


데이터 소개

- Kaggle에서 수집한 Human Stress Prediction 데이터셋
- 정신건강과 관련된 subreddit에 올라온 2838개의 게시물 
- 0은 스트레스가 없음 / 1은 스트레스가 있음을 의미
- 총 7개의 컬럼으로 구성
- 사람들의 게시물에는 심리적 스트레스를 받고 있는지 여부를 보여줄 수 있는 단어 존재

| | subreddit | post_id | sentence_range | text | label | confidence | social_timestamp |
|------|------------------|---------|----------------|---|-------|------------|---------------------|
| 0 | ptsd | 8601tu | (15, 20) | He said he had not felt that way before, sugge... | 1 | 0.800000 | 2018-03-21 15:39:13 |
| 1 | assistance | 8lbrx9 | (0, 5) | Hey there r/assistance, Not sure if this is th... | 0 | 1.000000 | 2018-05-23 02:23:37 |
| 2 | ptsd | 9ch1zh | (15, 20) | My mom then hit me with the newspaper and it s... | 1 | 0.800000 | 2018-09-03 09:46:45 |
| 3 | relationships | 7rorpp | [5, 10] | until i met my new boyfriend, he is amazing, h... | 1 | 0.600000 | 2018-01-20 15:25:55 |
| 4 | survivorsofabuse | 9p2gbc | [0, 5] | October is Domestic Violence Awareness Month a... | 1 | 0.800000 | 2018-10-18 05:43:25 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 2833 | relationships | 7oe1t | [35, 40] | * Her, a week ago: Precious, how are you? (I i... | 0 | 1.000000 | 2018-01-06 06:17:24 |
| 2834 | ptsd | 9p4ung | [20, 25] | I don't have the ability to cope with it anymo... | 1 | 1.000000 | 2018-10-18 10:50:12 |
| 2835 | anxiety | 9nam6l | (5, 10) | In case this is the first time you're reading ... | 0 | 1.000000 | 2018-10-11 23:48:32 |
| 2836 | almosthomeless | 5y53ya | [5, 10] | Do you find this normal? They have a good rela... | 0 | 0.571429 | 2017-03-08 10:55:43 |
| 2837 | ptsd | 5y25cl | [0, 5] | I was talking to my mom this morning and she s... | 1 | 0.571429 | 2017-03-08 02:58:36 |

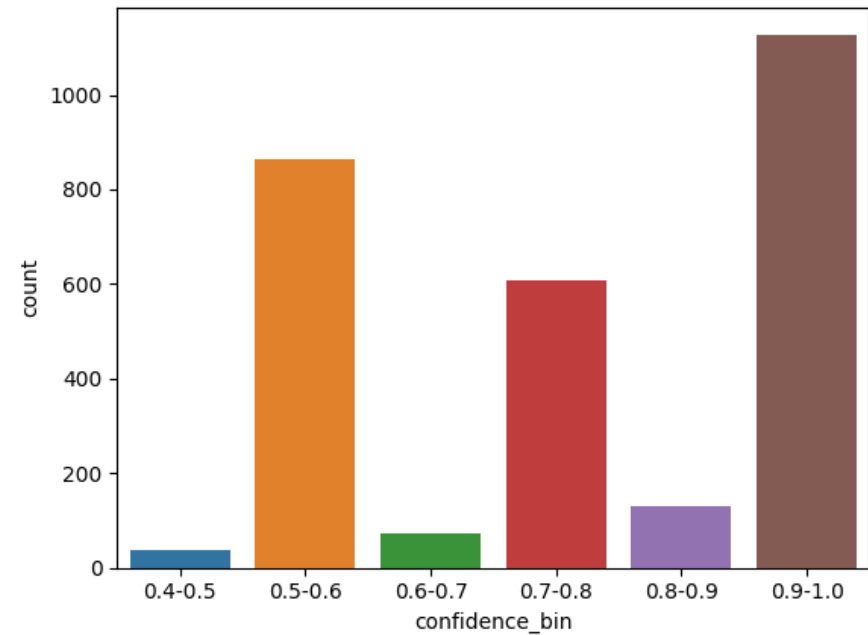
데이터 EDA

- 데이터 null 값 확인

| | |
|------------------|---|
| subreddit | 0 |
| post_id | 0 |
| sentence_range | 0 |
| text | 0 |
| label | 0 |
| confidence | 0 |
| social_timestamp | 0 |

