



Artificial Korean Natural Language Project

뉴스 기사를 기반으로 한 신조어 추출 프로젝트

한국소프트웨어산업협회 & 솔트룩스 채용 확정 교육

조장 노현진
김지현
박현태
안수빈

목차

Contents



01 프로젝트 개요

- 1 프로젝트 목표
- 2 신조어 정의
- 3 프로젝트 흐름도
- 4 시스템 아키텍처
- 5 사용도구 및 환경
- 6 프로젝트 관리
- 7 일정 관리
- 8 팀 소개

04 데이터 모델링

- 1 신조어 추출 프로세스 개요
- 2 신조어 후보 추출 세부 내용
- 3 신조어 후보 추출 결과 집계
- 4 최종 신조어 채택 및 산출물
- 5 단어 유사도 파악 개요
- 6 FastText 모델의 선정 이유
- 7 FastText를 활용한 단어 유사도 파악

07 참고 문헌

02 데이터 수집 및 적재

- 1 사용 DB 및 특징
- 2 추출 과정 및 적재

05 자바 웹 프로그래밍

- 1 템플릿 선택
- 2 웹 페이지 구조도
- 3 각 단계별 시각화
- 4 기타 페이지
- 5 시연

03 데이터 전처리

- 1 데이터 탐색 및 처리 방안
- 2 데이터 전처리 과정 요약
- 3 데이터 전처리 과정

06 결론 및 추후 과제

- 1 프로젝트 결론
- 2 결과를 통해 배운점
- 3 추후 과제



1. 프로젝트 개요

발표자 : 노현진

AI

1. 프로젝트 개요

1-1. 프로젝트 목표



- 프로젝트 목표 및 미션

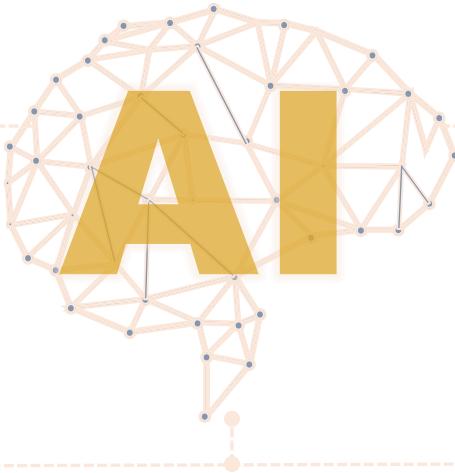
목표

뉴스 기사에 새로이 등장하는 신조어를 추출하고 이를 사전화 시킴으로써 한국어 NLP의 발전에 기여.



신조어 정의

- 신조어의 97%가 명사 또는 명사구로 이루어져 있기에 신조어의 품사를 **명사로 한정**.
- 분류가 되지 않을 가능성 있는 **복합 명사 또한 신조어로 정의**



미션

뉴스 기사 데이터를 수집, 정제하고 모델링을 통해 **신조어를 추출**. 해당 과정을 **시각화** 하여 웹으로 구현



역할 분담

데이터 베이스 관리 – 노현진
데이터 전처리 – 박현태
데이터 모델링 – 안수빈, 박현태
웹 프로그래밍 – 김지현



사전 지식

기존 연구 논문 탐색 및 스터디 관련 도서 구매

1. 프로젝트 개요

1-2. 신조어 정의

- 신조어에 대한 정의 및 예시



넓은 의미의 신조어

온라인 커뮤니티에서 사용하는 젊은 세대들만의 용어가 아닌 이미 예전부터 사용된 단어라도 대중적으로 사용되는 단어가 아니라면 신조어에 포함

일반적이지 않은 복합명사

형태소 분석을 통해 하나의 명사로 태깅 되지 않는 복합명사를 신조어 후보군으로 추출하지만 최종 후보군에는 기존 사전에 존재하지 않는 복합명사만을 신조어 후보군에 포함.



홑 따옴표 안의 명사

영화 이름이나 최근에 뜨고 있는 문구 같은 경우 홑 따옴표 안에서 띄어쓰기로 구분되어 있는 경우가 많아 따옴표 안의 띄어쓰기를 공백으로 제거하고 전체 기사에서 얼마나 자주 등장했는지 확인하여 신조어에 포함.

고유 명사

공인의 이름, TV 프로그램명, 회사명, 제품명 등 고유 명사 또한 신조어에 포함.

AI 1. 프로젝트 개요

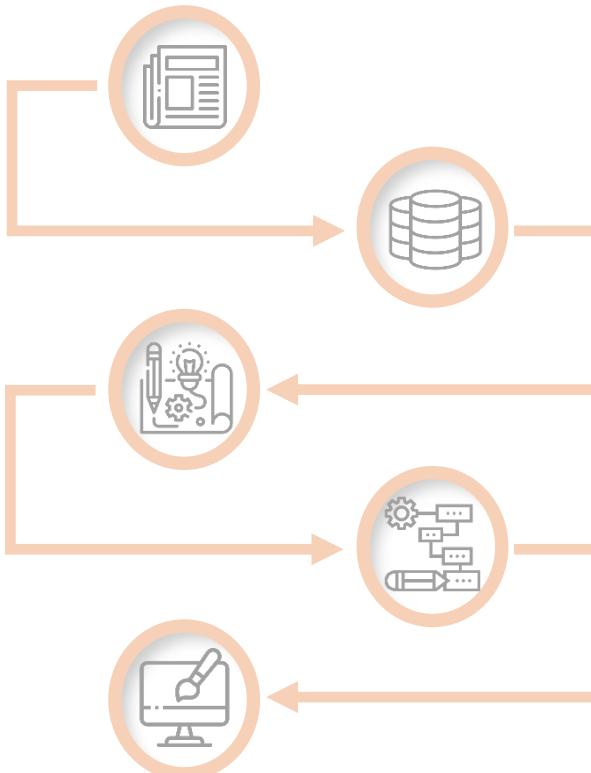
1-3. 프로젝트 흐름도



- 각 주요 단계별 특징

1. 기사 데이터 수집

- 파이썬의 **scrapy** 라이브러리를 활용.
- DAUM 기사에서 **종합 언론사 15개** 선택.
- 정치, 경제, 사회, 문화, IT 등 **총 10개 분야**.
- 수집된 기사는 NoSQL의 **MongoDB**에 적재.



2. 데이터 전처리

- 결측치가 하나라도 있는 행 **모두 제거**.
- 원문데이터에서 **영어 및 특수문자**로만 되어있는 기사(본문) 행 **제거**.
- **홑따옴표(') 안의 단어 및 문장** 고유명사로 판단하여 기존 본문데이터에 **컬럼으로 저장**.
- **정규표현식**을 활용하여 전처리.
- 전처리후의 데이터에서 **결측치**가 존재하는 행 **모두 제거**.
- 카테고리별로 분류.

3. 신조어 추출 및 단어 유사도 도출

- 카테고리와 주단위로 추출 기준 설정.
- **비지도 학습 기반 soynlp 패키지를** 활용하여 모든 명사 및 신조어 후보를 추출.
- 추출된 단어들은 모두 **사전과 비교**.
- 중복되지 않는 단어를 신조어 후보로 선정.
- 신조어 후보군 중 **최종 신조어 선택**.
- **fasttext**를 사용하여 단어 유사도 파악

4. Spring MVC를 통한 웹 구현

- 템플릿 선택(**Dashboard** 용)
- **Spring MVC 모델** 구현
- Mongo Db와 연동
- **Amchart** 등의 차트 라이브러리를 활용한 시각화

