감성 분류 말뭉치 빈도 분석

1. 데이터 개요

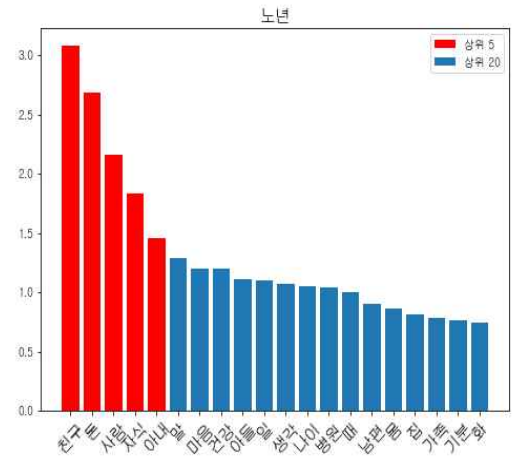
- 총 5130개의 데이터 포인트
- 각 데이터 포인트는 연령, 성별, 상황키워드, 감정_대분류, 감정_소분류, 사람문장 1~4, 시스템문장 1~4가 존재
- 사람문장3,4와 시스템문장3,4는 빈값이 존재

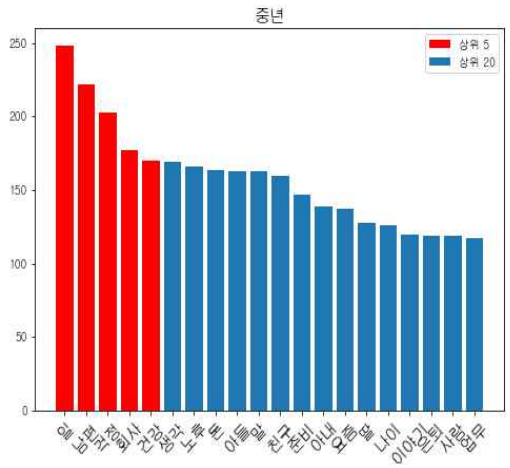
2. 목표 - 연령, 성별, 감정_대분류 분류항목에 따라 어떤 어휘적 특징이 있는지 확인

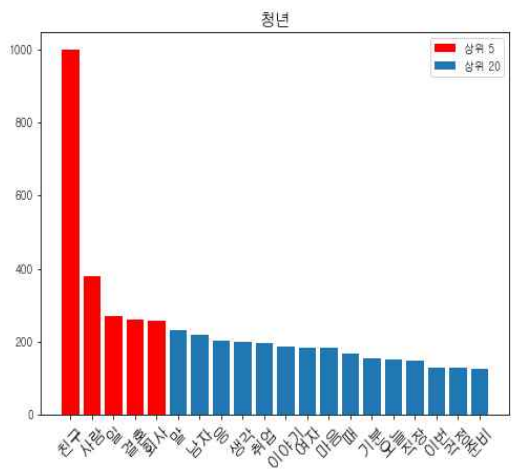
- 연령 4단계 구분(노년, 중년, 청년, 청소년)
- 성별 2단계 구분(남성, 여성)
- 감정_대분류 6단계 구분(기쁨, 당황, 분노, 불안, 상처, 슬픔)
- 각 분류항목별로 형태소 분석을 통해 어휘적 특징을 조사

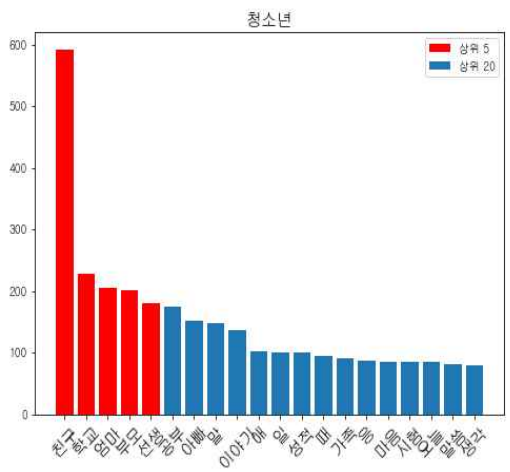
3. 가설1 - 명사태그와 어근을 추출한다면 각 분류항목별 관심대상을 알 수 있을 것이다.

