네이버 지식인 질문을 통한 텍스트 분석 : 우울증 키워드로 알아 본 사용자 특징과 언어

사람들은 때로 오프라인보다 온라인에서 진실을 드러낸다. 네이버 지식인은 네이버가 운영하는 QnA서비스로 사용자 간 질문과 답변을 하면서 지식,정보를 교류할 수 있다. 포털사이트 검색으로는 찾을 수 없었던 질문의 답을 사용자들의 집단지성이 모여 해결하는 것이 본래 서비스의 취지였으나 점차 사용자들은 단순 정보 교류를 넘어, 온라인의 익명성에 기대어 고민상담 질문을 올리고 있다.

한편, 우울증은 우리 사회에 널리 퍼져 있는 정신 질환 중 하나로 많은 현대인들이 고통 받고 있다. 최근 코로나 사태로 인해 집에 홀로 고립된 시간이 길어지며 우울증의 심각성은 더해지고 있다. 우울증은 개인이 질환을 인지하지 못하는 경우도 있고, 인지하더라도 정신과 방문을 꺼려해 제때 치료 시기를 놓치곤 한다.

소셜미디어의 발전에 따라 사람들은 온라인 공간을 일상생활과 자기 감정 표현의 공간으로 활용하고 있다. 특히 우울증 증세가 있는 사람들은 온라인에서 자신의 심리적 상황을 표현하고 다른 사람들에게 도움을 받기도 한다. 네이버 지식인은 익명성을 보장하며, 다양한 전문가의 도움을 받을 수 있는 플랫폼이다. 우울증과 같은 민감한 주제에 대해 정보 공유와 상담이 이뤄지기도 한다. 따라서 본 프로젝트는 네이버 지식인에 게시된 “우울증” 키워드 글을 분석하여 우울증 게시글 사용자 특징과 언어를 파악하고자 한다.

1. 데이터 수집

네이버 지식인에서 2022년 1월 1일 ~ 2022년 4월 30일 기간을 설정한 후 “우울증” 키워드로 검색하여 나타난 결과를 파이썬의 Beautifulsoup 패키지를 사용하여 수집하였다. 수집된 항목으로는 게시일자, 작성자(작성자 아이디 또는 공개/비공개 여부를 알 수 있는 정보), 작성제목, 작성내용, 태그(질문 분야)가 있으며, 데이터는 총 26,563개이다.

2. 전처리

- (컬럼 병합) 작성제목과 작성내용은 사용자가 작성한 본문이므로 분석 편의성을 위해 두 컬럼을 병합하였다.
- (결측치 제거) 수집된 데이터 26,563개 중 조사기간(2022.01.01~2022.04.30)에 벗어난 데이터(1,581건), 결측 데이터(3건), 중복 데이터(3,443건)을 제거하여 21,536건을 추출하였다.

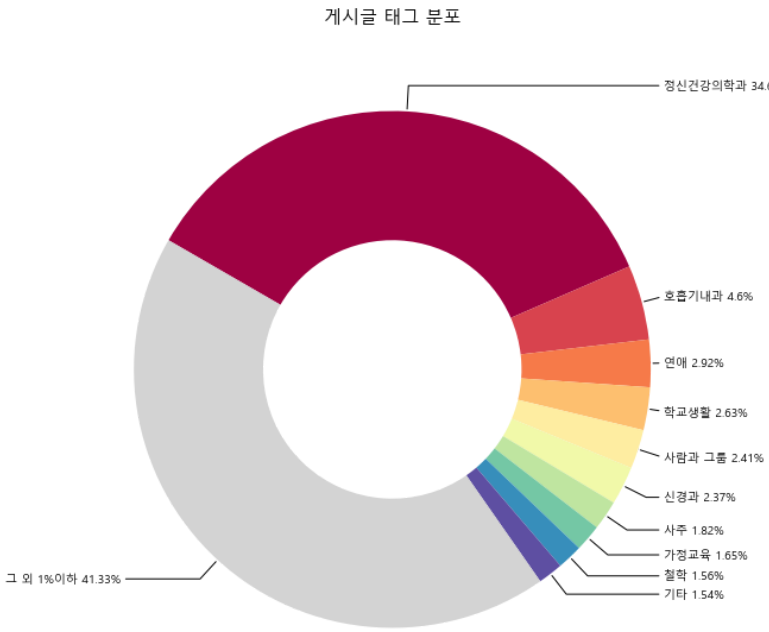
자연어 처리를 위해서 한국어 자연어 처리 패키지 KoNLPy의 Okt모듈을 사용하였다. Okt는 트위터에서 개발한 Twitter한국어 처리기에서 파생된 오픈 소스로 완벽하게 한국어의 형태소를 분리해내기 보다는 빠르게 분석하여 추출하는 것을 지향한다. 네이버 지식인은 대부분 소셜미디어에 익숙한 사용바가 편리하게 글을 게시하는 공간으로, 맞춤법, 띄어쓰기 등의 문법이 부정확하고 신조어, 이모티콘 등의 비문법적인 단어들이 쓰인다. 따라서 Twitter 한국어 처리기에서 파생된 Okt형태소 분석기를 사용하는 것이 적합하다고 판단하였다.

