아시아교육연구 20권 2호
Asian Journal of Education
2019, Vol. 20, No. 2, pp. 491-522.
https://doi.org/10.15753/aje.2019.06.20.2.491

토픽 모델링을 활용한 대학생의 서술형 강의평가 분석

곽민호(郭旻昊)*
민혜리(閔惠利)**
김미림(金美林)***

논문 요약

이 연구는 토픽 모델링(topic modeling)의 일종인 잠재 디리클레 할당(latent Dirichlet allocation, 이하 LDA)을 활용하여 S대학교의 학생들이 작성한 강의평가 응답을 분석함으로써 학생들이 갖고 있는 강의에 대한 생각을 보다 직접적으로 알아보고자 하였다. 이를 위해 2015년 1학기에 개설된 약 1,500개 강의에 대해 학생들이 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대해 서술한 약 47,000개의 응답 내용을 LDA를 활용해 분석하였다. 아울러, 6개의 단과대학(공과대학, 농업생명과학대학, 사범대학, 인문대학, 사회과학대학, 자연과학대학) 강의의 개선되어야 할 점, 좋았던 점에 대해 분석하였다.

분석 결과, 첫째, '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점' 모두 3개 주제 모형이 가장 적합한 것으로 나타났다. 먼저, '강의에서 개선되어야 할 점'은 1) 과제·실험·실습에 대한 개선사항, 2) 발표·토론에 대한 개선사항, 3) 시험·진도·수업내용에 대한 개선사항의 세 가지 주제로 나타났다. 다음으로, '강의에서 좋았던 점'은 1) 교수자·교수방법에 대한 긍정적 피드백, 2) 직접적 경험·실습에 대한 긍정적 피드백, 3) 강의내용에 대한 긍정적 피드백의 세 가지 주제로 나타났다. 둘째, 단과대학별 분석 결과, 단과대학별로 나타난 주제의 의미는 대체적으로 전체 대학 자료를 분석했을 때와 비슷했으나, 하나의 주제 정도가 단과대학의 특성을 반영하고 있는 것으로 나타났다.

이 연구는 강의평가의 선택형 문항 분석에 치중하였던 기존 연구와 달리, 토픽 모델링을 활용함으로써 대량의 서술형 강의평가 자료를 효율적으로 요약하였으며, 이를 통해 강의 전반에 대한 학생들의 인식을 보다 직접적이고 종합적으로 살펴볼 수 있었다는 의의를 갖는다.

주요어 : 잠재 디리클레 할당, 토픽 모델링, 텍스트 분석, 강의평가

* 제1저자, 조지아대학교 박사수료
** 공동저자, 서울대학교 교수학습개발센터 연구교수
*** 교신저자, 한국교육과정평가원 부연구위원

Ⅰ. 서론

강의평가는 노동시장으로의 인력 양성을 반영하는 취업률과 함께 고등교육기관의 핵심적인 평가 지표로 인식되어 왔다. 특히, 강의평가는 수요자인 학습자의 입장에서 평가한 강의의 내용, 방법, 목적 등에 대한 만족도라는 점에서 학습권과 관련하여 중요한 의미를 갖기 때문에(한신일, 김혜정, 이정연, 2005), 평가 이후의 강의 방향을 설정함에 있어 주요하게 고려되어야 할 요소라 할 수 있다. 그러나 일반적으로 대학에서 실시되는 강의평가는 리커트 척도를 활용한 선택형 문항이 주를 이루며, 응답을 선택할 때 수강생 수나 강의실 분위기 등 다양한 요소가 영향을 미칠 수 있다는 점에서 강의 자체에 대한 엄밀한 평가로 보기 어렵다(최정웅, 안동규, 2016). 더불어, 선택형 문항에 대한 학생들의 응답이 피상적이거나 무성의한 경우 강의평가 자료의 신뢰도가 크게 저해된다는 문제를 갖고 있다(권오영 외, 2014; 김학일 외, 2007). 이와 관련하여, 강의평가 문항에 대한 일관적 응답을 연구한 선행연구(김명화, 2005; 양길석, 2014; 홍경선, 2006)는 50% 이상의 학생들이 일관적 응답을 보였다고 공통적으로 보고하였다.

이러한 선택형 문항의 제한점을 극복하기 위해 대부분의 대학에서는 강의평가 마지막에 강의에 대한 전반적인 평가를 묻는 서술형 문항을 사용한다(한신일, 2003). 이러한 서술형 문항은 질적인 접근방법을 통해 데이터를 수집하는 방법이며, 학생들의 강의에 대한 인식을 직접적으로 드러내는 중요한 자료로 볼 수 있다. 또한, 서술형 문항은 선택형 문항에서 발생하는 신뢰도 및 타당도 문제에서 상대적으로 덜 취약하고, 자유로운 의견을 서술할 수 있다는 점에서 유익한 피드백을 제공할 수 있다(한신일, 2003). 따라서 학생들의 의견이 선택형 문항에서 얻어진 내용과 큰 차이를 보이지 않는다 하더라도, 이러한 결과가 갖는 신뢰도와 타당도는 보다 높을 것으로 예상된다. 같은 맥락에서, 강의평가 문항에서 영역별 중요도에 대해 묻지 않는 이상, 학생들이 강의평가 영역에서 어떠한 부분에 더 관심을 두는지를 알기는 쉽지 않다. 그러나 서술형 문항에서는 지면 제약 상 학생들이 주요하게 생각하는 부분에 대해 주로 서술할 것으로 생각되며, 이를 통해 학생들이 강의의 여러 영역 중에서 특히 어떤 부분에 관심을 두는지를 파악할 수 있을 것으로 기대된다.

그러나 서술형 문항에 대한 응답과 같은 텍스트 자료 분석은 단순한 빈도 분석 방법이나, 인터뷰 조사와 같은 질적인 방법을 활용하는 등 분석 방법이 제한되었던 것이 사실이다. 이러한 맥락에서, 최근 각광받는 데이터 마이닝(data mining) 기법 중에서 텍스트 마이닝(text mining)은 그간 분석이 제한되었던 텍스트 자료 자체를 다룰 수 있다는 점에서 주목할 만하다. 특히, 토픽 모델링(topic modeling)은 잠재 시멘틱 분석(latent semantic analysis; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), 자동 스코어링(automated scoring; Burstein, Leacock, & Swartz, 2001) 등 다양한 텍스트 분석 방법 중 가장 최근에 소개된 텍스트 분석 기법이다. 잠재 디리클레 할당(latent Dirichlet allocation, 이하 LDA; Blei, Ng & Jordan, 2003)으로 대표되는 토픽 모델링은

대규모의 텍스트 자료를 전체적으로 조망할 수 있는 분석 결과를 제시한다. 또한, LDA는 대규모 텍스트 자료를 몇 가지 주제의 집합으로 분석해 낼 수 있을 뿐만 아니라, 개별 문서의 주제 구성에 대한 정보를 제공한다는 점에서 전체적인 주제 구성은 물론, 각 문서의 주된 주제를 파악할 수 있다는 장점을 갖고 있다.

이에 따라 이 연구에서는 LDA를 활용하여 S대학교 대학생의 2015년 1학기 강의에 대한 서술형 강의평가 자료를 분석하였다. 참고로, S대학교의 강의평가는 총 2개의 서술형 문항을 포함하며, 각각 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대해 묻는다. 이 연구에서는 각 서술형 문항에 나타난 주요 주제를 찾아내고 이를 바탕으로 대량의 강의평가 응답을 효율적으로 요약하고자 하였다. 또한, 전체 대학의 강의평가 자료를 분석함과 동시에 단과대학별 강의평가 자료를 분석하여 단과대학에 따른 강의평가의 특성을 파악하고자 하였다. 이를 통해 S대학교 강의 전반에 대한 직접적 평가뿐만 아니라, 각 단과대학별 강의에 대한 직접적 평가를 보다 효율적으로 요약할 수 있을 것이며, 이는 추후 강의 개선에 있어 주요한 기초자료로서 활용될 수 있을 것으로 기대된다. 구체적인 연구문제는 다음과 같다.

첫째, '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 응답에서 나타난 주요 주제는 무엇인가?

둘째, 이러한 각 문항별 주요 주제는 단과대학에 따라 어떠한 양상을 나타내는가?

Ⅱ. 이론적 배경

1. LDA

이하에서는 먼저 LDA의 개념 및 모형 정의를 통해 LDA를 간단히 소개한 다음, 모수 추정 방법과 이와 관련한 사전정보 선택에 대해 정리하고, 마지막으로 주제 수를 결정하는 방식에 대해 다루고자 한다.

1) LDA의 개념 및 모형 정의

LDA는 Blei, Ng, & Jordan(2003)에 의해 제안된 토픽 모델링 방법이다. 이 분석 방법은 잠재 시멘틱 색인(latent semantic indexing; Deerwester et al., 1990) 등과 같은 기존의 텍스트 마이닝 방법과 달리 확률적인 모형이라는 점에서 큰 차이를 갖는다.

또한 LDA는 다항분포와 다항분포의 켤레 사전분포(conjugate prior distribution)인 디리클레 분

포가 모형 전반에서 핵심적으로 기능한다. 보다 구체적으로, LDA는 말뭉치(corpus)를 문서(document)의 집합으로 간주하고, 문서는 여러 주제의 집합으로 간주한다. 예를 들어, 강의평가에 대한 S대학교 전체 학생의 응답은 개별 학생 응답의 집합으로 볼 수 있고, 개별 학생이 생성한 문서는 여러 주제의 집합으로 나타낼 수 있다.

좀 더 구체적인 예를 제시하기 위해 다음과 같은 상황을 가정한다. 먼저, 전체 문서에서 총 3개의 주제가 추출되었고, 각 주제는 ① 과제와 실험에 대한 평가, ② 교수에 대한 평가, ③ 시험에 대한 평가로 정의되었다. 즉, S대학교 학생들의 강의평가는 크게 3개 주제로 요약된다. 또한, 개별 학생이 생성한 문서는 위의 세 주제가 서로 다른 비율로 구성되어 있을 수 있다. 예를 들어, 학생 A의 경우, 전체 내용의 50%를 과제와 실험에 대한 평가로 채우고, 교수에 대한 평가를 25%, 나머지 25%를 시험에 대한 평가로 구성하였을 수 있다. 반면 학생 B는 전체 내용의 5%만을 과제와 실험에 대한 평가에 할애하고, 교수에 대한 평가를 20%, 나머지 75%를 시험에 대한 평가로 구성하였을 수 있다.

이를 통계적으로 설명하면 다음과 같다. 먼저, 말뭉치, 문서, 주제, 단어는 특정한 위계를 이루고 있다. 보다 구체적으로, 하나의 말뭉치는 여러 문서의 집합이고, 하나의 문서는 여러 주제의 집합이며, 하나의 주제는 여러 단어와 각 단어에 대응하는 확률 질량(probability mass)으로 나타난다. 이를 다항분포(multinomial distribution)에 대응해보면, 각 주제가 지지집합(support)을 이루며, 그에 해당하는 확률 질량을 갖는다. 따라서 문서의 특성(예: 특정 주제에 대해 서술한 양)에 대한 판단은 각 주제에 대응하는 확률 질량에 따라 달라진다.