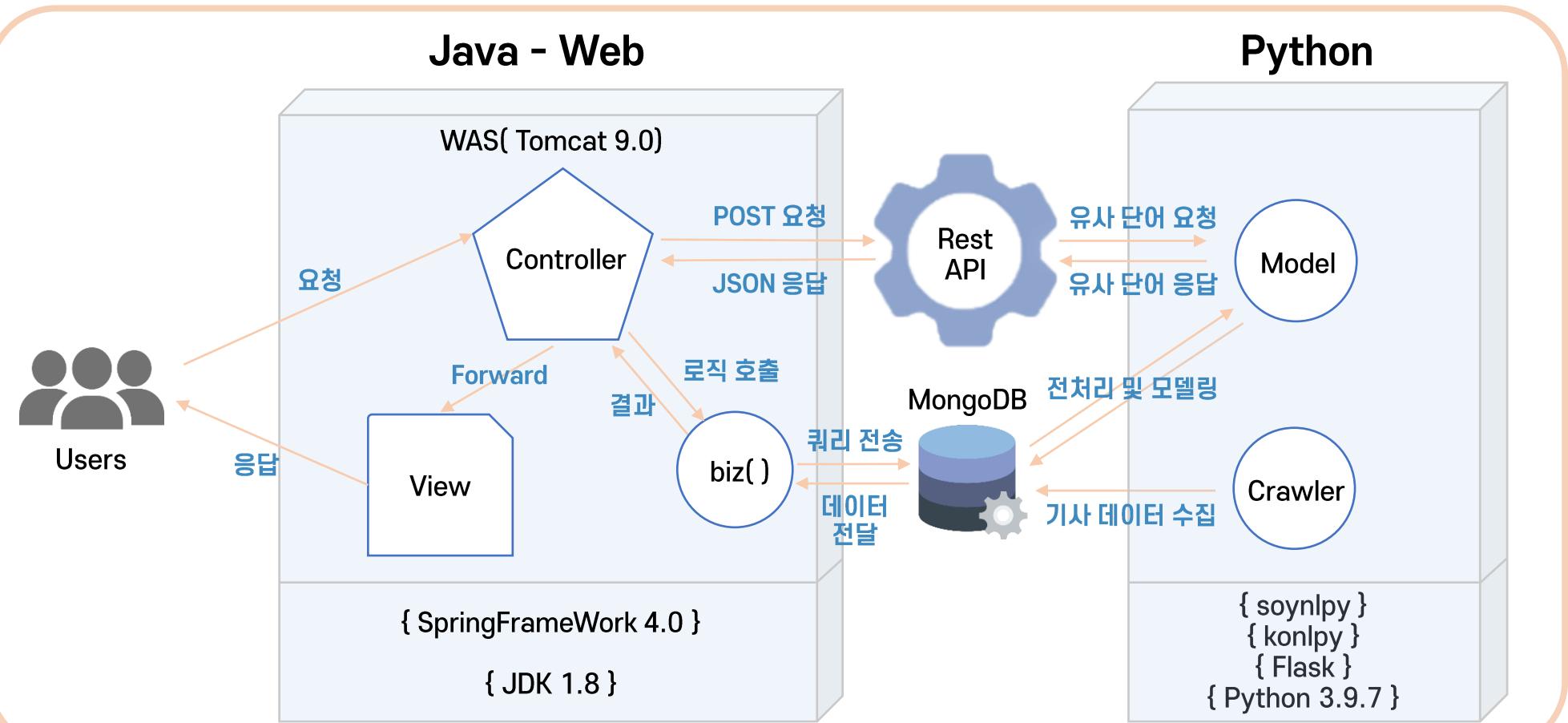
기사 크롤링 -> DB -> 전처리 -> 모델링 -> 웹 구현

1. 프로젝트 개요

1-4. 시스템 아키텍쳐



- 시스템 상세 아키텍쳐

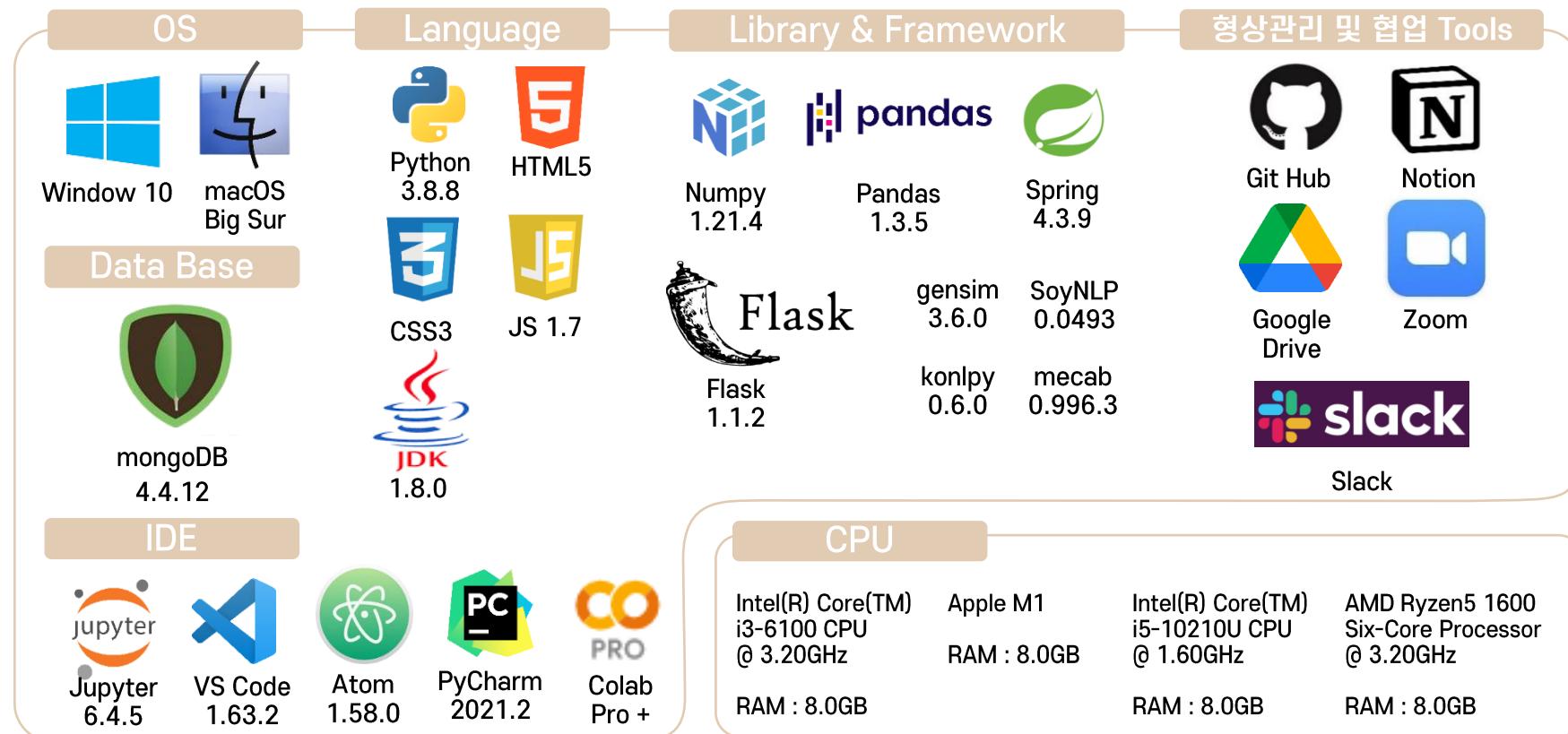


1. 프로젝트 개요

1-5. 개발도구 및 환경



- 개발 환경 및 각 도구 별 상세 버전



1. 프로젝트 개요

1-6. 프로젝트 관리(1)

- 깃허브를 이용한 소스 코드 형상 관리



Nohyunjin / ANLPK_Project Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 5 branches 0 tags Go to file Add file Code

HyuntaePark FastText 0b60e7d 44 minutes ago 91 commits

- Article_Data 1차 신조어 추출 18 days ago
- Python FastText 44 minutes ago
- java 모델링 차트 수정 17 hours ago
- .DS_Store 모델링 차트 수정 4 days ago
- index.html first commit 28 days ago
- mecab_python-0.996_ko_0.9.2_msvc-... 1차 신조어 추출 18 days ago
- new_word_temp_society_week1_ver2... 신조어 추출 : 사회분야, 20220119~20220125 20 days ago
- new_words_temp.csv 1차 신조어 추출 18 days ago
- new_words_temp2.csv 1차 신조어 추출 18 days ago
- new_words_temp_0224.csv 1차 신조어 추출 18 days ago
- new_words_temp_0224_num.csv 1차 신조어 추출 18 days ago
- new_words_temp_0224_num_ver2.csv 1차 신조어 추출 18 days ago
- new_words_temp_0224_num_ver3.csv 1차 신조어 추출 18 days ago

About No description, website, or topics provided.

0 stars 1 watching 3 forks

Releases No releases published Create a new release

Packages No packages published Publish your first package

Contributors 4 bini9788 wlglus9 Nohyunjin HyuntaePark

Help people interested in this repository understand your project by adding a README. Add a README

Languages

HTML 41.5%	CSS 29.3%
Python 15.4%	Java 6.4%
JavaScript 3.9%	SCSS 3.5%

A 1. 프로젝트 개요

1-6. 프로젝트 관리(2)

- 노션을 이용한 프로젝트 관리

TodoList			
제목	날짜	주요 Point	참고
4주차	2022년 3월 6일 ~ 2022년 3월 13일	최종 신조어 리스트 추출 단어 유사도 사용 여부 PPT	To do 후에 결정해야 할 것들. 1. 단어 유사도 사용 여부 2. 발표 내용 정리 및 연습
3주차	2022년 2월 28일 ~ 2022년 3월 6일	신조어 후보군 추출	To do 후에 결정해야 할 것들. 1. View단에서 표현할 것들
2주차	2022년 2월 21일 ~ 2022년 2월 27일	웹 구현 완성	
1주차	2022년 2월 14일 ~ 2022년 2월 20일	전처리 및 모델링 완료	To do 후에 결정해야 할 것들. 1. 지속적인 기사 수집 2. DB 저장

매주 To do list를 작성하여 해당 주에 완료되어야 할 업무를 구성원 모두가 숙지하고 이어서 진행될 이후 과제 또한 사전에 설정

매일 오전, 오후 회의를 통해 프로세스 진행도를 팀원 및 멘토와 함께 공유하고 회의록을 작성

회의록

모든 회의 ▾

- 3월 11일
- 3월 8일
- 3월 7일
- 3월 4일
- 3월 2일
- 3월 1일
- 2월 28일
- 2월 25일
- 2월 23일
- 2월 22일
- 2월 21일
- 2월 18일
- 2월 17일
- 2월 15일
- 2월 14일 오후 회의

+ 새로 만들기



속성 그룹화 필터 정렬 검색 ... 새로 만들기 ▾

- | | |
|----------|-----------------------|
| 지현정김민수빈 | 지난주 토요일 오후 4:39 |
| 현진정민지현수빈 | 지난주 토요일 오후 4:28 |
| 현진정민지현수빈 | 지난주 화요일 오전 10:02 |
| 지현정김민수빈 | 2022년 3월 4일 오후 6:07 |
| 현진정민지현수빈 | 2022년 3월 2일 오후 12:43 |
| 현진정민지현수빈 | 2022년 3월 2일 오후 12:43 |
| 현진정민지현수빈 | 2022년 2월 28일 오후 2:47 |
| 현진정민지현수빈 | 2022년 2월 25일 오전 11:22 |
| 현진정민지현수빈 | 2022년 2월 23일 오후 12:24 |
| | 2022년 2월 22일 오후 6:09 |
| 지현정김민수빈 | 2022년 2월 22일 오전 9:06 |
| 현진정민지현수빈 | 2022년 2월 20일 오후 7:43 |
| 지현정김민수빈 | 2022년 2월 17일 오후 6:30 |
| 수빈정진민지현 | 2022년 2월 15일 오후 6:38 |
| 현진정민지현수빈 | 2022년 2월 14일 오후 7:41 |

