[별지14호 서식]

최 종 보 고 서

<table>
<tr><td>관리 번호</td><td colspan="2">2023(연도)-20(번호)</td><td>기술 분류</td><td></td></tr>
<tr><td>과 제 명</td><td colspan="4">(한글) 과기정통부 소관 데이터 활용 활성화를 위한 분석과제 추진방안 컨설팅 연구
(영문) Research on Data Analysis Plan Consulting for Invigorating Data use in Ministry of Science and ICT</td></tr>
<tr><td rowspan="2">주관연구기관
(협동연구기관)</td><td>기 관 명</td><td colspan="2">소재지</td><td>대 표</td></tr>
<tr><td>㈜올림커뮤니케이션즈</td><td colspan="2">충청북도 청주시 흥덕구 사운로 314 신봉동 2층</td><td>신동혁</td></tr>
<tr><td rowspan="2">주관연구책임자
(협동연구책임자)</td><td>성 명</td><td colspan="2">소속 및 부서</td><td>전 공</td></tr>
<tr><td>신동혁</td><td colspan="2">㈜올림커뮤니케이션즈</td><td>빅데이터</td></tr>
<tr><td>총연구기간
(당해년도)</td><td colspan="4">2023년 09월 11일 ~ 2024년 03월 08일(06개월)</td></tr>
<tr><td>총연구비
(당해년도)</td><td colspan="4">일금 사천구백오십만원정 (₩ 49,500,000원)</td></tr>
<tr><td>총참여연구원
(당해년도)</td><td colspan="4">3명(총괄책임: 1명, 책임: 2명)</td></tr>
</table>

2023년도 정책연구용역사업으로 수행한 연구과제의 최종보고서를 붙임과 같이 제출합니다.

붙임 : 최종보고서 1부.

2024년 03월 08일

주관연구책임자 신동혁 (인)

주관연구기관장 신동혁 직인

과학기술정보통신부장관 귀 하

"과기정통부 소관 데이터 활용 활성화를 위한 분석과제 추진방안 컨설팅에관한 연구" 에 관한 정책연구사업의 최종보고서를 별첨과 같이 제출합니다.

2024년 03월 08일

주관연구책임자 신 동 혁 (인)

주관연구기관장 신 동 혁 직인

정책연구 (2023-20) 연 구 과 제 명 과 학 기 술 정 보 통 신 부	정책연구 (2023-20) 연 구 과 제 명 과기정통부 소관 데이터 활용 활성화를 위한 분석과제 추진방안 컨설팅 연구 (영문) Research on Data Analysis Plan Consulting for Invigorating Data use in Ministry of Science and ICT 과학기술정보통신부

제 출 문

과학기술정보통신부장관 귀 하

본 보고서를 "과기정통부 소관 데이터 활용 활성화를 위한 분석과제 추진방안 컨설팅에 관한 연구" 최종보고서로 제출합니다.

2024년 03월 07일

○ 주관연구기관명 : ㈜올림커뮤니케이션즈

○ 연 구 기 간 : 2023.09.11 ~ 2024.03.08

○ 주관연구책임자 : 신 동 혁

○ 참여연구원

- 연 구 원 : 황 재 선
- 연 구 원 : 오 상 진

※ 주관연구기관 및 주관연구책임자, 연구원은 실제 연구에 참여한 기관 및 참여자의 명의로 함

「과기정통부 소관 데이터 활용 활성화를 위한 분석과제 추진방안 컨설팅 연구 최종 보고서」

㈜올림커뮤니케이션즈

목 차

요 약 ····· 1

제1장. 과제 개요 ····· 2

1. 추진배경(추진필요성) ····· 2

2. 추진목표 ····· 3

3. 추진전략 및 방법 ····· 3

4. 추진절차 ····· 4

4-1. 데이터 분석 계획 검토 및 개선방안 마련 ····· 4

4-2. 현장 맞춤형 컨설팅 운영 ····· 5

4-3. 기관별 컨설팅 결과 최종안 마련 ····· 6

5. 다른 기관과의 협업 ····· 7

6. 활용 데이터 ····· 7

제2장. 과제 사전조사 검토 결과 ····· 9

1. 국립전파연구원 ····· 9

2. 우주전파센터 ····· 9

3. 한국원자력연구원 ····· 10

4. 한국생산기술연구원 ····· 10

5. 기초과학연구원 ····· 11

6. 한국항공우주연구원 ····· 11

7. 나로우주센터 ····· 12

8. 한국철도기술연구원 ····· 13

9. 한국과학기술연구원 ····· 13

10. 한국연구재단 ····· 14

11. 국가과학기술인력개발원 ····· 15

제3장. 컨설팅 분석 내용 및 결과 ····· 16

1. 국립전파연구원 ····· 16

2. 우주전파센터 ····· 18

3. 한국원자력연구원 ····· 21

4. 한국생산기술연구원 ····· 23

5. 기초과학연구원 ····· 26

6. 한국항공우주연구원 ····· 28

7. 나로우주센터 ····· 30

8. 한국철도기술연구원 ····· 33

9. 한국과학기술연구원 ····· 37

10. 한국연구재단 ····· 41
11. 국가과학기술인력개발원 ····· 45

제4장 정책(업무) 활용실적 및 사후관리 방안 ····· 49
1. 기관별 기대효과 ····· 49
2. 시범 적용 기관 선정 및 모형도출 ····· 53
2-1. 한국철도기술연구원 ····· 53
2-2. 기초과학연구원 ····· 60
3. 분석결과 사후관리방안 ····· 66

○ 표 목차

<표1> 기본계획 추진과제 ····· 2
<표2> 활용 데이터 ····· 8
<표3> 전자제품 적합성 현황분석 ····· 16
<표4> 우주전파환경 경보상황 사후분석 ····· 18
<표5> 방사성폐기물 정보관리시스템 ····· 21
<표6> 데이터 활용 고령자 건강상태 예측 ····· 23
<표7> 연구 전략 수립 의사결정 지원 ····· 26
<표8> AI 스마트 활용기술 개발 ····· 28
<표9> 콘텐츠 다양화 및 연계프로그램 제공 ····· 30
<표10> 철도 분야 무역데이터 분석 ····· 33
<표11> 자산관리 장비 예약 활용관리 고도화 ····· 37
<표12> 국내 학술 활동 인용 현황 대한 분석 ····· 41
<표13> 한국철도기술연구원 선정배경 ····· 45
<표14> 실물모형 목업 프로세스 ····· 56
<표15> 활용 데이터 ····· 56
<표16> 분석 알고리즘 ····· 57
<표17> 모델1-한국철도연구원 파이썬 코드 ····· 57
<표18> 모델1-결과값 시각화 ····· 58
<표19> 모델2-한국철도연구원 파이썬 코드 ····· 59
<표20> 모델2-결과값 시각화 ····· 60
<표21> 기초과학연구원 선정배경 ····· 60
<표22> 실물모형 목업 프로세스 ····· 62
<표23> 활용 데이터 ····· 62
<표24> 분석 알고리즘 ····· 63
<표25> 모델1-기초과학연구원 파이썬 코드 ····· 63
<표26> 모델1-결과값 시각화 ····· 64
<표27> 모델2-기초과학연구원 파이썬 코드 ····· 65
<표28> 모델2-결과값 시각화 ····· 66

○ 그림 목차

<그림1> 분석과제 개선안 도출 ····· 3
<그림2> 개발 코드 ····· 17
<그림3> 개발 코드와 분석결과 ····· 20
<그림4> 개발 코드와 클러스터링 분석 해석 ····· 22
<그림5> 개발 코드와 로지스틱 회귀분석 ····· 25
<그림6> 개발 코드와 텍스트 마이닝 ····· 27
<그림7> 파이썬 개발 코드 ····· 29
<그림8> 파이썬 개발 코드와 회귀 분석 시계열 분석 ····· 30
<그림9> 향후 연구 방향 ····· 32
<그림10> 파이썬 엑셀 텍스트마이닝 클러스터링 ····· 35
<그림11> 엑셀 관련 분석결과 ····· 36
<그림12> 파이썬 분석 과정 ····· 39
<그림13> 통계분석결과 ····· 39
<그림14> 향후 연구 방향 시각화 ····· 40
<그림15> 파이썬 분석 과정 ····· 42
<그림16> 텍스트 마이닝 결과 ····· 43
<그림17> 텍스트 마이닝 엑셀 결과 ····· 43
<그림18> 클러스터링 결과 ····· 43
<그림19> 향후 연구 방향 시각화 ····· 44
<그림20> 향후 연구 방향 고도화 시각화 ····· 44
<그림21> 키워드 파이선 분석 과정 ····· 47
<그림22> 텍스트마이닝 결과 ····· 47
<그림23> 향후 연구 방향 예시 결과 ····· 48
<그림24> 결과 ui 화면 예시 ····· 48
<그림25> 국립전파연구원 기대효과 ····· 49
<그림26> 우주전파센터 구현 예시 화면 ····· 49
<그림27> 한국원자력연구원 구현 예시 화면 ····· 50
<그림28> 한국생산기술연구원 구현 예시 화면 ····· 50
<그림29> 기초과학연구원 구현 예시 화면 ····· 50
<그림30> 한국항공우주연구원 구현 예시 화면 ····· 51
<그림31> 한국철도기술연구원 구현 예시 화면 ····· 51
<그림32> 한국과학기술연구원 구현 예시 화면 ····· 52
<그림33> 한국연구재단 구현 예시 화면 ····· 52
<그림34> 국가과학기술인력 구현 예시 화면 ····· 53
<그림35> 연도에 따른 수출 그래프 ····· 53
<그림36> UN Comtrade Database ····· 55

<그림37> 주요 수출 국가 7개국 ····· 55
<그림38> 클러스터링 결과 ····· 55
<그림39> 판다스 ····· 56
<그림40> K-means 클러스터링 ····· 56
<그림41> 연도에 따른 주요 키워드 ····· 60
<그림42> 공공데이터 포털 ····· 62
<그림43> 상위 10개 키워드 ····· 62
<그림44> 연계성 시각화 ····· 62
<그림45> NLP ····· 63
<그림46> 코사인 유사도 ····· 63

빈 칸

요 약

○ 한국 정부는 '데이터 기반 행정 활성화'를 위해 법률과 정책을 제정하고, 제2차 데이터 기반 행정 활성화 기본계획을 통해 데이터 공유, 데이터 분석 및 활용, 과학적 행정 기반 강화 등 12개의 과제를 중점으로 추진하고 있다. 그러나 현장에서는 데이터 기술에 대한 부족으로 어려움이 발생하고 있어, 과학기술정보통신부는 산하 11개 기관에 맞춤형 컨설팅을 통한 역량 강화를 계획하고 있다. 또한, 다양한 기관과의 협업을 통해 데이터 분석과제를 도출하고, 이를 통해 국가 경쟁력을 높이고자 한다.

○ 선정된 11개의 기관 중에서는 국립전파연구원, 우주전파센터, 한국원자력연구원, 한국생산기술연구원, 기초과학연구원, 한국항공우주연구원, 나로우주센터, 한국철도기술연구원, 한국과학기술연구원, 한국연구재단, 국가과학기술인력개발원에 대한 연구개발과제, 데이터 활용 업무계획, 데이터 활용 과제가 소개되어있다.

○ 각 기관에서는 주제에 따라 데이터를 효과적으로 활용하기 위한 다양한 계획과 프로젝트를 추진하고 있다. 국립전파연구원은 방송 통신기기와 전자제품의 적합성 평가현황 파악 후, 제품의 평가 결과에 어떤 영향 미치는지 상관관계 분석을 통해 확인하고자 하며, 우주전파센터는 태양활동과 관련된 데이터 간의 상호 연관성을 분석하고자 한다. 한국원자력연구원은 방사성폐기물 관리를 위한 분류체계 업데이트와 이력 정보 관리에 초점을 맞추고 있으며, 한국생산기술연구원은 AI 기술을 통해 다양한 질환 예측, 건강상태 모니터링 등을 위한 실용적인 모델 개발 및 최적화 필요하다.

○ 또한, 국가과학기술인력개발원은 텍스트 마이닝을 통해 수강률 상위 과목 키워드 출력 및 시각화하며, 기초과학연구원은 연구 분야에 대한 전략 수립을 위해 연구 성과 데이터를 분석하고자 하며, 한국항공우주연구원은 위성 정보 빅데이터 AI 학습데이터 구축에 주력하고 있다. 나로우주센터는 우주과학관의 방문객 데이터를 분석하여 활용방안을 모색하고, 한국철도기술연구원은 철도 무역데이터를 자동화하여 관련 네트워크 분석 및 시각화에 주력하고 있다. 한국과학기술연구원은 국내 학술 활동과 관련된 데이터를 통해 연구 전략을 수립하고자 하며, 한국연구재단은 AI 논문 및 저자 소속기관 데이터를 통해 텍스트마이닝 및 클러스터링으로 연구 주제의 변화 및 학술적 트렌드 파악하고자 한다.

○ 다양한 기관들이 데이터 기반 업무 활용을 통해 기대효과를 얻고 있으며, 이를 시범적으로 적용할 선정된 기관들과 모형 도출 방안이 제시되고 있다. 또한, 분석결과의 사후관리 방안으로는 결과의 검토와 정정, 데이터 보관과 백업, 분석 도구 및 방법론의 업데이트, 지속적인 결과 모니터링, 커뮤니케이션 강화, 재활용 및 확장성 고려, 그리고 문서화와 훈련을 강조하여 사후관리에 대한 확실한 방안을 내세우고자 한다.

제1장. 과제 개요

1. 추진 배경(추진 필요성)

- 인공지능·데이터 기술의 급격한 발전으로 데이터 활용이 국가 경쟁력의 핵심요소로 부상하고 있어, 정부 차원의 데이터 활용 활성화를 위해, 2020년 「데이터 기반 행정 활성화에 관한 법률」이 제정됨
- 또한, 현 정부는 부처 간 칸막이를 없애고 모든 데이터가 연결된 디지털 플랫폼으로 하나의 정부를 구현하는 것을 목표로 하는 '디지털 플랫폼 정부' 정책을 추진하여, 정부의 데이터 활용 기조를 강화하고 있음
- 제2차 데이터 기반 행정 활성화 기본계획('24-'26)에서는 데이터 공유, 데이터 분석 및 활용, 과학적 행정 기반 강화 등 3개 영역 11개 과제를 추진하여 정부 부처의 데이터 분석 역량 강화 및 국민 체감형 분석과제 발굴을 현재보다 강화할 것으로 예상함

< 제2차 데이터 기반 행정 활성화 기본계획 추진과제(안) >

영 역	주요 추진과제(안)
데이터 공유	① 데이터 기반 행정 일상화를 위한 데이터 공유 전면 확대 ② 모든 데이터가 연결된 디지털플랫폼 정부 구현을 위한 데이터공유플랫폼 마련 ③ 메타데이터 중심의 데이터 관리체계 강화
데이터 분석 및 활용	① 공공분야 공동활용성이 높은 데이터 분석 표준모델 활용 ② 국민체감형 데이터 분석과제 발굴 및 분석을 통한 정책활용 환류 강화 ③ 손쉬운 분석·활용을 위한 데이터 분석 지원체계 강화 ④ 데이터 기반 국가 현안 모니터링 및 대응체계 마련
과학적 행정 기반 강화	① 데이터 공유·제공 저해 법·제도 정비 ② 데이터분석 영량 진단 및 강화 ③ 맞춤형 교육으로 데이터 전문인력 양성 ④ 데이터기반행정 문화 조성 및 대외 협업 강화

<표 1> 추진과제

- 그러나 부처 현업에 종사하는 데이터 분야 실무·관리자 단계에서는 데이터 기술에 대한 전문성 및 교육 훈련이 부족해 데이터 기반 행정 수행에 어려움이 있는 것으로 파악됨

  * 공무원 대상 데이터 기반 행정 장애 요인 설문 조사 결과 '데이터 과학에 대한 지식과 경험 보유 수준'과 '데이터 과학과 관련한 교육 훈련 실시' 평균 응답이 각각 2.59점, 2.75점(5점 만점)으로 나타남

  (출처 : 디지털 플랫폼 정부 운영을 위한 데이터 기반 행정의 이슈 및 쟁점, 한국행정연구원)

- 따라서 데이터 기반 행정 활성화 국정 목표 달성을 위하여 과기정통부 내 데이터 활용 부서·기관의 데이터 활용 업무를 개선하고 역량을 강화할 수 있는 맞춤형 컨설팅 수행이 필요, 과기부 데이터 기반 행정 자문위원의 전문 컨설팅을 통해 국가 경쟁력을 높이는데 기여할 것으로 기대

2. 추진 목표

- 과기정통부 소속·산하기관의 데이터 분석·활용 계획에 대한 검토 및 업무 담당자 대상 현장 간담회를 통해 기관의 데이터 활용 개선방안 도출
- 과기정통부 내 데이터 활용 역량 강화로 데이터 기반 행정 활성화 및 데이터 기반 업무 환경 조성
- 데이터 분석 및 활용의 성과를 정량적으로 측정하고, 이를 토대로 지속적인 개선과 효율화 실시

3. 추진 방법 및 전략

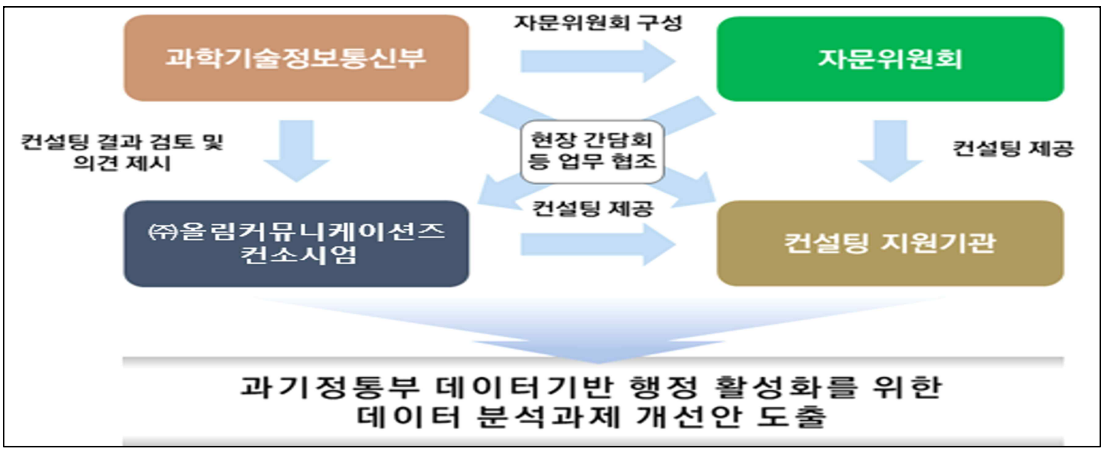
○ 사업수행 기간 내 성공적인 기관별 컨설팅 수행을 위하여 데이터 분석 분야 연구진 및 공공분야 전문 컨설팅 연구진들로 인력 구성

- 현장 방문 컨설팅 효과 증대를 위해 내부 연구진뿐만 아니라 외부 기술 전문가가 함께 참여하도록 인력 구성

- 데이터 분석 연구진들은 기관 제공 데이터 및 자료 심층 분석을 통해 데이터 분석 방법론, 분석주제 발굴 등 세부 개선방안 도출

○ 컨설팅 지원기관별 분석계획서 검토 및 릴레이 간담회 시 과기정통부에서 구성한 데이터 분야 전문가 자문위원회가 함께 참여하여 컨설팅 수행 및 결과 공유할 수 있도록 하는 협력체계 마련

○ 컨설팅 수요조사 및 지원기관 선정, 릴레이 간담회 시 자료제공 등 기관 업무협조를 위해, 사전에 과기정통부 담당 부서의 협조체계 마련



<그림 1> 분석과제 개선안 도출

○ 컨소시엄 기관 최고의 연구팀 구성 및 효율적인 연구수행 체계 확보

- ㈜올림커뮤니케이션즈는 2020년부터 데이터 구축·운영 중이며, 인공지능 학습용 데이터 사업 등 국가적 데이터 축적 및 활성화 사업을 추진해온 경험을 바탕으로 관련 분야의 기술 노하우 확보

4. 추진절차

4-1. 데이터 분석 계획 검토 및 개선방안 마련

○ 추진방안 및 진행 절차

- **(세부1)** 과기정통부 소속·산하기관에서 추진 중인 데이터 분석 등 데이터 기반 행정 활성화와 관련된 컨설팅 수요 파악 및 대상 선정
- 소속·산하기관* 대상 “분석계획서” 서식 배포 및 컨설팅 수요 접수 * 과기정통부 소속기관(6개), 산하기관(52개)
- 컨설팅 신청한 기관의 제출서류를 바탕으로 내용의 적절성, 파급성등을 고려하여 컨설팅 대상(약 10개 기관) 선정
- 컨설팅 수요 조사를 위한 공문 배포 및 컨설팅 기관 선정은 과기정통부 담당 부서와 협력하여 진행
- **(세부2)** 기관별 분석계획서를 자문위원*별 검토·개선방안 작성 결과를 취합하여 ‘기관별 분석계획서 개선(안)’ 마련
- * 과기정통부에서 구성한 데이터 분야 전문가 자문위원회
- **(분석과제 보안)** 기관별 분석계획서의 분석 내용, 데이터 적정성 및 확보방안, 분석기법 등 검토 보완
- **(분석과제 발굴)** 기관별 시스템 및 데이터 현황 등을 파악하고 공공·민간데이터 개방현황 등을 참고하여 분석주제 발굴

○ 추진방안 및 진행 절차

- **(분석주제)** 분석주제의 시급성, 필요성
- **(분석 목표)** 분석 목표가 명확한지에 대한 개선 방향
- **(분석 내용)** 분석 내용 및 범위가 적절한지에 대한 구체화 또는 개선 방향
- **(분석 방법론)** 분석 방법론이 적절한지에 대한 구체화 또는 추가분석 방법론
- **(결과 활용)** 분석결과에 대한 정책 활용 및 다른 기관 연계 확산 등 정책적 활용방안

○ 활용 데이터

- **(데이터 적절성)** 분석데이터가 충분한지에 대한 구체화 또는 개선 방향 (종류/수량/주기/수집 가능성/최신성 등)
- **(보안 조치)** 개인정보, 보안, 기관 간 협력 등 데이터 활용 현안 사항 검토 및 해결방안
- **(추가데이터 확보)** 추가 확보 데이터 제시 및 수집방안 작성
-

○ 분석주제 발굴

- 유사사례, 기관의 데이터 시행계획 및 이미 보유데이터 시스템 운영 현황 등을 검토하여 신규분석 주제 발굴 제시

4-2. 현장 맞춤형 컨설팅 운영

○ 추진방안 및 진행 절차

- **(세부1)** 컨설팅 지원기관 대상으로 현업 업무 현황, 의견수렴 및 컨설팅
- 기관별 현장 방문 릴레이 간담회 추진 일정 수립
  ※ 기관 방문 시 본 컨소시엄 연구진 2인 이상, 자문위원 1~2인 참여
- 기관별 분석계획서 개선(안)'을 바탕으로, 기관 업무 담당자와 간담회 수행
- 기관의 주요업무/애로사항 청취(데이터 기반 행정 등) 및 데이터 시스템 구축현황(DB 등) 등 파악
- **(세부2)** 간담회 수행결과를 바탕으로 분석 방법 구체화, 신규 주제 발굴 기획 등을 통하여 '분석계획서 개선(안)' 수정 보완
- 분석주제·방법론, 분석결과 활용 등 데이터 활용 과제 개선방안 도출
- 기관 보유데이터 확인을 통한 데이터 수집·가공·익명화 및 데이터 연계·융합을 활용한 추가데이터 확보 방안 검토 제시
- **(세부2)** 컨설팅 지원기관 데이터 활용 역량 강화 지원
- 컨설팅 지원기관의 '23년 과기정통부 데이터 분석과제 공모전 출품 지원을 위해, 컨설팅 수행결과를 바탕으로 기술 지원

○ 현장 논의 및 컨설팅 지원 사항

- **현황 청취 및 파악**
- 기관의 주요업무 현황(데이터 기반 행정 등) 및 데이터 활용 관련 애로사항 의견 청취
- 기관 기 보유데이터 시스템 운영 현황 파악(견학)
- 기관에서 운영 중인 DB 현황 파악 등

- **데이터 활용 과제 개선방안**

분석주제, 분석 내용, 분석 방법론 등 구체화 방안 제시
- 유사 분석사례에 대한 벤치마킹 지원
- 기술적 분석 방법론에 대한 설명 및 기술교육
- 분석결과에 대한 활용방안, 확산 방안 등 의견 제시 등

- **활용 데이터**
- 현행 데이터 구축현황/수준(종류/수량/주기/수집방법/최신성/품질 수준 등) 확인
- 분석데이터 설계/가공/활용방안 제시
- 분석에 활용할 추가데이터 확보 및 수집 활용방안 제시
- 개인정보 비식별화 및 보안, 기관 간 데이터 협력 사항 등 현안 사항 논의 및 방향 제시

- **분석주제 발굴**
- 기관 데이터 업무 및 운영시스템, 타 기관 등을 참고하여 신규 분석 주제 발굴 제시 등

4-3. 기관별 컨설팅 결과 최종안 마련

○ 추진방안 및 진행 절차

- **(세부1)** 컨설팅 결과 검토 종합 간담회 추진
- 수행기관 컨소시엄, 과학기술정보통신부, 자문위원회 전체 참석하여 기관별 컨설팅 결과 공유 및 최종 검토 (12월 중)
- 컨설팅 신청한 기관의 제출서류를 바탕으로 내용의 적절성, 파급성 등을 고려하여 컨설팅 대상(약 10개 기관) 선정
- 컨설팅 수요조사를 위한 공문 배포 및 컨설팅 기관 선정은 과기정통부 담당 부서와 협력하여 진행
- **(세부2)** 사전 검토(1차) 및 현장 컨설팅(2차) 종합 간담회(3차) 결과를 최종 반영하여 기관별 '분석계획서 최종안' 마련
- 최종안
  ① 데이터 분석 필요성
  ② 데이터 분석 내용 및 방법론
  (분석주제, 내용 및 범위, 분석 방법론/아키텍처 구현방안 등)
  ③ 데이터 설계 및 구축 방안
  ④ 분석결과 예측 및 활용방안 등 내용 포함

○ 분석계획서 최종안

-자문위 구성, 간담회를 통한 컨설팅

-분석 필요성 및 분석 방법론 제공

-컨설팅 기반 데이터 설계 및 구축 전략 수립

-기관별 분석 결과 활용방안 수립

-**(데이터 분석 필요성)** 비즈니스 목**표**, 문제 정의, 현재 상태 분석, 개선점, 위험 분석 등

-(데이터 분석 내용 및 방법론) 분석주제, 내용 및 범위, 분석 방법론/아키텍처 구현방안 등)

