의 관심사를 추출하는 방법은 별도의 설문을 진행하지 않아도 됨
  - 그러나 학생들이 직접 입력한 데이터가 아니며, 구글 스칼라에 논문 등록이 안된 학생들의 경우에는 관심사를 추출할 수 없음
  - 반면에 논문 추천 시스템을 통해 축적한 학생 관심사 데이터는 학생들이 직접 입력하였으며, 관심 주제가 명확하게 명시되어 있음
  - 별도의 학생들의 관심사를 물어보는 설문을 돌리지 않아도 되기 때문에 시간과 비용을 절약

타임스탬프	상세연구분야(종합)	[임의 수정 금지] 다음은 논문 추천에 사용된 대표 키워드 단어 5개입니다. (각 단어는 콤마로 구분)	이외 연구 분야를 대표하는 영어 키워드(최대5개)를 입력 부탁드립니다. (콤마로 구분 ex: A, B, C)	6. 관심 연구분야를 선택해주시기 바랍니다.	메일 수신일
8-3-2021 20:55:01	Agricultural and Biological Sciences (miscellaneous), Food ScienceHealth, nutritionBiochemistry, Genetics and	['vitamin', 'nutrient', 'intent', 'curcumin', 'biopolymer']	Vitamin B12, Vitamin C, 3D printing, emulsion, liposome	Agricultural and Biological Sciences	20210720
7-20-2021 20:45:23	Electrical and Electronic Engineering	['junctionless', 'walled', 'single', 'multilayer', 'noise']	Transistor, CNT, TMDS, FET	Engineering	20210720

02 데이터 구축 – 학생 입력 키워드(논문추천 시스템)
번역

- 상세연구분야(종합) 컬럼과 학생이 입력한 관심분야의 값 합쳐서 **GOOGLE TRANSLATE** 라이브러리를 사용해서 한글로 번역

```
student_text
```

```
['Agricultural and Biological Sciences miscellaneous Food ScienceHealth nutritionBiochemistry Genetics and Molecular Biology miscellaneous Ageing Biotechnologyvitamin nutrient intent curcumin biopolymer',
 'Electrical and Electronic Engineeringjunctionless walled single multilayer noise',
 'Artificial Intelligence Computer Vision and Pattern Recognition Signal Processingcolor coding video image algorithm',
 'Materials Science miscellaneous Electronic Optical and Magnetic Materials Surfaces Coatings and Filmsbonding emitter emission display flexible',
 'Inorganic Chemistry Organic Chemistryreaction chemistry organic synthesis microreactor',
```

```
translates
```

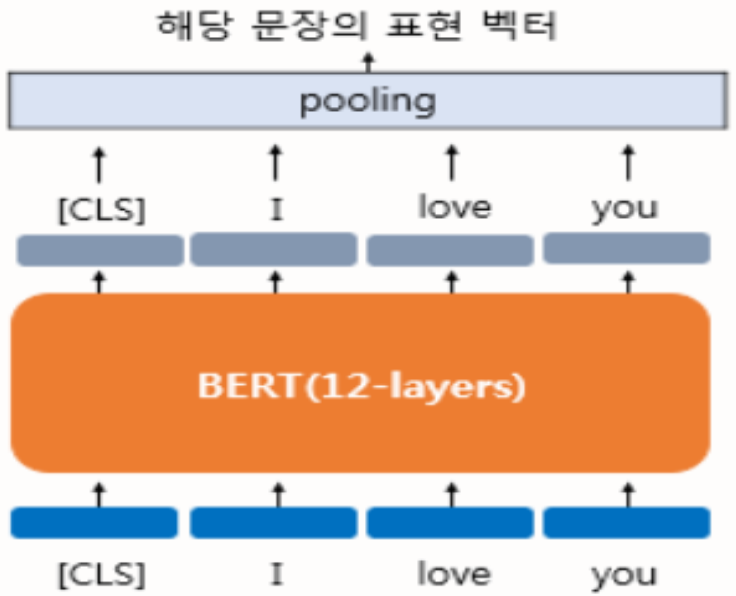
```
['농업 및 생물 과학 기타 식품 과학 건강 영양학 유전학 및 분자 생물학 기타 노화 생명 공학 비타민 영양 의도 Curcumin Biopolymer',
 '전기 및 전자 엔지니어링 접합이없는 벽화 단일 다층 노이즈',
 '인공 지능 컴퓨터 비전 및 패턴 인식 신호 처리 검색 비디오 이미지 알고리즘 코딩',
 '재료 과학 기타 전자 광학 및 자기 재료 표면 표면 코팅 및 필름 홀딩 이미 터 방출 디스플레이 유연성',
 '무기 화학 유기 화학 반응 화학 유기 합성 미세 반응기',
```

