본 연구에 사용된 모든 코드는 Github[1]에서 확인 및 결과 재현 가능

자살 유발 정보 디텍팅을 위한 빅데이터 기반 자연어 연구

고귀환 (성균관대학교 인공지능 융합학과 재학생)
E-mail : ie1914@g.skku.edu
Github : https://github.com/GwiHwan-Go

1. https://github.com/GwiHwan-Go/Detect_SC

연구 의도

- 한국 생명 존중 희망 재단에서 자살 유발 정보 모니터링[1]에 관한 활동은 매년 계속해서 이루어 지고 있음.

  → 하지만, **완벽한 모니터링은 불가능한 상황**

- 자살 관련 키워드에 대한 정보도 **매우 한정적**[2]

- 자살, 우울에 관한 이야기가 주로 오가는 곳의 정보를 수집해서 중요한 키워드를 찾아내보자.

  → 더 **효과적인 자살/자해 관련 정보 모니터링**

1. 활동에 관한 자세한 정보는 링크 참고 : https://sims.kfsp.org/usr/main/mainPage.do
2. 다음 페이지 참고.

주로 알려진 자살 관련 정보 키워드[1]

(2) 검색어 입력

#자살
#동반자살
#ㄷㅂㅈㅅ
#ㅂㄱㅌ
#자살계
#죽을래
#투신
#자살계
#자살쇼
#자살각
#자해계
#자해
#자해전시
#죽고싶다
⋮

→ 갖은 노력에도 제한된 키워드때문에,
일부의 자살 유발 정보만을 모니터링 할 수 있음.

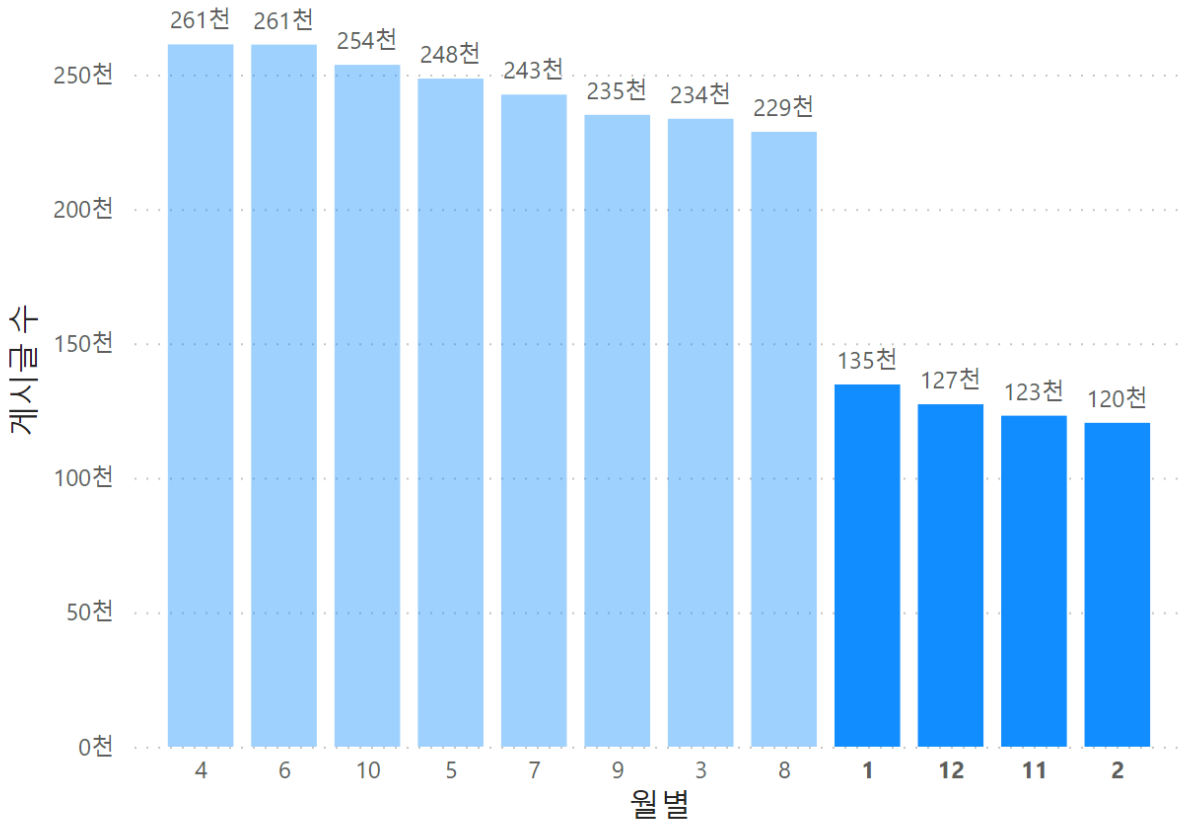
1. 자료 출처 : https://sims.kfsp.org/usr/main/mainPage.do [한국생명존중희망재단의 자살정보 모니터링단 교육자료]

