فاز اول پروژه درس مبانی پردازش زبان و گفتار

یگانه مرشدزاده

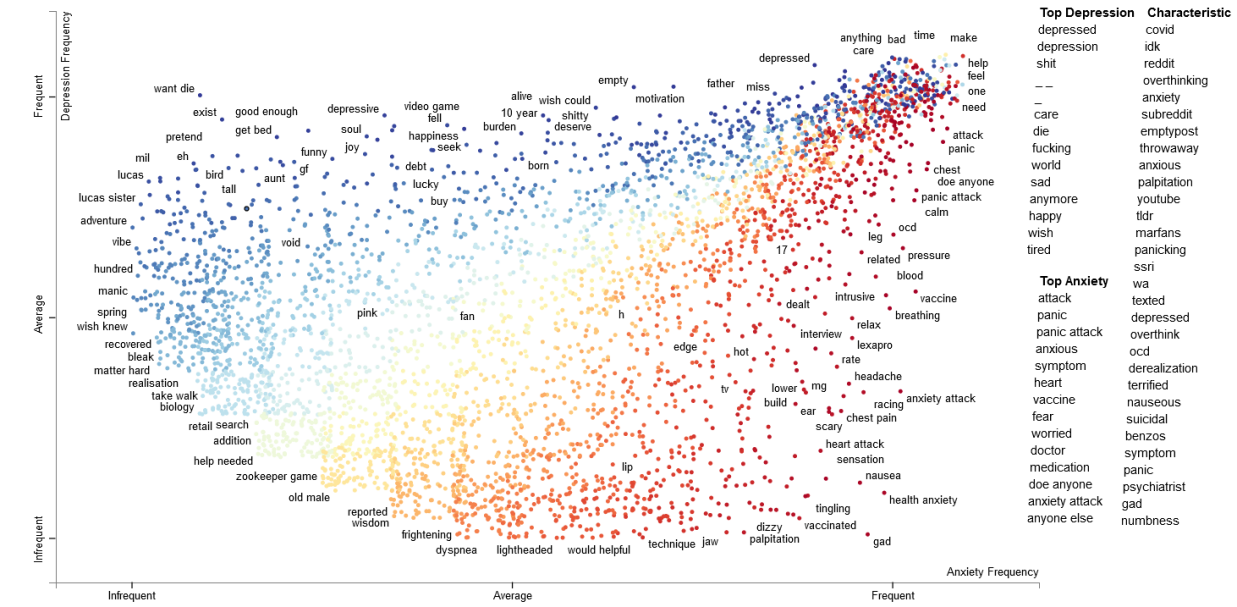
اردیبهشت ۱۴۰۰

موضوع داده

رابطه‌ی بین اضطراب و افسردگی با توجه به عبارات بیان شده توسط کاربران Reddit.

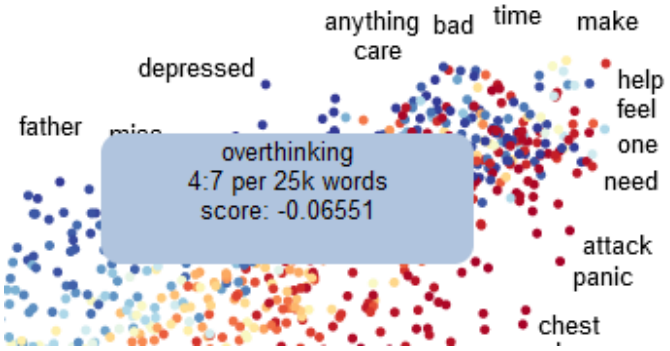
دو دسته کلی مورد بررسی اضطراب و افسردگی است.

نمای کلی آنچه در آخر بدست آمد:



در آخرین سلول در فایل 2_data_cleaning_preprocessing_EDA.ipynb این نمودار وجود دارد که بسیار مفید است.