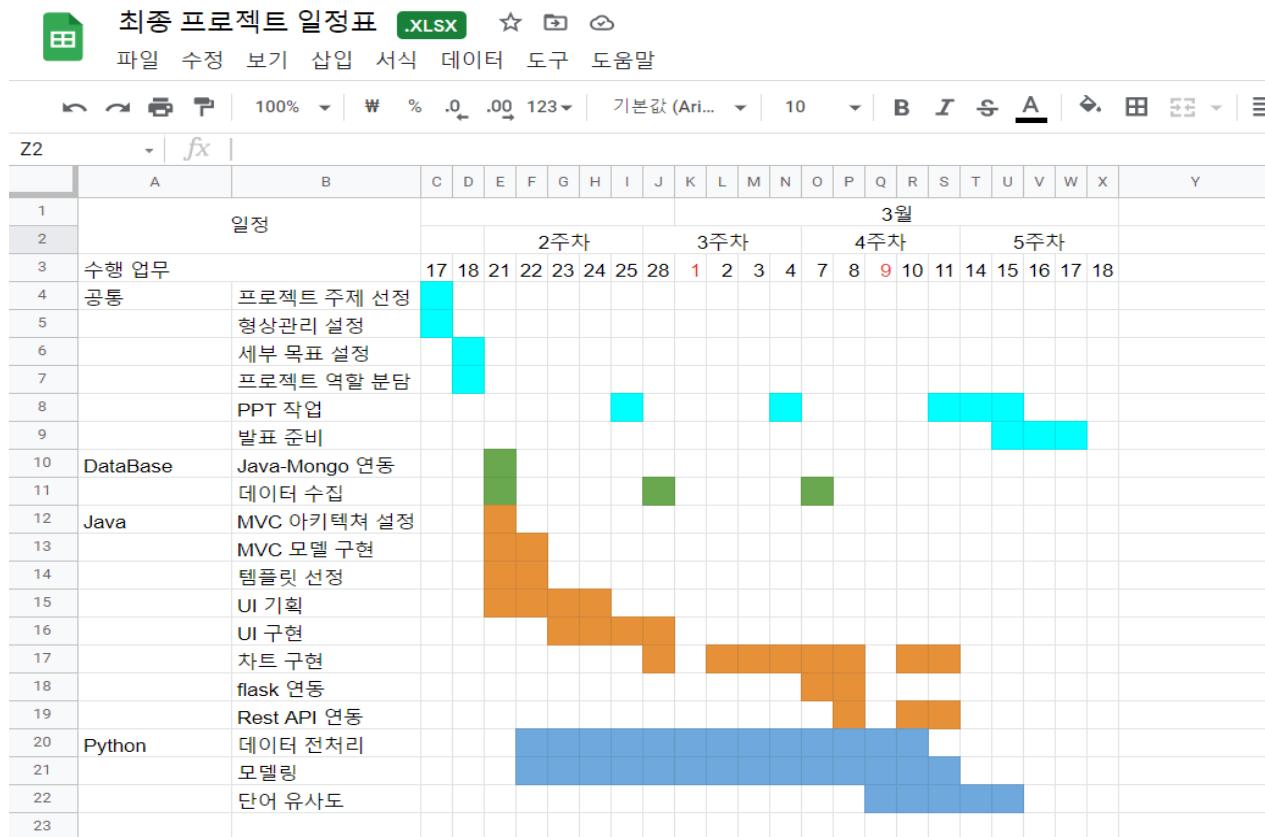
스탠드업

1. 프로젝트 개요

1-7. 일정 관리



- 구글 스프레드 시트를 이용한 일정 관리

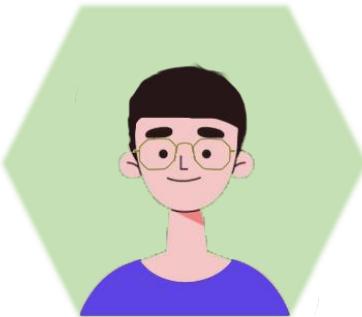


1. 프로젝트 개요

1-8. 팀 소개



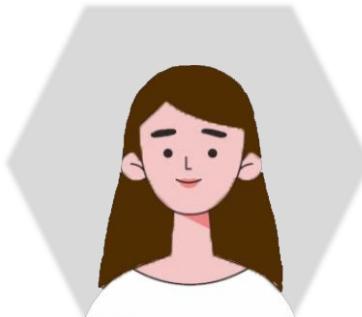
- Team Member



노현진
Manager

사용 언어 : SQL, Python
보유 역량 :
MySQL, Oracle,
MongoDB 등 DB 관리

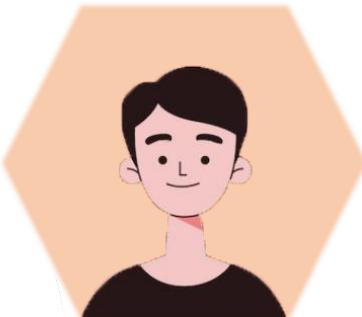
역할 : 데이터 베이스 담당



김지현
Developer

사용 언어 : JAVA
보유 역량 :
웹 개발
차트 시각화

역할 : 웹 프로그래밍 담당



박현태
Data -
Scientist

사용 언어 : Python
보유 역량 :
Data Engineering

역할 : 데이터 전처리 담당



안수빈
Data -
Scientist

사용 언어 : Python
보유 역량 :
Data Engineering

역할 : 데이터 모델링 담당



2. 데이터 수집 및 적재

발표자 : 노현진

2-1. 사용 DB 및 특징

- NoSQL의 특징 및 선택 이유



1. 도큐먼트 데이터베이스

필드와 값의 쌍으로 구성, 관계를 갖는 데이터를 중첩 도큐먼트와 배열을 사용하여 1개의 도큐먼트로 표현 가능

2. 유연한 스키마

도큐먼트들의 필드 집합이 동일하지 않고 같은 필드라도 데이터 타입이 다를 수 있음

3. 비 관계형 데이터베이스

테이블 간의 관계 개념이 없어 정렬, 분류, 탐색 속도가 빠름

4. 비 트랜잭션

데이터의 유효성을 보장하기 위한 ACID를 수행하지 않아 속도 개선

A 2. 데이터 관리

2-2. 데이터 추출 및 적재

- 기사를 크롤링 후 MongoDB에 저장

The diagram illustrates the process of extracting data from a news article and storing it in a MongoDB database.

News Article Source:

- Website: V.daum.net
- Section: News (뉴스)
- Category: Society (사회)
- Article Title: "[경향신문] '딸 걱정' 문 부순 부모 '무죄'"
- Author: 전현진 기자
- Date: 입력 2022. 01. 02. 21:36
- Comments: 댓글 6개
- Content Summary: A summary of the news article content, including the main headline, author, date, and a snippet of the article text.

MongoDB Storage:

- Result View: Shows the extracted document structure in JSON format.
- Document Structure (Extracted from the screenshot):


```
{
        "_id" : ObjectId("622551ddbbf8123a1098e059"),
        "article" : "[경향신문] ,방 안에 있던 딸이 자해할까봐 문손잡이를 부쉈다 재물손괴 혐의로 기소유예 처분을 받았지만 재판관은 '기소유예 처분을 취소해달라'며 검찰을 상대로 청구한 헌법소원심판 사건에서 재판관 전원일치 의견으로 인용 결정했다고 2일 밝혔다.",
        "category" : "사회",
        "date" : "20220102",
        "week" : "1주차",
        "name" : "전현진 기자",
        "source" : "경향신문",
        "title" : "\\"딸 걱정\\" 문 부순 부모 '무죄'",
        "url" : "https://news.v.daum.net/v/E2DZd1Chcu"
      }
```

Key Points:

- Crawling:** 파이썬 scrapy 라이브러리 활용
- 15개 언론사 :** 다음 뉴스 제휴 언론사 중 종합 언론사
(조선일보, 동아일보, 문화일보, 한겨례, 한국일보, 중앙일보, 연합뉴스, 세계일보, 서울신문, 문화일보, 대전일보, 뉴시스, 뉴스1, 국제신문, 국민일보, 경향신문, 강원도민일보)
- 10개 분야 :** 사회, 경제, 문화, 정치, 국제, 스포츠, 연예, 사설, 보도자료, IT



3. 데이터 전처리

발표자 : 노현진

3-1. 데이터 탐색 및 처리 방안(1)

- 기사 본문 속 불필요한 데이터 탐색



3-1. 데이터 탐색 및 처리 방안(2)



- 데이터 처리 방법

1. 정규 표현식

- 전체 데이터에서 공통적으로 드러나는 패턴에 대하여 정규표현식으로 제거

정규식을 이용한 기자 이름 제거 예시

```
re.split(r'^A-Za-z0-9가-]', str(df['name'][0]))
```

2. 데이터 컬럼 이용식

```
Result | Query Code | Explain |  
1  {  
2      "_id" : ObjectId("622551ddbbf8123a1098e059"),  
3      "article" : "[경향신문] ,방 안에 있던 땅이 자해할까봐 문손잡이를 부쉈다 재물손괴 협의로",  
4      "category" : "사회",  
5      "date" : "20220102",  
6      "week" : "1주차",  
7      "name" : "전현진 기자",  
8      "source" : "경향신문",  
9      "title" : "\"딸 걱정\" 문 부순 부모 '무죄'",  
10     "url" : "https://news.v.daum.net/v/E2DZd1Chcu"  
11 }
```

- “name”, “source”의 데이터를 이용하여 해당 열의 기사본문 데이터에서 불필요한 데이터를 제거



3. 데이터 전처리

19

3-3. 전체 전처리 요약



- 전체 전처리 과정 및 방식 정리

전처리 내용	방식	비고
한글 비중이 30% 이하인 기사	정규표현식	한글 비중이 30% 이하인 경우 데이터 완전 제거 후 인덱스 초기화
이메일 제거	정규표현식	
기자 이름 제거	정규표현식	
기자 이름 제거	데이터 컬럼 이용	기사 작성자 컬럼과 비교하여 본문과 일치하는 단어 제거
지역 + 언론사 제거	정규 표현식	특정 언론사에서 지역과 언론사에 대한 패턴이 등장함
홑 따옴표 안의 문자열 처리	정규표현식	띄어쓰기 제거, 전체 기사에서 10번 이상 등장한 경우만 (한 기사에서 여러 번 등장한 경우 1번으로 count)
특수 문자 제거	정규 표현식	특수문자를 먼저 제거할 경우 정규표현식이 제대로 역할을 수행하지 못함
언론사 제거	데이터 컬럼 이용	언론사 컬럼 데이터를 이용하여 본문속의 언론사 제거
한 음절 글자 제거	정규표현식	한 음절 글자가 발견이 안될 때 까지 반복 작업
숫자만 있는 글자 제거	정규표현식	한 음절 글자가 발견이 안될 때 까지 반복 작업
숫자 + 글자 제거	정규표현식	한 음절 글자가 발견이 안될 때 까지 반복 작업

3-4. 데이터 전처리 과정(1)



- 데이터 전처리 순서 및 프로세스 정의

1. 흄 따옴표(' ') 안의 단어나 문구를 리스트에 따로 저장 및 관리

1-1. 처리 방법

- 띠어쓰기 모두 제거하여 하나의 단어로 인식하게끔 변환

Ex) '어벤져스 : 인피니티 워' -> 어벤져스:인피니티워

1-2. 저장 및 관리방법

- 하나의 기사에서 여러 번 등장한 경우 전체 기사에 대해 1번 등장했다고 취급하고 같은 카테고리 안에서 10번 이상 등장한 경우 추출하여 저장 및 관리

2. 한글비중이 적은 데이터 처리

2-1. 처리 방법

- 본문에서 글자수로 한글 비중이 20% 이하인 본문에 대하여 데이터 완전 제거 선택 기준:

- 한글 비중 30% -> 25148개의 데이터 탈락
- **한글 비중 20% -> 24820개의 데이터 탈락**
- 한글 비중 10% -> 24663개의 데이터 탈락
- 한글 비중 5% -> 24635개의 데이터 탈락

2-2. 이후 전체 데이터에 대하여 결측치(any) 제거 후 인덱스 초기화 진행



3. 데이터 전처리

21

3-4. 데이터 전처리 과정(2)



- 데이터 전처리 순서 및 프로세스 정의

3. 이메일 제거

3-1. 처리 방법

- 정규표현식을 사용하여 이메일 패턴에 해당하는 글자 제거

Ex) hongildong@kosa.com

3-2. 처리 이유

- 기사 특성상 기사 작성자의 이메일을 기입하는 경우가 대부분

4. 기사 작성자 제거

4-1. 처리 방법

- 1차적으로 정규표현식, 2차로 컬럼의 값을 이용하여 기사 작성자 제거

4-2. 처리 이유

- 기사 특성상 기사 작성자의 이름과 직함 혹은 소속 등을 기입하는 경우

Ex) 흥길동 기자, 흥길자 특파원

- 정규표현식으로 제거가 안되는 글자에 대해 컬럼의 값을 이용하여 추가적으로 제거

Ex) 직함이 없이 이름만 있는 경우 -> 흥길동,

정규표현식 패턴에 맞지 않을 경우 -> 서울대 흥길자 교수



3. 데이터 전처리

22

3-4. 데이터 전처리 과정(3)



- 데이터 전처리 순서 및 프로세스 정의

5. 지역+언론사 제거

5-1. 처리 방법

- 정규표현식을 사용하여 (지역=언론사) 패턴에 해당하는 글자 제거
Ex) (서울=뉴스1) 흥길동 기자

5-2. 처리 이유

- 몇 개의 특정 언론사에서 위의 동일한 패턴 발견

6. 홑 따옴표 안의 문구와 단어 제거

6-1. 처리 방법

- 정규표현식을 이용하여 ‘ ’ 안의 문구나 단어 제거

6-2. 처리 이유

- 1번 과정에서 이미 필요한 데이터는 수집을 해 놓았기 때문에 불필요한 데이터로 취급



3. 데이터 전처리

23

3-4. 데이터 전처리 과정(4)



- 데이터 전처리 순서 및 프로세스 정의

9. 특수문자 제거

9-1. 처리 방법

- 정규표현식을 사용하여 ‘-’를 제외한 모든 특수문자 제거 후 제거된 글자들을 띄어쓰기로 취급

9-2. 처리 이유

- ‘K-방역’, ‘S-oil’과 같이 ‘-’로 중간에 연결된 단어를 다수 발견

10. 언론사명 제거

10-1. 처리 방법

- 컬럼안의 데이터를 이용하여 기사속에 일치하는 언론사명 제거

10-2. 처리 이유

- 위의 과정을 거치고도 기사 속에 남아있는 언론사명을 제거하기 위함



3. 데이터 전처리

24

3-4. 데이터 전처리 과정(5)



- 데이터 전처리 순서 및 프로세스 정의

11. 한 음절 글자 제거

11-1. 처리 방법

- 정규표현식을 사용하여 단어 양끝의 공백을 기준으로 한 음절일 경우 제거
- 단어 양끝의 공백을 기준으로 제거되기 때문에 복수의 작업 진행

Ex) 할 수 밖에 -> 수 밖에 -> 밖에

11-2. 처리 이유

- 신조어 특성상 한 음절로 구성된 의미 있는 단어는 없다고 판단

12. 숫자만 있는 글자 제거

12-1. 처리 방법

- 단어와 결합되지 않은 숫자 혼자 독립적으로 있는 경우 제거
- 단어 양끝의 공백을 기준으로 제거되기 때문에 복수의 작업 진행

Ex) 2022 03 14 -> 03 14 -> 14

12-2. 처리 이유

- 신조어 특성상 숫자로만 구성된 의미 있는 단어는 없다고 판단

3-4. 데이터 전처리 과정(6)



