**Ⅰ. 서론**

1. **연구 배경**

최근 세계 경제가 저성장 국면으로 들어서고 취업난이 지속되면서 대학생에 가중되는 물질적, 심리적 부담이 증가하고 있다. 이에 따라 각 주요 당권들은 대학생들을 위한 정책을 주요 공약으로 내세울 만큼 청년 문제의 해결에 관심이 많다. 한편, 대학생의 고민이나 문제 상황에 대한 파악은 현재 설문 조사나 초점집단인터뷰(FGI) 등 의존하거나, 경제적 상황 등의 외부 거시적 요인으로 사회․심리적 상태를 추측하는 방식으로 이뤄지고 있다. 이런 조사방법론의 한계점은 대상이 연구자를 의식하여 응답과정에서 오류가 있을 수 있다는 것이다. 뿐만 아니라, 특정 문제의식을 가지고 시작하는 조사 방법인 만큼, 응답자도 조사의 목적을 의식하여 특정 프레임(Frame)에 갇힌 응답만을 하게 된다는 문제를 갖고 있다. 따라서 대학생과 관련한 공감 또한 제한된 주제 내에서만 분절적으로 이뤄지기 때문에 사적인 수준에서의 공감은 잘 이뤄지지 않는 현실이다.

한편, 최근 소셜 네트워크 서비스(SNS)의 발달로 많은 대학생들이 페이스북(Facebook) 카카오톡(Kakaotalk) 등을 통해 타인과 소통하고 자신의 고민거리를 나누고 있다. 특히, 대학생들은 각 대학마다 존재하고 페이스북 페이지인 ‘대나무숲’을 이용해 정보와 사연을 공유하고 있다. 각 대학교마다 ‘OO대학교 대나무숲’이라는 형태로 있으며, 해당 대학의 학생 또는 또래의 학생들이 익명성을 기반으로 각자의 고민과 사연을 나누며 공감하고 있다. 해당 페이지에서는 익명성을 기반으로 하기 때문에 타인과 쉽게 나눌 수 없는 자신만의 고민거리를 나누거나 의견을 내는 정보의 장을 구성하고 있다.

1. **연구 목적**

SNS 이용 현황을 근거로 본 연구는 한국 대학생의 사적 영역의 주제 탐색을 위해 Facebook의 활용 가능성을 연구하고자 한다. 2016년 인터넷 이용 실태 조사(한국인터넷진흥원 2016)에 따르면 만 3세 이상 인구의 인터넷 이용률은 88.3%이고, 연령대별로 각 10대 100%, 20대 99.9%, 30대 99.8%, 40대 99.4%의 사용률을 보이는 등 대부분이 인터넷 이용자로 조사되었다. 특히, 20대에서 SNS의 이용률이 91.5%로 가장 높은 것을 확인 할 수 있다. 그리고 20대가 가장 많이 활용하는 SNS는 Facebook(81.8%)로 나타났다. 각각 2위 3위를 기록한 카카오스토리(64.5%) 네이버밴드(31.5%)의 경우는 Facebook과는 달리 블로그와 같이 개인의 사적 공간과 그 이웃들 간의 사적인 소통을 위해 이용되는 경우가 많으므로, 본 연구의 공적 담론 조사 목적과는 부합하지 않는다. 따라서 20대의 인터넷에서의 상호작용을 관찰하기 위한 대상으로 Facebook을 선정하였다. 연구 대상은 Facebook 대나무숲 페이지 이용자이며, 주로 해당 대학교 대학생 및 졸업자 혹은 또래로 구성되어 있다. 본 연구는 잠재의미분석(LSA)을 통해 대학생의 사적 담화 주제를 탐색적으로 살펴 볼 것이다. 이후에는 잠재디리크레할당(LDA) 예측 모형을 통해 대학생의 사적 담화와 주요 관심사를 자세히 분류하고, 나아가 새로운 데이터에 대해 자동으로 분류하는 태그정보로 이러한 데이터에 대한 마중물을 제공할 수 있는 방안을 논하고자 한다.

