**[SSAFY] 8기 멘티 활동 일지**

**작성일자: 2023년 03월 06일**

|  |  |  |  |
| --- | --- | --- | --- |
| **도메인** | **빅데이터 분산** | | |
| **멘토** | **홍승환** | **팀 코드** | **C206** |
| **진행 일시** | 2023년 03월 06일 13:40 ~ 14:10 | | |
| **진행 내용** | 머신러닝 연구와 분산 데이터 처리에 대한 연구를 수행 중이신 스캐터랩의 홍승환 멘토님과 프로젝트 기획과 관련된 면담을 진행할 수 있었습니다. 저희가 드렸던 사전 질문들에 대해, 다음의 답변들을 얻을 수 있었습니다.  **Q1. 상용화 하지 않는 프로젝트임에도 기사의 저작권을 침해할 가능성이 있을지 궁금합니다.**  A1. 저작권법 7조에 따르면, 사실의 전달에 불과한 시사보도는 저작권법에 의해 보호받지 못한다. 또한, 제28조에 따르면 공표된 저작물은 보도, 비평, 교육, 연구 등을 위해서는 정당한 범위 안에서 인용할 수 있기 때문에 교육적인 목적으로 사용하되, 출처 고지만 명확하게 한다면 큰 문제는 없을 것이다.  **Q2. 매일 기사 생산량이 6만 건 정도라고 하는데, 이 정도 규모에서 분산의 효율성을 체감할 수 있을지 궁금합니다.**  A2. 어느 정도의 디그리(Degree)로 분산하는지가 중요하다. 즉, 컴퓨터 2대 정도로 분산한다면 효율이 낮겠지만, 64대 정도로 분산한다면 빛의 속도로 처리될 것이다. 분석 과정에서 사용되는 알고리즘이나 시간 복잡도에 따라 효율 체감도가 낮겠지만 싱글 프로세스로 다루는 것보다는 확실히 효율적일 것이다.  **Q3. 한 번에 많은 데이터를 처리할 경우, 서버가 자주 터질 것 같은데 이런 경우에 대한 대응 방법이 궁금합니다.**  A3. 서버가 터지는 경우에 대한 대응 방법은 정말 많은데, 가장 간단하고도 단순 무식한 방법은 서버를 증설하는 것이다. 오토 스케일링(Auto-Scaling)을 활용하여, CPU 점유율이 10분 동안 80%이상을 유지할 경우 증설해주는 것이 바람직하다. AWS의 경우 EMR을 통해 오토 스케일링을 수행할 수 있고, 백 프레셔를 적용할 수도 있다. 입력으로 들어오는 데이터보다 처리되는 데이터가 적을 경우, 입력 속도를 줄이거나 큐잉을 적용하여 대응할 수 있다. 단, 실시간성이 중요한 서비스인 경우 비용을 들여서라도 서버를 증설하는 것이 바람직하다.  **Q4. 현업에서 데이터를 어떤 방식으로 수집하시는지, 크롤링이나 다른 수단을 활용하시는지 궁금합니다.**  A4. ‘CHAT-GPT’, ‘이루다’의 경우 모두 크롤링을 통해 데이터를 획득한다고 보면 된다. 트위터, Common 크롤러, Books Dataset을 활용하여 데이터를 최대한 확보할 수 있다.  **Q5. Hadoop에서 데이터 수정이 어려운걸로 알고 있는데, 정정 기사에 대해서는 어떻게 처리해야할지 궁금합니다.**  A5. 데이터가 어떻게 들어오고, 어떻게 처리한 다음 데이터 파이프라인을 통해 어떻게 출력하는지가 본 프로젝트의 핵심이다. 즉, 신규 업로드 기사와 정정 기사를 구분 해야하는데 ‘최종 업데이트된 시간’과 같은 컬럼을 추가한 다음, 정정 기사의 경우 기존의 기사를 덮어쓰는 방식으로 처리하면 될 것 같다. 이때, 트렌드의 분석 주기는 1시간 단위 혹은 하루 단위로 처리하면 될 것 같다.  **Q6. Apache Spark와 Apache Kafka의 차이와 둘을 함께 사용할 수 있는지 궁금합니다.**  A6. Kafka(Producer)는 스트리밍 데이터 처리를 위한 큐 역할을 수행하며, Spark(Consumer)는 데이터 스트림 처리를 위한 라이브러리 플랫폼 역할을 수행한다. 즉, 크롤러가 데이터 크롤링을 수행한 뒤, Spark는 Kafka로부터 데이터를 받아 처리한다. 해당 방식을 적용했을 경우, 입력 데이터의 속도와 출력 데이터의 속도가 달라도 된다는 이점이 있으며, 카프카의 대기열을 여러 개 만들 수 있어 확장성 또한 좋다.  **Q7. (추가질문) 스캐터랩의 대화형 AI 서비스 '이루다'는 챗GPT 등장 이전 대한민국의 큰 관심을 받았는데요, 뜨거운 관심과 함께 서비스와 관련된 논란도 일부 존재했습니다. 필터링 규칙을 교묘히 회피한 어뷰징에 의한 빅데이터 오염에 대한 대응 방법에 대해 질문드리고 싶습니다!**  A7. 일반적인 대화형 AI 서비스의 경우, 필터링 규칙을 통해 어뷰징을 적발한다. 다만, 필터링을 회피하는 교묘한 질문을 통해 악의적인 목적을 갖고 데이터를 제공하는 경우가 있는데, 이러한 경우에 대해서는 단순 키워드 형태로 분석하는 것이 아닌 문맥 전체를 고려하여 필터링 모델을 머신러닝시킨다. 시간이 지날수록 교묘해지는 어뷰징 행위에 대해서는 사람이 직접 적발하여 필터링 모델을 학습시킴으로써 인공지능이 오염된 데이터를 학습할 수 없도록 노력하고 있다. 본 프로젝트에서도 유사 언론에 대한 방지책과 단순 광고문을 배제할 수 있는 필터링 과정을 추천하는 것이 좋아 보인다. | | |