- 데이터 전처리 순서 및 프로세스 정의

13. 숫자+글자 정규표현식으로 제거

13-1. 처리 방법

- 정규표현식을 사용하여 단어 앞의 공백을 기준으로 숫자와 글자로 구성된 경우 제거
- 단어 양끝의 공백을 기준으로 제거되기 때문에 복수의 작업 진행

Ex) 13만5000원입니다 -> 5000원입니다 -> 입니다

13-2. 처리 이유

- 숫자+글자로 구성된 단어들을 확인해본 결과 보통 숫자+단위가 대다수 발견

14. ‘-’ 기호 음수로 판단될 경우 제거

14-1. 처리 방법

- 숫자나 단어 앞에 ‘-’가 있는 경우 정규표현식으로 제거

14-2. 처리 이유

- 음수를 제거하기 위함과 위의 과정을 거친 후 잔재하는 ‘-’와 결합된 단어들을 제거해주기 위함



4. 데이터 모델링

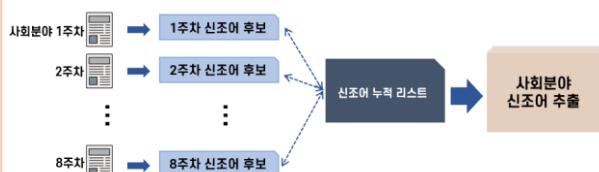
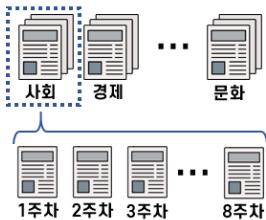
발표자 : 안수빈, 박현태

4-1. 신조어 추출 프로세스 개요

- 다음의 단계에 따라 신조어 추출을 진행

기사 데이터 분리

전처리 된 304,247개의 기사를
10개의 카테고리 별로 분리하고
 카테고리 내에서 다시 **8주**로 분리하여
카테고리 및 주별 신조어 추출



신조어 후보 추출

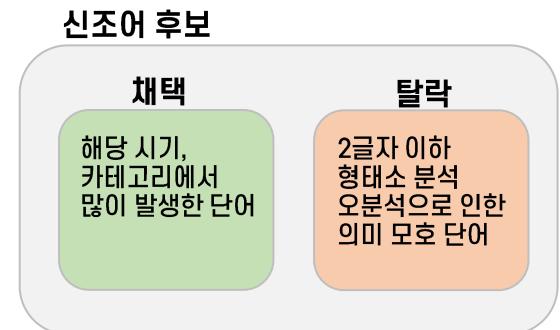
- a. 명사 가능성이 높은 모든 단어를 추출
- b. 사전에 등록되지 않은 단어만 추출
- c. 기사에서 자주 등장한 단어 채택
- d. 카테고리 내 이전 주 추출 신조어와 비교 후 전에 등장하지 않았던 새로운 단어만 추출



최종 신조어 채택

추출된 단어를 조사하여
시의적절한 신조어이면 선택하고,
 부적합한 신조어는 **불용어 사전**으로
 정의하여 탈락시켰으며,
 형태소 분석에서 오분석으로 인한
 잘못 추출된 단어는
단어 변환 사전 정의하여 적용

신조어 후보





4. 데이터 모델링

28

4-2. 신조어 후보 추출 세부 내용(1)



- 각 카테고리, 주차에 따라 신조어 후보가 추출되는 프로세스 중 1단계 명사 추출 내용

STEP 01 명사 가능성이 높은 모든 단어를 추출

STEP 02 사전에 등록되지 않은 새로운 단어만 추출

STEP 03 기사에서 자주 등장한 단어만 추출

STEP 04 이전에 등장하지 않았던 새로운 단어만 추출

카테고리 별
신조어 후보

기존 한국어 형태소 분석기의 한계

- 기존 형태소 분석기는 복합 명사를 [가장 작은 단위의 명사](#)로 추출
- 사전에 등록되지 않은 단어는 인식하지 못하기 때문에 [신조어 탐지](#)에 있어 성능 저하

Soynlp 패키지로 모든 단어 추출

- 복합명사 및 신조어를 하나의 명사로 온전하게 추출하기 위해 [비지도 학습](#) 기반 자연어처리 라이브러리인 [soynlp](#)를 사용하여 기사 내 명사 가능성이 높은 모든 단어들을 추출
- [soynlp](#)의 [NewsNounExtractor](#)를 사용하여 기사 내 모든 명사를 추출
- NewsNounsExtractor는 기존 명사 추출기를 뉴스 기사 내 명사 추출에 최적화시킨 추출기로서 전체 문서에서 [3번 이상 등장한 단어](#)를 추출

명사 추출 예시

<전처리 후 기사 원문> 미국 IT 전문 매체 테크크런치는 CES2022를 앞두고 흥미로운 기획을 선보였다 전인 CES2012에서 등장한 기술들을 회고한 것이다

Komoran	미국, 전문, 매체, 테크, 크런치, 흥미, 기획, 전인, 2012, 등장, 기술, 회고, 것
Kkma	미국, 전문, 매체, 테크, 테크크런치, 크런치, 2022, 흥미, 기획, 전인, 2012, 등장...
soynlp. NewsNounExtractor	미국, IT, CES2022, CES2012, 기술



4. 데이터 모델링

29

4-2. 신조어 후보 추출 세부 내용(1)



- 각 카테고리, 주차에 따라 신조어 후보가 추출되는 프로세스 중 1단계 명사 추출 내용

STEP
01 명사 가능성이 높은 모든 단어를 추출

STEP
02 사전에 등록되지 않은 새로운 단어만 추출

STEP
03 기사에서 자주 등장한 단어만 추출

STEP 이전에 등장하지 않았던
04 새로운 단어만 추출

카테고리 별
신조어 후보

Mecab 형태소 분석기의 Part-of-speech (POS) tag를 통한 명사 판별

- Soynlp로 추출한 단어들은 비지도 학습 기반으로 추출되어 학습 데이터에 따라 명사가 아님에도 명사로 추출되는 경우가 발생
- 명사가 아님에도 명사로 추출되는 mecab pos 패턴을 파악하고 불용 POS로 지정
- 복합 명사의 패턴을 갖는 단어만 추출할 수 있게 함

불용 POS 예시

- 유의미한 명사가 아니라고 판단하여 탈락시킨 단어 예시 및 불용 POS

탈락 단어 예시	불용 POS
발행한다고, 출범한다고	동사 + 연결어미
31일, 5만원, 75개	단위명사
-50, 2022-	숫자 + 기호
1km, 11L	숫자 + 영어
온라인으로	명사 + 부사격조사

4-2. 신조어 후보 추출 세부 내용(2)



- 각 카테고리, 주차에 따라 신조어 후보가 추출되는 프로세스 중 2단계 사전 비교 추출 내용

STEP
01 명사 가능성이 높은 모든 단어를 추출

STEP
02 사전에 등록되지 않은 새로운 단어만 추출

STEP
03 기사에서 자주 등장한 단어만 추출

STEP
04 이전에 등장하지 않았던 새로운 단어만 추출

카테고리 별
신조어 후보

사전에 포함되지 않은 단어만 추출하기

- 신조어 및 복합 명사 중에서도 사전에 등록되지 않은 명사만 추출하기 위해 사전과 비교
- 비교 사전 후보

1. Mecab 형태소 분석기 기본 사전

- 21세기 세종 계획 균형 말뭉치 기반으로 약 1,000만 어절로 구성
- Mecab 형태소 분석기의 명사 추출 결과와 비교

2. NIADic

- 국립국어원 우리말샘 사전과 및 SNS 분석기업 인사이트에서 자체 보유한 사전을 기반으로 최신 단어로 구성된 형태소 사전 (총 93만 단어)
- 표준어, 신어 및 생활어, 지역어, 전문용어, 브랜드, 유명인, 장소 등을 포함

3. New Korean Word Corpus

- 한국어 분석을 위하여 만들어진 신조어 코퍼스
- 2002년 05월 06일부터 2019년 09월 29일까지의 NAVER 뉴스 헤드라인에서 추출한 신조어



4. 데이터 모델링

31

4-2. 신조어 후보 추출 세부 내용(2)



- 각 카테고리, 주차에 따라 신조어 후보가 추출되는 프로세스 중 2단계 사전 비교 추출 내용

STEP
01 명사 가능성이 높은 모든 단어를 추출

STEP
02 사전에 등록되지 않은 새로운 단어만 추출

STEP
03 기사에서 자주 등장한 단어만 추출

STEP
04 이전에 등장하지 않았던 새로운 단어만 추출

카테고리 별 신조어 후보

비교 사전에 따른 추출 신조어 비교

Mecab + New Korean Word Corpus

- 형태소 분석 후 본래 복합 명사와 다른 형태의 명사는 모두 추출
- 표준 사전에 등재된 명사임에도 추출되는 복합 명사 다수 존재
(단어 예시) 불균형, 불확실, 이상적, 혼수상태
- 사회 분야 1주차 기사 기준 총 2994개 신조어 후보 추출

=> NIADic + New Korean Word Corpus 사전을 최종 사전으로 채택

NIADic + New Korean Word Corpus

- 사전에 등재된 주로 사용되는 복합 명사들이 일정 부분 필터링되어 추출
- Mecab으로 추출되지 않았던 틱톡, 힐링과 같은 신조어 후보들이 포함되어 추출
- 사회 분야 1주차 기사 기준 총 2884개 신조어 추출



4. 데이터 모델링

32

4-2. 신조어 후보 추출 세부 내용(3)



- 각 카테고리, 주차에 따라 신조어 후보가 추출되는 프로세스 중 3~4단계 추출 내용

STEP
01 명사 가능성이 높은 모든 단어를 추출

STEP
02 사전에 등록되지 않은 새로운 단어만 추출

STEP
03 기사에서 자주 등장한 단어만 추출

STEP
04 이전에 등장하지 않았던 새로운 단어만 추출

카테고리 별
신조어 후보

3단계 : 빈도수 집계 후 자주 등장한 단어만 추출

- 추출한 단어가 등장한 총 기사 수를 신조어의 빈도수로 정의
- 신조어가 등장한 기사 수가 많을수록 해당 신조어가 일부만 사용하는 단어가 아닌 범용적으로 사용되어 많은 사람이 신조어로서 인식할 수 있는 단어라고 판단하여 빈도수를 다음과 같이 정의
- 정의한 빈도수 기준 상위 25% 이상인 단어만 추출

4단계 : 주 별로 추출된 이전 누적 추출 신조어와 비교

- 주 별로 추출 시 이전 주에는 등장하지 않았던 단어만을 신조어 후보로 추가
- 단어가 처음 추출된 주를 해당 단어의 등장 기간으로 정의

4-3. 신조어 후보 추출 결과 집계



- 각 카테고리, 주차에 따라 추출된 신조어 후보 개수 집계

총 304,247개의 기사에서 soynlp로 588,892개의 단어 추출, 신조어 후보 12,688개 추출

추출 소요 시간 : 1h 27m 13s

문화 분야는 기사 수는 적지만 신조어 후보군이 기사 비율대비 많이 추출

전체 카테고리 기사 (22.01.02~22.02.26)										합계	
집계	카테고리	사회	정치	경제	스포츠	국제	연예	문화	IT	사설칼럼	보도자료
전체 기사 수	84,145	64,854	51,851	33,026	23,879	17,462	16,556	5,634	5,595	1,245	304,247
Soynlp 추출 단어 수	136,435	78,742	99,959	33,388	55,445	39,031	70,763	23,519	43,166	8,444	588,892
신조어 후보 수	3,758	1,830	3,205	563	791	438	1,254	399	276	174	12,688

