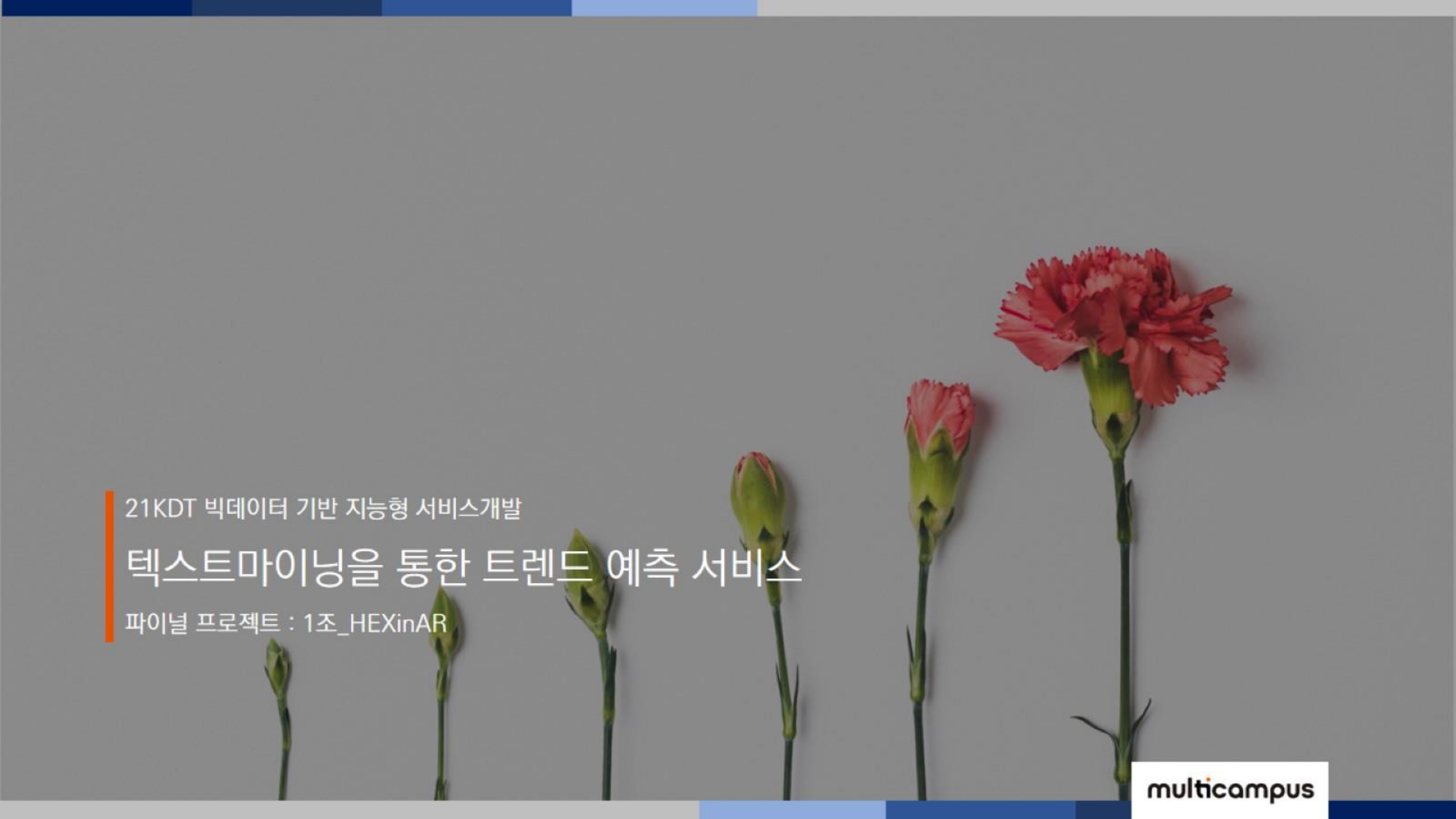


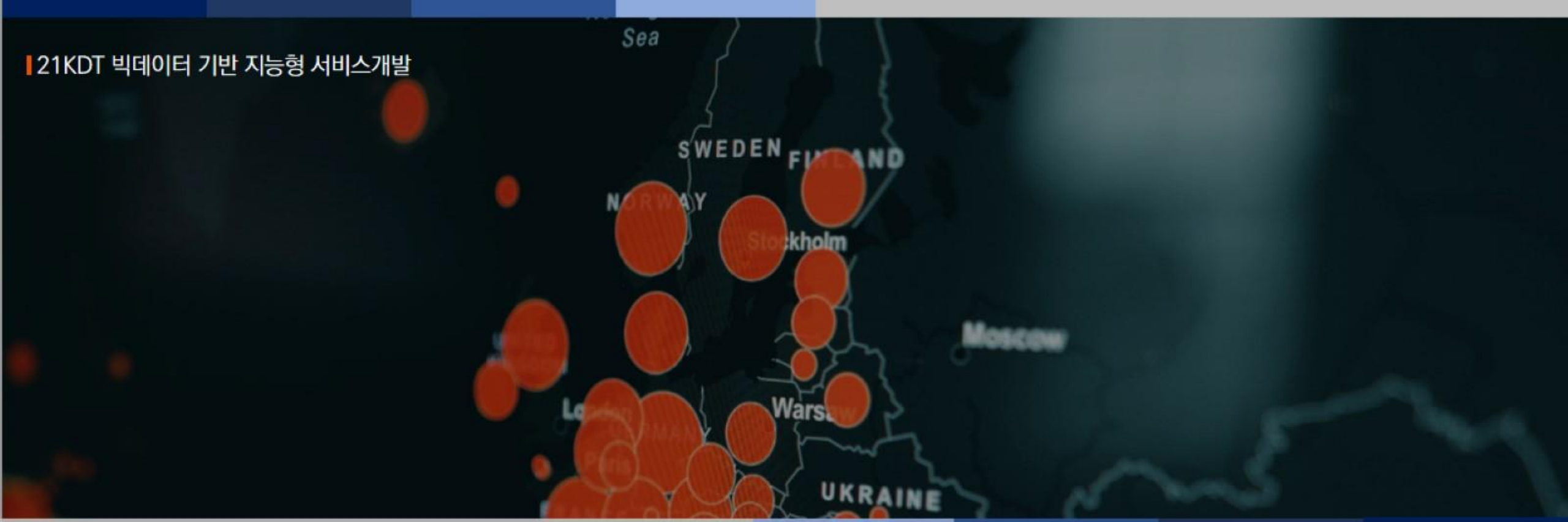
21KDT 빅데이터 기반 지능형 서비스개발

텍스트마이닝을 통한 트렌드 예측 서비스

파이널 프로젝트 : 1조_HEXinAR

multicampus

목 차



프로젝트 배경

multicampus

프로젝트 개발 배경

미래변화의 트렌드를 파악하고 미래의 핵심기술을 선별하기 위하여 주요 선진국과 컨설팅 기관들은 주기적으로 미래 트렌드를 분석해 결과를 발표하고 있다. 온라인에서 생산되는 텍스트 형태의 비정형 데이터가 실제로 경제 및 사회에 미치는 영향력이 높아짐에 따라 빅데이터를 활용한 미래예측 연구가 진행되고 있으나, 분석기술의 어려움으로 대부분 전문가의 지식과 의견에 따라 미래를 전망하는 방식에 의존하는 현실이다.

본 서비스는 이벤트 / 사고, 주요 정책과 이슈에 대한 미래신호를 탐지하여 예측모형을 제시하고자 한다. 해당 정보는 정부 및 지방자치단체의 정책 수립, 기업의 사업 전망 및 신규사업 발굴, 투자자들의 투자 아이디어 및 리스크 점검 등 다양한 분야에 사용될 것으로 예상되어 여러 의사결정의 나침반이 되고자 한다.

프로젝트 개발 배경

문제는 타이밍!

저성장 시대에 성장성이 높은 산업에 대한 프리미엄은 높아져가고, 성장이 정체된 산업은 소외되는 쏠림현상이 가속화되고 있다.

기업들은 생존을 위한 사업재편이 끊임없이 이루어져야 하는 환경에 처해있다. 이러한 변화의 흐름에 조금이라도 빠르게 대처하는 기업들은 생존할 것이며, 제때 대응하지 못하는 기업들은 따라잡기 위한 비용부담이 가중되면서 기업의 존폐까지 이르는 리스크가 존재할 것으로 예상된다.

대기업조차도 기존의 사업을 모두 없애고 획기적인 사업으로 재편될 만큼 세상은 어느 때보다 격변하고 있고, 이에 대한 적절하고, 적극적인 대응이 이루어지지 않으면 비용으로 전가될 것이다.

서비스에서 제공하는 미래 예측방법론

약신호(weak signal)는 현재에는 주목하기 힘들 정도로 노출 빈도나 강도가 낮지만, '미래에 가능한 변화의 징후'로 정의할 수 있다. 시간이 흐르면서 약신호는 강신호(strong signal)로, 강신호는 다시 트렌드(trend)나 메가트렌드(mega trend)로 발전할 수 있는 측면에서 약신호 분석(weak signal analysis)은 미래를 예측하는 가장 적극적인 기술이라 할 수 있다.

웹 뉴스의 문서를 수집하여 텍스트마이닝 분석을 통해 생성된 단어빈도(Term Frequency, TF)와 문서빈도(Document Frequency, DF)를 신호(signal), 이슈(issue)의 2차원의 공간으로 미래신호(future signal)를 설명한다. 단어빈도, 문서빈도, 발생빈도 증가율을 이용하여 KEM(Keyword Emergence Map)과 KIM(Keyword Issue Map)의 키워드 포트폴리오를 작성하고 이를 이용하여 약신호를 선별한다.

KEM은 가시성을 보여주는 것으로 DoV(Degree of Visibility)를 산출하고, KIM은 확산정도를 보여주는 것으로 DoD(Degree of Diffusion)을 산출한다.

$$DoV_{ij} = \left(\frac{TF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}$$

$$DoD_{ij} = \left(\frac{DF_{ij}}{NN_j}\right) \times \{1 - tw \times (n - j)\}$$

(*NN*: 전체 문서 수, *TF*: 문서빈도, *tw*: 시간가중치, *n*은 전체시간구간, *j*: 시점)

서비스 모델 미래예측 절차

미래신호 탐색 절차

1. 단어빈도(Term Frequency, TF)는 각 문서에서 단어별 출현 빈도를 산출한 후, 문서별 출현 빈도를 합산하여 산출할 수 있다.
문서빈도(Document Frequency, DF)는 특정 단어가 들어가는 문서의 수를 나타낸다.

