<한국어인공지능과 데이터과학, 최동혁교수님('22년도 1학기)>

토픽모델링을 활용한 코스닥 우량기업과 한계기업의 감사의견 비교

학번: 2021710232
이름: 신소희

1. 서론

국내 증권시장에 상장된 상장기업의 경우 투자자에게 합리적인 투자판단 자료를 제공하고 시장에서 공정한 가격형성이 이루어지도록 사업내용, 재무상황 등 기업활동 내용을 정기적으로 공시하고 있다. 상장기업은 회사의 회계처리의 투명성을 위해 회사로부터 독립된 외부감사인에게 회계감사를 받고 감사보고서를 공시하여야 한다. 감사보고서의 감사의견이 적정이었음에도 불구하고 분식회계, 횡령 등이 발생하여 선의의 투자자가 피해를 보는 상황이 발생하기도 하는데 이를 보완하기 위해 감사보고서 상 '핵심감사사항'을 추가하는 핵심감사제도가 2017년 도입되었다. 핵심감사사항은 외부감사인의 전문가적인 판단상 왜곡되어 해석될 위험, 유의한 위험이 있는 분야 등에 대한 감사인의 의견이며 이를 통해 투자자들은 수치데이터인 재무정보 뿐만 아니라 감사인의 의견을 통해 기업에 대한 심도깊은 이해가 가능해진다. 한계기업 모델링과 관련한 기존의 연구는 주로 재무상태를 나타내는 수치데이터를 사용하였다. 본 연구에서는 텍스트를 정보를 추가한 한계기업 모델링 연구의 초석이 될 수 있도록 우량기업과 한계기업 사이의 감사보고서 상 감사의견에 차이가 있는지 분석하고자 한다.

2. 배경지식

2.1 코스닥소속부

코스닥시장의 우량기업과 일부 부실기업들이 '코스닥'이라는 이름으로 동일시되며 디스카운트 현상을 보인 것을 해소하고자 우량기업과 부실기업이 구분될 수 있도록 2008년부터 코스닥 소속부제도를 시행하였다. 12월 결산업인 사업보고서 제출일을 기준으로 심사하며 도입 초기에는 벤처인증 유무에 따라 벤처기업과 일반기업으로 구분하였다. 하지만 상장폐지 실질심사 도입 및 외부감사인의 회계감사가 강화됨에 따라 투자자가 위험을 인지하지 못한 상황에서 관리종목 지정 및 상장폐지 등이 발생함에 따라 투자자들이 사전에 참고하여 신중한 투자를 할 수 있도록 우량기업부, 벤처기업부, 중견기업부, 신성장기업부 4개 소속부로 2011년 5월 개편되었다. 외국기업, 투자회사, 상장지수집합투자기구, 부동산투자회사, SPAC(Special Purpose Acquisition Company)은 기타이며 관리종목 및 투자주의환기종목은 소속부에 포함하지 않고 별도관리한다.

2.2 관리종목 및 투자주의환기종목

한국거래소는 상장기업이 상장폐지기준에 해당되면 원칙적으로 즉시 상장폐지되어야 하나 투자자보호를 위해 즉시 상장폐지하는 대신 관리종목으로 지정하여 일정기간 상장폐지를 유예하여 투자자에게는 투자위험을 환기시키고 상장기업에게는 상장폐지사유를 해소할 수 있는 시간적 기회를 준다. 상장폐지 사유발생, 50%이상 자본잠식, 자기자본 미달, 감사의견 비적정, 공시의무 위반 사업보고서 미제출 등의 사유로 지정된다. 코스닥시장의 기업계속성 및 경영투명성에 주의를 요하는 기업을 사전에 인지하기 어려워 투자주의환기종목을 지정하여 투자자들이 사전에 참고할 수 있도록 제도를 도입하였다.

2.3 핵심감사제도

기존의 감사보고서는 감사의견만을 제시하고 감사과정이나 감사인의 의견에 대한 정보가 부족하였다. 이를 보완하기 위해 금융감독원은 감사인의 전문적인 판단에 따라 당기의 재무제표 감사 중 가장 유의적인 내용에 대해 서술하도록 하는 핵심감사제를 2017년 도입하였다. 감사인은 감사에서 가장 유의적인 사항으로 결정한 사항과 그 이유에 대해 설명하여야 한다. 상장기업의 부담을 최소화하기 위해 2018년 사업보고서는 자산규모 2조 이상, 2019년도 사업보고서는 자산 1천억원 이상, 2020년 사업보고서부터는 전체 상장기업에 도입되었다. 본 연구에서는 전체 상장사에 적용된 2020년 이후인 2021년 사업보고서를 대상으로 한다.