- (문자열 정제) 토큰화를 진행하기 전, 문자열을 정제를 수행하였다. 한글과 영어를 제외한 모든 문자열(이모티콘, 구두점, 특수문자, 숫자 등)을 공백으로 대치하였다. 날짜 단위(년,월,일,달 등), 나이 단위(살, 대, 세대), 시간 단위(시,분, 시간), 학교 단위(고3,중3, 1학년 등)와 같은 자주 언급되지만 숫자를 제거할 경우 의미 없어지는 단어들을 제거하였다.
- (토큰화, 불용어 제거) 토큰화를 위해 게시글 본문별로 Okt 명사추출기를 사용해 단어를 분류하였고, 유의미한 단어를 선별하기 위해 분석에 도움 되지 않는 단어들(불용어*)을 제거하였다. 21536건의 데이터 중 명사가 추출되지 않은 본문 7건을 제외, 총 21529건을 분석을 위한 데이터셋으로 설정하였다.

  *보편적으로 제공하는 한국어 불용어 리스트(https://www.ranks.nl/stopwords/korean) 675개, 사용자가 임의로 추가한 단어 767개. 사용자가 설정한 불용어 사전의 단어수는 총 1442개입니다.

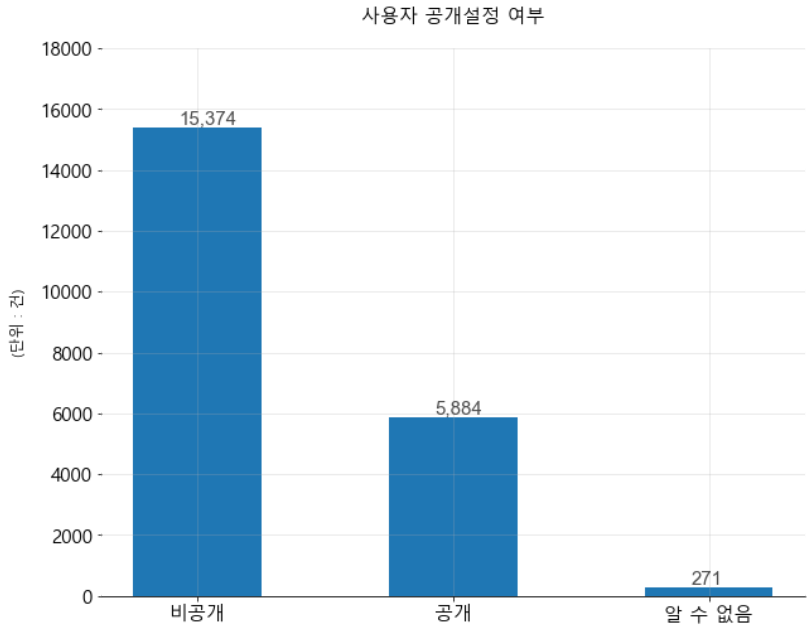
3. 데이터 분석

1) 게시글 분석

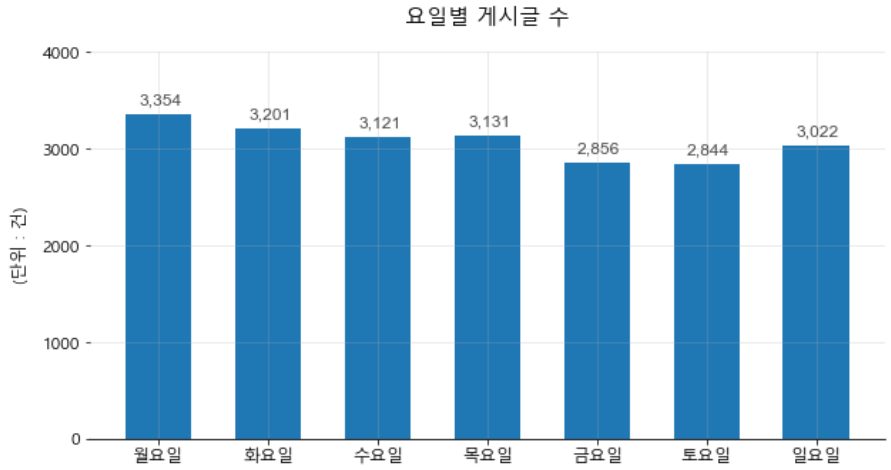


rank	tag	cnt(건)	pct(%)
1	정신건강의학과	7456	34.63
2	호흡기내과	991	4.60
3	연애	628	2.92
4	학교생활	567	2.63
5	사람과 그룹	519	2.41
6	신경과	510	2.37
7	사주	391	1.82
8	가정교육	356	1.65
9	철학	336	1.56
10	기타	332	1.54
-	그 외 1%이하	9131	41.33