**Ⅱ. 본론**

1. **데이터 수집**

**가. 분석대상 선정**

본 연구의 분석 대상은 Facebook의 모든 대학교 대나무숲 페이지를 선정하였다. 대나무 숲이라는 이름은 삼국유사에 실린 신라의 48대 왕인 경문왕의 비밀을 복두장이 대나무 숲에 털어놓는 이야기에 착안하여, 대학교 학생들이 자신의 이야기를 익명으로 제보할 수 있도록 만들어놓은 Facebook 커뮤니티이다. 대학교 대나무 숲 커뮤니티는 다음과 같은 특징을 가지고 있다. 1) 사용자의 대부분이 20대이다. 2)사람들의 상호작용이 매우 활성화 되어있다. 3) 사용에 특별한 권한이 필요하지 않다. 4) 좋아요 및 공유, 댓글을 통해 커뮤니티를 직접적으로 사용하지 않는 사람들에게도 담론이 공유 및 재생산된다. 5) 관리자가 있어 제보가 체계적으로 정리된 편이다. 6) 완전 익명 및 부분적 익명의 이용이 가능하다. 부분적 익명이라 함은, Facebook 계정이 가상공간의 프로필임을 이용해서 실재하는 자신의 정보를 숨길 수 있다는 것이다. 이런 특징들과 함께 우리나라 고등교육기관 진학률이 70%에 달함을 고려 할 때(통계청 2010), 20대의 담론을 가장 직접적이고 여과나 편향 없이 관찰하기 위한 커뮤니티로 대학교 대나무 숲 커뮤니티를 선정하였다. 집단에 따른 차이를 최소화하고 데이터의 양을 최대화하기 위해 5천명 이상의 팔로워(구독자)를 가진 대학교 48개 및 그에 준하는 대나무 숲 페이지 3개 총 51개를 선정하였다.

|  |
| --- |
| 성균관대학교, 연세대학교, 고려대학교, 한양대학교, 중앙대학교, 서강대학교,  서울대학교, 한국외국어대학교, 서울시립대학교, 숭실대학교, 건국대학교, 동국대학교, 홍익대학교, 경희대학교, 인하대학교, 한성대학교, 숙명여자대학교, 경기대학교,  상명대학교, 서경대학교, 상지대학교, 한국예술종합학교, 성신여자대학교, 광운대학교, 충남대학교, 가톨릭대학교, 항공대학교, 부경대학교, 강남대학교, 동덕여자대학교,  충북대학교, 동아대학교, 부산대학교, 서울과학기술대학교, 세종대학교, 아주대학교, 가천대학교, 공주대학교, 한밭대학교, 인천대학교, 김천대학교, 단국대학교, 경상대학교, 카이스트대신전해드립니다2, 한국외대글로벌캠퍼스, 순천대학교, 한국교통대학교, 서울여자대학교 백석예술대학교, 전국대학교대나무숲, 미대생대, 의학과/의예과대, |

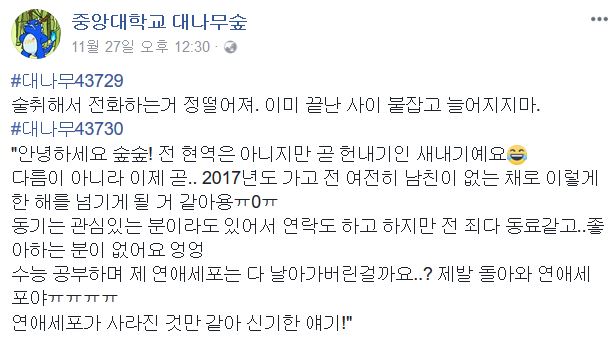
**나. 데이터 수집 및 전처리**

(1) 데이터 수집

데이터 수집은 통계 프로그래밍 언어인 R의 RFacebook 라이브러리를 이용하여 Facebook 커뮤니티의 게시글을 크롤링(Crawling) 하였다. 페이지의 개설일 부터 2017년 10월 16일까지 게시된 모든 게시글의 텍스트만을 크롤링 하였다. 총 51개 목표 페이지 중 49개의 수집에 성공하였으며, 총 107,317개의 게시글을 수집했다.