위 단락에서 설명한 예시를 다항분포를 이용해 다시 표현하면, 학생들이 서술한 문서의 특성은 3개 주제인 과제와 실험에 대한 평가, 교수에 대한 평가, 시험에 대한 평가에 확률 질량이 대응되는 방식에 따라 결정된다고 볼 수 있다. S대학교 학생들의 강의평가에 대한 문서별 주제 분포를 벡터로 표현하면 [그림 1]과 같다.

$$(\text{문서별 주제 분포}) = \begin{pmatrix} \text{'과제와 실험에 대한 평가' 주제에 대응되는 확률 질량} \\ \text{'교수에 대한 평가' 주제에 대응되는 확률 질량} \\ \text{'시험에 대한 평가' 주제에 대응되는 확률 질량} \end{pmatrix}$$

[그림 1] 문서별 주제 분포

이와 같은 벡터표기법을 사용했을 때, 학생 A의 문서별 주제 분포는 $(0.5, 0.25, 0.25)^T$로, 학생 B의 문서별 주제 분포는 $(0.05, 0.20, 0.75)^T$로 표현된다. 이러한 문서별 주제 분포는 LDA의 주요한 두 가지 모수 중 하나이며, 학생 수에 해당하는 만큼의 개수가 추정된다.

또 다른 주요 모수는 주제별 단어 분포이다. 앞서 LDA에서 주제가 정의되는 방식은 단어와 각 단어에 대응하는 확률 질량이라고 언급하였다. 이 역시 문서별 주제와 같이 다항분포를 이용하여

설명될 수 있다. S대학교 학생들의 강의평가에 대한 주제별 단어 분포를 벡터로 표현하면 [그림 2]와 같다.

$$('과제와 실험에 대한 평가'주제) = \begin{pmatrix} '과제'단어에 대응되는 확률 질량 \\ '실험'단어에 대응되는 확률 질량 \\ '평가'단어에 대응되는 확률 질량 \\ \cdots \\ '피드백'단어에 대응되는 확률 질량 \end{pmatrix}$$

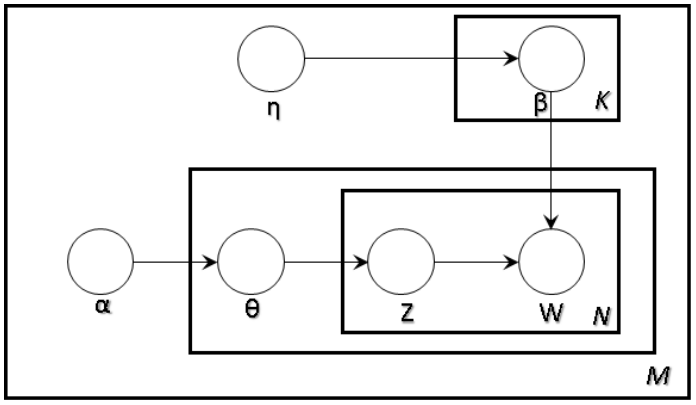
[그림 2] 주제별 단어 분포

'과제와 실험에 대한 평가' 주제에 속하는 단어를 확률 질량에 따라 오름차순으로 정리하면, [그림 2]와 같은 결과를 얻을 수 있다. 즉, '과제와 실험에 대한 평가'라는 주제를 구성하는 단어는 전체 말뭉치에 사용된 단어와 일치하지만, 그에 대응되는 확률 질량을 살펴보면, '과제', '실험', '평가'와 같은 단어에 높은 확률 질량이 대응되었고, 이를 바탕으로 이 주제를 '과제와 실험에 대한 평가'라고 주제를 특정 지을 수 있게 된다.

한편, LDA 분석에서 염두에 두어야 하는 세 가지 요소는 다음과 같다. 첫째, LDA의 주제 분석의 결과는 두 가지 모수인 문서별 주제 분포와 주제별 단어 분포 두 가지로 요약되며, 이 두 모수는 모두 다항분포로 나타난다는 점이다. 둘째, LDA 분석은 주제별 단어 분포와 문서별 주제 분포를 추출할 뿐, 주제명을 정해주지는 않는다. 이는 연구자가 주제별 단어 분포를 바탕으로 직접 결정해야 한다. 예를 들어, 위의 예시에서 사용된 세 가지 주제인 '과제와 실험에 대한 평가', '교수에 대한 평가', '시험에 대한 평가'는 실제 분석에서는 주제 1, 주제 2, 주제 3과 같이 표현되었으며, 연구자가 주제별 단어 분포를 참고하여 주제에 가장 적합한 설명을 부여한 것이다. 이 과정은 주제가 명확하게 추출되지 않을 경우, 모호성을 동반할 수 있다. 셋째, LDA 분석은 베이지언 방법에 기반을 두고 있으며, 사전정보(prior)를 요구한다. 이 사전정보를 어떻게 부여할 것인지에 대한 선행연구는 일관된 결과를 제시하지 못하였고(Chang, 2010; Thomas et al., 2014), 최근의 시뮬레이션 연구에서 기존에 제시된 사전정보의 타당성을 검토하였으나, 그 연구 결과의 적용은 표본의 크기와 사용된 어휘의 수에 따른 제한이 있다(Kwak, Kim, & Cohen, 2017).

LDA 모형은 생성적 모형(generative model)의 관점에서 정의될 수 있는데, 이때의 생성적 모형은 차별적 모형(discriminative model)과 대비되는 개념으로 사용된다. 생성적 모형은 은닉변수(hidden variable)가 주어진 상태에서 투입과 산출을 동시에 생성해내는 반면, 차별적 모형은 투입을 통해 산출을 추론해낸다. LDA, 가우시안 혼합 모형(Gaussian mixture model), 은닉 마르코프 모형(hidden markov model) 등이 생성적 모형에 포함되며, 로지스틱 회귀분석, 서포트 벡터 머신(support vector machine), 신경망분석이 차별적 모형에 포함된다.

일반적으로 생성적 모형은 생성 과정(generative process)과 판 표기법(plate notation)으로 표현된다. 먼저, LDA의 판 표기법은 [그림 3]과 같다.



[그림 3] LDA의 판 표기법(Blei, Ng, & Jordan(2003)에서 재구성)

이상의 판 표기법을 다음의 생성 과정으로 구체화시킬 수 있다. 생성 과정에서는 사전정보와 LDA의 주요 모수들이 실제 문서 생성에 관여하는 방식과 과정이 보다 상세히 제시된다.

(1) η를 모수로 하는 디리클레 분포에서 β_k 선택.

(2) α를 모수로 하는 디리클레 분포에서 θ_d 선택.

(3) 문서 d에 속하는 각 단어 $w_{d,n}$에 대해,

(3-1) θ_d를 모수로 하는 다항분포에서 주제 배당 $z_{d,n}$ 선택.

(3-2) β_k를 모수로 하는 다항분포에서 단어 배당 $w_{d,n}$ 선택.

먼저, 첫 번째 단계인 β_k를 선택하는 단계를 살펴보면 다음과 같다. β_k는 LDA의 주요한 두 가지 모수 중 하나인 주제별 단어 분포로, η를 모수로 하는 디리클레 분포에서 표집된다. β_k는 벡터로서, 말뭉치 전체가 총 V개의 단어로 이루어져있다면, β_{k_1}부터 β_{k_V}까지의 집합으로 볼 수 있다. 보다 간단히 표현하면 $\beta_k = (\beta_{k_1}, \ldots, \beta_{k_V})$와 같다. 즉, 각 β_{k_v}는 주제 k에 대한 V개 단어에 대응하는 확률벡터이다. 따라서 β_k는 전체 주제 수 K만큼 존재한다.

다음으로, 두 번째 단계인 θ_d를 선택하는 단계를 살펴보면 다음과 같다. θ_d는 LDA의 주요한 두 가지 모수 중 하나인 문서별 주제 분포로, α를 모수로 하는 디리클레 분포에서 표집된다. θ_d는 벡터로서, 말뭉치가 총 d개의 문서로 이루어져 있다면, θ_{d_1}값부터 θ_{d_K}값까지의 집합으로 볼 수 있다. 보다 간단히 표현하면 $\theta_d = (\theta_{d_1}, \ldots, \theta_{d_K})$와 같다. 즉, 각 η_{d_k}는 문서 d에 대한 K개 주제에 대응하는

확률벡터이다. 따라서 θ_d는 전체 문서 수 D만큼 존재한다.

이후, 선택된 두 다항분포에 따라 주제 배당과 단어 배당을 선택하는 단계를 수행한다. 먼저, θ_d를 모수로 하는 다항분포에서 주제 배당 $z_{d,n}$를 선택한다. 이 과정을 주제 할당만으로 이루어진 문서가 생성된다. 마지막으로, β_k를 모수로 하는 다항분포에서 단어 배당 $w_{d,n}$를 선택하여, 최종적인 문서 집합이 생성된다.

이를 요약하면, 두 모수 β_k와 θ_d는 각각 주제별 단어 분포, 문서별 주제 분포를 의미하며, LDA의 직접적인 결과이다. LDA는 베이지언 확률론에 기반을 둔 모형이기 때문에 두 모수의 추정은 데이터뿐만 아니라, 사전정보 α와 η에 의해서도 영향을 받는다. 구체적인 모수 추정 방법과 사전정보의 선택은 다음 절에서 보다 구체적으로 살펴보고자 한다.

2) 모수 추정 방법 및 사전분포 선택

LDA와 관련된 주요한 선행연구에서는 크게 볼록 기반 변분 알고리즘(convexity-based variational algorithm; Blei, Ng, & Jordan, 2003)과 깁스 표집(gibbs sampling; Griffiths & Steyvers, 2004)의 두 가지 모수 추정 과정을 제시하고 있다. 이 연구에서는 깁스 표집을 주된 모수 추정 방법으로 선택하였는데, 그 이유는 모형 평가에 필요한 정보를 보다 간단하게 얻을 수 있기 때문이며, 베이지언 모형의 사후분포(posterior distribution)를 추정하는 데 있어서 가장 일반적으로 사용되는 방법이기 때문이다.

Kwak, Kim, & Cohen(2017)의 연구에서 Heinrich(2008)가 유도한 깁스 표집기 유도 과정을 보다 간단하게 정리하였으며, 이는 아래와 같다.