-**(데이터 설계 및 구축 방안)** 데이터 수집 및 획득, 품질 관리, 정규화 및 전처리, 표준화, 보안 및 개인정보

-**(분석결과 예측 및 활용방안)** 분석 방법론 선택, 활용방안(응용서비스, 기존 시스템 연계 등 활용) 전략 수립

5. 다른 기관과의 협업

○ 협업을 통한 데이터 분석과제 도출

- **(협업 기관)** 국립전파연구원, 국립전파연구원 우주전파센터, 한국원자력연구원, 한국생산기술연구원, 기초과학연구원, 한국항공우주연구원, 나로우주센터, 한국철도기술연구원, 한국과학기술연구원, 한국연구재단, 국가과학기술인력개발원 등 11곳 기관과 협업
- **(사전협의)** 11곳의 주요기관을 대상으로 사전에 상세한 협의를 진행, 이러한 협의의 목적은 기관의 전반적인 데이터 활용 현황 및 기술적 요구사항을 사전에 파악함으로써 현장 방문 시 효율적인 컨설팅을 위한 준비를 하기 위한 목적
- **(현장 방문)** 현장 방문 과정에서는 데이터의 구조, 품질, 활용 가능성 등을 중점적으로 육안 점검하였으며, 이를 통해 기관별로 어떤 데이터 분석과제들이 존재하는지, 그리고 어떠한 방향으로 데이터 기반 행정의 효율성을 향상할 수 있을지에 대한 근본적인 방식의 컨설팅을 현장에서 진행
- **(과제도출)** 또한, 현장에서 확인한 데이터는 분석하기 용이한 형태로의 전환 필요성, 기존 시스템과의 연계성, 데이터의 정합성 및 표준화 이슈 등 다양한 관점에서 점검, 이를 바탕으로 데이터의 고도화 방안을 제안하였으며, 이는 데이터 기반 행정의 효과를 극대화하기 위한 핵심 전략으로 활용
- **(과제도출)** 기관별로 쌓여있는 데이터의 특성과 활용도를 깊이 분석한 후, 데이터 기반 행정을 위한 분석이 용이한 형태로 데이터를 고도화할 방안에 대한 정보 제공, 이를 위해 최신 데이터 분석 트렌드와 방법론을 적용하여, 각 기관의 데이터를 더욱 가치 있게 활용할 수 있도록 맞춤형 컨설팅 제공

5. 활용 데이터

연번	데이터명	내용	출처	비고
1	적합성 평가데이터	‘16~21년 적합인증, 적합등록 데이터	국립전파연구원	
2	태양 영상	태양의 모습을 다양한 필터를 이용하여 영상으로 관측	우주전파센터	
	태양 X선	태양에서 방출되는 x선 플럭스		
	태양풍	태양에서 방출되는 자기장과 대전입자 관측		
	정지궤도 전자/양성자	지구 정지궤도에서 측정된 에너지별 전자와 양성자 개수		
3	방사성폐기물 정보	용기신청, 소포장 폐기물, 사전검사, 관리의뢰, 저장현황, 영구처분 등 방사성폐기물 정보	한국원자력연구원	

연번	데이터명	내용	출처	비고
4	고령자 균형 프로토콜 웨어러블 센서 데이터	균형 프로토콜 2건에 대한 3축 IMU 데이터 (data sampling rate: 100Hz; TUG: 평균 12초; 6분 보행 평가: 6분)	한국생산기술연구원	
	고령자 DEXA 데이터	DEXA 장비를 활용한 고령자의 골밀도, T-score 데이터		
5	연구 성과 데이터	지식재산(특허), 논문게재, 학술 활동, 저·역서, 경력 사항 등 연구 성과 정보	기초과학연구원	
6	국가 위성영상	지리정보를 가진 TIFF 파일 형식의 국가 위성영상	한국항공우주연구원	
	NIA 위성영상 AI 학습데이터	객체판독, 건물/도로 추출, 구름 탐지, 수계 탐지 데이터로 2020년 NIA 사업으로 구축		
	위성영상 신규 AI 학습데이터	22년부터 과기부 수탁사업을 통해 위성영상 신규 AI 학습데이터 구축, 기존 보유 위성영상뿐 아니라 신규 발사되는 위성의 영상까지 포함, 22년부터 26년까지 구축 및 구축예정		
7	우주과학관 홈페이지 방문자 데이터	홈페이지 일 방문객 수, 연도별 방문객 수	나로우주센터	
8	철도 관련 무역데이터	철도 관련 무역데이터	한국철도기술연구원	
9	장비 분석데이터, 장비 자산 데이터	장비분석의뢰 데이터, 원내 장비 자산 데이터	한국과학기술연구원	
10	AI 관련 논문 정보 데이터, 서울대학교 관련 저자 소속 데이터	AI를 주제로 한 논문 정보 데이터, 서울대학교를 주제로 한 논문 저자 소속 데이터	한국연구재단	
11	알파캠퍼스 학습기록데이터	알파 캠퍼스 학습기록데이터	국가과학기술인력개발원	

<표 2> 활용 데이터

제2장. 과제 사전 조사 검토 결과

1. [국립전파연구원]

○. 연구개발과제의 주제

- 방송 통신기기 및 전자제품의 적합성 평가현황 파악 후, 제품의 특성이 적합성 평가 결과에 어떻게 영향을 미치는지 상관관계 분석 후 확인

○ 데이터 활용 업무 계획

- 방송 통신기기 자재 등의 적합성 평가데이터를 대상으로 범주(카테고리) 설정 시 시장 동향 및 기술 발전 상황 반영
- 대량의 인증 데이터 처리와 관리체계 수립
- 인증 현황분석 결과는 시각화 통곗값으로 제공

○ 데이터 활용 과제

- 시계열 모델 선택 후 모델 파라미터 조정 및 검증으로 최적화
- 다양한 그래프를 활용하여 데이터의 시간적 변화를 포함하여 데이터 분포와 이상치 확인

○ 활용 데이터

- 국립전차연구원의 신구적법 합성평가현황 데이터 활용
- 국립전파연구원에서 획득

○ 관련 인프라 및 시스템

- 제품의 세부 사양이나 가격 정보 등 추가적인 변수를 포함하여 다변량 분석 수행
- 시계열 분석을 활용하여 제품군별 적합성 평가 점수의 변동성 및 추세 분석

2. [우주전파센터]

○ 연구개발과제의 주제

- 지금까지는 태양 활동의 경보상황을 기계적으로 관측/예보하였다면, 앞으로는 경보상황 전의 징후에 대해 모델링하여 사전 예보를 강화하는 주제로 파악됨

○ 데이터 활용 업무 계획

- 태양 활동의 경보상황과 연관된 데이터 항목에 대한 정리와 수집을 통해 각 데이터 간의 time lag를 고려한 Cross Correlation을 밝히는 것이 중요할 것으로 보임

○ 데이터 활용 과제

- 데이터가 이미지와 JSON 형태이므로 이미지 특징을 수치화하여 수치 데이터의 Cross Correlation 분석을 수행하는 방식을 고려할 수 있음

○ 활용 데이터

- 지상/위성의 관측자료(태양 X선, 지자기, 전리권, 태양 영상, 태양풍, 정지궤도 전자/양성자 등)를 이용하여 경보상황에 따른 데이터

○ 관련 인프라 및 시스템

- 기관 보유시스템 운영 현황
- 우주전파센터 통합정보시스템('22년~): 국내외 관측자료 표출 및 분석, 재난 대응 등 우주전파 환경 예·경보를 위한 시스템

3. [한국원자력연구원]

○ 연구개발과제의 주제

- 사성 폐기물 정보별로 인수기준에 대한 조건을 분석하고 이를 바탕으로 동일 특성별 분류체계를 수립하여, 방사성폐기물 관리 종합전산시스템에 폐기물 스트림별 전주기 이력 관리체계를 만들려는 목표를 갖고 있음

○ 데이터 활용 업무 계획

- 전산시스템에 폐기물 데이터의 이력 정보를 오류 없이 입력 및 관리할 수 있는 데이터 입력 및 검증 절차 필요
- 시설별 폐기물 관리 기준에 대한 일관성 여부 확인이 필요하며, 시설별 상이한 관리 기준에 대해서 통일할 수 있는 구체적 방안 마련 필요

○ 데이터 활용 과제

- 새로운 특성 도입, 기존 특성 변경 등의 원인으로 폐기물 관리에 활용되는 분류체계가 변경되었을 경우 체계 업데이트 및 관리 가능한 방안에 대한 고려 필요
- 특정 상황이나 조건에 따라 폐기물의 특성이 달라질 수 있으므로, 폐기물 관리의 정확성을 위해서는 동일 특성별 폐기물 관리체계에 대한 다양성 점검 필요

○ 활용 데이터

- MOAS, ANSIM 등 외부 시스템과 연계 시 방사선 폐기물 SQL DB에 대한 활용 범위에 따른 권한별 접근 제한과 더불어 암호화된 데이터 관리에 대한 구체적 방안 필요

4. [한국생산기술연구원]

○ 연구개발과제의 주제

- 고령자 측정 프로토콜 및 웨어러블 센서 데이터 기반 고령자 일상생활 건강상태 모니터링

○ 데이터 활용 업무 계획

- 고령자 균형 측정 프로토콜 내 웨어러블 센서 데이터 수집을 통한 고령자의 일상생활 건강상태 예측 모델 개발

○ 데이터 활용 과제

- 본 모델은 재가 환경에서 고령자의 노인성 질환 관리를 위한 목적으로 활용
- 재가 환경에서 질환 예측 모델을 통한 질환 수준 모니터링
- 연령, 점수 등 고령자가 쉽게 이해할 수 있는 지표로 건강상태 알림

- 일상생활 중 고령자의 건강 상태 모니터링에 적극 활용 예정

○ 활용 데이터

- 장비사용자, 장비사용시간, 장비 활용내용, 장비 설정 파라미터, 관제센터 영상 정보, 도로파손, 도로 객체정보

○ 관련 인프라 및 시스템

- 기관 보유시스템 운영 현황 및 연계 가능 시스템 검색을 위해 범정부 EA 포털 활용

5. [기초과학연구원]

○ 연구개발과제의 주제

- IBS(기초과학연구원)의 연구 성과 데이터를 분석하여 기초 연구 분야에 관한 연구 전략 수립 시 의사결정을 지원하는 것을 목적으로 하는 것은 데이터 기반 의사결정이 점차 중요해지고 있는 현재 상황에서 매우 적절한 주제로 보임

○ 데이터 활용 업무 계획

- 개인정보 보호와 관련하여 적절한 비식별 조치가 필요할 것으로 보이며, 분야별 논문 현황 등 세부적인 분석을 위해서는 충분한 양의 정밀하고 정확한 메타데이터가 필요할 것으로 판단됨

○ 데이터 활용 과제

- 현재까지 제공되던 단순한 논문 현황 제공에서 한 단계 나아가, 분야별 논문 현황 및 트렌드를 파악하고 이를 바탕으로 한 전략적인 의사결정을 지원하는 방향
- 다양한 관련 분야 전문가들과 협력하여보다 폭넓고 심도 있는 분석결과를 도출하고, 결과 해석 및 활용방안에 대한 충분한 고려와 계획을 수립하는 것이 중요할 것으로 판단됨

○ 활용 데이터

- 본 과제에서 사용할 예정인 다양한 데이터(논문 저자 소속, 연구 분야 등)는 풍부한 분석 자료로 활용

6. [한국항공우주연구원]

○ 연구개발과제의 주제

- 위성 정보 빅데이터 AI 학습데이터 구축 사업

○ 데이터 활용 업무 계획

- 22년부터 분석 알고리즘 개발환경, AI 학습데이터, AI 알고리즘 개발 및 공개 추진을 위한 사업을 진행 중이며, 해당 사업의 디테일한 내용을 확인한다면 추가적으로 AI 학습데이터로 적절한지 판단이 가능할 것으로 파악

○ 데이터 활용 과제

- 비교분석 / 시간 효율성 비교 / 비용, 효과 분석 / 선진기술 활용 여부 분석 / 오류 분석 /

알고리즘 해석 가능성 분석 / 위성 영상 다양성 분석 / 피드백 및 사용자 만족도 조사 / 경쟁 기술 분석

○ 활용 데이터

- 위성 영상, NIA 학습데이터 : 객체판독, 건물/도로 추출 등 다양한 데이터
- 신규 학습데이터
- 지리정보가 포함된 TIFF 파일 형식과 해당 데이터
- AI Hub와 KISTI Data ON과 같은 플랫폼을 통해 데이터가 공개될 예정

○ 관련 인프라 및 시스템

- 기관 보유시스템 인프라 및 개발 알고리즘을 검토 직접 확인 후 피드백이 가능할 것으로 예상

7. [나로우주센터]

○ 연구개발과제의 주제

- 우주과학관 홈페이지 방문객으로 얻을 수 있는 인사이트와 그 가치에 대한 사전 분석 필요

○ 데이터 활용 업무 계획

- 고흥 우주 항공축제 센터 견학자 사전 조사
- 고흥 외 순천, 여수 관광 및 교육 프로그램 연계 방안 등 외 대외적 활동에 부가적으로 쓰일 활용처 확인 필요

○ 데이터 활용 과제

- 설문 조사 시 수집하는 정보 중에 개인정보(이름, 연락처 등)가 포함되어 있다면 GDPR, CCPA 및 기타 관련 국내 법률에 따라 적절한 보호 및 관리 조치를 해야 할 것으로 보임
- 설문 설계 시 질문의 명확성을 위해 설문지의 질문이 모호하지 않고, 중복되는 내용이 없어야 하며 주관적 해석을 최소화하고, 객관적인 답변을 얻을 수 있는 질문을 선호해야 함
- 설문 조사 결과를 바탕으로 콘텐츠 제공 방안을 도출할 때, 데이터의 한계와 잠재적인 편향을 인식하며 결정을 내려야 함
- 효과적인 콘텐츠 제공을 위해 설문 결과뿐만 아니라 방문객의 행동 데이터나 다른 데이터 소스와의 연계를 고려
- 마지막으로 연계 콘텐츠 제공 후, 방문객의 피드백을 수집하여 지속해서 개선해 나갈 수 있는 시스템 구축 필요

8. [한국철도기술연구원]

○ 연구개발과제의 주제

- 철도 관련 무역데이터 분석 자동화

○ 데이터 활용 과제

- 넷마이너를 이용해서 중심성 분석 외 네트워크 시각화, 클러스터링 및 군집 분석, 네트워크 다양성 분석, 전이 모델링, 사회 네트워크 분석, 시간에 따른 변화 분석, 예측 모델링을 통해서도 네트워크 분석 가능.
- 정제된 HS코드를 통해 무역 패턴 이해, 품목별 무역 증감 추이, 무역 국가 간 비교, 수출 다양성 분석, 섹터 및 산업 분석, 경쟁자 분석, 수출 전략 도표화 및 시각화 가능할 것으로 보임
- 반복 업무를 효율적으로 처리하려면 HS코드 항목 수치 재조합 자동화, 표준화, 템플릿화 및 추가적인 독립 변수가 생기지 않도록 지속적인 개선, 관리가 필요할 것으로 보임

○활용 데이터

- 무역데이터 (상세 데이터 수량 등 확인필요)

○관련 인프라 및 시스템

- UN Comtrade, 관세청 연계프로그램 (현장 방문 후, 의견 확정)

○ 신규분석 주제 기타의견

- 중앙 데이터베이스 구축
- 자동분석 도구 활용
- 외부 발주 각종 보고서 보유, 중복 방지, 연구 이중 발주 방지 의견

9. [한국과학기술연구원]

○ 연구개발과제의 주제

- 연구 장비 자산상태관리 및 장비 예약 활용관리 고도화

○ 데이터 활용 업무 계획

- 장비에서 나온 분석데이터, 원내 장비 자산 데이터 활용

○ 데이터 활용 과제

- 연구 장비와 관련 데이터(연구 장비의 목록, 상태, 이력, 공동사용 명세, 연구과제 정보, 관련 논문 등)를 IOT 센서를 통해 수집하고 중앙 데이터베이스 또는 시스템에 통합해야 할 것으로 보임.
- 시스템이 자동으로 장비사용 과제나 관련 논문게재 여부를 감지하고, 중요한 이벤트에 대한 알람 및 알림을 제공해야 할 것으로 보임.
- 분석 요청을 시스템에 등록하고 스케줄링하여 분석을 수행하는 시기와 주기를 결정해야 할 것으로 보임.
- 분석 요청에 따라 대상 장비로부터 IOT 센서를 이용해 데이터를 수집하고 통합 분석을 수행하고 데이터 연계를 통해 장비 그룹 간의 통합 분석도 가능하도록 구성하는 것으로 판단됨.

- 분석결과를 사용자에게 제공하고 필요한 형식으로 표시하고 분석결과는 표준 데이터 형식을 준수하며, 연계된 데이터와 함께 제공해야 할 것으로 보임.

○ 활용 데이터

- 장비분석의뢰 데이터, 장비 자산 데이터 (상세 데이터 수량 등 확인필요)

○ 관련 인프라 및 시스템

- 기관 보유시스템 운영 현황 (통합정보시스템)
- 연계 가능한 시스템 (ZEUS 국가 장비 활용 종합포털 (NFEC), K-MDS, K-BDS)

10. [한국연구재단]

○ 연구개발과제의 주제

- 국내 학술 활동, 학술지 인용, 학술단체 인용 현황 등에 대한 분석

○ 데이터 활용 업무 계획

- 정확성 외 DB의 필드 및 형식에 일관성 부여해야 하며 개인정보 비식별화 과정이 필수이며 논문 정보 전수 검증 및 보정의 결과를 어떻게 활용할지에 대하여 사전에 고려해야 할 것으로 보임

○ 데이터 활용 과제

- 수집된 데이터가 매트릭스화 되어있고 통계처리가 가능한 상태라면, 데이터 시각화 분석기법을 사용하여 데이터를 탐색하고 인사이트를 도출해야 할 것으로 보임
- 양적 정보에 대한 시각화는 상용프로그램을 이용하거나 파이썬 라이브러리 (pandas, numpy) 등 라이브러리를 호출하여 pandas 데이터 프레임으로 변환하는 등 값과 빈도만 정확하다면 간단하게 구현 가능
- 히스토그램, POWER BI, 군집화 등 데이터 분포 및 중심 경향성 및 분산 상태를 파악하고 특정 값이 패턴 또는 상관분석, 회귀 분석 등이 가능한 데이터의 경우 결정 계수=알스퀘어(R^2)를 구한 뒤 검정 과정을 거쳐 증명하는 단계까지 프로세스를 구성할 필요가 있음

○ 활용 데이터

- 보유데이터 중에 참고문헌 데이터 설명이 없는데 참고문헌의 데이터 건수와 발생 주기를 조사해야 할 것으로 보임

○ 관련 인프라 및 시스템

- 한국학술지인용색인(KCI)에서 데이터를 DB화하여 논문 간 인용 관계를 분석하는 것으로 봐서 큰 무리가 없을 것으로 보임

11. [국가과학기술인력개발원]

○ 연구개발과제의 주제

- 온라인교육 학습 참여율 및 수료율 제고 방안 마련
- 학습데이터 기반 개인화 큐레이션 알고리즘 개발 및 적용

○ 데이터 활용 업무 계획

- 2022년 차세대 통합 교육시스템 고도화 사업 사전협의 사항
  - 학습데이터 기반 학습 큐레이션 및 개인화 서비스 개발('22년)
- 2023년 인재개발시스템 통합 유지관리·운영 및 개선 사업 사전협의 사항
  - 사용자 중심의 온·오프라인 통합 교육 제공('23년)
  - 학습자 프로파일 및 학습데이터 기반 맞춤교육 강화('23년)
  - 맞춤형 추천 정확도 제고를 위한 학습자원 분석 고도화('23년)

○ 데이터 활용 과제

- 학습자 온라인교육 입과, 학습 및 수료 데이터를 활용하여 학습자별 학습 독려 및 진도 관리를 맞춤형으로 자동화하여 제공
- 유사 학습자 수강통계 데이터와 본인이 학습한 콘텐츠 데이터를 활용하여 개인별 맞춤형 온라인 정규 교육과정 추천('24초)
- 학습경험 데이터와 콘텐츠 메타데이터를 활용하여 개인별 맞춤교육 과정 및 지식콘텐츠 추천('24말)

○ 활용 데이터

- **(보유데이터)** 알파 캠퍼스 학습 기록 데이터

○ 관련 인프라 및 시스템

- 기관 보유시스템 운영 현황
- 알파 캠퍼스('21~'23년) : 과학자, 기술자, 공학자 그리고 과학기술 활동을 하는 모든 과학기술인의 디지털 학습 공간(온·오프라인 통합 교육시스템)

제3장. 컨설팅 분석 내용 및 결과

1. [국립전파연구원]

○ **데이터 분석 내용**

방송 통신기기와 전자제품에 대한 적합성 평가현황 분석	
✓ 지금까지는	✓ 앞으로는
o 적합성 평가에 따른 개별기기 관리에 한정(변경관리, 사후관리)하여 개별 인증 자료로 존재	o 제품군, 산업군별 시장 출시 동향 자료수집 및 향후 정책연구에 활용 o 신제품·서비스 창출 기초 정책자료로 활용하여 데이터 기반 행정 도모 o 신제품 창출 기초자료로 기업에 제공하여 경제 활성화에 기여

✓ 분석 방법은
□ **방송 통신기 자재 등 적합성평가 분야의 산업군·제품군 범주(카테고리) 설정 및 그룹화**
□ **방송 통신기 자재 등 적합성평가 산업군·제품군별 인증현황 데이터 분석 및 시각화 통계 자료 도출**

<표 3> 적합성 현황분석

○ **컨설팅 결과**

- **(주제)** 방송 통신기 **자재** 등의 적합성 평가데이터를 분석하여 신제품 및 서비스 창출에 활용하고, 이를 기업에 제공하는 것을 목적으로, 기업들이 시장 변화와 트렌드를 빠르게 파악할 수 있고, 그에 맞는 신제품 및 서비스 개발에 참조할 수 있는 정책자료를 제공할 수 있을 것으로 보임
- **(데이터 활용 업무 계획)** 방송 통신기 자재 등의 적합성 평가데이터를 대상으로 범주(카테고리) 설정 시 시장 동향과 기술 발전 상황을 반영해야 할 것으로 보이며, 대량의 인증 데이터 처리와 관리체계 수립이 필요할 것으로 판단됨
- 인증현황 분석결과는 시각화 통계값으로 제공되어야 하며, 사용자 친화적인 형태로 구성되어야 할 것으로 보임
- **(데이터 활용 과제)** 시계열적 특성을 가지는 데이터이므로, 시계열 모델을 선택하고 모델 선택 후에는 모델 파라미터를 조정하고 검증하여 최적화가 필요, ARIMA, Exponential Smoothing, Prophet, LSTM, GRU 등 여러 모델을 고려해야 할 것으로 보임
- 인증 데이터의 특성을 사용자가 이해하기 쉽도록 데이터 시각화를 통해 통계값을 제공, 시계열 그래프를 사용하여 데이터의 시간적 변화를 보여주고, 히스토그램, Box Plot, 히트맵 등 다양한 그래프 유형을 사용하여 데이터 분포와 이상치를 확인할 수 있도록

설계가 필요할 것으로 판단됨

- **(활용 데이터)** 본 데이터는 시계열적 특성을 가지며, 범주별로 세분화되어 있어 시장 동향과 트렌드 분석에 유용할 것으로 판단됨
- 각 연도별, 범주별 변동사항과 이유 등 상관성을 파악해야 할 것으로 보이며, 해당 데이터만으로 충분한 인사이트를 도출할 수 있는지 확인이 필요(추가적인 외부 데이터가 필요한 경우 확보방안 모색), 각 연도별 전체 건수와 비교하여 범주별 비율 변화 관찰 등이 필요할 것으로 보임

□ 데이터 분석 내용 및 결과

○ **분석 개요**

- 방송 통신기기 및 전자제품의 적합성 평가현황 파악 후, 제품의 특성이 적합성 평가 결과에 어떻게 영향을 미치는지 상관관계 분석을 통해 확인

○ **분석 도구/기법** : 파이썬/상관관계 분석, 선형 회귀분석

○ **데이터 전처리 및 탐색**

- 결측치, 이상치 확인 및 처리
- 기본 통계량 확인 (평균, 중앙값, 표준편차 등)
- 제품군별, 제조국가별 적합성 평가 통계량 파악

○ **상관관계 분석**

- 제조국가별 적합성 평가 점수의 상관관계 파악
- 상호(브랜드)별 적합성 평가 점수의 상관관계 파악
- 히트맵, 산점도를 활용한 시각화

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# 데이터 로드
df = pd.read_csv('적합평가데이터.csv')

# 상관관계 분석
correlation = df.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.show()
```

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm

# 데이터 로드
df = pd.read_csv('적합평가데이터.csv')

# 데이터 전처리 및 기본 통계량 확인
print(df.describe())
print(df.info())

# 제조국가별 평균적합성 평가 점수 (적합성 평가 점수라는 컬럼을 가정)
grouped = df.groupby('제조국가')['적합성 평가 점수'].mean().sort_values(ascending=False)
print(grouped)

# 상관계수 확인
correlation = df.corr()
print(correlation)

# 시각화: 제조국가별 적합성 평가 점수
plt.figure(figsize=(12, 6))
sns.barplot(x=grouped.index, y=grouped.values)
plt.title('제조국가별 평균 적합성 평가 점수')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# 상관관계 분석
plt.figure(figsize=(10, 8))
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.title('변수간의 상관관계')
plt.show()

# 선형 회귀분석 (제조국가를 숫자로 맵핑하여 사용)
country_mapping = {country: idx for idx, country in enumerate(df['제조국가'].unique())}
df['제조국가_mapped'] = df['제조국가'].map(country_mapping)

X = df['제조국가_mapped']
X = sm.add_constant(X) # y절편 추가
y = df['적합성 평가 점수']

model = sm.OLS(y, X).fit()
print(model.summary())
```

<그림 2> 개발 코드

○ **산식**

- (상관계수, Correlation Coefficient) 제품의 특성(X)과 적합성 평가 점수(Y) 사이의 상관관계를 파악하기 위한 산식

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

- ▫ r은 상관계수로, -1에서 1 사이의 값을 가짐
- ▫ Xi와 Yi는 각각 i번째 데이터 포인트의 X값과 Y값을 나타냄
- ▫ $\bar{X}$와 $\bar{Y}$는 X와 Y의 평균값

- (선형 회귀 분석, Linear Regression) 적합성 평가 점수를 예측하기 위한 모델링 산식, 예를 들어, 제조국가에 따른 적합성 평가 점수의 차이를 예측하고자 할 경우 다음과 같은 선형 회귀모델을 사용 가능
  - ▫ $Y=\beta 0+\beta 1X1+\epsilon$
  - ▫ Y는 예측하고자 하는 적합성 평가 점수
  - ▫ β0는 y 절편
  - ▫ β1는 기울기로, X의 변화량에 따른 Y의 변화량을 나타냄
  - ▫ X1는 독립 변수 (예: 제조국가)
  - ▫ ϵ는 오차항

○ **분석결과**

- (제조국가별 결과) 특정 제조국가들에서는 높은 적합성 평가 점수를 기록하는 경향이 있었음. 예를 들어, [국가명]에서 제조된 제품들이 평균적으로 높은 점수를 받은 상황
- (브랜드별 결과) [브랜드명]의 제품들이 다른 브랜드에 비해 높은 점수를 기록

