**CS 프로젝트 업무 일지**

|  |  |
| --- | --- |
| **프로젝트명** | “아픈 사람”커뮤니티 웹서비스 개발(토이인터랙티브) |
| **이름** | 유인근, 이정원, 이정민 |
| **기간** | 19.04.01~19.04.12 |
| **내용** | \* 연구 및 개발 주요 진행 사항 작성(2주 단위)  **[7-8주차 진행 예정 사항]**   * 7주차   + 크롤링 과정 구체화 방식 논의   + 중간고사 시험 기간으로 인해 전체적인 모델링 논의 위주로 진행할 것이며, 시험이 끝나는 대로 구현을 시작할 예정 * 8주차   + 진행계획 없음   **[5-6주차의 진행사항]**은 다음과 같습니다.   * **깃허브 private repository 생성 및 환경 연동**   + 각자 Branch 생성하여 모듈단위 작업을 수행하고 각자의 branch에 commit 하며 진행   + 주기적으로 각자의 코드에 대한 code review를 수행하고 merge하며 진행될 것   + Python: 3.7.3   + Beautifulsoup4: 4.7.1   + KoNLPy: 0.4.3 * **크롤링을 위한 html문서 구조 확인**   + 크롤링 대상 문서 구조는 정의, 원인, 증상 등의 카테고리가 <h3>태그로 구성되었고, 바로 이어지는 sibling tag <p>에 구체적인 텍스트가 담겨있다.        * **크롤링 과정 문제 및 해결방안**   + 초기에 시도했던 방식은 카테고리에 해당하는 h3태그와 내용에 해당하는 p태그를 읽는 방식을 채택   + bs객체에서 제공하는 findAll() 메서드를 활용하여 리턴된 리스트를 순차적으로 대응하며 태그를 적재함   + <h3> 태그는 있지만 <p> 태그가 없이 사진만 나오는 경우에 위의 방식에서 문제가 발생      * + Beautiful soup 공식 Docs를 확인해본 결과 HTML을 Tree 구조로 탐색하고 sibling tag를 탐색하는 기능을 제공하여 이 방식을 활용해야 할 것으로 보임   + <https://www.crummy.com/software/BeautifulSoup/bs4/doc/#navigating-the-tree>   + <https://www.crummy.com/software/BeautifulSoup/bs4/doc/#next-sibling-and-previous-sibling>   이하 내용 끝. |