- 일반명사(NNG), 고유명사(NNP), 어근(XR) 추출
- 등장빈도 기준 상위 5개 항목, 상위 20개 항목 관찰
  - 연령 분류항목 기준 등장빈도 상위 20개 항목

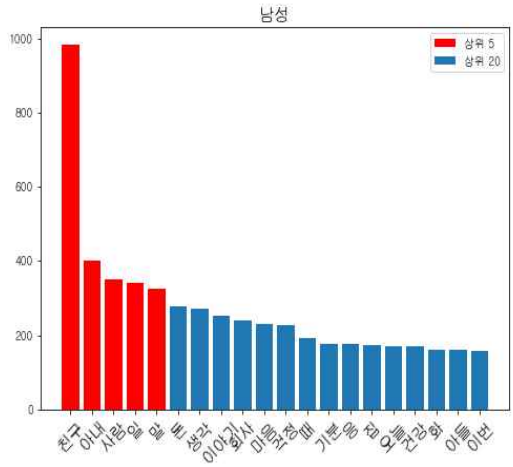


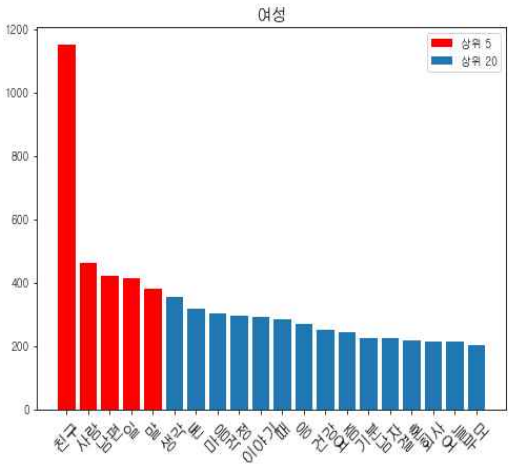




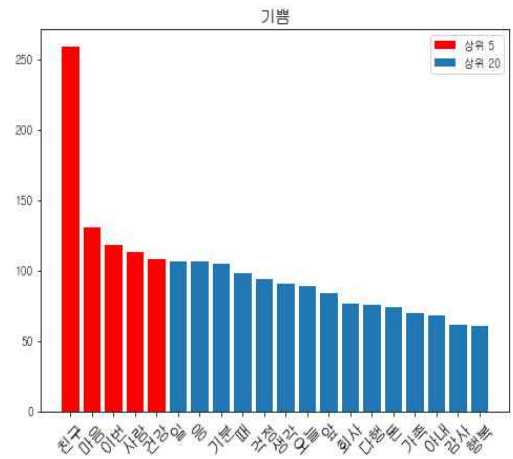


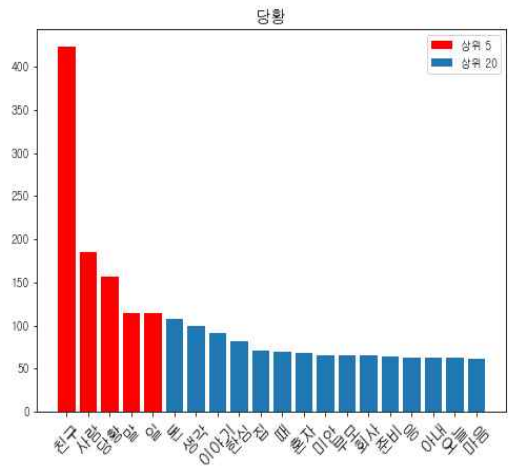
- 성별 분류항목 기준 등장빈도 상위 20개 항목

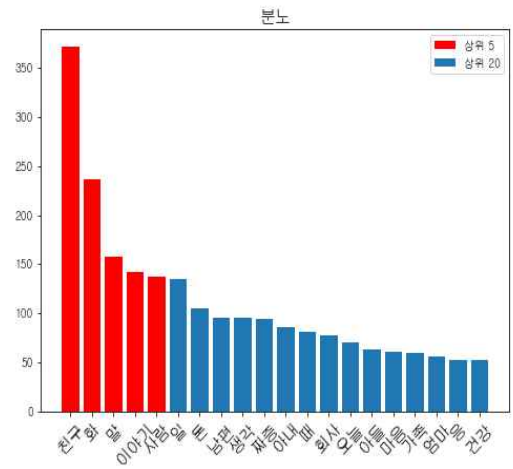


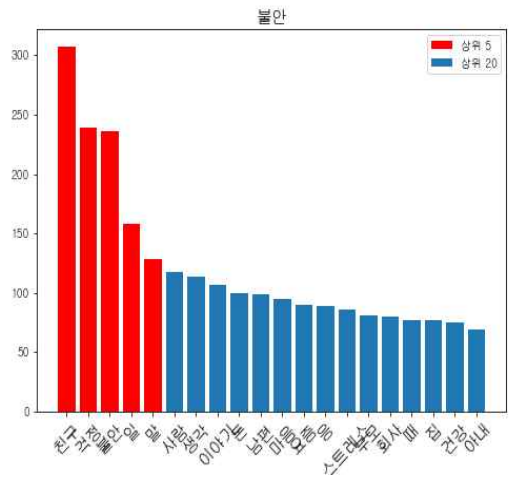