3. 연구방법

3.1 데이터

본 연구에서 우량기업 및 부실기업을 코스닥 소속부 기준으로 우량기업부 소속기업을 우량기업으로 관리종목 및 투자주의환기종목을 부실기업으로 보았다. 2021년도 사업보고서 공시 이후 소속부 심사가 끝난 2022년 5월 소속부를 기준으로 하였다. 데이터는 대상기업의 최근 1개 사업년도의 감사보고서의 감사인의 독립된 의견중 감사의견, 핵심감사사항, 강조사항, 기타사항이다. 데이터는 한국거래소 상장공시홈페이지(kind.krx.co.kr)에서 수집하였다.

3.2 데이터 전처리

텍스트 분석을 위해 다양한 전처리를 시행하였다. 먼저 텍스트 데이터에서 한글 문자를 제외한 숫자, 영어, 특수기호를 모두 삭제하였다. 본 연구에서는 수치데이터가 아닌 텍스트데이터를 다루기 때문에 감사의견 상의 수치는 제외하였다. 이후 Konlpy의 Okt 및 kiwi 라이브러리를 사용하여 텍스트를 단어단위로 토큰화하였다. 불용어 삭제 전 빈도수를 확인해보았으며 결과는 Figure 1과 같다. 텍스트데이터에서 주로 나타나는 Zipf 분포가 발견되었다.

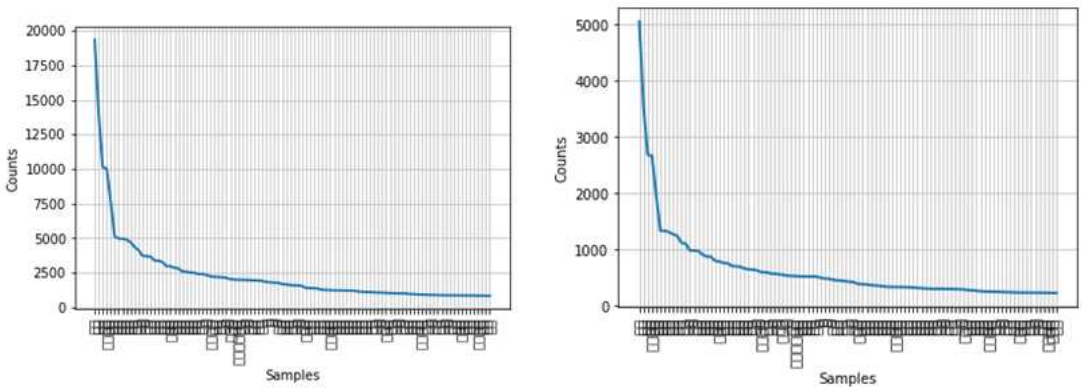

Figure 1. 우량기업의 단어 빈도수(좌측) 및 부실기업의 단어 빈도수(우측)

텍스트 분석에 유의미한 단어는 빈도수가 중간정도인 단어라고 알려져 있어 불용어사전을 만들어 제거하였고 1음절단어도 의미가 없다고 판단하여 제거하였다.

'감사', '연결', '우리', '재무제표', '사항', '대한', '표시', '관련', '의견', '책임', '회사', '기업', '평가', '대하', '수행', '해당', '보고서', '영진', '공시', '경우', '기구', '공시', '작성', '결론', '포함', '기간', '사용', '재무', '현재', '대해', '개별', '경제', '목적', '업무', '진과', '정보', '요약', '단락', '발행', '구심', '항상', '화가', '서일', '이사', '부로', '모든', '로부터', '여러', '백만원', '금창', '방법', '일자', '변수', '파악', '완전', '주식회사', '이하', '재무상태표', '자본변동표', '현금흐름표', '손익계산서', '회계', '정책', '주석', '동일로', '핵심'

Table 1. 불용어

4. 분석방법

4.1 워드클라우드

텍스트의 빈도 기반 주요 키워드를 파악하기 위해 우량기업과 부실기업 텍스트로 워드클라우드를 만들었으며 Figure 2와 같다. 워드클라우드 상으로는 우량기업과 부실기업의 핵심키워드가 잘 드러나지 않는다고 판단되는데 이는 토큰화 할 때 주요 키워드가 한단어로 묶이지 않고 토큰화되었기 때문이라고 생각된다. 예를 들면, 코스닥 상장폐지실질심사위원회의 경우 '코스닥', '상장', '폐지', '심사', '위원회'로 분리되어 단어 고유의 의미가 사라져 버리게 된다.