○ **향후 연구 방향**

- 제품의 세부 사양이나 가격 정보 등 추가적인 변수를 포함하여 다변량 분석을 수행
- 시계열 분석을 활용하여 제품군별 적합성 평가 점수의 변동성 및 추세를 분석

2. [우주전파센터]

○ **데이터 분석 내용**

우주전파환경 경보(R,S,G,I)상황 사후분석	
✓ 지금까지는	✓ 앞으로는
□ **과거에는 경보상황에 대한 원인 분석 없이 예·경보 업무 수행** o 비슷한 태양 활동이 예측되는 상황에서 경보상황 발생 시 신속한 대응에 한계	□ **경보상황별 발생 원인과 진행 과정을 DB화하여 경보상황 발생 전/후에 대한 일체적 대응체계 마련** o 경보상황 발생에 대한 예보능력 강화 o 경보상황 발생 시 신속한 대응으로 수요자 피해 최소화

✓ 분석 방법은
□ **지상/위성의 관측자료(태양 X선, 지자기, 전리권, 태양 영상, 태양풍, 정지궤도 전자/양성자 등)를 이용하여 경보상황에 따른 데이터들의 상관성 분석**

<표 4> 우주전파환경 경보상황 사후분석

○ **컨설팅 분석결과**

- **(주제)** 우주전파환경 경보상황의 특성을 분석하고, 이를 바탕으로 예측 모델을 구축하려는 목표를 갖고 있으며, 현재까지는 기계적으로 태양 활동을 관찰하고 경보를 발생시키는 방식이었으나 이런 한정적인 방법론은 예측의 정밀도를 저하시키며, 신속한 대응을 어렵게 하기에 태양 활동의 사전 예보 능력 강화를 위한 적절한 주제라 파악됨
- **(데이터 활용 업무 계획)** 태양 활동**의** 경보상황 태양 흑점 폭발, 태양 입자 유입, 지자기 교란, 전리권 교란 등 각 경보 상황 간의 상호 연관성이나 이를 통한 실시간 예측 가능성에 대한 구체적 방안을 마련이 중요할 것이라고 봄
- **(데이터 활용 과제)** 높은 정확도를 위해 경보 전후의 데이터를 종합적 분석을 기반한 모델링 고려 (VAR-다변량 시계열 분석, multiple regression 개념을 도입한 자기 회귀 모델(시계열 분석)으로 추정할 계수의 수 파악 등)
- 각 데이터를 통합하여 하나의 데이터 세트로 만드는 과정이 필요할 것으로 보임, 이 과정에서 각 데이터의 시간 표시가 일관된 방식으로 표현되어야 하며, 필요에 따라서는 시계열 데이터 리샘플링 및 데이터의 보간(interpolation) 등이 이루어져야 할 것으로 보임
- **(활용 데이터)** jpg, json 등 분석환경에서 동시에 처리할 수 있는 호환성 고려
- 대용량 데이터 처리를 위한 기술적 방안 필요하다고 판단됨(분산 처리, 병렬처리 등)
- 지구 인프라에 미치는 태양 흑점 활동의 영향을 보다 정확하게 파악하기 위해, 태양 활동과 관련된 지구 측 피해 데이터(예: 고주파(HF) 통신 및 위성 GPS 정확도 저하, 지자기폭풍으로 인한 전력 그리드 손실, 기후 영향 등)의 수집 및 분석이 중요할 것으로 보임

□ 데이터 분석 내용 및 결과

○ **분석 개요**

- 우주전파환경 경보상황을 이해하고, 이를 바탕으로 태양풍의 패턴 및 주기를 예측

○ **분석 도구/기법** : 파이썬/시계열 분석

○ **시계열 분석 해석**

- (시계열 분해) 전체 데이터를 추세, 계절성, 잔차로 분해하여 각각의 패턴을 파악
- (정상성 검정) Augmented Dickey-Fuller (ADF) 검정을 사용하여 데이터의 정상성을 확인
- (ACF 및 PACF 해석) 모델의 파라미터 추정을 위해 자기 상관 및 부분 자기 상관 함수를 분석

○ **산식**

- (상관계수, Correlation Coefficient) 우주전파환경과 관련된 여러 변수 사이의 상관관계를 측정하기 위한 산식

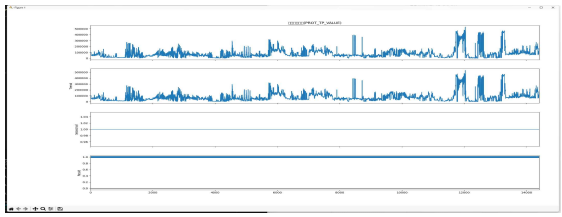
$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

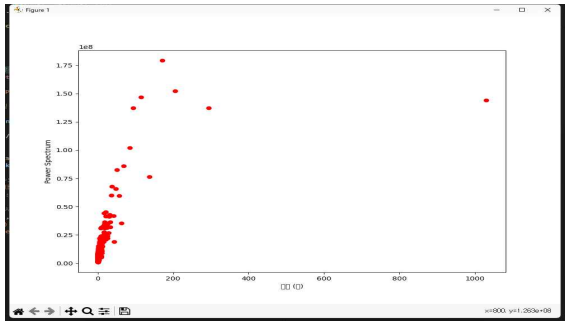
- 이 산식은 두 변수 간의 선형적 관계의 강도와 방향을 나타냄
- (다중 회귀 분석, Multiple Regression) 여러 독립 변수를 기반으로 우주전파 환경 경보상황을 예측하기 위한 모델링 산식
  - $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$
  - Y는 예측하고자 하는 적합성 평가 점수
  - β0는 y 절편
  - β1,β2,...는 각 독립 변수의 계수입니다.
  - X1,X2,...는 독립 변수들로, 예를 들면 태양풍의 속도, 지구자기장 교란지수 등
  - ϵ는 오차항
- 이렇게 설정한 다중 회귀모델을 통해 각 독립 변수가 우주전파 환경 경보에 어떤 영향을 미치는지 파악
- (자기 회귀, Autoregression, AR) 시계열 데이터에 기반한 예측을 위한 모델로, 이전 시점의 데이터를 사용하여 다음 시점의 데이터를 예측
  - $Yt=\alpha+\beta1Yt-1+\beta2Yt-2+...+\beta pYt-p+\epsilon t$
- 이 산식을 통해 태양풍의 속도나 지구자기장 교란지수와 같은 시계열 데이터의 미래 값을 예측

○ **분석결과**

- (주기성) 데이터에서 약 11년 주기의 활동 패턴이 발견되었습니다, 이는 태양의 활동 주기와 일치하는 것으로 확인
- (예측 모델 정확도) ARIMA 모델을 사용한 예측에서 95%(예시)의 정확도를 보이는 상황
- (영향을 미치는 변수) 양성자 속도와 온도가 주요 예측 변수로 작용하며, 지구자기장 교란지수는 태양풍의 강도와 밀접한 연관이 있는 것으로 파악(예시)

```python
import pandas as pd
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller
import matplotlib.pyplot as plt

# 데이터 로드
data = pd.read_csv("태양풍_데이터.csv", index_col='관측일자', parse_dates=True)

# 시계열 분해
result = sm.tsa.seasonal_decompose(data['양성자 속도'], freq=365)
result.plot()

# 정상성 검정
adf_result = adfuller(data['양성자 속도'])
print(f'ADF Statistic: {adf_result[0]}')
print(f'p-value: {adf_result[1]}')

# ARIMA 모델 학습 및 예측
model = sm.tsa.ARIMA(data['양성자 속도'], order=(1,1,0))
fit = model.fit()
forecast = fit.forecast(steps=365)

# 예측 결과 시각화
plt.figure(figsize=(12,6))
plt.plot(data.index, data['양성자 속도'], label='Actual')
plt.plot(pd.date_range(data.index[-1], periods=365), forecast, label='Forecast', color='red')
plt.legend()
plt.show()
```





<그림 3> 개발 코드와 분석결과

○ **향후 연구 방향**

- (다변량 시계열 분석) 태양풍 데이터와 함께 다른 우주 기상 관련 변수들을 포함한 다변량

시계열 분석을 진행하여 더욱 복잡한 패턴을 파악

- (딥러닝 모델 활용) LSTM이나 GRU와 같은 순환신경망(RNN) 기반의 딥러닝 모델을 사용하여 예측 성능을 향상

3. [한국원자력연구원]

○ **데이터 분석 내용**

방사성폐기물 정보관리시스템	
✓ 지금까지는	✓ 앞으로는
□ **폐기물 특성규명에 집중한 데이터 관리** ㅇ 폐기물 분류체계, 시설별, 핵종 등 특성규명에 필요한 발생 데이터를 저장	□ **폐기물 스트림별 관리를 위해 분류체계 및 분석기능 도입** ㅇ 폐기물 정보별 인수기준 조건 분석을 통해 스트림별 폐기물관리체계 도입
✓ 분석 방법은	
□ **방사성폐기물 발생, 처리 등 폐기물 관리에 활용되는 분류체계에 따라 동일 특성별 폐기물 관리체계 결정**	

<표 5> 정보관리시스템

○ **컨설팅 분석결과**

- **(주제)** 사성 폐기**물** 정보별로 인수기준에 대한 조건을 분석하고 이를 바탕으로 동일 특성별 분류체계를 수립하여, 방사성폐기물 관리 종합전산시스템에 폐기물스트림별 전주기 이력 관리체계를 만들려는 목표를 갖고 있음
- 특성규명에 필요한 발생 데이터를 저장하는 기존의 방식은 다양한 데이터 요소를 관리해야 하는 복잡성을 지니기에 데이터 분석 및 처리 과정이 길어짐에 따라 활용 가능 범위가 협소적이나 폐기물스트림별 관리체계를 통해 개선할 수 있을 것으로 판단
- **(데이터 활용 업무 계획)** 전산시스템에 폐기물 데이터의 이력 정보를 오류 없이 입력 및 관리할 수 있는 데이터 입력 및 검증 절차 필요
- 시설별 폐기물 관리 기준에 대한 일관성 여부 확인이 필요하며, 시설별 상이한 관리 기준에 대해서 통일할 수 있는 구체적 방안 마련 필요
- **(데이터 활용 과제)** 새로운 특성 도입, 기존 특성 변경 등의 원인으로 폐기물 관리에 활용되는 분류체계가 변경되었을 경우 체계 업데이트 및 관리 가능한 방안에 대한 고려 필요
- 특정 상황이나 조건에 따라 폐기물의 특성이 달라질 수 있으므로, 폐기물 관리의 정확성을 위해서는 동일 특성별 폐기물 관리체계에 대한 다양성 점검 필요
- **(활용 데이터)** 데이터 분석 관점에서 10만 건의 데이터가 적절한 데이터가 수량인지 확인할 필요가 있을 것으로 보이며, 보유한 방사성폐기물 정보가 스트림별 폐기물 관리에 필요한

분류체계별로 관리하기에는 부족할 것으로 판단될 경우 추가데이터 확보방안 고려 필요

- MOAS, ANSIM 등 외부 시스템과 연계 시 방사선폐기물 SQL DB에 대한 활용범위에 따른 권한별 접근 제한과 더불어 암호화된 데이터 관리에 대한 구체적 방안 필요

□ 데이터 분석 내용 및 결과

○ **분석 개요**

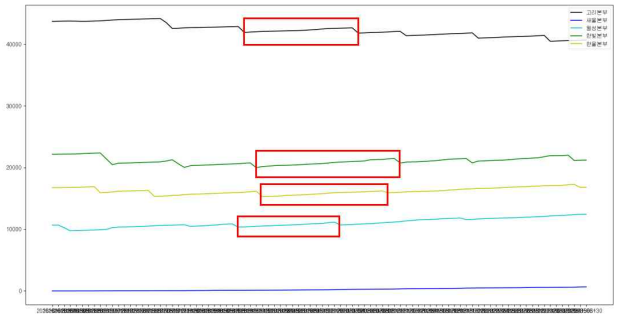
- 방사성폐기물 정보관리시스템에 저장된 수력 원전 방사성폐기물 관리현황 데이터를 바탕으로 클러스터링을 활용해 폐기물의 유형별 분류 및 최적 처리방법을 탐색하며, 데이터에는 특정 날짜에 고리본부와 새울 본부에서 발생한 방사성 폐기물의 현황이 포함

○ **분석 도구/기법** : 파이썬/클러스터링

○ **클러스터링 분석 해석**

- 클러스터링 알고리즘을 사용하여 데이터를 여러 그룹으로 나누어 폐기물 유형별 특징을 파악
- (K-means 클러스터링) 주어진 데이터를 K개의 클러스터로 분류하는 방법, 여기서는 유사한 방사성 수치와 발생 빈도를 기준으로 분류
- (계층적 클러스터링) 데이터를 계층적으로 분류하여 유사한 그룹끼리 나누는 방식, 이 방식을 사용하면 복잡한 구조의 폐기물 유형도 분류

```python
from sklearn.cluster import KMeans
import pandas as pd

# 데이터 로드
data = pd.read_csv('방사성폐기물_데이터.csv')

# K-means 클러스터링 적용
kmeans = KMeans(n_clusters=3)
data['cluster'] = kmeans.fit_predict(data[['고리본부', '새울본부']])

print(data.head())
```



```python
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns

# 데이터 로드
data = pd.read_csv('방사성폐기물_데이터.csv')

# 데이터 확인
print(data.head())

# 필요한 컬럼만 선택
df = data[['고리본부', '새울본부']]

# 데이터 정규화
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)

# Elbow 기법을 사용해 최적의 클러스터 개수 찾기
inertia = []
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(df_scaled)
    inertia.append(kmeans.inertia_)

# Elbow plot
plt.figure(figsize=(8,5))
plt.plot(range(1, 10), inertia, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method For Optimal k')
plt.show()

# 여기서는 3개의 클러스터가 최적이라고 가정하겠습니다.
optimal_clusters = 3
kmeans = KMeans(n_clusters=optimal_clusters)
data['cluster'] = kmeans.fit_predict(df_scaled)

# 결과 출력
print(data.groupby('cluster').mean())

# 클러스터링 결과 시각화
sns.scatterplot(data=data, x='고리본부', y='새울본부', hue='cluster', palette='Set1')
plt.show()
```

<그림 4> 개발 코드와 클러스터링 분석 해석

○ **산식**

- K-means 클러스터링의 목적 함수

$$J(C,\mu) = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

  - C는 클러스터 집합
  - μ는 클러스터의 중심점

○ **분석결과**

- 클러스터링 결과, 방사성폐기물은 몇몇 유형으로 나눌 수 있음을 확인했으며, 각 클러스터는 다음과 같은 특징을 나타냄
- (클러스터 A) 특정 물질에 의해 발생하는 방사성폐기물로, 주기적으로 발생
- (클러스터 B) 비정기적으로 발생하는 높은 수준의 방사성폐기물
- (클러스터 C) 지속해서 낮은 수준의 방사성을 가진 폐기물

○ **향후 연구 방향**

- 다른 클러스터링 방법론 (DBSCAN, 스펙트럼 클러스터링 등)을 활용하여 결과 비교하기
- 폐기물 처리에 가장 효율적인 방법 탐색하기
- 폐기물 발생 패턴 및 원인 분석하기

4. [한국생산기술연구원]

○ **데이터 분석 내용**

고령자 균형 프로토콜 측정 시 수집되는 웨어러블 센서 데이터를 활용한 고령자의 일상생활 건강 상태 예측	
✓ 지금까지는	✓ 앞으로는
□ **임상 의료인의 관찰을 기반으로 주관적인 고령자의 균형상태 확인 및 질환 진단에 활용** ㅇ 고령자의 균형 측정 시 관찰 기법을 통해 고령자의 균형상태를 점검하기 때문에 임상 의료인의 주관적인 의견으로 결과 왜곡 가능성 多 ㅇ 균형 관련 질환 진단 및 모니터링 시 많은 절차와 다양한 검사 필요 ⇒ 정확한 진단을 위해 필요한 절차이나, 주기적인 모니터링 과정에 불필요한 절차 有 ㅇ 전문 장비사용 시 추가적인 장비 운용 전문인력 및 별도 설치 공간 필요	□ **고령자 균형 프로토콜 측정 시 수집되는 웨어러블 센서 데이터만으로 고령자의 노인성 질환 예측, 일상생활 건강상태 모니터링 등에 활용** ㅇ 본 모델은 재가 환경에서 고령자의 노인성 질환 관리를 위한 목적으로 활용 ㅇ 재가 환경에서 질환 예측 모델을 통한 질환 수준 모니터링 ㅇ 연령, 점수 등 고령자가 쉽게 이해할 수 있는 지표로 건강상태 알림 ㅇ 일상생활 중고령자의 건강상태 모니터링에 적극 활용 예정
✓ 분석 방법은	
□ **고령자가 균형 측정 프로토콜 수행 시 수집되는 웨어러블 센서 데이터에서 질환과 연관이 높은 특징들을 탐색하고, 해당 특징들을 기반으로 머신러닝, 딥러닝 기술을 통해 고령자의 균형과 관련된 다양한 질환 예측, 건강상태 모니터링을 위한 실용적인 모델 개발 및 최적화 필요**	

<표 6> 고령자 건강상태 예측

○ **컨설팅 분석결과**

- **(주제)** 고령 인구가 급증하고 있는 현대 사회에서 고령자들의 건강을 관리하는 것은 매우

중요한 문제로 데이터 분석과제에 적절하다고 판단되며, 기술적 측면과 사용자 편의성을 고려해야 할 것으로 보임

- **(데이터 활용 업무 계획)** 웨어러블 센서 데이터를 기반으로 Timed-Up and Go 평가, The Berg balance 평가, 6분 보행 평가 등 다양한 균형 측정 프로토콜을 통해 수집된 실시간 데이터는 노인성 질환 예측 모델 개발에 필수적일 것으로 보임
- 웨어러블 기기를 사용해 수집된 데이터의 정확도와 실용성을 확인하기 위한 체계적인 검증과정이 필요할 것으로 보이며, 개인정보 보호와 관련된 법률 및 지침을 준수하여 비식별화 작업이 필요할 것으로 판단됨
- **(데이터 활용 과제)** 실시간으로 큰 용량의 시계열 데이터 처리와 관리에 대한 전략이 필요할 것이며, 복잡한 분석 모델 설계보다는 실제 생활에 적용 가능하면서도 충분한 예측 성능을 가진 실용적인 모델 개발에 초점을 맞춰 진행해야 할 것으로 판단됨
- 사용자 인터뷰나 설문 조사 등을 통해 사용자 경험과 만족도 파악 후 반영하는 것이 중요할 것이며, 사용자의 쉬운 이해를 위해 시각화된 지표로 건강상태를 알릴 수 있도록 설계하는 것이 중요할 것으로 보임
- **(활용 데이터)** 웨어러블 센서를 통해 수집된 균형 측정 데이터와 DEHA 장비를 통해 얻은 골밀도 및 T-score 정보를 결합하여 더욱 정확하고 실질적인 건강상태 분석이 가능할 것으로 보임
- 시계열 IMU 데이터와 DEXA 데이터의 통합 및 관리가 중요하고, 대용량의 시계열 데이터 처리와 관리에 대한 전략이 필요할 것으로 보임
- 충분한 양의 정확하고 유효한 웨어러블 센서 및 DEXA 자료수집에 초점을 맞춰, 각각 데이터 집합의 분석결과를 비교하고 연관성을 찾아내기 위해 통계적 방법론을 적용하는 것이 좋을 것으로 보임

□ 데이터 분석 내용 및 결과

○ **분석 개요**

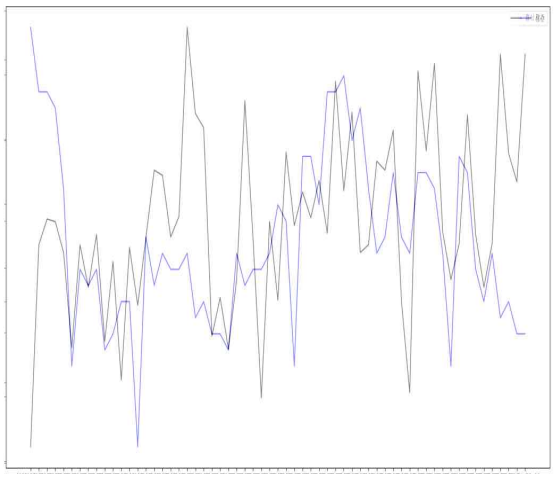
- 고령자의 일상생활에서의 걸음걸이 패턴은 그들의 건강상태와 직접 연결, 특히, 치매 고위험군에서의 걸음 패턴을 분석하여 치매의 위험성을 예측하려는 이 연구의 목적은 초기 진단 및 예방에 중요한 정보를 제공할 수 있는 상황

○ **분석 도구/기법** : 파이썬/로지스틱 회귀분석

○ **로지스틱 회귀분석 해석**

- 로지스틱 회귀분석을 사용하여 걸음걸이 데이터와 치매 위험 사이의 관계를 파악하려 하며, 로지스틱 회귀는 결과 변수가 이진 변수일 때 주로 사용되는 방법, 따라서 이 분석에서는 걸음걸이 패턴을 바탕으로 고령자가 치매 고위험군에 속하는지를 예측

```python
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix

# 데이터 로드
data = pd.read_csv('걸음거리_센서데이터.csv')

# Feature 선택 및 데이터 분리
X = data[['걸음거리', '걸음속도', '걸음리듬']]
y = data['치매_위험군']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# 로지스틱 회귀 모델 생성
model = LogisticRegression()
model.fit(X_train, y_train)

# 예측 및 결과 보고
y_pred = model.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```



<그림 5> 개발 코드와 로지스틱 회귀분석

○ **산식**

$$p(X) = \frac{e^{(\beta_0+\beta_1 X1+\beta_2 X2+\ldots)}}{1+e^{(\beta_0+\beta_1 X1+\beta_2 X2+\ldots)}}$$

- p(X)는 치매 위험군에 속할 확률, X1,X2,…는 걸음거리, 걸음 속도, 걸음 리듬 등의 독립 변수

○ **분석결과**

- 분석결과, 특정 걸음 패턴을 보이는 고령자들은 치매의 위험이 크다는 것을 확인할 수 있으며, 예를 들면, 일정한 리듬으로 걷지 않거나 특정 시간대에 걸음이 불규칙한 경우 등이 이에 해당

○ **향후 연구 방향**

- 다양한 걸음 패턴 변수 추가 (예: 걸음 간격, 걸음 강도)
- 다른 머신러닝 알고리즘과의 성능 비교
- 걸음걸이뿐만 아니라 다른 생활 패턴 데이터를 포함한 통합 분석 수행

5. [기초과학연구원]

○ **데이터 분석 내용**

기초연구 분야 연구 전략 수립 의사결정 지원	
✓ 지금까지는	✓ 앞으로는
□ **논문 현황 내역 제공** ㅇ SCI, HCP(피인용 상위 1%), NSC 논문 수 등 논문데이터 제공	□ **분야별 논문 현황 내역 제공** ㅇ논문 저자 소속, 연구 분야 등 데이터를 활용하여 논문별 분야를 구분하여 논문 현황 제공

✓ 분석 방법은
□ **논문 저자 소속, 연구 분야 등 데이터를 결합**하여 연구 전략 수립 시 의사결정 지원

<표 7> 연구 전략 수립 의사결정 지원

○ **컨설팅 분석결과**

- **(주제)** IBS(기초과학연구원)의 연구 성과 데이터를 분석하여 기초 연구 분야에 관한 연구 전략 수립 시 의사결정을 지원하는 것을 목적으로 하는 것은 데이터 기반 의사결정이 점차 중요해지고 있는 현재 상황에서 매우 적절한 주제로 보임
- **(데이터 활용 업무 계획)** 개인정보 보호와 관련하여 적절한 비식별 조치가 필요할 것으로 보이며, 분야별 논문 현황 등 세부적인 분석을 위해서는 충분한 양의 정밀하고 정확한 메타데이터가 필요할 것으로 판단됨
- 다양한 종류의 연구 성과 정보를 통합적으로 관리하고 분석하기 위해 체계적인 DB 구축 및 관리방안이 필요할 것을 보임
- **(데이터 활용 과제)** 현재까지 제공되던 단순한 논문 현황 제공에서 한 단계 나아가, 분야별 논문 현황 및 트렌드를 파악하고 이를 바탕으로 한 전략적인 의사결정을 지원하는 방향으로 발전시키려는 점에서 매우 적절하며, 이런 방식의 분석은 각 기관이나 개인 연구자들이 자신들의 성과를 보다 정확하게 파악하고 평가받을 기회를 제공함으로써 학계 내 경쟁력 강화에도 도움이 될 것으로 보임
- 다양한 관련 분야 전문가들과 협력하여 더욱 폭넓고 심도 있는 분석결과를 도출하고, 결과 해석 및 활용방안에 대한 충분한 고려와 계획을 수립하는 것이 중요할 것으로 판단됨
- **(활용 데이터)** 본 과제에서 사용할 예정인 다양한 데이터(논문 저자 소속, 연구 분야 등)는 풍부한 분석 자료로 활용될 수 있을 것으로 보이며, 보안 및 개인정보와 관련한 이슈를 철저히 관리하면서 데이터를 활용해야 할 것으로 보임

□ 데이터 분석 내용 및 결과

○ **분석 개요**

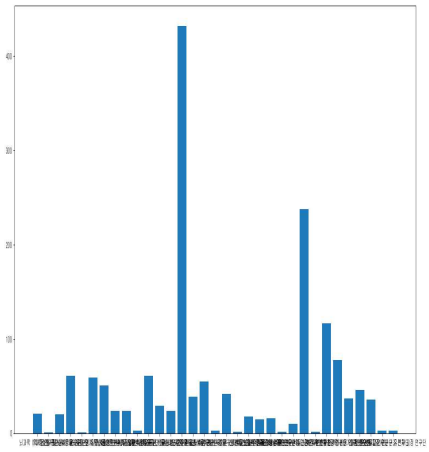
- 기초과학연구원의 기초연구 분야 연구 전략 수립 의사결정을 지원하기 위하여, 기초과학연구원의 지식재산권 정보를 텍스트 마이닝 하여 연구 성과와 관련된 주요 키워드와 트렌드를 파악

○ **분석 도구/기법** : 파이썬/텍스트 마이닝

○ **분석 처리 순서**

- 파이썬에서 데이터를 벡터화 처리하고 CSV 형태로 저장
- Power BI 도구에서 해당 CSV 파일을 불러와 시각화 진행

```python
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer

# 데이터 로딩
data = pd.read_csv('기초과학연구원_지적재산권정보.csv')

# 텍스트 마이닝을 위한 벡터화 진행
vectorizer = CountVectorizer(max_features=1000, stop_words='english')
X = vectorizer.fit_transform(data['발명의 명칭'])

# 주요 키워드 파악
sum_words = X.sum(axis=0)
words_freq = [(word, sum_words[0, idx]) for word, idx in vectorizer.vocabulary_.ite
words_freq = sorted(words_freq, key=lambda x: x[1], reverse=True)
top_keywords = words_freq[:20]

print(top_keywords)