2. 희귀한 단어일수록 더 높은 가중치를 부여하기 위해 역문서빈도(Inverse Document Frequency, IDF)를 적용한다. $IDF_j = \log_{10}(\frac{N}{DF_j})$

서비스 모델 미래예측 절차

미래신호 탐색 절차

3 미래신호 탐색

순위	TF		DF		TF-IDF	
	키워드	빈도	키워드	빈도	키워드	빈도
1	일자리	8,212	증세	8,059	일자리	6,328
2	증세	8,059	일자리	7,459	증세	5,940
3	세금	5,339	세금	5,339	복지급여	4,955
4	복지급여	4,520	복지급여	3,524	세금	4,890

단어빈도, 문서빈도, 단어의 중요도 지수를 고려한 키워드 분석

키워드	DoV			평균증가율	평균단어빈도
	1월	2월	3월		
일자리	0.142	0.168	0.186	0.147	2737
증세	0.158	0.237	0.063	-0.116	2686
세금	0.136	0.098	0.077	-0.246	1780
복지급여	0.11	0.067	0.094	0.008	1507

DoV 평균증가율과 평균단어빈도

키워드	DoD			평균증가율	평균문서빈도
	1월	2월	3월		
증세	0.172	0.251	0.07	-0.131	2686
일자리	0.138	0.163	0.186	0.159	2486
세금	0.148	0.104	0.085	-0.238	1780
복지급여	0.092	0.055	0.082	0.047	1175

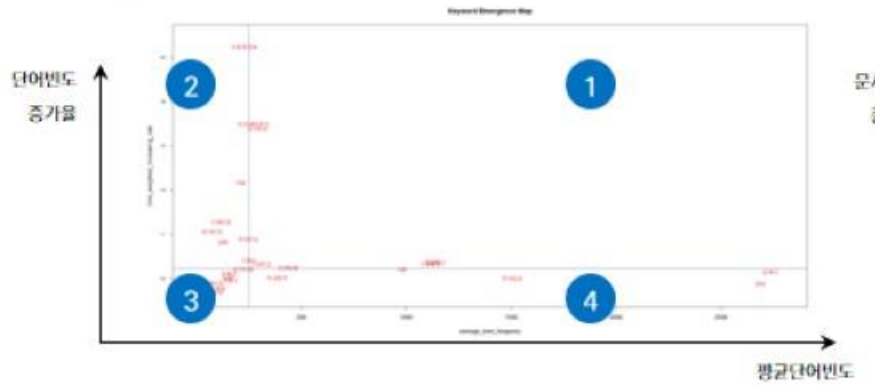
DoD 평균증가율과 평균단어빈도

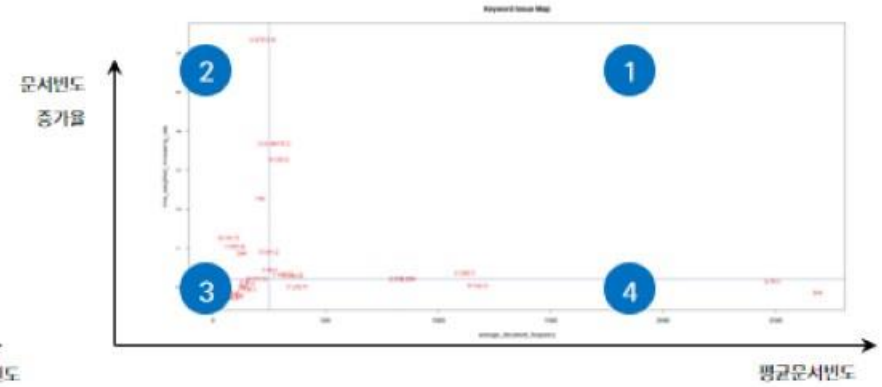
서비스 모델 미래예측 절차

미래신호 탐색 절차

1. 단어빈도(TF), 문서빈도(DF) 추출
2. 중요 정보 추출 (TF-IDF)
3. 미래신호 탐색 (DoV, DoD)
4. 데이터시각화 (KEM, KIM)
5. 미래신호 예측 (Weak Signal)

4 데이터 시각화





서비스 모델 미래예측 절차

미래신호 탐색 절차

5 미래예측 신호

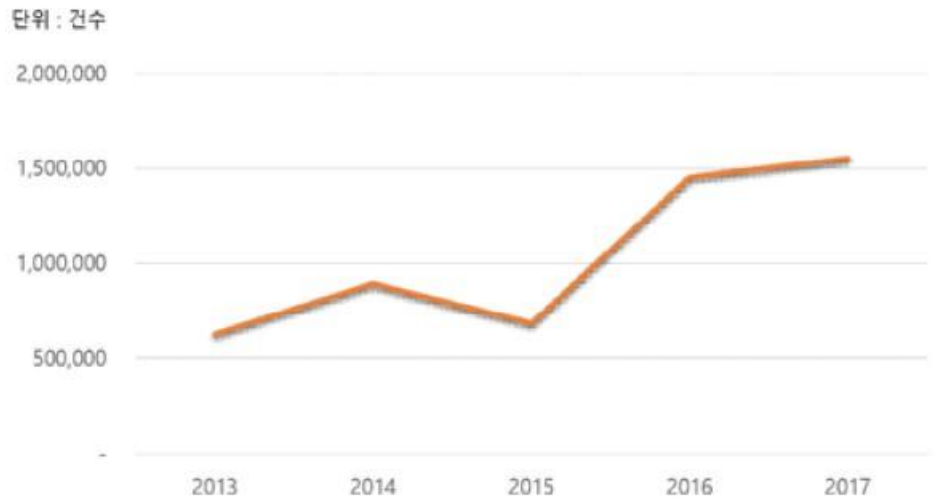
구분	3 잠재신호 (Latent signal)	2 약신호 (Weak Signal)	1 강신호 (Strong signal)	4 강하진 않지만 잘 알려진 신호 (Not strong but well known signal)
KEM	등록금, 보육, 부동산, 원격의료, 양극화, 중증질환	의료민영화, 자살, 가족친화, 환자안전, 담배, 보건산업, 국민연금	미래세대육성, 개인정보, 의료비, 기초연금, 건강보험, 결혼출산, 치료, 건강증진	무상정책, 복지급여, 세금, 일자리, 증세
KIM	등록금, 보육, 부동산, 원격의료, 양극화, 중증질환	의료민영화, 자살, 환자안전, 가족친화, 담배, 보건산업, 국민연금	미래세대육성, 개인정보, 의료비, 기초연금, 건강보험, 치료, 건강증진	무상정책, 결혼출산, 복지급여, 세금, 일자리, 증세
주요 신호	등록금, 보육, 부동산, 원격의료, 양극화, 중증질환	의료민영화, 자살, 환자안전, 가족친화, 담배, 보건산업, 국민연금	미래세대육성, 개인정보, 의료비, 기초연금, 건강보험, 치료, 건강증진	무상정책, 복지급여, 세금, 일자리, 증세

서비스 모델 유용성 검증

해외 감염병 Weak Signal 탐지 사례

2013~2017년 Guardian 외 21개 언론사를 기준으로 5,222,150건의 뉴스 추출

[2013~2017년 연도별 해외뉴스 건수]



키워드 사전을 통해 18,059건의 감염병 뉴스를 추출
(키워드사전은 법정감염병 제1~5군 및 지정감염병을 포함한 79여종과 해외에서 위험하다고 지목된 감염병을 추가해 총 126종의 감염병 키워드로 구성)

↓

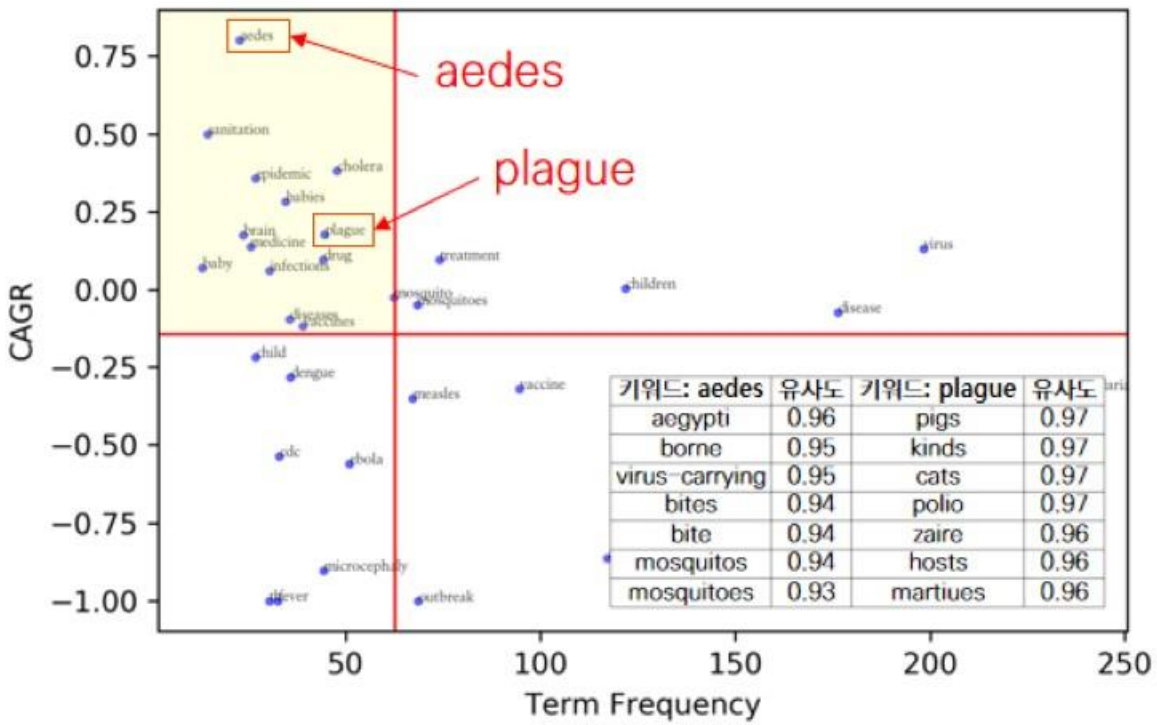
추출된 감염병 뉴스에서 죽음과 관련된 뉴스를 추출하기 위해 kill, kills, killed, death, die, died, dying, dead 등의 키워드를 활용하여 감염병 뉴스 중 죽음과 관련된 10,629건의 뉴스를 추출