Null값 존재하지 않음

- 게시물 신뢰도 값 분포 확인

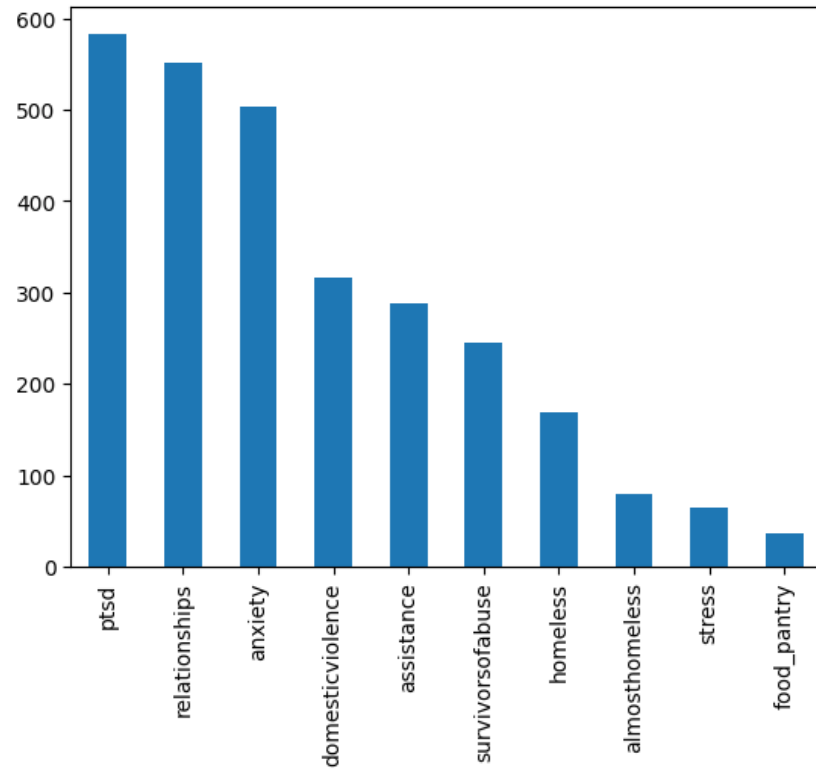


신뢰도가 0.5 이상인 글들만 분석에 사용

총 2836개의 게시물 사용

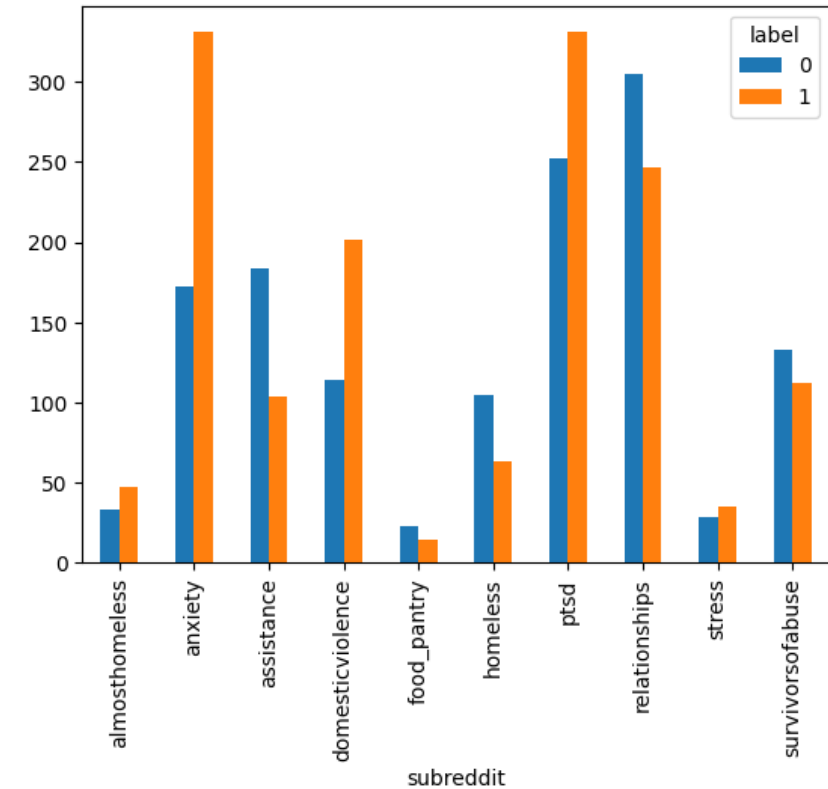
데이터 EDA

- Subreddit 별 text 개수 확인



전체 데이터 중 57%가
PTSD, Relationship, anxiety 게시판에 존재

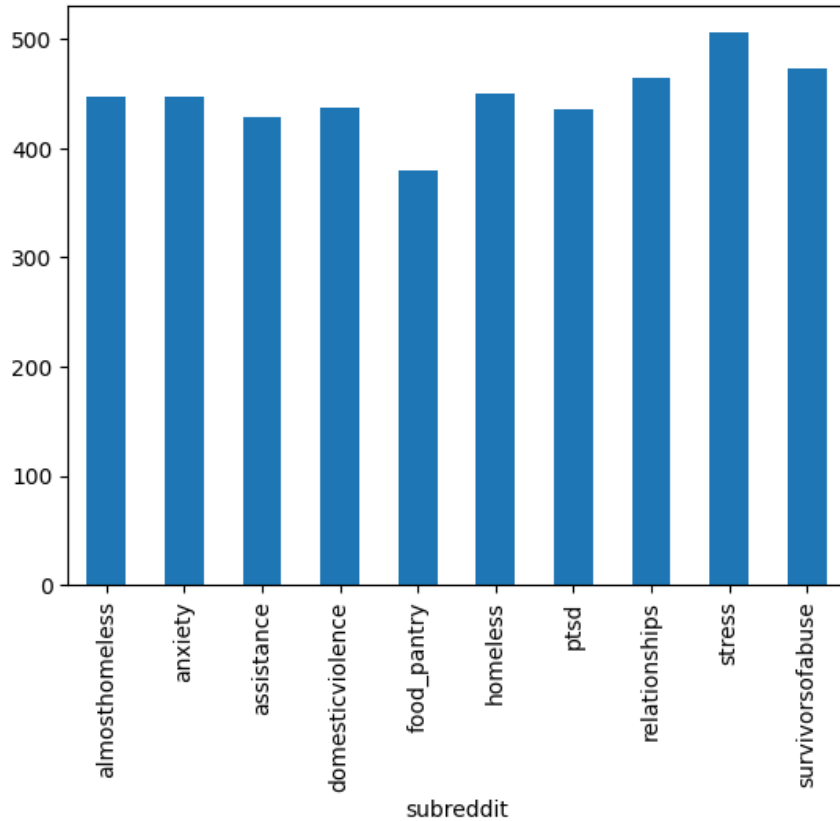
- Subreddit 별 stress 여부



Anxiety, domestic violence, ptsd 게시판에서
스트레스가 더 많은 것으로 확인

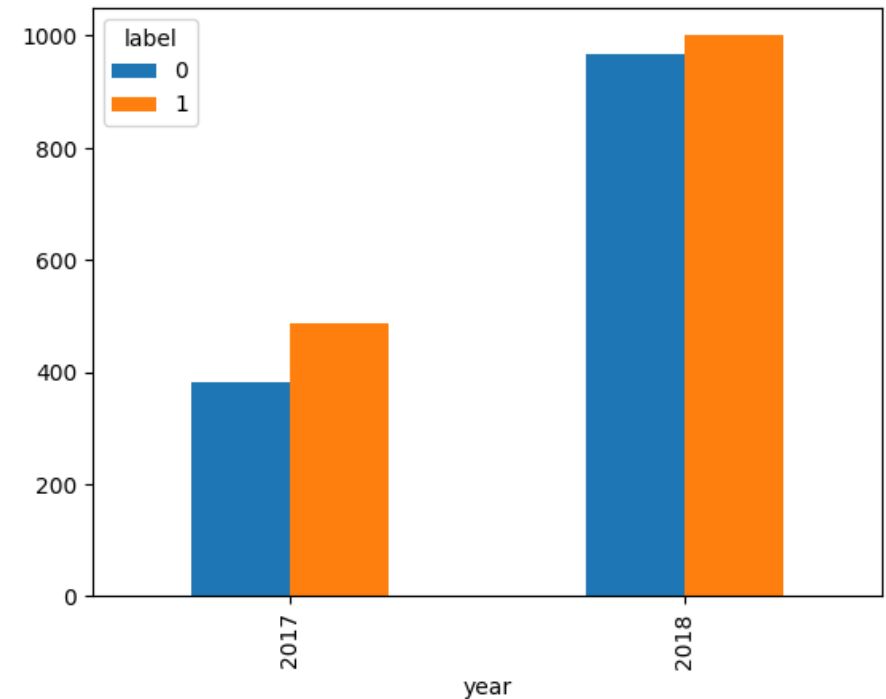
데이터 EDA

- Subreddit 별 text 평균 길이 확인



Food_pantry 게시판이 379로 제일 짧고,
stress 게시판이 505로 제일 길었지만
대체적으로 대부분의 게시판의 text 길이가 비슷한 편

- 연도별 게시물 수 확인



2017년 868개 - 0 : 382 / 1 : 486
2018년 1968개 - 0 : 968 / 1 : 1000

데이터 분석 - 전처리 전

- 불용어 제거, 소문자 변환, 기호 제거 등 하지 않은 상태에서 분석
- 텍스트 최대 길이 : 1639, 텍스트의 평균 길이 : 447.9492

1) 로지스틱 회귀

- Train - Test set 8:2 분리
학습 데이터 개수 : 2268
테스트 데이터 개수 : 568
- CounterVectorizer 사용

Precision : 0.70, Recall : 0.78