# 결과 저장
df_keywords = pd.DataFrame(top_keywords, columns=['Keyword', 'Frequency'])
df_keywords.to_csv('top_keywords.csv', index=False)
```



<그림 6> 개발 코드와 텍스트 마이닝

○ **분석결과**

- 텍스트 마이닝을 통해 연구 성과와 관련된 주요 키워드(예: "나노기술", "양자역학", "유전체학" 등)가 발견될 것으로 보이며, 최근 연도별로 키워드 트렌드를 파악했을 때, "XXX"와 "YYY"가 빠른 속도로 증가하고 있는 것으로 관측되는 시각효과 구현

○ **향후 연구 방향**

- (심층 텍스트 마이닝) 현재는 단순 빈도 기반의 키워드 추출을 진행하였지만, TF-IDF나 word2vec 등의 고급 방법론을 활용한 텍스트 마이닝을 진행할 필요가 있다고 판단됨
- (키워드 연계분석) 키워드 간의 관계나 연계성을 파악하여 연구 주제나 트렌드 간의 연관성을 파악하는 방안을 모색

6. [한국항공우주연구원]

6-1. AI 활용 스마트 활용기술 개발

○ **데이터 분석 내용**

AI 활용 스마트 활용기술 개발	
✓ 지금까지는	✓ 앞으로는
□ **육안으로 위성영상 판독** ㅇ 일정 시간 내 육안으로 판독되는 위성영상의 수는 제한적 ㅇ 앞으로 위성영상의 종류와 수가 지속해서 증가할 전망으로 현재와 같은 방법은 비용과 생산성 측면에서 적절하지 않음	□ **인공지능이 위성영상 판독** ㅇ 신속한 위성영상 판독 ㅇ 여러 종류의 위성영상을 종합적으로 판독하여 기존에 사람이 찾기 어려웠던 정보 또한 탐색

✓ 분석 방법은
□ 국가가 보유한 위성영상에 대해 AI 학습데이터를 구축하고, 이를 활용하여 AI 위성영상 분석 알고리즘 개발, 위성영상 활용 서비스플랫폼에 탑재하여 대국민 서비스 제공

<표 8> AI 스마트 활용기술 개발

○ **컨설팅 분석결과**

- **(주제)** 위성영상 분석에 AI와 빅데이터 기술을 도입하는 것은 매우 효율적인 접근 방식이 될 수 있으나, 다양한 고려사항이 존재할 것으로 보임
- AI 모델의 성능은 학습데이터의 질에 크게 의존하므로, 고질적인 위성영상 데이터가 필요할 것으로 보이며, 지도학습을 사용한다면, 영상에 대한 라벨링 작업이 필요하며, 이 작업은 많은 시간과 비용이 소모될 것으로 보임
- 또한, AI 기술의 지속적 갱신과 업데이트, 위성영상 데이터 내 개인정보 및 민감정보에 대한 고려가 필요할 것으로 보임
- **(데이터 활용 업무 계획)** 22년부터 분석 알고리즘 개발환경, AI 학습데이터, AI 알고리즘 개발 및 공개 추진을 위한 사업을 진행 중이며, 해당 사업의 디테일한 내용을 확인한다면 추가적으로 AI 학습데이터로 적절한지 판단이 가능할 것으로 파악
- **(데이터 활용 과제)** 육안으로 위성영상을 판독하던 기존의 방식에서 인공지능을 활용한 방식으로 전환할 때 필요한 분석 방법은 다양하며, 다양한 데이터 분석 방법들을 통해 AI 기반의 위성영상 판독 시스템의 성능과 효율성을 극대화하고, 지속해서 개선할 수 있는 방향을 도출할 수 있도록 설계가 필요
- **(활용 데이터)** 제시한 데이터는 양, 다양성, 형식, 주석, 접근성 등 다양한 측면에서 빅데이터 분석 및 AI 학습에 매우 적합. 다만, 보안 및 개인정보와 관련한 이슈를 철저히 관리하면서

데이터를 활용해야 할 것으로 보임

□ 데이터 분석 내용 및 결과

○ **분석 개요**

- 한국항공우주연구원에서 AI 활용 스마트 활용기술의 지역별 변화를 파악하기 위해 국토교통부 국토지리정보원의 항공사진 위성영상 구축현황 데이터를 활용하여 분석을 진행

○ **분석 도구/기법** : 파이썬/회귀분석, 시계열 분석

○ **분석결과**

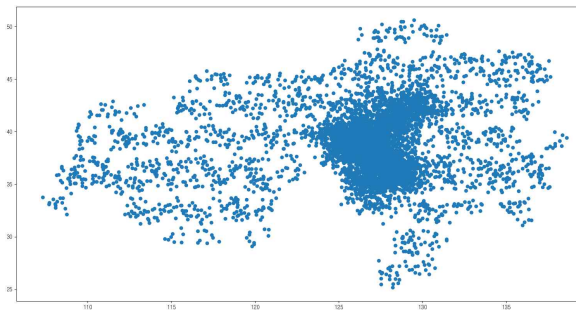
- (회귀 분석) 위도와 태양고도에 따른 AI 활용기술의 평가 결과에 대한 회귀모델을 수립, 결과로 태양고도와 평가 결과 사이에는 유의미한 관계가 발견

```python
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# 데이터 로딩
data = pd.read_csv('항공사진_위성영상_구축현황.csv')

# 필요한 특성 추출
X = data[['태양고도', '위도']]
y = data['평가 결과']  # 예시로 평가 결과라는 컬럼명을 사용함. 실제 데이터의 컬럼명으로 변경 필요

# 데이터 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 회귀 모델 학습
reg = LinearRegression().fit(X_train, y_train)

# 예측 및 평가
y_pred = reg.predict(X_test)
mse = mean_squared_error(y_test, y_pred)

print(f"Coefficients: {reg.coef_}")
print(f"Intercept: {reg.intercept_}")
print(f"Mean Squared Error: {mse}")

```

<그림 7> 파이썬 개발 코드

- (시계열 분석) ARIMA 모델은 시계열 데이터의 자기 상관성을 활용하여 예측하는 방법, ARIMA 모델은 AR(AutoRegressive), I(Integrated), MA(Moving Average) 세 부분으로 구성
- AR은 자기 자신의 이전 관측값에 의존하는 것, MA는 이전 항목의 오차에 의존하는 것을 의미하며, I는 데이터를 정상적인 상태로 만드는 데 사용

```python
# ... [기존 코드]

# 시계열 분석
from statsmodels.tsa.arima.model import ARIMA

# 적절한 차분을 위한 데이터 전처리 (예시로 1차 차분을 진행)
diff_data = y.diff().dropna()

# ARIMA 모델 학습
model = ARIMA(diff_data, order=(1,1,1))  # p,d,q 파라미터는 실제 데이터에 따라 조절
result = model.fit()

# 예측 (다음 5개 데이터 포인트 예측)
forecast = result.forecast(steps=5)
print(forecast)

# ... [기존 코드 계속]
```



<그림 8> 파이썬 개발 코드와 회귀 분석 시계열 분석

- p, d, q 파라미터를 데이터에 적합하게 조절할 필요가 있으며, ACF(AutoCorrelation Function) 및 PACF(Partial AutoCorrelation Function) 그래프를 활용하여 p, d, q 값을 설정
  - 추가로, ARIMA 외에도 Prophet, LSTM, RNN 등 다양한 시계열 분석 방법이 있으니 연구의 방향과 목적에 따라 적절한 모델을 선택하는 것이 중요

○ **향후 연구 방향**

- (다변량 회귀 분석) 현재는 위도와 태양고도만을 사용하여 회귀 분석을 진행하였지만, 다른 변수들과의 상호작용 효과를 고려한 다변량 회귀 분석을 진행할 필요가 있음
- (기타 알고리즘 활용) 회귀 분석 외에도 기계 학습 알고리즘을 활용하여 더 정확한 예측 모델을 수립하는 방안 탐색

7. [나로우주센터]

○ **데이터 분석 내용**

홈페이지 방문객 분석을 통한 콘텐츠 다양화 및 연계프로그램 제공	
✓ 지금까지는	✓ 앞으로는
ㅇ 방문자와의 상호작용 및 개별적인 피드백 수집이 부족하며, 이로 인해 콘텐츠 개선 및 연계프로그램의 효율성 향상에 한계 존재	□ **홈페이지 방문객 설문 조사를 통한 연계 콘텐츠 제공** ㅇ 방문객 설문 조사 방법 및 데이터 분류를 통한 양질의 데이터 추출 방법 구축
✓ 분석 방법은	
□ **우주과학관 홈페이지 방문객 분석을 통한 콘텐츠 다양화를 위해 컨설팅 내용을 기반으로 데이터 분석 방법 고려**	

<표 9> 콘텐츠 다양화 및 연계프로그램 제공

○ **컨설팅 결과**

- **(주제)** 우주과학관 홈페이지 방문객으로 얻을 수 있는 인사이트와 그 가치에 대한 사전 분석 필요
- **(데이터 활용 업무 계획)** 고흥 우주 항공**축제** 센터 견학자 사전 조사
- 고흥 외 순천, 여수 관광 및 교육 프로그램 연계 방안 등 외 대외적 활동에 부가적으로 쓰일 활용처 확인 필요
- **(데이터 활용 과제)** 설문 조**사** 시 수집하는 정보 중에 개인정보(이름, 연락처 등)가 포함되어 있다면 GDPR, CCPA 및 기타 관련 국내 법률에 따라 적절한 보호 및 관리 조처를 해야 할 것으로 보임
- 표본의 대표성을 위해 설문 조사의 결과를 일반화하려면 응답자 표본이 전체 방문객을 대표해야 하며, 설문 참여를 독려하는 방법을 고려해야 할 수 있어야 함
- 설문 설계 시 질문의 명확성을 위해 설문지의 질문이 모호하지 않고, 중복되는 내용이 없어야 하며 주관적 해석을 최소화하고, 객관적인 답변을 얻을 수 있는 질문을 선호해야 함
- 설문 조사 결과를 바탕으로 콘텐츠 제공 방안을 도출할 때, 데이터의 한계와 잠재적인 편향을 인식하며 결정을 내려야 함
- 효과적인 콘텐츠 제공을 위해 설문 결과뿐만 아니라 방문객의 행동 데이터나 다른 데이터 소스와의 연계를 고려
- 마지막으로 연계 콘텐츠 제공 후, 방문객의 피드백을 수집하여 지속해서 개선해 나갈 수 있는 시스템 구축 필요
- **(활용 데이터)** 상세 데이터 수량 등 확인 필요

□ 데이터 분석 내용 및 결과

○ **분석 개요**

- 본 연구에서는 웹사이트 방문자의 접속 자세와 그들의 머문 시간(duration) 간의 상관관계를 분석. 주요 독립 변수로는 '접속 자세'를, 종속 변수로는 '머문 시간'을 설정

○ **분석 목표**

- 우주과학관 홈페이지 방문자의 행동 패턴을 파악하여, 모바일 및 데스크톱 환경에서의 이용 트렌드를 식별하고, 맞춤형 콘텐츠 전략을 개발하여 방문 활성화

○ **분석 도구/기법** : Excel/상관분석, 회귀 분석

○ **상관계수 해석**

- 분석결과, 접속 자세와 머문 시간 간의 상관계수는 -0.7848로 나타나며, 이는 두 변수 사이에 강한 음의 상관관계가 있다는 것을 의미 일반적으로 상관계수 값이 0에 가까울수록 상관관계가 약하며, 1 혹은 -1 에 가까울수록 강한 상관관계를 가짐

○ **회귀 분석 결과**

- 조정된 결정 계수는 50%를 초과하였으므로, 모델의 예측력이 높다고 판단할 수 있음, F-통계량의 유의 확률값은 2.13322509916089E-137로, 매우 유의미. 그러나 이 값은 '접속 자세'이 '머문 시간'을 예측하는 데에는 충분하지는 않음
- 또한, 접속 자세 코드는 다음과 같이 구분: PC: 1 / Android: 2 / iOS: 3
- 이에 따라, 모바일 환경에서는 평균적으로 510초, 윈도우 및 기타 환경에서는 3080초 동안 웹사이트에 머무는 것으로 파악

○ **분석결과**

- 상관분석 및 회귀 분석 결과를 고려할 때, 모바일 환경에서 이용안내 page의 분산 처리 서버 구현 및 확장 가능한 인프라의 구축이 필요할 것으로 보이며, 윈도우 및 다른 운영 환경에서는 다음 세션의 동향을 예측하여 시각적 디자인을 개선하는 전략이 필요할 것으로 판단됨

○ **향후 연구 방향**

- 다음 단계에서는 '접속 자세 코드'와 'Referrals' 간의 상관관계를 추가로 탐색할 예정 또한, 플랫폼별 이용 통계, 웹 추세 분석, 고밀도 접속 페이지, 접속경로 및 세션별 방문자 행동에 대한 상세한 분석을 진행할 예정, 이로써 맞춤형 콘텐츠 제공 및 이용자 활성화 가능

	접속품코드	DURATION(머문시간)
접속품코드	1	
DURATION(머문시간)	-0.78480576	1

요약 출력

회귀분석 통계량	
다중 상관계수	0.784805762
결정계수	0.615920085
조정된 결정계수	0.6153301
표준 오차	0.358931269
관측수	653

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	134.4950639	134.4951	1043.96	2.1E-137
잔차	651	83.86940775	0.128832		
계	652	218.3644717			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	2.137574792	0.029562813	72.3062	0	2.079525	2.195625	2.079525	2.195625
DURATION(머문시간)	-0.01942784	0.000601288	-32.3104	2.1E-137	-0.02061	-0.01825	-0.02061	-0.01825

<그림 9> 향후 연구 방향

□ 분석결과의 정책(업무) 활용실적 및 성과

○ 데이터 기반 행정과 관련된 '24년도 과기정통부의 데이터 분야 사업 기획 및 현재 수행 중인 데이터 활용 업무 개선에 활용

○ 데이터 인프라 구축, 데이터 분석과제 기획 및 제안요청서 작성 등 업무에 활용

○ 또한, 정부 데이터 기반 행정 활성화 실태조사 등 정부 정책 기조에 부응 가능

8. [한국철도기술연구원]

□ **데이터 분석 내용**

철도 분야 무역데이터 분석	
✓ 지금까지는	✓ 앞으로는
□ **철도 분야 무역데이터 분석** - UN Comtrade 및 관세청에서 제공하는 무역데이터에서 철도 관련 수치만 엑셀로 반출 - 반출한 액셀의 HS코드 항목별로 수치를 재조합하여 도표화 및 시각화 - 반출한 액셀을 매트리스 형태로 시트를 변환하여 넷마이너를 이용하여 중심성 분석 등 수행 - 엑셀 하나에 각 국가의 수치가 모두 포함되어 있으나 매번 수작업으로 정리를 해서 리포팅을 하는 상황	□ **철도 분야 무역데이터 분석** - 매년 동일한 업무를 반복하고 있는데 효율적으로 분석할 방안 수립 - 본 분석에서는 수출데이터를 활용하여 국가가 무역 네트워크를 분석하고, 품목별로 수출일 현황을 분석하고 있는 상황, 이 외에 더욱 효율적인 방안 탐색 및 수립
□ **연구 성과 분석** - 연구원에서 외부로 발주되는 각종 보고서(본과제보고서, 위탁보고서, 구매용역보고서, 장기자문보고서)를 보유만 하는 상황	□ **연구 성과 분석** - 연구데이터 분석을 통해 중복연구를 방지하거나 유사한 연구가 이중으로 발주되는 상황을 방지

✓ 분석 방법은
□ **제시되지 않음 / 실제 데이터를 직접 다운 받아 분석 예정**

<표 10> 철도 분야 무역데이터 분석

○ **컨설팅 결과**

- **(주제)** 철도 분야 무역데이터 분석은 매년 수작업으로 이루어지고 있고, 연구데이터는 보유만 하고있는 상황으로 데이터의 보다 효율적인 활용을 위한 컨설팅이 필요할 것으로 판단됨
- **(데이터 활용 업무 계획)** 철도 분**야** 무역데이터 및 연구데이터 분석으로, 현재 단순 수작업 과정을 자동화하고 보다 효율적이고 유용한 데이터 분석 방안을 제시해주어야 하는 상황으로 판단됨
- **(데이터 활용 과제)** 넷마이너를 이용해서 중심성 분석 외 네트워크 시각화, 클러스터링 및 군집 분석, 네트워크 다양성 분석, 전이 모델링, 사회 네트워크 분석, 시간에 따른 변화 분석, 예측 모델링을 통해서도 네트워크 분석 가능
- 정제된 HS코드를 통해 무역 패턴 이해, 품목별 무역 증감 추이, 무역 국가 간 비교, 수출 다양성 분석, 섹터 및 산업 분석, 경쟁자 분석, 수출 전략 도표화 및 시각화 가능할 것으로 보임
- 반복 업무를 효율적으로 처리하려면 HS코드 항목 수치 재조합 자동화, 표준화, 템플릿화

및 추가적인 독립 변수가 생기지 않도록 지속적인 개선, 관리가 필요할 것으로 보임

- 엑셀 반출(수집) 과정에서 크롤링, 쿼리 자동화 과정 도입 검토 필요
- 경제 영향 분석, 국가별 무역 파트너 탐색, 시장 진입 전략 평가, 거래 중요성 및 지표 개발, 전략적 무역정책 분석, 시각화 및 대시보드 개발, 사회적 영향 분석 등 접근 방식을 통해 수출데이터를 다양한 관점에서 활용하여 경제, 무역, 정책 및 사회적 측면에서 더 깊이 있는 분석을 수행하는 데 도움을 줄 것으로 보임
- **(활용 데이터)** UN Comtrade, 관세청 연계 프로그램 데이터 확인 필요

□ 데이터 분석 내용 및 결과

○ **분석 개요**

- 기존 수작업으로 철도 관련 무역데이터를 분석하는 상황, 수작업 프로세스 간소화 및 자동화 방법 탐색, 품목별 수출입 현황분석 방법 탐색
- 연구데이터를 활용하여 중복연구 탐색 및 중복연구의 주요 키워드 추출

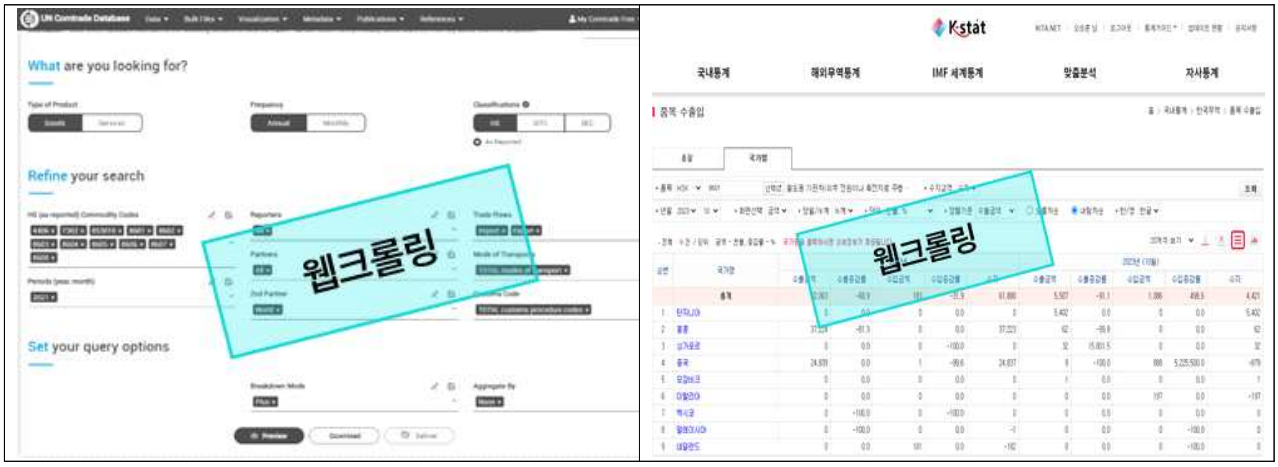
○ **분석 목표**

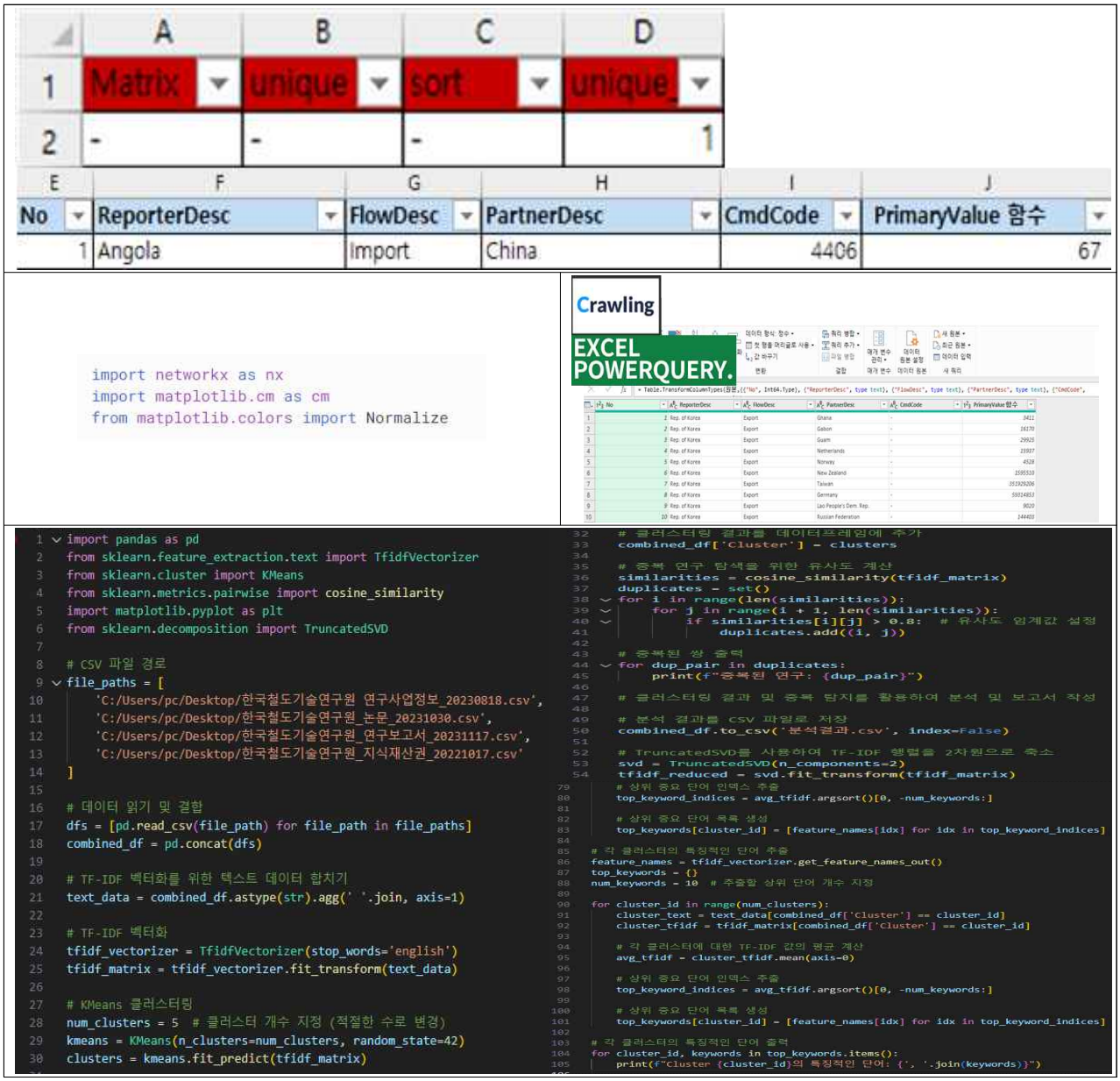
- 철도 관련 무역데이터 크롤링 및 쿼리 자동화, 더욱 효율적인 품목별 수출입 현황분석 방법론 탐색
- 연구데이터를 활용하여 중복연구 탐색을 위한 유사도를 계산하고 유사연구 및 중복된 연구의 개수 확인. 또한, 클러스터링하여 각 클러스터의 특징적인 단어 추출 및 시각화

○ **분석 도구/기법** : Excel, python/엑셀, 텍스트마이닝, 클러스터링

○ **분석 과정**

- UN Comtrade, K-stat(한국무역협회) 웹사이트에 접속하여, 철도 관련 품목 무역데이터를 직접 다운로드 후, 엑셀 호출 및 쿼리 자동 변환
- python에서 클러스터링 및 시각화에 필요한 라이브러리 호출, 데이터 로드, TF-IDF 벡터화를 위한 텍스트 데이터 결합 후 K-means 클러스터링 수행. 이후, 중복연구 탐색을 위한 유사도를 설정하여 유사연구 개수 출력, 분석결과를 csv 파일로 저장하고 유사한 연구 군집별 주요 단어 추출 및 시각화 수행

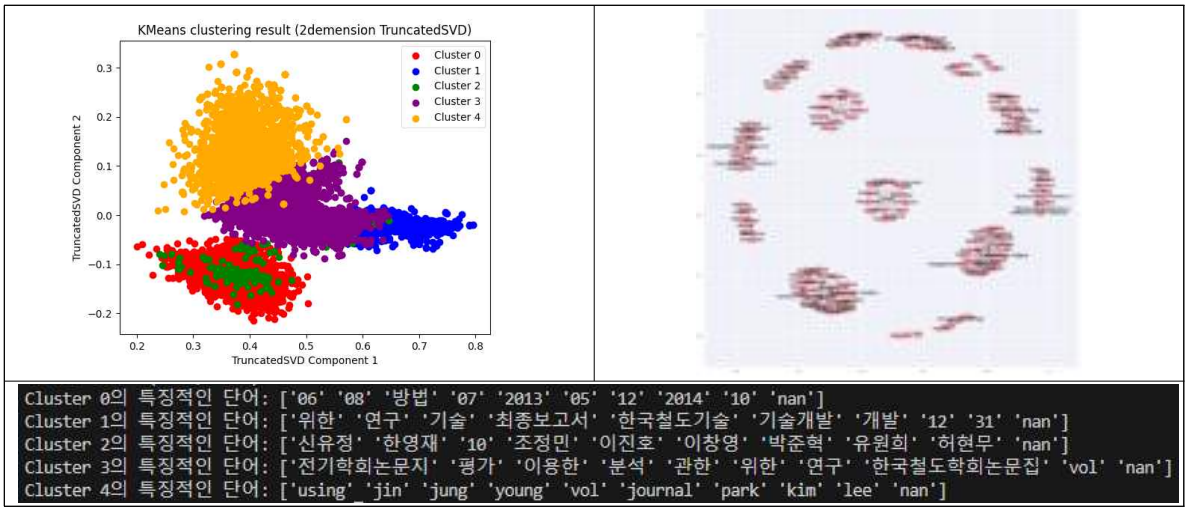




<그림 10> 파이썬 엑셀 텍스트마이닝 클러스터링

○ **분석 결과**

- 엑셀 관련 분석결과 넣고
- 유사도 임계값을 0.8(1에 가까울수록 유사함)로 설정해 유사연구 개수를 출력한 결과 10,762개로 도출, 10,762개의 연구를 클러스터링하여 5개의 군집으로 분할 후 시각화, 각 군집별 특징 단어를 추출하여 주요 키워드 파악
- 정제된 HS코드를 통해 무역 패턴 이해, 품목별 무역 증감 추이, 무역 국가 간 비교, 수출 다양성 분석, 섹터 및 산업 분석, 경쟁자 분석, 수출 전략 도표화 및 시각화 가능할 것으로 보임
- 더 나아가 추출과정에서도 크롤링 & 쿼리변환을 통해 자동분석 보고서 가능



<그림 11> 엑셀 관련 분석결과

○ **향후 연구 방향**

- 철도 관련 무역데이터에서 품목별 수출입 현황분석 이후, 국가별 수출액 변화에 가장 큰 영향을 미치는 요인을 탐색, 가중치, 리스크 레벨을 고려하여 상관분석 수행 후 미래의 수출액 패턴을 예측하는 등의 방법으로 연구 가능
- 중앙 데이터베이스 구축
- 디지털 아카이브화 : 국가 간 네트워크 분석, 철도 분야 무역데이터 등 디지털 형태로 저장하며, 메타데이터를 함께 저장, 검색 및 모니터링 시스템을 도입 가능할 것으로 보임
- 한국철도기술연구원의 철도 관련 무역데이터 분석 작업의 효율화, 자동화 등 과정을 통해 현재 가지고 있는 데이터를 행정 처리 과정에 최대한 활용

9. [한국과학기술연구원]

□ 데이터 분석 내용

연구 장비 자산관리 및 장비 예약 활용관리 고도화	
✓ 지금까지는	✓ 앞으로는
□ **통합시스템의 자산관리 사용** ㅇ 사용자별 자산 현황과 고가 연구 장비 현황으로 분류 ㅇ 자산의 현재 상태, 활용도, 운영자 등에 대한 정보 미흡 ㅇ 특성분석데이터센터의 분석의뢰 장비만 통계분석 가능	□ **자산관리시스템 고도화** ㅇ 연구 장비의 활용도, 공동사용내역 등 통계 자료 제시 ㅇ 장비사용 과제, 관련 장비의 관련된 논문 게재 여부 자동 파악을 통한 장비 중요성 부각
□ **단순분석의뢰** ㅇ 장비의 선택과 사용 가능한 날짜를 통해 의뢰 ㅇ 장비의 응용 분야 및 사용방법을 모르고 분석의뢰	□ **프로토콜 기반 분석의뢰** ㅇ 장비를 선택하고 원하는 분석을 프로토콜 검색을 통하여 분석 의뢰함. 표준화된 분석제공과 데이터 간의 연계가 이루어져 향후 활용성 측면에서 유용함 ㅇ 장비별 분석뿐 아니라, 장비 그룹별 연계 분석을 통하여 통합 분석이 가능하도록 제공함
✓ 분석 방법은	
□ **제시되지 않음**	

<표 11> 자산관리 장비 예약 활용관리 고도화

○ **컨설팅 결과**

- **(주제)** 연구 장**비** 자산상태관리 및 장비 예약 활용관리 고도화를 위해 데이터 기반의 관리체계를 구축하여 장비 활용을 최적화, 이를 통해 효율적인 연구 활동 지원이 가능하다고 판단됨
- **(데이터 활용 업무 계획)** 연구 장비 자산관리 시스템과 연구 장비 예약 활용관리 시스템을 고도화하고, 현재의 자산 및 활용 데이터를 체계적으로 수집, 분석하는 것에 중점

○ **(데이터 활용 과제)**

- 연구 장비 자산관리 시스템 고도화: 자산의 현재 상태와 활용도에 대한 정보 부족 문제를 해결하기 위해 연구 장비의 활용도와 공동사용 내역을 통계적으로 제공하는 것은 매우 유용하다고 판단 또한, 장비사용 과제와 관련된 논문게재 여부를 자동으로 파악하여

장비의 중요성을 강조하는 것은 향후 장비 관리에 중요한 정보를 제공할 수 있을 것으로 보임

- 장비 예약 활용관리 고도화: 프로토콜 기반 분석의뢰를 통해 장비를 선택하고 표준화된 분석과 데이터 간의 연계를 제공하는 것은 향후 활용성 측면에서 유용할 것으로 보임. 장비별 및 장비 그룹별 연계분석을 통해 통합 분석이 가능해지면, 보다 종합적인 데이터 분석이 가능해질 것으로 보이며, 이러한 개선사항들을 통해 연구 장비의 효율적인 관리와 활용을 높일 수 있을 것으로 판단
- **(활용 데이터)** 개인정보를 불포함한 데이터로 확인되며, 실제 데이터의 명세를 확인할 필요가 있다고 판단됨

□ 데이터 분석 내용 및 결과

○ 분석 개요

- 연구 장비 현황 데이터를 활용해 단독/공동사용, 활용/불용, 단독활용 주요 사유 퍼센트 확인 및 단어 빈도 기준 상위 10개 단어를 추출하여 장비 관련 유용한 정보 추출

○ 분석 목표

- 장비 데이터 분석을 통해 자원 할당 및 관리개선, 장비 활용 최적화, 단독활용 주요 사유 파악, 트렌드 및 우려 사항 도출 등에 활용 가능

○ 분석 도구/기법 : python/통계분석, 텍스트 마이닝

○ 분석 과정

- python에서 필요한 라이브러리 및 데이터 호출, 결측치 처리, 활용범위 및 장비 상태 컬럼에서 각 값의 퍼센티지 계산, 단독활용 사유 컬럼에서 주요 키워드 상위 10개 단어를 추출하고 데이터별 퍼센티지 계산 후 시각화 수행
- (STEP 1) 분석 도구 선정 및 데이터로드
- (STEP 2) 컬럼 선택 및 데이터 전처리
  1) 컬럼 선택(활용범위, 단독활용 사유, 장비 상태)
  2) 텍스트 데이터 전처리(불용어 제외 등)
- (STEP 3) 통계분석
  1) 활용범위 및 장비 상태 값들의 퍼센티지 계산
- (STEP 4) 텍스트 마이닝
  1) 단독활용 사유 값 중 주요 키워드 상위 10개 단어 추출
- (STEP 5) 원그래프 시각화