데이터 설명

- 사이트 : 디씨 인사이드(dcinside) 우울증 갤러리[1]
- 수집 기간 : 2021.03.03 ~ 2022.10.27
- 선정 이유 : 자살과 관련된 '우울'이란 감정을 가장 잘 내포하고 있을 것이라 추측.
- 수집 방법 : Python 을 이용한 데이터 크롤링[2]
- 수집된 데이터 규모 : 240M

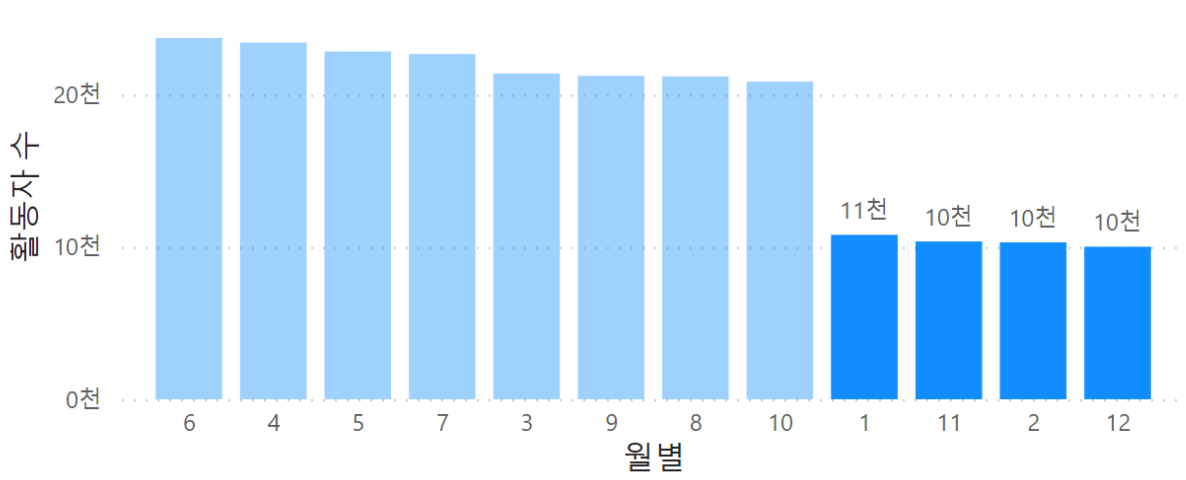
1. https://gall.dcinside.com/board/lists/?id=depression_new1
2. 코드는 깃헙 리파지터리에서 확인가능(https://github.com/GwiHwan-Go/Detect_SC)