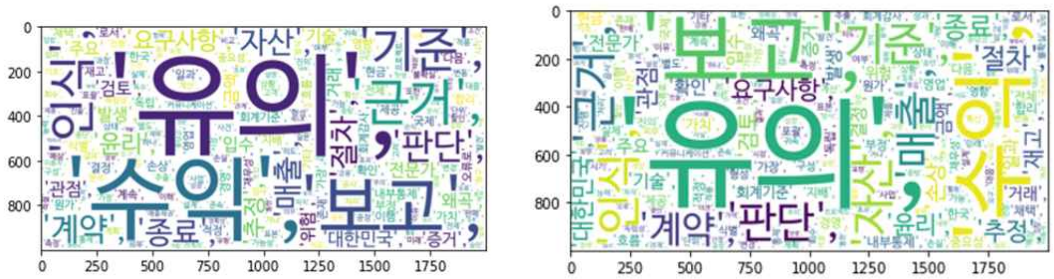

Figure2. 우량기업 텍스트의 워드클라우드(좌측) 부실기업 텍스트의 워드클라우드(우측)

4.2 sklearn의 LDA를 사용한 토픽모델링

sklearn에서 제공하는 LDA라이브러리를 사용하여 토픽모델링을 시행하였다. 전체 단어의 5%이상으로 많이 나오는 단어 및 개의 문서 미만으로 등장하는 단어는 제외하고 우량기업 및 부실기업의 토픽을 각각 5개씩 추출하였다. 이 경우에도 추출된 토픽이 기업의 성격을 나타낸다고 보기는 어려우며 이는 워드클라우드와 마찬가지로 텍스트 토큰화의 문제에 기인한다고 판단된다.

구분	토픽
우량기업	Topic #1 회생 건설 개시 법원 제품 소송 조사 감가상각 이내 기존 Topic #2 한정 억원 회차 담보 연구개발 주주 경상 최대 코스 재화 Topic #3 처분 충당 조사 실사 게임 부외 요청 실현 전환 입회 Topic #4 실체 회생 인수 합병 종결 신규 보통주 결의 무형 유상증자 Topic #5 원가 개발 의무 공사 건설 청구 용역 대손충당금 투입 신주인수권
부실기업	Topic #1 재화 억원 용역 처분 대손충당금 매입 대가 유통 신규 장비 Topic #2 실체 원가 한정 실현 공사 충당금 건설 청구 투입 가격 Topic #3 회생 건설 충당 개시 감가상각 종결 부회 관리제 거래처 의무 Topic #4 개발 소송 연구개발 경상 회차 주주 담보 최대 사채 바이오 Topic #5 회생 조사 처분 법원 인수 아래 신청 보통주 서술 지연