برای عبارت overthinking:



Top Depression	Characteristic
depressed	covid
depression	idk
shit	reddit
_ _	**overthinking**
_	anxiety
care	subreddit
die	emptypost
fucking	throwaway
world	anxious

Term: overthinking

Depression frequency:
4 per 25,000 terms
12 per 1,000 docs
Some of the 13 mentions:

Depression
wise detective idk happens everywhere live stray cat dog often go various house looking food shelter bad enough wander around first place even kind people give food water look shoulder every 2 fucking second jump away slightest sound hungry thirsty god know long ready give food fright might happen stay shit make anyone blink eye fact absolutely heartbreaking tell someone hear animal like care **overthinking** talking real problem overthinking half time people like made feel bad opening mouth even wa talking real stuff award making start questioning basic decency humanity go 4 legged friend look shoulder fearing u overthinking

Depression
ip hate regretting thing know stop thinking stuff know pointless think help thought back head usually forget stuff within day recently struggling stop ruminating **overthinking** past

Depression
coffee 35 people tell control brain control brain doe stop speaking control brain doe stop humiliating slightest imperfect thing control brain suffocates point breathe control brain hit hardest weak scared crumbled floor cry **overthinking** joke shit eats alive

Depression
1 lk bog depressed 3ish year gotten worse worse tried killing 4 different time thinking getting tired taking lot med anything make go bed give massive headache overthink everything point think people think walk **overthinking** stop processing people want process worried use issue starting kill pretty much trust one parent hold super high standard meet make feel like let compared family member super antisocial struggle even starting conversation get friend 2 people talk usually unavailable shit keep getting worse really want deal anymore already plan stuff scared tell people terrified could happen know really need help

Depression
wu ance moore hi venting subreddit since early year last month wa refreshing felt alive every 2 month feel overwhelming sexual frustration making depressed episode increasing every year suggestion tackle go therapy chance become much problem life kind phase proper approach fight thanking advance advice took lot courage write something sensitive private people usually take

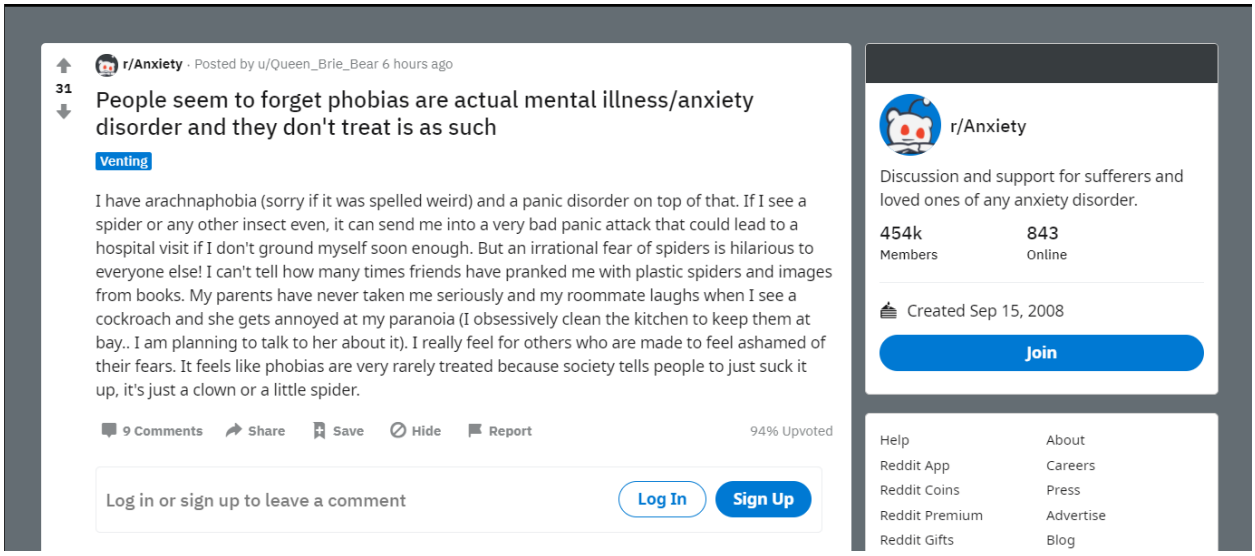
Anxiety frequency:
7 per 25,000 terms
25 per 1,000 docs
Some of the 26 mentions:

Anxiety
big dirt guy month **overthinking** existence meaningless life grand scheme thing sorta like feeling like simulation kinda thing trying keep level head seeing psychiatrist soon seem figure stem hard time getting enjoyment thing normally love still like listening music kinda calm make feel real grounded reality second keep getting anxious scared thing reality matter like watching movie make anxious reason start thinking show real matter simply keep intrusive negative thought normal thing normally pretty chill person considered stem super stressed full time job 21 move back parent house pandemic feel extreme anyone similar experience advice time anxiety existence