↓

연도별 단어빈도(TF)와 문서빈도(DF)를 추출하여 DoV 연평균증가율과, 평균단어빈도, DoD 연평균증가율과 평균문서빈도 도출

서비스 모델 유용성 검증

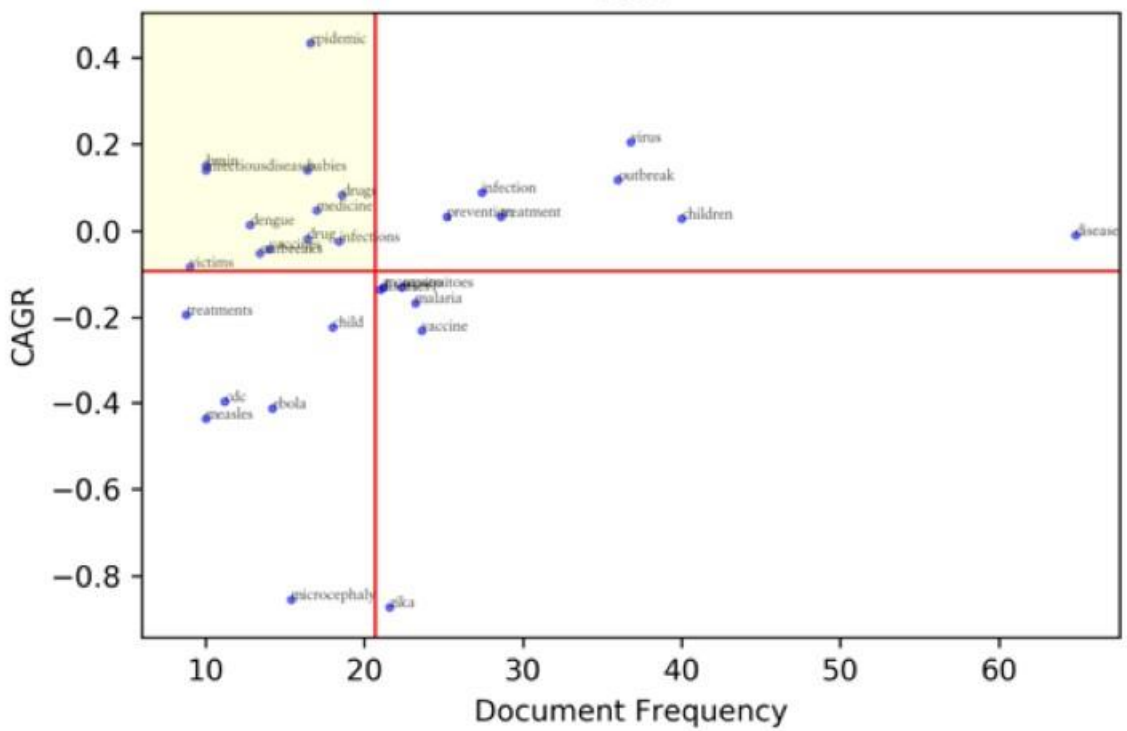
KEM



키워드: aedes	유사도	키워드: plague	유사도
aegypti	0.96	pigs	0.97
borne	0.95	kinds	0.97
virus-carrying	0.95	cats	0.97
bites	0.94	polio	0.97
bite	0.94	zaire	0.96
mosquitos	0.94	hosts	0.96
mosquitoes	0.93	martiues	0.96

aedes라는 키워드는 DoV 연평균증가율이 0.8로 높은 키워드 증가율을 보여주고 있으며, aedes는 숲모기라는 뜻을 갖고 있음. Word2Vec을 통해 추출한 유사도가 높은 키워드는 aegypti, borne, virus-carrying, bites, bite, mosquitos, mosquitoes 순으로 나타났으며, 지카바이러스를 옮기는 매개체인 것으로 나타나 강신호로 발전하기 이전에 바이러스를 옮기는 매개체에 대한 대비가 필요

KIM



plague라는 키워드는 사전적으로 페스트, 전염병 등의 다중의미를 포함하고 있으며, 따라서 Word2Vec을 통해 유사단어를 추출하자 pigs라는 키워드가 가장 큰 유사도를 보이고 있는 것으로 나타나 돼지전염병에 대한 대비가 필요

서비스 모델 유용성 검증

지카바이러스는 아직까지 백신이 개발되지 않아 여전히 위험성이 높은 감염병이며, weak signal 감지 후 6개월 이내 유행되는 패턴을 보여 실시간성이 반영될 수 있는 새로운 미래신호 탐색 모델을 구현해야 할 필요성이 있음

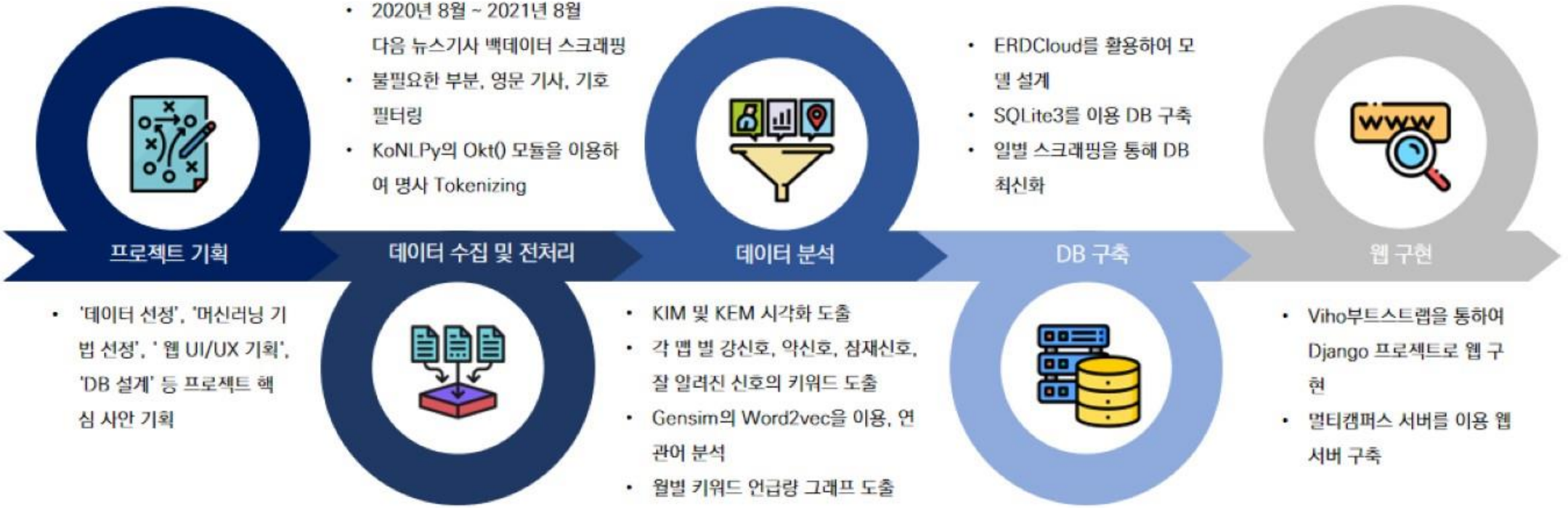
아프리카 돼지열병(ASF) 바이러스의 생존력은 매우 높아 냉장육 및 냉동육에서도 수개월~수년간 생존 가능하고, 가염건조나, 훈제, 공기건조된 육류에서도 생존이 가능한 것으로 알려져 오래된 육류는 매우 위험한 감염원으로 지적되고 있어 지속적인 모니터링이 필요

