

## 환자 증상정보 기반 희귀질환 조기 발견 보조시스템

최재민\* · 김선용\*\*

Early Detection Assistance System for Rare Diseases  
based on Patient's Symptom Information

Jae-Min Choi\* · Sun-Yong Kim\*\*

## 요약

희귀질환은 증상이 전형적이지 않고 진단정보가 부족하여 전문의들조차 증상을 기반으로 질환을 의심하거나 질환명을 떠올리는 데에 어려움을 겪는다. 따라서 증상이 시작한 시점에서부터 정확한 진단을 받기까지 많은 시간 및 비용이 발생하며, 이는 환자의 신체적, 정신적, 경제적 부담을 심각하게 초래한다. 환자의 증상정보를 통해 의심되는 희귀질환을 제시하여 의사의 진단에 활용할 수 있도록, 본 논문에서는 웹 크롤링 및 텍스트마이닝을 활용한 희귀질환 조기 발견 보조시스템을 제안하고 이를 구현한다.

## ABSTRACT

Untypical symptoms and lack of diagnostic records make it difficult for even medical specialists to detect rare diseases. Thus, it takes a lot of time and money from the onset of symptoms to an accurate diagnosis, which seriously results in physical, mental, and economic pressure on patients. In this paper, we propose and implement an early detection assistance system for rare diseases using web crawling and text mining, which can suggest the names of suspected rare diseases so that medical staffs can easily recall the disease names and make a final diagnosis of the rare diseases.

## 키워드

Rare Diseases Detection, Symptom Information, Web Crawling, Text Mining, KoNLPy  
희귀 질환 발견, 증상 정보, 웹 크롤링, 텍스트 마이닝, 코엔엘파이

## 1. 서론

희귀질환이란 유병(有病)인구가 2만 명 이하이거나 진단이 어려워 유병인구를 알 수 없는 질환을 의미한다. 국내에서 보고되고 있는 희귀질환 수는 현재 1,123 가지이며, 질병관리청의 2020년 희귀질환자 통계 연보

에 따르면 신규 희귀질환자 수가 약 5만 2천 명에 달한다 [1]. 희귀질환자들은 정확한 진단을 받기까지 여러 병원을 전전함에도 불구하고 타 질병으로 오진을 받은 경험이 있는 경우가 약 33%에 달하며, 증상이 나타난 후 병원 첫 방문부터 최종 진단까지 소요된 기간은 평균 21개월로 보고된다 [2]. 이러한 희귀질환은 비록 전

\* 동서대학교 학부생(woals987@gmail.com)

\*\* 교신저자 : 동서대학교 소프트웨어학과

• 접수 일 : 2023. 02. 16

• 수정완료일 : 2023. 03. 15

• 게재확정일 : 2023. 04. 17

• Received : Feb. 16, 2022, Revised : Mar. 15, 2023, Accepted : Apr. 17, 2023

• Corresponding Author : Sun-Yong Kim

Dept. of Software, Dongseo University

Email : sykim@dongseo.ac.kr

문의라고 할지라도 해당 희귀질환에 대한 진단 경험이 없으면 증상을 듣고도 어떠한 질환인지 쉽게 떠올리지 못하거나 다른 질환으로 오진을 내리기가 쉽다. 더욱이, 희귀질환자들이 직접 경험하는 증상을 통해 희귀질환 명을 검색하는 수단 또한 미비한 실정이다. 현재 질병 관리청에서 운영하는 희귀질환 헬프라인에서 희귀질환에 대한 정보를 열람할 수는 있지만, 질병명을 직접 입력하여 검색해야 하거나, 한국 표준 질병·사인 분류 (Korean Standard Classification of Disease) 코드를 통해 검색해야 하기에 의학 지식이 없는 일반인들이 검색할 방법이 만무하다) [3]. 이외에도 여러 병원 홈페이지에서 주어진 증상 중 하나를 선택하고 추가적인 질문지를 통해 자가진단을 하는 서비스가 존재한다. 하지만 주어진 증상 항목이 부족하고, 진단 결과로써 희귀질환보다는 보편적인 질병 한 가지만을 주로 출력해주기 때문에, 실제 희귀질환을 앓고 있는 환자에게는 큰 도움이 되기 어렵다.