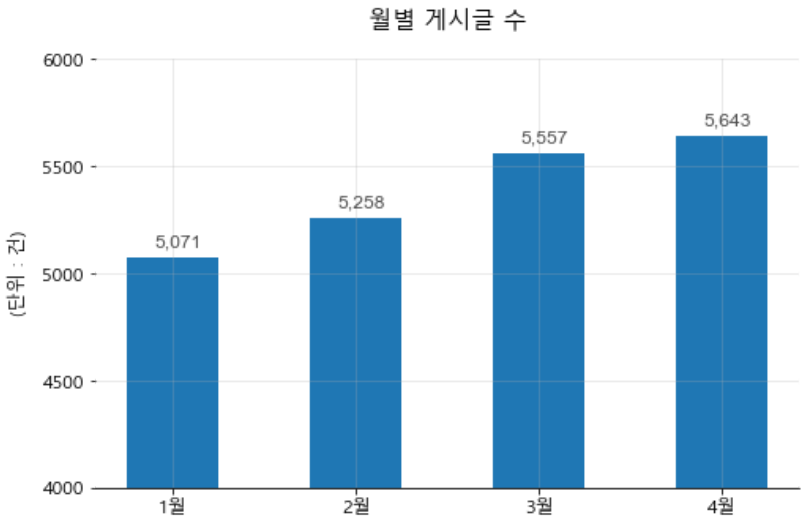
- (게시글 태그 분포) 네이버 지식인(QnA)에 질문을 등록할 때 더 빠르고 정확한 답변을 받기 위해 사용자가 질문에 맞는 질문 분야를 태그로 지정할 수 있다. 사용자가 등록한 여러 개의 태그 중 첫 번째로 지정된 태그를 대표 태그로 분류하여 데이터 21,529건의 분포를 살펴보았다. `정신건강의학과` 태그가 7456건(34.63%)으로 가장 많았다. 이어 `호흡기내과`(4.6%), `신경과`(2.37%) 를 포함하여 의료 분야 태그가 전체 데이터셋의 40%를 차지하였다. 그 다음으로는 `연애`(2.92%), `학교생활`(2.63%), `사람과 그룹`(2.41%), `가정교육`(1.65%)과 같은 관계와 관련한 대한 태그가 약 10%를 차지하였다.



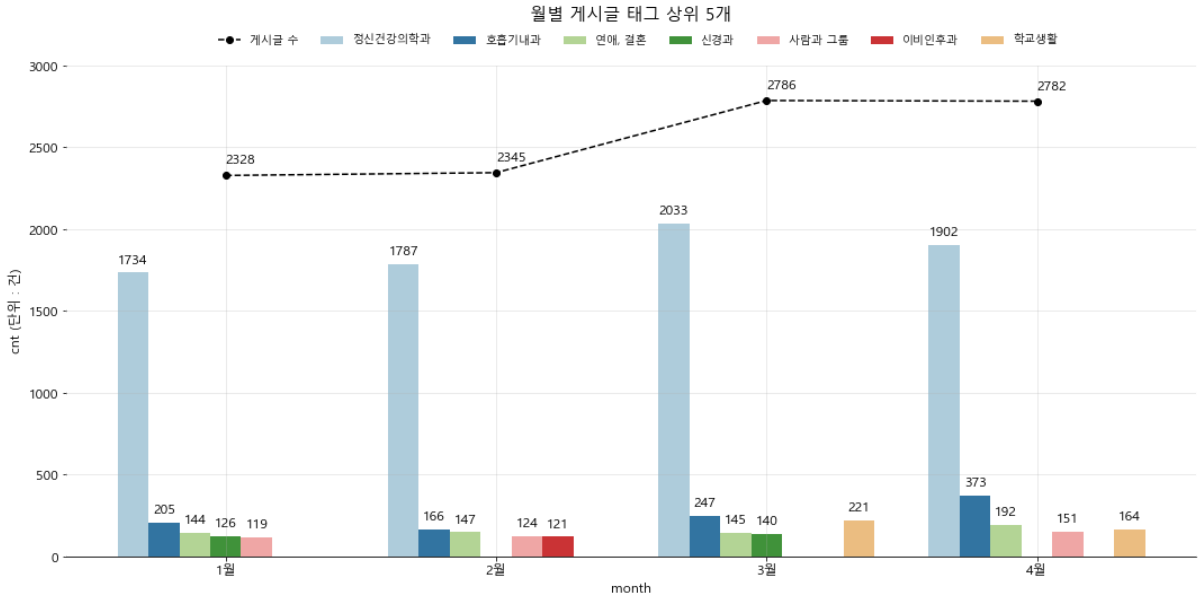
- (사용자 공개설정 여부) 게시글 등록 시 사용자 공개설정 여부에선 `비공개`(15, 374건)가 압도적으로 많았다. 이는 사용자 대부분이 익명으로 게시글을 작성하여 신상정보를 드러내고 싶어하지 않는 것을 알 수 있다. `알 수 없음` (271건)은 사용자가 네이버 계정을 탈퇴하거나 정보를 찾을 수 없는 계정인 것으로 보인다.