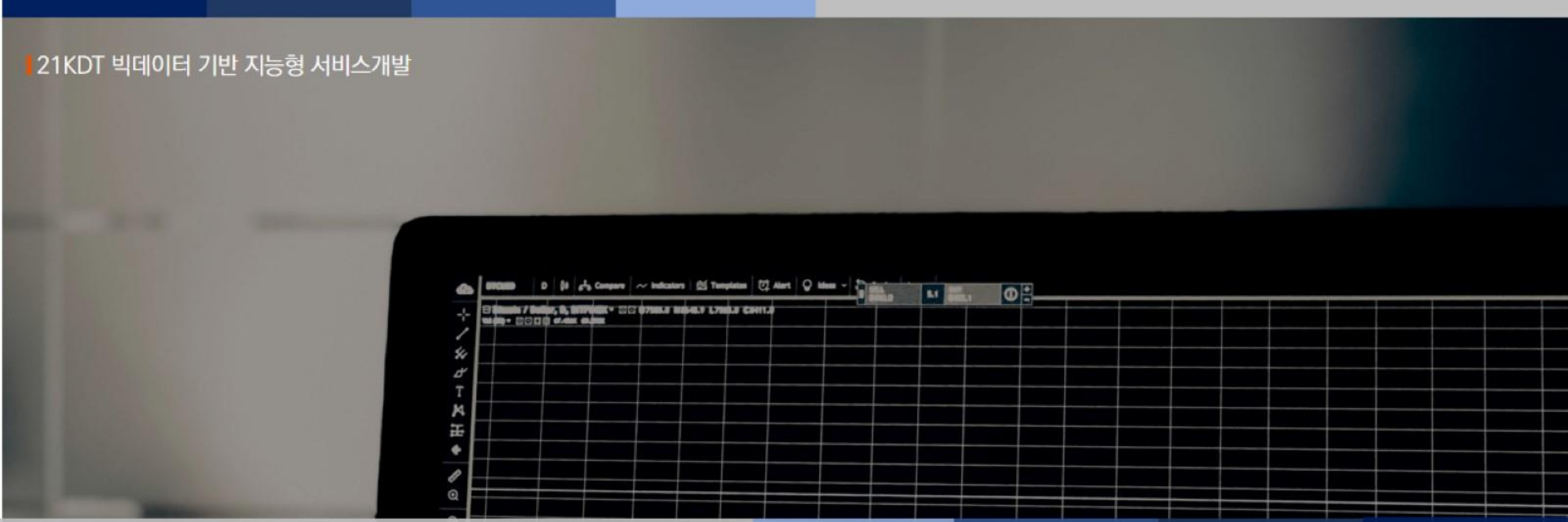
서비스 모델

※ 무료회원: 현재 데이터 열람 가능

※ 유료회원: 과거 데이터 열람(weak signal의 이동경로 파악), 데이터 다운로드 가능

프로젝트 절차


프로젝트 기획
• '데이터 선정', '머신러닝 기법 선정', ' 웹 UI/UX 기획', 'DB 설계' 등 프로젝트 핵심 사안 기획
데이터 수집 및 전처리
• 2020년 8월 ~ 2021년 8월 다음 뉴스기사 백데이터 스크래핑
• 불필요한 부분, 영문 기사, 기호 필터링
• KoNLPy의 Okt() 모듈을 이용하여 명사 Tokenizing
데이터 분석
• KIM 및 KEM 시각화 도출
• 각 맵 별 강신호, 약신호, 잠재신호, 잘 알려진 신호의 키워드 도출
• Gensim의 Word2vec을 이용, 연관어 분석
• 월별 키워드 언급량 그래프 도출
DB 구축
• ERDCloud를 활용하여 모델 설계
• SQLite3를 이용 DB 구축
• 일별 스크래핑을 통해 DB 최신화
웹 구현
• Viho부트스트랩을 통하여 Django 프로젝트로 웹 구현
• 멀티캠퍼스 서버를 이용 웹 서버 구축