한편, 텍스트 데이터 분석 기술이 발전함에 따라 웹 크롤링(Web Crawling), 텍스트마이닝(Text Mining) 기술 등이 다양한 응용 분야에 적용되어 정체되지 않은 텍스트 데이터로부터 유의미한 정보를 추출하거나 추출된 정보를 기반으로 새로운 서비스를 개발하는 연구가 활발히 진행되고 있다. 경영학 분야의 연구 동향을 살피기 위해 웹 크롤링 기반 KCI 등재지 주제

어 분석 기법이 연구되고 있으며, 해시태그 추천 시스템을 개발하기 위한 일환으로써 인스타그램 내 텍스트 데이터를 크롤링하는 연구가 수행되고 있다 [4, 5]. 텍스트마이닝 기술의 경우, 중고거래 리뷰 데이터 기반 소비자 인성 분석, 아동용 감정 사진 기반 아동 감정 분석 등을 위한 연구에 적용되고 있다 [6, 7].

본 논문에서는 환자의 증상정보를 기반으로 가장 의심되는 희귀질환명을 제시해 줄 수 있는 희귀질환 조기 발견 보조시스템을 제안한다. 환자가 직접 겪고 있는 증상을 입력하면 가장 의심되는 후보 희귀질환 명들을 유사도 순으로 제시함으로써 의사의 신속한 희귀질환 진단에 보조 역할을 수행할 수 있을 것으로 기대된다. 증상정보 데이터셋 구축 및 분석을 위해 웹 크롤링 및 텍스트마이닝 기술을 적용하며, 본 논문의 선행연구 내용을 개선 및 보완하였다 [8].

## II. 희귀질환 조기 발견 보조시스템 개발

그림 1은 구축하고자 하는 희귀질환 조기 발견 보조 시스템의 개념도이다. 희귀질환 증상정보 데이터셋을 구축하기 위하여 request 및 BeautifulSoup를 이용한 웹 크롤링을 통해 서울대학교 병원 및 질병관리청 웹페이지의 html에서 희귀질환 증상정보들을 추출 및 수집한다.

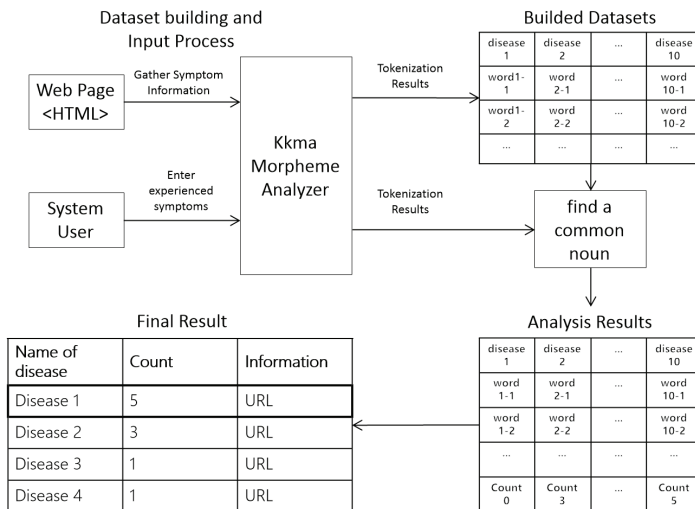


그림 1. 희귀질환 조기 발견 보조시스템 개념도  
Fig. 1 Conceptual diagram of early detection assistance system for rare diseases

1) <https://www.koicd.kr/>

수집한 정보를 KoNLPy의 Kkma 형태소 분석기를 통해 토큰화한 결과로 데이터셋을 구축한다. 이후에 시스템을 통해 이용자가 직접 경험한 증상을 자유로운 텍스트 형식의 입력 또한 Kkma 형태소 분석기를 통해 토큰화한다. 해당 결과와 기존에 구축한 데이터셋을 비교하여 공통된 명사를 찾고 개수를 센다. 마지막으로 이용자에게 분석 결과의 공통된 명사 개수를 내림차순으로 출력하여 결과를 쉽게 볼 수 있게하고, 질병관리청 회귀질환 헬프라인의 URL링크를 추가하여 해당 회귀질환의 정보를 바로 확인 할 수 있게 한다. 이 장의 각 세부 챕터에서는 제안하는 시스템 개발과정을 단계적으로 제시한다.