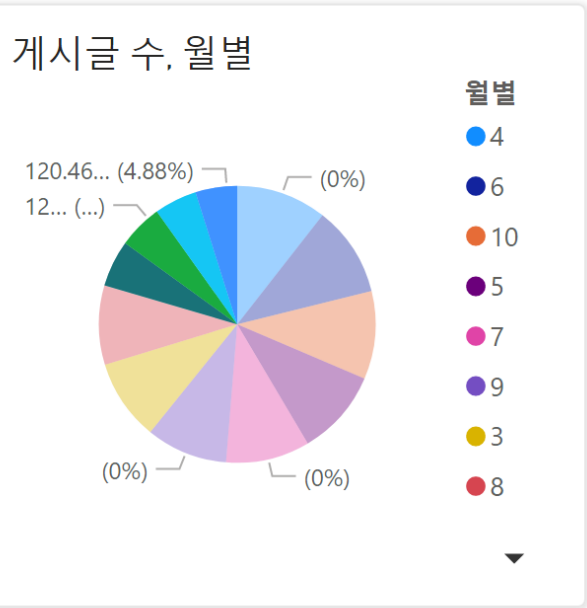
월 별 데이터 분포 시각화

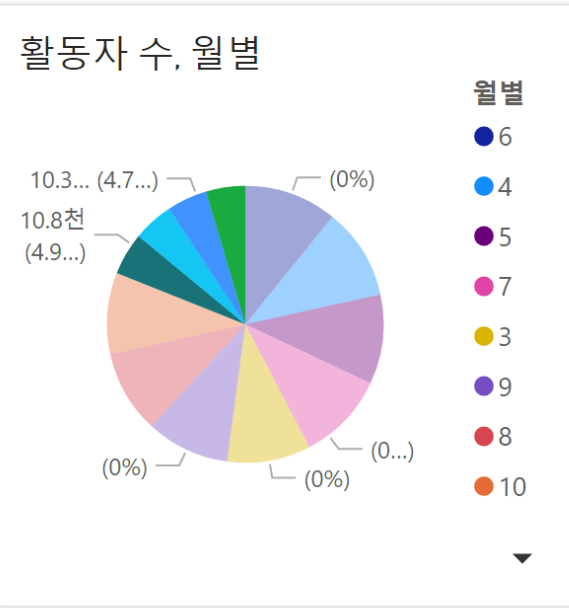


수집된 데이터의 기간 : **21.3.3 ~ 22.10.27**

11,12,1,2 월은 21년도와 22년도 사이에 겹치는 범위가 아니다.
이를 보여주듯, 11,12,1,2월의 데이터는 다른 달의 데이터에 비해 약, 50% 적다.