03

키워드추출 및 유사도계산

Sentence-BERT 개념

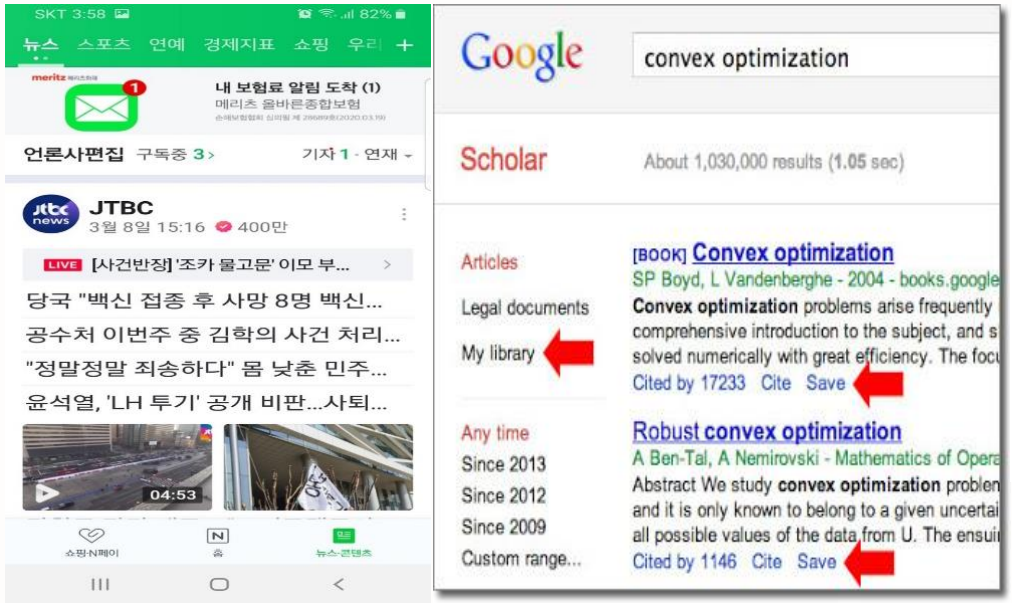
- **BERT의 문장/문서 임베딩의 성능을 우수하게 개선시킨 모델**
  - Sentence-BERT를 학습하는 방법은 문장 쌍 분류 태스크. 대표적으로는 NLI(Natural Language Inferencing) 문제를 해결함
  - NLI는 두 개의 문장이 주어지면 수반(entailment) 관계인지, 모순(contradiction) 관계인지, 중립(neutral) 관계인지를 맞춤
  - 기존의 BERT는 [CLS] 토큰의 표현 벡터를 문장 표현으로 사용하지만 Sentence-BERT는 BERT의 모든 단어의 표현 벡터를 평균 풀링하여 만든 벡터를 문장 표현으로 사용

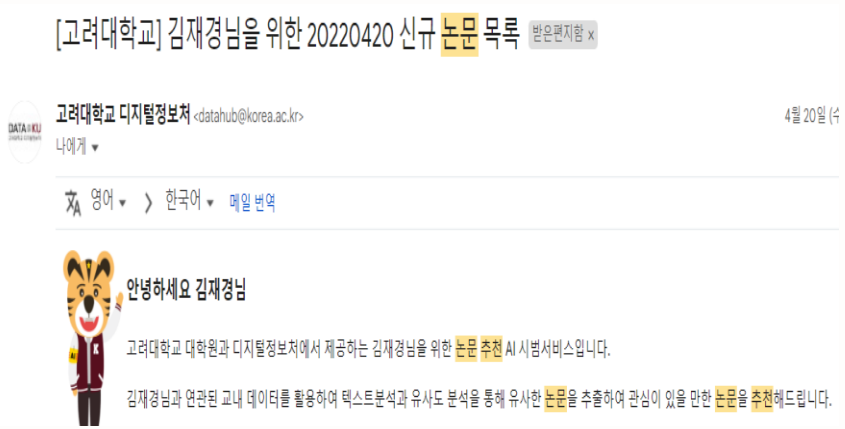


03 키워드 추출 및 유사도 계산

Sentence-BERT구현

- **sentence_transformers 패키지를 사용한 임베딩**
  - 사전 학습된 Sentence-BERT를 를 사용하여 텍스트를 벡터(vector)로 변환
  - 네이버 뉴스 기사, 구글 스칼라 논문 초록, 논문 추천 시스템에서 축적된 학생들의 관심 키워드들을 Sentence-BERT를 모델에 넣어서 벡터로 변환