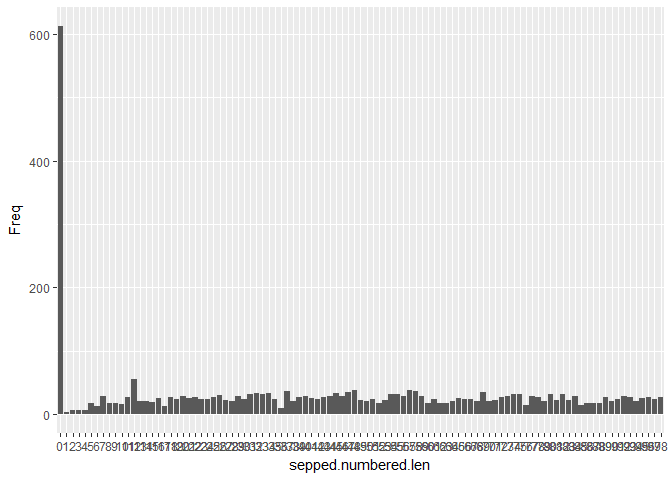
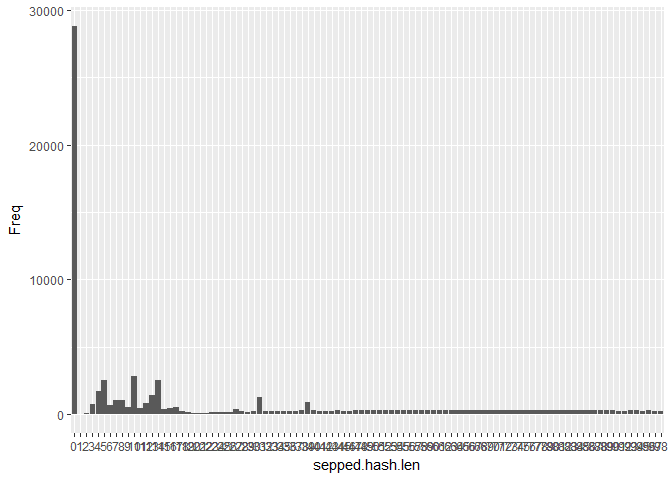
(2) 전처리 1 - 게시글의 제보(사연)별 분리

이후에 자세히 기술할 잠재의미분석, 잠재디리클레 할당은 모두 말뭉치(Corpus)에서 문서-단어행렬(DocumentTermMatrix)을 생성하여 문서 및 단어를 표현한다. 그러나, 대나무 숲 게시글의 경우에는 관리자가 업로드의 편의를 위해 다양한 제보(문서)를 하나의 게시글에 묶어서 업로드 하는 경우가 종종 있다 (그림1). 올바른 분석을 위해서는 이런 맥락과 작성 주체가 혼재된 게시글을 개별적인 문서들로 분리시켜야 할 필요가 있다.



우선, 제보 관리를 체계적으로 하지 않아 게시글을 자동으로 분리하기 어려운 7개 커뮤니티를 데이터에서 제외하였다. 남은 38개 페이지에 대해서 제보의 구분 방법을 조사한 결과, 17개는 해쉬태그 ‘#’로 제보를 구분하였고, 2개는 ‘(1)’, ‘(2)’의 번호로 제보를 구분하였다. 그리고 나머지 23개는 별도의 제보 구분이 필요하지 않았다. 각각의 분리 규칙에 맞게 제보들을 분리하였다.

이렇게 특정문자열(# 등)을 기준으로 글을 분리할 경우, 제보 구분의 목적 외에 사용된 #을 기준으로도 텍스트를 분리하기 때문에, 이를 제거하기 위한 과정이 필요하다. 즉, #와 # 사이에 제보가 아닌 문자열이 들어가는 경우를 제거하기 위해, #사이의 글자 수 빈도를 바탕으로, 히스토그램을 그리고 , 적정 글자 수 이하가 되는 데이터는 제거하였다. #로 분리한 경우<그림 2>는 적정 글자 수가 17개였고, (1)(2) 제보 번호로 분리한 경우<그림 3>는 1개였다.



(3) 전처리 2 – 띄어쓰기

SNS상의 비정형 텍스트 데이터 분석에 있어서 난점 중 하나는, 띄어쓰기가 제대로 되어있지 않다는 것이다. 띄어쓰기가 제대로 되어있지 않은 경우, 형태소분석기로 단어를 추출할 때 형태소 분리가 제대로 되지 않는다는 문제가 있다. 이 문제를 해결하기 위해 딥러닝 기반의 kospacing 패키지를 이용하였다.

(4) 전처리 3 – 불용어 처리

SNS의 비정형 텍스트 데이터에서는 형태소분석기로 단어별 품사를 명확하게 태깅하는 것이 가장 중요하다. 하지만 실제 SNS에 올라오는 글들은 표준어를 준수하지 않는 경우가 많기 때문에 의미가 보통명사만 깔끔하게 추출되지 않는다. 따라서 형태소분석기를 활용하여 보통명사만을 1차적으로 추출해내고, 보통명사가 아님에도 추출된 단어들을 따로 불용어(Stopwords)로 처리해야 한다. 본 연구에서는 형태소분석기에서 보통명사가 아님에도 보통명사로 분류된 단어를 불용어로 처리할 때 참고할 수 있도록 SNS 텍스트 비정형 데이터에 대한 불용어 사전 제작 가이드라인을 제시한다. 가이드라인은 (가)보통명사 외의 품사를 가진 단어 제거 (나)분석 데이터 특성이 반영된 단어 제거 (다)형태소 분석기의 한계로 인한 의미 불명의 단어 제거의 세 단계로 구성된다.