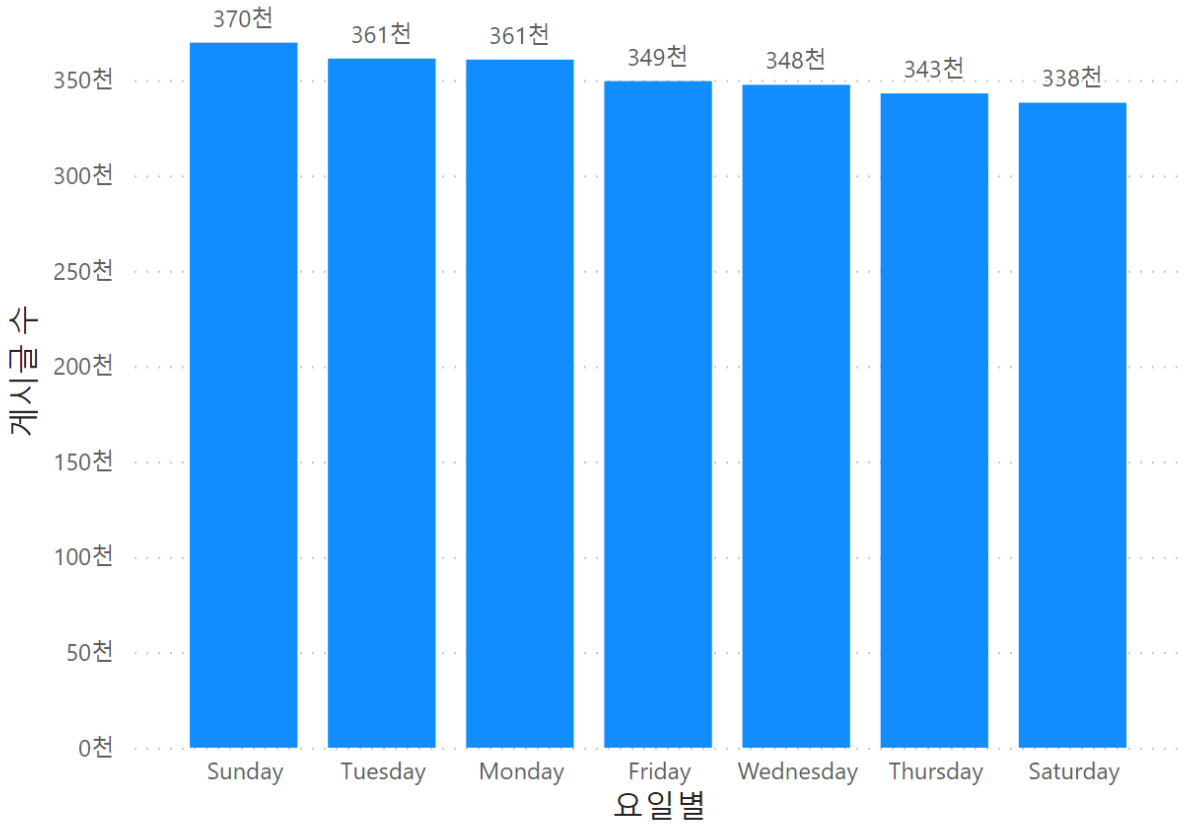






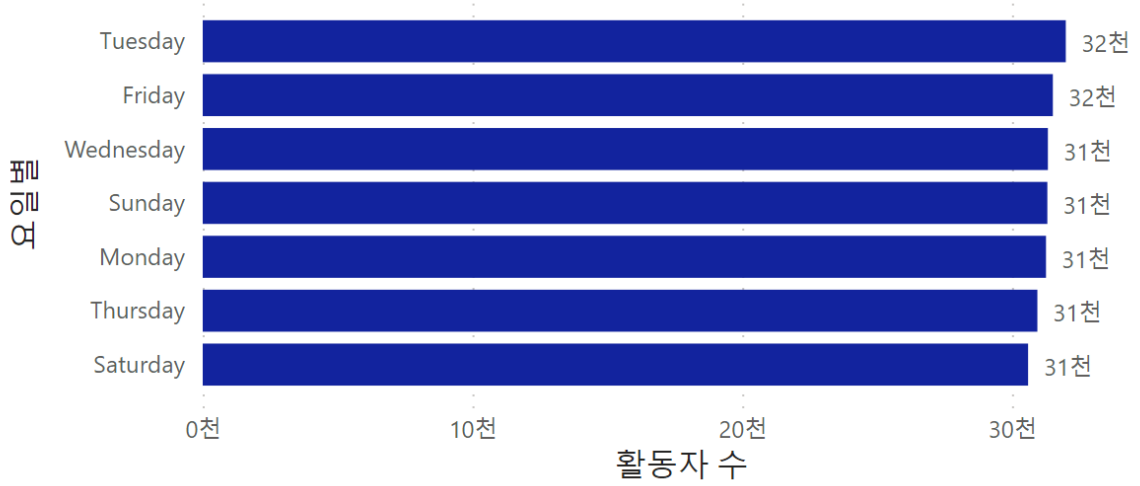
요일별 데이터 분포

게시글 수, 요일별

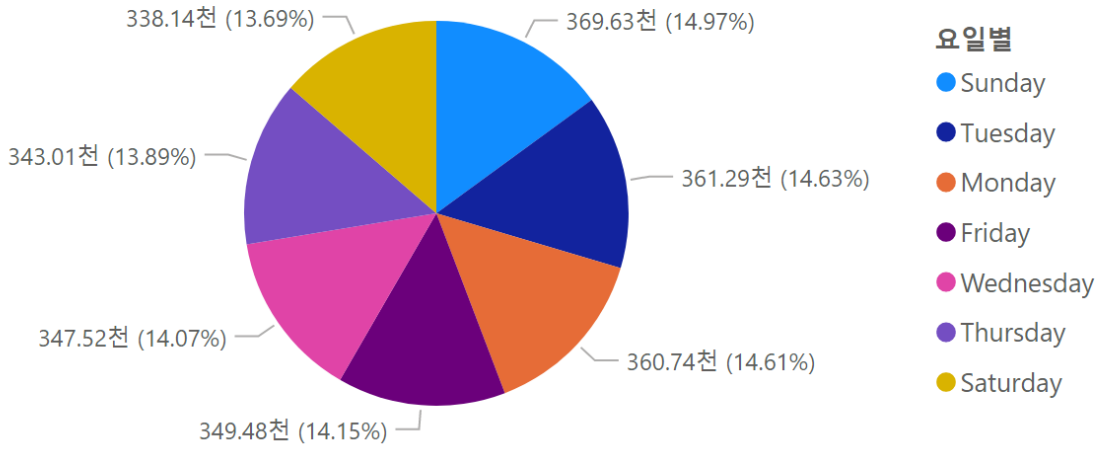


일요일이 다른 요일보다 **게시글의 수가 더 많음**을 알 수 있음.
하지만, **[요일 별 활동자 수]** 그래프를 보면, 일요일의 활동자 수는 많은 편에 속하지 않음.
==> 일요일엔 **특정 유저들의 활동이 활발함**을 알 수 있음.

활동자 수, 요일별



게시글 수, 요일별



데이터 분석 전문 용어 정리

- 단어 정제(preprocessing) : 조사, 숫자, 특수문자 제거
- 토큰화(tokenizing) : 문장을 형태소 혹은 단어로 나누는 행위, 여기서 나눠진 형태소 혹은 단어를 토큰(token)이라 함.
- 신조어 : 사전에 등재되지 않은 단어
- 비지도 학습방식 : 데이터 속에서 컴퓨터가 스스로 규칙을 찾아 학습하게 하는 방법
- 워드 클라우드 : 단어를 빈도수 별로 강조한 시각화 도구

자살 관련 키워드 추출 분석 방법[1]