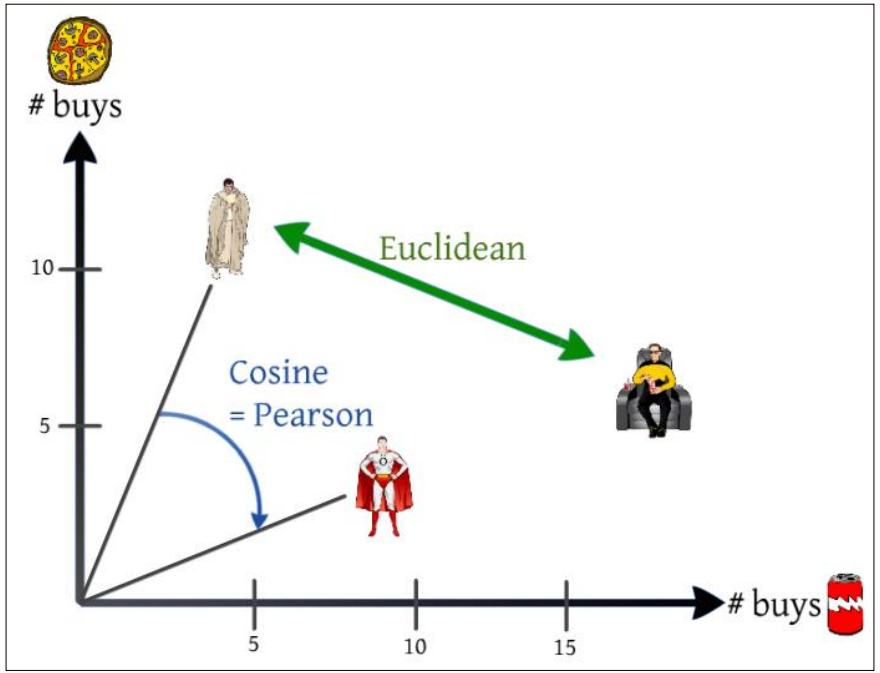




Embedding vector
[0.1 4.2 1.5 1.1 2.8]
[1.0 3.1 2.5 0.7 1.1]
[0.3 2.1 1.5 2.1 0.1]
[2.2 1.4 0.5 0.9 1.1]
[0.7 1.7 0.5 0.3 0.2]

유사도 개념

- **임베딩한 데이터 사이의 거리(유사도)를 구하는 다양한 방법들이 있음**
  - 임베딩 A, B의 유사도를 구하는 방법에는 유클리디안 유사도로 대표되는 거리 기반 유사도와 코사인 유사도로 대표되는 각도 기반 유사도가 있음
  - 거리 기반 유사도는 비슷하거나 가까운 좌표에 있는 점들이 유사도가 높다고 측정함
  - 각도 기반 유사도는 좌표를 기준으로 생각했을 때 x축과 (0, 0)에서 좌표까지 이르는 점선 주변에 있는 점들이 유사도가 높다고 측정함



코사인 유사도 개념

- **코사인 유사도를 활용하여 문서 간 유사도 비교 가능**
  - 코사인 유사도는 1 에 가까울수록 두 벡터가 유사하다고 해석하며, 0에 가까울수록 두 벡터가 유사하지 않다고 해석함
  - 문서의 길이가 다른 경우에도 비교적 공정하게 비교할 수 있다는 장점
  - 코사인 유사도를 활용하여 두 개의 벡터값들 간의 거리를 계산하여 거리가 가까운 벡터값을 가진 뉴스 기사들을 추천해주는 알고리즘을 개발