프로젝트 팀 구성 및 역할

multicampus

팀 명 및 프로젝트 조직

HEX in AR

백엔드 개발

권회동님

산업경영공학 전공

github.com/tecktonik08
superinssa@gmail.com

백엔드 개발

김수원님

경영학 전공

github.com/webdessin
webdessin@gmail.com

PM

김민성님

데이터사이언스 전공

github.com/msio900
msio900@gmail.com

프론트 엔드 개발

김하영님

호텔조리제과 전공

github.com/young-ha713
hayeongg386@gmail.com

프론트 엔드 개발

남승주님

전자공학 전공

github.com/SJ-16
qwer567103@gmail.com

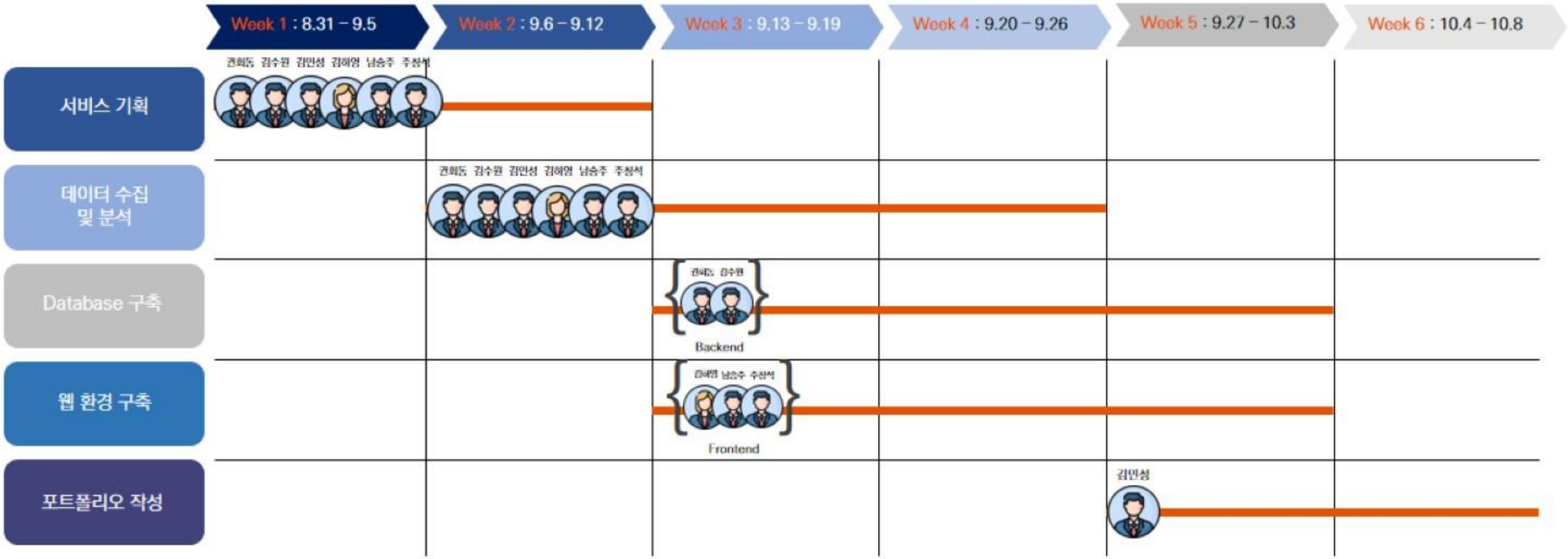
프론트 엔드 개발

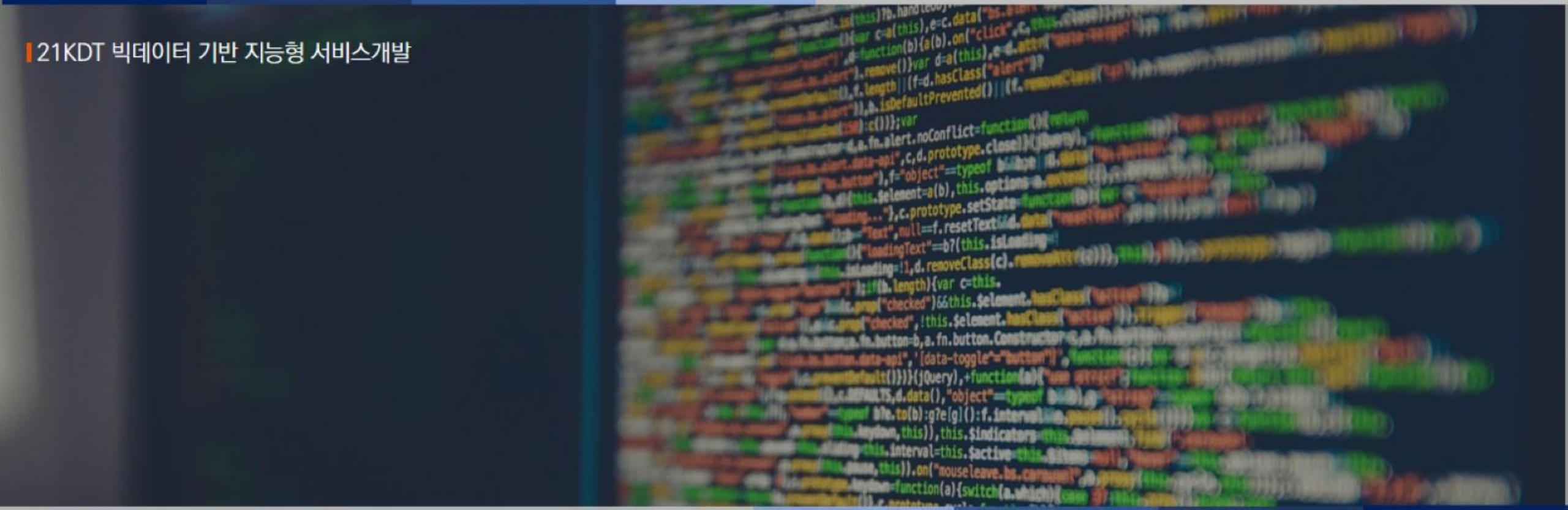
주창석님

체코어 전공

github.com/victoria2012
radomir843@gmail.com

프로젝트 추진 일정

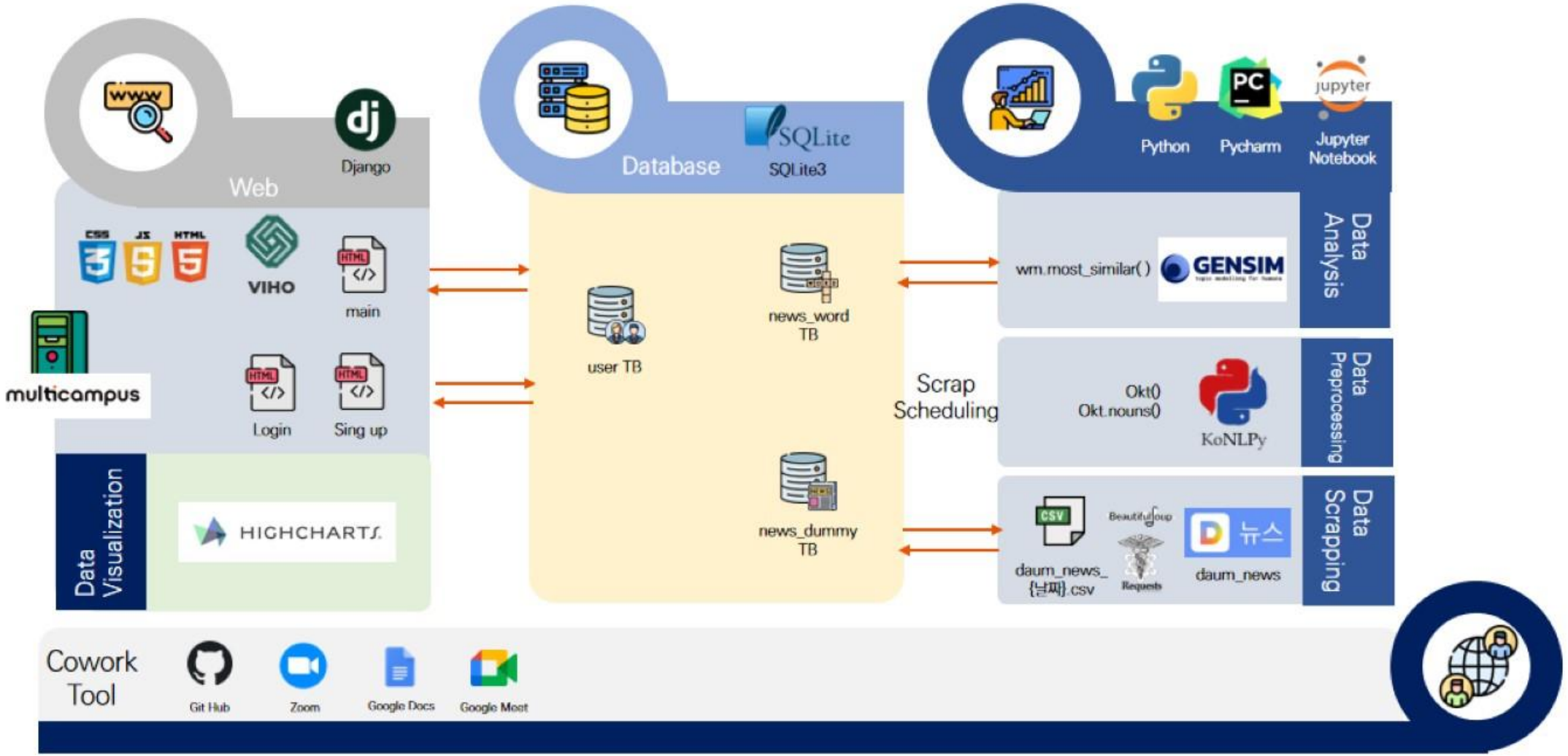

Week 1 : 8.31 – 9.5
Week 2 : 9.6 – 9.12
Week 3 : 9.13 – 9.19
Week 4 : 9.20 – 9.26
Week 5 : 9.27 – 10.3
Week 6 : 10.4 – 10.8
서비스 기획
데이터 수집 및 분석
Database 구축
웹 환경 구축
포트폴리오 작성
Backend
Frontend
김민성



프로젝트 수행 절차 및 방법

multicampus

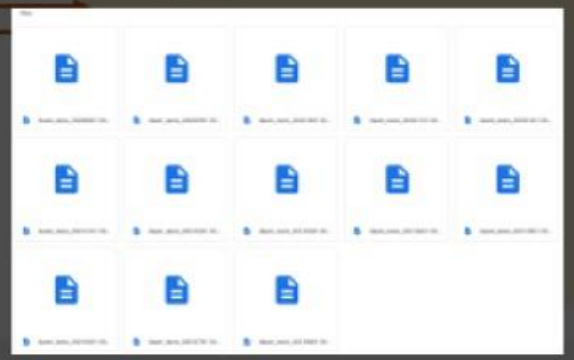
서비스 모델


Web
Django
VIHO
main
Login
Sing up
multicampus
Data Visualization
HIGHCHARTS
Database
SQLite
SQLite3
user TB
news_word TB
news_dummy TB
Scrap Scheduling
Python
Pycharm
Jupyter Notebook
wm.most_similar( )
GENSIM
Data Analysis
Okt()
Okt.nouns()
KoNLPy
Data Preprocessing
daum_news_{날짜}.csv
Requests
daum_news
Data Scrapping
Cowork Tool
Git Hub
Zoom
Google Docs
Google Meet

데이터 수집

- 다음 뉴스 스크래핑
  - 2020년 8월 부터 2021년 8월 까지 5,253,255건의 기사 스크래핑
  - BeautifulSoup과 Requests 모듈을 이용 하여 스크래핑한 기사를 월별로 CSV형태로 저장

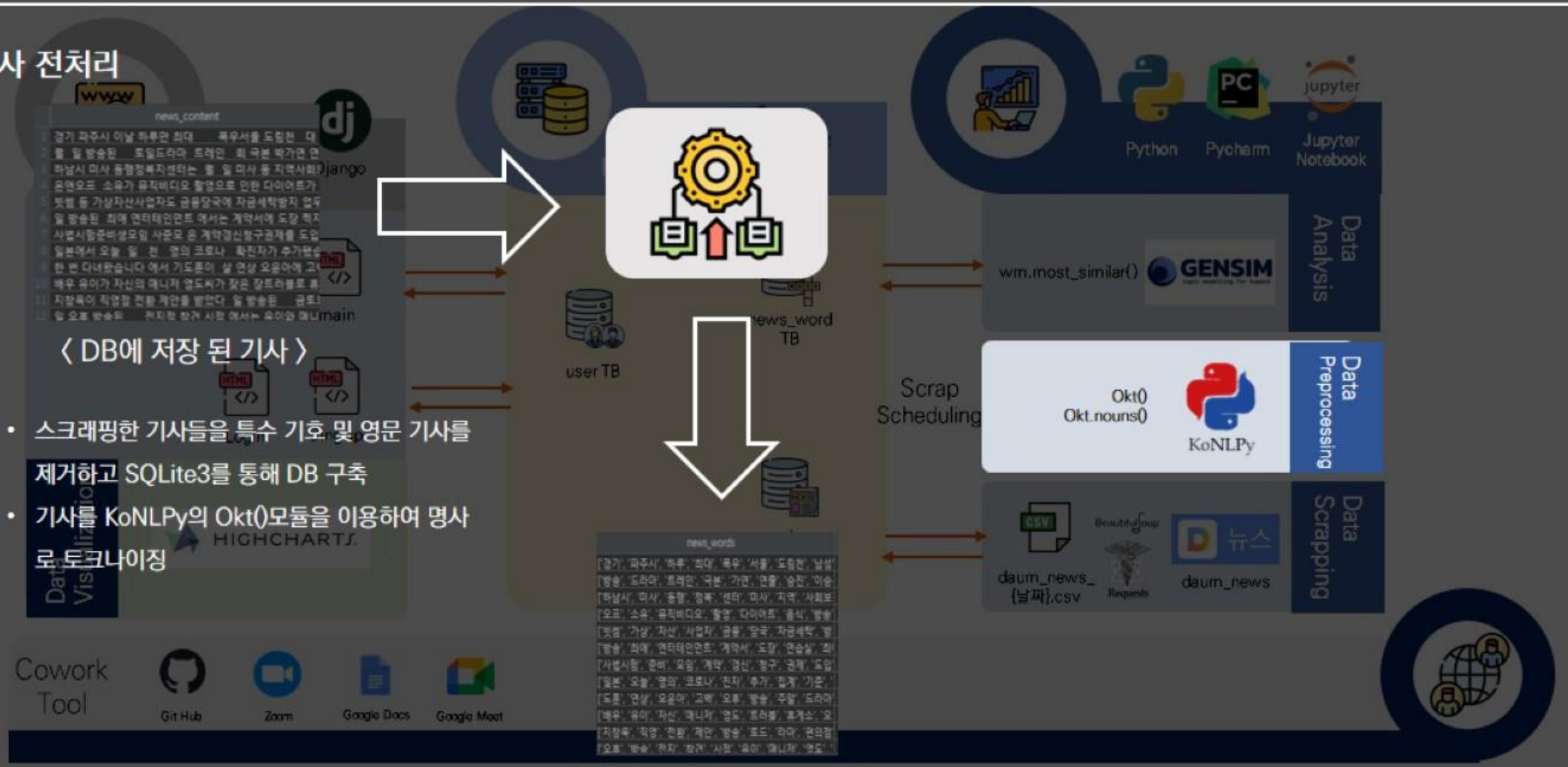
〈 다음 전체 기사 페이지 〉



〈 스크래핑 후 CSV로 저장 〉

데이터 전처리

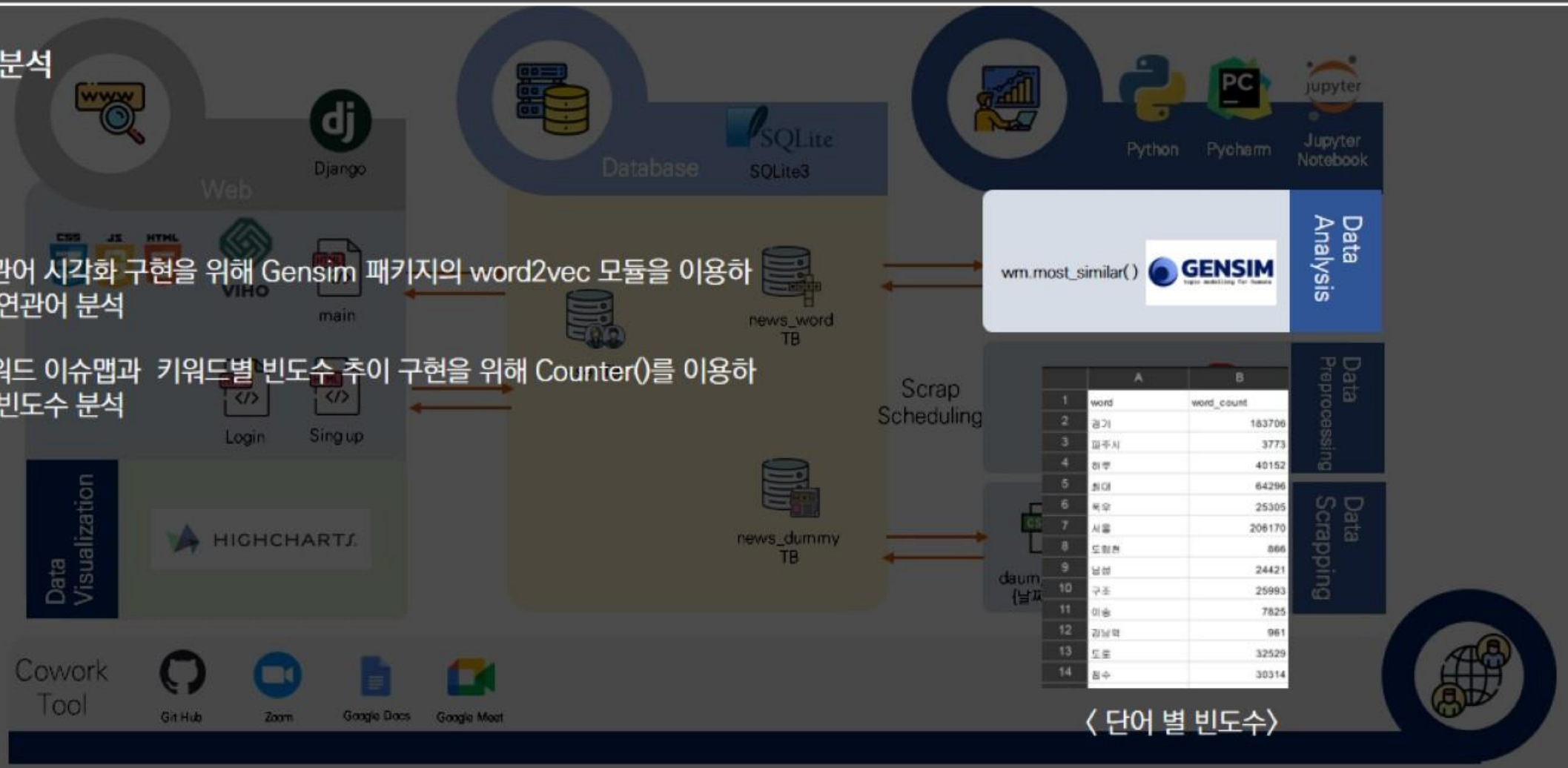
- 뉴스 기사 전처리



- 스크래핑한 기사들을 특수 기호 및 영문 기사를 제거하고 SQLite3를 통해 DB 구축
- 기사를 KoNLPy의 Okt()모듈을 이용하여 명사로 토크나이징