F1-score : 0.74, Accuracy : 0.72

2) BERT (bert-base-uncased 사용)

- Train - Test set 7:3 분리
학습 데이터 개수 : 1985
테스트 데이터 개수 : 851
- MAX_LEN = 450, optimizer=AdamW, epochs=4

Accuracy : 0.79

데이터 분석 - 전처리 전

- 로지스틱 회귀 모델을 이용해 스트레스 여부 판단에 영향을 미치는 상위 20개 단어 분석

1) 스트레스 있는 글 단어 상위 20개

```
tell (0.384)
hate (0.378)
fucking (0.363)
feel (0.320)
even (0.316)
do (0.269)
me (0.257)
past (0.247)
don (0.247)
no (0.239)
scared (0.237)
anxiety (0.235)
sick (0.228)
what (0.225)
just (0.221)
nothing (0.218)
myself (0.213)
because (0.205)
am (0.176)
why (0.167)
```

2) 스트레스 없는 글 단어 상위 20개

```
great (-0.118)
they (-0.121)
would (-0.125)
post (-0.125)
person (-0.126)
first (-0.139)
for (-0.141)
that (-0.158)
we (-0.173)
be (-0.179)
your (-0.180)
bit (-0.181)
her (-0.186)
good (-0.187)
let (-0.196)
others (-0.264)
you (-0.309)