[카테고리 예시] 사회 분야 (22.01.02~22.02.26)

단계 별 단어수	주차	1주차	2주차	3주차	4주차	5주차	6주차	7주차	8주차
Soynlp 추출		16,402	16,940	17,556	17,763	13,198	18,487	18,476	17,613
불용 POS 제거 후		14,157	14,625	15,106	15,300	11,460	15,931	15,938	15,292
사전 비교 후		2,880	3,158	3,316	3,348	2,058	3,616	3,426	3,443
빈도수 기준 추출 후		734	817	880	887	634	925	859	992
이전 신조어와 비교 후		740	556	518	478	281	447	334	404

4-4. 최종 신조어 채택 및 산출물



- 카테고리 별 신조어 후보 취합 후 최종 신조어 사전 구축 및 채택 과정에 따른 산출물 도출

신조어 사전

- 최종 신조어 719개를 추출하여 신조어, 해당 단어 등장 기사 수, 기사 카테고리, 기사 주차, 주차 해당 날짜를 포함한 사전 구축
- 해당 카테고리와 시기에 주로 쓰이는 신조어를 채택하여 사전으로 구축했다는 의미가 있음

new_word	freq	category	week	date1	date2
국민의힘	1329	정치	1주차	20220102	20220108
코로나19	617	사회	1주차	20220102	20220108
오미크론	445	국제	3주차	20220116	20220122
ESG	255	경제	1주차	20220102	20220108
자가검사키트	160	사회	8주차	20220220	20220226
CES2022	134	경제	1주차	20220102	20220108
희망적금	134	경제	8주차	20220220	20220226
MZ	129	경제	1주차	20220102	20220108

신조어 + 복합 명사 사전

- 신조어 사전에 포함된 719개 단어와 복합 명사를 포함한 총 9106개의 신조어 + 복합 명사 사전 구축
- 형태소 분석에서 오분석 될 수 있는 복합 명사를 하나의 단어로 추출하기 용이한 사전을 구축했다는 의미가 있음
- 유사 단어 추출**에 구축한 사전을 활용

word
진흥계획
책임연구원
청년월세
청년정책조정위원회
최종점검위원회
출입명부
태그리스
택지조성

4-5. 단어 유사도 파악 개요

- Word2Vec VS FastText



word2vec

- 구글에서 개발한 딥러닝 기반의 분산표현 워드임베딩, 자연어를 기계가 이해할 수 있는 숫자 벡터값으로 변경하는 기술

- 단순하지만 성능이 우수해서 많이 사용되지만 말뭉치 기반으로 학습을 진행하며 **미등록 단어를** 마주했을 때 성능이 떨어지는 치명적인 단점이 존재

- Word2Vec은 **미등록 단어가 등장했을 때 OOV(Out-of-Value) 에러가 발생하지만** FastText는 단어를 **subword로** **쪼개서 학습을 하기 때문에 위의 문제를 보완할 수 있음**

fastText

- 페이스북에서 개발한 Word2Vec 문제점을 보완하기 위해 나온 언어 모델

- FastText는 단어를 대할 때 단어를 **다시 n-gram으로 구성하여 내부적으로 나누는데**, 이를 **subword**라고 하며 이렇게 내부 단어를 사용하는 것을 “단어 이하 수준의 정보를 사용한다”라고 함



4. 데이터 모델링

36

4-6. FastText 모델의 선정 이유



- FastText를 활용한 유사 단어 추출

- Facebook의 AI Research lab에서 만든 단어 임베딩 및 텍스트 분류 학습 라이브러리
- 단어를 벡터로 표현하여 단어 간의 의미 파악 가능
- 내부 단어(subword)를 학습하여 모르는 단어에 대해서도 다른 단어와 유사도를 구할 수 있음
- 빈도수가 적었던 단어에 대해서도 일정 수준의 성능 보장
- 신조어의 특징을 고려하였을 때 유사 단어 추출에 적합한 라이브러리라 판단

코로나19 n-gram = 4일 때,

<코로나, 코로나1, 로나19, 나19>

글자 단위 n_gram을 지정하여 내부단어 학습

=> 신조어 학습에 용이

4-7. FastText를 활용한 단어 유사도 파악(1)

- 단어 유사도 분석 프로세스



01

사용자 정의 사전 추가

Komoran 형태소 분석기 안에 SoyNlp와 고유명사 추출 프로세스의 결과로 생성된 명사 사전과 한국정보화진흥원에서 배포한 NIADic의 명사 사전을 사용자 정의 사전으로 추가하여 복합명사를 효과적으로 추출 가능하게끔 함

03

토크나이징

말뭉치를 한 줄씩(문장 단위) 읽어와 1단계에서 정의한 사전에 명사로 토크나이징 진행
토크나이징된 데이터를 문장단위로 다시 문자열로 변환하여 새로운 txt파일로 저장

02

전처리

SoyNlp의 전처리 과정과 거의 일치하지만 효과적인 단어 유사도 파악을 위해 말뭉치를 문장 단위로 구분하고, 문장단위로 구분된 말뭉치를 토크나이징을 하기 위해 txt파일로 저장

04

FastText로 학습

토크나이징 된 말뭉치를 FastText 모델에 학습
학습에 활용된 파라미터는 다음과 같음
sg = 1(skipgram),
word_ngrams(연관 지을 주변 단어 윈도우 사이즈) = 3,
min_count(최소 등장 횟수 제한) = 20



4. 데이터 모델링

38

4-7. FastText를 활용한 단어 유사도 파악(2)



- 사전 구축 과정 예시

(원문) 그는 월드컵에서 노메달로 부진하다 세계선수권에서 준우승을 차지했다

1차 – (Komoran + SoyNlp로 뽑아낸 신조어사전)

-> 월드컵 / 노메달 / 부진 / 세계선 / 수권 / 준우승 / 차지

2차 – (Komoran + SoyNlp로 뽑아낸 신조어사전 + NIADic)

-> 월드컵 / 에서 / 노메달 / 부진 / 하다 / 세계선수권 / 에서 / 준우승 / 차지

3차 – (Komoran + SoyNlp로 뽑아낸 신조어사전(불용어추가 및 후처리) + NIADic 정제)

-> 월드컵 / 노메달 / 부진 / 세계선수권 / 준우승 / 차지

위 3가지 과정을 거치면서 원문에 대하여 불필요한 형태소(명사제외)가 제외됨을 확인.
최종적으로 Komoran 형태소분석기 베이스에 SoyNlp로 뽑아낸 불용어 제거와 **후처리가 진행된 신조어 사전**과 명사만 뽑아낸 NIADic을 **최종 사전으로 구축하였음**

4-7. FastText를 활용한 단어 유사도 파악(3)

- 단어 유사도 분석 프로세스



01 사용자 정의 사전 추가

Komoran 형태소 분석기 안에 SoyNlp와 고유명사 추출 프로세스의 결과로 생성된 명사 사전과 한국정보화진흥원에서 배포한 NIADic의 명사 사전을 사용자 정의 사전으로 추가하여 복합명사를 효과적으로 추출 가능하게끔 함

03 토크나이징

말뭉치를 한 줄씩(문장 단위) 읽어와 1단계에서 정의한 사전에 명사로 토크나이징 진행
토크나이징된 데이터를 문장단위로 다시 문자열로 변환하여 새로운 txt파일로 저장

02 전처리

SoyNlp의 전처리 과정과 거의 일치하지만 효과적인 단어 유사도 파악을 위해 말뭉치를 문장 단위로 구분하고, 문장단위로 구분된 말뭉치를 토크나이징을 하기 위해 txt파일로 저장

04 FastText로 학습

토크나이징 된 말뭉치를 FastText 모델에 학습
학습에 활용된 파라미터는 다음과 같음
sg = 1(skipgram),
word_ngrams(연관 지을 주변 단어 윈도우 사이즈) = 3,
min_count(최소 등장 횟수 제한) = 20



4. 데이터 모델링

40

4-7. FastText를 활용한 단어 유사도 파악(4)



- 전처리 전후 과정 예시

전처리 전 기사 원문 예시

[경향신문] 방 안에 있던 딸이 자해할까봐 문손잡이를 부쉈다 재물손괴 혐의로 기소유예 처분을 받은 양모에 대해 헌법재판소가 검찰의 기소유예 처분을 취소해야 한다고 판단했다. 현재는 A씨가 "기소유예 처분을

불필요한 데이터들의 제거와 문장단위 구분이 필요

전처리 후 기사 원문 예시

방 안에 있던 딸이 자해할까봐 문손잡이를 부쉈다 재물손괴 혐의로 기소유예 처분을 받은 양모에 대해 헌법재판소가 검찰의 기소유예 처분을 취소해야 한다고 판단했다↓
현재는 A씨가 기소유예 처분을 취소해달라 검찰을 상대로 청구한 헌법소원심판 사건에서 재판관 전원일치 의견으로 인용 결정했다고 밝혔다↓

A씨는 지난해 의붓딸이 방문을 열어주지 않자 문손잡이를 부쉈다가 재물손괴 혐의로 기소유예 처분을 받았다↓

검찰에서 A씨는 의붓딸이 과거에 자해를 적이 있는데 사건 당일 방에서 오랜 시간 인기척이 나지 않아 확인하려고 펜치로 문을 두드린 것이라고 했다↓

검찰은 서로가 오해한 일이며 처벌을 원하지 않는다는 취지의 처벌불원서를 의붓딸로부터 받고 A씨를 기소유예 처분했다↓

기소유예는 범죄에 해당되지만 여러 정황을 참작해 재판에 넘기지 않는 것을 말한다↓

기소유예 처분은 법원의 재판을 통해서 유무죄 판단을 받을 없고 현재의 헌법소원심판을 통해서만 취소할 있다↓

전처리 진행 후 불필요한 데이터가 제거되었고 문장단위로 구분이 됨

4-7. FastText를 활용한 단어 유사도 파악(5)

- 단어 유사도 분석 프로세스



01 사용자 정의 사전 추가

Komoran 형태소 분석기 안에 SoyNlp와 고유명사 추출 프로세스의 결과로 생성된 명사 사전과 한국정보화진흥원에서 배포한 NIADic의 명사 사전을 사용자 정의 사전으로 추가하여 복합명사를 효과적으로 추출 가능하게끔 함

03 토크나이징

말뭉치를 한 줄씩(문장 단위) 읽어와 1단계에서 정의한 사전에 명사로 토크나이징 진행
토크나이징된 데이터를 문장단위로 다시 문자열로 변환하여 새로운 txt파일로 저장

02 전처리

SoyNlp의 전처리 과정과 거의 일치하지만 효과적인 단어 유사도 파악을 위해 말뭉치를 문장 단위로 구분하고, 문장단위로 구분된 말뭉치를 토크나이징을 하기 위해 txt파일로 저장

04 FastText로 학습

토크나이징 된 말뭉치를 FastText 모델에 학습
학습에 활용된 파라미터는 다음과 같음
sg = 1(skipgram),
word_ngrams(연관 지을 주변 단어 윈도우 사이즈) = 3,
min_count(최소 등장 횟수 제한) = 20



4. 데이터 모델링

42

4-7. FastText를 활용한 단어 유사도 파악(6)