Anxiety
name sq j feeling impending doom death anxiety amp like certain time mind freak start getting deja vu feel like impending doom derealization put together dont know make stop nothing really come randomly tomorrow birthday get thought head gonna die 12 00 probably sound stupid genuinely know anymore getting chest pain leg amp foot start cramping anxiety ha literally taken life tell anyone scared three invalidate problem bad people think looking attention know feel point today woke feeling really shitty afternoon started getting intrusive suicidal thought anything usually self harm something irrational angry felt productive got room clean amp anxiety feel like skyrocket felt like sugar wa high diabetic felt shitty took way much insulin probably gonna fucked later tonight amp everything freaking like numb used feeling honestly sick want live life without **overthinking** analyzing anxiety derealization intrusive thought health problem amp everything else used spiritual person anxiety come religious ocd kid wa jehovas witness seeing lot christian around everything really confusing scared make wrong choice yk mean started overanalyze started thinking every thought head wa judged god thought mine wa someone communicating ik sound insane wanna back head 24 7 self conscious get really paranoid public impending doom thing try fight consumes complete fear thought take know know anymore everything changing worst night pain add insult injury breath start thinning start getting scared heart stopping heart attack stop breathing always worse case scenario getting really tough handle told mom tried comfort rely every time plus kind embarrassing ask help problem since wa elementary school start spiraling feel like real world think tragedy death really really fuck head carry day noticed getting violent violent thought feel disconnected everything family friend world lost motivation interest thing used love like gaming basketball working pretty much heavily addicted porn past year normal slowly

1. منبع دقیق داده طوری که بازیابی آن با روش مشابه برای یک محقق دیگر قابل انجام و راست آزمایی باشد.

داده‌ی جمع‌آوری شده دارای دو دسته‌ی اضطراب و افسردگی است که به ترتیب از دو community در reddit با آدرس reddit.com/r/Anxiety و reddit.com/r/depression استخراج شده‌اند.