데이터 분석

- 데이터 분석

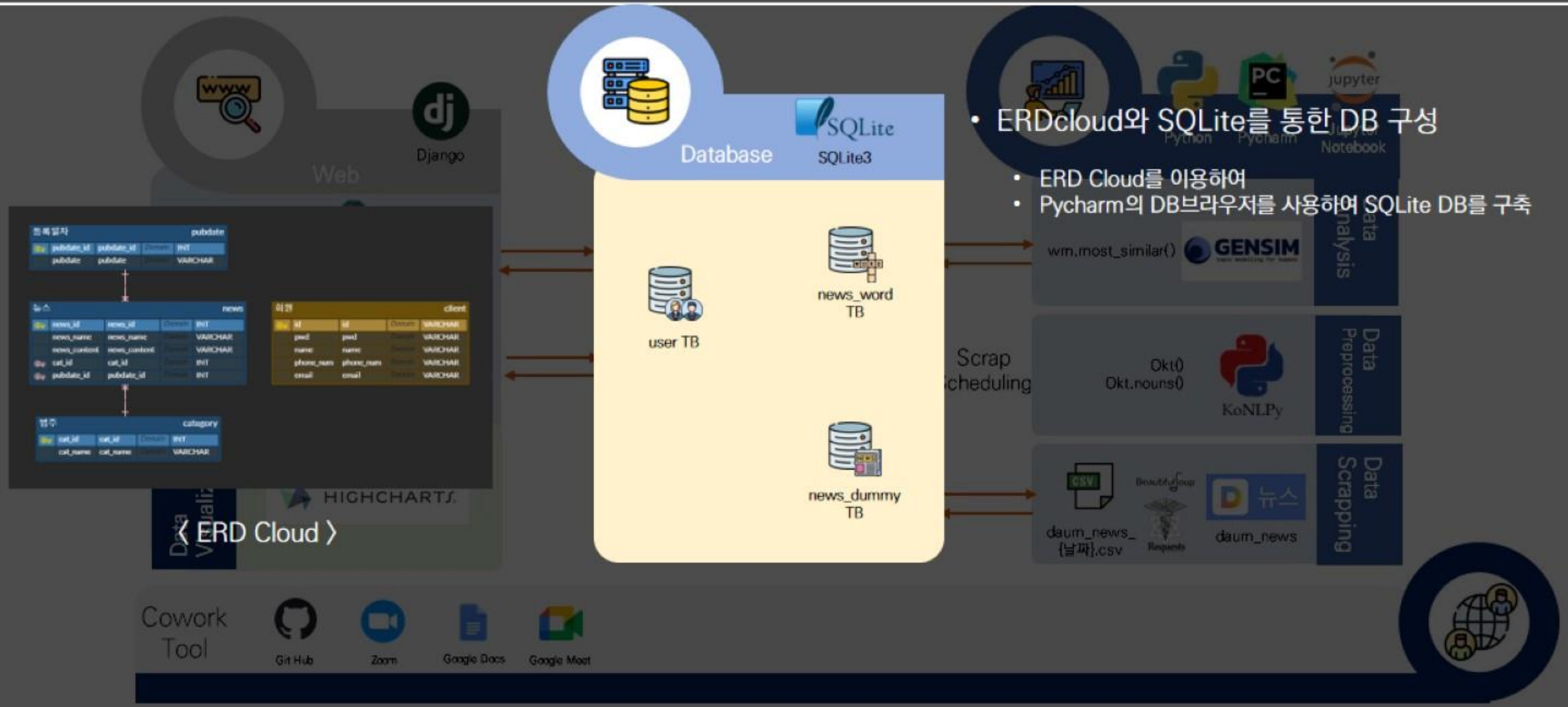


- 연관어 시각화 구현을 위해 Gensim 패키지의 word2vec 모듈을 이용하여 연관어 분석
- 키워드 이슈맵과 키워드별 빈도수 추이 구현을 위해 Counter()를 이용하여 빈도수 분석

	A	B
1	word	word_count
2	경기	183706
3	파주시	3773
4	한쿵	40152
5	최대	64296
6	목우	25305
7	서울	208170
8	도원천	866
9	남성	24421
10	구조	25993
11	이송	7825
12	김남택	961
13	도로	32529
14	화수	30314