## 2.1 request 및 BeautifulSoup4를 활용한 웹 크롤링

```
<h3 class="stress" id="TABLE_OF_CONTENT3">증상</h3>
<p class="txt">주 증상은 운동시 호흡곤란과
[The main symptoms are difficulty breathing during exercise and]
<a href="/entry.naver?docid=927462&amp;ref=y"
onclick="clickcr(this, 'bdk.creference', '', '', event)">
만성 기침</a>으로, [chronic cough]
<a href="/entry.naver?docid=927458&amp;ref=y"
onclick="clickcr(this, 'bdk.creference', '', '', event)">
청색증</a> [cyanosis]
(<a href="/entry.naver?docid=927457&amp;ref=y"
onclick="clickcr(this, 'bdk.creference', '', '', event)">
저산소증</a>에 의해 입술주변이 파랗게 질리는 현상).
[(the phenomenon of turning blue around the lips due to hypoxia)]
<a href="/entry.naver?docid=927455&amp;ref=y"
onclick="clickcr(this, 'bdk.creference', '', '', event)">
곤봉지</a>(만성적인 저산소증에 의해 손가락 끝이 둥글게 되는 현상)
등이 동반될 수 있다. 폐 섬유화가 진행될수록 이런 현상들이
심해지며 만성적인 호흡곤란과 저산소증을 겪게 된다.
[clubbing finger
(the phenomenon of rounded fingertips due to chronic hypoxia)
can be accompanied. As pulmonary fibrosis progresses,
these phenomena worsen and
suffer chronic respiratory distress and hypoxia.]
</p>
증상 Symptoms
<p>대부분 만성적으로 1-2년에 걸쳐 서서히 진행되는
마른 기침과 운동시 호흡곤란이 주 증상이고,
병이 진행할수록 호흡곤란이 심해져 일상생활도 힘들게 된다.
말기에 폐동맥 고혈압이 합병되면 심장에 부담이 가시
누우면 심해지는 호흡곤란과 전신부종 등의 우심실 부전증상이 나타납니다.
진찰상 양쪽 폐의 아랫 부위에서 특징적인 숨기말 수포음이 들리고,
말단 곤봉지가 관찰되기도 하며, 말기에는 간비대,
말초부종 등의 우심부전 소견이 관찰됩니다.</p>
[The main symptoms are dry cough and
difficulty breathing during exercise,
which mostly progress chronically over 1-2 years.
As the disease progresses, breathing difficulties worsen,
making daily life difficult.
When pulmonary hypertension is merged in the end,
right ventricular insufficiency such as dyspnea and systemic edema,
which become severe when lying down due to a burden on the heart.
On examination, characteristic end-of-intake blisters
are heard in the lower parts of both lungs,
terminal clubbing finger are observed,
and right heart failure such as
hepatic hypertrophy and peripheral edema is observed at the end.]
```

그림 2. 특발성 폐섬유화증 증상정보 웹 크롤링 결과  
Fig. 2. Crawling results from idiopathic pulmonary fibrosis

request는 urllib 패키지 내의 존재하는 모듈 중 하나로써 웹페이지에서 html 소스 코드를 가져올 수 있

게 해준다. 그리고 BeautifulSoup4는 request를 통해 가져온 html 소스 코드로부터 원하는 정보를 가공 및 추출할 수 있게 해주는 패키지이다<sup>2)</sup>. 이를 통하여 수많은 양의 웹 문서를 사람이 일일이 구별하여 모으지 않고 자동으로 수행하는 웹 크롤링을 할 수 있다. 해당 기술을 통해 질병관리청의 회귀질환 헬프라인 및 서울대학교병원 의학 정보 페이지에서 회귀질환 10가지에 대한 증상정보를 추출 및 수집했다. 그림 2는 크롤링을 통해서 특발성 폐섬유화증에 대한 증상 표현만을 가져온 결과이다. html 문서를 txt 파일 형태로 변환하여 저장하고, 이후 형태소 분석기를 통해 전처리 및 토큰화를 진행하여 데이터셋을 구축하였다.