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

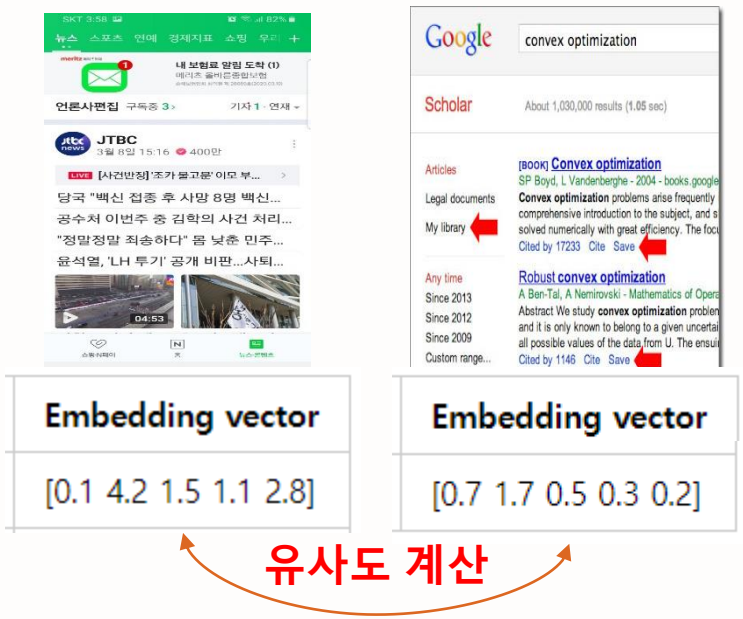
문서1 : 저는 사과 좋아요
문서2 : 저는 바나나 좋아요
문서3 : 저는 바나나 좋아요 저는 바나나 좋아요

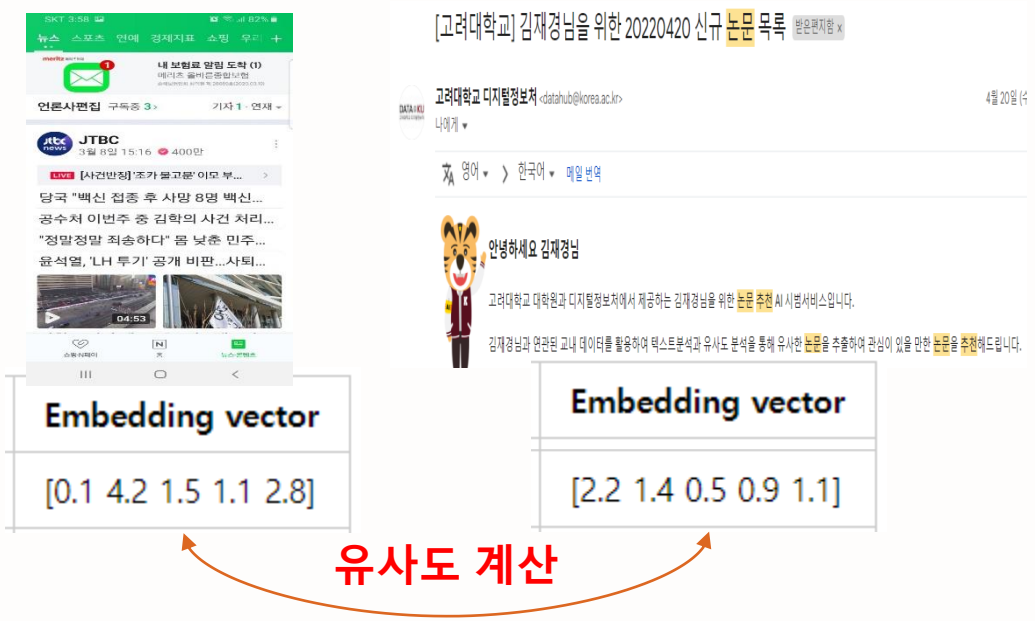
-	바나나	사과	저는	좋아요
문서1	0	1	1	1
문서2	1	0	1	1
문서3	2	0	2	2

```python
def cos_sim(A, B):
  return dot(A, B)/(norm(A)*norm(B))

doc1 = np.array([0,1,1,1])
doc2 = np.array([1,0,1,1])
doc3 = np.array([2,0,2,2])

print('문서 1과 문서2의 유사도 :',cos_sim(doc1, doc2))
print('문서 1과 문서3의 유사도 :',cos_sim(doc1, doc3))
print('문서 2와 문서3의 유사도 :',cos_sim(doc2, doc3))

# 문서 1과 문서2의 유사도 : 0.67
# 문서 1과 문서3의 유사도 : 0.67
# 문서 2과 문서3의 유사도 : 1.00
```