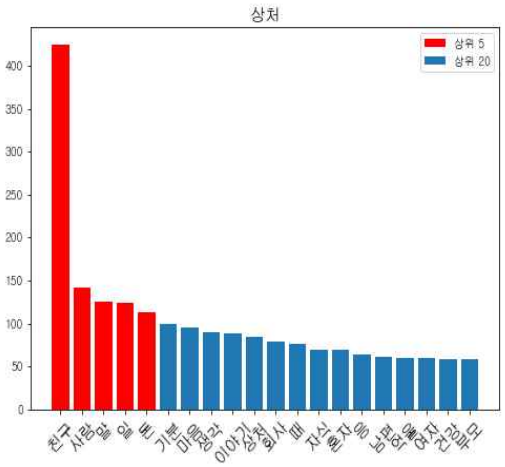
- 연령 분류항목 기준 등장빈도 상위 20개 항목

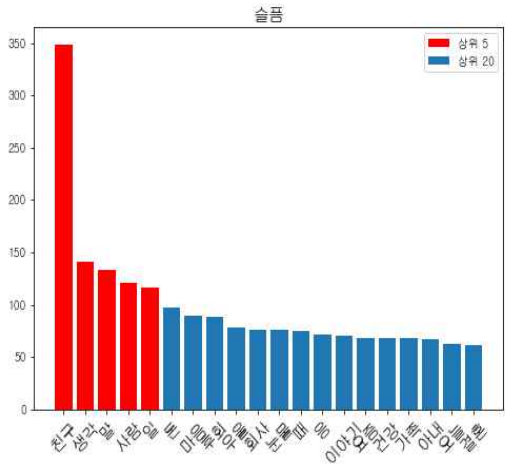












3.1 결과

- 연령 분류항목에 대한 빈도 분석

각 연령 분류별로 관심사에 대한 내용을 찾아볼 수 있다. 중년을 제외한 나머지 항목에 대해서는 친구 키워드가 가장 높게 나타났다. 노년층에서는 자식 키워드를 볼 수 있고, 중년층에서는 일, 회사, 노후 키워드에서 두드러진 특징을 보인다. 청년층과 청소년층에서는 친구 키워드의 빈도가 압도적으로 높고, 청소년층에서는 학교, 부모, 선생 키워드가 높은 빈도가 나타났다.