## 2.2 KoNLPy 패키지 형태소 분석기의 명사 추출 결과 비교 및 선정

질병에 대한 증상 부위 및 표현의 기준이 머리, 배, 팔다리, 두통, 복통, 구토와 같이 주로 명사로 표현되기에 토큰화 단위는 명사로 하였다. 본 논문에서 제안하는 시스템을 개발하기 위해 여러 가지 한국어 형태소 분석기 패키지 중 하나인 KoNLPy 패키지내의 Kkma, Okt, Komoran 중 어떤 형태소 분석기가 적절한지 알아보기 위해서 웹 크롤링으로 수집한 쿠싱 증후군 증상정보에서 명사를 추출하였다<sup>3)</sup> [9, 10]. 추출하는 과정에서 주로 다른 품사로 사용되지만, 다의어 이기에 다른 명사로 사용되기도 하는 52가지의 명사를 불용어 처리한 후 비교하였다. 그 결과는 표 1과 같다.

표 1. 형태소 분석기별 명사 추출 성능 비교  
Table 1. Comparison of noun extraction performance by morpheme analyzer

Model	Independent	Unintentionally extracted	Total
Kkma	43	2	206
Okt	30	7	195
Komoran	12	1	175

Model은 형태소 분석기의 모델을 의미하며, Independent는 각 형태소 분석기로 추출한 명사가 독립적으로 존재하는 개수를 의미한다. Unintentionally extracted란 부자연스러운 명사가 추출되거나 문장에서

2) <https://pypi.org/project/beautifulsoup4>

3) <https://konlpy.org/>

사용된 명사의 뜻으로 추출되지 않고 다른 뜻의 명사로 추출된 경우이다. 부자연스럽거나 의도한 뜻으로 추출된 명사가 많을 시 데이터셋을 구축할 때 후처리가 힘들다. 또한 시스템 이용자의 입력을 받아 분석 시 성능이 떨어지게 되는 결과를 초래할 수 있다. Total은 각 형태소 분석기에서 추출한 명사의 총 개수이다.

Kkma 형태소 분석기가 추출한 명사는 206개로 세 형태소 분석기 중 가장 많은 명사를 추출하였다. 하지만 문장에서 사용된 의미가 아닌 다른 의미로 사용된 명사를 추출하기도 하였다. 첫 번째로, ‘팔과 다리가...’라는 문장에서 증상 부위인 ‘팔’을 추출해야 하지만 여덟 가지 한방 치료 과목이라는 뜻의 ‘팔과’라는 명사를 추출하였다. 두 번째로, 범위에 드는 둘 이상의 대상을 모두 나열한 뒤, 복수(複數)임을 나타내는 말인 ‘등’을 추출하여 신체 부위로 혼란을 주었다. 이 두 경우를 제외하면 문장에서 의도하지 않은 뜻을 가진 명사를 추출한다는 등의 문제는 전혀 보이지 않았다.

Okt 형태소 분석기는 총추출한 명사의 수는 195개로 두 번째로 많이 추출했다. 하지만 Okt 형태소 분석기 또한 Kkma 형태소 분석기와 같이 복수(複數)임을 나타내는 말인 ‘등’을 추출하여 신체 부위로 오해를 불러일으키는 경우가 발생하였다. 그리고 “길항하는...”, “길항적으로...”라는 문장에서 “길항”이라는 명사를 추출하지 않고 관련이 없는 ‘항하’, ‘항작’이라는 명사를 추출하는 등 문장에서 의도하지 않은 뜻을 가진 명사가 총 7개가 추출되었다.

Komoror 형태소 분석기가 추출한 명사는 175개로 세 가지 형태소 분석기 중 가장 적은 수를 추출하였다. Kkma 형태소 분석기 및 Okt 형태소 분석기와 같이 오해를 불러일으킬 수 있는 ‘등’을 추출하였고 이외에는 문장에서 의도하지 않은 뜻을 가진 명사를 추출하는 문제는 보이지 않는다.

