COSE362 Final-term report

2021320024 김민서
2021320303 정지원
2021320322 윤민서
2021320323 허준환

1. 문제 정의

이번 Term Project는 고려대학교 학우들의 강의 평가 데이터를 통해 학부 강의들을 어떤 유형으로 구분지을 수 있을지에 대한 문제를 해결하는 것을 목표로 한다. 대학생들이 생각하기에 가장 보편적이고 대표적인 강의 유형에는 후술할 '명강'과 '꿀강'이 있는데, 이번 프로젝트의 우선적인 목표는 이와 같이 최소 2개 이상의 강의 유형을 찾아내는 것이다. 또한, 분석 과정 및 결과에서 2가지의 강의 유형 이외에도 새로운 유형이 발견된다면, 그 유형을 새롭게 정의하고 특징을 분석하고자 한다.

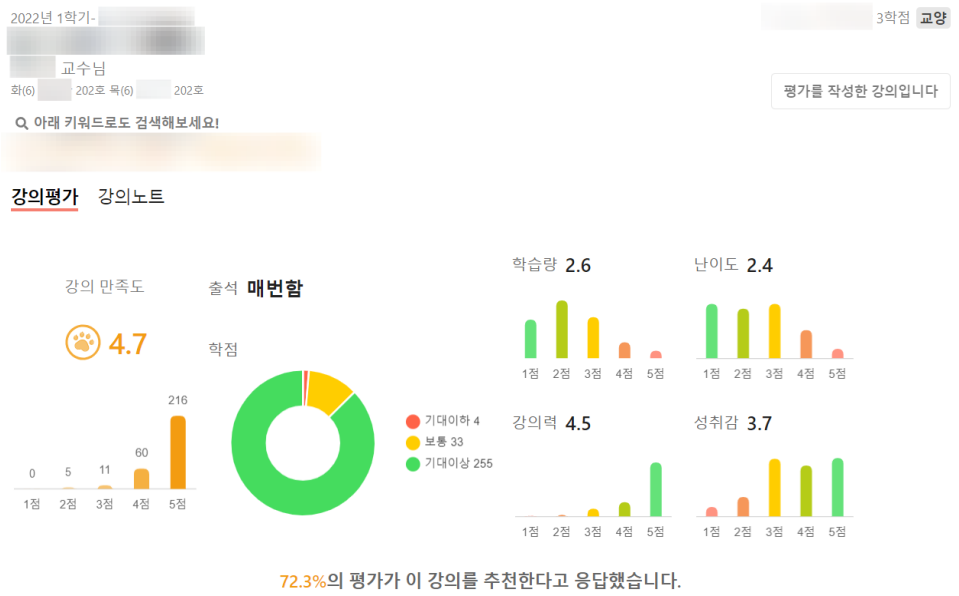
'명강'은, 성적이 잘 안 나올지라도 교수님이 전달해 주시는 내용과 학문 자체에서 얻어갈 것이 많은 강의이며, '꿀강'은 강의 내용이 수강자에게 알차지 않거나 이미 배운 내용이 대부분 포함되어 있음에도 불구하고 단순히 최종 성적만을 바라보는 신청을 하는 강의이다. 이외에도, 따로 이름이 붙지는 않았으나 성적 산출이나 강의 진행, 강의력 등의 이유로 들으면 좋지 않은 강의 등이 있다.

대부분의 대학생들은 강의를 구분지을 때 강의평, 그리고 이에 주어지는 난이도, 학점, 학습량 등 여러가지 요소들을 참고하고 고려하면서 강의 유형을 파악한다. 하지만 이러한 유형이 있음에도 불구하고, 이 유형들에 대한 구분 기준은 명확하게 존재하지 않고 있다. 현재 4학기를 재학 중인 시점에서 앞으로의 전공 과목을 이러한 구분 기준을 바탕으로 자신에게 적합한 강의를 더욱 신중하게 선택해야 할 필요가 있다. 따라서 이번 Term Project에서는 대학생들의 강의평 데이터를 바탕으로 강의 유형이 어떤 기준으로 나눠질 수 있는지 알아보고자 한다.

