

## 캡스톤 18조 장우혁 멘토링 활용 보고서

프로젝트명	O24Sec	팀명	멜러리를 찾아서
팀 멘토링	* 간단히 팀의 공통사항에 대한 미팅 주제와 개인별로 느낀점을 적었습니다.		
	<b>[04/30] Ceeya 미팅 - 문성익 멘토님</b>		
	<ul style="list-style-type: none"> <li>● (Q) 프로젝트의 범위를 회사의 요청에 따라 조금 축소하여 진행하기로 하였습니다. 그래서 80번 포트에 대한 데이터 분석을 진행하게 되었는데 현재 이에 대해 디코딩된 80번포트의 페이로드를 사용하게 되어 거의 자연어처리와 유사한 방법으로 진행하고 있습니다. [현재 프로젝트의 진행 상태에 대한 공유]</li> </ul>		
	<ul style="list-style-type: none"> <li>● (멘토) 확실히 80번포트만 두고 보았을 때 자연어와 유사하지만 그래도 컴퓨터적 언어가 많이 사용 되고 말씀하신 처리 방법 중에 특수문자를 모두 제외하는 방법은 SQL INJECTION등 공격 패턴에서 자주 사용 되는 스트링이 제외 될 수 도 있으니 위험할 것 같습니다. 공격에 대한 탐지라면 특수문자를 어느정도 처리해주는 것이 좋을 것 같기도 합니다.</li> </ul>		
	<ul style="list-style-type: none"> <li>● (멘토) 또 예시로 Spam filter 방법 중에 이전에 알려진 방법으로basian filter 방법이 있었는데 이거는 성능이 생각보다 안 좋았고 그 이후에 다른 사람이 실수로 잘못 만든 filter 방법이 있었는데 그게 성능이 더 좋았습니다. 그때 두 필터의 차이는 다른 사람이 실수로 특수문자에 대한 예외처리를 안해줘서 특수문자가 포함된 String을 읽게 되는 것이 었는데 이게 생각보다 결과가 좋았다고 합니다. 때문에 일단 다 제외하는 것도 좋지만 가능성은 열어두고 여러 가지 방식으로 Tokenization을 해보면서 결과를 보는 것이 좋을 것 같습니다.</li> </ul>		
	<ul style="list-style-type: none"> <li>● (Q) 현재 tf-idf벡터까지 만들어 놓은 상태이고 이 벡터를 통해서 유사도를 찾으려고 하는데 각 데이터를 다른 모든 데이터와 유사도를 비교하게 되면 시간 복잡도가 <math>O(N^2)</math>이 되게 됩니다. 이는 상당한 복잡도라고 생각해서 다른 방법으로 랜덤으로 하나 샘플링 하면서 뭉쳐서 데이터 개수를 점차 줄여가면서 비교를 하는 방법을 하려고 하는데 이거 외에도 해볼만한 방법이 있을까요? [실험 전 진행 방향 확인]</li> </ul>		
	<ul style="list-style-type: none"> <li>● (멘토) tf-idf 벡터까지 나와 있다면 많이 사용 되는 클러스터링 방법으로 유사도를 찾는 것도 좋을 것 같습니다.</li> </ul>		
	<ul style="list-style-type: none"> <li>● (멘토) 기존 텐서플로 Clustering 라이브러리 중에 하이퍼파라미터를 제안해주는 라이브러리가 있으니 이걸 이용해 봐도 좋을 것 같고 상황상 그게 안될 것 같다면 말씀하신 알고리즘을 통한 방법도 좋을 것 같습니다. 과정에 대한 근거를 다 보여줄 수 있을 것 같네요 또 다른 방법으로 하이퍼파라미터를 데이터가 어떤 상황일 때 어떤식으로 넣어줄지 가정하고 실험한 결과를 같이 주는 것도 좋은 방법일 것 같습니다.</li> </ul>		
	<ul style="list-style-type: none"> <li>● (Q) 유사도 알고리즘을 찾아보니 여러 종류가 있었습니다. 현재 저희는 많이 사용되는 코사인 유사도를 통해 결과를 만들려고 하는데 혹시 현업에서 여러 유사도 알고리즘을 적용해 보셨을 때 큰 차이가 있으셨나요? [추천받은 알고리즘인 코사인 유사도를 통해 프로젝트를 진행하고 있지만 이외에도 현업에서 주요하게 많이 사용되는 알고리즘이 있는지 조언을 받고자함]</li> </ul>		
	<ul style="list-style-type: none"> <li>● (멘토) 사실 그 부분에 대한 확답은 어려운데 피쳐특징에 따라 다를 것 같습니다. 코사인 유사도 같은 각도 기반과 유클리디안 같은 거리기반의 유사도 알고리즘이 이름부터 차이가 있듯이 거리는 멀지만 같은 각도를 가진 피쳐들이 있을 경우 이 두가지의 결과는 크게 다를 것이니 말이죠 하지만 제 생각에는 이런 경우가 있긴 하지만 많지는 않을 것 같기 때문에 각 알고리즘을 다 적용해 보면서 최적의 결과가 나오는 것을 사용 하는게 좋을 것 같습니다. 그리고 만약 결과가 모두 비슷하다면 자원소모가 적은 것을 골라 사용 하는게 결과적으로 좋겠죠.</li> </ul>		

## 팀 멘토링

- (Q) 사람들에게 발표를 통해 내용을 전달하는 것 관련된 내용인데, 저희가 중간발표 평가에서 내용이 너무 어렵다는 피드백을 받아서 이 내용을 좀 더 쉽게 많은 사람들에게 전달하고자 할 때 어떤 방법을 통해 전달하면 전문적인 지식이 없는 일반인도 이해시키면서 설명을 할 수 있을까요? [중간 발표 피드백에서 가장 중요한 부분이 다음 발표 때 청중이 이해하기 쉽게 발표하자는 내용이었기 때문에 이부분에 대해서 우리보다 많이 프레젠테이션을 해오신 멘토님께 의견을 구하고자 함]

- (멘토) 발표에는 기본적으로 3가지 요소가 중요하다고 생각합니다 What : 무엇에 대한 설명인지 Why : 왜 그것을 만들게 되었는지 How : 어떻게 만들었는지, 이 세가지를 통해 발표를 할 때 What과 Why는 배경에 대한 설명이기 때문에 일반인도 흥미롭게 들을 수 있습니다. 때문에 이부분에 대해서는 일반인이 이해할 정도로 쉽게 말하는 것이 좋고 How 는 기술적인 부분이기 때문에 자신들이 얼마나 노력을 했는지 보여 주려면 일반인은 자더라도 열심히 준비한 것을 보여주는 게 나쁘다고 생각하지 않습니다. 그리고 발표자료에 최소의 내용과 이해를 도와주는 그림을 넣으면서 너무 왔다 갔다 하지 않고 흐름적으로 쉽게 보여줄 수 있게 구성하는 게 좋을 것 같습니다.

=> 느낀점: 현재 3차 목표인 클러스터링 부분이 성능 평가 및 여러 방법 활용으로 프로젝트 범위가 변경 및 축소된 상황인데, 이에 맞는 해결법을 찾기위해 스스로 노력하고 그것을 멘토님과 공유하며 여러 의견을 나눌 수 있었습니다.