Okt 형태소 분석기는 문장에서 의도하지 않은 뜻을 가진 명사를 많이 추출하여 해당 연구에 적합해 보이지 않는다. Komoror 형태소 분석기는 의도하지 않은 뜻을 가진 명사를 많이 추출하지는 않는다. 하지만, Kkma 형태소 분석기가 Okt, Komoror 형태소 분석기보다 ‘피하출혈’, ‘털과다증’, ‘근력저하’와 같이 더 많은 양의 복합 명사를 추출함으로써 좀 더 명확한 증상을 확인할 수 있다. 이외에도 Kkma 형태소 분석기는 다른 두 형태소 분석기에 비해 띄어쓰기의 오류에 덜 민감한 형태소 분석기라는 특징도 있기 때문에 이용자가 증상을 입력할

때 발생하는 띄어쓰기 오류에 대비할 수 있어 본 논문에서 개발하는 시스템에 적합하다<sup>4)</sup> [11]. 따라서 본 논문에서는 웹 크롤링으로 가져온 증상정보를 토큰화할 때, 입력단계에서 이용자가 입력한 증상을 토큰화할 때 모두 Kkma 형태소 분석기를 이용한다.

### 2.3 증상 명사 데이터셋 구축

쿠싱 증후군을 제외한 나머지 9가지 희귀질환에 대해서도 Kkma 형태소 분석기를 통해 토큰화하였다. 이후 Python의 Pandas 패키지의 데이터 프레임(Dataframe)을 통해 표(Table) 형태로 처리하였다. 첫 행에 질환명을 저장하였고, 두 번째 행부터는 질환명에 해당하는 열에 토큰화한 결과를 저장하여 증상 명사 데이터셋을 구축했다. 그림 3은 구축한 결과의 일부를 나타낸 것이다. 10가지 희귀질환의 평균 명사 추출 개수는 약 196개이다. 명사를 가장 많이 추출한 크론병의 경우에는 288개, 가장 적게 추출한 특발성 폐섬유화증은 68개이다. 특발성 폐섬유화증의 경우에는 주 증상이 마른기침과 호흡 곤란, 청색증, 곤봉지뼀이며 추가적인 증상은 폐동맥 고혈압 합병증으로 인해 나타난다. 추가로 폐의 아래 부위에서의 수포음, 말기에는 간비대, 말초부종 등은 진찰을 통해서 알 수 있다. 이처럼 유난히 추출된 개수가 적은 질환이 존재하는 이유는 해당 질환이 증상 부위 및 종류가 다른 질환에 비해 유난히 적기 때문이다.

### 2.4 증상 입력 및 분석 단계 설계

본 논문에서 제안하는 모델의 입력단계에서 시스템 이용자가 입력한 증상 또한 데이터 구축단계와 같이 Kkma 형태소 분석기가 명사를 추출하게끔 설계하였다. 입력받은 증상에서 추출한 명사들을 기존에 구축한 데이터셋과 하나씩 매칭하여 공통된 명사를 발견할 때마다 해당 질환의 count에 +1을 더하는 방식으로 설계하였다. 이후 count가 높은 순서대로 희귀질환 리스트를 출력하고, 해당 희귀질환에 대한 정보를 볼 수 있는 웹 페이지의 URL 링크를 통해 이동할 수 있게 하였다.

## III. Case Study

본 논문에서 제안하는 환자 증상정보 기반 희귀질

4) <http://kkma.snu.ac.kr/>

환 조기 발견 보조시스템의 개발 결과를 확인하기 위해 희귀질환인 쿠싱 증후군 및 크론병을 대상으로 테스트를 진행하였다.

### 3.1 쿠싱 증후군 증상정보 입력 기반 분석

쿠싱 증후군의 증상정보를 통한 분석을 위해 SBS 좋은아침 5,261회에 출연한 가수 이은하 씨가 언급한 쿠싱 증후군의 증상들을 활용하였다.

증상을 자세히 입력해주세요.  
(Please describe your symptoms in detail.)