2. 방법론

1) 데이터 설명

KLUE의 강의평가에서 평가 요소로 작용하는 것은 <Figure 1>에서 확인할 수 있는 것과 같이 수강 학점, 강의 구분, 교수, 요일 및 교시, 강의실 위치, 강의 만족도, 출석 빈도, 최종 성적의 만족도, 학습량, 난이도, 강의력, 성취감, 강의 추천 응답 비율 등이 있다. 아래는 인터페이스를 포함한 데이터의 예시이다.



<Figure 1: 고려대학교 전용 강의평가 페이지 KLUE에 있는 강의>

수집한 데이터의 범위는 코로나19로 인하여 강의의 경향이 크게 바뀌기 시작한 2020년 1학기부터 점차 비대면 수업 이전으로 복구되어 갔던 2022년 1학기까지이다.

2) 데이터 수집 방법

KLUE 사이트의 경우 스크롤을 내려야 후기평들이 갱신되는 동적페이지이다. 따라서 selenium 라이브러리와 chrome의 웹 드라이버를 사용하여 강의 사이트의 리뷰 데이터들을 동적 크롤링하였다. 이후 beautiful soup 라이브러리를 통해 각 중요 평가요소를 parsing하여 데이터셋을 만들었다.

다만, KLUE 사이트의 강의 평가 항목이 2020년도 2학기를 기준으로 변경되었다. 2020년 1학기 까지는 학습량, 난이도, 학점, 성취감의 4가지 평가 항목이 있었으나, 2020년 2학기부터 학점이 강의력으로 바뀌면서 평가 항목은 학습량, 난이도, 강의력, 성취감이 되었다. 최근 데이터를 반영하기 위해 2020년 2학기부터 2022년 1학기까지의 데이터를 먼저 수집하였고, 2020년 1학기 데이터 또한 사용하기 위해 없는 평가 항목인 강의력을 prediction한 후 사용하였다.

2-1) 2020년 2학기 이후 강의평의 학점 feature 추가

2020년 2학기부터의 강의평은 학점 대신 강의력을 1, 2, 3, 4, 5점 중에 하나 선택할 수 있는 rating으로 바뀌었으며 기존의 학점 feature는 기대 이하, 보통, 기대 이상 중 하나를 선택하는 형식으로 바뀌었다. 기대 이하를 1로, 보통을 3으로, 기대 이상을 5로 설정하여 이전의 5점 척도의 학점 feature를 대신하여 사용하였다.

2-2) 2020년 1학기 강의평의 강의력 예측

2020년 2학기부터 2022년 1학기까지의 강의평을 토대로 2020년 1학기의 강의력 feature를 prediction하였다. 2020년 1학기의 강의력 feature를 prediction하기 위해 먼저 2020년 2학기부터 2022년 1학기까지의 강의 리뷰를 단어 단위로 분리하였다. 이후 3번 이상 등장한 단어들을 포함한 단어 집합에 대하여 Tokenizer model을 fitting하였다. 이때 parameter의 개수가 지나치게 많아지는 것을 방지하기 위하여 리뷰의 길이를 최대 200단어로 제한하였다. 리뷰의 길이가 200단어를 초과하는 강의평은 전체에서 6.2%만 차지하고 있기

때문에, 큰 성능 저하가 우려되지 않는 것으로 보인다. 가공된 dataset에 대한 강의 리뷰로 Word Embedding - LSTM model을 사용하여 강의력을 1, 2, 3, 4, 5점으로 변환시켜 데이터프레임에 추가해 주었다. 이 과정에서 p = 0.3의 dropout을 수행하여 generalization 성능을 개선하였다.

