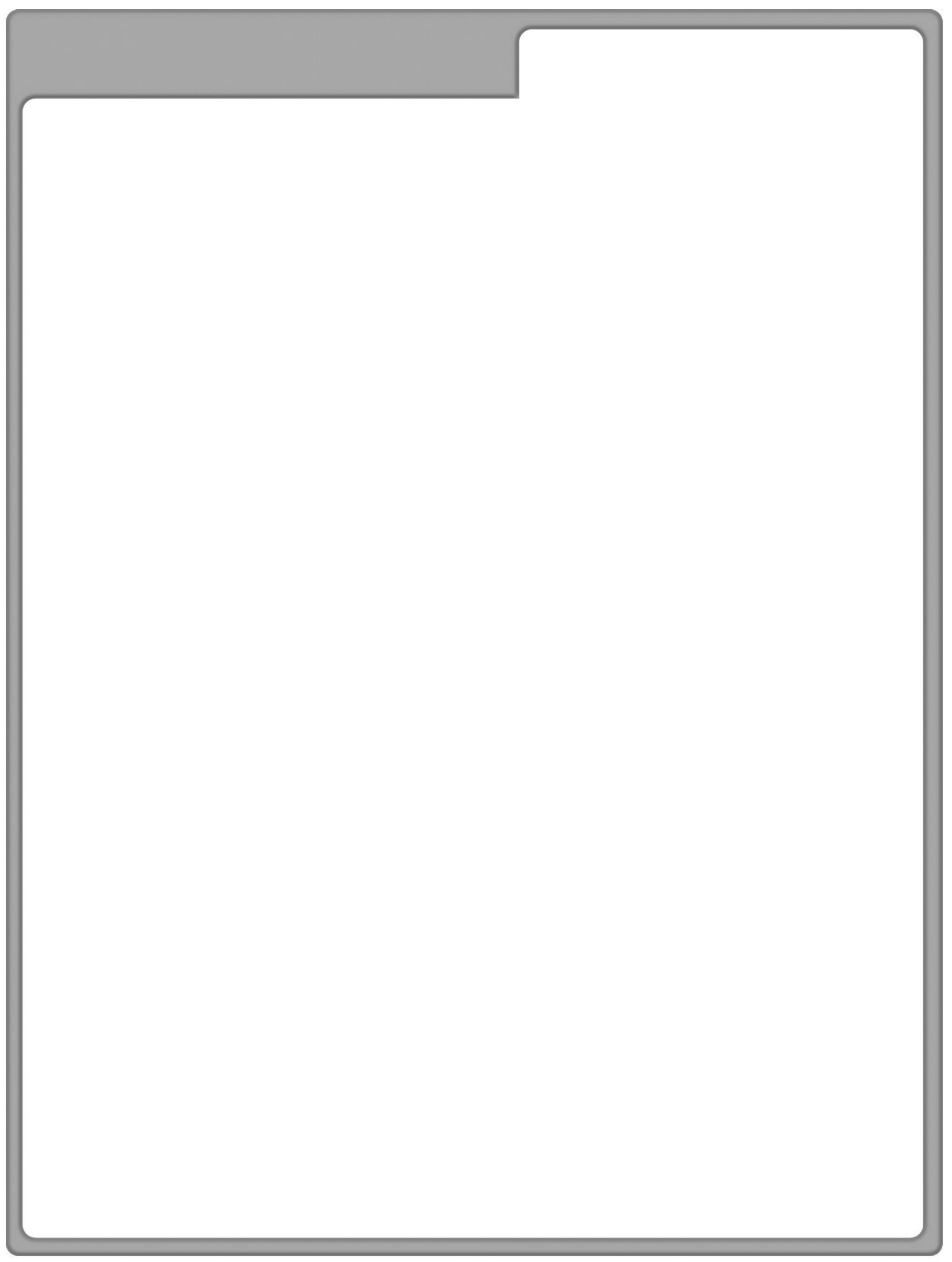
|  |
| --- |
| 프로젝트기반 빅데이터 분석 과정 8기  프로젝트 계획서 |

