| **빅데이터 분석 정의서** |
| --- |

| **주제 :** | **AI기술 도입 희망 기업의 주요사업별**  **최적의 데이터셋 및 AI 모델 추천 서비스** |
| --- | --- |

| **2022.12.15** |
| --- |

**Ⅰ. 개요**

1. 아이디어 주제

: AI기술 도입 희망 기업의 주요사업별 최적의 데이터셋 및 AI 모델 추천 서비스

(세상의 모든 AI)

2. 개발 목표

: 웹 서비스를 통한 솔루션 패키지 추천 및 AI 모델 개발

3. 개발 내용

: 데이터 구축

: 고객들이 입력할 데이터 형태 확인(텍스트)

: 데이터 양이 부족한 경우 확인한 데이터를 기준으로 가상데이터 추가

: 데이터 전처리 및 인공지능 모델 개발

: 입력된 데이터(텍스트)를 토대로 Word2vec을 활용하여 패키지 추천

: 패키지 추천 후 입력된 데이터를 토대로 AI기술 추천 및 개발

**Ⅱ. 기능별 빅데이터 분석 명세서**

| **기능명** | 데이터셋 추천 패키지 |
| --- | --- |
| 1. **데이터 준비** | |
| **데이터 정의** | AI Hub 내 384개의 데이터셋 |
| **데이터 획득 방법** | AI Hub에서 데이터셋 제목 추출 |
| **2. 전처리** | |
| **전처리 과정** | AI Hub에서 데이터셋 제목을 추출하여 엑셀파일에 정리  통계청 및 국세청에서 사용하는 실제 업종명 대, 중, 소분류 가져오기 |
| **3. 모델 생성 및 학습** | |
| **모델링 목표** | 업종명과 가장 유사한 데이터셋을 word2vec을 사용하여 찾기 |
| **모델링 가능 알고리즘** | word2vec, 코사인 유사도, FastText |
| **학습** | 코사인 유사도, FastText, word2vec을 모두 사용하여 업종명과 가장 유사한 데이터셋을 찾도록함. |
| **4. 검증** | |
| **모델링 검증 방안** | 임의로 업종명을 입력하여 예측 모델 검증증 |
| **모델링 평가 결과** | 정확도를 나타낼 명확한 지표가 없음음 |

| **기능명** | 주요사업별 최적의 데이터셋 및 AI 모델 추천 서비스 |
| --- | --- |
| 1. **데이터 준비** | |
| **데이터 정의** | AI Hub 내 384개의 데이터셋 |
| **데이터 획득 방법** | AI Hub에서 데이터셋들을 크롤링  AI Hub에서 데이터셋 제목 추출  커리어 사이트에서 회사이름 및 주요사업 내용 크롤링 |
| **2. 전처리** | |
| **전처리 과정** | AI Hub에서 데이터셋 제목 및 사용한 AI 모델을 추출하여 엑셀파일에 정리  통계청 및 국세청에서 사용하는 실제 업종명 대, 중, 소분류 가져오기 |
| **3. 모델 생성 및 학습** | |
| **모델링 목표** | 업종명과 커리어 사이트내에 있는 주요사업의 내용을 가지고 업종과 가장 유사도가 높은 AI 모델 추천 |
| **모델링 가능 알고리즘** | Ko-Sentence-Transformers, Ko-sroberta-Multitask |
| **학습** | 크롤링한 회사 홈페이지 내의 헤더, 커리어 사이트 내의 회사이름과 주요사업 내용과 가장 유사한 AI 모델을 찾도록 학습. |
| **4. 검증** | |
| **모델링 검증 방안** | 모델에 임의의 회사명, 주요사업 내용을 넣어보기 |
| **모델링 평가 결과** | Ko-sroberta-Multitask의 정확도 도식화 |

| **기능명** | 올인원 패키지 |
| --- | --- |
| 1. **데이터 준비** | |
| **데이터 정의** | AI Hub 내 384개의 데이터셋 |
| **데이터 획득 방법** | AI Hub에서 데이터셋들을 크롤링  AI Hub에서 데이터셋 제목 추출  커리어 사이트에서 회사이름 및 주요사업 내용 크롤링 |
| **2. 전처리** | |
| **전처리 과정** | AI Hub에서 데이터셋 제목 및 사용한 AI 모델을 추출하여 엑셀파일에 정리  통계청 및 국세청에서 사용하는 실제 업종명 대, 중, 소분류 가져오기 |
| **3. 모델 생성 및 학습** | |
| **모델링 목표** | 업종명과 커리어 사이트내에 있는 주요사업의 내용을 가지고 업종과 가장 유사도가 높은 데이터셋과 AI 모델 추천 |
| **모델링 가능 알고리즘** | Ko-Sentence-Transformers, Ko-sroberta-Multitask |
| **학습** | 크롤링한 회사 홈페이지 내의 헤더, 커리어 사이트 내의 회사이름과 주요사업 내용과 가장 유사한 데이터셋과 AI 모델을 찾도록 학습. |
| **4. 검증** | |
| **모델링 검증 방안** | 모델에 임의의 회사명, 주요사업 내용을 넣어보기 |
| **모델링 평가 결과** | Ko-sroberta-Multitask의 정확도 도식화 |