- **텍스트 전처리**
  - **신조어의 추출을 위해** 사전 기반이 아닌 **SoyNLP 의 비지도 학습방식의 토큰화 도구 이용해서 수집한 데이터를 학습시킴.**
  - **실제 사용 예 : "야이개색야" – 기존 토크나이저 : '야', '이개', '색', '야' | 본 연구 : '야', '이', '개색', '야'**

- **텍스트 분석 방법**
  - Positive Point Mutual Information(PMI) 이용
  - **단어와 문맥의 상관성을 측정하는 방법**

주의 : 본 연구 결과에는 비속어/ 저속어가 다수 포함되어 있습니다.
자료 출처 : https://github.com/lovit/soynlp/blob/master/tutorials/pmi_usage.ipynb

자주 쓰이는 단어들

-정제된(preprocessed) 단어들임.

전체 데이터에 관한 워드클라우드

주의 : 본 연구 결과에는 비속어/ 저속어가 다수 포함되어 있습니다.

연구 결과 설명

- 분석의 시작이 되는 단어는 3페이지의 단어들로 구성함.

- 문맥 단어
  - 특정 단어와 같이 등장한 단어.
  - => 같이 쓰이는 단어들이 어떤 것이 있는지

- 유사 단어
  - 같은 문맥을 가지기 때문에 유사하다고 여겨지는 단어.
  - => 어떤 단어들이 대체해서 쓰이는 지, 같은 맥락에서 쓰이는지

- 시각화 된 결과를 곱씹어 볼 수록 많은 것을 느낄 수 있을 것이다.

주의 : 본 연구 결과에는 비속어/ 저속어가 다수 포함되어 있습니다.

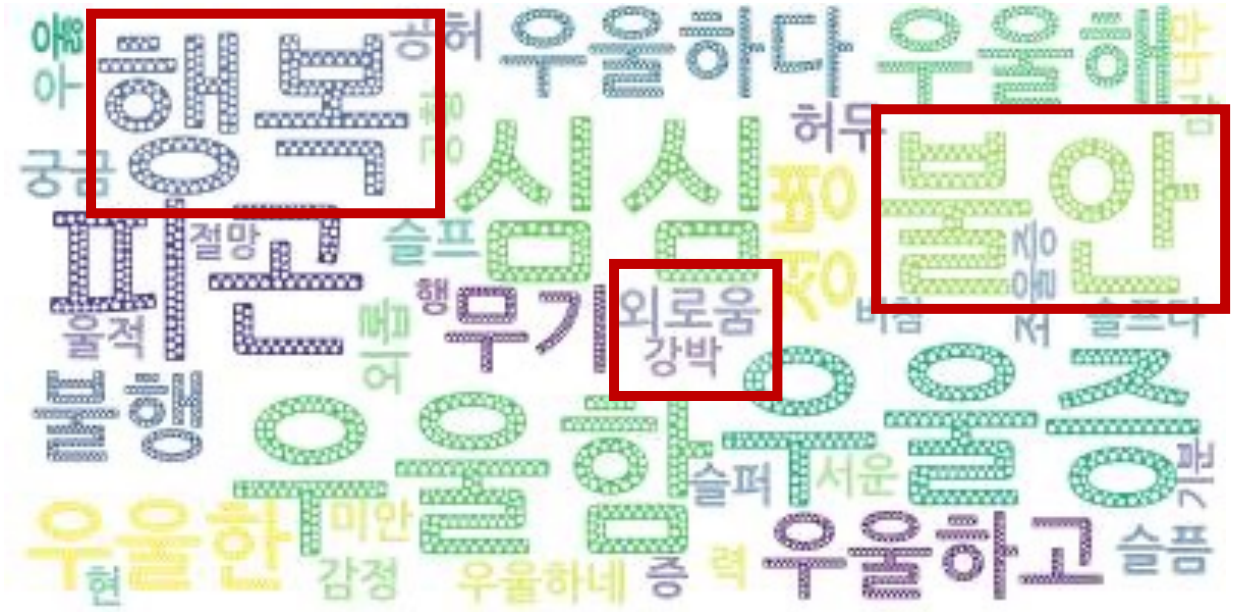
동반: 문맥 단어⬆️ 유사 단어⬇️

동반은 주로 '입대'라는 단어와 연관을 보이며,
'자살'과는 큰상관을 보이지 않는다.

자살 : 문맥 단어⬆️ 유사 단어⬇️

'자살' 이란 단어는 '살고싶다'라는 단어와 연관이 크다
사실 그 누구보다 살고 싶어하는 사람들인 것이다.

