**[SSAFY] 9기 멘티 활동 일지**

**작성일자: 2023년 9월 5일**

|  |  |  |  |
| --- | --- | --- | --- |
| **도메인** | **빅데이터(분산)** | | |
| **멘토** | **김성현** | **팀 코드** | **D202** |
| **진행 일시** | 2023.09.05(화) | | |
| **진행 내용** | 데이터 – 크롤링 등의 방식을 통해 수집.  Elk 등의 프레임워크를 사용해서 방향을 확장할 필요가 있다.  웹이랑 앱이랑 연동해서 확장하는 방법도 있다.  1. batch process – 주기별 / 양 별  현업에서는 양을 측정하기는 객관적이지 않다.  그래서 시간 단위로 한다.  양 단위로 하기에는 분석하기도 힘들다.  2. 자연어 처리 : 키워드 배열에 사용, 로그 쌓을 때도 사용자에 대한 정보를 넣기도 한다.  어떻게 활용을 하냐에 따라 다르다  분산 프레임워크 : elk, kafka, 하둡이 쓰이는데 기업마다 다르다.  금융권이나 실시간적으로 처리가 필요한 곳은 카프카가 많이 쓰인다.  하둡, elk, kafka 비교해보고 써보는것도 괜찮음  3. 분산 -> 속도 향상, 저장할 때 데이터 관리 속도가 높아짐. 하지만 데이터 구성에 따라 달라지기도 하기 때문에 case by case이다.  4. 보통 elk, kafka 많이 쓰이나 elk 라이선스 이슈가 있다.  다른 프레임워크 리서치하는 시간을 가지는것도 좋음.  5. 현재 상태로는 분산 기술 적용이 적절하지 않다.  데이터 수집 -> 메타데이터 변환 -> 분산 의 파이프라인이 들어가야하는데 현재 상황에서는 이게 제대로 되지 않음  데이터 활용 사이트 들어가서 조사를 해야할 필요성이 있다.  6. 데이터 관련 직업 : 분석, 수집, 처리  수집 : 하둡, 카프카 솔루션 제대로 알고있으면 됨  분석 : 분석은 기본적으로 nlp 학위가 있는게 유리하다. 전문가님은 자연어처리를 전공하셨다.  7. CI : 깃 : 소스관리  전문가님은 PM이라서 다른 파트원들이 merge request하면 읽어보고 수락을 하는 위치.  git으로 하면 결재까지 가능하다. 그래서 대부분 git을 쓴다.  CD : 깃 연동해서 젠킨스를 사용한다. 젠킨스를 사용하는 곳도 있고 그렇지 않은 곳도 있다.  요즘은 클라우드화가 대세라서 도커 이미지도 많이 만듦.  스크립트 이용해서 이미지 만듦.  쿠버네티스 쓴다!  추가 질문  Q. 텍스트로 적혀있을 때 정형화 되어있지 않은 부분에 있어서 전처리랑 분산 어캐하는지?  A: NLP가 들어가야한다. 하지만 이게 힘들다면 정형화 시키는 것이 낫다.  은행 앱 - 정형화되어있다.  굳이 자연어처리에 시간을 쓸 필요는 없다.  Q. 자연어처리를 한다면 어디에 하는게 좋을까!?  A. < 20 front squats 155 / 105 lb >  이 경우 숫자가 뒤에 나올 수도 있다.  그럼 숫자는 중량 아니면 횟수다  이러면 처리가 너무 어렵기 때문에 그냥 입력 폼을 정형화 시키는 것이 좋을 것 같다.  전문가님이 사용자라면 걍 select 박스로 할 것같다.  그게 좀 더 명확할 것 같다.  여기보다 빅데이터 부분을 확장하는게 좋을 것 같다.  기획이 아무리 좋더라도 결과물이 중요하다. | | |