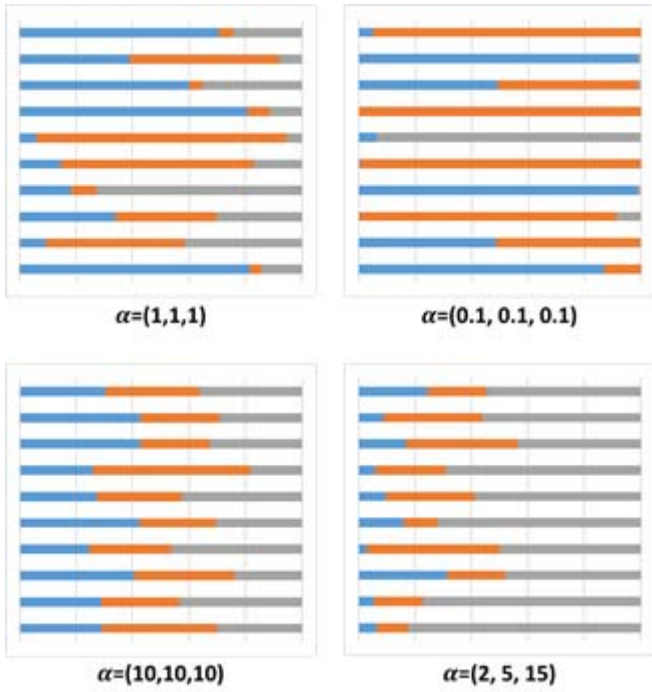
$$p(z_i = k|z_{\neg i}, w) \propto \left[\frac{n_{k,\neg i}^{(v)} + \eta_v}{\left[\sum_{v=1}^{V} n_{k,\neg i}^{(t)} + \eta_v \right]} \right] \left[\frac{n_{d,\neg i}^{(v)} + \alpha_k}{\left[\sum_{v=1}^{K} n_{d,\neg i}^{(k)} + \alpha_k \right] - 1} \right]$$

이때, $p(z_i = k|z_{\neg i}, w)$는 현재 i번째 주제 배당에 대한 조건부 확률이며, 구체적으로 이는 현재 i번째 주제 배당을 제외한 나머지 주제 배당과 현재까지의 모든 단어 배당이 주어진 상태를 조건으로 했을 때 현재 주제 배당이 주제 k로 배당될 확률을 의미한다. $n_{k,\neg i}^{(t)}$는 현재 i번째 배당을 제외한 이전 배당들에서 주제 k로 배당된 단어 v의 빈도를 의미한다. 따라서 $\sum_{v=1}^{V} n_{k,\neg i}^{(v)}$는 현재 i번째 배당을 제외한 이전 배당들에서 주제 k로 결정된 단어들의 총합을 의미한다. $n_{d,\neg i}^{(k)}$는 현재 i번째 배당

이전에 문서 d에서 k 주제로 결정된 단어의 수를 의미한다. 따라서 $\sum_{k=1}^{K} n_{d,\neg i}^{(k)}$는 i번째 배당을 제외하고, 문서 d에서 주제가 결정된 단어의 수를 의미한다.

LDA 모형에서 사전분포는 디리클레 분포이며, 각 분포는 α와 η에 의해 결정된다. 일반적으로 디리클레 분포는 다항분포의 켤레분포로써 혼합 문항반응이론(mixture item response theory)에서도 널리 사용된다. LDA 모형에서 α와 η의 영향을 직관적으로 설명하면 다음과 같다. 먼저, α와 η는 각각 문서별 주제 분포 및 주제별 단어 분포에 영향을 미치며, 비록 다른 분포이나 영향을 미치는 방식은 같다. 다시 말해, α와 η가 커지거나 작아짐에 따라 문서별 주제 분포, 주제별 단어 분포에 같은 영향을 미치게 된다. 만약 α와 η가 커지게 되면, 다항분포의 각 카테고리에 해당되는 확률 질량의 값이 유사해진다. 이와 반대로 α와 η가 작아지게 되면, 다항분포의 각 카테고리에 해당되는 확률 질량의 값이 매우 달라진다. 이를 그림으로 표현하면 [그림 4]와 같다. [그림 4]에서는 디리클레 모수 α에 의해 10번의 다항분포의 표집에 따른 확률 질량 분포 결과가 어떻게 변하는 지를 설명하고 있다. 왼쪽 하단부의 그림처럼 α가 $(10, 10, 10)$과 같이 매우 큰 값을 갖게 되면, 파랑, 주황, 회색으로 구분된 세 범주의 비율이 거의 비슷하게 표집되는 것을 볼 수 있다. 반면, 오른쪽 위의 그림처럼 α의 값이 $(0.1, 0.1, 0.1)$과 같이 매우 작은 값을 갖게 되면, 세 범주의 비율이 매우 다르게 되고, 특정 범주가 매우 높거나 낮은 분포의 모양을 가지게 된다.

이를 LDA의 문서별 주제 분포, 주제별 단어 분포에 비추어 설명하면 다음과 같이 설명할 수 있다. 먼저, 문서별 주제 분포의 경우, 이에 해당하는 사전분포 모수 α가 큰 값(예: 10, 15 등)을 가지게 되면, 하나의 문서 내에서 여러 주제의 비율을 유사하게 만들어 낸다. 반면에, 사전분포의 모수 α가 작은 값(예: 0.1, 0.001 등)을 가지게 되면, 하나의 문서가 특정 하나의 주제로 대부분 채워지는 형태의 분포를 만들어 낸다. 즉, 사전분포의 모수 α가 크면, 하나의 문서가 다양한 주제로 구성되게 되고, 사전분포의 모수 α가 작으면, 하나의 문서가 극히 제한된 주제로 구성되게 된다.



[그림 4] 디리클레 분포의 모수에 따라 달라지는 세 카테고리 분포의 비교
(Frigyik, Kapila, & Gupta(2010)에서 재구성)

주제별 단어 분포에서 η의 영향 역시 기본적으로 α의 영향과 같다. 만약 η의 값이 크면, 하나의 주제 내에서 단어들의 확률 질량이 서로 유사한 값을 갖게 된다. 반면에, η의 값이 작으면, 하나의 주제 내에서 단어들의 확률 질량이 매우 다르게 된다. 즉, η의 값이 클수록 주제들 간의 유사성이 높게 되고, η의 값이 작을수록 주제들이 서로 명확히 구분된다고 볼 수 있다.

이처럼 LDA에서 사전분포의 모수 α와 η는 주제별 단어 분포와 문서별 주제 분포의 추정에 명확히 영향을 미친다. 그러나 어떠한 사전분포 정보가 적절한지에 대해서는 선행연구마다 조금씩 다르다. 또한 Chang(2010)과 Thomas et al.(2014)의 연구에서도 이에 대해 명확한 이론적 접근이나 관련 시뮬레이션 연구가 수행되지 않았다고 밝히고 있다. 따라서 이 연구에서는 기존 연구(Blei, Ng, & Jordan, 2003; Canini, Suh, & Pirolli, 2011; Griffiths & Steyvers, 2004)에서 그간 대표적으로 사용되었던 $50/K$와 $200/V$를 각각 α와 η 모수로 사용하였다.

3) 모형 평가

전통적 통계방법의 하나인 탐색적 요인분석(exploratory factor analysis)의 경우 요인의 수를 미리 설정하지 않고, 탐색적 접근을 통해 데이터를 가장 잘 설명해내는 모형을 찾음으로써 최적의 요

인 수를 찾는다. 이때, 같은 데이터라 할지라도 어떤 식으로 고유값을 추출했는지, 추출해낸 고유값과 그에 해당되는 고유벡터를 어떤 식으로 조합하고 회전시켰는지에 따라 가장 적합한 요인 수가 달라질 수 있다. 예를 들어, 고유값을 추출해내는 과정에서 1보다 큰 값을 사용할 수도 있고 스크리 도표(scree plot)를 사용할 수 있으나, 어떠한 방법도 다른 방법에 비해 명확한 우위가 있다고 말하기 어렵다. 또한, 경우에 따라 앞서 말한 여러 기준에 의해 추출된 요인의 수와 다르더라도, 해석이 용이하다면 추출하여 사용하는 경우도 있다.

토픽 모델링에서도 유사한 방식을 사용하여 주제의 수를 추출해낸다. 선행연구에서는 최적 주제의 수를 산출하기 위해 우도나 혼잡도(perplexity; Grün & Hornik, 2011)를 주로 사용했다. 그러나 Chang(2010)의 연구에서는 혼잡도에 기반한 모형 선택의 경우, 실제 훈련된 연구 참여자가 주제를 유추하여 부여한 결과와 큰 차이를 보이는 것으로 나타났다. 이는 혼잡도가 모형 적합에 대한 적절한 평가를 하는 데 한계를 가지고 있다고 볼 수 있다. 이에 이 연구에서는 혼잡도를 배제하고, 우도와 DIC를 모형 적합도 판단에 사용하였다. 후자는 여러 선행연구(Kwak, Kim, & Cohen, 2017; Lauderdale & Clark, 2014; Sizov, 2012)에서 사용되었다. LDA 모형을 사용하는 대부분의 경우 모수의 수가 관측의 수를 초과하기 때문에 AIC(Akaike, 1974), BIC(Schwarz, 1978), AICc(Hurvich & Tsai, 1989), ABIC(Sclove, 1987)와 같은 빈도주의적 접근(frequentist approach) 기반의 모형 적합도보다 유효한 모수의 수를 고려하는 DIC(Spiegelhalter et al., 2002)가 더 유용한 모형 적합도로 간주된다(Kwak, Kim, & Cohen, 2017).

구체적으로 본 연구에서 DIC는 다음과 같이 계산되었다. 먼저, Spiegelhalter et al.(2002)은 DIC를 아래와 같이 정의하였다.

$$DIC = D(\hat{\theta}) + 2p_D$$

위 식의 $D(\hat{\theta})$와 관련하여, $D(\theta)$는 편차를 의미하며 아래와 같이 정의된다.

$$D(\theta) = -2\log(p(y|\theta)) - 2\log f(y)$$

이때, y는 데이터, θ는 모수, 그리고 $p(y|\theta)$는 사후 분포를 의미한다. $-2\log f(y)$는 데이터에 종속된 상수이므로, 모델 간 비교를 할 때 상쇄된다. 다음으로, 유효한 모수 수인 p_D는 아래와 같이 정의된다.

$$p_D = \overline{D(\theta)} - D(\hat{\theta})$$

이때, $\overline{D(\theta)}$는 평균 편차를 의미하고, 이는 MCMC 과정에서 계산되는 모든 마디(node)에서의 편차 평균으로 계산된다. $D(\hat{\theta})$은 사후분포의 편차를 의미하며, 조건부 확률의 조화평균으로 근사된다(Griffiths & Steyvers, 2004). 구체적인 DIC 계산 과정은 [부록]에 제시하였다.

2. 서술형 강의평가

연구자에 따라 강의평가의 목적은 달리 정의되나, 일반적으로는 총괄적(summative) 목적과 형성적(formative) 목적으로 분류된다(류춘호, 이정호, 2003; 박인우, 2012; 한신일, 김혜정, 이정연, 2005). 구체적으로는, 총괄적 목적은 강의 능력을 정량화하여 교수자의 승진 및 재임용, 정년보장의 등의 의사결정의 근거를 제시하기 위함이고, 형성적 목적은 교수자에게 학생들의 강의에 대한 피드백을 제공하여 강의의 질을 개선하기 위함이다. 이외에도, 학생들에게 교과목 선택을 위한 참고자료를 제공하기 위한 목적도 있다(Marsh, 1984). 그러나 우리나라의 경우, 대부분의 대학은 강의평가 결과를 학교 행정당국 및 일부 교수에게 제한적으로 공개하고 있어서 이 단계에서 활용되고 있다고 보기는 어렵다(홍경선, 2006).