3) 데이터 전처리

강의를 구분짓는 기준에는 위에서 언급한 '명강', '꿀강'뿐만 아니라 각각의 강의를 특징지을 수 있는 무수히 많은 기준이 포함될 수 있다고 보았다. 즉, 명확한 기준이 존재하지 않는 강의 유형의 특징 상 label이 존재하지 않을 수밖에 없다. 따라서 기계학습의 분야 중 label이 필요한 supervised learning이 아니라, label이 필요 없는 unsupervised learning이 더 적절하다고 볼 수 있다. 또한 어떤 강의 유형이 존재하고, 어떠한 기준으로 나눠지는지 확인하기 위해서는 Clustering을 이용하는 것이 적절하다.

이뿐만 아니라, 앞서 말했듯 KLUE에 있는 강의평들을 통해 한 강의 데이터 당 얻을 수 있는 feature의 수는 14개 이상이 존재한다. 만일 이 feature들을 모두 사용한다면, dimension 수가 매우 커져 데이터를 처리하는 데 연산이 복잡해진다. K-Means Algorithm은 Clustering 중에서도 구현이 간단하고, 계산 속도가 빠르다는 장점이 있다. 고차원의 feature들을 다루면서 생기는 연산 복잡도의 문제로 인해, Clustering 방법론 중에서도 간단하고 빠른 K-Means Algorithm을 선택하였다.

column에 대한 설명

1. year: 연도-학기
2. type: 학수번호
3. subject: 과목명
4. professor: 교수자
5. day: 요일, 교시
6. satisfy: 강의 만족도
7. writer: 강의평 작성자
8. ratings0: 학습량
9. ratings1: 난이도
10. ratings2: 학점
11. ratings3: 성취감
12. comment: 강의 리뷰
13. helpful: 강의평이 받은 추천의 개수
14. ratings4: 강의력

Clustering을 진행하기 위하여 먼저 유의미하지 않은 강의평 작성자와 강의평이 받은 추천의 개수, csv 파일을 정리하는 과정에서 생긴 index 등의 feature를 삭제하는 과정을 진행하였다. 이후 Clustering에는 유의한 변수일지 모르나, 데이터 전처리가 어려운 변수들 또한 제외하였다. 과목명 및 학수번호, 교수자는 문자열로 이루어져 있어 문자열 자체로 사용할 수 없어 categorical variable로 설정해야 한다. 학수번호, 교수자의 경우 int 자료형으로 encoding하기 위해 one-hot encoding을 통하여 0 또는 1의 int 자료형으로 변환하였다. 하지만, 이와 같은 방식으로 인해 variable의 수가 100개 이상으로 급격히 늘어나게 되어 Clustering 분석이 매우 어려워졌다. 그래서 학수번호와 교수자 feature 또한 제외시켰다. 요일 및 시간은 categorical variable로 처리되어 있지만, 요일과 시간을

분리하여 저장하면 요일은 categorical variable, 시간은 numeric variable로 취급할 수 있는 feature이기 때문에 요일과 시간을 분리하였다. 요일은 월요일부터 금요일까지의 variable을 만들고 난 후, 월요일 및 수요일에 수업이 있으면 월요일과 수요일 변수만 1이고 나머지는 0으로 저장되게끔 int 자료형으로 변환하였다.

Variable 후처리 과정을 거쳐 최종적으로 다음과 같은 column들이 선정되었다.

	satisfy	ratings0	ratings1	ratings2	ratings3	ratings4	mon	tue	wed	thu	fri	period
0	5	2	2	4.0	5	1	0	1	0	1	0	6
1	3	2	3	4.0	2	1	0	1	0	1	0	6
2	4	4	3	4.0	4	2	0	1	0	1	0	6
3	5	4	4	3.0	4	5	0	1	0	1	0	6
4	5	3	4	4.0	4	3	0	1	0	1	0	6