finally (-0.325)
url (-0.486)
met (-0.534)
```

데이터 분석 - 전처리 후

- 불용어 제거, 소문자 변환, 기호 제거, 어간 추출, 원형 추출 등 전처리 진행
- 텍스트 최대 길이 : 882, 텍스트의 평균 길이 : 223.9495

1) 로지스틱 회귀

- Train - Test set 8:2 분리
학습 데이터 개수 : 2268
테스트 데이터 개수 : 568
- CounterVectorizer 사용

Precision : 0.69, Recall : 0.72
F1-score : 0.70, Accuracy : 0.69

2) BERT (bert-base-uncased 사용)

- Train - Test set 7:3 분리
학습 데이터 개수 : 1985
테스트 데이터 개수 : 851
- MAX_LEN = 250, optimizer=AdamW, epochs=4

Accuracy : 0.75

데이터 분석 - 전처리 후

- 로지스틱 회귀 모델을 이용해 스트레스 여부 판단에 영향을 미치는 상위 20개 단어 분석

1) 스트레스 있는 글 단어 상위 20개

```
scare (0.568)
hate (0.555)
fuck (0.443)
feel (0.437)
sick (0.437)
even (0.387)
cri (0.357)
tell (0.357)
anxieti (0.293)
know (0.262)
attack (0.261)
happen (0.238)
noth (0.235)
get (0.228)
abus (0.222)
trigger (0.188)
stop (0.188)
past (0.186)
dr (0.185)
school (0.182)
```