노년		중년		청년		청소년	
친구	3.08	일	1.98	친구	7.0	친구	7.33
돈	2.69	남편	1.77	사람	2.67	학교	2.83
사람	2.16	회사	1.41	일	1.9	엄마	2.55
자식	1.83	노후	1.33	회사	1.79	부모	2.5
아내	1.46	돈	1.31	남자	1.54	선생	2.23

- 성별 분류항목에 대한 빈도 분석

성별에 차이에 따른 빈도차이는 뚜렷하게 나타나지 않았다. 친구, 사람, 일, 말이 공통적으로 유사하게 나타났으며 같은 의미를 지낸 아내, 남편 키워드가 각각 나타났다.

남성		여성	
친구	4.58	친구	4.46
아내	1.88	사람	1.8
사람	1.65	남편	1.64
일	1.59	일	1.61
말	1.42	말	1.4

- 감정 분류항목에 대한 빈도 분석

감정에 따른 빈도 분석에서는 전체적으로 친구에 관련된 내용이 많다. 각 감정에 따른 키워드 등장빈도는 전체적으로 유사하게 나타났다.

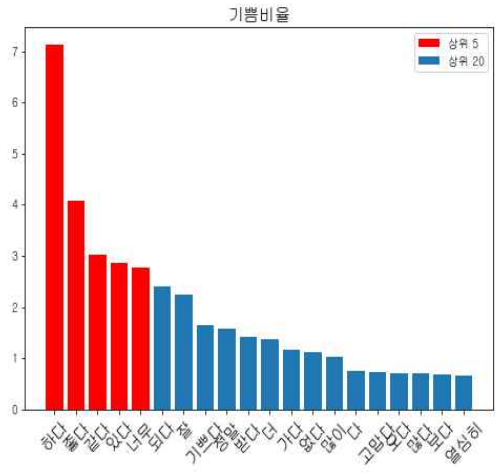
기쁨		당황		분노		불안		상처		슬픔	
친구	3.32	친구	5.53	친구	4.56	친구	3.69	친구	5.55	친구	4.52
마음	1.69	사람	2.42	화	2.93	일	1.91	사람	1.87	말	1.65
이번	1.52	일	1.49	말	1.86	걱정	1.58	일	1.63	사람	1.58
사람	1.45	돈	1.41	사람	1.69	사람	1.42	말	1.58	일	1.52
일	1.38	말	1.35	일	1.67	말	1.39	돈	1.49	생각	1.28