코사인 유사도 구현

- **임베딩값들 간의 유사도 계산**
  - 네이버 뉴스 기사, 구글 스칼라 논문 초록, 논문 추천 시스템에서 축적된 학생들의 관심 키워드 들을 Sentence-BERT를 모델에 넣어서 벡터로 변환
  - 네이버 뉴스 기사에 대한 임베딩값 – 구글 스칼라 논문 초록에 대한 임베딩값에 대한 유사도 계산
  - 네이버 뉴스 기사에 대한 임베딩값 – 학생 입력 관심 키워드에 대한 임베딩값에 대한 유사도 계산
  - 모든 뉴스 기사 간의 코사인 유사도를 계산하고 그 중 유사도 값이 높은 상위 20개 뉴스를 추천 결과로 선정





04

실험 결과 및 결론

실험 결과

- **구글 스칼라 데이터를 활용한 기업 추천 결과**
  - 구글 스칼라를 통해 구축한 키워드를 활용하여 학생 별 관심 기업 및 뉴스 기사 추천 알고리즘 진행
  - 학생과 뉴스 기사 간의 코사인 유사도를 계산한 후 유사도 값이 높은 상위 20개 뉴스를 추천 결과로 선정

- **인공지능 분야로 논문을 작성한 뇌공학과 학생 2명의 기업 추천 결과 예시**

[학생 A]

title	url	company	similarity
네이버 클로바, 세계 1위 머신러닝 학회 논문 12건 채택	http://www.seg	NAVER	0.76435
대학 개발자 멘토링 'SKT AI Fellowship' 성료...우수연구팀 4개 팀 선정	http://www.aiti	SK텔레콤	0.760769
LG AI연구원, 국제인공지능학회서 논문 발표...수준급 SW 역량 입증	http://www.aju	LG	0.728547
해외논문·알고리즘 해석은 기본? '고딩 주식러' 떴다	https://www.he	신세계	0.720013
KB저축은행, 대규모 인력 채용...하반기 '차세대 시스템' 완성도 높인다	https://www.aju	KB금융	0.706542
[AI 전환] 채용·목축·주행...AI 기술로 일상 바꾸는 스타트업들	https://www.aju	LG	0.702872

[학생 B]

title	url	company	similarity
KB저축은행, 대규모 인력 채용...하반기 '차세대 시스템' 완성도 높인다	https://www.aju	KB금융	0.722801
LG, 최고 권위 AI학회서 '초거대AI' 기술력 선봬...7편 논문 발표	https://www.yn	LG	0.693125
카카오브레인, 글로벌 학회에 AI 관련 논문 9건 등재	https://view.asi	카카오	0.688626
네이버 클로바, 세계 1위 머신러닝 학회 논문 12건 채택	http://www.seg	NAVER	0.681574
대학 개발자 멘토링 'SKT AI Fellowship' 성료...우수연구팀 4개 팀 선정	http://www.aiti	SK텔레콤	0.673739
라온피플, 신사업 수주 확대 '수익성 개선' 다양한 사업 범주 AI 솔루션 수	http://www.new	미래에셋증권	0.67106

실험 결과

- **학생 입력 데이터를 활용한 기업 추천 결과**
  - '논문 추천 서비스'를 통해 구축된 학생들의 관심 분야 데이터 활용 관심 기업 및 뉴스 기사 추천 알고리즘 진행
- [Electrical and Electronic Engineering]**과** [Agricultural and Biological Sciences (miscellaneous), Food ScienceHealth, nutritionBiochemistry, Genetics and Molecular Biology (miscellaneous), Ageing, Biotechnology]**를 관심 분야로 작성한 학생의 추천 결과**

[학생 A]

title	url	company	sim
현대자동차, 해외 석박사급 인재 상시채용	http://www.g	현대자동차	0.694755
SKE&S, 美에너지솔루션기업 투자..."에너지 신사업 강화"	http://www.e	LS	0.681094
SK이노베이션, 차세대 배터리 개발 인력 수시 채용	https://maga	SK	0.640378
SK이노베이션, 차세대 배터리 개발 인력 수시 채용	https://maga	SK이노베이션	0.640378
SKE&S, 美에너지솔루션기업 투자..."에너지 신사업 강화"	http://www.e	SK	0.626645
윤종혁 DGIST 교수, TSMC R램 공정 적용한 PIM 프로세서 개발...'2021	http://www.a	삼성전자	0.62042
서진시스템 컨테이너박스 생산도 신사업에, 전동규 사업다각화 분주	https://www.	HMM	0.61923