*보톡스를 맞은 듯 달덩이 처럼 부은 얼굴이됩니다. 손목은 가늘어지고 팔뚝은 오동통해지구요, 하루에 한끼만 먹는데도 임신부처럼 배가 나왔습니다. 다리도 발목은 가늘어지고 아물다리같이 부어있어요. (My face became swollen like a moon ball as if I had been hit by Botox. Wrist gets thinner, forearms get plump, and despite eating only one meal a day, belly comes out like a pregnant woman. My ankles are getting thinner, but my calves are swollen.)*

그림 4. 환자의 쿠싱 증후군 증상 입력  
Fig. 4 Cushing's Syndrome Symptom Input

Name of disease	Count	Information
쿠싱증후군(Cushing's syndrome)	5	https://helpline...
길랭-바레 증후군(Guillian Barre syndrome)	4	https://helpline...
다발성 경화증(Multiple sclerosis)	2	https://helpline...
모야모야병(MoyaMoya)	1	https://helpline...
혈우병(제 8,9인자 결함)(Haemophilia)	1	https://helpline...

그림 5. 쿠싱 증후군 증상 기반 분석 결과  
Fig. 5 Results of Cushing's Syndrome Symptom-Based Analysis

그림 4는 이은하 씨가 언급한 증상들을 기반으로 자연스러운 문장을 생성하여 시스템에 입력하는 화면이며, 그림 5는 분석 결과를 출력한 화면이다. 쿠싱 증후군의 증상정보를 입력하였기 때문에 쿠싱 증후군의 Count가 가장 높게 출력됨을 확인할 수 있었다.

### 3.2 크론병 증상정보 입력 기반 분석

또 다른 희귀질환인 크론병의 증상정보를 통한 분석을 위해, 크론병의 주 증상인 설사, 복통, 체중감소 등의 키워드로 문장을 생성하여 그림 6과 같이 시스템에 입력하였다.

증상을 자세히 입력해주세요.  
(Please describe your symptoms in detail.)

*배가 자주 아프고 설사를 자주합니다. 그러다 보니 체중도 줄 떨어졌습니다. 그리고 변을 볼 때 종종 피랑 고름이 섞어나오기도 해요. 지난번에 치료받은 항문 쪽에 치루가 다시 생긴 것 같아요 (My stomach hurts often and I often have diarrhea. As a result, I lost weight, too. Blood and pus are often mixed in the stool. The anal fistula that was treated last time has formed again.)*

그림 6. 환자의 크론병 증상 입력  
Fig. 6 Crohn's Disease Symptom Input

Name of disease	Count	Information
크론병(Crohn's disease)	6	https://helpline...
다발성 경화증(Multiple sclerosis)	3	https://helpline...
혈우병(제 8,9인자 결함)(Haemophilia)	2	https://helpline...
전신홍반루푸스(Systemic Lupus Erythematosus)	2	https://helpline...
쿠싱증후군(Cushing's syndrome)	1	https://helpline...

그림 7. 크론병 증상 기반 분석 결과  
Fig. 7. Results of Crohn's Disease Symptom-Based Analysis

그림 7은 입력된 증상 문장을 분석한 결과를 출력한 화면이며, 크론병의 Count가 가장 높게 출력됨을 확인할 수 있었다.

## IV. 결론 및 향후연구

본 논문에서는 전문의이더라도 희귀질환에 대해 진단 경험 부족 등의 원인으로 인해 발생하는 정확한 진단의 어려움 및 일반인들이 직접 겪는 증상을 통해 희귀질환명 검색 수단의 미비점을 보완하기 위하여 환자가 자연스러운 문장으로 입력한 증상 정보를 통해 희귀질환명을 검색할 수 있는 시스템을 제안 및 구현하였다. 기존에 존재하는 질병관리청 희귀질환 헬프라인 및 서울대학교병원 의학정보의 희귀질환 증상들을 Python의 request와 BeautifulSoup4를 이용한 웹 크롤링을 통해 수집 및 추출하였으며, 형태소 분석기 패키지인 KoNLPy에서 본 연구에 적절한 Kkma 형태소 분석기로 명사를 기준으로 토큰화하여 증상 부위 및 종류를 추출하고 데이터셋을 구축하였다.