2023.12.28

|  |  |  |
| --- | --- | --- |
| **프로젝트 명** | **강화학습을 이용한 국립도서관 도서 추천 시스템** | |
| **개발 기간** | **2023.12.22.~2024.2.6** | |
| **팀 장** | **태정수** | **xowjdtn@gmail.com** |
| **팀 원** | **김대양** | **kenken2021320012@gmail.com** |
| **서동윤** | **sdy9716@gmail.com** |



**Ⅰ. 프로젝트 개요**

1. 프로잭트 개요 1

2. 추진배경 및 필요성 1

3. 개발범위 2

4. 기대효과 3

**Ⅱ. 추진 방안**

1. 추진목표 4

2. 추진전략 4

3. 추진체계 5

4. 추진일정 5

**Ⅲ. 개발 내용**

1. 개발내용 개요 6

2. 상세 요구사항 7

**Ⅳ. 평가방법**

1. 평가방법 개요 10

2. 세부 평가 항목 10

3. 검수 방법 11

목 차

|  |  |  |
| --- | --- | --- |
| **Ⅰ** |  | **프로젝트 개요** |

|  |  |
| --- | --- |
| 1 | 프로젝트 개요 |

❍ 프로젝트 명: **강화학습을 이용한 국립도서관 도서 추천 시스템**

❍ 개발 기간: **2023.12.22.~2024.2.6**

❍ 주요 개발 내용

- 수집·정제된 데이터 분석

- 데이터 탐색적 분석 및 결과 도출

- 통계적 데이터 분석

- 강화학습을 통한 도서 추천 모델 구현

- 활용 방안 제시

|  |  |
| --- | --- |
| 2 | 추진배경 및 필요성 |

**① 추진배경**

❍ 기존의 도서 추천 시스템은 이용자의 선호도나 도서관의 소장 도서 목록 등을 기반으로 도서를

추천하는 방식이었다. 이러한 방식은 이용자의 관심사나 취향에 따라 추천 도서의 적중률이

떨어질 수 있다는 단점이 있다.

❍ 도서관에 많은 도서가 있어 항상 갈 때마다 고민을 하게 되고, 어떤 책을 읽어야할지 선택하기

어려움을 느낀다.

**② 필요성**

❍ 이용자가 원하는 도서를 보다 쉽게 찾을 수 있도록 도와주어, 도서관 이용률을

증가시킬 수 있다.

❍ 이용자의 관심사와 취향에 맞는 도서를 보다 정확하게 추천하여, 이용자의 만족도를

향상시킬 수 있다.



|  |  |
| --- | --- |
| 3 | 개발범위 |

**① 빅데이터 자료 수집 방법**

1. 공공 데이터 활용

- 국립중앙도서관 대출데이터(문화빅데이터플랫폼)

- 국립중앙도서관 대출상승 도서목록(문화빅데이터플랫폼)

- 국립중앙도서관 도서별 키워드 정보(문화빅데이터플랫폼)

- 국립중앙도서관 베스트 셀러(문화빅데이터플랫폼)

- 국립중앙도서관 연령별 키워드 검색데이터(문화빅데이터플랫폼)

- 국립중앙도서관 인기 대출데이터(문화빅데이터플랫폼)

**② 빅데이터 자료 정제 방법**

1. 1번 또는 2번 이하의 빈도를 가진 대출 데이터는 삭제한다 (선호도 또는 이용객의 특징을 잘 반영 못 할 수도 있기때문).
2. 데이터분석을 통해서 , 학습에 사용이 가능할 만한 의미있는 특성을 선택한다.
3. 이용객의 특징을 반영 할 수 있도록 , 데이터의 형태를 이용객의 ID를 기본키로 기준을 설정하여 , 필요한 특성을 나열하는 방식으로 데이터를 가공.
4. 실시간 모델 업데이트를 위하여 , 데이터 가공 파이프라인을 구축한다.