강의평가의 구체적인 내용은 대학마다 다르나, 한신일, 김혜정, 이정연(2005)의 연구에 따르면 대학에 따라 대학 전체 차원에서 하나의 설문을 공통으로 사용하거나, 수업 특성에 따라 다른 설문을 사용하는 경우도 있었다. 단일평가 설문지의 경우, 1) 한 학기 동안의 강의에 대한 총평, 2) 학습자에 대한 질문, 3) 교수자에 대한 질문, 4) 수업에 대한 질문으로 구분됐다. 구체적으로는 수업에 대한 질문이 가장 많은 비중을 차지했으며(74.5%), 그중에서도 특히 수업조직(설계)(13.4%)과 과제·시험·평가(12.48%)에 대한 질문 빈도가 높았다. 반면, Marsh(1984)는 요인분석 결과를 바탕으로 SEEQ(Students' Evaluations of Educational Quality)의 9개 구인을 제시했는데, 학습/가치(Learning/Value), 교수자의 열정(Enthusiasm), 조직(Organization), 상호작용(Group Interaction) 등 대부분의 구인이 수업과 관련된 것으로 나타났다.

강의평가의 실시는 대체로 학기 말에 온라인에서 해당 강의를 수강한 학생들을 대상으로 3~5점의 리커트형 설문에 응답하도록 하는 형식을 취하고 있다(양미경, 2008). 일반적으로, 강의평가 참여율을 높이기 위해 강의평가에 응답하지 않을 경우 성적 조회 제한을 두거나 수강신청을 제한하는 방식으로 의무적인 참여를 요구한다(김성열 외, 2001; 한신일, 김혜정, 이정연, 2005). 그러나 이러한 강제적인 참여가 학생들의 무성의한 답변을 유발한다는 지적이 있었다(김학일 외, 2007; 류춘호, 이정호, 2003; 박인우, 2012; 한경수, 최숙희, 박재철, 2011). 예를 들어, 한경수, 최숙희, 박재철(2011)의 연구에서 5개 학기의 강의평가를 분석한 결과, 평균 20.4%의 학생이 모든 과목에서 동일하게 응답했으며, 평균 30.8%의 학생이 5과목 이상에서 동일하게 응답한 것으로 나타났다.

무성의한 응답은 강의평가의 신뢰도와도 관련이 있다(권오영 외, 2014; 김학일 외, 2007). 강의평

가의 신뢰도는 보통 내적일관성 신뢰도를 나타내는 Cronbach's α로 대변되는데, 학생의 무성의한 응답은 학생이 문항에 일관적으로 응답하고 있다고 해석될 수 있으므로 신뢰도 계수값을 높이는 결과를 초래한다. 따라서 강의평가의 높은 신뢰도 계수가 실제로 강의평가 도구의 안정성을 의미하는지, 혹은 많은 학생들이 무성의하게 일관적으로 응답했기 때문인지는 따져보아야 할 문제이다. 김명화(2005)의 연구에서도 50.3%의 학생이 모든 문항에 같은 평정을 한 것으로 나타난 특정 학기의 강의평가 결과를 사용해 신뢰도를 계산한 결과, Cronbach's α가 .77~.99 수준으로 높게 나타났다.

이처럼 강의평가의 선택형 문항으로부터 기인하는 문제로 인해, 대부분의 대학은 강의평가 마지막에 학생이 자유롭게 의견을 제시할 수 있는 서술형 문항을 제공하고 있다. 서술형 문항에서 수집되는 학생들의 의견은 리커트 척도에서 계산된 총점보다 유익한 피드백을 제공한다(한신일, 2003). 동일한 맥락에서, 홍경선(2006)은 학생의 무성의한 응답을 피하기 위해 질적 강의평가의 도입을 주장했다. 한경수, 최숙희, 박재철(2011) 역시 전북대학교의 사례를 소개하며, '수업에서 인상 깊고 유익했던 사항'과 '수업내용과 방법의 개선을 위한 제안'과 같은 서술형 문항에 많은 학생들이 성실하게 응답했음을 명시했다.

한편 서술형 강의평가 분석과 관련하여, 한신일(2003)은 모 대학의 강의평가 결과가 상·하위 10%에 해당하는 54개 강좌를 대상으로 학생들의 건의사항을 주제별로 약호화하여 분석했다. 구체적으로는, 학생들의 서술형 응답을 SEEQ의 10개 영역(기존의 9개 영역과 '전체만족도' 포함)으로 분류하고, 이에 해당하지 않는 응답은 새로이 추가한 4개 영역으로 분류하였다. 이때 5명의 교육학 분야의 연구자가 직접 응답을 유목화했다. 한신일(2003)의 연구는 서술형 강의평가 응답을 분석하여 학생들의 의견을 유형별로 나누었다는 점에서 의의가 있으나, 연구 방법의 한계로 인해 54개의 강의평가를 분석하는 데 그쳤다. 이에 반해, 이 연구에서 사용하는 LDA는 대량 문서를 한 번에 분석할 수 있다는 점에서 기존 질적 연구 방법의 한계를 보완하고 있다.

최근 들어 텍스트 마이닝에 대한 관심이 급증하면서 텍스트 마이닝 기법을 활용하여 강의평가의 서술형 응답을 분석한 연구 역시 보고되었으나(이해듬, 남민우, 2018; 최정웅, 안동규, 2016), 그 수는 매우 제한적이다. 이해듬과 남민우(2018)는 강의평가 총점을 활용하여 상위 30%의 과목을 분류하고, 이들의 10년 간 매학기별 강의평가의 서술형 문항을 활용하여 전공계열별로 좋은 강의의 특성과 패턴을 분석하였다. 최정웅과 안동규(2016)는 텍스트 마이닝을 통해 '학습자 상호작용'과 관련된 강의의 특성을 추출하고, 이들 키워드의 점수와 선택형 검사에서 얻어진 강의평가 점수를 비교하였다. 두 연구는 모두 텍스트 마이닝 방식을 활용하였다는 점에서 이 연구와 비슷하다고 할 수 있다. 그러나 두 연구 모두 '좋은 강의'나 '학습자 상호작용'과 같은 제한적인 주제를 다룬 반면, 이 연구는 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 텍스트 분석을 통해 강의에 대한 긍정적·부정적 피드백을 종합적으로 볼 수 있다는 차별성을 갖는다.

Ⅲ. 연구 방법

1. 연구 대상

이 연구의 대상은 S대학교 학생들의 2015년 1학기에 개설된 약 1,500개의 강의에 대한 서술형 강의평가, 즉 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대해 서술한 약 47,000개의 응답이다. '글쓰기의 기초'와 같이 많은 학생들이 듣는 수업에는 626명이 의견을 남겼으며, '18·19세기 영국소설'과 같이 특정 전공 학생들이 듣는 수업의 경우 14명이 응답을 남겼다. 학생들은 평균적으로 약 10단어 내외로 응답하였으며, 한 강의에 남긴 모든 학생들의 의견들을 하나의 개별 문서로 간주하였다. 따라서 이 연구에서는 하나의 강의가 최소 분석 단위가 되었고, '강의에서 개선되어야 할 점'은 1,587개의 강의(문서)에 대해 분석되었으며, '강의에서 좋았던 점'은 1,547개의 강의(문서)에 대해 분석되었다.

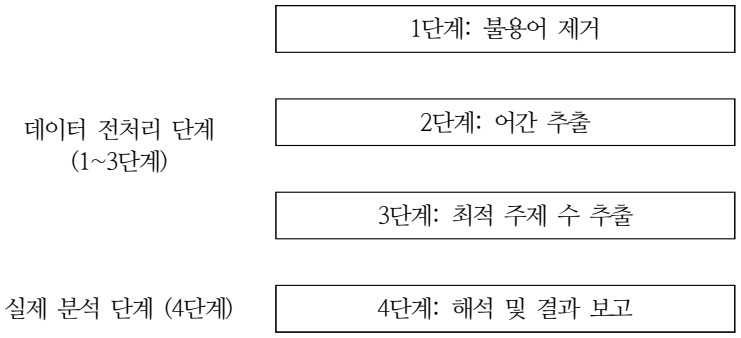
이 연구에서는 먼저 전체 대학 자료를 활용하여 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대해 분석하였고, 다음으로 전체 대학 분석에 활용된 8개 단과대학 중 표본 크기가 확보된 공과대학, 농업생명과학대학, 사범대학, 사회과학대학, 인문대학, 자연과학대학에 대해 단과대학별로 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대해 분석을 실시하였다(〈표 1〉 참조). Kwak et al.(2018)에서 사전정보에 강건하고(robust) 상대적으로 안정적인 추정값을 제공하는 최소 사례 수(문서 수) 크기가 150으로 제시되었기 때문에, 미술대학과 음악대학은 단과대학별 분석 대상에서 제외하였다.

〈표 1〉 단과대학별 사례 수

	개선되어야 할 점	좋았던 점	합계
공과대학	219	228	447
농업생명과학대학	219	149	368
미술대학	110	116	226
사범대학	256	263	519
사회과학대학	151	148	299
음악대학	116	118	234
인문대학	362	366	728
자연과학대학	154	159	313
합계	1,587	1,547	3,134

2. 분석 방법

LDA를 이용하여 토픽 모델링을 실시하기 위해서는 우선 두 가지의 전후 처리 단계, 1) 불용어(stopword) 제거와 2) 어간추출(stemming) 단계가 필요하다. 이 단계들의 목적은 말뭉치를 일종의 분석 가능한 형태로 가공하는 데 있다. 이 가공된 데이터를 바탕으로 최적 주제 수를 결정하고, 이 주제 수를 바탕으로 한 모형을 데이터에 적합시키면, 최종적인 분석 결과인 주제별 단어 분포, 문서별 주제 분포 모수가 추정된다. 그러나 문서별 주제 분포 분석이 의미를 갖기 위해서는 응답자의 인구통계학적 특성이나 학업적 동기, 흥미, 자아존중감과 같은 교육심리적 특성과 연결해야 하므로, 연구의 범위가 매우 광범위해진다. 이 연구의 목적은 S대학교 학생들의 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 응답에 대한 전반적인 분석이므로, 문서별 주제 분포는 생략하고, 주제별 단어 분포를 제시하였다.



[그림 5] 텍스트 마이닝의 전후 처리 및 실제 분석 단계

[그림 5]에서 볼 수 있듯이, 데이터 전처리의 첫 번째 단계는 불용어의 제거이다. 불용어는 *나는, 우리는, 입니다, 습니다* 등과 같이 특별한 의미는 없으나, 말뭉치 내 대부분의 문서에서 빈번하게 등장하는 단어들을 의미한다. 불용어는 문서 전반에 걸쳐 반복적으로 나타나기 때문에 분석에서 제외되거나 매우 낮은 가중치를 부여하는 방식으로 해결되어야 한다.