시스템 이용자가 증상 입력단계에서 증상을 여러 가지 표현으로 입력할 수 있다. 하지만 본 연구에서는 서울대학교 의학 정보와 질병관리청 희귀질환 헬프라인에서 웹 크롤링을 통해 가져온 정보만으로 시스템을 구축

하였기에 한글 표현의 다형성(Polymorphism) 및 중의성(Ambiguity)을 극복하지는 못하였다 [12]. 향후 연구로써 신뢰도가 높은 질환 증상정보들을 추가로 수집하고 질환 증상 유의어 사전을 구축을 계획 중이다. 이를 통해 증상을 여러 가지 표현으로 입력하여도 정확한 결과를 도출하여 희귀질환 환자들이 여러 병원을 전전하며 소비되는 시간을 아끼고, 정확한 진단을 받아 건강 악화를 최소화하는 데 기여할 수 있을 것으로 기대한다.

**감사의 글**

본 연구는 2020년 동서대학교 학술연구조성비 지원에 의하여 이루어진 것임. (DSU-20200023)

**References**

[1] E. Choi and J. Lee, "Key findings from 2020 annual report on rare disease patients in Korea: incidence, mortality and medical service utilization," *Annual Report*, 2022.

[2] E. Choi, "Investigation of unmet needs for making diverse policy measures rare diseases," *Report*, 2021.

[3] Y. Na, N. Jang, and J. Won, "Statistical analysis of domestic laboratory accidents using classification criteria of KCD 7 and OIICS," *Journal of the Korean Society of Safety*, vol. 34, no. 3, 2019, pp. 42-49.

[4] K. Kang and S. Park, "Keyword analysis of KCI journals on business administration using web crawling and machine learning," *Korean Journal of Business Administration*, vol. 32, no. 4, 2019, pp. 597-615.

[5] S. Kim and D. Cho, "Design and implementation of hashtag recommendation system based on image label extraction using deep learning," *The Journal of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 4, 2020, pp. 709-716.

[6] Y. Jung and Y. Ju, "Analysis of text mining of consumer's personality implication words in review of used transaction application <Daangn Market>," *The Journal of the Korea Contents Association*, vol. 21, no. 11, 2021, pp. 1-10.

[7] Y. Park, S. Hwang, S. Lee, M. Kim, and S. Kim, "Text mining-based child sentiment analysis using child sentiment dictionary," In *Proceedings of Korean*

*Institute of Communications and Information Sciences Conference*, Jeju, South Korea, 2022, pp. 1602-1603.

[8] J. Choi, D. Kim, M. Kang, and S. Kim, "Early detection assistance system for rare diseases using web crawling and text mining," In *Proceedings of Korean Institute of Communications and Information Sciences Conference*, Jeju, South Korea, 2022, pp. 1605.

[9] E. Park and S. Cho, "KoNLPy: Korean natural language processing in python," In *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*, Chuncheon, South Korea, 2014, pp. 133-136.

[10] J. Lee, H. Han, K. Cho, and H. Lee, "Characteristics analysis of news articles using KoNLPy," In *Proceedings of Korean Institute of Information Technology (KIIT) Conference*, Cheongju, South Korea, 2020, pp. 628-629.

[11] B. Choe, I. Lee, and S. Lee, "Korean morphological analyzer for neologism and spacing error based on sequence-to-sequence," *The Journal of the Korea Institute of Information Scientists and Engineers*, vol. 47, no. 1, 2020, pp. 70-77.

[12] Y. Park, "A study on lexical ambiguity resolution of Korean morphological analyzer," *The Journal of the Korea Institute of Electronic Communication Sciences*, vol. 7, no. 4, 2012, pp. 783-787.

**저자 소개**

**최재민(Jae-Min Choi)**



2017년~현재 동서대학교 소프트웨어학과 재학

※ 관심분야 : 빅데이터, 텍스트마이닝

**김선용(Sun-Yong Kim)**



2014년 한동대학교 전산전자공학부 졸업(공학사)

2019년 University of Maryland, Baltimore County 방문연구원  
2020년 GIST 전기전자컴퓨터공학부 졸업(공학박사)

2020년 GIST AI연구소 연구원

2020년~현재 동서대학교 소프트웨어학과 조교수

※ 관심분야 : 스마트그리드, 강화학습