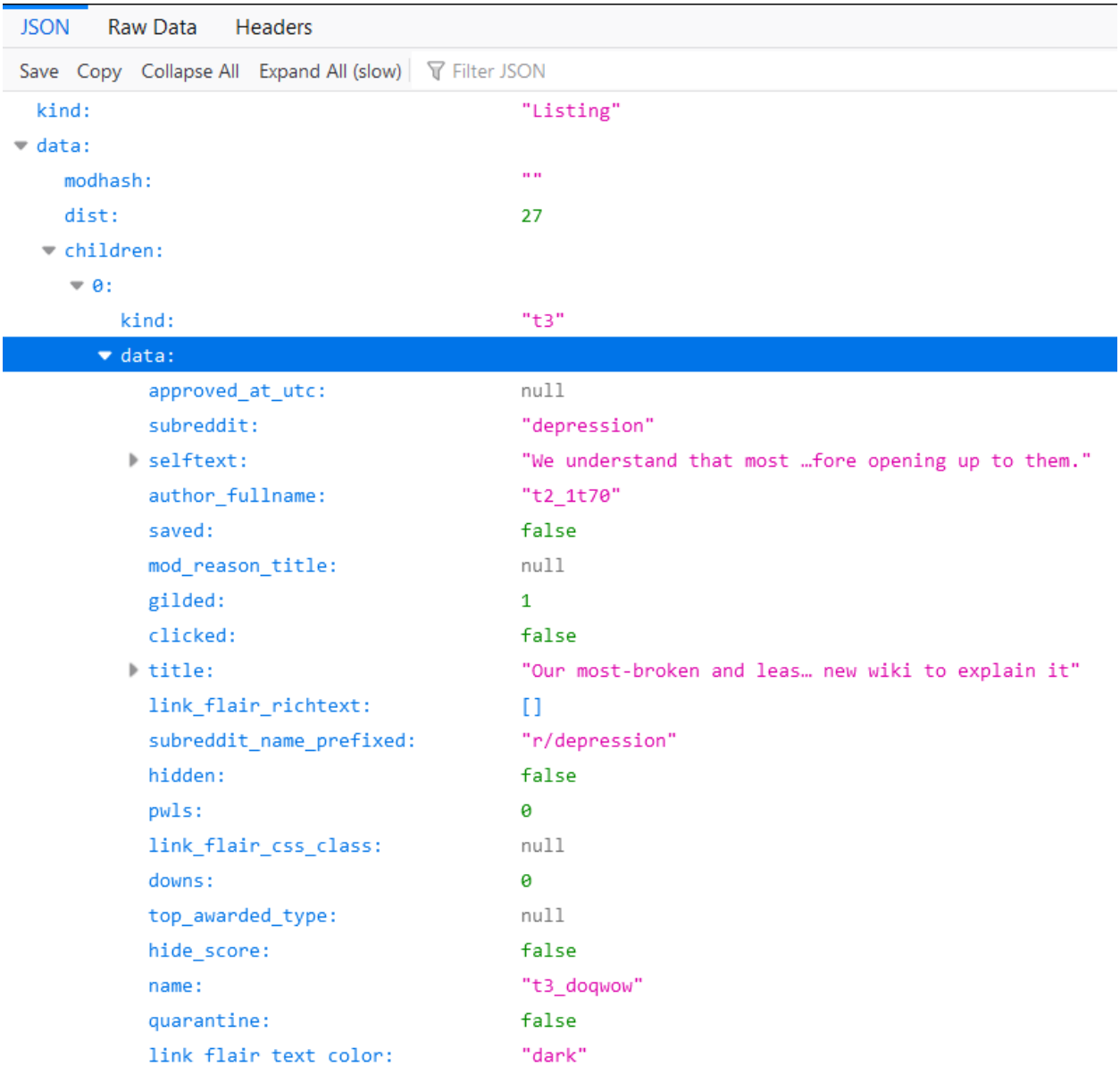


2. روش جمع آوری، مراحل و ابزارهای استفاده شده برای جمع آوری داده.

پست های دو subreddits با نام های r/Anxiety و r/depression با استفاده از Reddit's API جمع‌آوری می‌شوند.

هنگام جمع آوری داده ها از سرورها، یک تأخیر تصادفی بین درخواست ها به عنوان ملاحظه‌ای به سرورهای Reddit و کارکنان امنیتی ایجاد می‌شود.

داده‌ها به صورت فایل JSON جمع‌آوری می‌شود.

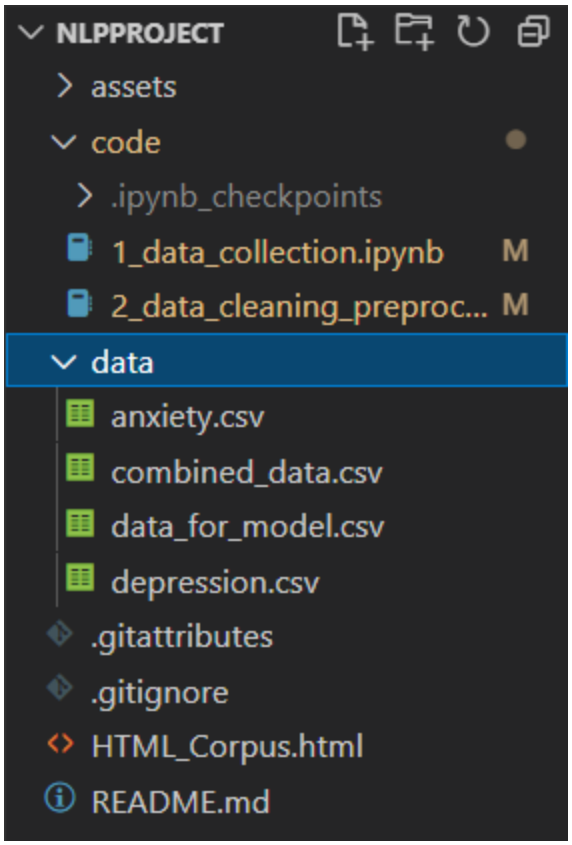


داده ها به صورت دیکشنری هایی ساماندهی شده اند.

با استفاده از آخرین پست و فیلد after آن به عنوان QUERY STRING که نشان دهنده کلید پست بعدی است، یک صفحه از پست های reddit را scrape میکنیم. در پایان هر بار لحظاتی را مکث می‌کنیم تا ملاحظه سرورهای Reddit را کرده باشیم.

لازم به ذکر است برای اینکه دریافت تعداد scrap ها به درستی و کامل صورت بگیرد نیاز است از VPN استفاده شود.