이 연구에서 불용어는 TF-IDF 방법(Grün & Hornik, 2011; Manning, Raghavan, & Schutze, 2008)을 기반으로 설정되었다. TF-IDF 점수는 각 단어마다 부여되는 일종의 가중치로서, 이 값이 작을수록 불용어에 가까운 단어로 간주할 수 있다. 본래 선행연구(Grün & Hornik, 2011)에서는 전체 단어 분포의 중위값을 기준으로 하여, 그보다 낮은 TF-IDF 점수를 보이는 단어들을 불용어로 간주하는 것을 추천하였다. 그러나 이는 이론적·경험적 연구가 부족한 결과에서 최소 가이드라인으로 제시된 기준이며, 최근의 연구에서는 단어의 수가 제한적인 경우에 좀 더 낮은 퍼센타일의 값을 쓰는 것이 경험적으로 적절하다는 보고가 있다(Kwak, Kim, & Cohen, 2017). 또한, 대개 30개 정도

의 불용어가 적절하다는 연구(Manning, Raghavan, & Schütze, 2008)에 따라 이 연구에서는 약 30개의 불용어를 추출할 수 있는 TF-IDF 점수인 10.9%를 사용하여 이보다 작은 값을 갖는 단어를 불용어로 분류하고 제외한 데이터를 분석에 사용하였다.

두 번째 단계는 어간 추출 단계이다. 이 과정은 단어나 어구를 기본형으로 통일시키는 것을 의미한다. 예를 들어, *좋았습니다, 좋습니다, 좋다*는 모두 같은 의미를 지니고 있기 때문에 좋다로 통일되어야 한다. 이 단계에서는 R 패키지인 KoNLP(Jeon & Kim, 2016)을 사용하여 명사 추출 작업을 수행하였다.

위의 두 과정은 모형의 모수공간을 합리적으로 축소시키고, 표본 크기가 상대적으로 작을 수밖에 없는 교육학적 맥락에서 주제별 단어 분포를 보다 명확하게 추정하는데 도움이 된다는 점에서 매우 중요한 단계라고 볼 수 있다.

세 번째 단계이자, 데이터 전처리의 마지막 과정이라 할 수 있는 단계는 적합한 주제 수를 결정하는 단계이다. 이를 위해 DIC와 우도에 자연로그를 취한 값, 즉 log(우도)를 이용하였고, 총 2개부터 5개의 주제가 탐색되었다. 일반적으로 컴퓨터 과학 분야에서 사용되는 모형의 경우 몇 백 개 이상의 주제가 고려되기도 하지만, 이는 웹에서 무작위로 추출된 문서를 대상으로 할 때이며, 지금처럼 어휘의 수가 적게 사용된 경우에는 주제의 수가 제한된다. '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점' 모두 3개 주제가 적절한 것으로 나타났다. 각 문항별 세부사항은 연구 결과에서 자세히 다룰 예정이다.

추가적으로, 앞선 불용어 처리의 기법으로 언급한 TF-IDF의 경우 키워드 분석에 빈번하게 사용되는 기법 중 하나이다. 불용어 처리를 위해 문서에 사용된 모든 단어의 TF-IDF 값을 계산하였고, 연구 결과에서 이를 활용한 키워드 분석 결과를 간략히 제시하였다.

Ⅳ. 연구 결과

1. 기술통계

응답 자료에 대한 기술통계치를 제시하면 〈표 2〉와 같다. 전처리 전과 전처리 후 값이 모두 제시되어 있다. 전처리 후 자료의 특성에 대해 살펴보면, '강의에서 개선되어야 할 점'의 경우, 응답 수는 1,539개였고, 응답 길이의 평균은 39.12이었으며, 응답 길이의 표준편차는 70.73이었다. 어휘 수는 4,375개였으며, 총 단어 수는 60,203개였다. '강의에서 좋았던 점'의 경우, 응답 수는 1,534개였고, 응답 길이의 평균은 52.47이었으며, 응답 길이의 표준편차는 82.25이었다. 어휘 수는 3,858개였으며, 총 단어 수는 80,489이었다.

〈표 2〉 기술통계치

	개선되어야 할 점		좋았던 점	
	전처리 전	전처리 후	전처리 전	전처리 후
응답 수	1,587	1,539	1,547	1,534
응답 길이 평균	68.47	39.12	90.41	52.47
응답 길이 표준편차	120.66	70.73	136.10	82.25
어휘 수	10,627	4,375	11,554	3,858
총 단어 수	108,659	60,203	139,867	80,489

앞서 연구 방법 단계에서 언급한 TF-IDF 분석을 통해, 개선되어야 할 점과 좋았던 점에 대해 각각 상위 30개 단어를 추출하여 〈표 3〉에 제시하였다. 개선되어야 할 점의 경우, *개괄, 자유, 방목, 오티, 필수, 지양, 본질적, 전달과정, 오탈자, 최신자료* 등의 단어가 높은 TF-IDF 값을 나타냈다. 좋았단 점의 경우, 강의내용 및 학과와 밀접한 관련이 있는 *재즈, 자전거, 체조, 한국가곡, 곤충, 애니메이션, 유화, 장단, 생체, 맹자, 토양* 등의 단어가 높은 TF-IDF 값을 나타냈다.

〈표 3〉 TF-IDF 키워드 빈도 분석

	개선되어야 할 점		좋았던 점	
	단어	TF-IDF 값	단어	TF-IDF 값
1	개괄	2.71	집단	1.13
2	자유	1.53	재즈	0.98
3	방목	1.46	자전거	0.97
4	오티	1.44	체조	0.94
5	필수	1.44	한국가곡	0.87
6	지양	1.27	곤충	0.72
7	본질적	0.96	애니메이션	0.72
8	상상	0.96	유화	0.68
9	규격	0.82	재미	0.64
10	전달과정	0.80	장단	0.64
11	불공정	0.76	주거	0.64
12	그냥	0.76	단소	0.64
13	오탈자	0.75	생체	0.62
14	최신자료	0.72	중요	0.58
15	악기	0.64	맹자	0.57
16	안좋아요	0.61	토양	0.53
17	선호	0.60	재생	0.53
18	긴장	0.59	농구	0.53
19	못듣겠음	0.58	예보	0.49
20	인신공격	0.53	시장	0.49
21	무지한	0.53	회계	0.49
22	이티엘	0.50	연기	0.49
23	대나무	0.49	판화	0.48
24	테이크홈	0.47	다른사람	0.47
25	리더십	0.44	육상	0.46
26	힘듦	0.42	식품	0.45
27	계열	0.42	광고	0.45
28	어플	0.41	합창	0.44
29	조별활동	0.41	배웠다	0.43
30	자주	0.40	방사선	0.43

2. 최적 주제 수 결정

‘강의에서 개선되어야 할 점’과 ‘강의에서 좋았던 점’에 대한 각각의 말뭉치에 적합한 최적 주제 수를 결정하기 위하여 DIC와 log(우도)를 활용하였다. DIC의 경우 최솟값을 보여준 경우, log(우도)

의 경우 최댓값을 보여준 주제 수의 모형을 선택하였다(〈표 4〉 참조). '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점' 두 말뭉치에 대해 DIC와 log(우도)를 통해 선택된 모형은 3주제 모형으로 일치하였다.

〈표 4〉 서술형 강의평가별 말뭉치의 모형 적합도

주제 수	강의에서 개선되어야 할 점		강의에서 좋았던 점	
	DIC	log(우도)	DIC	log(우도)
2	826,857	−413,427	1,033,930	−516,964
3	826,593	−413,295	1,025,736	−512,866
4	828,858	−414,427	1,026,573	−513,284
5	835,747	−417,870	1,032,828	−516,411

보다 구체적으로, '강의에서 개선되어야 할 점'에 대한 말뭉치의 DIC는 3주제 모형이 가장 낮은 수치를 보였다. log(우도)의 경우 역시 3주제 모형이 가장 높은 수치를 보였다. '강의에서 좋았던 점'에 대한 말뭉치의 DIC 역시 3주제 모형이 가장 낮은 수치를 보였다. 또한, log(우도) 역시 3주제 모형이 가장 높은 수치를 보였다. 따라서 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점' 모두 3주제 모형이 가장 적합한 것으로 판단하였다.

3. 주제 분석

이하에서는 '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 응답의 주제 분석 결과를 각각 제시하였다. 또한 김지은과 백순근(2016)의 연구에서 제시한 바와 같이, 주제와 함께 해당 주제가 가장 대표적으로 나타난 문서의 수 및 비율을 제시하였다.

1) '강의에서 개선되어야 할 점'에 대한 주제 분석

'강의에서 개선되어야 할 점'에 대한 주제 분석은 〈표 5〉와 같다. 〈표 5〉에서 제시된 주제별 단어 분포를 살펴보면, 각 주제에 적절한 의미를 부여할 수 있다. 먼저, '주제 1'의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *시간, 생각, 발표, 방식, 진행, 토론, 참여, 평가, 과제, 필요, 방법, 부족, 준비, 다양, 지루, 기회, 글쓰기* 등이다. 즉, 주제 1은 *발표, 토론, 과제*에 대한 코멘트로 볼 수 있으며, 이와 함께 등장한 단어들이 *방식, 진행, 참여, 평가, 필요, 부족, 지루, 기회*인 것으로 보아, 발표 및 토론 등의 수업 방식이 지루하고, 참여 기회가 부족하며, 평가에 대한 불만이 있다는 것을 유추할 수 있다. 따라서 '주제 1'은 '강의·발표·과제에 대한 개선사항'으로 명명하였다.

〈표 5〉 주제별 단어 분포: 강의에서 개선되어야 할 점

빈도순	주제 1		주제 2		주제 3	
	단어	빈도	단어	빈도	단어	빈도
1	시간	1,080	과제	709	내용	1,573
2	학생	875	시간	370	시험	902
3	교수님	691	학생	331	교수님	864
4	생각	488	실험	284	이해	600
5	발표	395	진행	260	문제	581
6	방식	310	조교	252	설명	514
7	진행	258	도움	193	진도	434
8	토론	241	필요	162	생각	405
9	참여	230	학점	155	공부	367
10	평가	217	실습	152	자료	328
11	주제	199	사람	151	학생	304
12	과제	179	부담	150	영어	220
13	필요	175	점수	148	기말고사	189
14	방법	169	학기	146	고사	178
15	부족	165	성적	141	교재	178
16	준비	163	이론	131	감사	171
17	피드백	159	부족	126	집중	166
18	다양	122	과목	126	목소리	149
19	지루	117	프로젝트	118	말씀	143
20	이야기	114	수준	109	출석	135
21	세미나	105	로드	107	난이도	130
22	사람	104	난이도	100	정리	121
23	기회	104	기준	98	질문	120
24	글쓰기	98	레포트	97	범위	116
25	학기	97	사용	95	중요	115
26	전체	95	채점	90	개념	115
27	효과	94	시작	82	후반	111
28	전공	90	퀴즈	81	마지막	110
29	과목	89	만족	80	필기	109
30	논문	87	공지	79	어려움	95
관련 문서 수 (%)	530 (34.44)		440 (28.59)		401 (26.06)	

다음으로, ‘주제 2’의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *과제, 시간, 학생, 실험, 조교, 도움, 실습, 부담, 점수, 부족, 프로젝트, 수준, 로드, 기준, 레포트, 채점, 퀴즈, 만족, 공지* 등이다. 즉, ‘주제 2’는 *과제, 실습, 프로젝트, 레포트, 퀴즈* 등이 높은 빈도를 보이고 있으며,

이와 함께 등장한 단어들이 *부담, 부족, 개선, 로드* 등인 것으로 보아 학생들이 *과제, 실습 및 레포트*에 대한 부담을 느끼고 있는 것으로 유추할 수 있다. 따라서 '주제 2'는 '과제 및 실습에 대한 부담'으로 명명하였다.