<Figure 2. variable 후처리 결과>

column에 대한 설명
1. satisfy: 강의 만족도
2. ratings0: 학습량
3. ratings1: 난이도
4. ratings2: 학점
5. ratings3: 성취감
6. ratings4: 강의력
7. mon: 월요일
8. tue: 화요일
9. wed: 수요일
10. thu: 목요일
11. fri: 금요일
12. period: 시작 교시

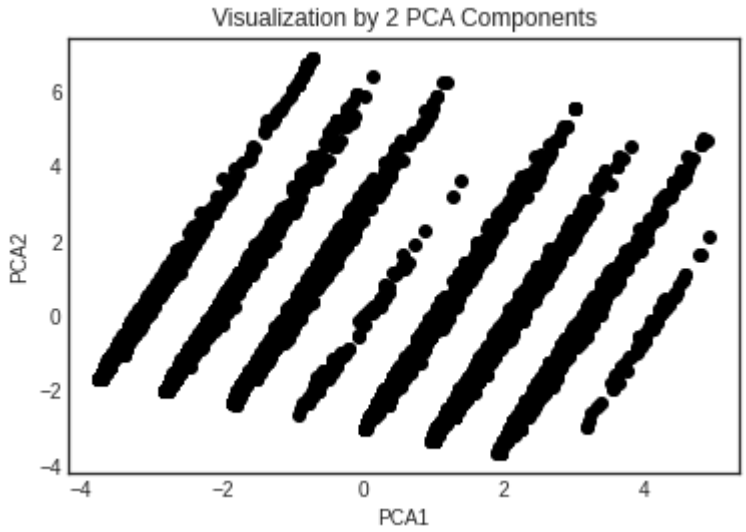
3. 결과와 해석

본격적으로 K-Means algorithm을 적용하기에 앞서서 MinMaxScaling을 통하여 모든 데이터의 값을 0과 1 사이에 위치하게 하였다. Scaling이 진행되지 않았을 경우에는 데이터의 값이 너무 크거나 너무 작은 경우에 model에 K-Means algorithm을 fitting하는 과정에서 0으로 수렴하게 되거나 무한대로 발산하게 되는 현상이 일어날 수 있기 때문에 전처리에서 크게 중요한 과정으로 간주하였다.

이후 Cluster 개수의 범위를 1부터 10으로 설정한 뒤 각각의 Cluster 개수에 대하여 Inertia를 계산하였다. 계산된 Inertia는 x축을 K, y축을 Inertia로 한 꺾은선 그래프로 시각화하여 Cluster의 개수를 정하는 데에 사용하였다.

Clustering의 결과를 2차원 공간에 시각화 하기 위하여 주성분 분석(Principle Component Analysis, 이하 PCA)을 통하여 component의 개수를 2와 3으로 감소시켰다. PCA의 결과로 새롭게 만들어진 두 개의 축을 각각 PCA1, PCA2, PCA3으로 이름을 정한 뒤 PCA1을 x축, PCA2를 y축, PCA3를 z축으로 한 scatter plot을 그려서 Clustering이 잘 진행되었는지 점검하는 과정을 수행하였다.

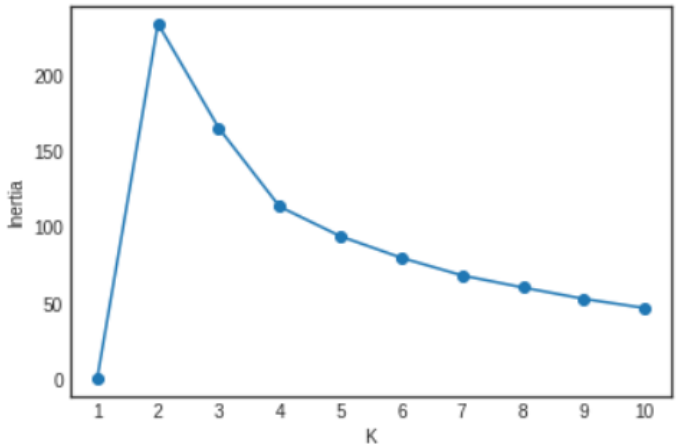
Clustering을 진행하기 앞서, 데이터의 분포를 살펴보기 위해 <Figure 3>에서 주성분 분석을 통해 데이터를 2차원에 projection 하였다.