[학생 B]

title	url	company	sim
푸드플랜 수립 연구용역 '제안서 평가위원 모집' 등	https://www.	GS	0.793719
푸드플랜 수립 연구용역 '제안서 평가위원 모집' 등	https://www.	금호석유	0.774082
푸드플랜 수립 연구용역 '제안서 평가위원 모집' 등	https://www.	한화솔루션	0.774082
녹차유산균 위 손상 억제...아모레퍼시픽, 국제저널에 논문 게재	http://theviev	아모레퍼시픽	0.709115
아모레퍼시픽, 녹차유산균 엑소솜 피부 항염효과 논문 게재	http://www.te	아모레퍼시픽	0.695834
아모레퍼시픽, 녹차유산균 위손상 개선 효과 논문 발표	http://biz.nev	아모레퍼시픽	0.675178
SK바이오팜, 세노바메이트 연구 논문 국제 학술지 게재	http://www.n	SK바이오사이언스	0.659675

결론

- 기업 추천 서비스는 졸업을 앞둔 대학원 연구자들에게 하나의 **지원 통합형 솔루션**을 제공함
- **최신 취업 시장 트렌드까지 반영**한 인공지능 기반 추천 서비스를 통해 본교의 취업 시장 경쟁력 제고 뿐만 아니라 학생들의 재학 만족도도 증진시킬 수 있음
- 설문조사와 같은 학생들의 정보 제공을 기반으로 작동하던 **기존의 방법론의 한계를 극복**
- 학생 이메일 주소만을 사용하여 구글 스칼라로부터 학생에 관한 연구 키워드를 자동적으로 추출한 후 **개인 맞춤형 취업/진로 추천 서비스 제공**
- 현재 시행중인 '논문 추천 서비스'를 통해 구축된 학생들의 관심 분야 데이터도 추가적으로 활용 가능
- 기계지능 연구실의 학생들의 **구글 스칼라**와 '**논문 추천 시스템**'으로 구축된 관심 분야 데이터를 활용하여 기업 추천 서비스 연구를 진행하여 사용자의 주요연구에 기반한 최적의 기업 추천 및 연관 기사 결과를 얻을 수 있었음

05

향후 계획

05 향후 계획

데이터 적용 범위 확장 및 데이터 베이스화

- **적용범위 확장**
  - 뉴스 크롤링 대상 기업 : 기존 100개 기업에서 다양한 산업군의 기업 추가
  - 구글 스칼라 크롤링 대상 이메일 : 기존 기계지능 연구실 졸업생 대상에서 타 학과 및 연구실로 확장
  - 논문 추천시스템 대상 : 기존 500여명 데이터에서 확장하여 다양한 전공 학생 추가

- **다양한 데이터 연계 활용 방안**
  - 구글 크롤링 데이터와 논문 추천 시스템 입력 데이터가 모두 존재하는 학생들에 대해서 어떻게 두 정보를 활용할 수 있는지에 대한 방안 모색 (ex 다중 모달리티 융합 추천 알고리즘)

- **데이터 베이스 구축**
  - 기업 추천에 활용할 수 있는 최신 정보 기반의 데이터 정제
  - 향후 다른 과제에도 적용가능한 신뢰성 있는 대규모 데이터베이스 구축

- **자동화 API 모듈**
  - 데이터 수집부터 추천까지의 전과정을 한번에 수행가능한(End-to-end) 자동화 API 모듈 구현 및 서비스화