마지막으로, '주제 3'의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *내용, 시험, 이해, 문제, 설명, 진도, 영어, 기말고사, 고사, 집중, 목소리, 난이도, 범위, 중요, 어려움* 등이다. 즉, '주제 3'은 *시험, 설명, 기말고사, 진도*에 대한 의견으로 볼 수 있으며, 이와 함께 등장한 단어들이 *이해, 난이도, 어려움*인 것으로 보아, 수업에서 설명된 내용을 이해하는 것이나 기말고사 등의 시험이 어려웠으며, 수업 진도나 시험 범위에 대한 불만이 있다는 것을 유추할 수 있다. 따라서 '주제 3'은 '시험·진도·수업내용에 대한 개선사항'으로 명명하였다.

각 주제별로 해당 주제가 가장 대표적으로 나타난 문서 수를 보면, 주제 1의 경우 530개(34.44%), 주제 2의 경우 440개(28.59%), 주제 3의 경우 401개(26.06%)개의 문서에서 해당 주제가 가장 대표적으로 나타났다.

2) '강의에서 좋았던 점'에 대한 분석

'강의에서 좋았던 점'에 대한 주제 분석은 〈표 6〉과 같다. '주제 1'의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *교수님, 설명, 학생, 내용, 감사, 친절, 감사, 도움, 질문, 열정, 과제, 시험, 노력, 개념, 만족, 재밌게, 진행, 흥미, 예시* 등이다. 즉, '주제 1'은 *교수님, 설명, 과제, 시험* 등에 대한 코멘트로 볼 수 있으며, 이와 함께 등장한 단어들이 *친절, 도움, 열정, 만족, 재밌게, 흥미*인 것으로 보아, 교수님이 친절하고 열정적이었으며, 강의가 흥미롭고 재미있었다는 것을 유추할 수 있다. 따라서 '주제 1'은 '교수자·교수방법에 대한 긍정적 피드백'으로 명명하였다.

다음으로, '주제 2'의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *학생, 실험, 조교, 시간, 경험, 진행, 도움, 이론, 실제, 자유, 기회, 실습, 분위기, 참여, 친절, 과정, 피드백, 영어, 방법, 연구, 열정, 실력, 필요* 등이다. 즉, '주제 2'는 *실험, 경험, 발표, 실제, 실습, 참여* 등이 높은 빈도를 보이고, 이와 함께 등장한 단어들이 *도움, 피드백, 친절* 등인 것으로 보아 직접적·실제적으로 경험하고 참여할 수 있었던 실험, 발표, 경험 등에 긍정적인 반응을 보이는 것으로 유추할 수 있다. 따라서 '주제 2'는 '직접적 경험·발표·실험에 대한 긍정적 피드백'으로 명명하였다.

마지막으로, '주제 3'의 경우, 높은 빈도를 보이고 있으며 동시에 의미를 가진 단어들이 *생각, 다양, 내용, 흥미, 교수님, 지식, 토론, 유익, 주제, 자료, 시간, 분야, 이야기, 역사, 영화, 작품, 의견, 철학, 시각, 관점, 평소, 관심, 접근, 영상* 등이다. 즉, '주제 3'은 *생각, 지식, 토론, 의견, 시각, 관점*에 대한 코멘트로 볼 수 있으며, 이와 함께 등장한 단어들이 *다양, 흥미, 유익, 자료, 이야기, 역사, 영화, 작품, 영상*인 것으로 보아, 다양한 생각을 유도하는 각종 자료에 대한 긍정적인 반응을 보이는 것으

로 유추할 수 있다. 따라서 '주제 3'은 '다양한 관점을 부여하는 자료에 대한 긍정적 피드백'으로 명명하였다.

〈표 6〉 주제별 단어 분포: 강의에서 좋았던 점

빈도순	주제 1		주제 2		주제 3	
	단어	빈도	단어	빈도	단어	빈도
1	교수님	4892	학생	691	생각	1166
2	설명	1415	실험	454	다양	1146
3	이해	1256	조교	422	내용	978
4	학생	1138	시간	413	흥미	563
5	내용	1118	경험	339	교수님	470
6	감사	810	진행	304	지식	438
7	친절	646	도움	281	시간	436
8	공부	511	이론	279	토론	390
9	준비	457	자유	274	유익	383
10	도움	453	발표	260	주제	331
11	질문	418	실제	249	자료	321
12	열정	411	기회	237	분야	309
13	과제	361	실습	227	기회	305
14	자세	274	분위기	218	이야기	290
15	시험	269	참여	211	역사	289
16	학기	235	전공	195	방식	284
17	모습	230	학기	189	영화	226
18	노력	227	교수님	178	작품	219
19	개념	199	친절	177	의견	189
20	문제	196	과정	175	철학	183
21	만족	178	피드백	171	시각	177
22	재밌게	174	영어	169	관점	176
23	진행	158	방법	159	음악	170
24	경제	141	연구	149	자체	155
25	과목	141	열정	143	평소	152
26	기초	141	논문	130	관심	140
27	흥미	140	활동	116	학기	136
28	교재	137	방식	112	접근	133
29	수학	133	실력	106	교양	131
30	예시	125	필요	105	영상	130
관련 문서 수 (%)	626 (40.47)		293 (18.94)		646 (41.76)	

각 주제별로 해당 주제가 가장 대표적으로 나타난 문서 수를 보면, 주제 1의 경우 626개(40.47%), 주제 2의 경우 293개(18.94%), 주제 3의 경우 646개(41.76%)개의 문서에서 해당 주제가 가장 대표적으로 나타났다.

4. 단과대학별 주제 분석

단과대학별 분석 결과의 경우, 〈표 5〉나 〈표 6〉과 같은 단어별 빈도는 지면의 한계로 생략하였고, 각 주제별 주요 단어 및 주제에 대한 적합한 의미 부여 결과만을 제시하였다. 앞서 언급했듯이 충분한 표본 크기가 확보된 6개 단과대학인 공과대학, 농업생명과학대학, 사범대학, 사회과학대학, 인문대학, 자연과학대학에 한하여 단과대학별 분석을 실시하였다.

단과대학별 분석 결과에서는 다음의 세 가지 특징을 찾을 수 있었다. 첫째, 단과대학마다 조금씩 다르지만 대개 3~4개 주제로 분류되었다. 둘째, 전반적인 주제의 의미는 전체 대학을 대상으로 분석한 결과와 매우 유사하였다. 셋째, 6개 단과대학들은 단과대학의 특성에 따라 유사한 주제 분포를 나타내었기 때문에, 크게 두 가지 유형인 이공계열과 인문사회계열로 구분할 수 있었다. 공과대학, 농업생명과학대학, 자연과학대학은 이공계열에 해당되고, 사범대학, 사회과학대학, 인문대학은 인문사회계열에 해당된다.

1) '강의에서 개선되어야 할 점'에 대한 주제 분석

이하에서는 각 계열(이공계열, 인문사회계열)의 '강의에서 개선되어야 할 점'에 대한 전반적인 주제 분포를 제시한 후, 단과대학별로 각 주제에 해당되는 대표 단어들을 제시하였다.

이공계열 대학의 '강의에서 개선되어야 할 점'에 대한 대표적인 세 가지 주제는 다음과 같다: ① 시험·진도에 대한 개선사항, ② 실험·실습·프로젝트·과제에 대한 개선사항, ③ 강의에 대한 개선사항.

보다 구체적으로, 첫째, 단과대학별 '시험·진도·과제에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *내용, 교수님, 시험, 문제, 진도*, 농업생명과학대학의 경우 *내용, 교수님, 시험, 문제, 이해, 생각, 진도, 난이도*, 자연과학대학의 경우 *시험, 문제, 과제, 내용, 학생, 진도, 공부, 난이도* 등의 단어가 높은 빈도를 보였다.

둘째, 단과대학별 '실험·실습·프로젝트·과제에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *설명, 실험, 과제, 조교, 시간, 프로젝트, 실습*, 농업생명과학대학의 경우 *시간, 과제, 실험, 조교, 프로젝트, 실습*, 자연과학대학의 경우 *실험, 조교, 레포트, 내용, 학생, 시간, 학점* 등의 단어가 높은 빈도를 보였다.

셋째, 단과대학별 '강의에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *학*

생, 생각, 시간, 세미나, 진행, 이해, 자료, 방식, 농업생명과학 대학의 경우 *학생, 진행, 세미나, 시간, 생각, 필요, 방식*, 자연과학대학의 경우 *교수님, 내용, 학생, 시간, 이해, 설명, 생각* 등의 단어가 높은 빈도를 보였다.

인문사회계열 대학의 '강의에서 개선되어야 할 점'에 대한 대표적인 세 가지 주제는 다음과 같다: ① 시험·진도에 대한 개선사항, ② 발표·토론·과제에 대한 개선사항, ③ 강의 방식 및 내용에 대한 개선사항.

보다 구체적으로, 첫째, 단과대학별 '시험·진도에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *내용, 생각, 시험, 교수님, 과제, 이해*, 사회과학대학의 경우 *내용, 교수님, 문제, 공부, 설명, 이해, 진도*, 인문대학의 경우 *시험, 학생, 진도, 교수님, 문제, 어려움, 범위, 난이도* 등의 단어가 높은 빈도를 보였다.

둘째, 단과대학별 '발표·토론·과제에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *학생, 교수님, 발표, 진행, 평가, 토론, 과제, 준비, 피드백*, 사회과학대학의 경우 *학생, 생각, 교수님, 내용, 발표, 토론, 시간, 자료, 필요, 참여*, 인문대학의 경우 *시간, 학생, 과제, 생각, 교수님, 발표, 내용, 글쓰기, 토론* 등의 단어가 높은 빈도를 보였다.

셋째, 단과대학별 '강의 방식 및 내용에 대한 개선사항' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *시간, 부족, 필요, 개선, 학생, 사람, 방식, 부담*, 사회과학과학의 경우 *과제, 시간, 사람, 부족, 평가, 설명, 방법, 부담*, 인문대학의 경우 *내용, 교수님, 이해, 자료, 설명, 진행, 영화, 지루* 등의 단어가 높은 빈도를 보였다.