〈 단어 별 빈도수〉

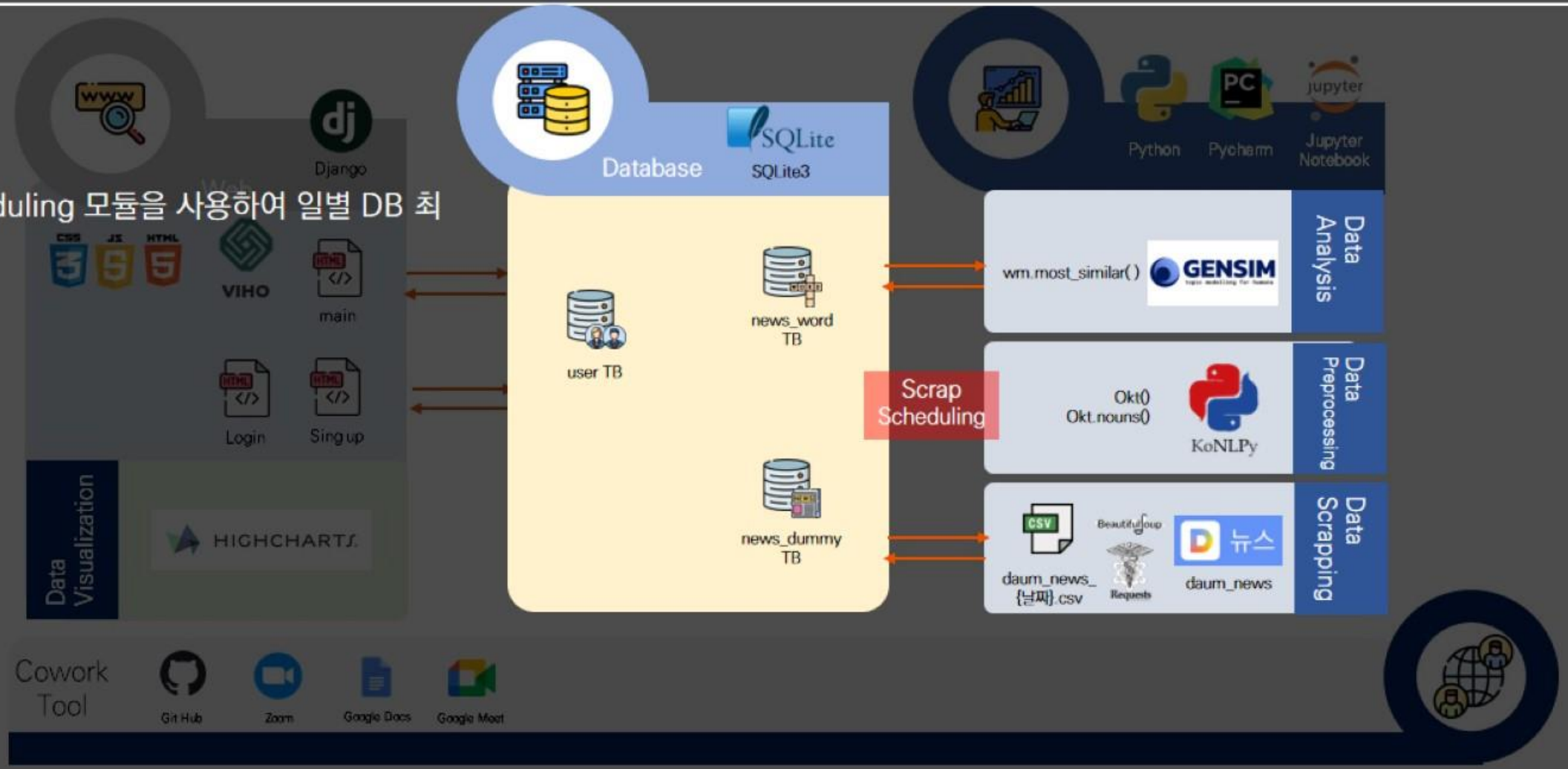
DB 구성 모델



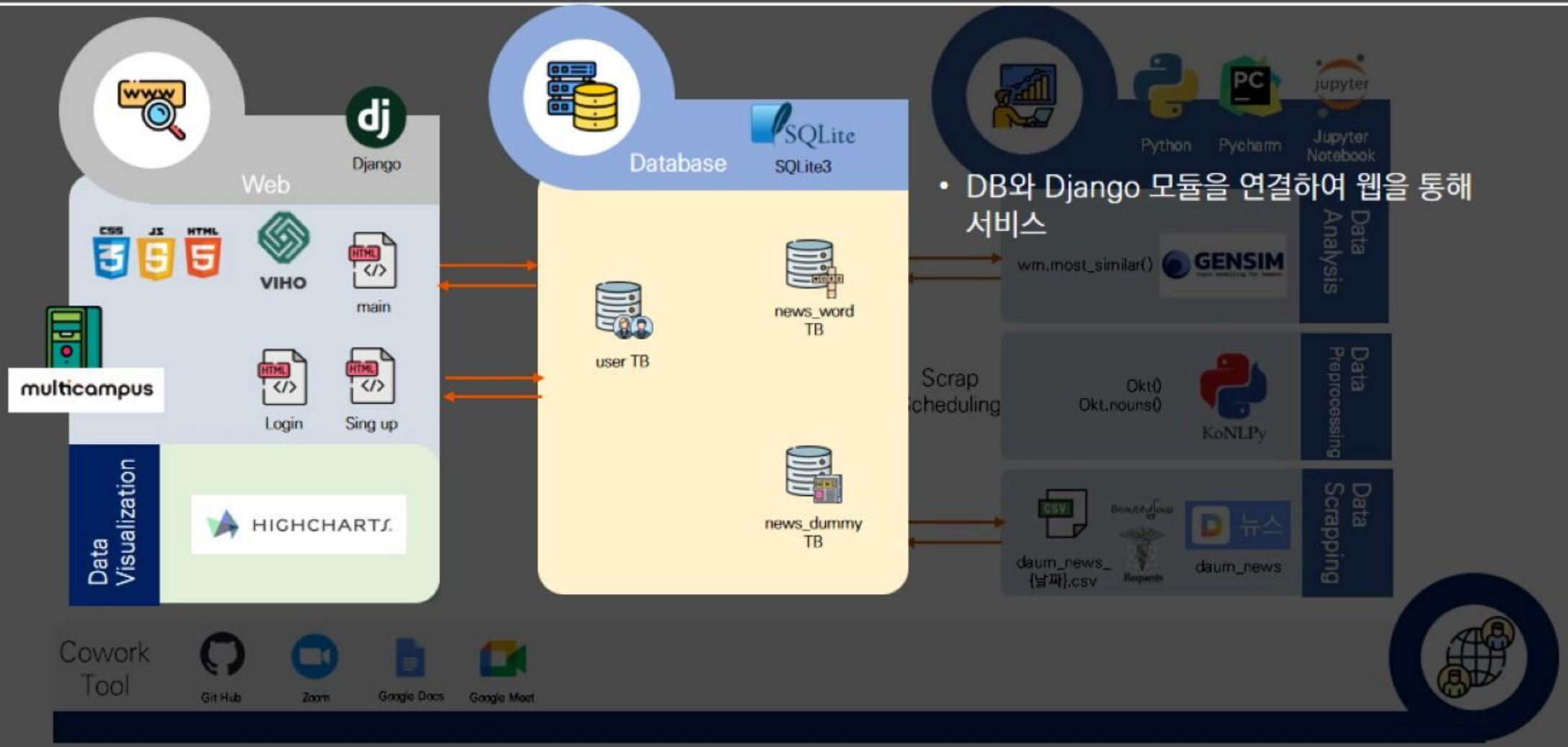
- ERDcloud와 SQLite를 통한 DB 구성
  - ERD Cloud를 이용하여
  - Pycharm의 DB브라우저를 사용하여 SQLite DB를 구축