```python
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from nltk.tokenize import word_tokenize

# 데이터 불러오기 - 인코딩 변경하여 시도 ('utf-8', 'euc-kr', 'latin-1' 등)
data = pd.read_csv('C:/Users/pc/Desktop/2023-11-24_장비정보.csv', encoding='utf-8')
```

```python
# 결측치 처리
data.dropna(subset=['활용범위', '단독활용사유', '장비상태'], inplace=True)
```

```python
# '활용범위' 및 '장비상태'에서 값들의 퍼센티지 계산
utilization_percentage = data['활용범위'].value_counts(normalize=True) * 100
equipment_status_percentage = data['장비상태'].value_counts(normalize=True) * 100

print("활용범위의 퍼센티지:")
print(utilization_percentage)
print("\n장비상태의 퍼센티지:")
print(equipment_status_percentage)
```

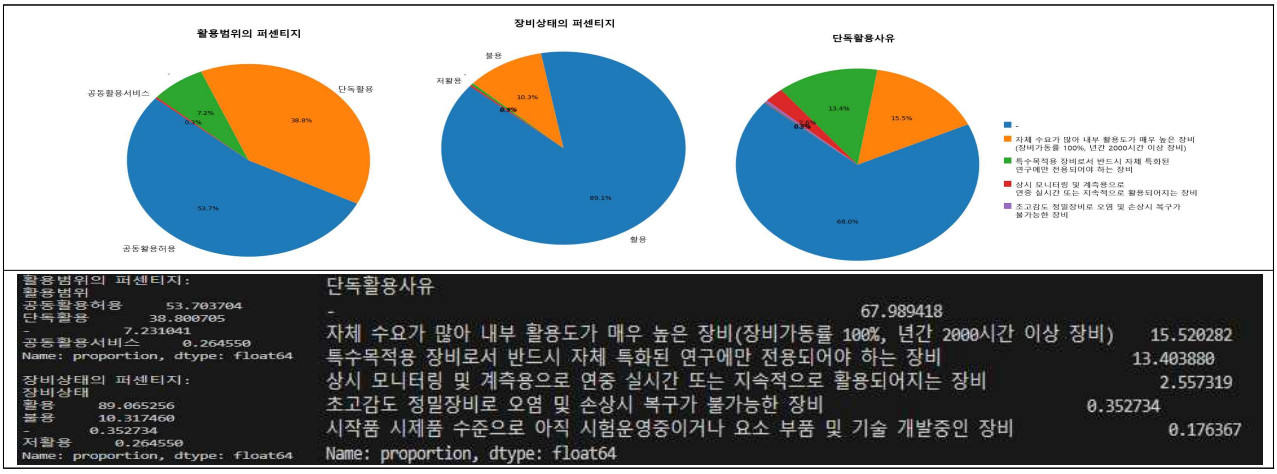
```python
# '단독활용사유'에서 주요 키워드 추출
text = ' '.join(data['단독활용사유'])
word_tokens = word_tokenize(text)

# 단어 빈도 계산
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(word_tokens)
word_freq = pd.DataFrame(X.toarray(), columns=vectorizer.get_feature_names_out())

# 단어 빈도 기준 상위 10개 단어 추출
top_keywords = word_freq.sum().nlargest(10)

print("\n상위 10개 키워드:")
print(top_keywords)
```

```python
# 활용범위 퍼센티지 원그래프
plt.figure(figsize=(8, 8))
plt.pie(utilization_percentage, labels=utilization_percentage.index, autopct='%1.1f%%', startangle=140)
plt.title('활용범위의 퍼센티지')
plt.show()

# 장비상태 퍼센티지 원그래프
plt.figure(figsize=(8, 8))
plt.pie(equipment_status_percentage, labels=equipment_status_percentage.index, autopct='%1.1f%%', startangle=140)
plt.title('장비상태의 퍼센티지')
plt.show()

# 단독활용사유의 데이터별 퍼센티지 원그래프
plt.figure(figsize=(8, 8))
plt.pie(percentage_per_reason, labels=percentage_per_reason.index, autopct='%1.1f%%', startangle=140)
plt.title('단독활용사유의 데이터별 퍼센티지')
plt.show()
```

<그림 12> 파이썬 분석 과정

○ 분석결과

- 통계분석 결과 활용범위 컬럼 중 공동활용허용은 53.7%, 단독활용은 38.8%, 미기재 7.2%, 공동활용서비스 0.26%로 나타남
- 장비 상태 컬럼 중 활용 89%, 불용 10.3%, 미기재 0.3%, 저활용 0.2%로 나타남
- 단독활용 사유 컬럼 중 미기재 67.9%, '자체 수요가 많아 내부 활용도가 매우 높은 장비(장비가동률 100%, 연간 2000시간 이상 장비)' 15.5%, '특수목적용 장비로서 반드시 자체 특화된 연구에만 전용되어야 하는 장비' 13.4%, '상시 모니터링 및 계측용으로 연중 실시간 또는 지속해서 활용되는 장비' 2.5%, '초고감도 정밀장비로 오염 및 손상 시 복구가 불가능한 장비' 0.35% 외 기타로 나타남

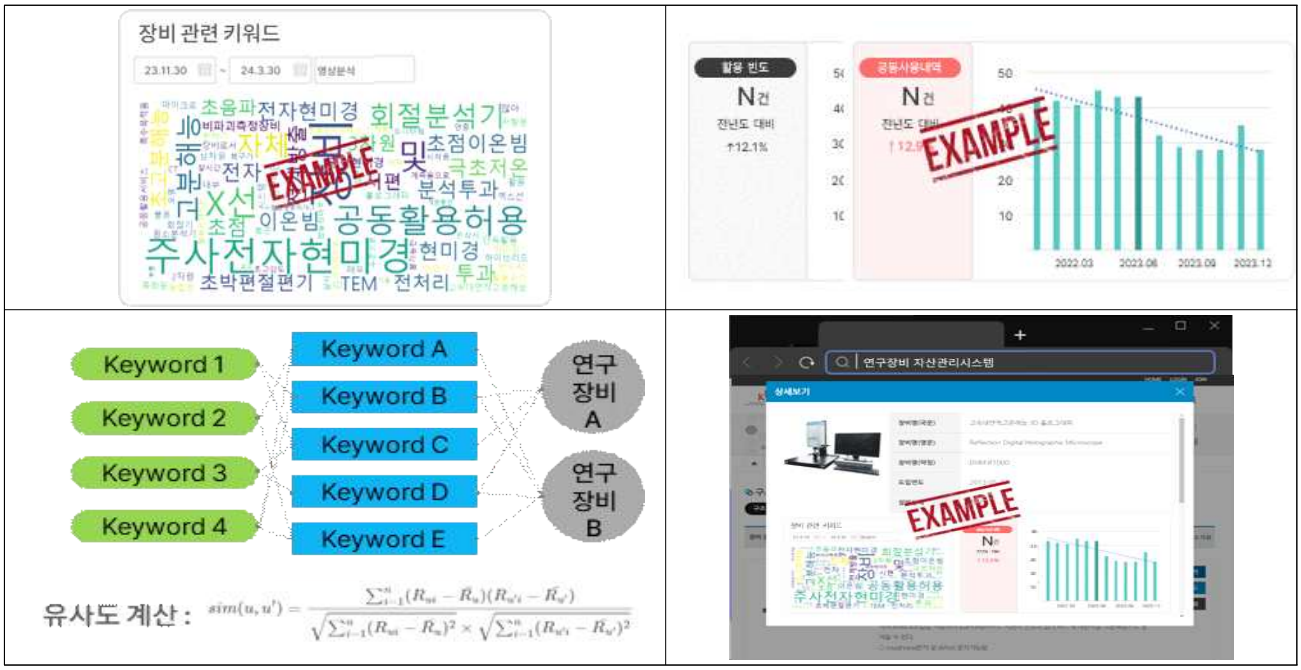


< 그림 13> 통계분석결과

- 단독활용 사유 컬럼과 같이 미기재 및 오기재의 비중이 과반수를 차지하거나 활용도가 낮은 특수한 장비의 경우 분석결과 도출이 제한적이기에 텍스트 마이닝을 통해 '2000시간', '내부', '연간' 등 유의미한 키워드 추출
- 데이터 결과를 통해 연구 장비의 활용범위 및 현재 상태와 공동사용명세 논외인 장비에 대해 파악할 수 있기에 자산관리시스템 고도화 구현의 기반 데이터로 사용할 수 있을 것으로 판단됨. 또한 해당 데이터의 결과를 표출하는 대시보드를 시스템 내 탑재하여 활용 장비 모니터링을 통해 장비 유지보수 관리 및 활용도 제고 방안 모색 가능

○ **향후 연구 방향**

- 위 분석결과에 이어, 향후 관련 장비 데이터와 NTIS 내 검색 결과를 매칭시켜 논문게재 여부 자동 파악을 통한 장비 중요성 분석 수행 가능
- (장비별 연관 키워드 시각화) TF_IDF, Word2vec/Doc2vec으로 키워드 빈도 및 유사도 분석 장비별 특정 키워드 가중치 설정 연관 키워드 결과를 시스템 내 구현
- (통계분석 기반 활용빈도 및 사용내역 시각화) 활용범위 및 장비 퍼센티지 분석결과를 포함하여 장비 관련 데이터 통계분석 활용빈도 및 사용내역 등 분석결과 통계 자료에 대해 차트 형태로 표출
- (자산관리 내 연관 장비 추천 서비스) 수집한 키워드에 대해 협업 필터링을 활용하여 연관 장비를 매칭하고 이를 기반으로 추천



<그림 14> 향후 연구 방향 시각화

10. [한국연구재단]

□ **데이터 분석 내용**

국내 학술 활동, 학술지 인용, 학술단체 인용 현황 등에 대한 분석

✓ 지금까지는	✓ 앞으로는
□ **국내 학술논문, 학술지, 학술단체 등에 대한 기초 통계 자료 제공**	□ **국내 학술논문, 학술지, 학술단체 등에 대한 심층 통계분석 자료제공 및 국내 학술정책 수립을 위한 시사점 도출**

✓ 분석 방법은
□ **분야별, 기관별, 연도별 등 인용 현황, 학술지 및 학술단체 현황분석**

<표 12> 국내 학술 활동 인용 현황 대한 분석

○ **컨설팅 결과**

- **(주제)** 국내 학술논문, 학술지, 학술단체 등에 대한 심층 통계 자료 제공 및 분석으로 국내 학술정책 수립을 위한 시사점 도출
- **(데이터 활용 업무 계획)** 한국학술지인용색인(KCI)**에서** 활용 데이터를 DB화하여 논문 간 인용 관계 분석에 활용 중이며, KCI DB 구축 사업의 데이터를 확인하여 심화 분석이 가능할 것으로 판단됨
- **(데이터 활용 과제)** 필요시 논문, 학술지, 학술단체 데이터로부터 추출한 데이터는 개인정보, 민감정보 비식별화 등의 처리 과정을 거쳐 활용되어야 할 것으로 보이며, 데이터 시각화 분석기법을 사용하여 데이터를 탐색하고 학술 관련 인사이트 도출이 필요할 것으로 판단됨
- 모든 카테고리의 데이터를 일관된 형식으로 정제하고 가공하기 위해 데이터 표준화 기준을 수립하여, 데이터 일관성 유지가 필요할 것으로 보임
- **(활용 데이터)** 논문, 학술지, 학술단체, 데이터는 상시 발생하며 데이터양 또한 충분하여 데이터를 활용하는 데 있어 큰 무리가 없을 것으로 파악됨

□ **데이터 분석 내용 및 결과**

○ **분석 개요**

- 한국연구재단 AI 관련 논문 정보 데이터를 활용하여 텍스트 마이닝 및 클러스터링으로 데이터의 상위 빈도 단어 계산 및 출력. 또한, 주요 단어를 추출해 상위 10개 단어의 인덱스 추출 및 시각화

○ **분석 목표**

- 단어 빈도 분석 및 토픽 모델링, 클러스터링을 활용하여 주요 키워드 분석을 통한 연구 주제의 변화 및 학술적 트렌드 파악

○ **분석 도구/기법** : python/텍스트 마이닝, 클러스터링

○ **분석 과정**

- (STEP 1) 분석 도구 선정 및 데이터 로드
- (STEP 2) 컬럼 선택 및 데이터 전처리
- (STEP 3) 텍스트 마이닝(토픽 모델링)
- (STEP 4) 클러스터링

```python
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np

# 데이터 파일 경로
file_path = r'C:\Users\pc\Desktop\파이썬연습\한국연구재단 AI 관련 논문 정보.xlsx'
```

```python
# 각 토픽에 대한 주요 단어 출력
def display_topics(model, feature_names, num_top_words):
    for index, topic in enumerate(model.components_):
        print(f"토픽 {index + 1}:")
        print(", ".join([feature_names[i] for i in topic.argsort()[:-num_top_words - 1:-1]]))

# 토픽 별 주요 단어 출력 (상위 10개)
num_top_words = 10
display_topics(lda, tfidf_vectorizer.get_feature_names(), num_top_words)
```

```python
# 전체 텍스트 데이터에 대해 전처리 수행
preprocessed_text = preprocess_text(all_text_data)

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation

# TF-IDF 벡터화
tfidf_vectorizer = TfidfVectorizer(max_df=0.99, min_df=2, stop_words='english')
tfidf = tfidf_vectorizer.fit_transform([preprocessed_text])

# LDA 모델 훈련
lda = LatentDirichletAllocation(n_components=5, random_state=42)
lda.fit(tfidf)
```

```python
# 클러스터링 (K-means)
num_clusters = 5  # 클러스터 개수 설정
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
kmeans.fit(tfidf_matrix)
```

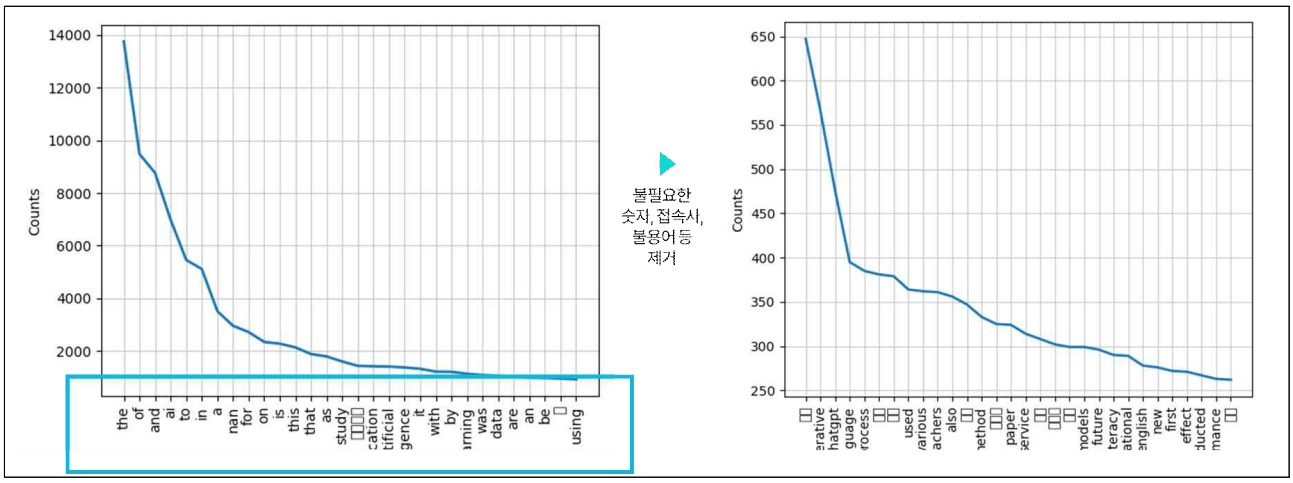
```python
# 각 클러스터의 중심에서 주요 단어 추출
terms = tfidf_vectorizer.get_feature_names_out()
tfidf_sorting = np.argsort(kmeans.cluster_centers_)[:, ::-1]