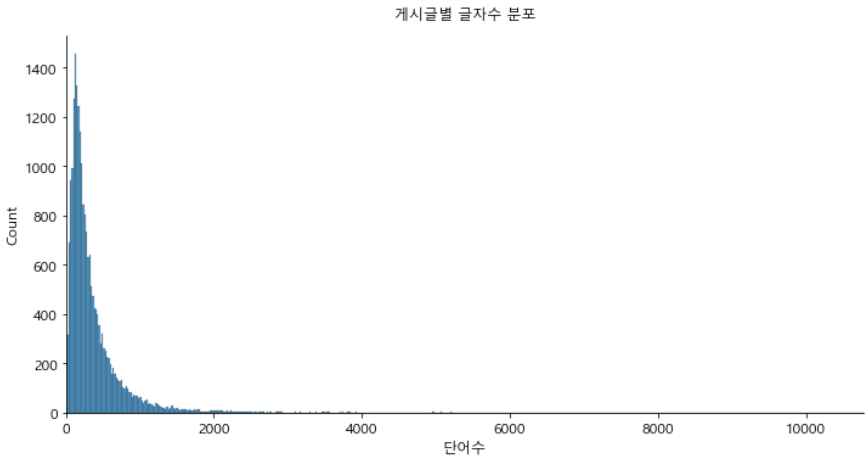
- (요일별 게시글 수) 요일별 게시글 수를 보면 대게 요일별 차이가 크지는 않지만, `월요일`이 3,354건으로 가장 많고 `토요일`이 2,844건으로 가장 적다. 평일보다는 주말에 게시글의 수가 적은 것을 보아 평일에 우울감을 더 느낄 수도 있다고 추측할 수 있다.



- (월별 게시글 수) 월별 게시글은 조사기간(1월~4월) 중 1월이 5,071건으로 가장 적고 4월이 5,643건으로 가장 많다. 연초에는 비교적 우울증 관련글이 적으나 봄이 시작되는 3월부터 게시글 수가 증가하고 있다. 이는 계절별, 환경적 요인이 우울증에 영향을 미칠 수도 있다고 추측할 수 있다.

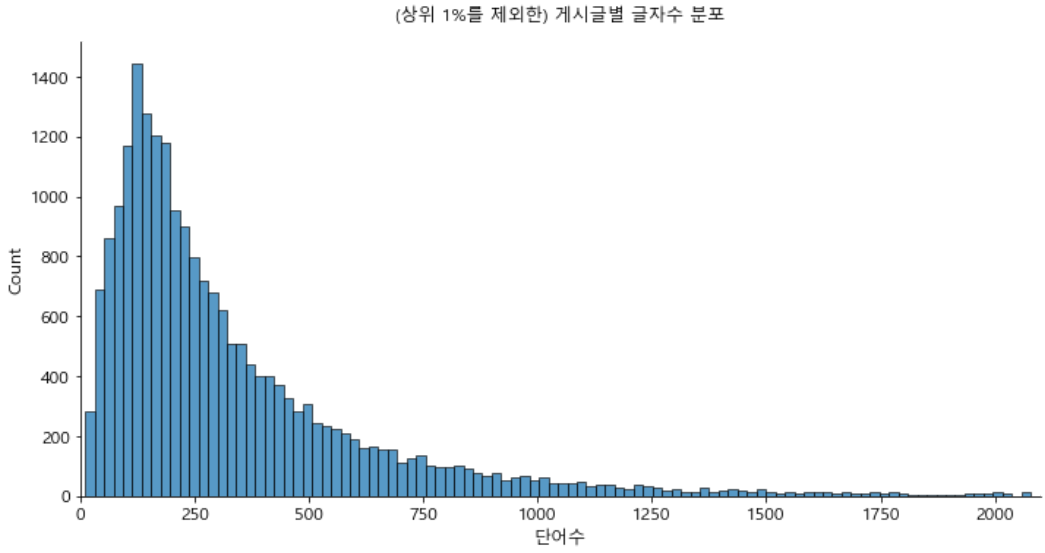


월별 게시글 태그 상위 5개를 보았을 때, 1~4월 모두 정신건강의학과 태그가 압도적으로 많다. 이어서 호흡기내과, 연애, 결혼 태그가 1~4월까지 2,3위를 차지한다. 3월은 전월에 비해 게시글 수가 크게 증가했으며, 특히 정신건강의학과 태그 글이 가장 많다. 또한 3, 4월에 학교생활 태그가 상위 5개 순위권에 포함 된 것으로 보아 10~20대 사용자들이 증가한 것으로 예상된다.