3. فرمت داده ها (فایل و ساختار پوشه). ساختار هر فایل به چه صورت است و برچسبهای مختلف چگونه از هم متمایز هستند.



داده ها را میتوان به dataframe تبدیل کرد و به راحتی بر روی آن ها محاسبات را انجام داد.

داده بدست آمده از depression حدود ۱۰۳ ستون دارد و ما نیز جهت راحتی تشخیص label ها یک ستون به نام is_anxiety اضافه میکنیم که برای داده های depression برابر با صفر و برای anxiety برابر با ۱ است.

	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	link_flair_richtext	...	parent_whitelist_status
0	None	depression	We understand that most people who reply immed...	t2_1t70	False	None	1	False	Our most-broken and least-understood rules is ...	[]	...	no_ads
1	None	depression	Welcome to /r/depression's check-in post - a p...	t2_1t70	False	None	0	False	Regular Check-In Post, with important reminder...	[]	...	no_ads
2	None	depression	I'm so low rn I can't even type anything coher...	t2_8oa0yyky	False	None	0	False	Low	[]	...	no_ads
3	None	depression	When I wake up after 8 hours of decent sleep I...	t2_8bk84r51	False	None	0	False	I'm always amazed at how much energy healthy p...	[]	...	no_ads
4	None	depression	I guess i have always been depressed but never...	t2_bzoskmwx	False	None	0	False	30 and never lived a day in my life	[]	...	no_ads

5 rows × 104 columns

تعداد ستون ها برای anxiety بیشتر است که به صورت زیر است:

	approved_at_utc	subreddit	selftext	author_fullname	saved	mod_reason_title	gilded	clicked	title	link_flair_richtext	...	stickied	ur
0	None	Anxiety	Hello everyone! Welcome to the r/Anxiety month...	t2_6l4z3	False	None	0	False	Monthly Check-In Thread	[]	...	True	https://www.reddit.com/r/Anxiety/comments/myl0..
1	None	Anxiety	With the subreddit continuing to grow we're lo...	t2_5uptt	False	None	0	False	Looking for new mods!	[]	...	True	https://www.reddit.com/r/Anxiety/comments/modo..
2	None	Anxiety	The company that I worked for: "Hey it's menta...	t2_7grjc3lq	False	None	0	False	It's so frustrating when society wants to be a...	[]	...	False	https://www.reddit.com/r/Anxiety/comments/n6ze..
3	None	Anxiety	It's pretty simple and may seem obvious, but j...	t2_50q4oaru	False	None	0	False	My therapist recently taught me a trick that h...	[]	...	False	https://www.reddit.com/r/Anxiety/comments/n6kr..
4	None	Anxiety	I'm proud as fuck of myself. It's hard. Really...	t2_ziole	False	None	0	False	I'm 31 years old and have been depressed and h...	[]	...	False	https://www.reddit.com/r/Anxiety/comments/n6r5..

5 rows × 108 columns

تفاوت ستون ها در زیر آورده شده است در بررسی های صورت گرفته به این نتیجه رسیدیم که این فیلد ها برای کار ما اهمیت کمتری دارند و در مرحله پاکسازی داده از ستون های کمتری که اطلاعات مهمتری دارند استفاده میکنیم.

```
Index(['link_flair_template_id', 'post_hint', 'preview', 'thumbnail_height',
       'thumbnail_width'],
```

این کار را برای هر دو دسته انجام می‌دهیم و سپس آن ها به صورت فایل .csv در آدرس data/anxiety.csv/ و data/depression.csv/ ذخیره می‌کنیم.

4.پیش پردازش های انجام شده

i. روش/ابزار تفکیک جملات

برای این کار از nltk.tokenize و sent_tokenize استفاده شده است.

ii. روش/ابزار تفکیک توکن ها/کلمات

برای این کار از RegexpTokenizer در nltk.tokenize استفاده شده است. Regular expression را به گونه ای تنظیم میکنیم که تنها کلمات و حروف را در این کار در نظر بگیرد و از اعداد صرف نظر کند.

iii. روش/معیارهای تمیز کردن داده

با توجه به اطلاعاتی که درباره فایل های JSON در لینک های reddit/wiki/JSON و reddit.com: api documentation آورده شده است، ستون های "","score", "author", "selftext", "title" ""num_comments", "is_anxiety", "url" انتخاب شده اند که اطلاعات بیشتری در اختیار ما قرار می‌دهند.