2) '강의에서 좋았던 점'에 대한 주제 분석

이하에서는 각 계열(이공계열, 인문사회계열)의 '강의에서 좋았던 점'에 대한 전반적인 주제 분포를 제시한 후, 단과대학별로 각 주제에 해당되는 대표 단어들을 제시하였다.

이공계열 대학의 '강의에서 좋았던 점'에 대한 대표적인 세 가지 주제는 다음과 같다: ① 교수자·교수방법에 대한 긍정적 피드백, ② 다양한 경험·실험·실습에 대한 긍정적 피드백, ③ 강의내용에 대한 긍정적 피드백. 특히, 두 번째 주제의 경우 공통적으로 실험과 조교의 도움 등의 대해 긍정적인 피드백을 언급하였다. 다만, 농업생명과학대학의 경우 강의내용에 대한 긍정적 피드백과 교수자 및 방법에 대한 긍정적 피드백이 혼재되어 나타났다.

보다 구체적으로, 첫째, 단과대학별 '교수자·교수방법에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *교수님, 설명, 학생, 열정, 친절, 강의력*, 농업생명과학대학의 경우 *교수님, 설명, 내용, 이해, 학생, 공부, 친절*, 자연과학대학의 경우 *교수님, 설명, 이해, 내용, 학생, 감사, 도움 열정, 친절* 등의 단어가 높은 빈도를 보였다.

둘째, 단과대학별 '직접적 경험 및 실험·실습에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *다양, 실험, 지식, 분야, 전공, 경험, 조교, 기회, 실습, 유익, 도움*, 농업생명과학대학의 경우 *다양, 전공, 실습, 실험, 흥미, 기회*, 자연과학대학의 경우 *실험, 조교, 친절, 다양, 시간, 진행, 내용, 실습, 감사* 등의 단어가 높은 빈도를 보였다.

셋째, 단과대학별 '강의내용에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 공과대학의 경우 *이해, 내용, 감사, 생각, 교수님, 학생, 질문, 설명*, 농업생명과학대학의 경우 *교수님, 내용, 지식, 도움, 감사, 생각, 분야, 유익*, 자연과학대학의 경우 *내용, 흥미, 과학, 분야, 방식, 유익, 다양* 등의 단어가 높은 빈도를 보였다.

인문사회계열 대학의 '강의에서 좋았던 점'에 대한 대표적인 세 가지 주제는 다음과 같다: ① 교수자·교수방법에 대한 긍정적 피드백, ② 토론·발표에 대한 긍정적 피드백, ③ 강의내용에 대한 긍정적 피드백을 언급하였다. 다만, 사회과학대학의 경우 강의내용과 교수자의 긍정적 피드백이 혼재되어 나타나고, 과제에 대한 긍정적 피드백이 나타났다.

보다 구체적으로, 첫째, 단과대학별 '교수자·교수방법에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *교수님, 학생, 친절, 열정, 시간, 학기, 운동, 감사, 분위기, 진행, 도움, 흥미*, 사회과학대학의 경우 *내용, 설명, 이해, 흥미, 강의력, 자료, 학생, 친절, 감사*, 인문대학의 경우 *교수님, 학생, 친절, 설명, 도움, 감사, 공부, 이해, 준비, 열정, 진행* 등의 단어가 높은 빈도를 보였다.

둘째, 단과대학별 '토론·발표에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *생각, 다양, 교수님, 학생, 기회, 토론, 경험, 발표, 실제*, 사회과학대학의 경우 *교수님, 생각, 다양, 토론, 철학, 기회, 작품, 발표*, 인문대학의 경우 *토론, 의견, 방식, 발표* 등의 단어가 높은 빈도를 보였다.

셋째, 단과대학별 '강의내용에 대한 긍정적 피드백' 주제별 단어 분포는 다음과 같다. 사범대학의 경우 *교수님, 내용, 설명, 이해, 감사, 학생, 공부, 자료*, 사회과학대학의 경우 *과제, 학생, 시간, 논문, 이론, 도움, 생각, 감사*, 인문대학의 경우 *내용, 교수님, 흥미, 역사, 영화, 이해, 생각, 지식, 설명* 등의 단어가 높은 빈도를 보였다.

V. 요약 및 논의

이 연구는 텍스트 마이닝 방법 중의 하나인 LDA를 활용하여 S대학교의 2015년 1학기 강의에 대한 강의평가의 서술형 문항('강의에서 개선되어야 할 점'과 '강의에서 좋았던 점')에 대한 응답을 분석함으로써 대량의 서술형 응답을 효율적으로 요약했으며, S대학교 강의 전반에 대한 학생들의 인식을 보다 직접적이고 종합적으로 살펴볼 수 있었다는 의의를 갖는다. 주요 연구 결과를 요약하면

다음과 같다.

첫째, 전체 대학의 강의평가 자료를 분석한 결과, ‘강의에서 개선되어야 할 점’에 대한 응답과 ‘강의에서 좋았던 점’에 대한 응답은 모두 세 가지 주제로 분류되었다. 먼저 ‘강의에서 개선되어야 할 점’은 ‘강의·발표·과제에 대한 개선사항’, ‘과제 및 실습에 대한 부담’, ‘시험·진도·수업내용에 대한 개선사항’의 세 가지 주제를 갖고 있는 것으로 나타났다. 다음으로, ‘강의에서 좋았던 점’은 ‘교수자·교수방법에 대한 긍정적 피드백’, ‘직접적 경험·발표·실험에 대한 긍정적 피드백’, ‘다양한 관점을 부여하는 자료에 대한 긍정적 피드백’의 세 가지 주제를 갖고 있는 것으로 나타났다.

‘강의에서 개선되어야 할 점’에 대한 분석 결과에서 특기할 만한 부분은, 시험 및 진도와 과제에 대한 학생들의 의견이 큰 부분을 차지하고 있다는 점이다. 즉, 실제 수업내용 자체와 전달되는 방식 및 교수방법도 학생들에게 중요한 의미를 갖지만, 그 내용의 평가 및 과제로 부여되는 방식 역시 학생들에게 큰 의미를 갖는 것으로 볼 수 있다. 이는 수업내용이 수업목표 및 과정에 적절하고 의미를 갖는다 하더라도, 학생이 과제나 시험이 수업내용과 크게 연관이 없다고 인식할 경우 학생에게 긍정적인 평가를 받기가 어려움을 시사한다. 따라서 강의자가 수업을 기획하는 데 있어, 실제 수업내용과 과제 및 평가가 긴밀하게 연결되어 있는지의 여부에 관심을 갖고 신중하게 접근해야 할 필요가 있다. 이는 최근의 대학교육 효과에 대한 연구에서 학생들의 고차원적 사고를 강화하기 위해서는 ‘평가와 피드백’이 중요하다는 점을 강조하고 있는 점과도 일맥상통한다(김선희, 2017; 오은주, 2009; 이희원, 민혜리, 2013). 지금까지 우리나라 대학에서 주로 사용되는 교육방식인 강의식 교육방식은 학습자를 수동적으로 만들고 고차적 인지능력 강화에는 도움이 되지 않으므로(Bigger & Tang 2011; Plasher et al. 2007; Wittrock, 1977), 학생에게 지적 자극을 주는 시험과 과제에 대한 피드백(Carless, 2006; Tinto, 2005)을 통해 학생을 높은 수준의 학습으로 이끄는 방식이 요구된다는 것이다.

‘강의에서 좋았던 점’에 대한 분석 결과에서 특히 주목할 부분은, 다양한 관점을 갖도록 도와주는 자료에 대해 학생들이 의미 있게 느끼고 있다는 점이다. 학생들이 기존에 가지고 있었던 생각이나 지식, 이론에 대해 새로이 접근할 수 있는 다양한 시각을 제공하는 부분을 인상 깊게 받아들이고 있다고 볼 수 있다. 이러한 결과는 기존의 강의평가에서 주요하게 다루어졌던 ‘평가 및 피드백’(김경언 외, 2018), ‘교수자가 강의에 임하는 태도’ 및 ‘교수자가 학생을 대하는 태도’(전영미 외, 2014), ‘수업의 난이도’(이해듬, 남민우, 2018) 등과는 다소 다른 영역이다. 다시 말해, ‘다양한 관점을 부여하는 자료’는 이 연구를 통해 새롭게 추출된 ‘좋은 강의’의 특성이라고 할 수 있다.

둘째, 단과대학별 강의평가 자료를 분석한 결과, 단과대학별로 나타난 주제의 의미는 대체적으로 전체 자료를 분석했을 때와 비슷하게 나타났으나, 하나의 주제 정도가 단과대학의 특성을 반영하고 있는 것으로 나타났다. 보다 구체적으로, 이공계열 단과대학들의 경우 실험 및 실습 등과 관련된 피드백 주제가 등장하였고, 인문사회계열 단과대학들의 경우 토론 및 발표와 관련된 주제가 나타났다.

한편, 전체 학생의 데이터를 사용한 결과와 단과대학별 주제 추출 결과가 상당히 일치하는 결과를 보인다는 점에서, LDA가 안정적으로 주제를 추출한 것으로 간주할 수 있다. 비록 사전정보와 TF-IDF 점수에 근거한 불용어 제거 등의 전처리 과정이 다소 달랐지만, 전체적으로 일관된 주제 추출 결과를 보인다는 점에서 LDA가 적절하게 기능한 것으로 볼 수 있다. 또한, 각 계열별로 주제들이 조금씩 달라진 경우에도 계열별 특징을 잘 나타냈다는 점 역시 LDA가 S대학교의 서술형 강의평가 자료를 분석함에 있어 적절한 방법이었음을 뒷받침한다.

이 연구의 제한점과 후속연구를 위한 제언은 다음과 같다. 첫째, 이 연구에서는 자료의 한계로 인해 강의평가의 서술형 문항에 대한 응답과 학생 특성 간의 관련성을 밝히지 못하였다. 강의평가의 선택형 문항의 경우, 백순근과 신효정(2008)은 학생 수준의 변인(학년, 전공, 수강동기 등)이 강의평가에 유의한 영향을 미친다고 보고하였다. 이러한 맥락에서 강의평가의 서술형 문항에 대한 응답 역시 학생 수준의 변인에 영향을 받을 것으로 예상된다. 따라서 이후 연구에서는 성별, 연령 등 인구통계학적 변수와 학업적 동기, 흥미, 자아존중감과 같은 교육심리적 특성 등을 활용하여 학생의 하위 집단을 구분하고, 하위 집단 간 긍정적인 피드백 혹은 부정적인 피드백의 주제 추출에 차이가 나타나는지의 여부를 파악할 필요가 있다. 여러 변수 중에서도 특히 학업성취도는 실제 강의의 목적 및 결과와 긴밀한 관련이 있다는 점에서, 학업성취도에 따른 학생들의 강의평가 결과 차이는 강의를 평가할 때 기능하는 학습자의 학습 전략, 심리 등을 탐색해 볼 수 있는 기회가 될 것으로 기대된다.