- 토크나이징 전 vs 후 예시

전처리 된 기사 예시

방 안에 있던 딸이 자해할까봐 문손잡이를 부쉈다 재물손괴 혐의로 기소유예 처분을 받은 양모에 대해 헌법재판소가 검찰의 기소유예 처분을 취소해야 한다고 판단했다↓
현재는 A씨가 기소유예 처분을 취소해달라 검찰을 상대로 청구한 헌법소원심판 사건에서 재판관 전원일치 의견으로 인용 결정했다고 밝혔다↓
A씨는 지난해 의붓딸이 방문을 열어주지 않자 문손잡이를 부쉈다가 재물손괴 혐의로 기소유예 처분을 받았다↓
검찰에서 A씨는 의붓딸이 과거에 자해를 적이 있는데 사건 당일 방에서 오랜 시간 인기척이 나지 않아 확인하려고 펜치로 문을 두드린 것이라고 했다↓
검찰은 서로가오해한일이며처벌을원하지않는다는 취지의 처벌불원서를 의붓딸로부터 받고 A씨를 기소유예 처분했다↓
기소유예는 범죄에 해당되지만 여러 정황을 참작해 재판에 넘기지 않는 것을 말한다↓
기소유예 처분은 법원의 재판을 통해서 유무죄 판단을 받을 없고 현재의 헌법소원심판을 통해서만 취소할 있다↓

토크나이징 후 기사 예시

방 안 딸 자해 문손잡이 재물 손괴 혐의 기소 유예 처분 양모 대해 헌법 재판 소가 검찰 기소 유예 처분 취소 판단↓
현재 씨 기소 유예 처분 취소 해달 검찰 상대 청구 한 헌법소원심판 사건에서 재판관 전원일치 의견 인용 결정↓
씨 지난해 의붓딸 방문 열어 주지 문손잡이 재물 손괴 혐의 기소 유예 처분↓
검찰에서 씨 의붓딸 과거 자해 적 사건 당일 방에 오래 시간 인기 척이 나지 확인 하려 펜치 문 두드리 것↓
검찰 서로 가오 해한 일이 처벌 취지 처벌 불원 서 의붓딸 씨 기소 유예 처분↓
기소 유예 범죄 해당 되지 정황 참작 해 재판 것 말↓
기소 유예 처분 법원 재판 통해 유무죄 판단 재의 헌법소원심판 통해 취소↓

4-7. FastText를 활용한 단어 유사도 파악(7)

- 단어 유사도 분석 프로세스



01 사용자 정의 사전 추가

Komoran 형태소 분석기 안에 SoyNlp와 고유명사 추출 프로세스의 결과로 생성된 명사 사전과 한국정보화진흥원에서 배포한 NIADic의 명사 사전을 사용자 정의 사전으로 추가하여 복합명사를 효과적으로 추출 가능하게끔 함

03 토크나이징

말뭉치를 한 줄씩(문장 단위) 읽어와 1단계에서 정의한 사전에 명사로 토크나이징 진행
토크나이징된 데이터를 문장단위로 다시 문자열로 변환하여 새로운 txt파일로 저장

02 전처리

SoyNlp의 전처리 과정과 거의 일치하지만 효과적인 단어 유사도 파악을 위해 말뭉치를 문장 단위로 구분하고, 문장단위로 구분된 말뭉치를 토크나이징을 하기 위해 txt파일로 저장

04 FastText로 학습

토크나이징 된 말뭉치를 FastText 모델에 학습
학습에 활용된 파라미터는 다음과 같음
sg = 1(skipgram),
word_ngrams(연관 지을 주변 단어 윈도우 사이즈) = 3,
min_count(최소 등장 횟수 제한) = 20



4. 데이터 모델링

44

4-7. FastText를 활용한 단어 유사도 파악(8)



- FastText 하이퍼파라미터 튜닝 (n_grams : {3,4,5,6,7} , min_count : {3,5,10,20})

입력 단어 : 중앙수습사고본부

	1	2	3	4	5
Ngram=3,min_count =3	중수본	손영래	반장	사고수습대책본부	복지부
Ngram=4,min_count =3	중수본	손영래	사고수습대책본부	반장	복지부
Ngram=5,min_count =3	중수본	손영래	반장	사고수습대책본부	복지부
			⋮		
			중략		
			⋮		
Ngram=6,min_count =20	중수본	손영래	반장	사고수습대책본부	복지부
Ngram=7,min_count =20	중수본	손영래	반장	사고수습대책본부	복지부



4. 데이터 모델링

45

4-7. FastText를 활용한 단어 유사도 파악(9)



- FastText 하이퍼파라미터 튜닝 (n_grams : {3,4,5,6,7} , min_count : {3,5,10,20})

입력 단어 : 한국형반값임대프로젝트

	1	2	3	4	5
Ngram=3,min_count =3	리스너프로젝트	코로나시대의경제대책	구상안	헛공약	정책공약
Ngram=4,min_count =3	리스너프로젝트	코로나시대의경제대책	헛공약	폭탄공급	구상안
Ngram=5,min_count =3	리스너프로젝트	구상안	헛공약	코로나시대의경제대책	가상자산개미투자자안심투자
⋮ ⋮ 중략 ⋮ ⋮					
Ngram=6,min_count =20	인생횟집	정책공약	족발집	노후아파트	우리동네공약
Ngram=7,min_count =20	정책공약	코로나피해실질보상촉구및정부규탄대회	족발집	노후아파트	노동공약

4-7. FastText를 활용한 단어 유사도 파악(10)



- FastText 최적화 파라미터

n_gram = 3, min_count = 20 일때, 단어간 유사도가 적합하게 도출되어
최종 파라미터로 선택

	1	2	3	4	5
스포츠유ти리타차	스포츠유ти리티차량	스포티지	쉐보레	픽업트럭	스타리아
이재명게이트	대장동녹취록	윤석열게이트	녹취록	정영학녹취록	김만배
중앙수습사고본부	중수본	손영래	반장	사고수습대책본부	복지부
코로나19	감염증	코로나	코로나바이러스	신종	대유행
한국형반값임대프로젝트	코로나피해실질보상촉구및정부규탄대회	윤석열정부	노후아파트	변화와쇄신	정책공약
오미크론	변이	오미크론변이	전파력	변이바이러스	우세종화
멸콩	달파멸콩	멸공챌린지	멸공인증	일베놀이	멸공



5. 자바 웹 프로그래밍

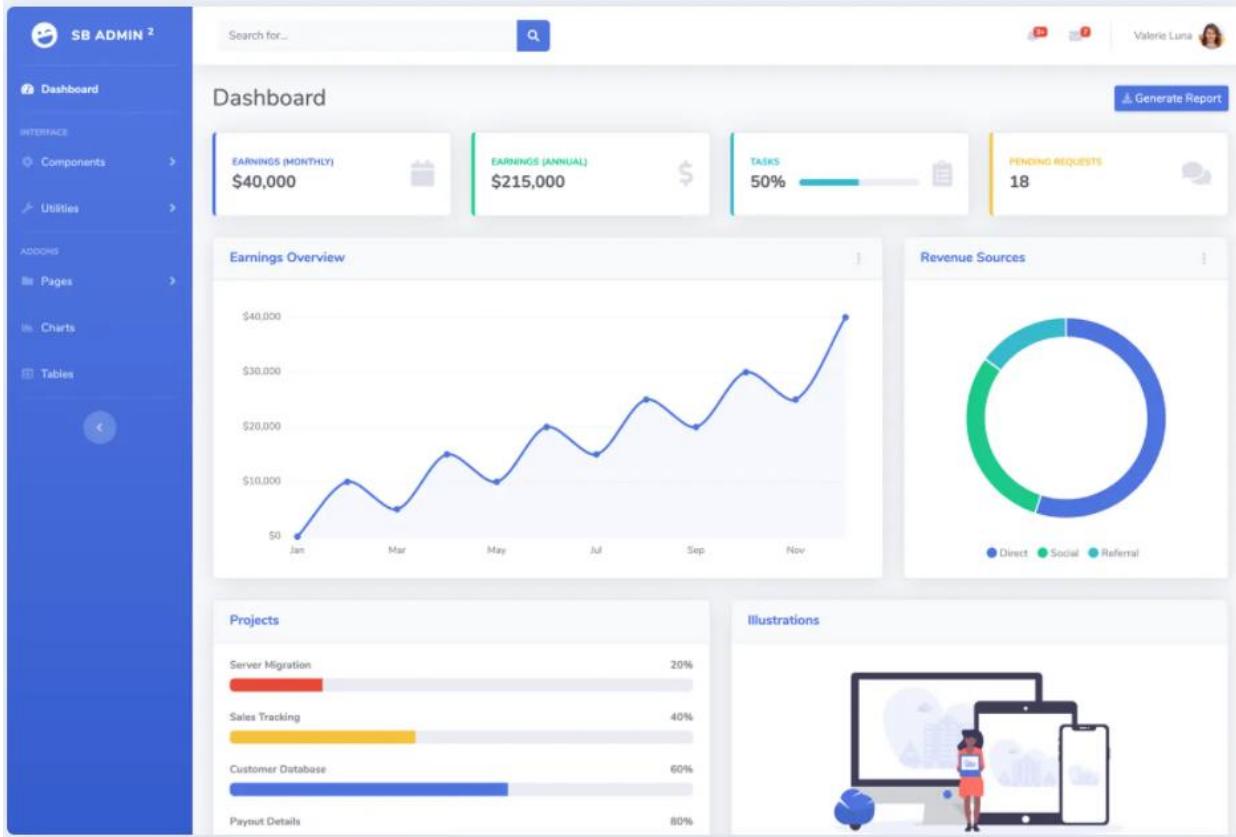
발표자 : 김지현

5. 자바 웹 프로그래밍

5-1. 템플릿 선택



- 스타트 부트스트랩에서 DashBoard용 템플릿 선택



데이터 수집부터 전처리 신조어 추출까지 각 과정에서 진행된 프로세스를 요약 정보로 표현하기 위해 Dashboard 형 템플릿 선택

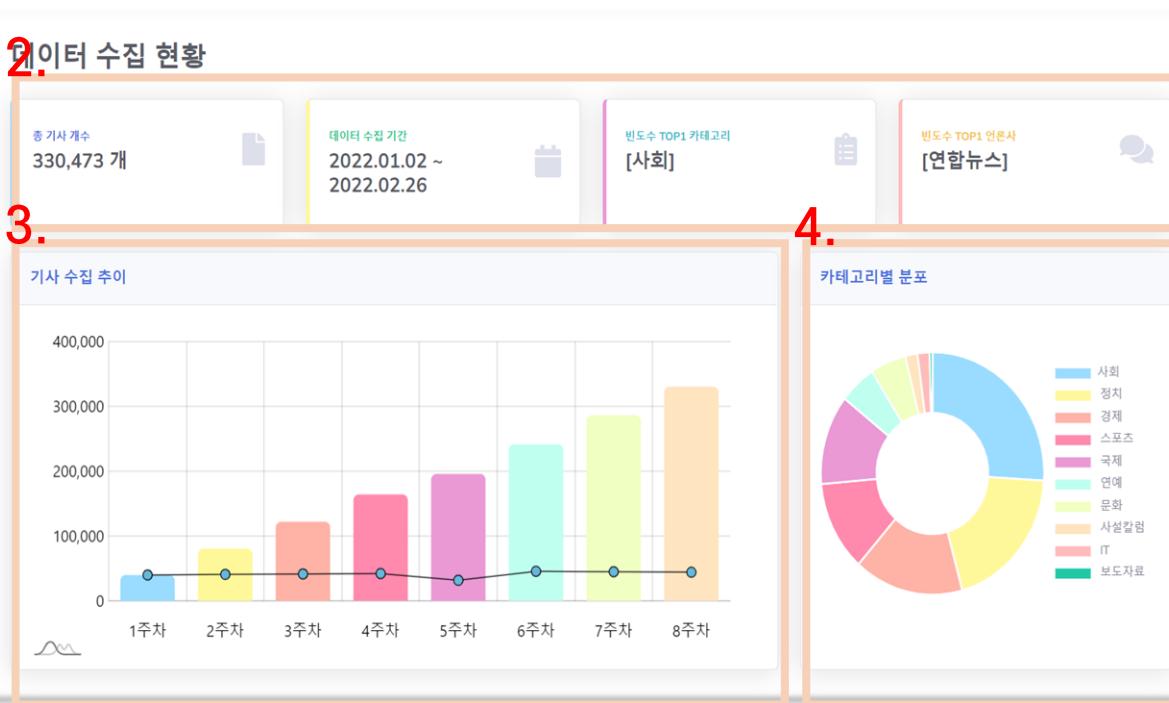
5-2. 웹 페이지 구조도

- 상세 웹 페이지 구조

1.