<Figure 3: dataset의 2차원 좌표평면 시각화 결과>

Cluster의 개수를 정하기 위해 계산한 Inertia의 값과 시각화 결과는 다음과 같다.

```
K: 2    Inertia: 234.38855643867885    Difference: -234.38855643867885
K: 3    Inertia: 165.52634019348037    Difference: 68.86221624519848
K: 4    Inertia: 113.7545025707991     Difference: 51.77183762268126
K: 5    Inertia: 94.19533545830546     Difference: 19.55916711249364
K: 6    Inertia: 80.014207295622       Difference: 14.181128162683464
K: 7    Inertia: 68.36643574234218     Difference: 11.647771553279824
K: 8    Inertia: 60.41336909861742     Difference: 7.953066643724753
K: 9    Inertia: 52.91950648599951     Difference: 7.49386261261791
K: 10   Inertia: 46.89418941193796     Difference: 6.0253170740615545
```



<Figure 4: K값에 따른 Intertia의 변화>

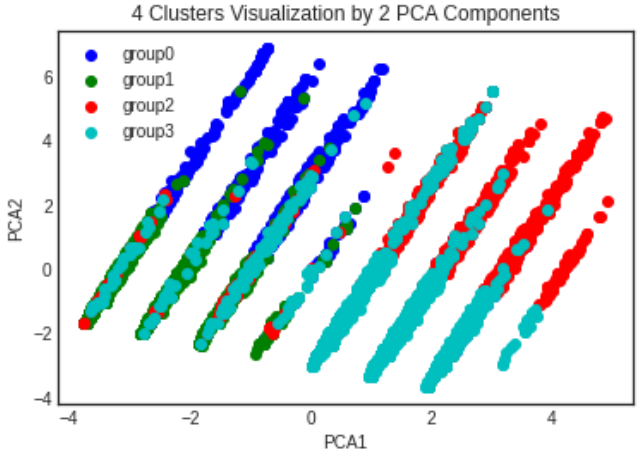
Elbow method를 통하여 K값을 결정된 Cluster의 개수는 4이다.

결정된 Cluster의 개수를 기반으로 하여 K-Means algorithm을 통하여 model을 fitting 하였다. fitting한 model을 바탕으로 강의평 별로 Cluster를 prediction 한 뒤에 원래의 데이터프레임에 cluster column을 추가하였다. 아래는 출력 예시이다.

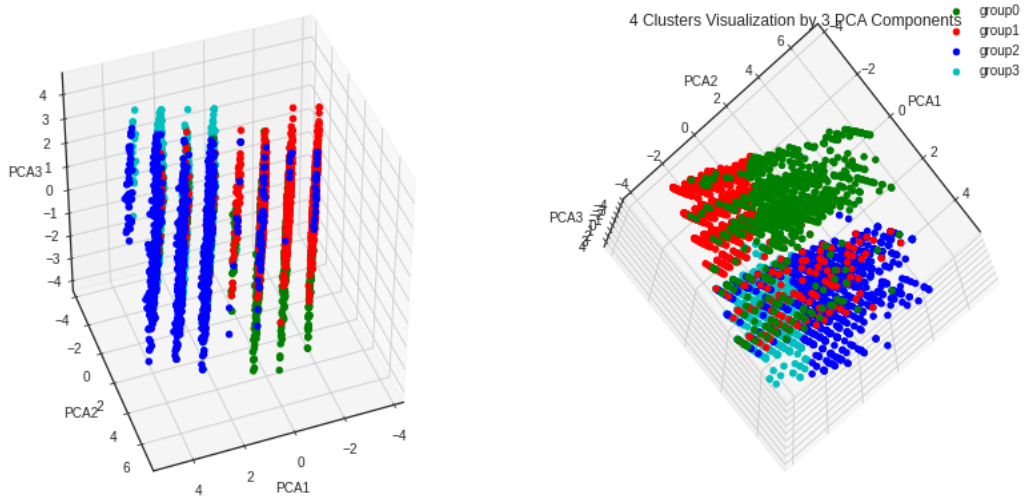
	satisfy	ratings0	ratings1	ratings2	ratings3	ratings4	mon	tue	wed	thu	fri	period	cluster
0	5	2	2	4.0	5	1	0	1	0	1	0	6	0
1	3	2	3	4.0	2	1	0	1	0	1	0	6	0
2	4	4	3	4.0	4	2	0	1	0	1	0	6	0
3	5	4	4	3.0	4	5	0	1	0	1	0	6	1
4	5	3	4	4.0	4	3	0	1	0	1	0	6	1