2) 스트레스 없는 글 단어 상위 20개

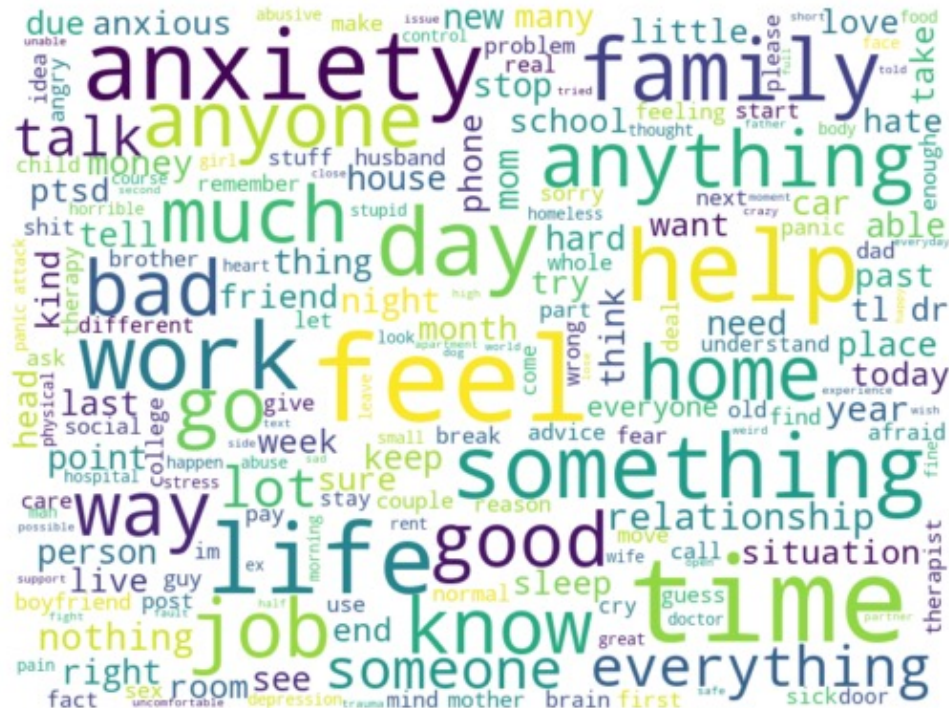
```
share (-0.117)
see (-0.120)
studi (-0.129)
person (-0.165)
way (-0.180)
support (-0.200)
togeth (-0.203)
would (-0.205)
good (-0.206)
guy (-0.225)
first (-0.235)
final (-0.250)
love (-0.263)
post (-0.267)
survey (-0.301)
experi (-0.317)
other (-0.319)
thank (-0.471)
met (-0.628)
url (-0.675)
```