Table 2. 우량기업 및 부실기업의 토픽

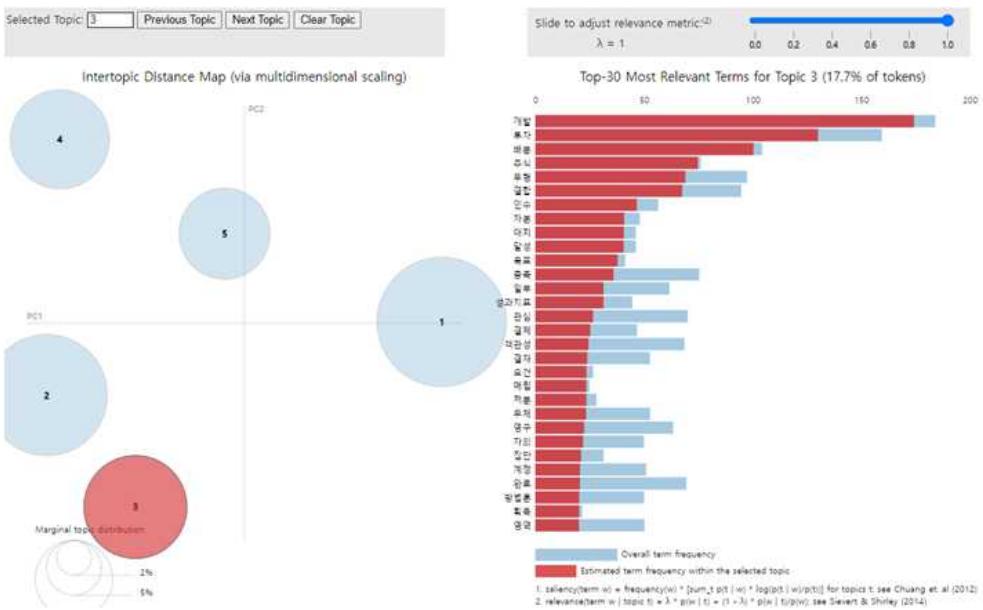

Figure3. LDA 시각화 예시(우량기업)

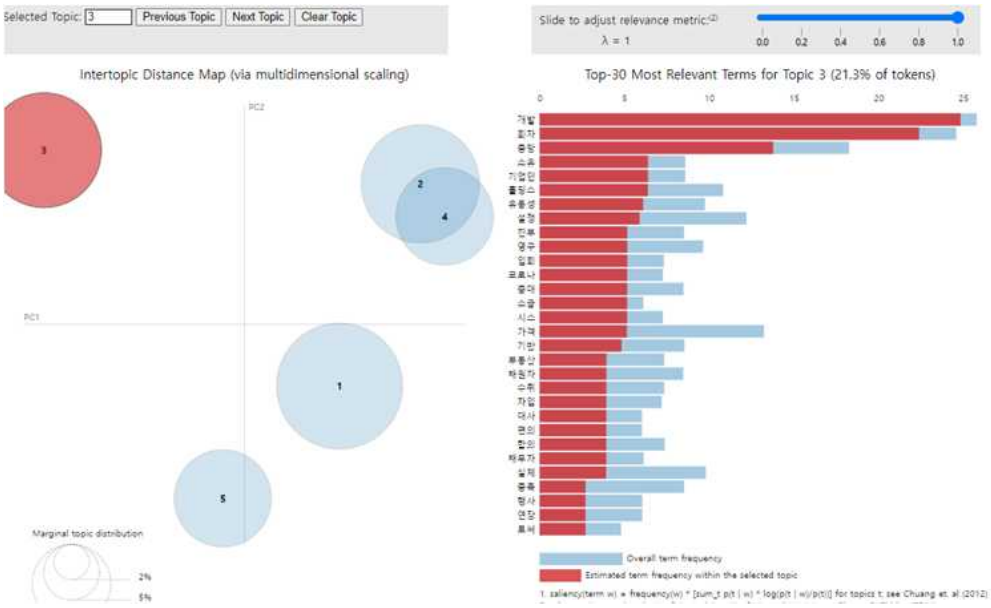

Figure4. LDA 시각회 예시(부실기업)

4.3 topotopy의 DMR을 사용한 토픽모델링

tomotopy는 토픽모델링과 관련하여 다양한 라이브러리를 제공하며 명목변수(우량기업, 부실기업)에 따라 토픽비율이 어떻게 다른지 판단하기 위해 DMR(Dirichlet Multinomial Regression)을 사용하여 분석하였다. 추출된 토픽을 살펴보면 1번째 토픽은 계속기업 불확실성과 관련된 주제이며 부실기업의 비율이 높았다. 2번째 토픽은 경영진 커뮤니케이션과 관련된 주제이며 우량기업의 비율이 높았다. 3번째 토픽은 기업활동 전반과 관련된 주제이며 우량기업의 비율이 높았다. 4번째 토픽은 계약가치 변동과 관련된 토픽이며 부실기업의 비율이 높았다. 5번째 토픽은 주요사업의 손익과 관련된 주제이며 부실기업의 비율이 높았다. 6번째 토픽은 기업활동과 관련된 주제로 우량기업의 비율이 높았다. 7번째 토픽은 수익인식과 관련한 주제로 우량기업의 비율이 높았다. 8번째 토픽은 자본흐름과 관련된 주제로 우량기업의 비율이 높았다. 9번째 토픽은 자본잠식과 관련된 주제로 부실기업의 비율이 높았다. 10번째 토픽은 기업윤리와 관련된 주제이며 부실기업의 비율이 높았다. 2, 3, 8, 10번째 토픽의 경우 우량기업과 부실기업의 비율차이가 크지 않으나 나머지 토픽에서는 큰 차이를 보여 두 기업군의 감사의견상 차이가 있다고 볼 수 있다.