우울: 문맥 단어⬆️ 유사 단어⬇️



살자 : 문맥 단어⬆️ 유사 단어⬇️

자해: 문맥 단어⬆️ 유사 단어⬇️

자살 관련 키워드가 검열 된다는 것은 공공연히 잘 알려진 사실이다.
따라서 커뮤니티 회원들은 이를 변형 시켜서 사용한다.
본 분석은 커뮤니티 회원들이 변형 시켜서 사용하는 단어들을 성공적으로 추적하였다. [좃해, 자혜, ㅈㅎ, 죳해]

인생 : 문맥 단어⬆️ 유사 단어⬇️

같이 쓰인 단어가 '인생 ' 이란 것을 감안할 때
많은 것을 느끼게 해준다.

죽고싶다: 문맥 단어⬆️ 유사 단어⬇️

죽을래 : 문맥 단어⬆️ 유사 단어⬇️

죽고싶다의 다른 표현으로
'버릴래', '죽구싶다' 도 사용되고 있다.
비표준어는 모니터링에 가장 방해가 되는 요소 중 하나이다.
본 분석은 비표준어 문제를 잘 해결하고 있다.

죽음: 문맥 단어⬆️ 유사 단어⬇️

죽음 + '의끝' 은 의미심장한 단어로, 주의 관찰할 필요가 있다.
'편안함', '행복' 등이 '죽음'과 같이 쓰이면,
마찬가지로 주의 관찰할 필요가 있다.

투신 : 문맥 단어⬆️ 유사 단어⬇️

투신 + 'ㅈㅏㅅㅏ' 의 형태로 커뮤니티 회원들이 사용하고 있는 것을 알 수 있다.
본 분석을 통해서 커뮤니티 회원들이 주로 사용하는 검열 회피 방식을 추적할 수 있다.

워드 클라우드 결과 종합

- '자살, 죽음, 자해' 등이 '편안, 행복, 살고 싶다.' 등의 긍정적 단어와 같이 등장한다.
  - 모니터링에서 '살고 싶다', '편안', '행복' 등의 키워드도 고려해볼 필요가 있다.
  - 죽을 생각이 없는 사람은 '살고 싶다.' 라는 생각을 하지 않기 때문이다.

- 자동 검열을 회피하기 위해서 사용되는 비표준 단어 추적

  ⇒자해 대신 [좃해, 쟈혜] , 자살 대신 'ㅈㅏㅅㅏ', 죽고싶다 대신 '죽구싶다' 등의 의도적인 비표준어 사용 을 성공적으로 추적.

  ⇒연구결과에 등장한 상식적으로 이해가 되지 않는 단어들 역시 특정 단어의 변형 형태일 가능성이 있다.

Graph 데이터 설명

- Graph 생성 알고리즘

```
until recursion boundary :
  Seed -> Context_word_list
    for Context_word in Context_word_list
      build graph
```