05 향후 계획

웹사이트 채용 공고 크롤링

- 추가 데이터 구축을 위해, 취업 및 진로와 관련성이 높은 웹 페이지 상의 채용 공고를 활용한 방안 모색

- 채용 사이트의 경우, 모집 기간이 지나면 채용 공고 확인이 어렵기 때문에 포털 사이트의 기업 채용 공고를 검색하여 해당 정보를 확보하고자 함

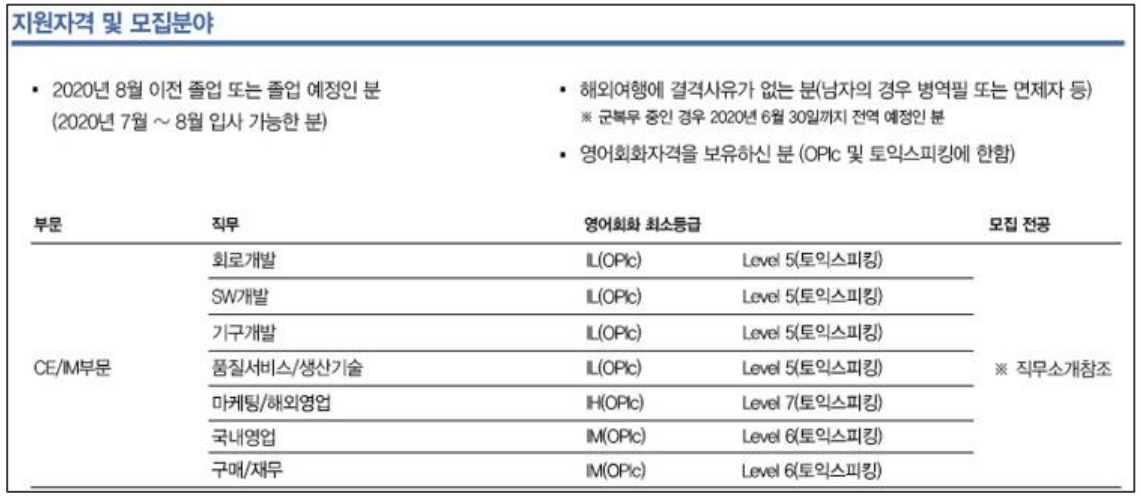
- 포털 사이트를 통한 웹크롤링 한계점
  - 웹 페이지들이 다양한 형식의 html을 가지고 있기 때문에 규칙성을 찾기 어려움
  - 네이버 기사와는 달리, 채용 공고는 문장 형태로 기재되어있지 않음
  - 채용 공고문 안에는 텍스트 이외의 이미지 형식의 정보도 존재하여 후처리가 용이하지 않음

- 따라서, 포털 사이트인 구글로부터 이미지 형식의 채용 공고문을 크롤링한 후, 광학 문제 인식(OCR, Optical Character Recognition) 기술을 통한 후처리 진행
  - 뷰티풀숍(BeautifulSoup) 패키지를 통해 채용 공고의 URL 및 이미지 크롤링
  - OCR을 활용해서 이미지를 텍스트로 변환하여 취업 및 진로 관련 데이터 확보

05 향후 계획

웹사이트 채용 공고 크롤링

- **확보된 채용 공고문들로부터 다양한 운영체제에서 활용 가능한 공식 OCR인 테서랙트(Tesseract) 패키지를 활용하여 텍스트 변환 진행**
  - 텍스트 위주로 구성된 채용 공고문을 사용했음에도 불구하고, 해당 텍스트들을 정확히 추출해내지 못한다는 기술적 한계 확인



지원자격 및 모집분야

- 2020년 8월 이전 졸업 또는 졸업 예정인 분 (2020년 7월 ~ 8월 입사 가능한 분)
- 해외여행에 결격사유가 없는 분(남자의 경우 병역필 또는 면제자 등) ※ 군복무 중인 경우 2020년 6월 30일까지 전역 예정인 분
- 영어회화자격을 보유하신 분 (OPIc 및 토익스피킹에 한함)

부문	직무	영어회화 최소등급		모집 전공
CE/IM부문	회로개발	IL(OPIc)	Level 5(토익스피킹)	※ 직무소개참조
	SW개발	IL(OPIc)	Level 5(토익스피킹)	
	기구개발	IL(OPIc)	Level 5(토익스피킹)	
	품질서비스/생산기술	IL(OPIc)	Level 5(토익스피킹)	
	마케팅/해외영업	IH(OPIc)	Level 7(토익스피킹)	
	국내영업	IM(OPIc)	Level 6(토익스피킹)	
	구매/재무	IM(OPIc)	Level 6(토익스피킹)	