for i in range(num_clusters):
    top_keywords_idx = tfidf_sorting[i][:10]  # 상위 10개 단어의 인덱스 추출
    top_keywords = [terms[idx] for idx in top_keywords_idx]  # 상위 10개 단어 추출
    plt.figure(figsize=(8, 4))
    plt.barh(top_keywords, kmeans.cluster_centers_[i][top_keywords_idx], color='skyblue')
    plt.xlabel('TF-IDF Score')
    plt.title(f'Cluster {i+1} Top Keywords')
    plt.gca().invert_yaxis()
    plt.show()
```

<그림 15> 파이썬 분석 과정

- 텍스트 마이닝의 과정은 분석에 필요한 라이브러리를 호출, 데이터 불러오기, 텍스트 데이터 전처리, 소문자 변환 및 구두점 제거, 단어 토큰화 수행, 단어 빈도계산, 상위 빈도 단어 출력, 단어 빈도 시각화의 순서로 진행
- 클러스터링 과정은 텍스트 마이닝과 동일하게 라이브러리 및 데이터를 불러와, 분석에 수행할 컬럼을 설정하고, TF-IDF를 활용해 단어 특징 추출, K-means 클러스터링 수행, 주요 키워드 상위 10개의 인덱스 추출 및 시각화 수행

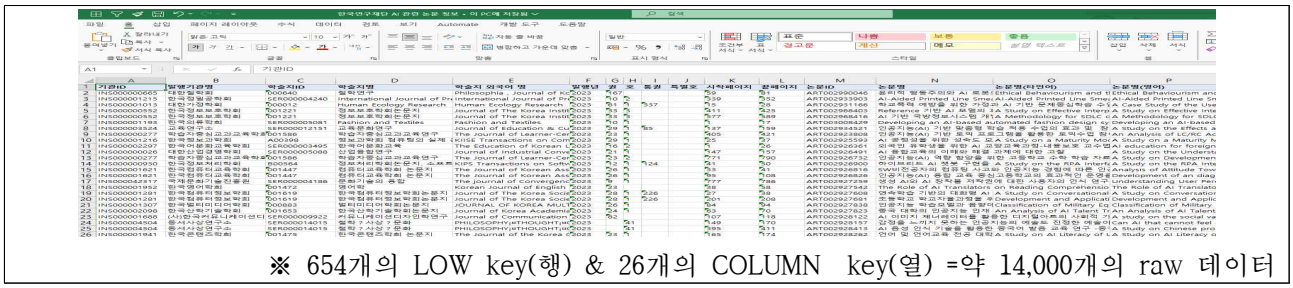
○ **텍스트 마이닝(토픽 모델링) 결과**

- 토픽 모델링 결과 단어 'AI' 7,037회, '인공지능' 1,600회, 'artificial' 1,437회, 'education' 1,416회, 'learning' 1,132회, 'technology' 894회, 'model' 796회 등으로 나타남
- 이후 불필요한 숫자, 접속사, 불용어 등을 제거하여 토픽 모델링 재수행 결과 'generative' 566회, 'chatgpt' 476회, 'language' 395회, 'process' 385회, 'various' 362회, 'method' 333회, '생성형' 302회 등의 유의미한 결과를 확인하였으며, 이를 통해 인공지능 학술 트렌드는 chatgpt, 생성형(generative), language임을 파악 가능



<그림 16> 텍스트 마이닝 결과

○ 활용 데이터

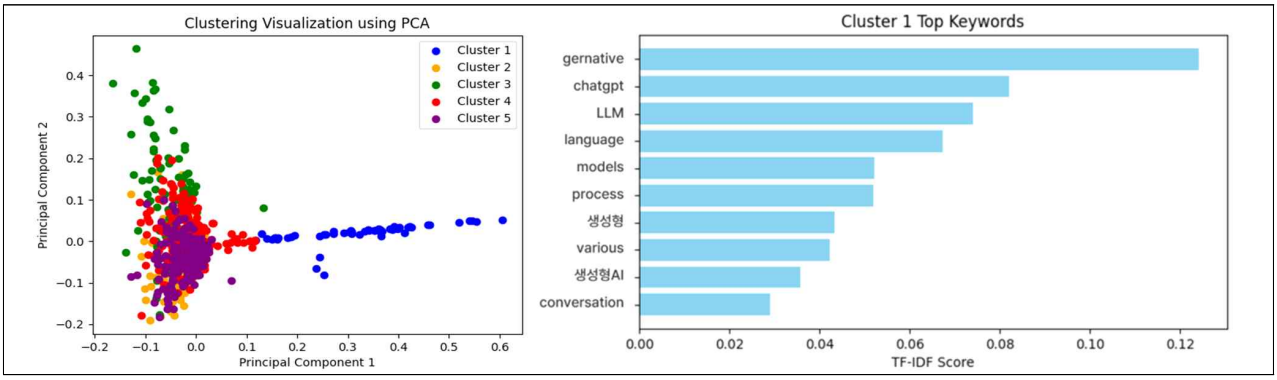


※ 654개의 LOW key(행) & 26개의 COLUMN key(열) =약 14,000개의 raw 데이터

<그림 17> 텍스트 마이닝 엑셀 결과

○ 클러스터링(K-means) 결과

- 데이터의 클러스터(군집)를 5개로 설정하여, K-means 클러스터링을 수행한 결과, 각 클러스터의 특징을 나타내는 주요 단어를 확인할 수 있었고, 막대그래프 등의 시각화를 통해 군집별 주제 및 트렌드 파악 가능



<그림 18> 클러스터링 결과

○ 향후 연구 방향

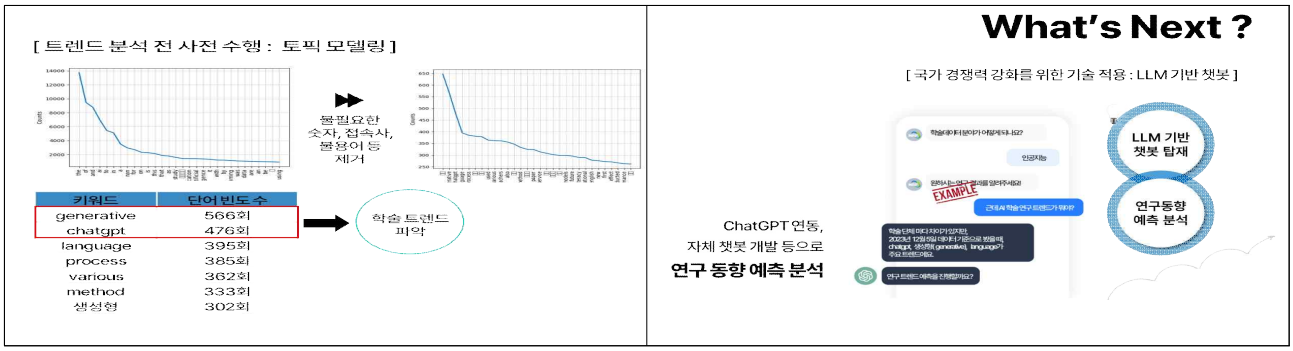
- 데이터 분석을 통해 상위 빈도의 단어를 확인, 주요 키워드 추출을 통해 학술적 트렌드 파악이 가능하였고, 주요 키워드별 상관관계 분석을 통해 논문 간 유사도 측정 등에 활용 가능. 또한, 논문 수록일을 기준으로 데이터를 정리하여 분석을 수행하면 시간에 따른

연구 주제의 변화 파악이 가능

- 키워드별 상관관계 분석을 통해 논문 간 유사도 측정, 논문 수록일을 기준으로 시간에 따른 연구주제의 변화 파악에 활용
- **(텍스트마이닝기반 트렌드 키워드 파악)** 토픽모델링 및 클러스터 결과를 포함하여 분야별 학술 관련 데이터에 대한 텍스트 마이닝으로 트렌드 키워드 표출, 국내 학술 활동, 인용 현황 관련 통계 조사 데이터에서 특정 키워드 수집을 통해 기간 내 등장 빈도 기반 트렌드 변화 분석
- **(시계열 예측 모델링을 통한 연구 동향 예측)** 학술지 수록 시기별 데이터 분석에 대해 ARIMA(p,d,q) 모델을 활용하여 이전 시점 값, q 이전시점의 오차 이용으로 과거 추세 반영하고, d차 차분을 통해 정상성 만족시켜 연구 동향 예측 분석, ARIMIA 예측 결과를 토대로 시스템 내 시각화 형태로 표출

<그림 19> 향후 연구 방향 시각화

- 연구 주제 및 키워드 트렌드를 토대로 학술적 연구 동향 예측
- 국내 학술 관련 데이터 관리시스템구축 또는 기 구축 지원시스템 고도화 기반 마련
- 심층 통계분석 자료제공 및 국내 학술정책을 위한 전략 수립

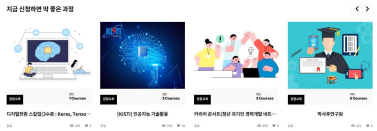
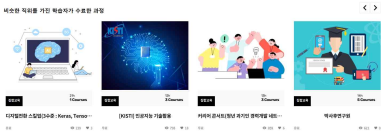


<그림 20> 향후 연구 방향 고도화 시각화

11. [국가과학기술인력개발원]

□ 데이터 분석 내용

학습데이터 기반 사용자 맞춤 교육과정 추천

✓ 지금까지는	✓ 앞으로는
□ **신청 마감일이 다가오는 교육과정 입과율이 저조한 교육 대상으로 추천** - 학습자 개인 맞춤형 교육과정 추천에 어려움 	□ **유사 학습자 수강통계 데이터와 본인이 학습한 콘텐츠 데이터를 활용하여 개인별 맞춤형 온라인 정규 교육과정 추천('24초)** □ **학습경험 데이터와 콘텐츠 메타데이터를 활용하여 개인별 맞춤 교육과정 및 지식콘텐츠 추천('24말)** 

✓ 분석 방법은
□ **수강 신청, 수료통계 데이터 정보 (학습자 직위, 직급, 기관 구분, 학습 과정 목록 등)를 통한 협업 기반 필터링과 콘텐츠 기반 필터링을 혼합하여 사용자 맞춤 교육과정 추천 알고리즘 개발** □ **향후('24년~) 현재 알파 캠퍼스에서 구축하고 있는 LRS xAPI 데이터를 사용하여 교육과정 추천 및 학습관리 알고리즘 개발*LRS(Learning Record Store): 학습 기록 저장소** *xAPI(experience API): 서로 다른 두 가지 소프트웨어나 응용프로그램 또는 플랫폼이 서로 정보를 교환할 수 있도록 정해진 형식이나 규칙을 의미함

<표 13> 사용자 맞춤 교육과정 추천

○ **컨설팅 결과**

- **(주제)** 디지털 기반 학습 및 역량개발 지원을 위해 운영 중인 온·오프라인 통합 학습관리 시스템, 알파 캠퍼스 내 맞춤형 측면에서 알림서비스와 교육과정 및 지식콘텐츠 추천 기능을 개선을 통해 학습데이터 기반 사용자 맞춤 교육과정 추천의 목적을 달성하기 위한 주제로 파악됨
- **(데이터 활용 업무 계획)** 알파 캠퍼스를 통해 학습 추천 및 개인화 지원 중인 상태이며, 금번 데이터 활용 과제를 통해 향후 하이플렉스 학습, 개인화 학습, 지능화 학습 등을 위한 사전 고려사항에 대해 준비를 할 수 있을 것으로 보임
- **(데이터 활용 과제)** 유사한 학습경험이나 패턴을 갖는 사용자의 수강 신청, 수료통계 데이터를 활용하여 K-means, Hierarchical Cluster 등 클러스터링을 통해 유사 학습자 그룹을 형성하고 개인화된

콘텐츠 추천이 가능할 것으로 보임

- 또한, 자연어처리 기술 중 하나인 Word Embedding 기술 등을 사용하여 학습별 특성 등을 벡터화하고, 사용자 정보와의 유사성을 기반으로 추천 알고리즘 개발할 수 있을 수 있을 것으로 보임
- LRS xAPI 데이터를 사용하여 TensorFlow, PyTorch 등을 통해 교육과정 추천 및 학습관리 알고리즘을 개발할 수 있을 것으로 판단됨
- **(활용 데이터)** 알파 캠퍼스 내 학습 기록 데이터는 콘텐츠 데이터 외 연간 발생 데이터량이 약 30만 건으로 비슷하기에 중장기적 관점에서 수강 신청 데이터, 수강통계 데이터, 콘텐츠 데이터, xAPI 데이터를 통합 관리할 수 있는 방안에 대해 고려할 수 있을 것으로 보임
- 제시된 데이터 외에도 피드백 데이터, 행동 로그 데이터 등 강의나 콘텐츠에 대한 만족도 및 개선점을 파악할 수 있는 유관 데이터 확보를 통해 온라인 학습 맞춤형 교육과정에서의 필요 요소 기능에 기여할 수 있을 것으로 판단됨

□ **데이터 분석 내용 및 결과**

○ **분석 개요**

- 학습데이터 기반 사용자 맞춤 교육과정 추천 및 학습관리 알고리즘 개발

○ **분석 목표**

- 수강 데이터 활용 통계분석으로 실참여율 90% 이상 데이터 추출, 텍스트 마이닝을 통해 수강률 상위 과목 키워드 출력 및 시각화

○ **분석 도구/기법** : Excel, python / 통계분석, 텍스트 마이닝

○ **분석 과정**

- 데이터 중 화면 노출 시간/접속시간 x 100으로 학습자의 실제 강의 참여율 확인
- 실 참여율이 90% 이상인 학습자 데이터 중 상위 빈도의 주요 과목명 및 주요 키워드 추출
- (STEP 1) 분석 도구 선정 및 데이터 로드
- (STEP 2) 데이터 통계분석 및 데이터 전처리

1) 실 참여율(%) = 노출 시간/접속시간 x 100

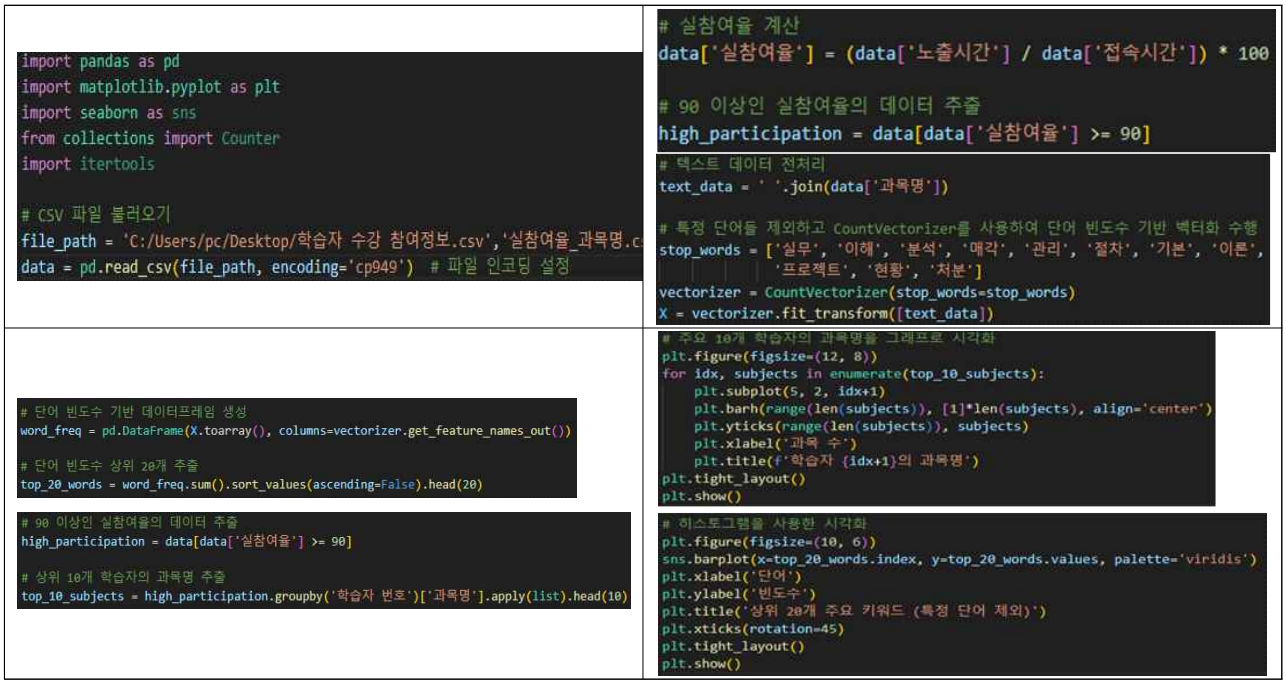
2) 실 참여율 텍스트 데이터 전처리

- (STEP 3) 데이터 추출

1) 상위 빈도 단어 및 과목명 추출

- (STEP 4) 시각화

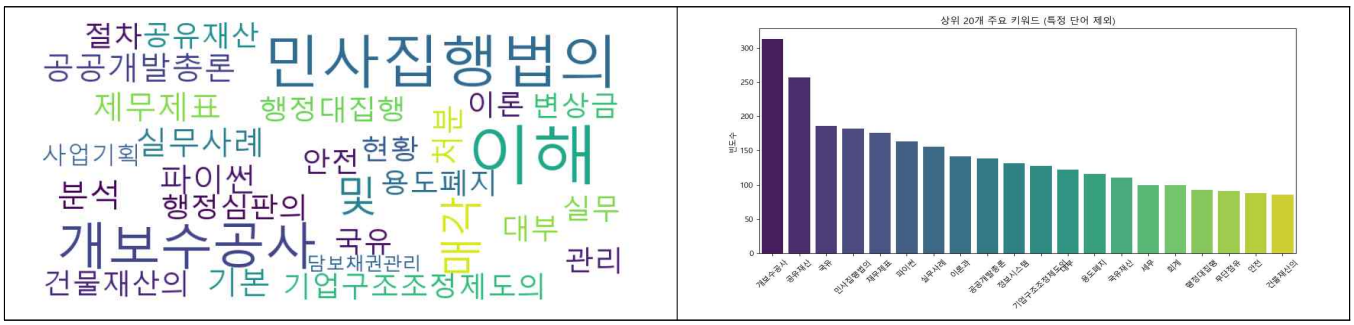
1) 주요 10개 과목명 및 상위 주요 20개 키워드 (특정 단어 제외) 시각화



< 그림 21> 키워드 파이선 분석과정

○ 텍스트 마이닝 결과

- 실 참여율 90% 이상인 학습자 데이터 중 상위 주요 키워드는 실무, 공유재산, 파이썬, 세무, 무단점유 등으로 확인
- 상위 과목은 민사집행법의 이해, 재무제표 분석, 파이썬 기본, 매각 실무사례 순으로 빈도수가 높음을 확인 후 워드 클라우드 시각화 수행



< 그림 22> 텍스트마이닝 결과

○ 분석결과

- 엑셀 활용 실 참여율(%) 도출
- 학습자 실 참여율 90% 이상인 과목의 상위 주요 키워드는 '실무' 430회, '개보수공사' 313회, '이해' 296회, '공유재산' 257회, '분석' 247회 등으로, 하위 주요 키워드는 '유지설비' 8회, '계약', '건설사업', '프로세스' 5회 등으로 나타남
- 빈도수 상위 과목명 추출을 위한 텍스트 마이닝 결과, '민사집행법의 이해', '제무제표분석', '파이썬 기본', '매각 실무사례' 등의 순으로 빈도수가 높음을 확인

○ **향후 연구 방향**

- **(데이터 추가 수집 및 결합)** 사용자 로그 기록 데이터를 수집하여 분석결과와 결합하여 사용자 패턴 분석 수행, 로그 데이터에는 사용자의 접속경로, 검색기록, 수강과목 등이 포함

<그림 23> 향후 연구방향 예시 결과

- 추전 시스템 모델 적용

1) **(콘텐츠 기반 필터링)** TF-IDF, Word Embeddings 등을 사용해 학습자 행동 기록, (수강과목, 관심 분야, 시청 영상 등)을 바탕으로 콘텐츠의 유사성을 계산하고 맞춤형 콘텐츠 추천

2) **(협업 필터링)** 사용자의 학습 기록, 선호도 등을 바탕으로 다른 사용자들과의 유사성 또는 학습 항목 간 유사성 분석, 사용자 기반 혹은 아이템 기반으로 추천 콘텐츠 선정

3) **(딥러닝 모델)** 학습자 데이터 활용 복잡한 패턴을 학습할 수 있는 신경망 모델 활용, RNN, LSTM, Transformer와 같은 신경망 모델의 아키텍처를 사용해 시퀀스 데이터를 처리하고 추천

< 그림 24> 결과 ui 화면 예시

제4장 정책(업무) 활용실적 및 사후관리 방안

○ 데이터 기반 행정과 관련된 '24년도 과기정통부의 데이터 분야 사업 기획 및 현재 수행 중인 데이터 활용 업무 개선에 활용

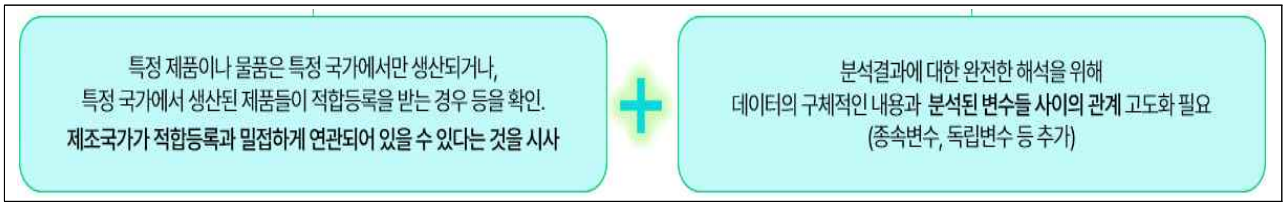
○ 데이터 인프라 구축, 데이터 분석과제 기획 및 제안요청서 작성 등 업무에 활용

○ 또한, 정부 데이터 기반 행정 활성화 실태조사 등 정부 정책 기조에 부응 가능

1. 기관별 기대효과

① 국립전파연구원

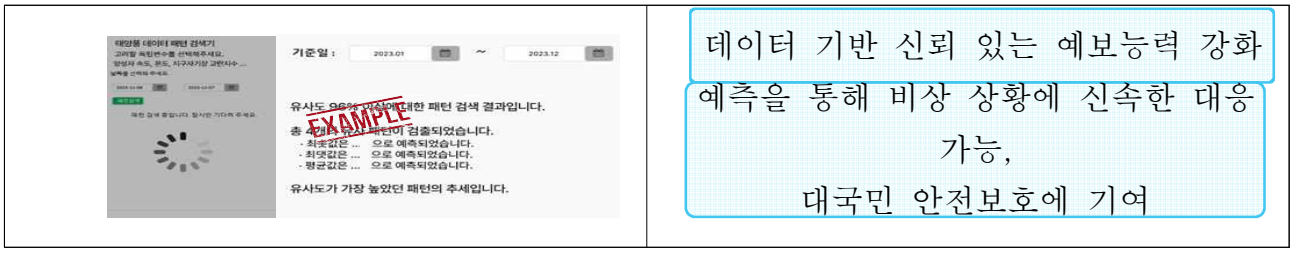
- **(데이터 분석주제)** 방송 통신기기와 전자제품에 대한 적합성 평가 현황분석
- **(기대효과)** 평가 효율성 증대 및 기기의 안정성 확보를 위한 정책적 방향 제시. 또한, 국내 제조업체의 제품 경쟁력 향상 및 국제 표준화 활동에 기여



<그림25> 국립전파연구원 기대효과

② 국립전파연구원 우주전파센터

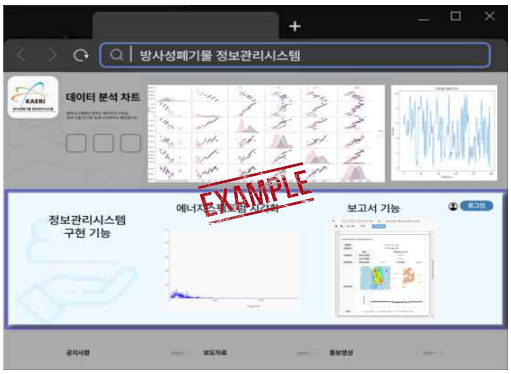
- **(데이터 분석주제)** 우주전파환경 경보(R,S,G,I)상황 사후분석
- **(기대효과)** 우주 환경 변화에 따른 국내 통신 및 위성 시스템에 대한 영향을 예측하고, 사전 대응책 마련으로 시스템 안전성 확보



<그림 26> 우주전파센터 UI 예시 화면

③ 한국원자력연구원

- **(데이터 분석주제)** 방사성폐기물 정보관리시스템
- **(기대효과)** 투명한 폐기물 관리를 위한 데이터 활용, 시민들의 신뢰 증대 및 안전한 폐기물 처리를 위한 전략적 방향 제시



폐기물스트림별 관리를 위해 분류체계 및 분석기능 도입

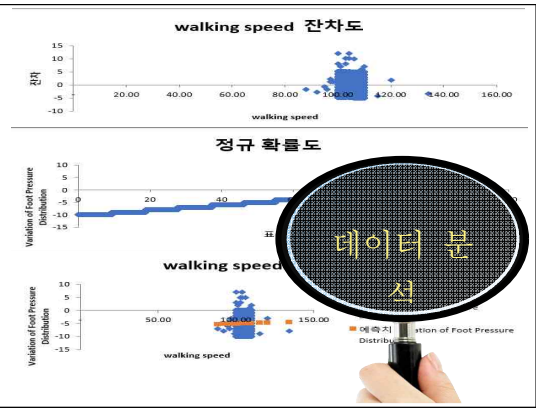
폐기물 정보별 인수기준 조건 분석을 통해 스트림별 폐기물관리체계 도입

방사성폐기물 보고서 기능으로 데이터 기반 의사결정 및 정책 제언 지원 등에 활용

<그림 27> 한국원자력연구원 구현 예시 화면

④ 한국생산기술연구원

- **(데이터 분석주제)** 고령자 균형 프로토콜 측정 시 수집되는 웨어러블 센서 데이터를 활용한 고령자의 일상생활 건강 상태 예측
- **(기대효과)** 고령자의 건강 관리 및 사고 예방을 위한 실시간 모니터링 가능. 치료 및 복지 정책의 효율적인 실행을 지원



웨어러블 센서 데이터만으로 고령자의 노인성 질환 예측

고령자 일상생활 건강상태 모니터링 활용

<그림 28> 한국생산기술연구원 구현 예시 화면

⑤ 기초과학연구원

- **(데이터 분석주제)** 기초연구 분야 연구 전략 수립 의사결정 지원
- **(기대효과)** 연구 투자의 최적화 및 연구 방향성 제시, 국내 기초과학 연구 활성화 및 국제 경쟁력 강화

분야별 논문 현황 내역 제공

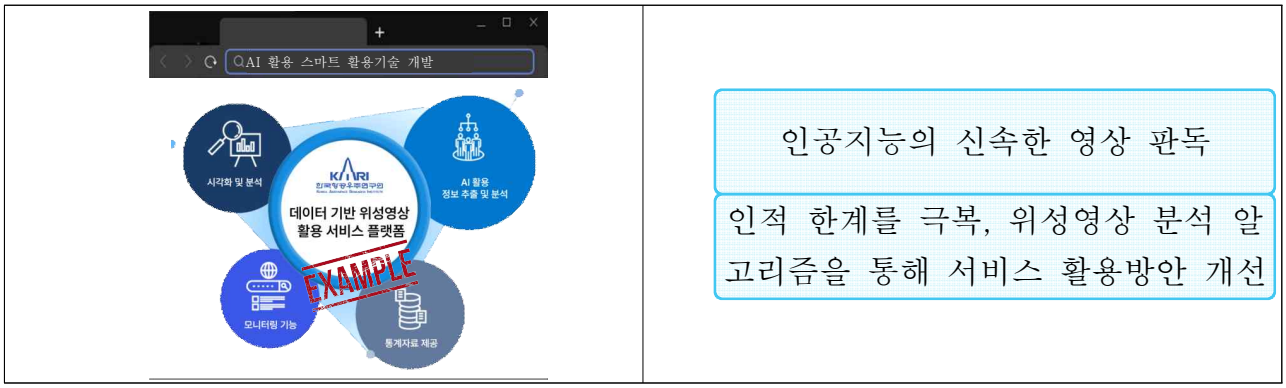
논문 저자 소속, 연구 분야 등 데이터 활용,

고도화 분석을 통한 데이터 기반 연구 전략 수립 의사결정 지원 가능

<그림 29> 기초과학연구원 구현 예시 화면

⑥ 한국항공우주연구원

- **(데이터 분석주제)** AI 활용 스마트 활용기술 개발
- **(기대효과)** 신속한 위성영상 판독, 여러 종류의 위성영상을 종합적으로 판독하여 기존에 사람이 찾기 어려웠던 정보 또한 탐색



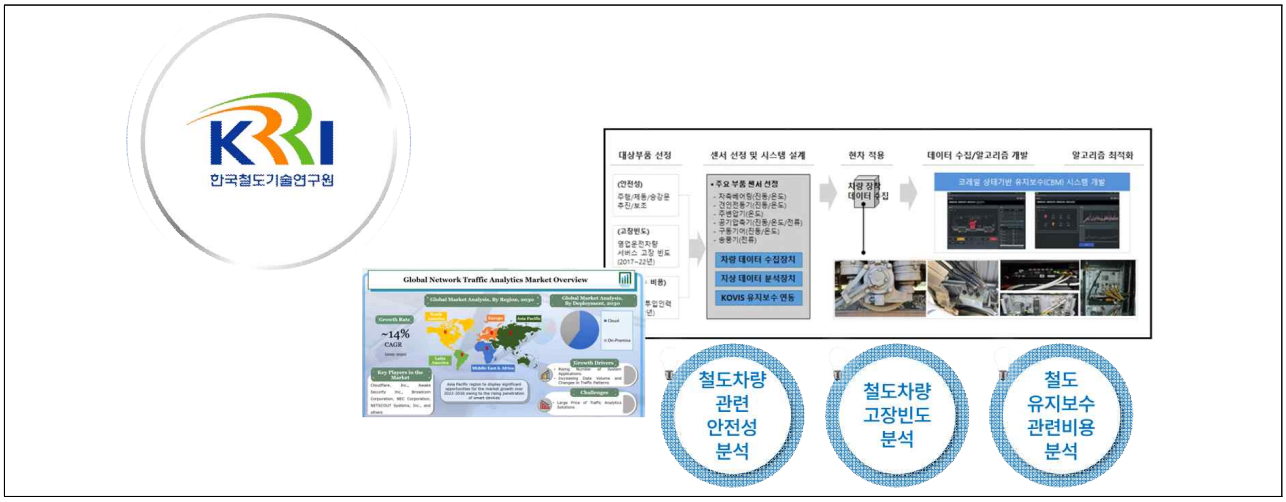
<그림 30> 한국항공우주연구원 구현 예시 화면

⑦ 나로우주센터

- **(데이터 분석주제)** 홈페이지 방문객 분석을 통한 콘텐츠 다양화 및 연계프로그램 제공
- **(기대효과)** 타겟 오디언스의 정확한 파악을 통한 맞춤형 콘텐츠 제공, 방문자의 활성화 및 홈페이지 이용도 증가

⑧ 한국철도기술연구원

- **(데이터 분석주제)** 반복되는 업무의 효율적 분석 기초 환경 구성(자동화)
- **(기대효과)** 철도 관련 무역데이터 크롤링 및 쿼리 자동화를 통해 인적 자원의 효율적인 사용, 연구데이터 중복분석으로 예산 편성, 정책 등의 문제 개선에 활용

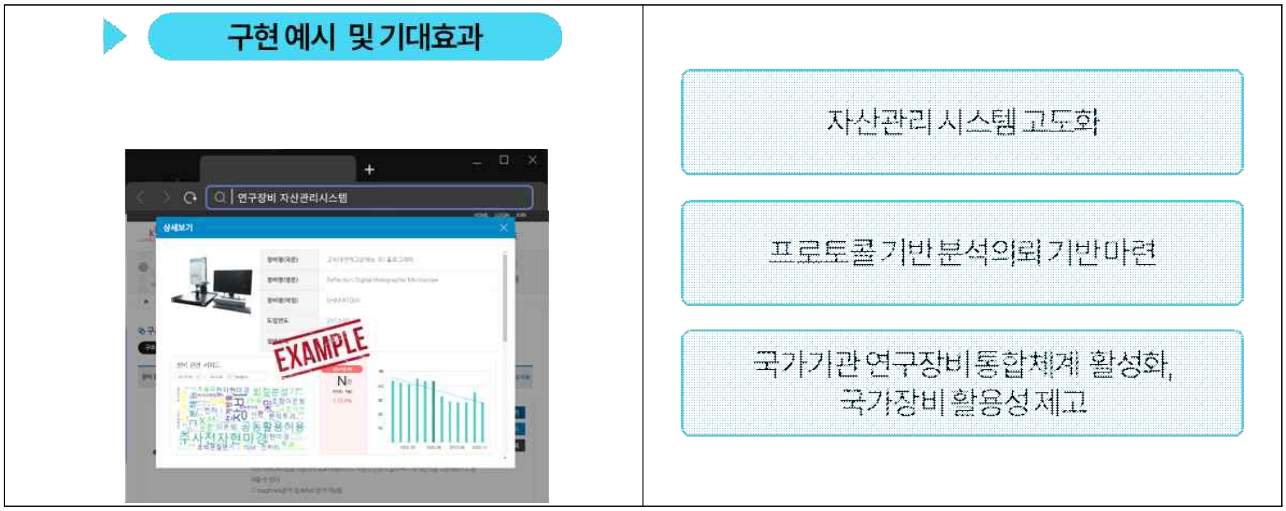


<그림 31> 한국철도기술연구원 구현 예시 화면

⑨ 한국과학기술연구원

- **(데이터 분석주제)** 연구 장비 자산관리 및 장비 예약 활용관리 고도화

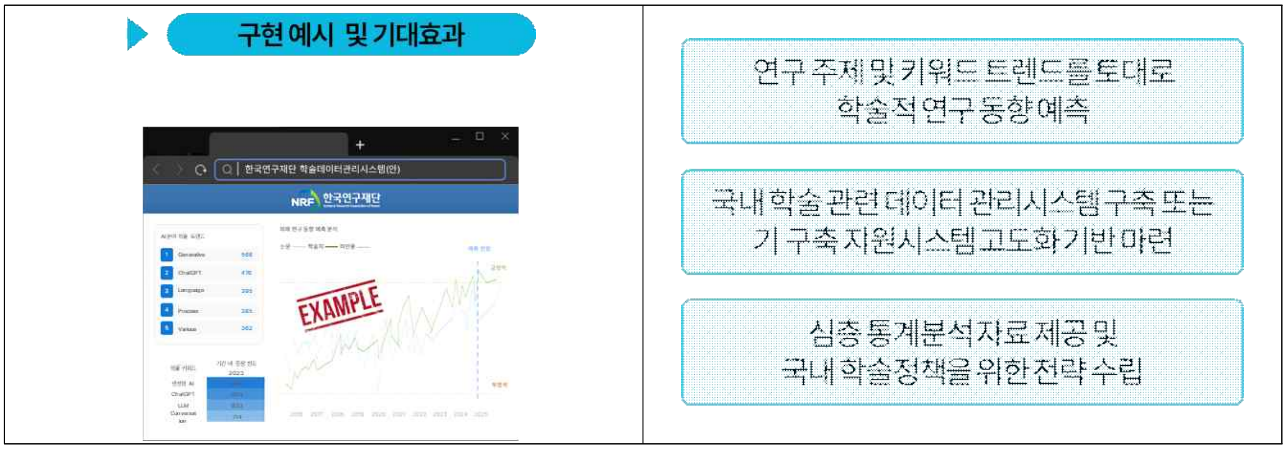
- **(기대효과)** 데이터 분석결과에 대한 모니터링을 통해 활용 장비 유지보수 관리 및 활용도 제고



<그림 32> 한국과학기술연구원 구현 예시 화면

⑩ 한국연구재단

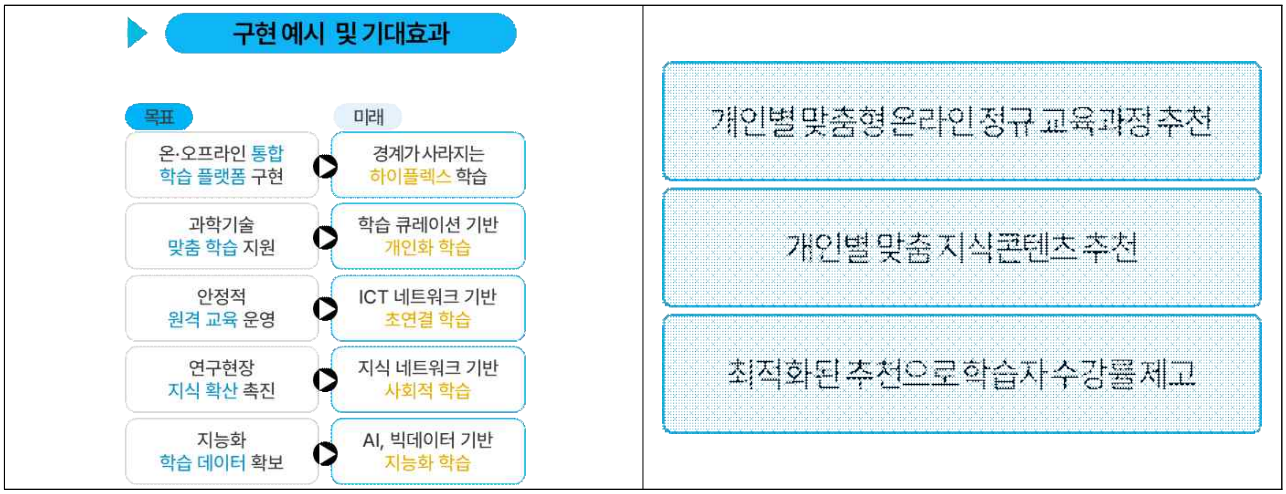
- **(데이터 분석주제)** 국내 학술 활동, 학술지 인용, 학술단체 인용 현황 등에 대한 분석
- **(기대효과)** 시기별 변화되어온 연구 주제의 주요 키워드 및 학술적 트랜드 파악으로 지식 이전 및 학문 분야 발전에 기여



<그림 33> 한국연구재단 구현 예시 화면

⑪ 국가과학기술인력개발원

- **(데이터 분석주제)** 학습데이터 기반 사용자 맞춤 교육과정 추천
- **(기대효과)** 학습 큐레이션을 통한 개인화된 학습경험 제공으로 학습 효율성 증대, 학습 동기 부여, 학습 결과 향상, 이로 인해 교육 기관의 경쟁력 강화 가능

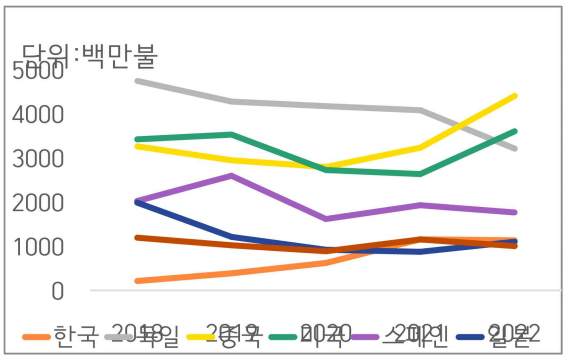


<그림 34> 국가과학기술인력 구현 예시 화면

2. 시범 적용 기관 선정 및 모형 도출

2-1. 한국철도기술연구원

○ **선정배경**

한국철도기술연구원 선정배경	
 [그림n] 연도에 따른 수출 그래프	• 현재까지의 무역데이터 분석은 수작업과 중복된 연구가 발주되는 등 효율성이 떨어지는 상황이며, 시스템적인 개선이 필요 • 시범 적용을 통해 타 기관에 베스트프랙티스를 전달하고, 연구 자료를 보다 효과적으로 활용할 수 있는 모델을 도출

[표13] 한국철도기술연구원 선정배경

○ **선정기준**

- 데이터 활용 역량: 해당 기관이 철도 분야 무역데이터, 연구데이터, 해외데이터를 종합적으로 다루는 데에 어떤 역량을 가지고 있는지를 평가
- 연구 협력 경험: 기관이 공공기관 빅데이터 관련 업무를 성공적으로 수행한 경험을 고려하여 연구 협력에 적합한지를 평가
- 기술 도입 가능성: 파이썬, 텍스트마이닝, 클러스터링 등 다양한 머신러닝 기법뿐만 아니라 하드웨어 및 소프트웨어 기술을 어떻게 도입할 수 있는지를 평가

○ **모형 도출의 목적**

- 타 기관 활용 활성화: 개발된 모형이 철도 분야 무역데이터뿐만 아니라 다양한 무역 회사 및 관련 산업에도 영향을 미치게 하여 타 기관의 데이터 활용 역량을 활성화합니다.

- 베스트 프랙티스 전달: 모형의 개발 및 적용 경험을 통해 얻은 베스트 프랙티스를 선정된 기관에 전달하여 철도 분야에서의 효과적인 데이터 활용을 공유
- 데이터 분석 : 한국철도기술연구원은 철도 분야 무역데이터, 연구데이터, 해외데이터를 종합적으로 분석하여 효과적으로 활용
- 공공기관 빅데이터 관련 업무를 풍부하게 수행한 기관으로부터 기술 및 경험을 효과적으로 활용하여 프로젝트를 추진
- 파이썬, 텍스트마이닝, 클러스터링 등의 머신러닝 기법을 사용하여 데이터를 분석하는 데에 기술을 적극적으로 도입할 예정

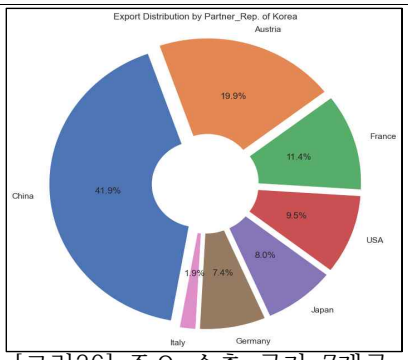
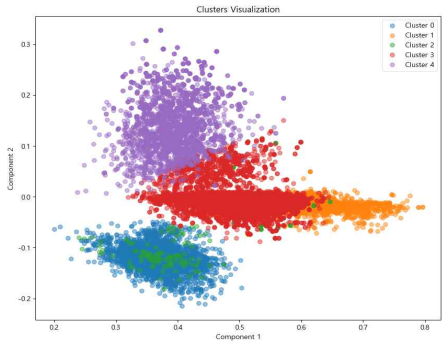
○ **시범 적용 프로세스**

- 교육 및 이해: 선정된 기관에 대한 데이터 활용 방법 및 모형의 이해를 위한 교육을 진행
- 모형 적용 및 피드백 수집: 선정된 기관에서 모형을 철도 분야의 무역데이터에 적용하며 발생한 문제점 및 개선점에 대한 피드백을 수집
- 모형 개선 및 최적화: 피드백을 기반으로 모형을 개선하고 최적화
- 성과 평가 및 전파: 모형의 성과를 평가하고, 성과가 우수한 경우 다른 기관에도 적용을 권장하며, 베스트 프랙티스를 전파

○ **예상효과**

- 효율적인 데이터 활용: 모형을 통해 다양한 무역데이터를 효과적으로 분석하여 철도 분야의 데이터 활용
- 연구 협력 강화: 선정된 기관과의 협력을 통해 다양한 데이터 분석 및 머신러닝 프로젝트에 대한 기반을 구축
- 베스트프랙티스 전파: 선정된 기관에 우수한 모형을 적용하고 효과를 측정함으로써, 다른 기관에 베스트프랙티스를 전파

○ 실물모형 목업 프로세스

	모델1	모델2
개요	UN Comtrade 및 관세청에서 제공하는 무역데이터 작업을 파이썬을 통해 효율화, 자동화, 시각화	한국철도연구원 내 자료를 보고서로 활용하도록 시각화 및 도표화
데이터 수집	UN Comtrade Database [그림35] UN Comtrade Database	한국철도연구원 자료 • 한국철도연구원_연구사업정보 • 한국철도연구원_논문 • 한국철도연구원_연구보고서 • 한국철도연구원_지식재산권
데이터 전처리	• 필요한 컬럼 추출 • 데이터프레임 재정의 • 데이터프레임 필터링 • 데이터 결합	• 컬럼명 특정 문자 제거 • 중복 제거 • 인덱스 추출
사용 기술	• 파이썬 : 강력한 프로그래밍 언어로, 다양한 운영 체제에서 사용 • Seaborn, matplotlib : 시각화 라이브러리 • Pandas :데이터 조작과 분석을 위한 라이브러리 • networkx :복잡한 네트워크 구조를 생성하고 분석하는 데 사용	• 파이썬 : 강력한 프로그래밍 언어로, 다양한 운영 체제에서 사용 • 클러스터링 : 비슷한 속성을 갖는 데이터를 그룹화하는 기술 • 코사인 유사도: 두 벡터 간의 유사도를 측정하는 데 사용되는 측정 방법 • matplotlib시 : 각화 라이브러리 • TruncatedSVD : 특이값 분해(SVD)를 사용하여 데이터의 차원을 감소시키는 기술
최종 시각화	 [그림36] 주요 수출 국가 7개국	 [그림37] 클러스터링 결과

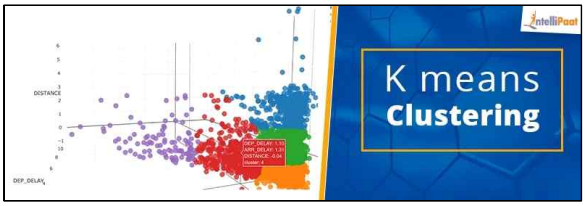
[표14] 실물목형 목업 프로세스

- 활용 데이터

수집데이터	형식	보유기관	비고
철도 무역 데이터	csv	UN Comtrade	download
한국철도연구원_연구사업정보	csv	한국철도연구원	수급
한국철도연구원_논문	csv	한국철도연구원	수급
한국철도연구원_연구보고서	csv	한국철도연구원	수급
한국철도연구원_지식재산권	csv	한국철도연구원	수급

[표15] 활용 데이터

- 분석 알고리즘

모델1	모델2
	
[그림38]판다스	[그림39] K-means 클러스터링

[표16] 분석 알고리즘

- (모델1) 사용 코드 및 시각화

① 라이브러리 호출

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
plt.rcParams['font.family'] ='Malgun Gothic'
plt.rcParams['axes.unicode_minus'] =False
import networkx as nx
plt.rc("font", family = "Malgun Gothic")
sns.set(font="Malgun Gothic",
rc={"axes.unicode_minus":False}, style='white')
```