(가) 보통명사 외 품사를 가진 단어

보통명사는 같은 종류의 사물에 두루 쓰이는 명사로서, 특정한 의미단위를 가지고 보편적으로 쓰이는 단어이기 때문에 토픽 모델링의 주요 분석대상이 된다. 하지만 분석에 활용된 데이터가 표준어를 준수하지 않았을 경우 형태소분석기의 분류 정확도가 낮아지기 때문에 종종 실제 단어가 보통명사가 아님에도 보통명사(NC)로 태깅되는 경우가 있었다. 형태소 분석 함수 Simplepos22에서 보통명사로 분류되었지만 실제 보통명사가 아닌 단어에는 다음과 같은 것들이 있었다.

|  |  |  |  |
| --- | --- | --- | --- |
| 5언 | 9품사 | | 예시 |
| 체언 | 명사 | 의존명사 | 것, 수, 만큼, 바, 지, 등, 들, 뿐, 대로 |
| 대명사 | 인칭대명사 | 나, 난, 나는, 저, 본인, 우리, 저희, 너, 너희, 당신, 그대, 그, 그녀, 그들, 자기 |
| 지시대명사 | 그것, 그걸, 무엇, 뭐, 어디, 이리, 언제쯤, 누구, 누군가 |
| 수사 | 양수사 | 하나(한), 둘(두), 셋(세), 일, 이, 삼 |
| 서수사 | 첫째, 둘째, 셋째, 제일, 제이, 제삼 |
| 용언 | 동사 | 그래도, 합니다, 했는데, 가고, 같아요, 알고 | |
| 형용사 | 사소한, 이상한, 그렇게, 이렇게, 다른, 좋은, 큰 | |
| 수식언 | 관형사 | 새, 헌, 정, 구, 대 | |
| 부사 | 제발, 빨리, 많이 | |
| 관계언 | 조사 | 을, 를, 이, 가, 와, 과, 에서, 에게, 로서, 부터, 까지 | |
| 독립언 | 감탄사 | . | |

이 외에도 접속어(그리고, 그래서)를 불용어로 추가하였다. 또한, ‘동안’처럼 문법적으로 보통명사가 맞지만 단어만으로는 의미 파악이 명쾌하지 않은 단어 역시 고려해야 한다.

(나) 분석데이터의 특성을 반영한 단어

첫째, 게시글 머리말 태깅(Tagging)으로 인해 반복 등장하는 문구를 불용어로 추가하였다. 대다수의 대나무 숲 페이지에서 첫 번째 줄에 게시글 번호, 학교 특성을 담은 문구, 날짜와 시간을 기재하였는데, 이 중 숫자 데이터는 따로 삭제했기 때문에 문자 부분만 삭제하였다. 학교 특성을 담은 문구에는 날갯짓, 아우성, 포효, 송골매, 선구자, 사자후, 울림 등이 있었다. 날짜와 시간에는 월~일요일, 오전, 오후 등이 있었다.

둘째, 언급 횟수는 많으나 맥락을 알 수 없는 경우가 많았던 대학교 이름은 모두 불용어로 추가하였다. 고려대, 광운대, 동아대, 부경대, 부산대, 상명대, 서강대, 서울여대, 숙명여대, 아주대, 연대, 연세대, 한밭대 등의 단어를 삭제하였다.

셋째, 커뮤니티에서 자주 사용되는 용어를 불용어로 추가하였다. 제보자, 관리자님, 익명제보, 게시글, 대나무숲, 대나무, 대숲과 같이 페이지 자체나 운영 전반에 관한 단어들과 경희숲, 아대숲, 충남대숲, 한예숲, 홍숲 등 특정 대학교의 대나무 숲 페이지를 줄여서 일컫는 말을 삭제하였다.

넷째, ㅜ, ㅠ, ㅋ, ㅎ와 같이 커뮤니티에서 일상적으로 사용 되나 문장부호, 영문, 숫자가 아니어서 걸러지지 않는 단어를 불용어로 추가하였다.

(다) 형태소분석기의 성능 한계로 인한 의미 불명의 단어

신조어와 유행어로 인해 형태소분석기가 일반명사를 제대로 추출해내지 못하는 경우가 종종 있었다. 예를 들어, ‘안녕하세용’과 같은 단어의 경우 ‘안녕하세+용’으로 분리되어 ‘용’이라는 단어가 추출되었다. 여기서의 용은 동서양 신화에 등장하는 상상의 동물일 수 있지만, ‘안녕하세용’과 같은 대화체에서 추출된 경우도 있기 때문에 이와같은 단어를 고려하여 삭제해야 한다. 유, 식, 로, 용, 기, 채, 세, 죠, 치, 썰, 하, 두, 주, 갈 등의 단어가 삭제되었다.