**③ 분석 방법 및 구현 방법**

1. 탐색적 데이터 분석

- 데이터의 기술 통계량을 검토하거나, 데이터의 종류(명목형, 순서형 등)에 따른 시각화를 통해서 해당 데이터를 파악

* 상관분석을 통해 독립변수들 간에 강한 상관관계가 존재하는 현상을 파악하여 처리

2. 통계적 데이터 분석

- 데이터 분석에서의 패턴이나, 새롭게 만든 파생변수의 영향력 등에 대해서, 가설을

세우고 가설검정을 한다.

3. 모델 구현

- 학습을 한 뒤에 가장 영향력이 있는 모델을 선정, 그 후에 모델의 성능이 오르는지, 학습을 모니터링 하면서 모델의 하이퍼패러미터를 튜닝하는 작업을 한다.

|  |  |
| --- | --- |
| 4 | 기대효과 |

❍ 이용자가 원하는 도서를 찾아주어 이용자의 만족도를 향상시킬 수

있고, 도서관 이용률을 증가시킬 수 있다.

❍ 도서 추천 시스템은 이용자의 선호도를 분석하여, 도서관의 소장

도서 목록을 보다 효율적으로 관리할 수 있습니다.



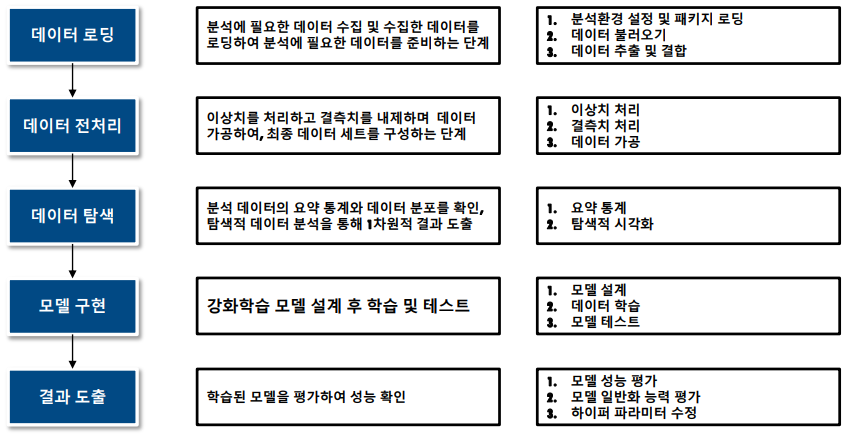
|  |  |  |
| --- | --- | --- |
| **Ⅱ** |  | **추진 방안** |

|  |  |
| --- | --- |
| 1 | 추진목표 |

❍ 도서 추천 시스템을 이용하여 선호도 요인 추출

❍ 이용자의 관심사와 취향에 맞는 도서를 정확하게 추천하여 시스템 개발

|  |  |
| --- | --- |
| 2 | 추진전략 |



|  |  |
| --- | --- |
| 3 | 추진체계 |

❍ 참여 연구원 및 Github 정보

|  |  |  |  |
| --- | --- | --- | --- |
| 성명 | 태정수 | 생년월일 | 1999.07.03 |
| 소속/과정 | 선문대학교 AI소프트웨어학과 /학사 | | |
| 연락처 | 010-3925-7132 | 이메일 | xowjdtn@naver.com |
| GitHub ID | taejungsu | | |
| 저장소 URL | https://github.com/K-Digital-Project01/kkh3363 | | |

|  |  |  |  |
| --- | --- | --- | --- |
| 성명 | 김대양 | 생년월일 | 2002.01.24 |
| 소속/과정 | 선문대학교 AI소프트웨어학과 /학사 | | |
| 연락처 | 010-7195-7238 | 이메일 | **kenken2021320012@gmail.com** |
| GitHub ID | Dae-yangKim | | |
| 저장소 URL | https://github.com/K-Digital-Project01/kkh3363 | | |