2.



3.

카테고리별 분포

1. Side-Bar :

- DashBoard 페이지(index.jsp)
- 신조어리스트 페이지(table.jsp)
- 전처리과정 페이지(prepro.jsp)

2. 수치로 표현 가능한 데이터 :

- 카드 형식으로 정보 제공

3. Amcharts5.js :

- 누적그래프
- Word Cloud

4. Chart.js :

- Pie Chart
- Bar Chart

5-3. 각 단계별 시작화(1)

- 데이터 수집 현황

데이터 수집 현황



1. 수집에 대한 요약 정보 :

- 수집된 기사의 총 개수, 기사 수집 기간, 빈도수가 가장 높은 카테고리와 언론사.

2. 기사 수집 추이 :

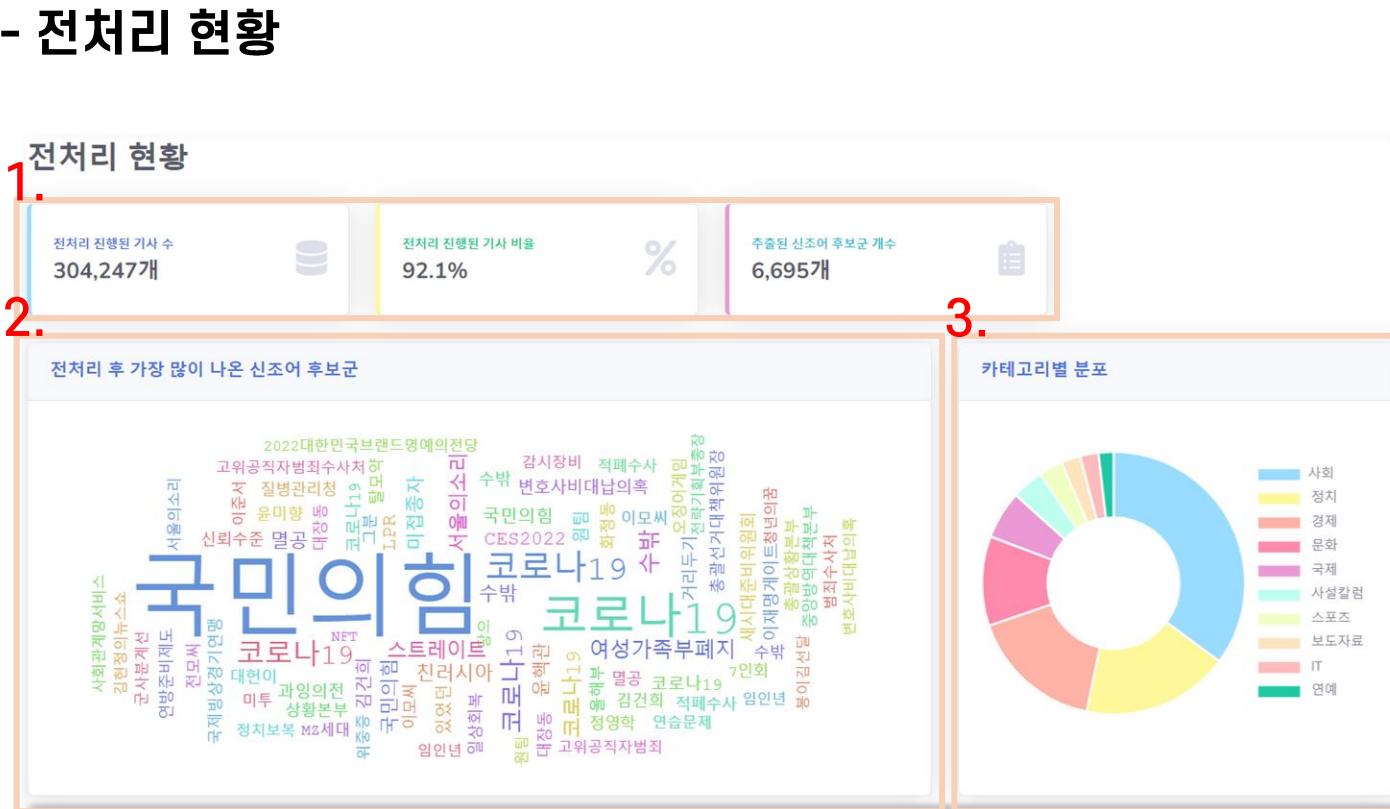
선 그래프 – 주차별로 수집된 기사의 수
막대 그래프 – 수집된 기사의 누적합

3. 카테고리별 분포 :

- 총 수집된 기사에서 카테고리의 분포도를 확인 가능.

5. 자바 웹 프로그래밍

5-3. 각 단계별 시각화(2)



1. 전처리의 요약 정보

- 수집된 기사 중 전처리가 진행된 기사의 수(영문 기사, 포토 뉴스 제외 등)와 그 비율, 전처리 과정을 통해 추출된 신조어 후보군의 개수

2. 워드 클라우드 :

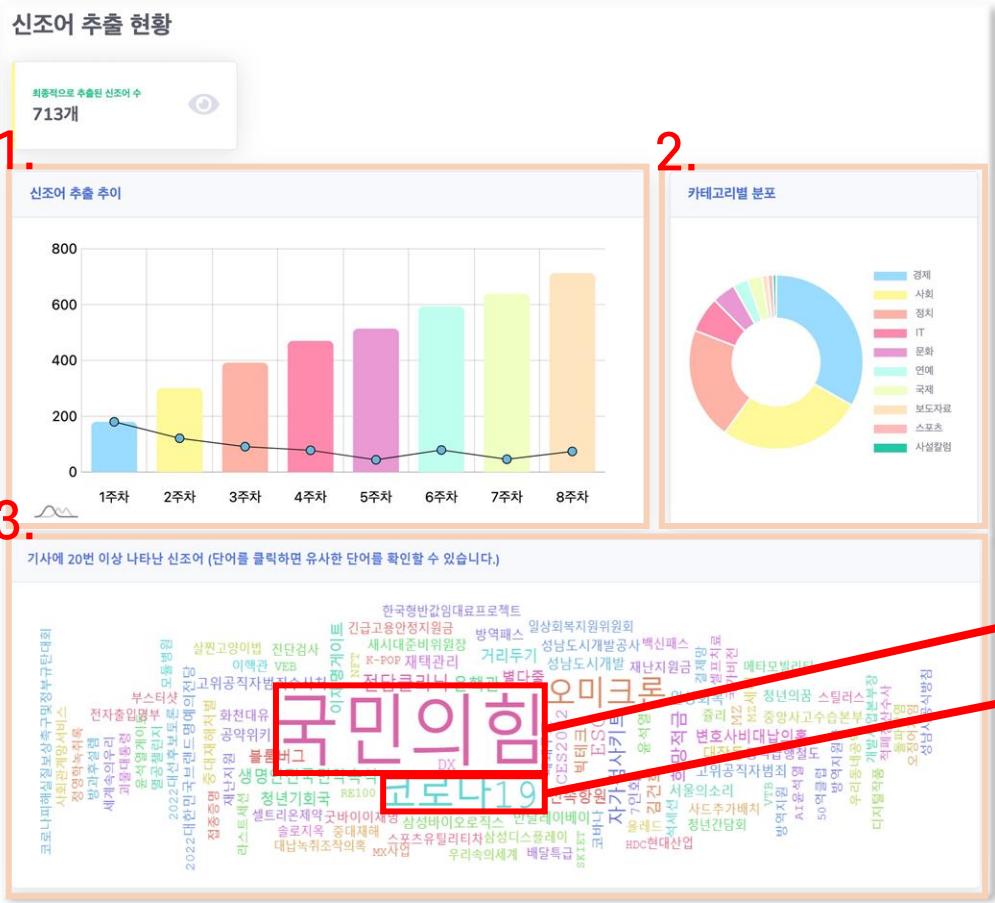
- 신조어 후보군의 빈도수를 시각화
-> 오미크(오미크론), 서비스(서비스), 여론조사(여론조사) 등 마지막 글자를 조사로 인식하여 제대로 추출되지 않은 것을 확인, 전처리 과정을 수정하거나 불용어 사전에 추가

3. 카테고리별 분포 :

- 추출된 신조어 후보군의 카테고리
분포를 시각화

5-3. 각 단계별 시각화(3)

- 신조어 추출 현황



1. 신조어 추출 추이 :

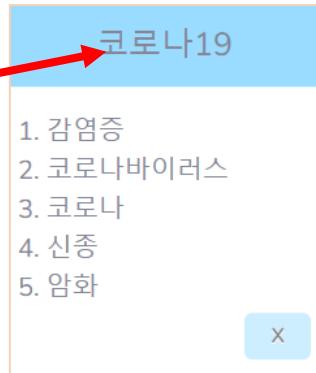
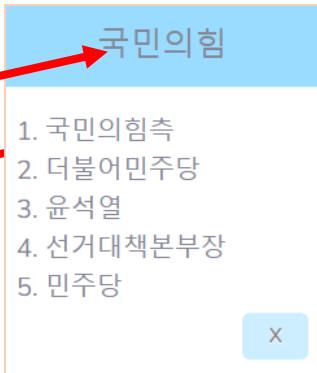
- 주차별로 추출된 신조어의 개수와 그 누적합을 시각화
- 1주차 이후로 신조어의 수가 점차적으로 줄어드는 것을 통해 구축된 신조어 사전이 잘 적용되고 있다는 것을 확인.

2. 워드 클라우드 :

- 한 카테고리에서 20번 이상 등장한 신조어를 시각화,
- 단어 클릭 시 유사도 높은 단어를 모달 창으로 표시.

3. 카테고리별 분포 :

- 추출된 신조어의 카테고리 분포를 시각화.





5. 자바 웹 프로그래밍

53

5-4. 기타 페이지(1)

- 신조어 리스트

신조어 리스트					
신조어	빈도수	카테고리	단어가 등장한 주차	해당 주차의 시작 날짜	해당 주차의 끝 날짜
2022신년정치대개혁토론회	10	정치	1주차	20220102	20220108
50억클럽	25	사회	1주차	20220102	20220108
CES2022	134	경제	1주차	20220102	20220108
DX	115	경제	1주차	20220102	20220108
DX부문	16	경제	1주차	20220102	20220108
ESG	255	경제	1주차	20220102	20220108
ESG경영	16	경제	1주차	20220102	20220108
GV80	12	경제	1주차	20220102	20220108
K-푸드	13	경제	1주차	20220102	20220108
MZ	129	경제	1주차	20220102	20220108

Showing 1 to 10 of 356 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [36](#) Next

추출된 신조어 리스트를 보여주는 페이지

컬럼을 클릭할 시 sort 가능

Search를 통해 신조어 검색 가능

- 전처리 및 모델링 과정

전처리 과정

기사 원문이 전처리 되는 과정을 순차적으로 보여드리는 페이지입니다.
전처리가 완료된 후에는 모델링 버튼을 통해 신조어가 추출되는 과정을 차트로 확인할 수 있습니다.



전처리 및 모델링 과정을 보여주는 페이지

Test 버튼을 통해 기사가 어떻게 전처리 되어가는지 확인 가능.