채용 공고 이미지의 크롤링 결과 중 일부

```
부문 직무 영 어 회화 최 소 등급 모집 전공
회 로 개발 uo 16 디 토 익 스 피킹)
s\ 애 발 Lor 10 더 토 익 스 피킹)
기 구 개발 uo 100 여 디 토 익 스 피킹)
ceM 부 문 품 질 서 비 스 / 생 신기술 Lor 10 서 더 토 익 스피킹) ※ 직 무 소 개 참조
마 케 팅 / 차 외 영업 HORA 10 셔 7( 토 익 스 피킹)
국 내 영업 Moro 10 8 토 익 스 피킹)
구 매 / 재 무 MOPS) 10001 6( 토 익 스 피킹)
회 로 설 계 / 신 호 및 시 스 템 설 계 / 평 가 및 분석 Lor) L 이 더 토 익 스 피킹)
반도체 공 정 설 계 / 빈 도체 공 정 기술 uo 1000 이 토 익 스 피킹)
패 키 지 개 발 / 기 구 개 발 /9 애 발 uoro L 이 티 토 익 스 피킹)
0s 부 분 . ㄷ 46 시 클 레 이 센 / 설 비 기 술 / 인 프 라 기술 Lor 10 녀 더 토 익 스 피킹) ※ 직 무 소 개 참조
색 산 관 리 / 환 경 안 전 Mor 10 6 토 익 스 피킹)
영 업 마케팅 HORE) 10 녀 7( 토 익 스 피킹)
재 무 / 경 영 지 원 ( 일 반 )/ 인 사 MOR 10 6 토 익 스 피킹)
※ 상세 모 집 분 아는 하 단 의 첨 부 파 일 을 반드시 참 조 해 주시기 바랍니다. 청 각 장 애 인 을 위한 별도 어학 평 가 기준 :
```

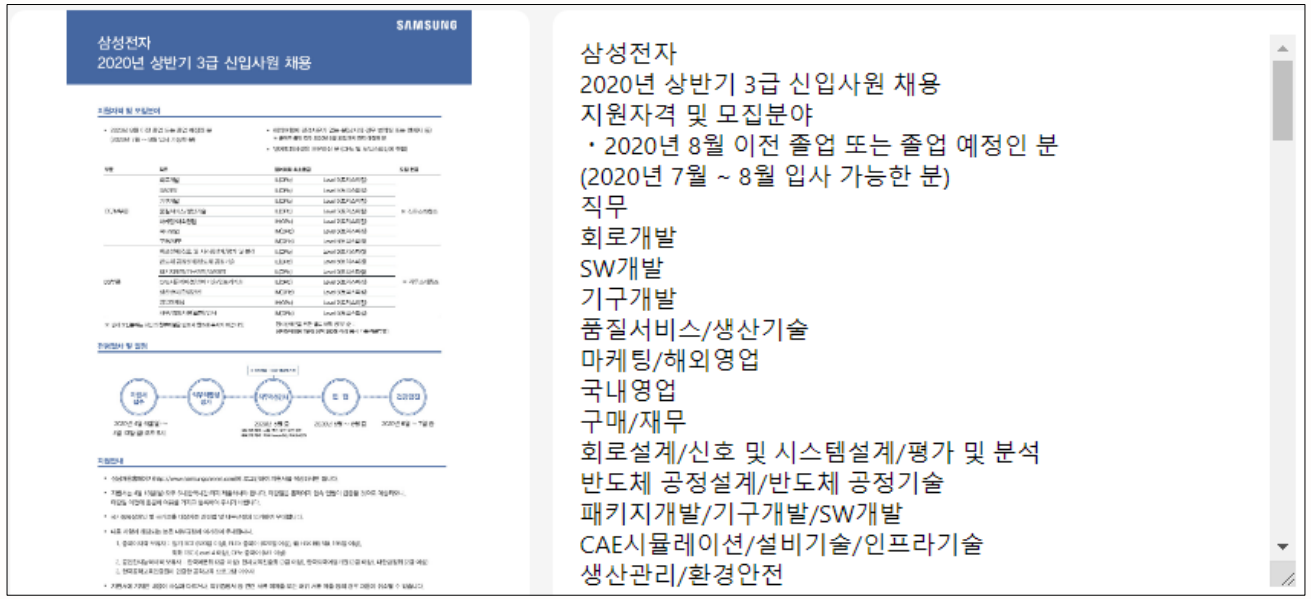
테서랙트를 활용한 채용 공고 이미지의 텍스트 변환 결과

정확도 측면에서 텍스트 인식률이 저조하다는 문제 존재

05 향후 계획

웹사이트 채용 공고 크롤링

- **OCR 텍스트 스캐너(Text Scanner) 사이트에서 공식적으로 활용 가능한 텍스트 변환 프로그램 활용**
  - 다양한 유형으로 구성된 채용 공고문임에도 불구하고, 다른 OCR 패키지들에 비해 원활히 변환되는 것을 확인
  - 해당 사이트를 활용하거나, 채용 공고문에 이미지 처리 등을 적용한 후에 테서랙트 패키지를 활용한다면 텍스트 인식률을 높일 수 있을 것으로 예상



무료 사이트를 활용한 채용 공고 이미지의 텍스트 변환 결과

(사이트 링크: https://www.cardscanner.co/ko/image-to-text)

Thank you