|  |  |  |  |
| --- | --- | --- | --- |
| 성명 | 서동윤 | 생년월일 | 1997.02.12 |
| 소속/과정 | 백석대학교 정보보호학과 / 학사 | | |
| 연락처 | 010-9105-2766 | 이메일 | sdy9716@gmail.com |
| GitHub ID | dongyunseo | | |
| 저장소 URL | https://github.com/K-Digital-Project01/kkh3363 | | |

|  |  |
| --- | --- |
| 4 | 추진일정 |

❍ 사업기간 :2023.12.22 ~ 2024.2.6

| **추진과제** | **12월 22일** ~ **2월 6일** | | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1주** | | | | **2주** | | | | | **3주** | | | | | **4주** | | | | |
| **1. 모델링 과정** | | | | | | | | | | | | | | | | | | | |
| ▪ 데이터 수집 및 파악 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 데이터 가공 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 데이터 분석 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 특성 엔지니어링 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 모델링 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **2. 결론 도출 및 정리** | | | | | | | | | | | | | | | | | | | |
| ▪ 모델 평가 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 정리 자료 제작 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **3. 보고회 등** | | | | | | | | | | | | | | | | | | | |
| ▪ 계획서 보고/중간/최종보고회 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ▪ 검수 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

※ 세부 일정은 사업추진상황에 따라 변경 가능

|  |  |  |
| --- | --- | --- |
| **Ⅲ** |  | **개발 내용** |

|  |  |
| --- | --- |
| 1 | 개발내용 개요 |

❍ 주요 프로젝트 내용

- 수집·정제된 데이터 분석

- 데이터 탐색적 분석 및 결과 도출

- 통계적 데이터 분석

- 강화학습을 통한 도서 추천 모델 구현

- 활용 방안 제시

❍ 공공 데이터

- 국립중앙도서관 대출데이터(문화빅데이터플랫폼)

- 국립중앙도서관 대출상승 도서목록(문화빅데이터플랫폼)

- 국립중앙도서관 도서별 키워드 정보(문화빅데이터플랫폼)

- 국립중앙도서관 베스트 셀러(문화빅데이터플랫폼)

- 국립중앙도서관 연령별 키워드 검색데이터(문화빅데이터플랫폼)

- 국립중앙도서관 인기 대출데이터(문화빅데이터플랫폼)

❍종합한 정보를 토대로 요약 시각화(대화형 대시보드) 진행

|  |  |
| --- | --- |
| 2 | 상세 요구사항 |

| **본류** | **번호** | **요구사항 명칭** |
| --- | --- | --- |
| 분석과제 파악 | IAT-001 | 과제 정의 및 요구사항 분석 |
| IAT-002 | 분석 데이터 확보 및 정의 |
| 데이터 분석 | DA-001 | 데이터 로딩 및 전처리 |
| DA-002 | EDA(탐색, 분석, 시각화) |
| DA-003 | 데이터 분석 |
| 기능 요구사항 | SFR-001 | 선별된 시각화 요약 대시보드 개발 |
| 프로젝트 관리 요구사항 | PMR-001 | 산출물 종류 및 보고종류, 시기 |

**(1) 분석과제 파악**

| **요구사항 고유번호** | | IAT-001 |
| --- | --- | --- |
| 요구사항 분류 | | 분석과제 파악 |
| 요구사항 명칭 | | 과제 정의 및 요구사항 분석 |
| 요구사항  상세설명 | 정의 | 분석 대상 정의 및 분석 방법 등 세부 시나리오 정의 |
| 세부내용 | • 분석과제 목표 및 범위 설정  - 각 과제의 추진 목표 및 추진 방향 설정  - 분석 방법 및 기법 정의  • 분석 시나리오 정의  - 분석 요구사항에 적합한 분석 주제 및 대상 정의  - 분석 방법 및 기법 정의  - 분석 주제 및 대상을 선정하고 이를 수행하기 위한 절차 수립 |