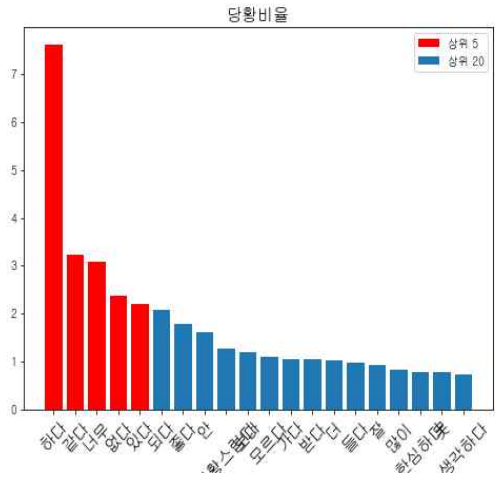
3.2 결론

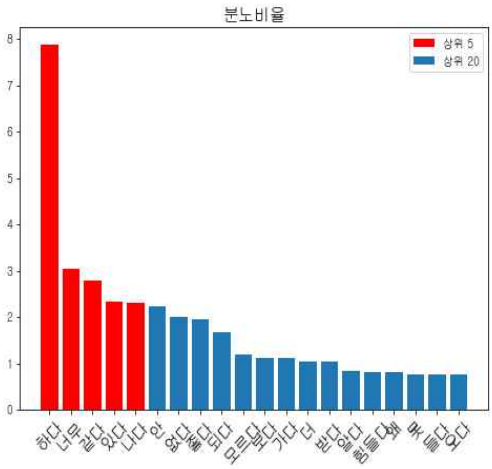
- 명사와 어근을 이용하게 형태소 빈도분석을 했을 때, 관심 분야에 대한 분석이 어느 정도 가능했다. 하지만 감정 분류항목에 대해서는 어떠한 감정을 가지고 있는지 분석하는 데에 충분한 정보가 제공되지 않는다. 감정을 나타내는 표현이 주로 용언에서 등장하기 때문에 이를 배제된 상황에서 나타난 결과라고 생각할 수 있다.

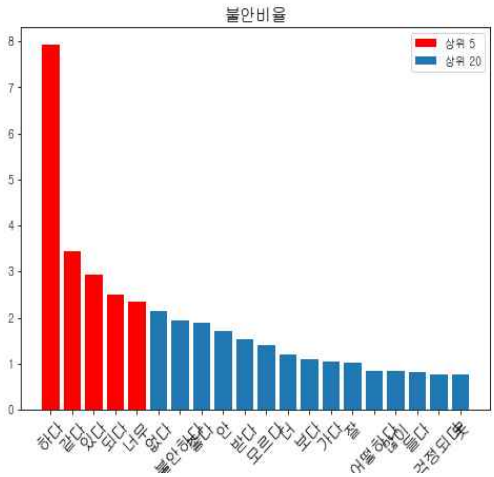
4. 가설2 - 용언(동사, 형용사)과 일반부사를 추출한다면 행위나 감정을 파악할 수 있을 것이다.

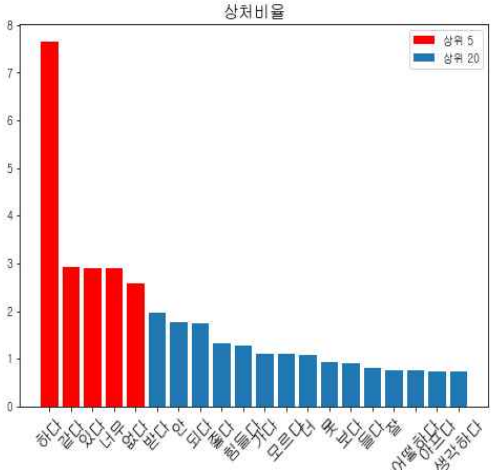
- 동사(VV), 형용사(VA), 일반부사(MAG) 추출
- 감정 분류항목에 대해 등장빈도 기준 상위 5개, 상위 20개 항목 조사

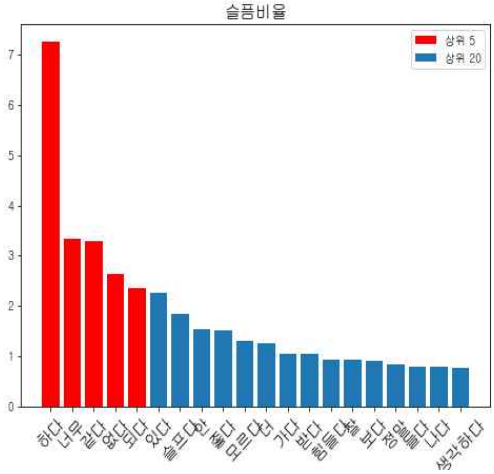












4.1 결과

- 감정 분류항목에 대한 빈도 분석

동사, 형용사, 일반부사에 대해서 감정 분류항목에 대한 키워드 등장 빈도는 하다, 같다, 있다 등의 공통된 용언 비율이 높다. 이를 통해서 얻을 수 있는 정보는 제약적이며 이는 사전에 불용어로 처리하지 않아 발생한 결과로 여겨진다. 기쁨과 당황에서는 그 감정을 그대로 담은 기쁘다와 당황스럽다가 등장하였으며, 나머지 감정에 대해서는 부정의 단어인 안이 등장하였다.