	word_num
count	21529
mean	362
std	435
min	11
25%	135
50%	234
75%	433
max	10777

- (게시글 글자수) 게시글의 평균 글자수는 362자이고 최소 길이는 11자, 최대 길이는 10,777자이다. 글자수 분포는 극단값의 차이가 매우 크고 왼쪽으로 치우쳐져 있다. 글자수의 통계량을 보면, 3분위수(75%)의 값이 433자인 반면 최대값은 10,777자로 이상치로 인해 글자수의 분포를 제대로 파악할 수가 없다. 따라서 글자수 상위 1%를 제외하고 99%의 분포를 다시 살펴보았다.



게시글 99%가 2000자 이내로 작성되었으며, 대게 500자 이내인 것을 확인할 수 있다. 또한 많은 사용자들이 250자 이내로 게시글을 작성하며 비교적 상세하게 게시글을 작성하나 500자 이상의 긴 글을 작성하지 않는 것을 확인할 수 있다.

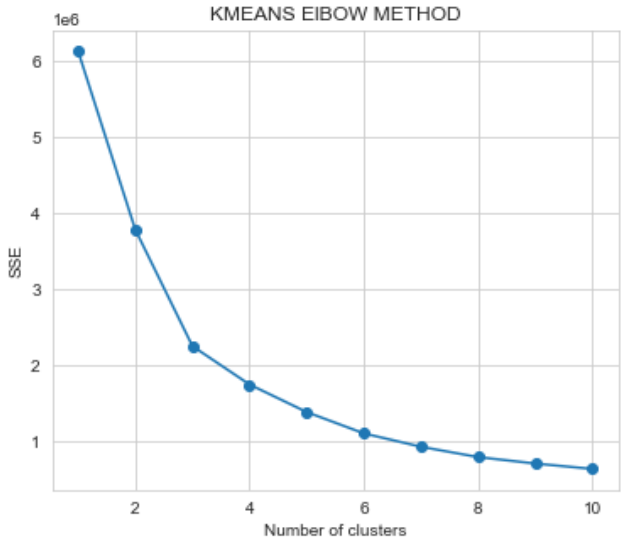
rank	word	cnt(단위 : 건)	rank	word	cnt(단위 : 건)
1	우울증	23819	11	치료	4938
2	생각	14349	12	공부	4793
3	말	10750	13	정신	4555
4	친구	9368	14	부모님	4293
5	사람	9157	15	스트레스	4272
6	약	7131	16	학교	4011
7	엄마	6819	17	조루증	3804
8	증상	5599	18	느낌	3601
9	병원	5486	19	잠	3557
10	집	5024	20	문제	3369

- (단어 빈도수) 게시글 본문에 사용된 단어들(토큰화 한 명사)의 빈도수를 살펴보면, `우울증`이 23,819건으로 가장 많이 사용되었다. 이어 `생각`(14,349건), `말`(10,750건), `친구`(9,368건) 가 뒤를 잇는다. 빈도수가 많은 단어들의 특징을 살펴보면, `엄마`가 6,819건으로 많이 언급되는 반면 `아빠`라는 단어는 3,329건으로 `엄마` 빈도수의 1/2로 적은 편이다. 또한 우울증 다음으로 가장 많이 언급되는 질환으로 조루증(3,804건)이 있다. 여성 질환 단어에 비해 남성 질환 단어들은 빈도수가 많은 편이다. `친구`, `공부`, `학교` 같은 단어들이 많이 언급되는 것으로 보아 10,20대 학생들이 주 사용자인 것을 확인할 수 있다.

2) 군집분석

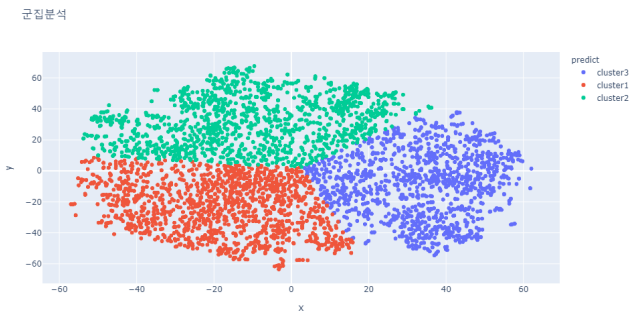
군집분석이란 개체들을 분류하기 위한 기준이 없는 상태에서 주어진 데이터들의 속성값을 고려해 유사한 개체끼리 그룹화하는 방법이다. 군집분석 유형에는 비계층적 군집분석과 계층적 군집분석이 있는데 본 프로젝트는 계산복잡도가 작고 대량의 데이터에 유리한 비계층적 군집방법을 채택하여 분석을 진행하였다.