DB 구성 모델

- Scheduling 모듈을 사용하여 일별 DB 최신화

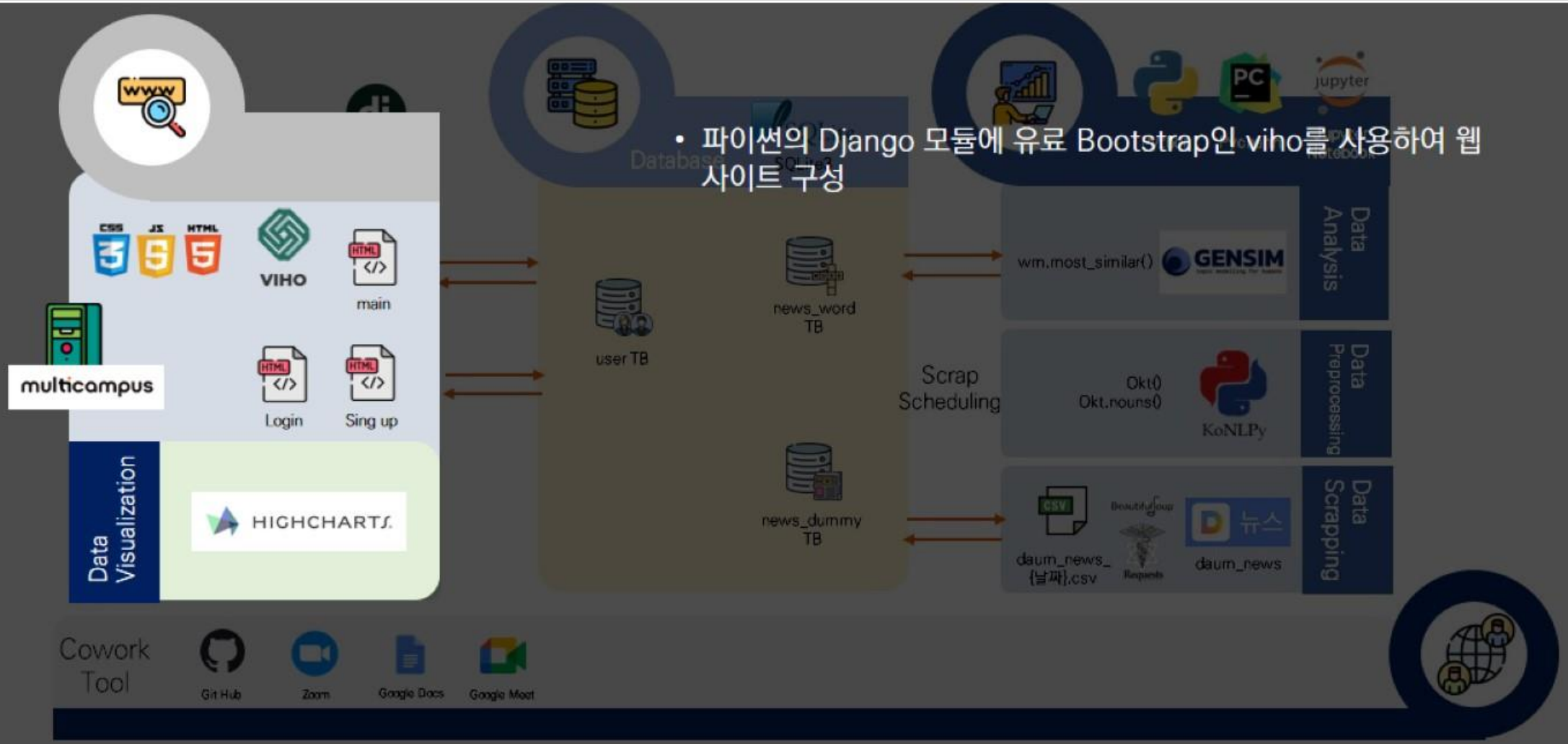


웹 구현 모델



- DB와 Django 모듈을 연결하여 웹을 통해 서비스

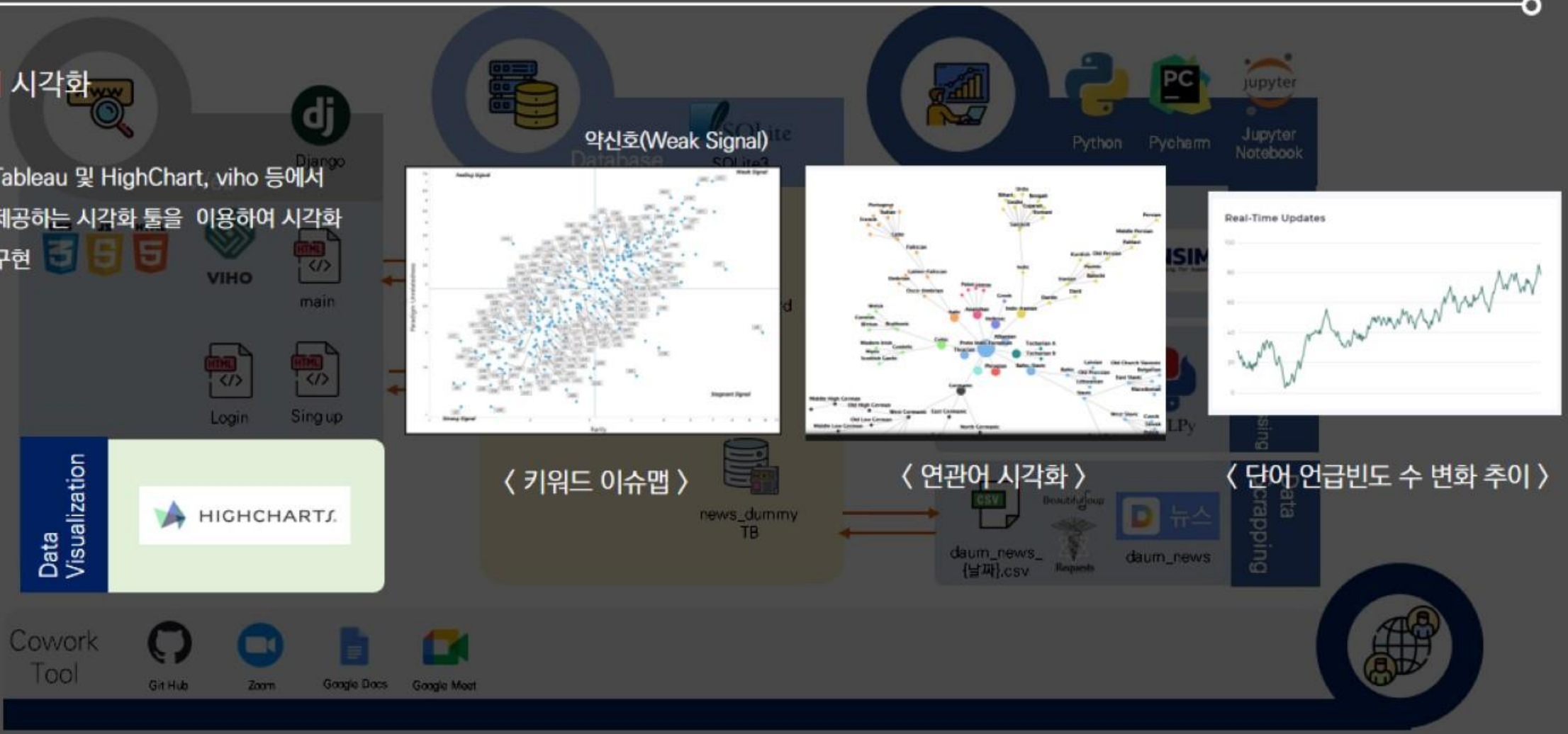
웹 구현 모델



- 파이썬의 Django 모듈에 유료 Bootstrap인 viho를 사용하여 웹 사이트 구성

웹 구현 모델

- 데이터 시각화
  - Tableau 및 HighChart, viho 등에서 제공하는 시각화 툴을 이용하여 시각화 구현



협업 방식 소개

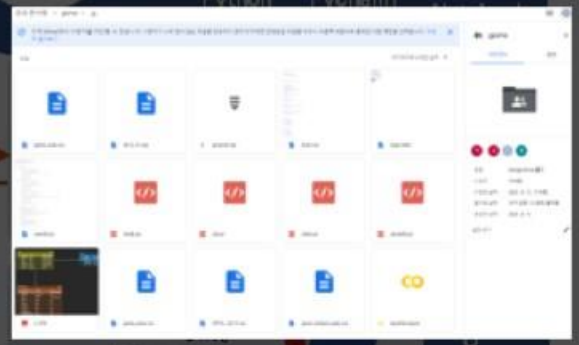
〈 Git Hub 〉

기본적인 분산 버전 관리 협업 툴로, django 코드와 Bootstrap, database 등 많은 코드들을 호스팅하여 관리함

〈 Zoom 〉

원격으로 화상회의를 하여 의사소통에 있어 중요한 표정, 몸짓 등 직접 만나지 않아도 원활한 소통이 가능하게 해주는 협업도구

Scrap
Scheduling

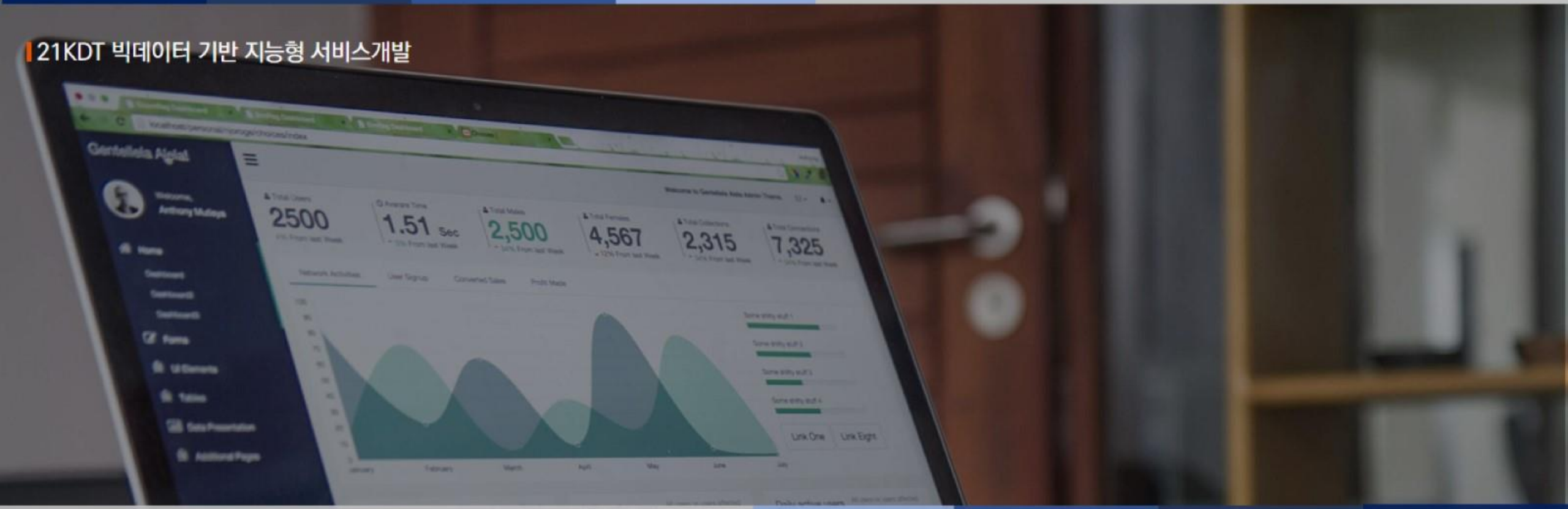


〈 Google Docs 〉

Git Hub와 연동이 가능하며, 비교적 큰 용량을 업로드하여 협업이 가능하기에 프로젝트 전체를 공유하여 협업이 가능함

Cowork Tool

Git Hub　Zoom　Google Docs　Google Meet



프로젝트 수행결과 및 기대효과

multicampus

수행결과 소개

〈 데이터 수집 및 전처리 과정 시연 〉

〈 웹 서비스 시연 〉

프로젝트의 개선할 사항

1 뉴스기사의 정제 기술 추가

- 해외 언론사들은 온라인에서만 운영되는 인터넷 기반의 디지털 뉴스미디어가 거의 많지 않은 반면, 국내 인터넷 언론사업체 수는 2,800여 개, 인터넷 신문의 매체 수는 4,300여 개에 이르고 있음
- 대부분 직접취재가 아닌 보도자료성 뉴스가 대부분이어서 중복되는 토픽의 뉴스가 과다한 측면이 있어 이를 정체할 수 있는 추가적인 연구가 필요

2 섹터 분류 및 관련 키워드 사전 구축

- 정치 / 사회 / 보건 / 복지 / 안전 / 재난 / 경제 / 금융 / 부동산 / 비즈니스 / 인물 / 문화 등 다양한 섹터로 구분하여 관련된 키워드 사전을 구축하여 weak signal의 유용성이 높은 섹터 발굴
- 예측 정확성을 높이기 위해 불용어 사전, 술어 사전 등 지속적인 한국어에 특화된 사전구축이 필요

개발 후기 및 느낀 점

multicampus

개발후기

남승주 3초 전

스크래핑할 때 시간이 너무 오래 걸려 프로그래밍용 컴퓨터도 성능이 좋아야 한다는 것을 다시금 느끼게 되었으며 용어를 통일한다거나 주석문을 다는 습관을 들이는 등 협업시에 어떤 행동이 도움이 되는지에 대해 어느정도 알게 된 것 같습니다. 프로젝트 과정이 때론 힘들기도 했지만 좋은 조원들을 만나서 웃으면서 할 수 있었고 함께 얘기를 나누며 몰랐던 사실들을 많이 배우고 가는것 같아 감사했습니다.

파이널 프로젝트가 교육과정동안 배운 것들을 모두 써서 최종 결과물을 만들어 내는거라 생각했기에 잘하고 싶은 마음이 컸는데 아직은 역량이 조금 부족하다고 느꼈고, 앞으로 더 노력하여 역량을 기르는 것이 남은 과제인 것 같습니다.

2.2천 답글

주창석 5초 전

1, 2차 세미 프로젝트에서 데이터수집(스크랩핑), 딥러닝을 맡아 작업을 했기 때문에 이번 프로젝트에서는프론트엔드를 맡았지만 스크랩핑부터 여러 과정에 걸쳐 조금씩 참여하면서 다양한 경험을 했습니다. 최종 프로젝트는 세미프로젝트 때보다 기간이 길어서 팀원들과 공동 작업하는 시간이 많아 팀웍을 다지고 서로 가르쳐주고 배울 수 있어서 좋았습니다. 구현해보고 싶은 것은 많았지만 시간상의 부족과 부족한 실력으로 현실에 타협할 수 밖에 없는 부분이 아쉬워서수료 후에도 이번 프로젝트를 더 발전시켜보자는 논의를 하고 있습니다.

2.2천 답글

권회동 5초 전

느낀점 : 스크래핑과 전처리에 대해 더 자세하게 배울 수 있는 기회였지 않았나 싶습니다. 다양한 모듈을 활용해 볼 수 있었고 여러가지 시행착오를 통해 코딩을 잘 이해하고 사용할 수 있게 되었습니다.

불편한 점 : 지정과제 스크래핑을 할 때, 페이지의 양식이 모두 동일하지 않아 스크래핑할 태그가 다르거나 하여 까다로운 점이 있었고 전처리의 경우, 같은 방식이라도 데이터 타입에 따라 코드를 바꿔줘야 하는 어려움이 있었습니다.

개선 방안 : 코드 수정을 적게 할 수 있도록 초기에 계획을 확실하게 세우거나 변경 가능성에 대해 충분히 생각해보면 개발 시간을 줄일 수 있다고 생각합니다.

2.5천 답글

개발후기

김하영 6초 전

이번 개발을 하면서 파이널 프로젝트인 만큼 난이도가 높았다는 생각이 들었고 또, 담원분들이 협동해서 문제를 해결해 나가는 모습을 보고 개인의 역량 또한 중요하지만 의사소통기술 역시 중요한 부분 이라는것을 깨달았습니다. 개발을 하면서 다른 분들은 이 분야에 대한 어느 정도의 지식을 갖추어 비교적 짧은 시간에 해결하는 반면에 저는 기본적인 지식이 부족해 같은 문제여도 시간이 오래 걸리는 부분이 프로젝트 진행상 불편했습니다.그렇기 때문에 기본기를 다지는 공부를 통해 개선해 나갈 생각입니다.

1.2천 답글

김수원 16초 전

다양한 의견을 조율하여 문제 해결을 한 부분이 큰 경험이 되었습니다.

2.2천 답글

김민성 4초 전

이번 파이널 프로젝트는 프로젝트 기획에서 부터 머신러닝 및 딥러닝 스터디, 데이터 수집 및 전처리 웹으로 구현하고 데이터를 DB를 통해 관리 하는 법까지 여러가지 어려운 점이 많았습니다. 특히 불편했던 점은 조금만 이 쪽으로 지식이 조금만 더 있었더라면 헤매는 시간을 줄일 수 있었을텐데…하고 생각을 갖곤했습니다. 하지만 함께 한 담원 분들 덕에 이러한 어려움들을 극복할 수 있었고 개선해 나갈 수 있었습니다. 이번 프로젝트를 통해 나온 산출물은 아직 부족한 점이 너무 많습니다. 하지만 꾸준한 관리와 디버깅을 통해 잔정한 미래 예측 시스템을 구현할 수 있도록 노력하겠습니다.

2.2천 답글

감사합니다.

Q&A