데이터 분석 - 전처리 후

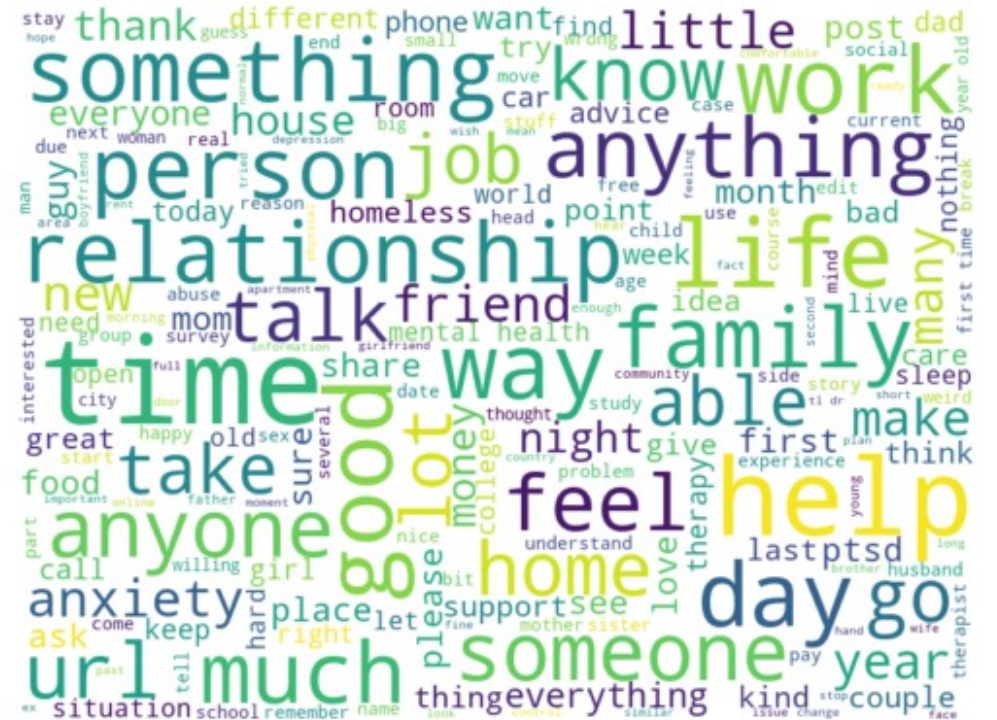
- 스트레스가 있는 집단과 없는 집단의 text 단어 빈도 분석

1) 토큰화 2) 품사 태깅 3) 명사, 형용사, 동사 추출 / 형용사 추출
4) 원형 찾기 5) 불용어 제거 6) 워드클라우드 시각화

1. 스트레스 있는 집단(명사, 형용사, 동사)



2. 스트레스 없는 집단(명사, 형용사, 동사)

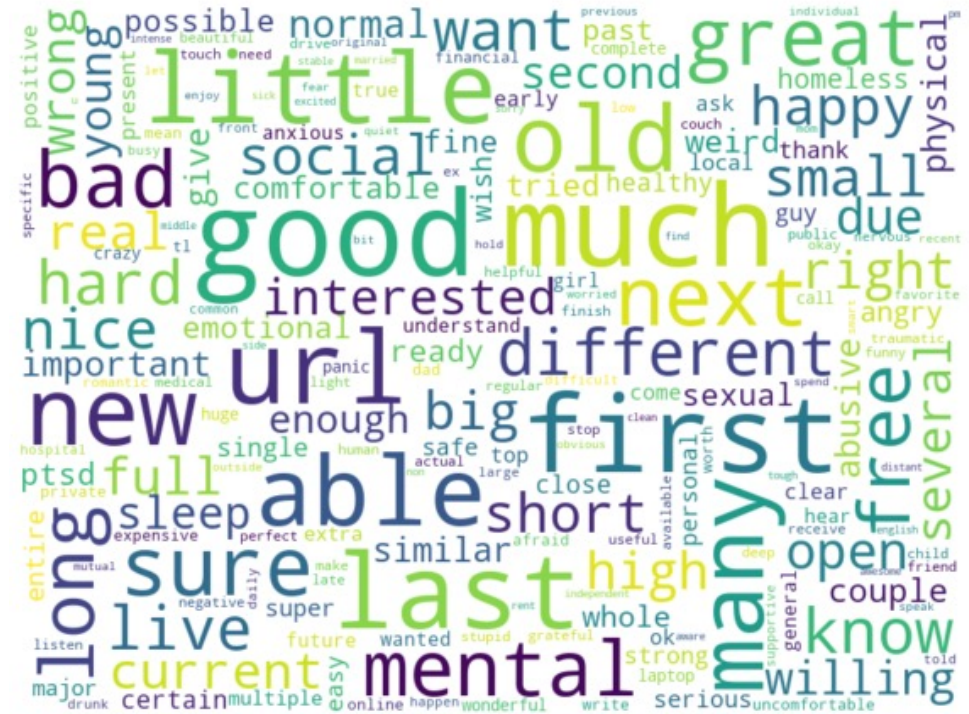


데이터 분석 - 전처리 후

1. 스트레스 있는 집단(형용사)