<Figure 5: Cluster가 배정된 데이터프레임>

Cluster의 번호는 0번부터 4번까지이다. 결과를 2차원 좌표평면과 3차원 좌표공간에 표시하기 위하여 PCA를 통하여 새로운 2개의 축을 만들었고 scatter plot을 나타낸 결과는 아래와 같다. legend의 group0~group6은 각각 0번부터 4번까지의 Cluster를 나타내는 것이다.



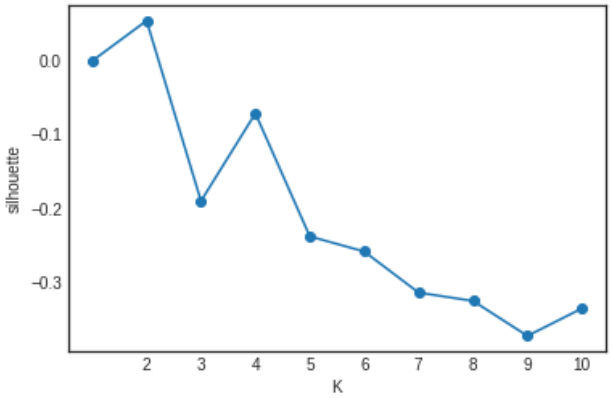
<Figure 6: Clustering 결과를 시각화한 2차원 scatter plot>



<Figure 7: Clustering 결과를 시각화한 3차원 scatter plot>

scikit learn의 K-Means algorithm의 hyperparameter 중 algorithm의 default 값인 Lloyd algorithm을 사용한 결과이다. 이후 Elkan algorithm을 사용해 보았으나 Lloyd algorithm의 결과와 다르지 않았기 때문에 결과는 생략한다.

K-Means algorithm을 사용한 Clustering 이외에도 Spectral Clustering을 시도해 보았다. K값을 정하기 위하여 silhouette 계수를 계산하였다. 그러나 silhouette 계수가 가장 큰 시점이 K = 2일 때 약 0.055로 유의미한 값을 보이지 않기 때문에 Spectral Clustering은 진행하지 않는 것으로 결정하였다. 해당 알고리즘은 Fully connected된 Graph에 대한 문제일 경우에만 분석이 유의미해질 수 있기 때문인 것으로 생각된다.



<Figure 8: K값에 따른 silhouette 계수의 변화>

최종 Model로 Lloyd Algorithm을 사용한 Clustering Model을 선정하였고 각 Cluster에 해당하는 강의평 수는 Cluster 0, Cluster 1, Cluster 2, Cluster 3이 각각 1092, 2084, 1007, 1737개이다. K-Means algorithm은 각 Cluster의 데이터를 가장 잘 표현하는 reference vectors를 중심으로 Clustering을 반복해나간다. 즉, Cluster의 centroid를 통해서 그 Cluster의 특징을 살펴볼 수 있다. 4개의 Cluster의 reference vector는 다음과 같다.