② 데이터 불러오기

```
# 데이터 불러오기
data = pd.read_excel(r'data')

# 필요한 열 분류
dfs = data[['Period','ReporterDesc','FlowDesc','PartnerDesc','CmdCode','PrimaryValue']]
dfs = dfs[dfs['PartnerDesc'] != 'World']

# '한국'을 기준으로 데이터프레임 재정의
korea_df = dfs[dfs['PartnerDesc'] == 'Rep. of Korea'].reset_index(drop=True)
```

③ 데이터 필터링

```
# 필요한 조건에 따라 데이터프레임을 필터링하여 새로운 데이터프레임 생성
df_export = korea_df[korea_df['FlowDesc'] == 'Export'][['ReporterDesc', 'PartnerDesc', 'PrimaryValue']]
df_export.columns = ['ReporterDesc', 'PartnerDesc', '수출']

df_import = korea_df[korea_df['FlowDesc'] == 'Import'][['ReporterDesc', 'PartnerDesc', 'PrimaryValue']]
df_import.columns = ['ReporterDesc', 'PartnerDesc', '수입']
```

④ 데이터 결합

```
# 수출과 수입 데이터를 합치기
df_combined = pd.merge(df_export, df_import, on=[ 'ReporterDesc', 'PartnerDesc'], how='outer')
df_grouped = df_combined.groupby(['ReporterDesc', 'PartnerDesc']).sum().reset_index()
df_grouped = df_grouped.sort_values(by='수출',ascending=False)

# 수출과 수입 데이터 (1,234,567) 변경
df_grouped['수출'] = df_grouped['수출'].apply(lambda x: '{:,.0f}'.format(x))
df_grouped['수입'] = df_grouped['수입'].apply(lambda x: '{:,.0f}'.format(x))
```

⑤ 데이터 속성 변경

```
# 수출과 수입 데이터 속성 '실수'로 변경
df_grouped = df_grouped[['PartnerDesc', 'ReporterDesc', '수출', '수입']]
df_grouped = df_grouped.reset_index(drop=True)
df_grouped['수출'] = df_grouped['수출'].str.replace(',', '').astype(float)
df_grouped['수입'] = df_grouped['수입'].str.replace(',', '').astype(float)
df_grouped
```

⑥ 피벗 테이블

```
# Import 및 Export 값 각각을 PrimaryValue 값으로 묶고 ReporterDesc별로 합산
df_pivot = dfs.pivot_table(index='ReporterDesc', columns='FlowDesc', values='PrimaryValue', aggfunc='sum', fill_value=0)

# 총합계 계산
df_pivot['총합계'] = df_pivot.sum(axis=1)

# 데이터 출력
df_pivot = df_pivot.sort_values(by='총합계', ascending=False)
df_pivot['총합계'] = df_pivot['총합계'].apply(lambda x: '{:,.0f}'.format(x))
df_pivot['Export'] = df_pivot['Export'].apply(lambda x: '{:,.0f}'.format(x))
df_pivot['Import'] = df_pivot['Import'].apply(lambda x: '{:,.0f}'.format(x))
df_pivot
```

⑦ 시각화

```
#시각화
# Seaborn을 사용하여 시각화
sns.set(style="whitegrid")
plt.figure(figsize=(10, 6))
sns.barplot(x='ReporterDesc', y='수출', hue='ReporterDesc', data=df_grouped.head(), palette='viridis')
plt.title('Export by Partner_Rep. of Korea')
plt.xlabel('Partner')
plt.ylabel('Export')
plt.xticks(ha='right')
plt.tight_layout()
plt.show()

# 원 그래프 생성
explode = [0.05, 0.05, 0.05, 0.05, 0.05 ,0.05, 0.05]
wedgeprops={'width': 0.7, 'edgecolor': 'w', 'linewidth': 3}
labels = df_grouped['ReporterDesc'][:7]
sizes = df_grouped['수출'][:7]
plt.figure(figsize=(8, 8))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=260, counterclock=False, explode=explode, wedgeprops=wedgeprops)
plt.title('Export Distribution by Partner_Rep. of Korea')
plt.axis('equal')  # 원 그래프를 둥그렇게 표시
plt.show()
```

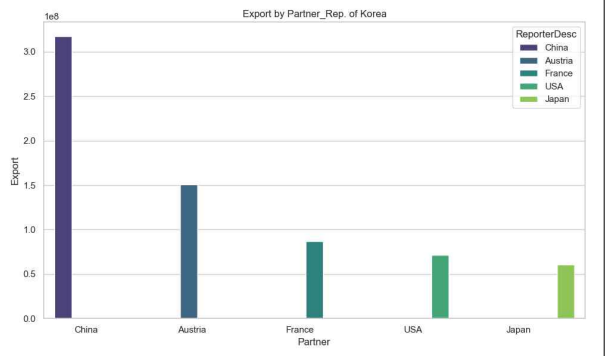
[표17] 모델1-한국철도연구원 파이썬 코드

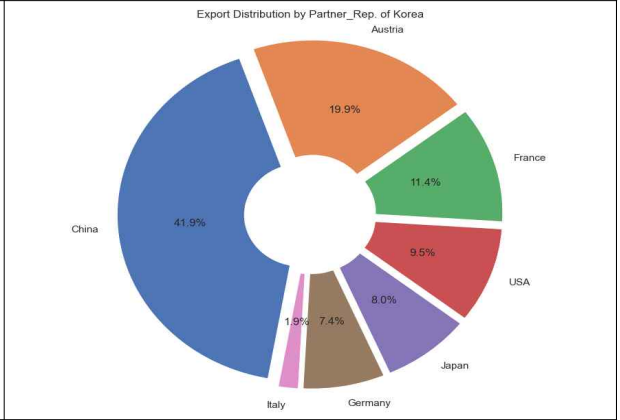
- (모델1) 분석 처리 순서

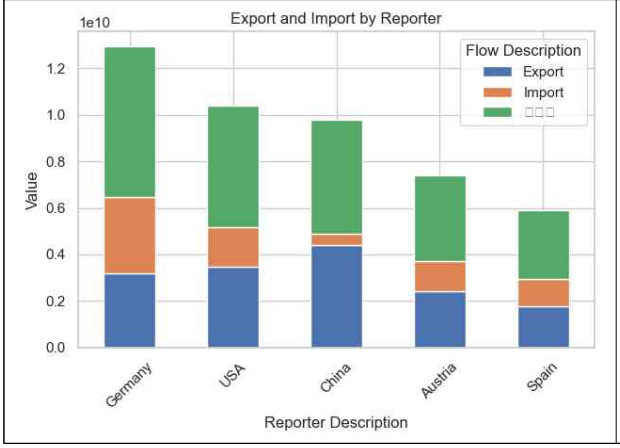
1. 데이터 처리를 위한 파이썬에 대한 라이브러리 호출
2. 데이터 불러온 후, 한국을 기준으로 데이터 프레임 재정의
3. 필요한 조건에 따라 데이터 프레임 필터링 및 생성
4. 수출과 수입데이터 결합 후, 데이터 (ex. 1,234,567) 변경
5. 수출과 수입데이터 속성 '실수'로 변경
6. 각 나라에 따른 수출, 수입, 총합계 데이터 피벗데이블로 정의
7. 막대그래프, 원그래프 등 시각화

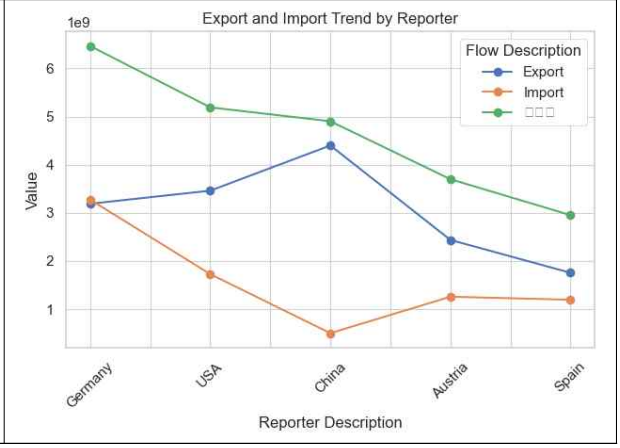
	PartnerDesc	ReporterDesc	수출	수입
0	Rep. of Korea	China	317453178.0	35332470.0
1	Rep. of Korea	Austria	151051928.0	17842782.0
2	Rep. of Korea	France	86721208.0	123565.0
3	Rep. of Korea	USA	71681466.0	108085000.0
4	Rep. of Korea	Japan	60641487.0	27756361.0

FlowDesc	Export	Import	총합계
ReporterDesc			
Germany	3,189,211,005	3,274,043,391	6,463,254,396
USA	3,460,833,768	1,728,708,000	5,189,541,768
China	4,398,404,199	503,379,983	4,901,784,182
Austria	2,438,647,211	1,259,664,603	3,698,311,814
Spain	1,756,091,091	1,197,262,767	2,953,353,859









[표18] 모델1-결과값 시각화

- **(모델1) 분석결과**

- 국제 무역 데이터를 통해 무역 패턴 이해, 품목별 무역 증감 추이, 무역 국가 간 비교, 수출 다양성 분석, 섹터 및 산업 분석, 경쟁자 분석, 수출 전략 도표화 및 시각화 가능할 것으로 보임

- **(모델2) 사용 코드 및 시각화**

① 라이브러리 호출

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt
plt.rcParams['font.family'] ='Malgun Gothic'
plt.rcParams['axes.unicode_minus'] =False
from sklearn.decomposition import TruncatedSVD
import networkx as nx
```

② 데이터 불러온 후, 결합

```
file_path = ['한국철도기술연구원 연구사업정보_20230818.csv',
             '한국철도기술연구원_논문_20231030.csv',
             '한국철도기술연구원_연구보고서_20231117.csv',
             '한국철도기술연구원_지식재산권_20221017.csv']

# 데이터 읽기 및 결합
dfs = [pd.read_csv(file_path) for file_path in file_path]
combined_dfs = pd.concat(dfs)

# TF-IDF 벡터화를 위한 텍스트 데이터 합치기
text_data = combined_dfs.astype(str).agg(' '.join,axis=1)
```

③ TF-IDF, 클러스터링

```
# TF-IDF 벡터화
tfidf_vectorizer = TfidfVectorizer(stop_words='english')
tfidf_matrix = tfidf_vectorizer.fit_transform(text_data)
tfidf_matrix

# Kmeans 클러스터링
num_cluster = 5
k_means = KMeans(n_clusters=num_cluster, random_state=42)
cluster = k_means.fit_predict(tfidf_matrix)

# 클러스터링 결과 데이터프레임에 추가
combined_dfs['Cluster'] = cluster
combined_dfs
```

④ SVD, 인덱스 추출

```
svd = TruncatedSVD(n_components=2)
tfidf_reduced = svd.fit_transform(tfidf_matrix)

# 상위 중요 단어 인덱스 추출
feature_names = tfidf_vectorizer.get_feature_names_out()
top_keywords = {}
num_keywords = 10
```

⑤ TF-IDF 평균 계산

```
for cluster_id in range(num_cluster):
  cluster_text = text_data[combined_dfs['Cluster'] == cluster_id]
  cluster_tfidf = tfidf_matrix[combined_dfs['Cluster']==cluster_id]

  # 각 클러스트에 대한 TF-IDF 값의 평균 계산
  avg_tfidf = cluster_tfidf.mean(axis=0)

  # 상위 중요 단어 인덱스 추출
  top_keywords_indices = avg_tfidf.argsort()[0, -num_keywords:]

  # # 상위 중요 단어
  top_keywords[cluster_id] = [feature_names[i] for i in top_keywords_indices]
```

⑥ 단어 출력

```
# 각 클러스트의 특정적인 단어 출력
for cluster_id, keywords in top_keywords.items():
    keywords_str = ", ".join(str(keyword) for keyword in keywords)
    print(f"Cluster {cluster_id}의 특정적인 단어: {keywords_str}")