- 그래프 전문 용어
  - 페이지 랭크[1] : 중요도를 표시하는 방식
  - 노드 : 단어를 가리킴
  - 엣지 : 단어와 단어 간의 연결선

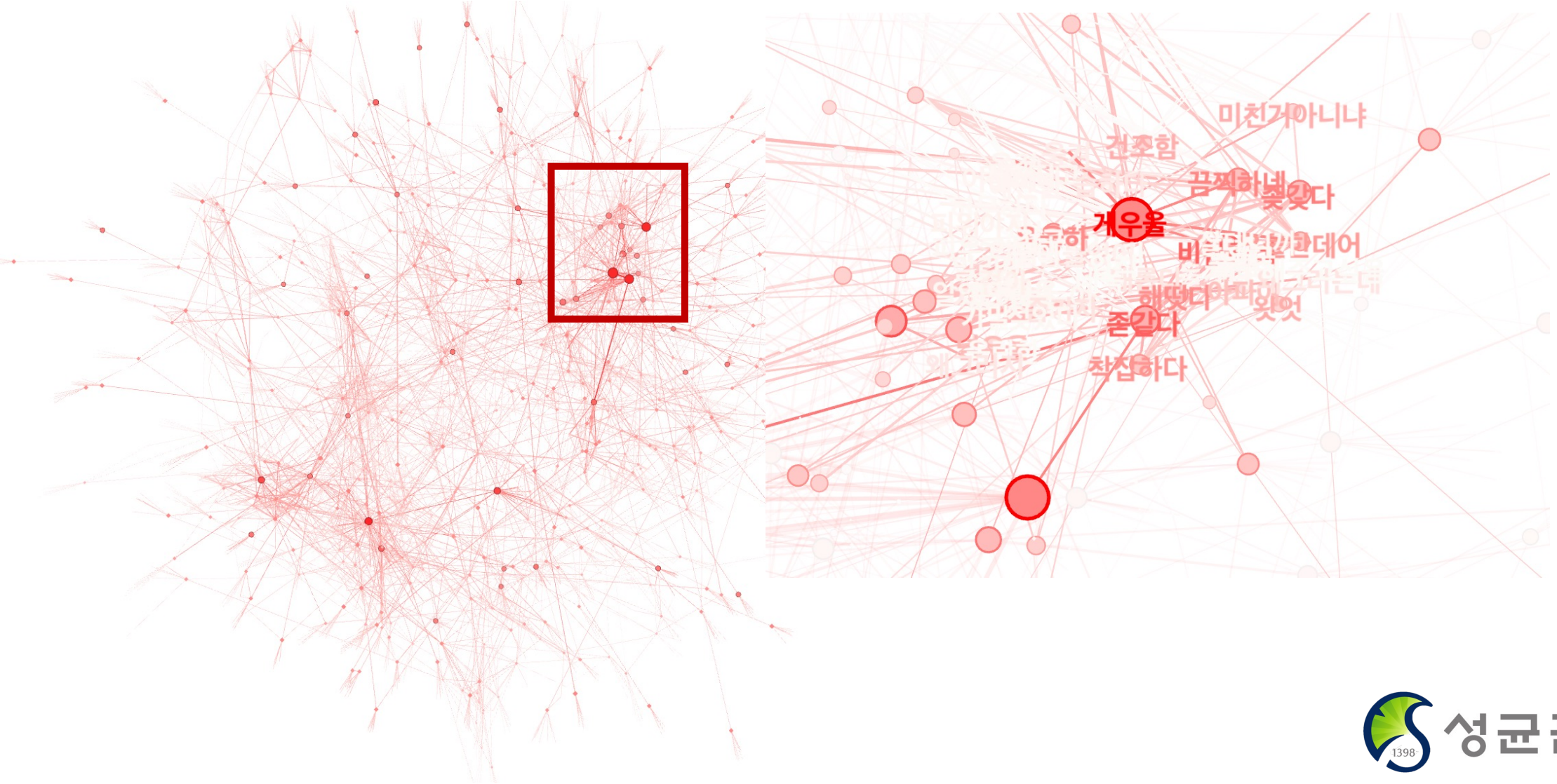
성균관대학교

1. http://infolab.stanford.edu/~backrub/google.html

그래프 데이터 설명

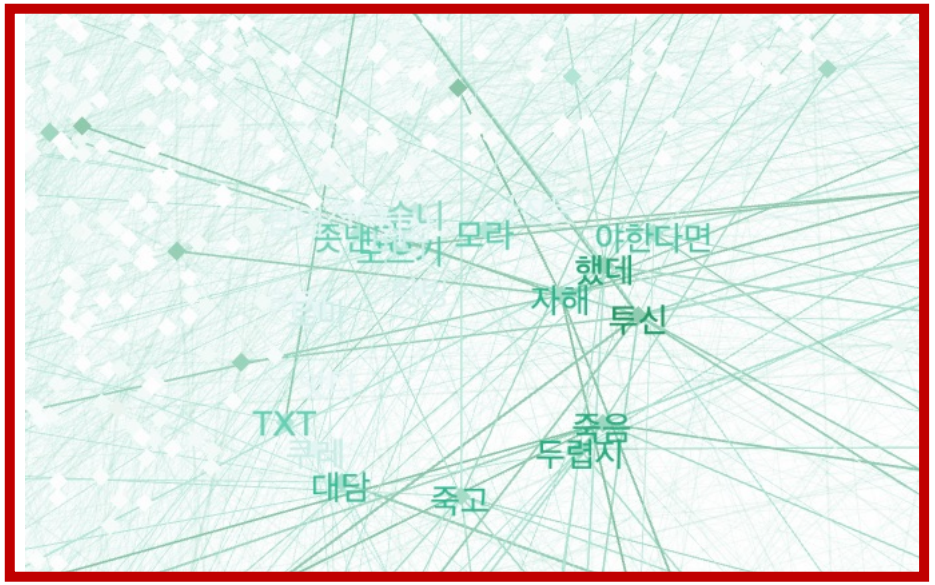
- 문맥 단어 그래프
  - 단어와 문맥 단어들 간을 연결한 그래프
- 유사 단어 그래프
  - 단어와 유사 단어들 간을 연결한 그래프
- 각, 그래프당 4만 여개의 노드와 20만 여개의 엣지가 존재.
- 효과적 분석을 위해서 페이지 랭크를 기준으로 필터링 수행.