(5) 전처리 4 – 명사 추출

명사 사전은 KONLP 패키지의 NIADic을 사용했다. 명사 추출 함수는 여러 명사 추출 명령어에 대한 성능 비교를 해본 결과 Simplepos22를 선택했다. 107,317개의 문서에 대해 1,424,577개의 단어가 추출됐다.

(6) 전처리5 – 문서­단어 행렬 (Document­Term Matrix)

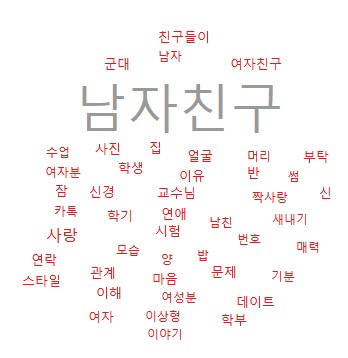
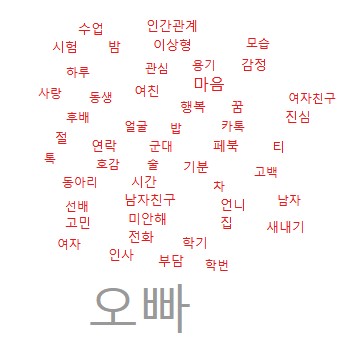
tm패키지의 DocumentTermMatrix 명령어를 이용해 문서-단어 행렬을 구성했다. 행렬을 구성할 때 2글자에서 20글자, TfI-df 가중치를 적용해 성능을 높이고자 하였다. 행렬의 희소성(Sparsity)을 줄이기 위해 문서 대비 단어의 빈도가 지나치게 낮은 단어들을 제거하였다.

**Ⅲ. 분석결과**

**1. 현황 분석(빈도분석)**

결과로 도출된 문서-단어 행렬의 단어 빈도를 분석했다. 그 결과, 연애, 남자친구, 사랑, 마음의 순서대로 가장 높은 빈도수를 기록했다. 이외에도 연락, 오빠, 여자친구, 동아리, 여자, 남자 등의 단어가 높은 빈도수를 기록했다.

**2. 담화 분석-1 (잠재의미분석)**

잠재의미분석(LSA)은 문서-단어 행렬을 분석하여 의미를 파악하는 방법 중의 하나이다. 단어들이 어떠한 맥락 속에서 나타나는지에 대한 정보를 분석하여 의미를 파악하게 되는데, 각 단어들이 어떤 맥락에서 출현했는지에 대한 정보도 포함되어 있다.

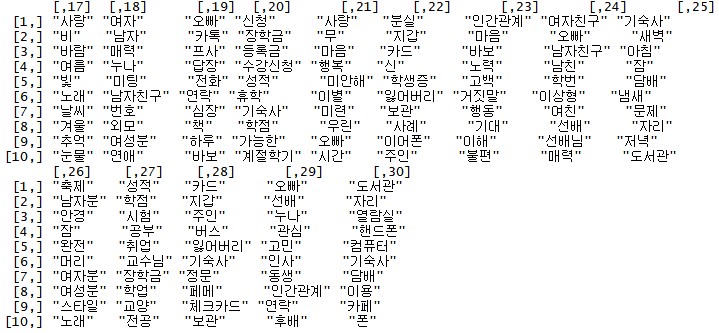


대학생 담화 주제의 다양성을 고려해 5차원에서 적절한 잠재의미분석 결과를 도출했다. 그 결과 차원 1은 ‘자신의 남자친구에 대한 여성의 고민’, 차원 2는 ‘이성적인 관심을 갖고 있는 연상의 남성에 대한 여성의 고민’, 차원 3은 ‘잃어버린 물건에 대한 제보’, 차원 4는 ‘사랑과 연애 자체에 대한 고민’, 차원 5는 ‘자신의 여자 친구에 대한 남성의 고민’ 으로 나타났다. 이를 통해 대학생들의 개인적인 고민은 주로 연애와 사랑에 대한 것임을 확인할 수 있었다. 특히, 대나무 숲에서 여성의 활동이 더 많음을 추론할 수 있었다. 이는 SNS의 특성과도 관련이 있는 것으로 보인다. 연애와 사랑 이외에는 대학생들이 대나무 숲이라는 페이지를 교내의 분실물을 찾기 위한 용도로 활용하고 있는 것을 알 수 있었다. 이는 학내 커뮤니티가 발달하지 않은 많은 학교에서 익명의 사연을 제보하는 대나무 숲 페이지가 분실물을 제보 하고 찾는 기능을 맡고 있음을 알 수 있다.