علت در نظر گرفتن url برای هنگامی است که نیاز شد به آن پست نگاه عمیق تری انداخته شود.

همچنین score نشان دهنده واکنش هایی که به صورت مثبت یا منفی نسبت به آن پست صورت گرفته است و همین طور تعداد کامنت ها میتواند اطلاعات مفیدی درباره اهمیت یک پست به ما بدهد.

	title	selftext	author	score	num_comments	is_anxiety	url
0	Our most-broken and least-understood rules is ...	We understand that most people who reply immed...	SQLwitch	2319	175	0	https://www.reddit.com/r/depression/comments/d...
1	Regular Check-In Post, with important reminder...	Welcome to /r/depression's check-in post - a p...	SQLwitch	312	1136	0	https://www.reddit.com/r/depression/comments/m...
2	Low	I'm so low rn I can't even type anything coher...	RagingFlock89	263	43	0	https://www.reddit.com/r/depression/comments/n...
3	I'm always amazed at how much energy healthy p...	When I wake up after 8 hours of decent sleep I...	cezzzie	1281	120	0	https://www.reddit.com/r/depression/comments/n...
4	30 and never lived a day in my life	I guess i have always been depressed but never...	ApprehensiveYou2385	36	5	0	https://www.reddit.com/r/depression/comments/n...
...	...	...	...	...	...	...	...
1925	Any tips on how not to panic during a midterm?	I have an Applied Statics midterm tomorrow. I ...	Anomalistic_Username	2	0	1	https://www.reddit.com/r/Anxiety/comments/n3p5...
1926	I find myself apologizing really often, checki...	I've recently decided to stop smoking weed (so...	zedhenson	3	1	1	https://www.reddit.com/r/Anxiety/comments/n3n1...
1927	I typed out my anxiety attack and thought I sh...	I recently got into a little habit where when ...	Tree-Nui-Tee	2	4	1	https://www.reddit.com/r/Anxiety/comments/n3oz...
1928	something happened that just triggered my anxi...	I need someone to vent to please	Capzfan5	8	5	1	https://www.reddit.com/r/Anxiety/comments/n3gc...
1929	Fear of Unknown Person	Everytime I go to my Grandma's old house (she'...	PanamaPhys_	2	0	1	https://www.reddit.com/r/Anxiety/comments/n3ot...

همچنین برای رکورد هایی که مقدار selftext آن ها Null بود مقدار "emptypost" را جایگزین کردیم.

داده های تا این مرحله را مجدد در آدرس '../data/combined_data.csv' ذخیره میکنیم جهت backup داشتن.

در پیش پردازش متن را به lowercase تبدیل کرده، علائم نگارشی حذف شده اند و کلمات با استفاده lemmatizer به یک کلمه‌ی متداول تر تبدیل شده اند و کلمات کمتر دارای اهمیت حذف شده اند.

برای این کار از nltk.stem و به طور خاص WordNetLemmatizer و همچنین از stopwords زبان انگلیسی nltk.corpus استفاده شده است.

بنابراین پس از این مرحله داده های اعداد و علائم نگارشی و emoji و کلماتی مانند we our دیده نمیشود و در نهایت پس از این پاکسازی مجدد یک استرینگ طولانی از این درست میکنیم و به عنوان selftext_clean در dataframe قرار میدهیم.