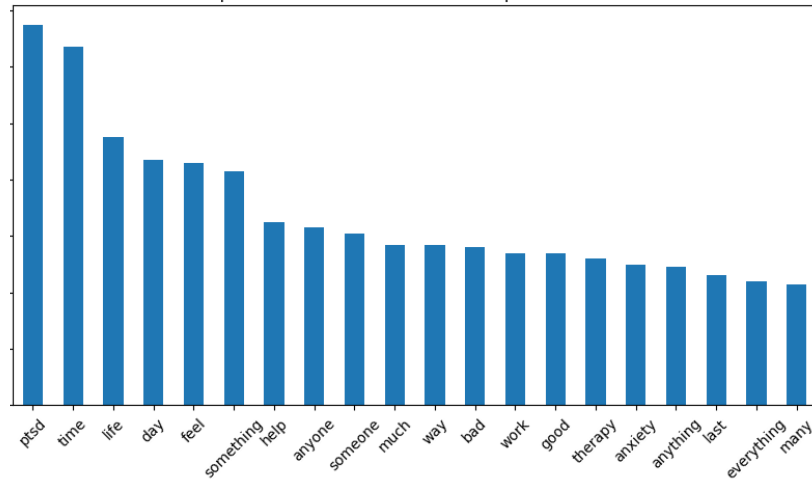
2. 스트레스 없는 집단(형용사)



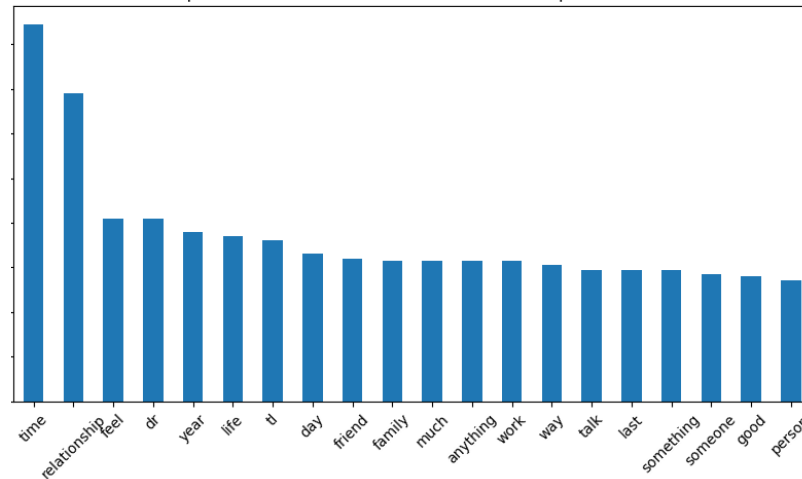
데이터 분석 - 전처리 후

- Subreddit 중 게시물이 가장 많았던 ptsd, relationships, anxiety 게시판의 text 단어 빈도 분석

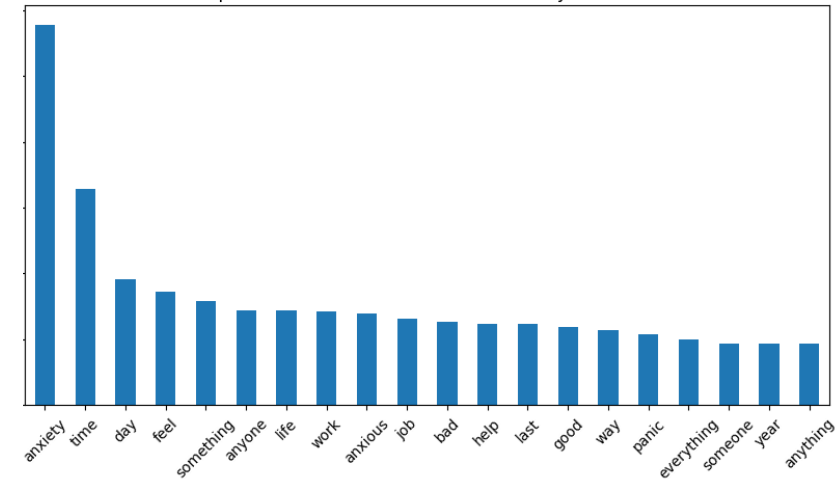
Top 20 most common words in the ptsd subreddit



Top 20 most common words in the relationship subreddit



Top 20 most common words in the anxiety subreddit



- 각 게시판마다 게시판 이름(ptsd, relationships, anxiety) 단어가 상위권에 존재
- time, day, feel, bad, good 이란 단어는 모든 게시판마다 공통적으로 등장
- relationship 게시판은 family, friend, talk, person과 같은 사람과 관련된 단어 많이 등장
- ptsd, anxiety 게시판에서는 anxious, anxiety와 같은 심리적 표현 단어 등장