기쁨		당황		분노		불안		상처		슬픔	
하다	7.12	하다	7.6	하다	7.87	하다	7.92	하다	7.64	하다	7.25
좋다	4.08	같다	3.22	너무	3.04	같다	3.44	같다	2.92	너무	3.34
같다	3.02	너무	3.09	같다	2.78	있다	2.93	있다	2.91	같다	3.3
있다	2.87	없다	2.38	있다	2.34	되다	2.5	너무	2.9	없다	2.64
너무	2.77	있다	2.2	나다	2.3	너무	2.34	없다	2.59	되다	2.35

4.2 결론

- 많은 의미를 담지 않고 있는 하다, 같다, 있다 등의 단어가 정보를 얻는데 방해를 하고 있다. 해당 단어들을 제외하는 처리를 해주어야 보다 뚜렷한 정보를 얻을 수 있을 것으로 생각된다. 또한 각 감정별로 뚜렷하게 등장하는 단어들을 많이 보이지 않고 해당감정을 그대로 서술한 단어만이 보인다.

5. 가설3 - 성별 분류항목에서 조사를 추출하여 비교하면 특정 성별에서 빈도가 높은 어투를 찾을 수 있을 것이다.

- 형태소 분석을 통해 모든 조사를 추출한다.
- 등장하는 모든 항목에 대하여 빈도가 상대성별의 빈도의 1.5배 이상이거나, 한쪽성별에서만 등장하는 형태소를 찾아 비교한다.

5.1 결과

- 등장하는 형태소에 따른 남성과 여성에서의 단어의 빈도 비율

형태소	남성	여성
서부터	0.01	x
로부터	0.03	0.02
아	x	0.01
부터	0.07	0.04
처럼	0.21	0.41
이서	0.03	0.02
더러	x	0.02
으로써	x	0.01
로서	0.03	0.02
하고	0.09	0.06
으로서	0.03	0.01

- 전체 문장에서 조사에 해당하는 형태소는 40개가 등장하였다. 이중 상위 20개 항목에서는 서로의 비율이 1.5배 이상 차이 나는 내용이 존재하지 않았다.
- 처럼을 제외한 형태소는 등장 비율이 0.1%를 넘기는 항목이 없었다.

5.2 결론

- 조사를 통해서 남성과 여성을 확실히 구분 짓는 방법은 제한적이다. 등장 비율의 차이가 있다고 하더라도 실제 등장 비율이 낮기 때문에 이를 통해서 특정 성별에 특징이 있다고 판단하기에는 무리가 있다. 해당 데이터에서 문장들이 문어체에 가깝게 서술되었기 때문에 이 특징이 나타나지 않는 것으로 추정된다.

6. 가설4 - 성별 분류항목에서 어미를 이용해서 중복되는 형태소를 제거하면 성별을 구분하는 특징을 발견할 수 있을 것이다.

- 성별 항목에서 어미에 해당하는 형태소를 추출한다.
- 중복되는 형태소의 비율을 남성 여성 서로 비교한다.
- 비율의 차이가 1% 이하인 형태소들은 배제한다.

6.1 결과

- 겹치지 않는 형태소가 등장하거나, 비율의 차이가 1% 이상인 단어 예시

남성에만 등장하는 형태소	등장 횟수	여성에만 등장하는 형태소	등장 횟수
으나	7	든가	6
라네	5	ㄴ다며	5
거니	3	느니	4
다야	3	다던데	4
라더니	2	느라고	4
구만	2	리	3
ㄴ단	2	어요	3

6.2 결론

- 전체 형태소 개수 약 73000개에서 해당 수치는 너무 미약한 수치이므로 이를 통해서 남녀는 구분해내는 것은 쉽지 않다. 이러한 결과 역시 해당 데이터에서 문장들이 문어체에 가깝게 서술되었기 때문에 이 특징이 나타나지 않는 것으로 추정된다.