전처리가 끝난 후에는 모델링 버튼 활성화



5-4. 기타 페이지(2)

- 모델링 과정

전처리 과정 (1)

전처리 과정 (2)

전처리 과정 (3)

전처리 과정 (4)

전처리 완료

미국인이 가장 우려하는 외교 현안은 북한 미사일 발사인 것으로 나타났다 러시아의 우크라이나 침공이든 행정부가 우크라이나 사태에 집중하고 있지만 유권자들은 미국 본토를 겨냥할 있는 북한 미사일 시간 미국 폭스뉴스는 유권자 조사한 결과 응답이 나타났다고 밝혔다 북한 미사일 발사가 였다 반면 을 선정된 북한 미사일 발사와 러시아의 우크라이나 침공 위기 가운데 미국 유권자는 북한 미사일을 걱정 한에 대해 우려된다는 응답은 북·미 대화가 본격화되기 전인 조사와 비슷한 수준으로 나타났다 조사에 발사가 우려된다고 답했다 최근 북한 미사일 방어 체계를 무력화시킬 있는 극초음속 미사일 발사에 사일 실험 중단 조치를 철회하겠다고 밝히면서 미국인은 한반도 정세가 이전으로 돌아갈 가능성이 있 해 가장 우려된다고 꼽은 이슈는 인플레이션 이었다 지난달 물가가 만에 최고 수준으로 급등한 데다 소 19 오미크론 변이 확산으로 물가상승이 장기화할 것이라는 관측에 따른 것으로 풀이된다 높은 범죄율 경지대 이민자 유입 유권자 억압 부정선거 등을 북한 미사일 발사나 우크라이나 사태보다 낮았다 바이 밝힌 가운데 질문에 바이든 대통령을 뽑겠다고 답한 반면 다른 사람을 뽑겠다는 응답은 였다 이는 도널 응답자 다음 대선에서 트럼프 대통령 아닌 다른 후보를 뽑겠다고 답한 것보다 높다 바이든 대통령 국정 로나19 대응엔 했고 외교정책과 경제정책은 각각 답했다

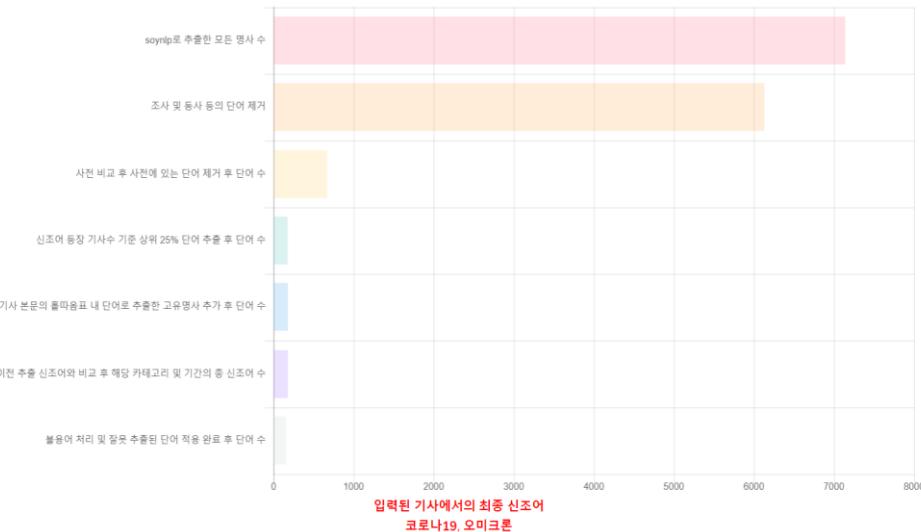
모델링

활성화 된 모델링 버튼

클릭 시 모델링 과정을 차트로 표현



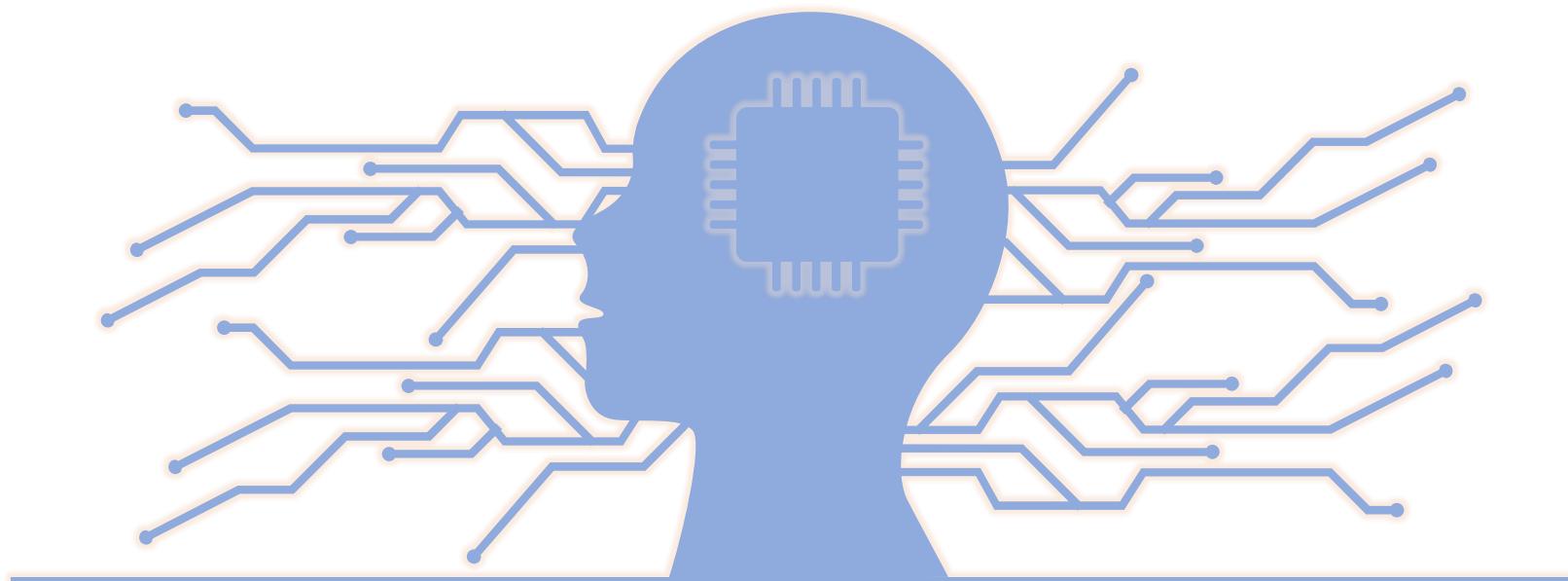
국제 분야의 4주차 기사 (총 2873개)
20220123 ~ 20220129



각 과정별로 추출되는 단어의 수를 막대 그래프 형식으로 시각화
차트 상단에는 해당 기사가 등장한 분야와 날짜를 표현
차트 내부에는 모델링이 진행되면서 줄어드는 단어 수를 표시
차트 하단에는 최종적으로 추출된 신조어 출력

5-5. 시연

- 시연하러 가기



시연 링크



6. 결론 및 추후 과제

발표자 : 노현진

6-1. 프로젝트 결론

- 각 단계 별 결과 및 최종 결론



데이터 수집 및 적재

다음 뉴스 기사에서 한달간 총 330,473건의 기사를 수집. 기사 내에서 어떤 요소가 필요할지 모르기 때문에 최대한 많은 부분을 수집. 데이터를 적재할 때 충분한 회의를 통해 팀원들이 필요한 데이터를 적시에 확인할 수 있도록 데이터를 정제.

33만 여개 수집 기사 전처리

복합명사를 추출하기 위해 전처리 과정에서 정규표현식과 데이터 컬럼 정보를 활용하여 불필요한 데이터를 제거 후 NewsNounExtractor에 돌릴 수 있는 형태로 변환시킴.

신조어 사전 구축 및 유사 단어 추출

카테고리 및 주 별로 기사 데이터를 통해 총 12,688개의 신조어 후보를 추출. 이를 통해 719개 단어의 신조어 사전과 9106개 단어의 신조어+복합 명사 사전 구축. 구축한 사전을 통해 형태소 분석에서 신조어 토큰을 추출. FastText를 사용하여 신조어와의 단어 유사도를 찾기 위해 신조어와 유사한 단어들을 도출.

전체 프로세스 및 결과 대시보드 제작

전체 수집 데이터 집계 및 신조어 추출 프로세스 차트 시각화. ajax를 통해 flask와 연동 후 전처리 및 모델링 과정을 통해 받아온 데이터를 워드 클라우드로 시각화. Rest API를 사용하여 기사 원문 전처리 및 모델링 과정 시각화.

뉴스 기사 데이터를 수집, 정제하고 모델링을 통해 **신조어를 추출**. 해당 과정을 **시각화** 하여 웹으로 구현
Mission Complete



6. 결론 및 추후 과제

58

6-2. 결과를 통해 배운 점

- 각 구성원 별 프로젝트 진행 후 배운 점

노현진

- 각 구성원들에게 업무를 분담하고 진행 상황을 파악하며, 스케줄 관리 및 팀원 관리를 통한 리더 역량 강화
- 몽고DB 쿼리 사용법 숙지
- 몽고DB와 자바, 파이썬 연동 숙련

박현태

- 전처리 과정에서 시행착오를 거치며 패턴 파악
- 전처리 순서의 중요성 알게 됨
- 자연어 임베딩 과정에서 전처리, 토크나이징 및 모델링 프로세스의 중요성 알게 됨
- 목적에 맞는 결론을 도출하기 위해 전체 프로젝트의 프로세스 숙지 및 각 단계 정리의 필요성을 실감

안수빈

- 다양한 형태소 분석기의 성능 및 특징에 이해
- 비지도 학습 기반 한국 형태소 분석 패키지 soynlp의 이해 및 nounextractor나 tokenizer를 적용 방법 습득
- 형태소 분석기에 사용자 지정 사전을 추가하면 복합명사 및 신조어를 하나의 명사 토큰으로 추출할 수 있는 형태소 분석기로 변형이 가능하다는 것을 알게 되었음

김지현

- RestAPI 사용법 습득
- ajax를 통한 비동기 요청 방식을 활용할 수 있는 역량 강화
- JSON 형태의 데이터 구조를 다룰 수 있는 역량 강화
- MongoDB 사용법 습득
- amchart, chart.js와 같은 차트 라이브러리 사용법 습득 및 데이터 시각화 역량 강화



6. 결론 및 추후 과제

59

6-3. 추후 과제



- 최종 목표 달성을 위해 추가적으로 필요한 과제

프로젝트의 목적

- 뉴스 기사에 새로이 등장하는 신조어를 추출하고 이를 사전화 시킴으로써
한국어 NLP의 발전에 기여.

1. 일회성이 아닌 지속적인 신조어 리스트의 업데이트가 필요
2. 기존 데이터에 적용된 방식이 아닌 신규 데이터 추가에 따른 신조어 자동 추출
시스템 구축을 위한 알고리즘 고도화 필요
3. 신조어 관련 다양한 요인을 활용하여 신조어 예측 모델 구축



7. 참고 문헌

60

참고 문헌 리스트



논문

1. 포털 뉴스 기사를 이용한 신조어 목록 자동 추출. 한국HCI학회 학술대회
김정욱, 정지완, 차미영(2020)
2. 국내 온라인 커뮤니티 게시글에 기반한 신조어 추출 방법 및 형태소 분석 적용에
관한 실증적 연구. 연세대학교 정보대학원 석사 학위 논문
김한준(2018)
3. 신조어 연어의 형성 원리(서울대학교 기초교육원), 장경현(2011)
4. 다양한 임베딩 모델들의 하이퍼 파라미터 변화에 따른 성능 분석. 제30회 한글 및
한국어 정보처리 학술대회 논문집, 이상아 외(2018)

도서

1. Deep Learning in Natural Language Processing(자연어 처리와 딥러닝) -
리덩, 양 리우 지음(2021)
2. 한국어 임베딩 - 이기창 저(2019)