	selftext	selftext_clean
0	We understand that most people who reply immediately to an OP with an invitation to talk privately mean only to help, but this type of response usually leads to either disappointment or disaster. it usually works out quite differently here than when you say "PM me anytime" in a casual social context. \n\nWe have huge admiration and appreciation for the goodwill and good citizenship of so many of you who support others here and flag inappropriate content - even more so because we know that so many of you are struggling yourselves. We're hard at work behind the scenes on more information and resources to make it easier to give and get quality help here - this is just a small start. \n\nOur new wiki page explains in detail why it's much better to respond in public comments, at least until you've gotten to know someone. It will be maintained at /r/depression/wiki/private_contact, and the full text of the current version is below.\n\n*****\n\n###Summary###\n\n**Anyone who, while a...	understand people reply immediately op invitation talk privately mean help type response usually lead either disappointment disaster usually work quite differently say pm anytime casual social context huge admiration appreciation goodwill good citizenship many support others flag inappropriate content even know many struggling hard work behind scene information resource make easier give get quality help small start new wiki page explains detail much better respond public comment least gotten know someone maintained r depression wiki private_contact full text current version summary anyone acting helper invite accepts private contact e pm chat kind offsite communication early conversion showing either bad intention bad judgement either way unwise trust pm anytime seems like kind generous offer might perfectly well meaning unless solid rapport ha established wise idea point consider offer accept invitation communicate privately posting supportive reply publicly help people op respons...
1	Welcome to /r/depression's check-in post - a place to take a moment and share what is going on and how you are doing. If you have an accomplishment you want to talk about (these shouldn't be standalone posts in the sub as they violate the "role model" rule, but are welcome here), or are having a tough time but prefer not to make your own post, this is a place you can share.\n\nWe try our best to keep this space as safe and supportive as possible on reddit's wide-open anonymity-friendly platform. The community rules can be found in the sidebar, or under "Community Info" in the official mobile apps. If you aren't sure about a rule, please ask us. \n\n********\n\n**Please keep in mind that no activism, i.e. advocating or fundraising for social change or raising awareness of social issues, is ever allowed here.** \n\nIt's not that we're against activism. We're strongly in favour of it. But we've learned the hard way that it doesn't work within a dedicated support space, so with re...	welcome r depression check post place take moment share going accomplishment want talk standalone post sub violate role model rule welcome tough time prefer make post place share try best keep space safe supportive possible reddit wide open anonymity friendly platform community rule found sidebar community info official mobile apps sure rule please ask u please keep mind activism e advocating fundraising social change raising awareness social issue ever allowed activism strongly favour learned hard way work within dedicated support space regret allow thanks understanding please report fundraising awareness raising petition call participation post mainly cause issue rather request personal support despite fact always maintain sticky post still seeing lot violation private contact policy please read r depression wiki private_contact report violation see widespread disregard important rule cause trouble helper vulnerable ops waste time delay improvement making community resource

همچنین در ستون های دیگری selftext_broken_sentences و selftext_broken_words را قرار میدهیم.

	title	selftext	author	score	num_comments	is_anxiety	url	selftext_clean	selftext_broken_sentences	selftext_broken_words	title_clean
0	Our most-broken and least-understood rules is "helpers may not invite private contact as a first...	We understand that most people who reply immediately to an OP with an invitation to talk private...	SQLwitch	2319	175	0	https://www.reddit.com/r/depression/comments/doqwow/our_mostbroken_and_leastunderstood_rules_is/	understand people reply immediately op invitation talk privately mean help type response usually...	[we understand that most people who reply immediately to an op with an invitation to talk privat...	[understand, people, reply, immediately, op, invitation, talk, privately, mean, help, type, resp...	broken least understood rule helper may invite private contact first resort made new wiki explain
1	Regular Check-In Post, with important reminders about the No Private Contact and No Activism rules	Welcome to /r/depression's check-in post - a place to take a moment and share what is going on a...	SQLwitch	312	1136	0	https://www.reddit.com/r/depression/comments/m246c4/regular_checkin_post_with_important_reminders/	welcome r depression check post place take moment share going accomplishment want talk standalon...	[welcome to /r/depression's check-in post - a place to take a moment and share what is going on ...	[welcome, r, depression, check, post, place, take, moment, share, going, accomplishment, want, t...	regular check post important reminder private contact activism rule
2	Low	I'm so low rn I can't even type anything coherent. I just want to express I'm at one of my lowes...	RagingFlock89	263	43	0	https://www.reddit.com/r/depression/comments/n728cp/low/	low rn even type anything coherent want express one lowest point stupid thing thanks listening	[i'm so low rn i can't even type anything coherent. i just want to express i'm at one of my lowe...	[low, rn, even, type, anything, coherent, want, express, one, lowest, point, stupid, thing, than...	low

مراحل یاد شده را متناسب با نوع داده بر روی author و title نیز اعمال میکنیم و به ترتیب در author_clean و title_clean در dataframe قرار میدهیم.

author	author_clean
khaellar	kha ella r
justborgia	borgia
Dependent_Ad_8358	dependent ad 8358
DumpsterPuff	dumpster puff
hawaiianshirtYT-1300	hawaiian shirt 1300
Anomalistic_Username	mali stic username
zedhenson	zed henson
Tree-Nui-Tee	tree nui tee
Capzfan5	cap z fan 5
PanamaPhys_	panama ph