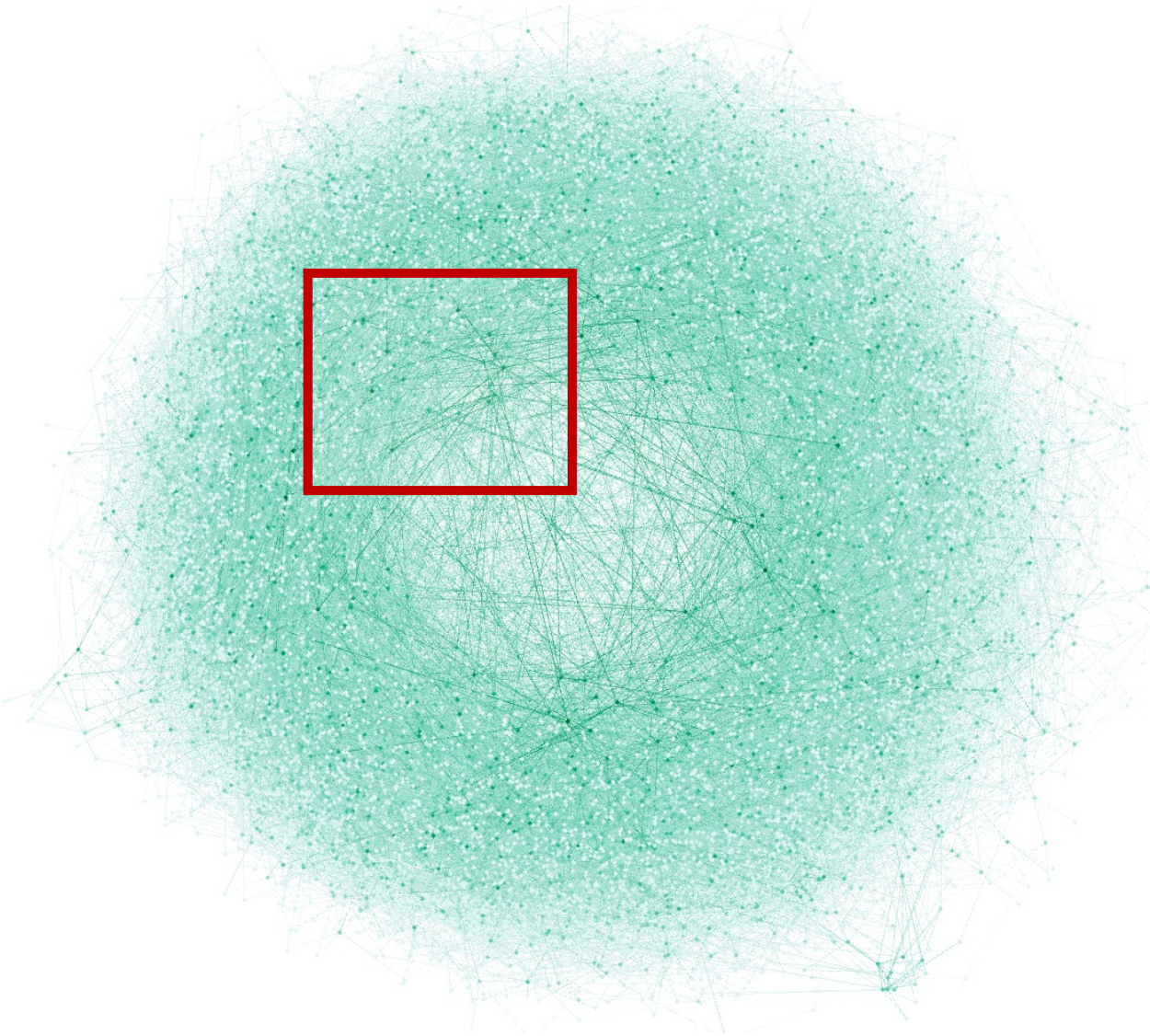
데이터 오버뷰 (유사 단어 그래프)



성균관대학교

데이터 오버뷰 (문맥 단어 그래프)





그래프 데이터 활용 방안

1. 유사 단어 그래프와 문맥 단어 그래프를 조합

2. 아래의 테스크를 수행하는 분류기 빌드

    - 특정 게시물, 혹은 작성자의 우울 지수 평가

    - 자살 관련 정보 분류기

Unique Origin Unique Future

자세한 코드는 깃헙에서 확인 및 연구결과 재현 가능
https://github.com/GwiHwan-Go/Detect_SC