Topic #0(계속기업 불확실성) 전기, 영향, 반영, 강조, 미치다, 불확실
Topic #1(경영진 커뮤니케이션) 중요, 지배, 계속, 커뮤니케이션, 합리, 경영진
Topic #2(기업활동) 판매, 고객, 측정, 정확, 계산, 처리, 활동
Topic #3(계약가치 변동) 계약, 변동, 금액, 추정, 발생, 산정, 확실
Topic #4(주요 사업 손익) 공정, 측정, 사업, 수준, 손익, 외부, 전문가, 적격
Topic #5(기업활동) 충분, 계속, 증거, 범위, 표명, 백만, 적합, 거래
Topic #6(수익인식) 매출, 수익, 인식, 귀속, 추출, 표본, 이전, 시점, 고객, 발생
Topic #7(자본흐름) 손상, 현금, 창출, 할인, 가치, 추정, 가정, 경영진, 흐름

Topic #8(자본잠식) 자산, 가능, 관계, 회수, 검토, 계산, 확인, 검증, 장부, 금액 Topic #9(기업윤리) 거래, 관점, 동일, 기준, 다루다, 지다, 종료, 윤리

Table 3. 추출된 토픽(파란색은 우량기업 비율이 높은 것, 빨간색은 부실기업 비율이 높은 것)

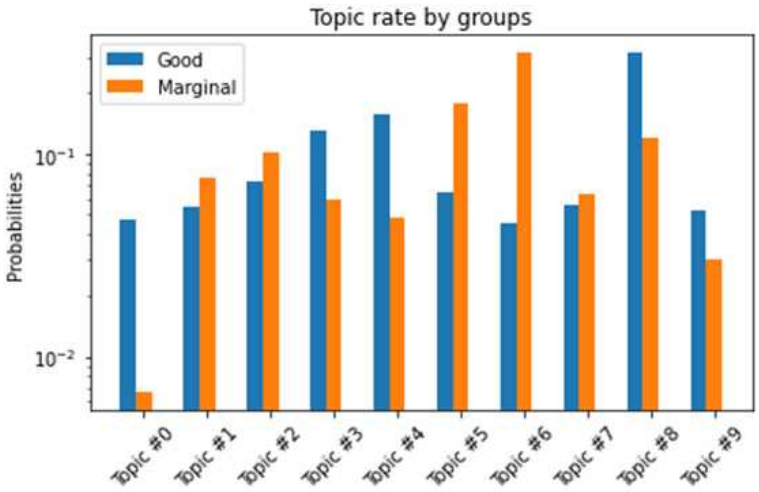

Figure 5. 명목변수(우량기업, 부실기업)에 따른 토픽비율 비교

4.4 Linear Regression

수집 데이터중 수치데이터는 기업 주식의 액면가와 상장주식수이다. 이론적으로 둘 간의 관계가 있는 것은 아니나 액면가가 상장주식수를 설명하는 유의한 변수인지 확인하기 위해 Linear Regression 분석을 시도하였다. 분석 결과 상장주식수를 설명하는데 액면가는 적절한 변수가 아님을 알 수 있다. 모형의 설명력이 낮으며 액면가 변수의 p-value가 매우 높아 액면가 변수의 계수가 0이라는 귀무가설을 기각하지 못한다.

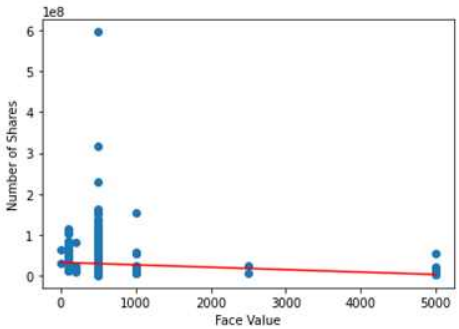

Figure 6. 액면가와 상장주식수의 Scatter Plot

OLS Regression Results

Dep. Variable:	상장주식수	R-squared:	0.005
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	2.674
Date:	Thu, 02 Jun 2022	Prob (F-statistic):	0.103
Time:	03:26:48	Log-Likelihood:	-9758.0
No. Observations:	516	AIC:	1.952e+04
Df Residuals:	514	BIC:	1.953e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.294e+07	2.59e+06	12.734	0.000	2.79e+07	3.8e+07
액면가	-5906.4564	3611.976	-1.635	0.103	-1.3e+04	1189.596