cluster	satisfy	ratings0	ratings1	ratings2	ratings3	ratings4	mon	tue	wed	thu	fri	period
0	3.134615	3.393773	3.139194	3.158425	2.636447	2.749084	0.415751	0.446886	0.431319	0.440476	0.178571	5.512821
1	4.575816	3.903071	3.576775	4.180422	4.259597	4.439539	0.356526	0.543186	0.414107	0.536948	0.126200	5.929942
2	3.006951	3.287984	3.147964	2.995035	2.478649	2.718967	0.425025	0.568024	0.444886	0.552135	0.129096	1.859980
3	4.442142	3.873921	3.617732	4.029361	4.187680	4.305699	0.454807	0.527346	0.477260	0.500864	0.082902	2.064479

<Figure 9: 각 Cluster의 centroid>

<Figure 9>를 살펴보면, mon, tue, wed, thu, fri column에 있는 값은 크게 차이가 나지 않아 강의가 진행되는 요일은 크게 영향이 없었던 것을 알 수 있다.

Cluster 0은 평균적으로 5교시에 시작하는 수업으로 이루어져 있음을 알 수 있다. Cluster 0의 강의 만족도는 평균적으로 3.1이고, 학습량은 3.4, 난이도는 3.1, 학점은 3.2, 성취감은 2.6, 강의력은 2.7 정도이다. 강의 만족도, 학습량, 난이도, 학점, 성취감, 강의력은 1부터 5까지이므로, Cluster 0은 다섯가지 항목 모두 보통 정도에 해당한다고 볼 수 있다. 결론적으로, 오후에 시작하는 보통 수준의 강의로 이루어진 하나의 Cluster가 있음을 알 수 있다.

이와 마찬가지로 Cluster 1~3도 비슷하게 해석할 수 있다. Cluster 1에 해당하는 수업은 평균적으로 6교시 이후에 시작된다. 평균적으로 강의 만족도는 4.6, 학습량은 3.9, 난이도는 3.6, 학점은 4.2, 성취감은 4.3, 강의력은 4.4이다. 다른 Cluster들에 비해 학습량이 가장 많은 편에 속하고 취득 학점과 강의 만족도가 높기에 Cluster 1은 앞에서 언급하였던 꿀강과 명강 모두, 특히 꿀강에 가깝게 해당한다고 볼 수 있다. 즉, 일주일 중 화, 목요일 오후에 한 번 수업을 진행하고, 공부량과 학점 및 강의 만족도가 모두 높은 Cluster가 있음을 알 수 있다.

Cluster 2는 보통 1~2교시에 진행되고, 평균적으로 강의 만족도는 3.0, 학습량은 3.3, 난이도는 3.1, 학점은 3.0, 성취감은 2.5, 강의력은 2.7 정도인 강의로 이루어져 있다. Cluster 0과 비교했을 때 강의 자체의 특성은 비슷하지만 오후 시간대가 아닌 오전 시간대에 진행한다는 차이점이 존재한다는 사실을 알 수 있다. Cluster 0과 마찬가지로 확실하게 강의 유형을 구분지을 수 없는 보통 수준의 강의이지만 오전 기상이 상대적으로 힘들다는 것을 고려하여 1, 2교시 강의를 피하려는 경향이 있기 때문에 Cluster 0과는 유의미한 차이점을 보인다고 해석할 수 있다.

Cluster 3은 보통 2교시에 진행되고, 평균적으로 강의 만족도는 4.4, 학습량은 3.9, 난이도는 3.6, 학점은 4.0, 성취감은 4.2, 강의력은 4.3 정도인 강의로 이루어져 있다. Cluster 1과 비교했을 때 학습량과 난이도는 비슷하지만 강의 만족도, 학점, 성취감, 강의력 항목이 다소 떨어지는 것을 볼 수 있다. Cluster 1에 비해서 성취감은 0.1이 차이나고 학점은 0.2가 차이나기 때문에, 즉 성취감에 비해서 학점이 비교적 더 많이 떨어지기 때문에 강의에서 요구되는 것에 비해 최종 학점이 잘 나오지 않는다는 것을 알 수 있다. 이러한 사실이 있음에도 불구하고 모든 항목이 높은 점수에 속하기에 Cluster 3은 앞에서 언급하였던 꿀강과 명강 모두에, 특히 명강에 가깝게 해당한다고 볼 수 있다.