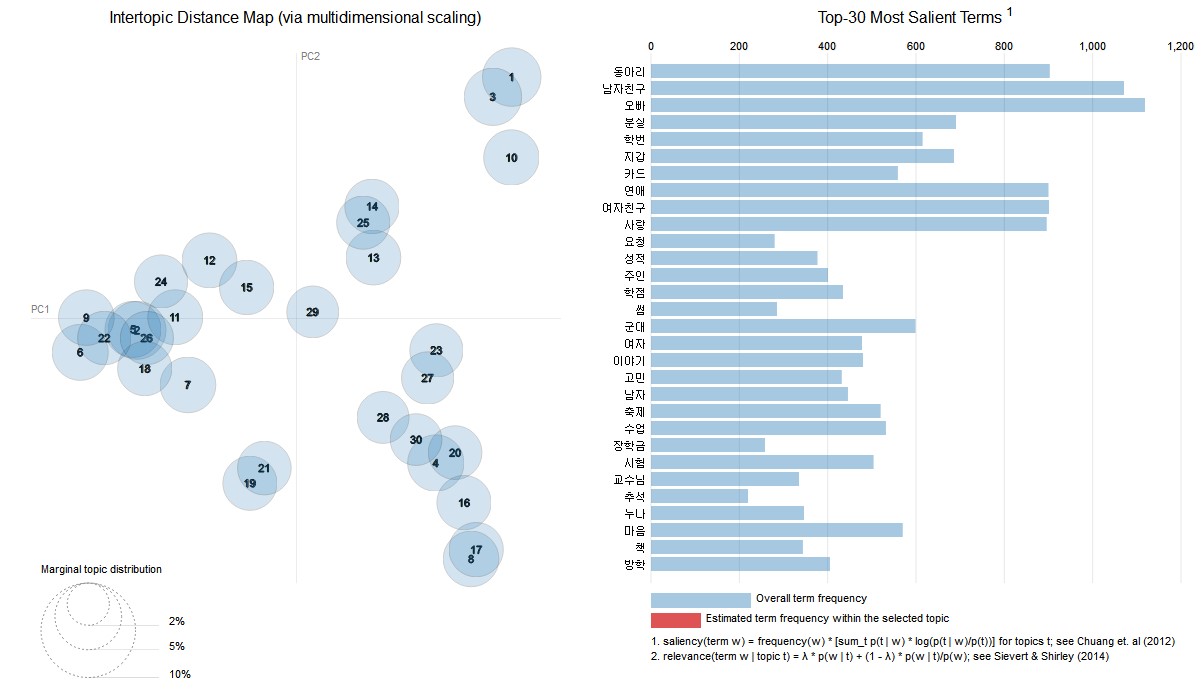
**3. 담화 분석-2 (잠재디리크레할당)**

잠재디리크레할당(LDA)는 말뭉치로부터 토픽을 추출하는 토픽모델링(Topic Modeling) 기법 가운데 하나이다. 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률모형이며, 주제별 단어의 분포, 문서별 토픽의 분포를 모두 추정해내는 머신러닝(Machine Learning) 기법이다. 이는 시간의 흐름에 따른 주제의 변화, 새로운 문서에 대한 주제 분류가 가능하다는 것이 강점이다. 혼잡도(Perplexity) 지표를 이용해 약 30개의 주제를 상정하고 LDA를 실시했다. 전체 깁스 샘플러 방법은 변수간의 의존성이 커서 수렴이 늦은 단점이 있다. Θ와 Φ 적분하여 제거하는 것을 통해 변수간의 의존성을 줄일 수 있어 붕괴 깁스 샘플러 방식을 사용했다. (Griffiths and Steyvers, 2004). 사후확률의 업데이트 회수는 5000회, 제외하는 초기값의 개수는 1000개, alpha = 0.01, eta = 0.01로 설정했다.