둘째, 이 연구에서는 한 개 대학의 강의평가를 활용했기 때문에 연구 결과를 일반화하기 어렵다는 한계를 갖고 있다. 대학의 고유한 특성에 의해 강의평가 결과가 좌우될 수 있음을 고려할 때, 여러 대학의 자료를 분석하고 종합할 수 있는 후속 연구가 요구된다.

참고문헌

권오영, 박영태, 황일규, 안태원, 김경숙(2014). 강의평가 시스템의 신뢰도 향상 방안 연구. **공학교육연구, 17**(2), 35-41.

김경언, 우혜정, 김지영, 김우철(2018). 대학 강의평가 도구 개선 및 타당화 연구 -K대학 사례를 중심으로-. **한국교육문제연구, 36**(4), 1-26.

김명화(2005). 강의평가의 타당도와 신뢰도. **아시아교육연구, 6**(3), 1-24.

김선희(2017). 강의평가 결과 분석을 통한 교육의 질 제고 방안 탐색: A 대학교 강의평가를 중심으로. **사회과학연구, 30**(1), 147-174.

김성열, 박재완, 김종철, 강현석, 박형민, 정은희(2001). **대학 학사과정 강의평가제 실태분석을 통한 교육업적 평가모형 개발연구**. 서울: 교육인적자원부.

김지은, 백순근(2016). 텍스트 빅데이터 분석 기법을 활용한 대학구조개혁 평가의 쟁점 분석. **아시아교육연구, 17**(3), 409-436.

김학일, 김성숙, 권오양, 이천, 노경호(2007). 이공계 강의평가 결과의 실증적 분석을 통한 강의평가 제도 개선방안. **공학교육연구, 10**(4), 58-77.

류춘호, 이정호(2003). 대학의 강의평가에 영향을 미치는 학생관련 요인에 관한 연구. **경영학연구, 32**(3), 789-807.

박인우(2012). 대학 강의평가에서 무성의 응답에 대한 학생의 자기평가의 영향에 관한 연구. **교육방법연구, 24**(1), 257-281.

백순근, 신효정(2008). 위계선형모형을 활용한 대학생의 강의평가 분석 -S대학교 교양강의를 중심으로-. **교육평가연구, 21**(2), 1-24.

양길석(2014). 대학 강의평가 일관적 응답의 경향성과 영향력 분석. **교육평가연구, 27**(2), 255-278.

양미경(2008). 학생의 평정에 의거한 대학 강의평가의 의의와 한계. **교육원리연구, 13**(1), 93-122.

오은주(2009). 강의평가 실태조사를 통한 강의평가 개선 방향 연구. **교육방법연구, 21**(2), 1-20.

이해듬, 남민우(2018). 대학 강의평가 주관식 결과의 텍스트마이닝을 통한 전공 계열별 좋은 수업 특성 분석. **한국유아교육연구, 20**(2), 21-41.

이희원, 민혜리(2013). 수업 개선을 위한 강의평가 결과 활용 방안 탐색. **열린교육연구, 21**(3), 257-283.

전영미, 김연희, 유정아, 박경문(2014). 강의평가 공통문항 개발을 위한 기초연구. **교육컨설팅연구, 2**(1), 1-24.

최정웅, 안동규(2016). 데이터분석을 이용한 서술형 강의평가 연구. **디지털융복합연구, 14**(11),

101-106.

한경수, 최숙희, 박재철(2011). 강제적인 대학 강의평가의 문제점. **한국통계학회논문집, 18**(1), 35-45.

한신일(2003). 대학생들의 서술형 강의평가내용 분석. **교육행정학연구, 21**(3), 359-378.

한신일, 김혜정, 이정연(2005). 한국대학의 강의평가실태 분석. **교육행정학연구, 23**(3), 379-403.

홍경선(2006). 대학교 강의평가에 나타난 일관적 응답 분석. **교육정보미디어연구, 12**(2), 97-127.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Bigger, J., & Tang, C. (2011). *Teaching for quality learning at university.* Buckingham, UK: McGraw-Hill International.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Burstein, J., Leacock, C., & Swartz, R. (2001). *Automated evaluation of essays and short answers.* Princeton, NJ: Educational Testing Service. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1790/1/burstein01.pdf

Canini, K.R., Suh, B., & Pirolli, P. (2011). Finding credible information sources in social networks based on content and social structure. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing,* 1-8.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219-233.

Chang, J. (2010). Not-so-latent Dirichlet allocation: Collapsed Gibbs sampling using human judgments. In C. Callison-Burch & M. Dredze (Eds.), *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT) 2010 Workshop on creating speech and language data with Amazon's mechanical turk* (pp. 131-138). Stroudsburg, PA: Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). *Introduction to the Dirichlet Distribution and Related Processes.* UWEE Technical Report Number UWEETR-2010-0006. Washington, DC: University of Washington, Department of Electronic Engineering.

Retrieved from https://pdfs.semanticscholar.org/775e/5727f5df0cb9bf834af2ea2548a696c27a38.pdf?_ga=2.69567060.592048754.1555215321-1579652232.1549010012

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Heinrich, G. (2008). *Parameter estimation for text analysis* (Technical Note). Germany: Fraunhofer IGD. Retrieved from http://www.arbylon.net/publications/text-est.pdf

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307.

Jeon, H., & Kim, T. (2016). KoNLP: Korean NLP package. R package version 0.80.1.

Kwak, M., Kim, S., & Cohen, A. S. (2017, January). *Mining students' constructed response answers.* Paper presented at the International Conference on Writing Analytics, Tampa, FL.

Kwak, M., Xiong, J., Kim, S., Choi, H.-J., & Cohen, A. S. (2018, July). *Dirichlet priors for latent Dirichlet analysis of constructed response items.* Paper presented at the International Meeting Psychometric Society, New York, NY.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science, 58*(3), 754-771.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Scoring, term weighting and the vector space model. *Introduction to Information Retrieval, 100*, 2-4.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

Plasher, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning: A Practice Guide* (NCER2007-2004). Washington, DC: National Center for Education Research, Institute of Education Science. Retrieved from https://files.eric.ed.gov/fulltext/ED498555.pdf

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333-343.

Sizov, S. (2012). Latent geospatial semantics of social media. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(4), 64.

Spiegelhalter, D., Best, N., Carlin, B., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64*(4), 583-639.

Thomas, S. W., Adams, B., Hassan, A. E., & Blostein, D. (2014). Studying software evolution using topic models. *Science of Computer Programming, 80*, 457-479.

Tinto, V. (2005, January). *Taking student success seriously: Rethinking the first year of college.* Paper presented at the Ninth Annual Intersession Academic Affairs Forum, Fullerton, FL.

Wittrock, M. C. (1977). The generative process of memory. In M. C. Wittrock, J. Beatty, J. E. Bogen, M. S. Gazzaniga, H. J. Jerison, S. D. Krashen, R. Nebes, & T. J. Teyler (Eds.), *The human brain* (pp. 153-184). Englewood Cliffs, NJ: Prentice-Hall.

* 논문접수 2019년 2월 7일 / 1차 심사 2019년 3월 11일 / 2차 심사 2019년 5월 9일 / 게재승인 2019년 6월 7일

* 곽민호: 서울대학교 농업생명과학대학 식물생산과학부를 졸업하고, 동대학원 농산업교육학과에서 직업심리 전공으로 석사 학위를 취득하였다. 조지아대학교 통계학과에서 석사 학위를 취득하고, 교육심리학과에서 양적 연구방법 전공으로 박사 과정을 수료하였다.
* E-mail: minho.kwak25@uga.edu

* 민혜리: 이화여자대학교 사범대학 교육학과를 졸업하고, 동대학원 교육학과에서 교육사회학 전공으로 석사 및 박사 학위를 취득하였다. 현재 서울대학교 교수학습개발센터 연구교수로 재직 중이다.
* E-mail: hrmin82@snu.ac.kr

* 김미림: 서울대학교 사범대학 교육학과를 졸업하고, 동대학원 교육학과에서 교육측정 및 평가 전공으로 석사 학위를 취득하였으며, 박사 과정을 수료하였다. 조지아대학교 교육심리학과에서 양적 연구방법 전공으로 박사 학위를 취득하였다. 현재 한국교육과정평가원 부연구위원으로 재직 중이다.
* E-mail: mrkim@kice.re.kr

[부록] DIC 계산을 위한 R 코드

```
# inspect topic range
sequ <- seq(2, 10, 1)

# defining harmonicMean
harmonicMean <- function(logLikelihoods, precision=2000L) {
  library("Rmpfr")
  llMed <- median(logLikelihoods)
  as.double(llMed - log(mean(exp(-mpfr(logLikelihoods,
                                       prec = precision) + llMed))))
}

# extract logliks from each topic
logLiks_many <- lapply(test, function(q) q@logLiks[-c(1:(burnin/keep))])

# compute d_theta_bar
logLiks_many_new <- lapply(test, function(q) q@logLiks[-c(1:iter)])
hm_many2 <- sapply(logLiks_many_new, function(h) harmonicMean(h))
average<-function(v){mean(unlist(logLiks_many[v]))}
t<-sequ-1
dbar<--2*sapply(t, function(g) average(g))

# compute Pd
Pd<-dbar-(-2*hm_many2)
dic<--2*hm_many2+2*Pd

# inspect(dic)
plot(sequ,dic, type="l")
```

Abstract

Analysis of Students' Open-Ended Course Evaluation Using Topic Modeling

Minho Kwak*
Hyeree Min**
Meereem Kim***

This study used a latent Dirichlet allocation (LDA) which is one of the types of topic modeling to analyze the students' open-ended responses regarding the survey of course evaluations. The survey asked the students to describe the unsatisfied and desirable aspects of the courses. 47,000 responses for 1,500 courses in the first semester of 2015 at S University were analyzed. The college-level analysis was performed as well.

The results of this study are as follows. First, the model selection results based on DIC suggested that three-topic model was most suitable for both students' positive and negative feedbacks about course evaluations. Specifically, the three topics of negative feedback are as follows: 1) Improvements of tasks, experiments, and exercises; 2) Improvements of presentations and discussions; 3) Improvements of the test, progress, appear. Also, the analysis of the desirable aspects suggested three main topics: 1) positive feedback on teaching and teaching methods, 2) positive feedback on direct experience and practice, and 3) positive feedback on the quality of the lecture. Second, for the result of the college-level analysis, except a topic reflecting characteristics of each college, the extracted topics from the responses of each college were similar to the topics extracted from the whole combined responses.

This study differs from the previous studies focusing on the analysis of the selected responses regarding course evaluation in that it mainly focused on the open-ended responses of the students. As a result, The study might reveal the positive and negative feedbacks of the course evaluation which the previous studies have not suggested yet.

Key words: Latent Dirichlet Allocation (LDA), Topic Modeling, Text Analysis, Student Course Evaluation

* First author, Ph.D. candidate, University of Georgia
** Research professor, Seoul National University
*** Corresponding author, Associate research fellow, Korea Institute for Curriculum and Evaluation