Cluster 별로 정량적인 분석은 완료되었으나 어떤 강의들이 속해 있는지 더 정확하게 분석하기 위해 정성적인 분석 방법이 필요하였고 데이터 프레임 전처리 과정에서 삭제한 comment feature를 이용하여 Wordcloud를 Cluster 별로 그리는 방법을 채택하였다. Cluster 별 Wordcloud 시각화 결과는 다음과 같다.

<Figure 10: Cluster 0의 Wordcloud 시각화 결과>

<Figure 11: Cluster 1의 Wordcloud 시각화 결과>

<Figure 12: Cluster 2의 Wordcloud 시각화 결과>

<Figure 13: Cluster 3의 Wordcloud 시각화 결과>

모든 Cluster의 강의평에 공통적으로 자주 등장할 가능성이 높으며 Wordcloud에 큰 비중으로 들어가 있는 단어인 '강의', '수업' 등의 단어를 삭제하는 처리 과정을 거쳤다. 하지만 삭제 과정을 여러 차례 거쳤음에도 불구하고 다시 같은 현상을 보이는 경향이 컸기

때문에 더이상 제외하지 않기로 결정했다. 이러한 부분으로 인해 '시간', '정도', '설명' 등의 단어들이 모든 Cluster에서 가장 큰 부분을 차지하게 되었고 완전히 용이한 분석이 진행되지 않았다는 점은 아쉽지만 '어려운' 등의 형용사나 '파이썬', '통계', '프로젝트' 등의 명사의 존재를 확인할 수 있다는 점에서 Cluster 별의 정성적인 특징을 유의미하고 빠르게 확인할 수 있다.

4. 본 연구를 바탕으로 향후 가능한 추가 연구 개발 방향

본 연구에서는 몇 가지 한계점이 존재한다. 우선, 강의평 데이터의 비일관성으로 인해 prediction을 통해 강의평 데이터를 통일하였는데, prediction이 완벽하지 않아 필연적으로 데이터 자체에 결함이 발생하였다. 데이터셋을 2020년 2학기 이후와 1학기 이전으로 나누어 각각 Clustering을 진행한다면 이러한 문제를 해결할 수 있을 것이다. 또한, 2차원 PCA를 시각화 한 그래프를 보았을 때 대략 8개의 Cluster가 있는 것으로 보이지만, 실제 K-Means 및 Spectral Clustering을 진행한 결과는 그렇지 않았다. 후속 연구에서 본 연구에서 적용했던 2가지 방식이 아닌, 다른 Clustering Algorithm을 여러가지 도입하여 비교한다면 좀 더 유의미한 Clustering 결과를 얻을 수 있을 것이다.

지금까지 본 연구에서는 강의평을 토대로 강의를 분류하고, 분류된 강의들의 특징을 분석하였다. 최근에는 대학뿐만 아니라 많은 단체와 기업에서 다양한 형태의 강의 서비스를 제공함에 따라 어떤 강의가 본인과 맞는 강의인지 구분할 필요성이 증가하고 있다. 강의 후기를 기반으로 강의를 분류할 수 있다면 수강생들의 강의 만족도와 성적 향상을 기대할 수 있을 것이다. 강의평가에 대한 분석 뿐만 아니라 다양한 분야의 후기에 대한 평가 또한 응용될 수 있다. 긍정적, 부정적을 나누는 감성 분석에 대한 연구에서 더 나아가 단순한 좋고 나쁨을 나누는 것이 아닌 어떻게 좋고 어떻게 나쁜지를 구분하여 이용자의 필요와 성향에 따른 추천 서비스를 제공할 수 있다.