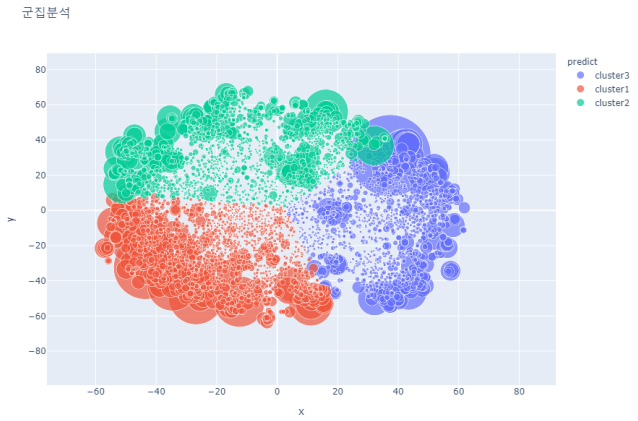
군집분석을 수행하기 앞서 자연어처리 라이브러리인 Gensim을 사용하여 토큰별로 Word2Vec모델로 벡터화시켜 3889x100로 변환시켰다. 이때 Word2Vec은 CBOW로 진행하였다. 3889 x 100로 이루어진 고차원으로 군집분석을 할 경우 모델의 성능이 저하되고 overfitting의 가능성이 우려되어 차원축소를 통해 feature 수를 줄였다. 차원축소 방법에는 PCA와 t-SNE가 있는데 텍스트 마이닝의 차원축소 방법으로 많이 사용되는 t-SNE를 적용하여 2차원으로 축소(3889 x 2)하였다. 텍스트 유사성에 다른 군집을 분류하고자 scikit-learn 라이브러리를 활용하여 k-means 클러스터링을 수행하였다. 최적의 군집 수(k값)를 산정하기 위해 elbow기법을 활용하였고 그래프를 통해 k값에 따른 오차제곱합을 구하였다.



그래프에 따라 k값이 3일때부터 ss가 줄어드는 비율이 급격하게 작아지고 있다. 이 지점을 elbow point라고 판단하여 최적의 군집수를 3으로 설정하여 군집분석을 실시하였다.

3개의 군집으로 데이터를 분류한 후 빈도 수가 많은 단어들을 군집별로 나열하여 특징을 살펴보았다. 군집1은 '생각', '친구', '사람', '엄마', '스트레스', '눈물', '기분', '걱정' 등으로 인간관계와 감정표현에 관한 단어들이 분포되었다. 군집2는 '약', '증상', '잠', '몸', '코로나', '복용' 등으로 건강상태를 설명하는 단어들이 분포되었다. 군집3은 '우울증', '병원', '치료', '정신과', '상담', '상태' 등으로 질병, 치료에 관한 단어들이 분포되었다. 군집별로 주제를 설정하면 인간관계, 건강상태, 질병상담 으로 분류할 수 있다.