| **요구사항 고유번호** | | IAT-002 |
| --- | --- | --- |
| 요구사항 분류 | | 분석과제 파악 |
| 요구사항 명칭 | | 분석 데이터 확보 및 정의 |
| 요구사항  상세설명 | 정의 | 분석 대상 데이터 확보, 세부 현황 및 분석 범위, 분석 흐름도 작성 |
| 세부내용 | • 과제 수행에 필요한 분석 데이터 확보 및 정의  - 국립중앙도서관 대출데이터(문화빅데이터플랫폼)  - 국립중앙도서관 대출상승 도서목록(문화빅데이터플랫폼)  - 국립중앙도서관 도서별 키워드 정보(문화빅데이터플랫폼)  - 국립중앙도서관 베스트 셀러(문화빅데이터플랫폼)  - 국립중앙도서관 연령별 키워드 검색데이터(문화빅데이터플랫폼)  - 국립중앙도서관 인기 대출데이터(문화빅데이터플랫폼)  ※ 과업 수행 중 필요데이터 추가 발생 시 협의를 통해 최대한 반영  • 데이터 분석 흐름 작성 및 분석 데이터 정의 (Work flow)  - 각 데이터의 생성부터 활용단계에 이르는 과정 도식화  - 분석에 적합한 데이터를 생성하기 위한 분석 데이터 정의 |

**(2) 데이터 분석**

| **요구사항 고유번호** | | DA-001 |
| --- | --- | --- |
| 요구사항 분류 | | 데이터 분석 |
| 요구사항 명칭 | | 데이터 로딩 및 전처리 |
| 요구사항  상세설명 | 정의 | 데이터 로딩 및 분석/시각화 기법에 따른 전처리 |
| 세부내용 | • 데이터 로딩  - 확보한 데이터셋 로딩  - 데이터 추출 및 결합  • 분석/시각화 기법 선정  - 각 데이터에 적합한 시각화/분석 기법 고려 후 선정  • 선정한 분석/시각화 기법에 따른 전처리  - 선정한 분석/시각화 기법에 따른 데이터 전처리 시행  - 이상치, 결측치 처리  - 파생변수 생성 |

| **요구사항 고유번호** | | DA-002 |
| --- | --- | --- |
| 요구사항 분류 | | 데이터 분석 |
| 요구사항 명칭 | | EDA(탐색, 분석, 시각화) |
| 요구사항  상세설명 | 정의 | 탐색적 데이터 분석 시행 |
| 세부내용 | • 데이터 특성 파악  - 통계형/시계열 특성 파악  - 적절한 분석/시각화 기법 도출  • 데이터 분석/시각화  - 시행할 기법에 따른 추가적 데이터 전처리  - 분석/시각화 시행  • EDA 결과 기술  - 분석/시각화 과정에서 습득한 정보 정리  - 분석/시각화 결과에서 유추 가능한 의미 파악 |

| **요구사항 고유번호** | | DA-003 |
| --- | --- | --- |
| 요구사항 분류 | | 데이터 분석 |
| 요구사항 명칭 | | 데이터 분석 |
| 요구사항  상세설명 | 정의 | 필요 데이터 분석 시행 |
| 세부내용 | • 통계적 데이터 분석  - 데이터 분석에서의 패턴이나, 새롭게 만든 파생변수의 영향력 등에 대해서, 가설 세우고 가설검정  • 결과 기술  - 분석/시각화 과정에서 습득한 정보 정리  - 분석/시각화 결과에서 인사이트 도출 |

**(3) 기능 요구 사항**

| **요구사항 고유번호** | | SFR-001 |
| --- | --- | --- |
| 요구사항 분류 | | 기능 요구사항 |
| 요구사항 명칭 | | 인터랙티브 대시보드 구현 |
| 요구사항  상세설명 | 정의 | 대화형 대시보드를 프로젝트 산출물로서 구현 |
| 세부내용 | • 시각화 선별  - EDA 및 분석 결과 시각화 중 필수적인 정보 선별  • 인터랙티브 대시보드 구현  - 수집한 데이터를 활용하여 시각화  - 데이터 특성을 고려한 그래프 선정  - 그래프와 레이아웃을 이용하여 대시보드 구현  - Plotly 라이브러리를 활용하여 대화형 시각화로 변환 |