분석 결과, LSA에서와 같이 주요 주제는 연애와 사랑, 그리고 분실물 제보와 관련된 것이었다. 그러나, LDA 분석을 통해 새롭게 드러난 몇 가지의 차원이 있었다. 3번째 차원에서는 대학생들의 등록금에 대한 고민을 확인할 수 있었다. 주요 단어로는 ‘등록금’,’신청’,’장학금’,’휴학’ 등으로 나타났다. 13번째 차원에서는 대학생들의 성적과 관련된 주제가 뚜렷하게 드러났다. 주요 단어로 ‘성적’,’학점’,’학업’,’공부’ 등이 나타나며 대학생들의 성적에 대한 관심을 확인할 수 있었다. 교내에서의 불만사항에 대한 주제들도 새롭게 등장하였는데 구체적으로는 28번째의 기숙사 불만사항이 있었다. 학생들은 ‘새벽’, ‘아침’, ‘냄새’, ‘담배’ 등을 언급하며, 새벽과 아침 시간의 소음문제와 담배냄새에 대한 불만사항을 이야기하고 있었다. 11번째의 도서관 불만사항 등에 대한 것이었다. ‘열람실’, ‘자리’, ‘핸드폰’, ‘컴퓨터’ 등을 언급하며 열람실에서 핸드폰이나 컴퓨터 사용으로 인해 발생하는 소음문제에 대한 불만사항이 주를 이루었다. 이외에도 29번째 차원에서는 ‘엄마’, ‘아빠’, ‘아버지’, ‘생일’, ‘선물’ 등의 단어가 언급됐다. 이는 생일을 맞이한 부모님의 선물거리를 고민하는 대학생들의 모습을 보여주고 있었다. 마지막으로 8번째 차원에서는 인간관계에 대한 주제가 확인되었다. 주요 언급 단어는 ‘인간관계’, ‘거짓말’, ‘노력’, ‘불편’ 등이 나타난 것으로 보아 인간관계로 고민과 갈등을 겪고 있는 대학생들의 생각을 읽을 수 있었다.



LDA 분석 결과의 시각화를 위해 json 파일을 이용해 인터랙티브(Interactive) 시각화를 수행했다. 이를 통해 각 주제가 전체에서 차지하고 있는 비중, 주제 내에서 단어의 비중, 주제 간의 거리 등을 확인할 수 있다.



**Ⅳ. 결론 및 제언**

**1. 결론**

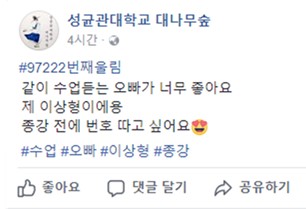
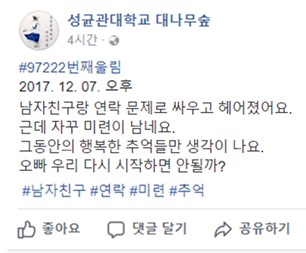
‘이성과의 관계’에 대한 주제는 여러 차원으로 구성되어 있었다. 잠재의미분석(LSA)의 결과, ‘이성과의 관계’의 범주에 속하는 주제 차원이‘차원1: 자신의 남자친구에 대한 여성의 고민’, ‘차원2: 이성적인 관심을 갖고 있는 연상의 남성에 대한 여성의 고민, ‘차원 4: 사랑과 연애 자체에 대한 고민’, ‘차원 5: 자신의 여자친구에 대한 남성의 고민’ 등 4개의 차원으로 나왔다. 이는 동일한 주제로 고려되는 담론이라도, 상호작용의 주체와 형태에 따라 여러 차원을 가진다는 결론을 내릴 수 있다.

이와 같은 내용은 잠재 디리크레 할당(LDA)의 결과에서도 동일하게 나타났다. 그러나, 토픽의 개수를 30개로 설정했기 때문에 더욱 다양한 주제가 드러났다. 구체적으로는 연애에서의 이별, 갈등 뒤의 사과, 연락, 이상형, 후회 등으로 나타났다. 이외에 등장한 주제로는 수업, 성적, 기숙사와 도서관에 대한 불편사항, 부모님의 생일선물, 신입생 적응, 동아리, 인간관계 등이 있었다. 대학생의 주요 관심사 및 고민 이외에도 주제에서 드러난 단어들을 분석한 결과 이러한 SNS 활동에 여성의 비율이 더 높음을 알 수 있었다. 또한, 분실물 제보 용도로 페이스북 페이지를 사용하고 있어 기존의 학교에서 담당하지 못하고 있는 기능을 일부 담당하고 있는 것으로 나타났다.