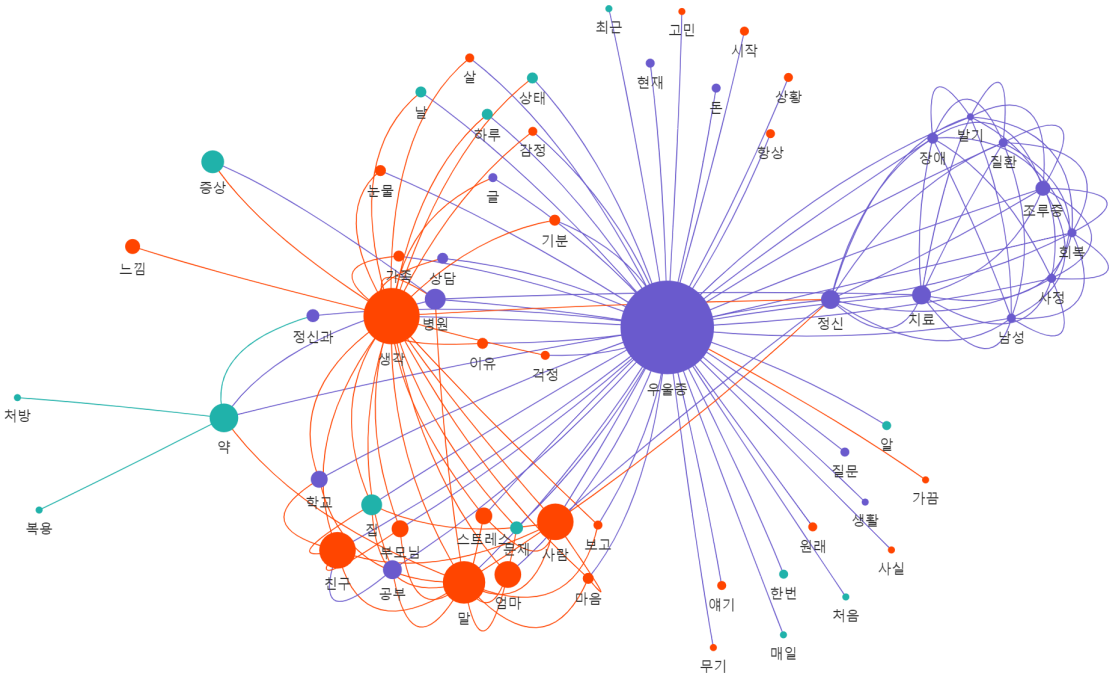


cluster	cluster1		cluster2		cluster3	
column	word	cnt	word	cnt	word	cnt
0	생각	14349.0	약	7131.0	우울증	23819.0
1	말	10750.0	증상	5599.0	병원	5486.0
2	친구	9368.0	집	5024.0	치료	4938.0
3	사람	9157.0	잠	3557.0	공부	4793.0
4	엄마	6819.0	문제	3369.0	정신	4555.0
5	부모님	4293.0	머리	2858.0	학교	4011.0
6	스트레스	4272.0	하루	2683.0	조루증	3804.0
7	느낌	3601.0	상태	2672.0	정신과	3266.0
8	아빠	3329.0	몸	2651.0	상담	2944.0
9	마음	2950.0	날	2535.0	장애	2551.0
10	눈물	2737.0	소리	2535.0	검사	2517.0
11	이유	2642.0	방법	2338.0	질문	2241.0
12	가족	2621.0	숨	2312.0	질환	2198.0
13	기분	2570.0	알	2252.0	돈	2142.0
14	걱정	2457.0	꿈	2163.0	사정	2141.0
15	살	2366.0	한번	2141.0	회복	2139.0
16	얘기	2231.0	매일	1959.0	남성	2114.0
17	감정	2183.0	코로나	1954.0	글	2072.0
18	보고	2152.0	오늘	1903.0	현재	2040.0
19	시작	2062.0	복용	1842.0	발기	1936.0

3) 연관분석

연관분석은 한 문장 안에 A라는 단어와 B라는 단어가 동시에 들어가 있을 때 두 단어는 연관성이 있다고 판단한다. 동시 출현하는 빈도수가 높을수록 그 두 단어의 연관성은 높아진다. 각 데이터를 토큰화하여 명사인 단어만 추출한 뒤 apriori알고리즘으로 지지도(support)가 0.04 이상, 향상도(lift)가 1 이상을 기준으로 하여 연관분석을 수행하였다. 지지도는 집합 내에서 얼마나 자주 나타나는지에 따른 빈도수를 가리키며, 향상도는 규칙이 우연에 의해 발생한 것인지 아닌지 판단하는 연관성 정도의 척도로 1 이상은 양의 상관 관계를 나타낸다. 또한 단일 단어에 대한 연관성을 살펴보기 위해 항목수를 2개로 제한하였다. 이를 통해 148개의 단어 쌍을 구하였으며 시각화를 위해 네트워크 그래프를 사용하였다. 60개의 노드와 148개의 엣지를 가지고 있으며, 각 노드의 크기는 전체 데이터셋에서 단어의 등장 횟수, 노드의 색깔은 군집 분석에 따라 분류된 군집별 색상이다.

먼저, 가장 눈에 띄는 키워드는 '우울증'으로 60개의 노드 중 55개의 노드와 연관성을 가지고 있다. 이어 '생각'은 29개의 노드, '말'은 16개의 노드를 가진다. 군집3의 단어들은 연관 정도가 매우 높은데 반해 군집2의 단어들은 연관 정도가 낮다. 특히, 군집3의 '조루증', '사정', '발기', '남성', '질환', '회복', '장애', '치료' 이 8개의 단어들은 단어들끼리 연관성이 매우 크게 형성되어 있다. 또한 군집1과 군집3은 군집 간 연관성이 보이지만, 군집2는 군집 내/군집 간 연관성이 낮아 보인다. 이는 군집2의 주제인 건강 상태는 개개인의 주관적인 설명이기 때문으로 보인다.



4. 결론 및 한계

네이버 지식인의 우울증 키워드를 설정하여 데이터를 수집한 뒤 우울의 심각성을 보이는 글을 미리 파악하고 예방하는 분석을 진행하고자 하였다. 하지만 분석을 실시한 결과 우울과 관련된 글들은 대게 본인이 처한 상황과 감정설명, 건강상태로 일상적인 주제의 글들과 구별이 쉽지 않아 보였다.

게시글 유형 분석을 통해 사용자의 특성을 살펴보고, 군집분석을 통해 우울증 글에 자주 사용되는 단어들과 군집별 특징을 찾았다. 연관분석을 통해 특정 단어가 출현했을 경우 연관성 높은 단어를 파악함으로써 우울증 게시글의 특성을 파악하였다.

