|  |  |
| --- | --- |
| **오픈 소스기반**  **빅데이터 분석**  **프로그래밍 과정** | **Final Project 수행계획서** |

**1.팀정보**

|  |  |  |  |
| --- | --- | --- | --- |
| **팀 명** | | 1조 | |
| **팀구성** | | **성 명** | **소 속** |
| **1 팀** | **팀장** | 남명민 |  |
| **팀원1** | 안상후 |  |
| **팀원2** | 류재욱 |  |
| **팀원3** | 김남혁 |  |

**2. 프로젝트 정보**

|  |  |
| --- | --- |
| **주제영역** | AI 및 웹스크랩핑을 활용한 포털 정보 신뢰도 분석 |
| **기술분야** | 웹스크랩핑, 크롤링, AI, Django, MySQL, HTML5, Docker |
| **성과목표** | 분석결과 보고 및 결론 도출 |
| **프로젝트명** | 포털정보 신뢰도 딥러닝 분석 |
| **수행기간**  **(예상)** | 2020년1월5일~2020년1월27일 |

**3. 수행계획**

**1.프로젝트 개요**

1. **추진배경 및 필요성**
2. 포털/검색 엔진을 통해 정보를 검색하는 경우 원하지 않는 광고성 글로 인해 시간이 낭비되는 경험
3. 포털이나 검색엔진의 콘텐츠 들이 키워드 별로 얼마나 광고 성격을 가지고 있는 지, 정보들이 어느 정도의 신뢰성을 가지고 있는지 알아보기 위함
4. 본 시스템은 포털이나 검색엔진의 콘텐츠들이 키워드 별로 얼마나 광고 성격을 가지고 있는지, 정보들이 어느 정도의 신뢰성을 가지고 있는지 알아보기 위한 서비스를 목표 .

**나.프로젝트 세부내용**

|  |  |  |
| --- | --- | --- |
| **기능** | | **세부내용** |
| **1** | **데이터 수집** | 네이버 블로그들의 공통 요소 탐색 |
| **2** | **데이터 수집** | 네이버 블로그 본문내용 웹스크랩핑(이미지, 본문 등) |
| **3** |  |  |
| **4** |  |  |

**2.장비(기자재/프로그램) 활용 계획**

**가. 필요 기자재(기자재/프로그램/클라우드서버)**

|  |  |  |  |
| --- | --- | --- | --- |
| **번호** | **품목** | **활용계획** | **비고** |
| **1** | **데이터 수집** | **포털 및 검색엔진 키워드 검색 URL 패턴 수집**  **Request, Response 처리를 통해 데이터를 수집하고, DB에 저장** |  |
| **2** | **딥러닝 모델링** | **수집된 데이터의 자연어 처리 딥러닝 분석 모델링을 통해**  **분석결과 도출**  **광고성 인지 실제 정보성인지 여부를 파악하기 위한 팩터를**  **도출하고 딥러닝 모델링 결과에 포함해 정확도 분석** |  |
| **3** | **키워드 신뢰성 웹서비스** | **키워드별로 정보 신뢰성 정확도를 Django 웹서비스를 통해**  **시각화 서비스** |  |
| **4** | **R 분석 리포트** | **데이터 수집 및 모델링을 통해 도출된 데이터 통해 유의미한**  **통계결과를 도출하고, 분석결과를 워드 클라우드 및 그래프를**  **통해 시각화** |  |

**3. 프로젝트 수행내용**

**가.팀원 업무분담**

|  |  |  |  |
| --- | --- | --- | --- |
| **번호** | **구분** | **이름** | **담당업무** |
| 1 | 팀장 | 남명민 | 팀장, 코드 작성/관리 |
| 2 | 팀원1 | 안상후 | 데이터 처리(전처리, 정제, 분석) |
| 3 | 팀원2 | 류재욱 | 웹설계, 웹디자인, 시각화 |
| 4 | 팀원3 | 김남혁 | 데이터 수집, 자료 검색 |

**나.프로젝트 추진일정**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **구분** | **추진내용** | **추진 일정** | | | | |
| **1일** | **2일~5일** | **6일~10일** | **10일~20일** | **발표** |
| **계획** | 역할분담 및 계획 | ○ |  |  |  |  |
| **데이터수집** | 특정 사이트 데이터  웹그롤링 이용하여 수집 |  | ○ |  |  |  |
| **데이터 정제** | 파이썬의 함수를 이용하여  데이터 저장 및 정제 |  |  | ○ |  |  |
| **DJANGO** |  |  |  |  | ○ |  |
|  |  |  |  |  |  |  |

**4. 성과목표 및 기대효과**

|  |  |
| --- | --- |
| **성과목표** | * **포털이나 검색엔진에서 제공하는 검색 서비스의 정보 제공에 대한 신뢰성 비교를 통해 객관적인 지표를 가질 수 있게 됨.** * **주기적인 청정 정보 신뢰도 공표를 통해 상업적 콘텐츠 운영사 및 포털-검색엔진 업체에 경각심 기대.** * **서비스를 통해 콘텐츠 소비 주체인 소비자의 권리 신장에 도움.** |