**(4) 프로젝트 관리 요구사항**

| **요구사항 고유번호** | | PMR-001 |
| --- | --- | --- |
| 요구사항 분류 | | 관리 요구사항 |
| 요구사항 명칭 | | 산출물 종류 및 보고종류, 시기 |
| 요구사항  상세설명 | 세부내용 | • 분석대상 분석기법 결과해석 및 검토 등의 중요 추진 내용은 프로젝트 매니저의 컨펌 이후 작성  (실무회의는 별도 협의하여 구성 운영함)  • 보고종류 및 시기  그림입니다. 원본 그림의 이름: clip_image001.png 원본 그림의 크기: 가로 566pixel, 세로 152pixel  • 보고서 작성 내용  - 일일 개발 일지는 업무 및 개발 내역 상세히 기술  - 회의록은 안건과 회의 결과 진행 상황 등을 기록  - 중간 및 최종보고 시기는 추진일정 중 상호 협의하여 결정  • 모든 산출물에 대한 안정적으로 형상관리 수행 |
| 산출물 | | 일일 개발일지, 회의록, 계획서, 최종보고서 등 |

|  |  |  |
| --- | --- | --- |
| **Ⅳ** |  | **평가 방법** |

|  |  |
| --- | --- |
| 1 | 평가 방법 개요 |

| **평가 항목** | **평가요소(기준)** | **배점** |
| --- | --- | --- |
| 추진목표 달성도 | ❍ 추진목표  - 연구 추진목표의 달성도를 전체(%)로 계산하여 배점 | 50 |
| 요구사항 달성도 | ❍ 요구사항   * 요구사항의 달성도를 전체(%)로 계산하여 배점 | 50 |
| **합계** | | **100** |

|  |  |
| --- | --- |
| 2 | 세부 평가 항목 |

1. **평가항목 및 배점**

| **평가 부문** | **평가 항목** | **평가요소** | **배점** | |
| --- | --- | --- | --- | --- |
| 전략 및  방벙론 | 프로젝트 절차 | * 전처리, EDA, 분석, 결과 도출 및 활용 방안 절차의 적절성 | 15 | 30 |
| 분석/시각화 | - 적용기술 및 사용 도구의 적절성  - 기술 사용법의 타당성(분석 의도에 타당한지) | 15 |
| 기술 및  기능 | 데이터 분석/시각화 | * 데이터 현황조사, 수집, 전력의 구체성 * 데이터 수집, 정제, 변환, 전환 역량 * 데이터 처리방법의 타당성(결측치, 이상치, 파생변수 등) * 분석기법의 적정성, 데이터 분석의 효과성 | 15 | 35 |
| 시스템 구현 | - 시각화 결과 대시보드로 구현  - 인터랙티브 기능 구현  - 분석 결과 및 프로젝트 전체 의의를 일목요연하게 표현 | 10 |
| 구현 테스트 | -구현한 기능 정상 작동 테스트   * 레이아웃 및 디자인 적절성 * 직관성 * 정보 전달성 | 10 |
| 프로젝트 관리 | 일정계획 | - 세부활동 도출 및 기간의 타당성(스케줄링이 적절한지)   * 세부활동 배열의 합리성(작업 프로세스가 적절했는지) | 15 | 15 |
| 결론 | 인사이트  도출 및 활용 방안 | -도출한 결론의 신뢰성 여부  -결론을 활용해 제시한 방안의 적절성 | 20 | 20 |
| **합계** | | | **100** | |

|  |  |
| --- | --- |
| 3 | 검수 방법 |

❍ 추진 목표 및 요구사항 달성도로 평가

❍ 프로젝트 산출물의 완성도 고료

❍ 모든 산출물 및 프로젝트 진행은 팀장 담당 하에 검수