# 각 클러스터의 Truncated SVD 결과를 추출합니다.
cluster_reduced = {}
for cluster_id in range(num_cluster):
    cluster_reduced[cluster_id] = tfidf_reduced[combined_dfs['Cluster'] == cluster_id]
```

⑦ 시각화

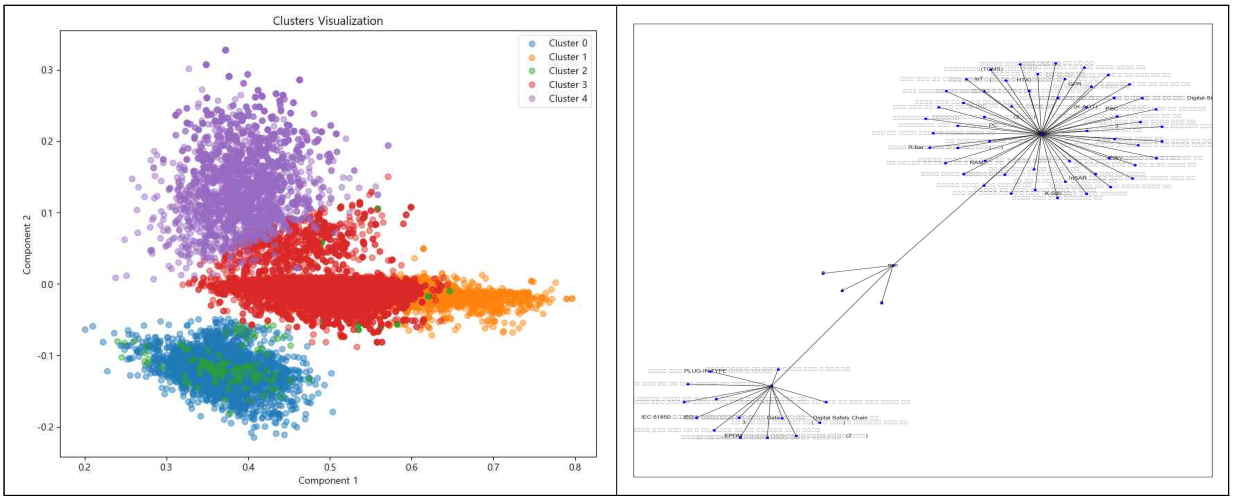
```
# 각 클러스터의 Truncated SVD 결과를 추출합니다.
cluster_reduced = {}
for cluster_id in range(num_cluster):
    cluster_reduced[cluster_id] = tfidf_reduced[combined_dfs['Cluster'] == cluster_id]

# 각 클러스터를 산점도로 시각화합니다.
plt.figure(figsize=(10, 8))
for cluster_id, cluster_data in cluster_reduced.items():
    plt.scatter(cluster_data[:, 0], cluster_data[:, 1], label=f'Cluster {cluster_id}', alpha=0.5)

plt.title('Clusters Visualization')
plt.xlabel('Component 1')
plt.ylabel('Component 2')
plt.legend()
plt.show()
```

```
#Network analysis
g = nx.Graph()
g = nx.from_pandas_edgelist(combined_dfs, source = '연구사업명', target = 'Cluster')

plt.figure(figsize=(20, 20))
pos = nx.spring_layout(g, k = 0.15)
nx.draw_networkx(g,pos, node_size = 25, node_color = 'blue')
plt.show()
```

[표19] 모델2-한국철도연구원 파이썬 코드

- (모델2) 분석 처리 순서

1. 파이썬/NLP(자연어처리)에 대한 라이브러리 호출
2. 데이터 호출 후 데이터 결합
2. TF-IDF 벡터화 및 클러스터링 적용, 예측
4. 벡터화 데이터 SVD 적용 및 인덱스 추출
5. 각 클러스터 간 TF-IDF 값의 평균 계산, 상위 중요 단어 추출
6. 각 클러스터의 특정 단어 출력, SVD 결과 출력
7. 출력한 SVD 결과물 스캐터를 통해 시각화



[표20] 모델2-결과값 시각화

- (모델2) 분석결과
- 단순 빈도 기반의 키워드 추출을 진행을 넘어서 TF-IDF나 word2vec 등의 고급 방법론을 활용한 텍스트 마이닝 진행
- 클러스터링 사용하여 각 클러스터의 관계 유사도 비교 및 주요 단어 출력
- 키워드 간의 관계나 연계성을 파악하여 연구 주제나 트렌드 간의 연관성을 히트맵으로 시각화하여 파악

2-2. 기초과학연구원

○ 선정배경

- 데이터 기반 의사결정의 중요성: 현재의 연구 환경에서는 데이터 기반 의사결정이 중요성을 더해가고 있습니다. 주제로 선정된 IBS(기초과학연구원)의 연구 성과 데이터 분석은 이러한 흐름에 부응하여 중요한 주제로 간주

기초과학연구원 선정배경	
 [그림40] 연도에 따른 주요 키워드	• 연구 전략 수립을 위한 필요성: 기초과학 분야에서 연구 전략을 수립하는 것은 학계 내 경쟁력 강화 및 연구 성과의 정확한 평가에 도움 • 데이터 분석 결과를 활용하여 전략 수립은 미래의 연구 방향성을 결정하는 데 도움이 될 것으로 기대 • 텍스트 마이닝을 통한 키워드 및 트렌드 파악: 텍스트 마이닝을 활용한 데이터 분석을 통해 주요 키워드와 최근 트렌드를 파악하고, 이를 토대로 전략적인 의사결정을 지원하는 것이 목표

[표21] 기초과학연구원 선정배경

○ 선정기준

- 데이터의 다양성과 품질: 선정된 기관은 다양한 연구 분야에서 풍부한 데이터를 보유 및 데이터의 품질과 정확성이 보장 가능해야 됨

- 연구 성과의 중요성: 선정된 기관의 연구 성과가 기초과학 분야에서 중요하며, 해당 성과가 전략적 의사결정에 영향을 미칠 것으로 기대 가능

- 협력 가능성: 선정된 기관은 다양한 관련 분야 전문가들과의 협력이 가능한 기관이며 협력을 통해 더욱 폭넓고 심도 있는 분석결과를 기대 가능

○ **모형 도출의 목적**

- IBS 연구 성과의 인사이트 제공: 모형 도출은 IBS(기초과학연구원)의 연구 성과에 대한 깊은 이해를 제공하여, 특정 연구 분야에서의 키워드, 트렌드, 및 상호 연관성을 도출하는 것을 목적으로 함
- 연구 전략 수립의 뒷받침: 모형 도출은 연구 전략의 수립에 필요한 정보를 제공하여, 현재의 데이터 기반 의사결정 흐름을 반영하여 IBS가 미래에 어떤 연구 분야에 주력해야 하는지를 결정하는 데 도움
- 향후 연구 방향성 제안: 모형 도출은 단순히 현재의 연구 성과를 분석하는 데 그치지 않고, 미래의 연구 방향성을 제안하는 데에도 목적, 텍스트 마이닝 결과를 토대로 향후 주목받을 것으로 예상되는 주제나 기술 동향을 도출하여 IBS의 미래 비전을 구체화.

○ **시범적용 프로세스**

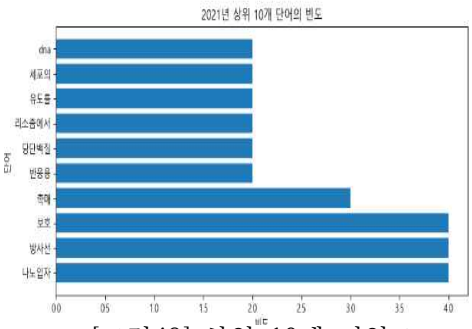
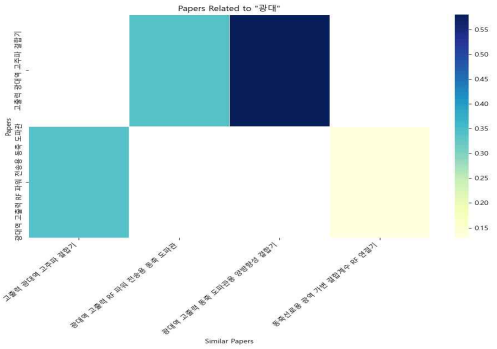
- 데이터 수집 및 전처리: 선정된 기관에서 다양한 연구 성과 데이터를 수집하고, 개인정보 보호에 주의하여 전처리를 수행
- 텍스트 마이닝 분석: 파이썬을 활용하여 데이터를 벡터화하고, 텍스트 마이닝을 통해 키워드와 트렌드를 파악
- 시각화 및 해석: Power BI 툴을 활용하여 CSV 파일을 불러와 시각화를 진행하고, 텍스트 마이닝 결과를 해석
- 심층 텍스트 마이닝 및 키워드 연계분석: 단순 빈도 기반의 텍스트 마이닝을 넘어 TF-IDF나 word2vec 등의 고급 방법론을 활용하여 심층적인 분석을 진행하며, 키워드 간의 연계성을 파악

○ **예상효과**

- 전략적 의사결정 강화: 선정된 기관의 연구 성과를 기반으로 한 데이터 분석은 IBS(기초과학연구원)가 연구 전략을 수립하고 미래 방향을 결정하는 데 큰 도움이 될 것입니다.
- 연구자들의 성과 평가 강화: 결과적으로, 각 연구자는 자신들의 성과를 보다 정확하게 파악하고 평가받을 수 있어, 학계 내 경쟁력을 강화하는 데 이바지할 것으로 예상합니다.

○ 실물모형 목업 프로세스

- 프로세스

	모델1	모델2
개 요	• 기초과학연구원의 지적 재산권 정보를 텍스트 마이닝하여 연구 성과와 관련된 주요 키워드와 트렌드를 파악	• 기초과학연구원의 지적 재산권 정보를 키워드 간 관계나 연계성을 파악하여 연구 주제 간의 연관성 파악
데이터 수집	DATA 공공데이터포털 .GO.KR [그림41] 공공데이터 포털	
데이터 전처리	• 필요 데이터 생성 • 불용어 처리 • 그룹화	• 데이터 중복 제거 • 필요 데이터 생성 • 불용어 처리
사용 기술	• 파이썬 : 고급 프로그래밍 언어로, 간결하면서도 가독성이 좋은 문법 • NLP : 기계가 인간의 언어를 이해하고 처리하게끔 하는 인공지능의 한 분야 • CountVector : 자연어처리에서 텍스트 데이터를 처리하기 위해 사용	• 파이썬 : 고급 프로그래밍 언어로, 간결하면서도 가독성이 좋은 문법 • TF-IDF : 어떤 단어가 얼마나 중요한지를 평가하는 통계적인 방법 • word2vec : 단어를 벡터로 표현하는 방법의 하나로, 단어 간의 의미적 유사성을 보존하는 효과적인 방법 • t-SNE : 차원 데이터를 저차원으로 축소하여 시각화하는 데 사용되는 기술 • 코사인 유사도 : 벡터 간의 코사인 각도를 이용하여 유사성을 측정하는 방법
최종 시각화	 [그림42] 상위 10개 키워드	 [그림43] 연계성 시각화

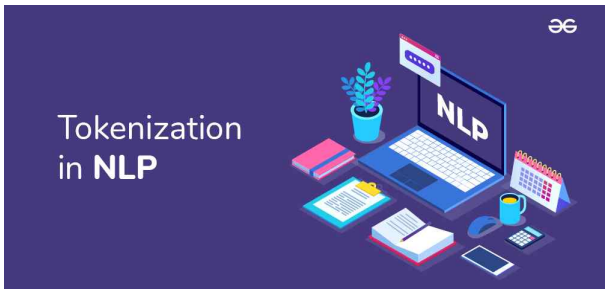
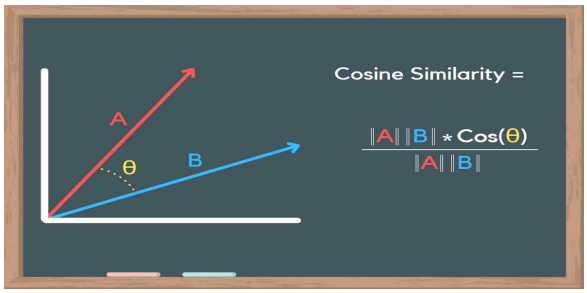
[표22] 실물모형 목업 프로세스

- 활용 데이터

수집데이터	형식	보유기관	비고
기초과학연구원_특허정보시스템	csv	공공데이터 포털	download

[표23] 활용 데이터

- 분석 알고리즘

NLP(Natural Language Processing, 자연어처리)	코사인 유사도
	
[그림44] NLP	[그림45] 코사인 유사도

[표24] 분석 알고리즘

- (모델1) 사용 코드 및 시각화

① 라이브러리 호출

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.family'] ='Malgun Gothic'
plt.rcParams['axes.unicode_minus'] =False
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
import matplotlib.pyplot as plt
import nltk
nltk.download('punkt')
```

② 데이터 불러오기

```
file_path = r'data'
data = pd.read_excel(file_path)
text_data = ' '.join(data['발명의명칭'])
exclude_keywords = ['방법','이의','제조','포함하는','이용한']
```

③ 데이터 전처리

```
vectorizer = CountVectorizer(stop_words=exclude_keywords)
X = vectorizer.fit_transform(example['발명의명칭'])

# 토큰화된 단어 데이터 프레임 전환
word_df = pd.DataFrame(X.toarray(), columns=vectorizer.get_feature_names_out())

# 데이터프레임에 연도 추가
word_df['연도'] = example['연도']
```

④ 데이터 시각화

```
# 각 연도별로 주요 키워드 출력과 가로 막대그래프 생성
fig, axs = plt.subplots(nrows=len(example['연도'].unique()), figsize=(10, 6 * len(example['연도'].unique())))
fig.subplots_adjust(hspace=0.5)

for i, year in enumerate(example['연도'].unique()):
    subset = word_df[word_df['연도'] == year]
    subset_sum = subset.drop('연도', axis=1).sum().sort_values(ascending=False)
    top_10_words = subset_sum.head(10).index
```

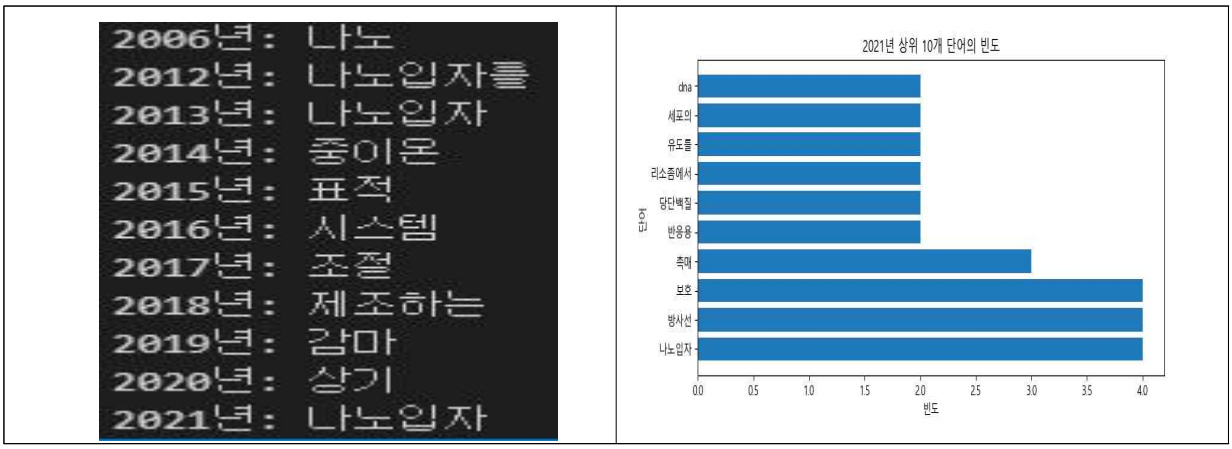
⑤ 데이터 시각화

```
    # 주요 키워드 출력
    print(f'{year}년: {top_10_words[0]}')

    # 가로 막대그래프 생성
    axs[i].barh(top_10_words, subset_sum.head(10))
    axs[i].set_title(f'{year}년 상위 10개 단어의 빈도')
    axs[i].set_xlabel('빈도')
    axs[i].set_ylabel('단어')

plt.show()
```

[표25] 모델1-기초과학연구원 파이썬 코드

- **(모델1) 분석 처리 순서**

1. 파이썬/텍스트마이닝에 대한 라이브러리 호출
2. 데이터 호출 후 불용어(필요 없는 단어) 선정
3. '발명의 명칭' 벡터화를 통해 토큰으로 변환 및 연도 추가
4. 토큰으로 변환된 데이터 연도에 따른 주요 키워드 출력
5. 연도별로 10개의 주요 키워드 그래프 시각화 및 가장 높은 빈출 키워드 출력



[표26] 모델1-결과값 시각화

- **(모델1) 분석결과**

- 텍스트 마이닝을 통해 연구 성과와 관련된 주요 키워드(예: "나노", "감마", "나노입자" 등)가 발견될 것으로 보이며, 최근 연도별로 키워드 트렌드를 파악했을 때, "방사선"과 "나노입자"가 빠른 속도로 증가하고 있는 것으로 관측되는 시각효과 구현

- **(모델2) 사용 코드 및 시각화**

① 라이브러리 호출

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.family'] ='Malgun Gothic'
plt.rcParams['axes.unicode_minus'] =False
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from gensim.models import Word2Vec
from nltk.tokenize import word_tokenize
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
import nltk
nltk.download('punkt')
```

② 데이터 불러오기

```
file_path = 'data'
data = pd.read_excel(file_path)
text_data = ' '.join(data['발명의명칭'])
exclude_keywords = ['방법','이의','제조','포함하는','이용한']

data['출원날짜'] = pd.to_datetime(data['출원일자'])
data['연도'] = data['출원날짜'].dt.year

example = data[['발명의명칭','연도']]
```

③ 벡터화 및 코사인 유사도 적용

```
# TF-IDF 기반 벡터화
tfidf_vectorizer = TfidfVectorizer(stop_words=exclude_keywords)
tfidf_matrix = tfidf_vectorizer.fit_transform(example['발명의명칭'])

# 발명의 명칭 간의 코사인 유사도
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

# word2vec 모델 학습
tokenized_text = [word_tokenize(sentence.lower()) for sentence in example['발명의명칭']]
word2vec_model = Word2Vec(sentences=tokenized_text, vector_size=100, window=5, min_count=1, workers=10)
```

④ 벡터 유사도 계산

```
# 키워드 간의 word2vec 벡터 유사도 계산
word2vec_sim = []
for i in range(len(example['발명의명칭'])):
    for j in range(len(example['발명의명칭'])):
        if i != j:
            vec1 = sum(word2vec_model.wv[word] for word in tokenized_text[i])
            vec2 = sum(word2vec_model.wv[word] for word in tokenized_text[j])
            similarity = cosine_similarity([vec1], [vec2])[0][0]
            word2vec_sim.append((i, j, similarity))
```

⑤ TF-IDF 연계성

```
# 결과 출력 및 저장
print("TF-IDF 기반 키워드 간의 연계성:")
tfidf_results = []  # List to store results for CSV

for i in range(len(cosine_sim)):
    keyword = example['발명의명칭'][i]
    print(f"{keyword}:")

    similar_keywords = [(example['발명의명칭'][j], round(cosine_sim[i][j], 2)) for j in range(len(cosine_sim[i])) if i != j]
    similar_keywords.sort(key=lambda x: x[1], reverse=True)

    for j, (similar_keyword, similarity) in enumerate(similar_keywords[:5]):
        print(f"  - {similar_keyword}: {similarity}")
        tfidf_results.append([keyword, similar_keyword, similarity])

    print("\n")
```

⑥ 특허 중복 제거 및 저장

```
# 중복제거 후 csv 파일로 저장
tfidf_csv_filename = '논문간_유사도_중복제거.csv'
df_tfidf_results = pd.DataFrame(tfidf_results, columns=['논문', '유사 논문', '유사성/연관성'])
df_tfidf_results.drop_duplicates(subset=['유사 논문'], keep='first', inplace=True)
df_tfidf_results.to_csv(tfidf_csv_filename, index=False, encoding='utf-8')

print(f"\nTF-IDF 결과가 '{tfidf_csv_filename}' 파일로 저장되었습니다.")
```

⑦ TSNE 2차원 축소

```
# t-SNE를 사용하여 2차원으로 축소
tsne_model = TSNE(n_components=2, random_state=42, init='random')
tsne_values = tsne_model.fit_transform(tfidf_matrix)

# 시각화
for i, (x, y) in enumerate(tsne_values):
    if i < len(example):  # DataFrame의 길이를 확인
        plt.scatter(x, y)

plt.title('Word2Vec 연계성 시각화 (t-SNE)')
plt.xlabel('t-SNE 차원 1')
plt.ylabel('t-SNE 차원 2')
plt.show()
```

⑧ 저장 파일 불러오기

```
# csv 파일 읽기
csv_filename = '논문간_유사도_중복제거.csv'
df_tfidf_results = pd.read_csv(csv_filename)

# 원하는 키워드
target_keyword = '고출력 광대역 고주파 결합기'

# Extract data for the target keyword
keyword_data = df_tfidf_results[df_tfidf_results['논문'].str.contains(target_keyword)]
```

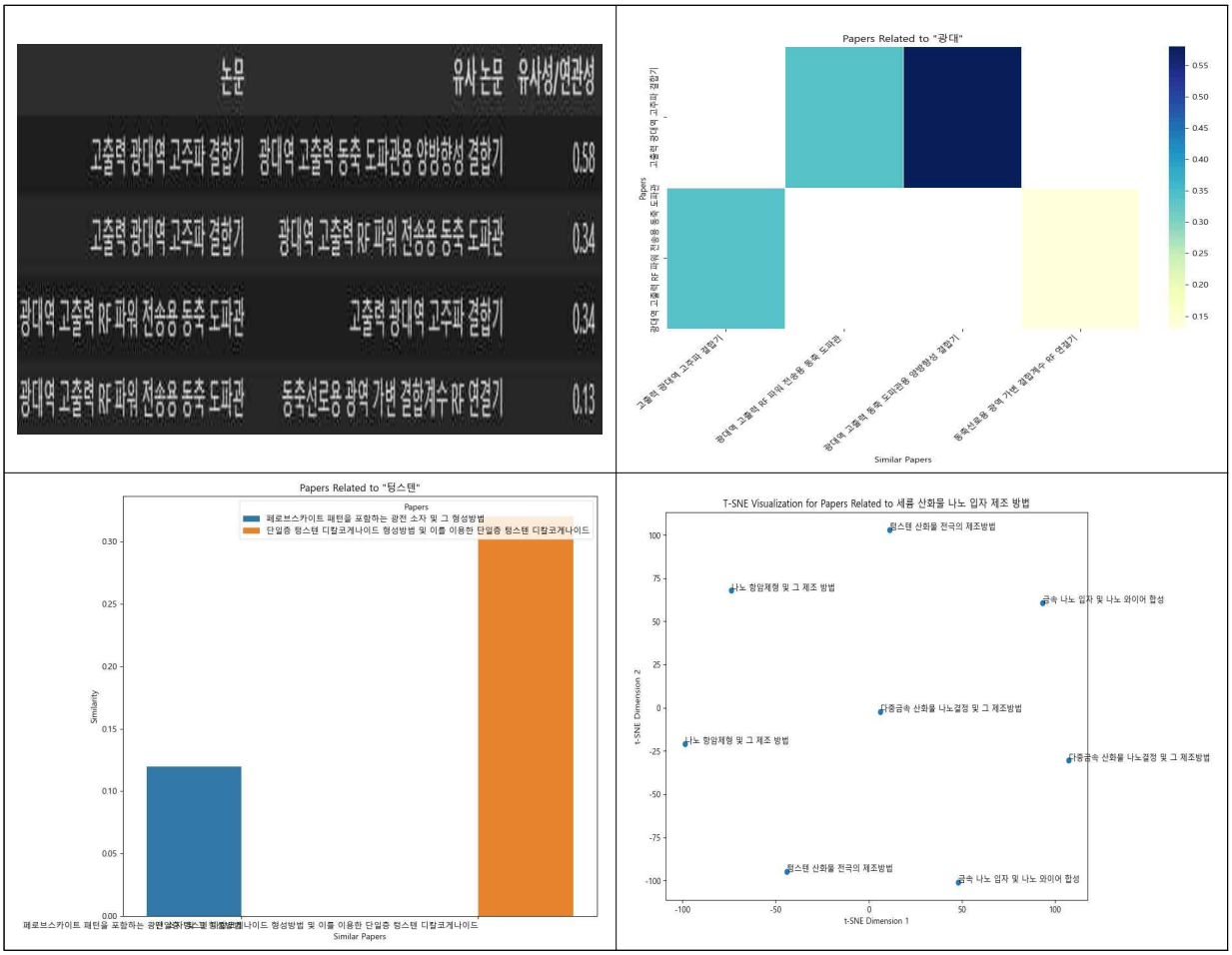
⑨ 시각화

```
# Pivot the data for heatmap
heatmap_data = keyword_data.pivot(index='논문', columns='유사 논문', values='유사성/연관성')

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(heatmap_data, cmap='YlGnBu', annot=False, fmt=".2f", linewidths=0.5)
plt.title(f'Papers Related to "{target_keyword}"')
plt.xlabel('유사 논문')
plt.ylabel('논문')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better readability
plt.tight_layout()  # Adjust layout to prevent overlap
plt.show()
```

[표27] 모델2-기초과학연구원 파이썬 코드

- (모델2) 분석 처리 순서

1. 파이썬/NLP(자연어처리)에 대한 라이브러리 호출
2. 데이터 호출 후 컬럼 생성 및 전처리
2. '발명의 명칭' TF-IDF 기반 벡터화 및 코사인 유사도 적용
4. 키워드 간의 word2vec 벡터 유사도 계산
5. TF-IDF 기반 키워드 간의 연계성 출력 (발명의 명칭, 유사한 명칭, 유사성)
6. 중복 제거 후 CSV 파일로 저장
7. t-sne 2차원으로 축소 및 연계성 시각화
8. 파일 불러온 후, 원하는 키워드 입력 후 키워드에 따른 데이터 추출
9. 추출한 데이터 기반 히트맵 시각화



[표28] 모델2-결과값 시각화

- **(모델2) 분석결과**
- (심층 텍스트 마이닝) 단순 빈도 기반의 키워드 추출을 진행을 넘어서 TF-IDF나 word2vec 등의 고급 방법론을 활용한 텍스트 마이닝 진행
- (키워드 연계분석) 키워드 간의 관계나 연계성을 파악하여 연구 주제나 트렌드 간의 연관성을 히트맵으로 시각화하여 파악

3. 분석결과 사후관리방안

○ **(분석결과의 검토 및 정정)** 초기 분석결과를 주기적으로 재검토하여 오류나 누락된 정보를 확인하고 수정, 새로운 데이터나 방법론이 나타날 경우, 분석 업데이트 수행

○ **(데이터 보관 및 백업)** 분석에 사용된 원시 데이터와 처리된 데이터를 안전하게 보관, 주기적인 백업과 복원 절차를 통해 데이터 손실 방지

○ **(분석 도구 및 방법론의 업데이트)** 분석 도구의 버전이나 사용된 방법론에 변경이 생길 경우, 해당 변화를 반영하여 분석을 재진행하거나 확인 필요

○ **(분석결과의 지속적 모니터링)** 분석결과가 현실 세계의 변화에 얼마나 적응하는지를 지속해서 모니터링, 이를 통해 분석의 유효성과 현장 적용성을 지속해서 확인

○ **(커뮤니케이션 강화)** 분석결과에 대한 피드백을 주기적으로 수집, 이를 통해 분석결과의 적용과 관련된 문제점이나 개선사항 파악 및 대응

○ **(재활용 및 확장성 고려)** 현재의 분석결과를 기반으로 다른 분야나 상황에서의 활용 가능성을 탐색, 분석 모델이나 방법론의 확장성을 고려하여 다양한 적용 사례 개발에 활용

○ **(문서화 및 트레이닝)** 분석 과정, 결과, 방법론 등을 체계적으로 문서화하여 지식의 전달 및 재사용 촉진에 활용