Omnibus:	730.504	Durbin-Watson:	2.066
Prob(Omnibus):	0.000	Jarque-Bera (JB):	167238.297
Skew:	7.319	Prob(JB):	0.00
Kurtosis:	89.973	Cond. No.	1.06e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.06e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 7. Linear Regression 결과

5. 결론

본 연구에서는 기존의 재무정보 등 수치데이터로 한계기업을 모델링 한 것과 달리 텍스트 정보로 우량기업과 부실기업을 구분할 수 있는지 알아보고자 토픽모델링을 실시하였다. 텍스트 데이터에 대해 전처리를 시행한 후 sklearn의 LDA와 topotopy의 DMR을 실시하였다. LDA 분석 결과 추출된 토픽이 기업의 속성을 잘 대변한다고 볼 수 없으며 이는 토큰화로 인해 특별한 의미를 갖는 단어가 여러 단어로 쪼개졌기 때문이라고 볼 수 있다. DMR 분석 결과 두 기업군 간에 토픽 비율의 차이가 있는 토픽들을 볼 수 있었다. 토픽모델링 분석 결과 우량기업과 부실기업의 감사의견 텍스트 상 약간의 차이가 있다고 볼 수 있다. 본 연구에서 전처리 과정상 불필요한 토큰화를 방지하기 위해 사용자사전을 구축하지 못한 한계가 있다. 또한 다양한 형태소분석기를 사용하지 못한 한계도 있다. 추후 연구에서는 여러 형태소분석기로 토큰화를 시도해보고 섬세한 전처리를 통해 정교한 토픽모델링을 실시할 필요가 있다.

수업에 대한 피드백

(좋았던 점) 데이터 분석에 관련하여 전반적인 내용을 배운 점이 좋았습니다. 다른 과목에서 알려주지 않는 '가려운 부분'을 많이 가르쳐 주셨습니다. 유닉스 언어, 깃헙 사용법, 정규표현식 등을 배워서 평소 알고 싶었던 내용을 많이 배웠습니다. 조별과제로 프로젝트를 진행한 점도 정말 많이 배울 수 있는 부분이었습니다. 조별로 어려움이 있을 때마다 개인시간을 많이 내어주셔서 도와주신 점도 정말 감사했습니다.

(개선되면 좋을 것 같은 사항)

①실습과 과제가 많은 수업이다 보니 성대에서 추천하는 flipped learning 방식도 좋을 것 같습니다. 전반적인 내용과 이론에 대해서는 미리 읽거나 동영상자료를 시청하고 오고 수업시간에는 모르는 내용질문과 실습 위주로 하면 오류해결도 잘 되고 자신감도 더 얻을 수 있을 것 같습니다.

② 정말 많은 영역을 커버해 주셔서 한국어 텍스트 분석과 관련된 내용을 시간관계상 많이 다루지 못한 점이 아쉽습니다. 형태소분석기, 단어임베딩, 사전구축 등의 내용이 다뤄지면 추후 텍스트 분석 문제를 해결할 때 큰 도움이 될 것 같습니다. 키워드 추출, 토픽모델링, 감성분석 등의 모델도 배우면 더더욱 좋을 것 같습니다.

③ 조별활동이 있을 것이라고 처음부터 말씀해주셔서 준비할 수 있었습니다. 막판에 몰아서 하지 않도록 중간에 작은 과제를 던져주시고 동기부여 해주셔서 도움이 많이 되었습니다. 교수님의 노력에도 불구하고 저를 포함해서 많은 학우들이 마지막에 집중해서 한 것 같은데요, 주제 선정을 학기 초에 하고 중간발표 때 어느 정도 진행된 것을 발표하고 어려웠던 점을 교수님께서 도와주시면 학기말에 보다 더 완성도 높은 프로젝트가 나올 것 같습니다.

** 교수님 덕분에 몰랐던 부분도 많이 알게 되고 앞으로 갈 길이 멀지만, 자신감도 많이 생겼습니다. 텍스트 분석에도 관심이 많이 생겨 앞으로도 수업도 듣고 논문이나 프로젝트에도 참여해보고 싶은 생각이 많이 들었습니다. 이렇게 또 여러 학생들을 변화시켜주시네요! 지금까지 대학원생활에서 가장 묵직한 경험을 하게 해주셔서 감사드립니다.