**2. 연구 의의 및 제언**

대학생들의 SNS상의 담론을 잠재의미분석과 잠재 디리크레 할당을 통해 내재한 독립적 주제들을 탐사함으로써, 대학생들의 솔직한 고민거리, 주요 관심사 등을 직접적으로 확인 할 수 있었다. 여론조사나 기존의 연구에서 이뤄진 대학생의 불안 이나 갈등요소에 대한 연구는, 대학생이 응답을 하는 과정에서 응답편향, 회상의 오류, 사회적 당위성에 의해서 오류가 있는 응답을 할 수 있다. 이에 대해 본 연구는 대학생 담론 분석의 대안을 제시했다는 의의가 있다. 그리고 동일한 범주로 생각되는 주제에 대해, 잠재한 세분화된 주제가 존재함을 밝힘으로써, SNS상의 대학생의 상호작용에 대해 보다 다양한 연구 가설을 세울 수 있는 가능성을 제시하였다. 개인화 되고 파편화 되어 알기 어려운 20대들의 솔직한 고민거리를 상대적으로 쉽게 파악할 수 있는 것이다.

또한, 머신러닝(LDA)을 통해 대학생의 대나무숲 제보에 대해 학습함으로써, 새로운 데이터에 대해서 주제분류가 가능하다. 새로운 데이터에 대한 분류 기능은 분석하기 어려웠던 SNS 비정형 텍스트 데이터에 대한 태그정보를 부여하여 접근성을 높일 수 있다. 자동 주제 분류기를 만들 수 있는 것이다. 예를 들어, 새로운 글에 대해 특정한 토픽이 일정 확률 이상을 갖는다면 해당 토픽 내의 단어들과 일치하는 단어들을 태그정보로 자동 주제 분류해 마중물을 제공할 수 있다. 이를 통해 비정형 데이터를 좀 더 잘 정리된 메타데이터와 함께 분류하고 관리하게 가능하게되어, 대학생들을 모수로하는 각종 연구 및 마케팅의 활용할 수 있는 기초 자료로서 페이스북 대나무숲 게시글의 활용 가능성을 제시했다.



<그림 11> 자동 주제 분류기로 분류된 제보 예시

**3. 한계점**

첫째로 문서-단어행렬을 통한 분석의 한계는, 단어를 문서의 맥락에 기반하여 분석하기 때문에 문서 맥락 외적인 관점에서 단어의 의미를 파악 할 수 없다는 것이다. 즉, 본 연구의 분석 과정에 있어서 문서 외적인 관점에서 불필요하다고 여겨지는 불용어들을 제외하는데 연구자의 수작업이 필요하다는 것이다. 따라서, 데이터가 달라져 문서의 맥락이 바뀌게 된다면, 불용어 사전이 새롭게 작성되어야 하는 것이다. 또한, 문장에서 단어를 분리 할 때 사용한 형태소분석기는 학습모델에 의해 형태소를 추출하는 것이 아니기 때문에 동일한 의미의 단어를 서로 다른 단어로 추출한 경우가 있어 데이터의 희소성(Sparsity)를 증가시켰다. 이는 정보의 손실을 의미하며 올바른 분석결과 해석에 어려움을 준다. 뉴럴네트워크(NN)를 이용한 Word2vec등의 알고리즘을 이용한다면 위 문제를 개선 할 수 있을 것으로 기대된다.

둘째로 컴퓨팅 성능의 한계로 다양한 파라미터(Parameter) 조정을 통한 분석의 한계가 있다. 희소성(Sparsity)를 줄이기 위해 문서 대비 단어의 빈도에 대한 기준을 설정하였으나 컴퓨터 성능의 한계로 문서대비 0.2% 보다 빈도가 낮은 단어의 경우 제거할 수밖에 없었다.

마지막으로 표본 선정에서의 한계가 있다. Facebook 게시글을 수집하는데 Facebook API를 이용했는데, Facebook에서는 과도한 트래픽을 줄이기 위해 Facebook 정보 접근에 대해 일정한 기준으로 제한을 하고 있다. 최근 2달을 제외하면 이전의 게시물 정보에 대해서는 매우 적은 양의 정보만을 추출할 수 있었다.

**[참고문헌]**

An Introduction to Latent Semantic Analysis. TK Landauder, Peter W. Foltz & Darrell Laham. Discourse Processes. Volume 25. 1998. Issue 2-3.

Finding Scientific Topics. Thomas L. Griffiths & Mark Steyvers. 2004.

Latent Dirichlet Allocation. David M. Blei, Andrew Y. Ng & Michael I. Jordan. Journal of Machine Learning Research. (2003). 993-1022

LSA 모형에서의 다의어 의미의 표상. 이태헌, 김청택. 한국인지과학회. 인지과학 15(2). 2004.6. 23-31

한국 미래창조과학부 외. (2016). 인터넷은 '소통'이다. 2016년 인터넷이용실태조사. 한국인터넷진흥원.

한국교육개발원 교육통계분석자료집, 통계청(2010) 장래추계인구