- (게시글 유형 분석) 우울증 키워드로 검색되는 글의 약 1/3은 `정신의학과` 분야로 정신과 질병, 상담과 같은 글이 대다수였으며 다른 분야에 비해 압도적으로 태그 수가 많았다. 그 다음으로 `호흡기내과` 분야의 글이 많은데 이는 공황장애, 우울증 질환에서 비롯된 증상과 호흡곤란으로 숨쉬기 어려운 상황을 설명하는 글이 대다수였다. 특히, 호흡곤란으로 생활의 불편함을 겪거나 알 수 없는 이유로 숨을 쉴 수 없어 원인이 무엇인지 상담하는 글이 많았는데, 이는 본인의 질병을 인지하지 못한 채 도움을 청하는 글로 보인다.
  요일별 게시글 수를 보면, 주말보다 평일에 더 많은 글이 작성된다. 월요일과 일요일은 약 500건 가량 게시글 수의 차이가 나는데, 이는 회사 생활, 학교생활과 같은 사회 활동이 우울증에 영향을 미치는 것으로 보인다.

  또한, 1월~4월까지 월별 게시글 수가 점차 증가하고, 특히 3월 게시글 수가 급증하며 `학교생활` 태그가 상위를 차지하는 것으로 보아 10대,20대 학생들이 개학 이후 친구 및 학업 문제 등으로 네이버 지식인을 이용하는 것을 알 수 있다.
- (군집 분석) 수집 된 데이터에 담겨진 단어들을 3개의 군집으로 나누었을 때, `군집1 인간관계`, `군집2 건강상태`, `군집3 질병상담` 으로 분류되었다. 세부적으로 군집1은 '친구', '엄마', '부모님' 등의 관계를 칭하는 단어와 '스트레스', '눈물', '기분', '감정' 등의 감정 표현 단어들이 있다. 군집2에는 '약', '증상', '건강', '잠', '머리', '숨' 등 몸 상태를 설명하는 단어나 최근 신체적 증상에 대한 단어들이 포함되어 있다. 군집3은 '우울증', '치료', '정신과', '상담' 등 병명과 치료에 관한 단어들이 등장했다. 또한 군집3은 남성 질환과 증상들을 설명하는 단어들이 많은 것으로 보아 남성들의 건강관련 상담글이 우울증과 관련하여 많이 작성되는 것을 알 수 있다.
- (연관 분석) 하나의 게시글에 등장하는 단어들의 연관성을 살펴보았을 때, '조루증', '사정', '발기', '남성', '질환', '회복', '장애', '치료' 의 남성 질환 단어들이 높은 향상도를 가지며 강한 연관성을 가졌다. 하지만 위 단어들은 강한 연관성에 비해 전체 데이터에 등장하는 횟수는 적은 편이다. '우울증', '생각'은 게시글에서 가장 많이 등장하는 단어로 많은 단어들과 연관성을 가지지만 낮은 향상도를 가진다. 연관분석 결과를 군집분석에서 분류된 군집으로 나누어 보았을 때, 군집3의 단어들은 강한 연관성을 가지는 반면, 군집1과 군집2의 연관성은 크지 않은 편이다. 인간관계(군집1)와 건강상태(군집2)는 개개인의 상황과 설명에 따라 다양한 단어들을 사용하게 되므로 단어별 연관도가 높지 않은 것으로 보인다.

본 프로젝트에서 우울증 관련 게시글을 선별하여 데이터를 수집하는데 한계가 있었다. 소셜미디어의 우울 분석의 경우 "우울증 환자"의 소셜미디어 텍스트를 분석하는 방식으로 진행된다. 하지만 본 프로젝트의 경우 우울증 환자를 특정할 수 없어 임의로 "우울증"이라는 키워드로 우울 경향 글을 수집하였다. 따라서 우울과 관련이 없거나, 우울증 환자와 관련 없는 글도 데이터에 포함되어 분석이 진행되었다. 또한 데이터 전처리 과정에서 불용어 처리에 미흡하였다. 'ㅈㅏㅅㅏㄹ', 'ㅈㅅ'처럼 단어를 분해하거나 축약해 놓은 단어와 비속어, 인터넷 용어 등을 사용자 사전을 구축하여 구축하였어야 했으나 시간상의 여건으로 수행하지 못하였다.

후속 프로젝트에서 데이터 수집 시 우울 경향 글을 선별하는데 더욱 적합한 방법을 적용하고, 사용자의 특성(예를들어, 성별, 이용시간, 답변 수, 답변 분야 등)을 포함하여 다양한 정보를 수집, 정확한 데이터를 추출하기 위해 사용자 사전을 구축한 텍스트 전처리를 수행한다면 더욱 유의미한 결과를 얻을 것이다.