همچنین مقادیر null موجود در dataframe را در پایان بررسی میکنیم که به صورت زیر است:

```
1 combined_data.isnull().sum()
```

```
title                          0
selftext                       0
author                         0
score                          0
num_comments                   0
is_anxiety                     0
url                            0
selftext_clean                 0
selftext_broken_sentences      0
selftext_broken_words          0
title_clean                    0
author_clean                   0
dtype: int64
```

iv. اندازه داده قبل/بعد تمیز کردن داده

تعداد کل پست ها ثابت مانده است زیرا داده ای را به طور کامل حذف نکرده ایم اما در فیلد های انتخابی با پردازش های صورت گرفته، در متن هر پست از تعداد کلمات کاسته شده است.

5.واحد برچسب گذاری (جمله، توییت، صفحه وب، ...) و روش برچسب گذاری

واحد برچسب گذاری هر پست مجزایی است که در هر کدام از community های مورد بررسی قرار دارند، است. به طور مشخص وضعیت هر پست در ستون is_anxiety مشخص شده که این مقدار برای پست های اضطراب ۱ و برای پست های افسردگی ۰ است.

6. آمار داده به تفکیک برچسب در قالب جدول و نمودار

i. تعداد «واحد» داده

```
print("All posts:", len(combined_data))

print("Depression posts:", len(combined_data[combined_data["is_anxiety"] == 0]))
print("Anxiety posts:", len(combined_data[combined_data["is_anxiety"] == 1]))
```

```
All posts: 1930
Depression posts: 932
Anxiety posts: 998
```

ii. تعداد جملات، تعداد کلمات و تعداد کلمات منحصر به فرد

```
All sentences count: 19646
All words count: 162773
All unique words count: 10578
```

```
depression sentences count: 10216
depression words count: 80588
depression Unique words: 7551
```

```
anxiety sentences count: 9430
anxiety words count: 82185
anxiety Unique words: 7492
```

iii. تعداد کلمات منحصر به فرد مشترک و غیرمشترک بین برچسب‌ها

```
intersection words count: 4465
difference words count: 6113
```

iv. ۱۰ کلمه پرتکرار غیر مشترک هر برچسب

Depression top 10 difference words

	word	count
2187	eh	22
4119	mil	17
7093	ward	17
3929	lucas	17
628	ba	16
1671	dd	12
2174	education	12
5638	scar	12
4773	pity	11
793	bird	11

Anxiety top 10 difference words

	word	count
2822	gad	37
6564	tingling	36
243	abortion	21
1046	caffeine	20
4297	nauseous	19
6930	vaccinated	17
4313	needle	16
2086	doom	15
4609	palpitation	15
3363	impending	14

v. ۱۰ کلمه مشترک برتر هر برچسب نسبت به برچسبهای دیگر بر اساس معیار زیر:

$$\text{RelativeNormalizedFrequency}(w_i) = \frac{\frac{\text{count}(w_i)}{\sum_{w \in \text{Label1}} \text{count}(w)}}{\frac{\text{count}(w_i)}{\sum_{w \in \text{Label2}} \text{count}(w)}}$$

```
[('exist', 30.594505385417182),
 ('existence', 21.41615376979203),
 ('pretend', 18.35670323125031),
 ('creative', 15.297252692708591),
 ('faith', 15.297252692708591),
 ('bullied', 14.277435846528018),
 ('halfway', 13.257619000347447),
 ('mirror', 12.237802154166875),
 ('subject', 11.2179853079863),
 ('destroy', 11.2179853079863),
 ('scholarship', 10.198168461805729),
 ('apps', 10.198168461805729),
 ('happiness', 9.382314984861269),
 ('earth', 9.178351615625155),
 ('approach', 9.178351615625155),
 ('void', 9.178351615625155),
```

vi. ۱۰ کلمه برتر هر برچسب بر اساس TF-IDF (در اینجا یک داکیومنت برابر است با تمام داده های متناظر با یک برچسب)

Depression top 10 TFIDF words

	word	count
10141	wa	0.310940
5517	like	0.300926
3707	feel	0.292664
10178	want	0.207043
5334	know	0.206042
5495	life	0.184761
4151	get	0.182258
9462	time	0.168238
3438	even	0.168238
10497	year	0.133439

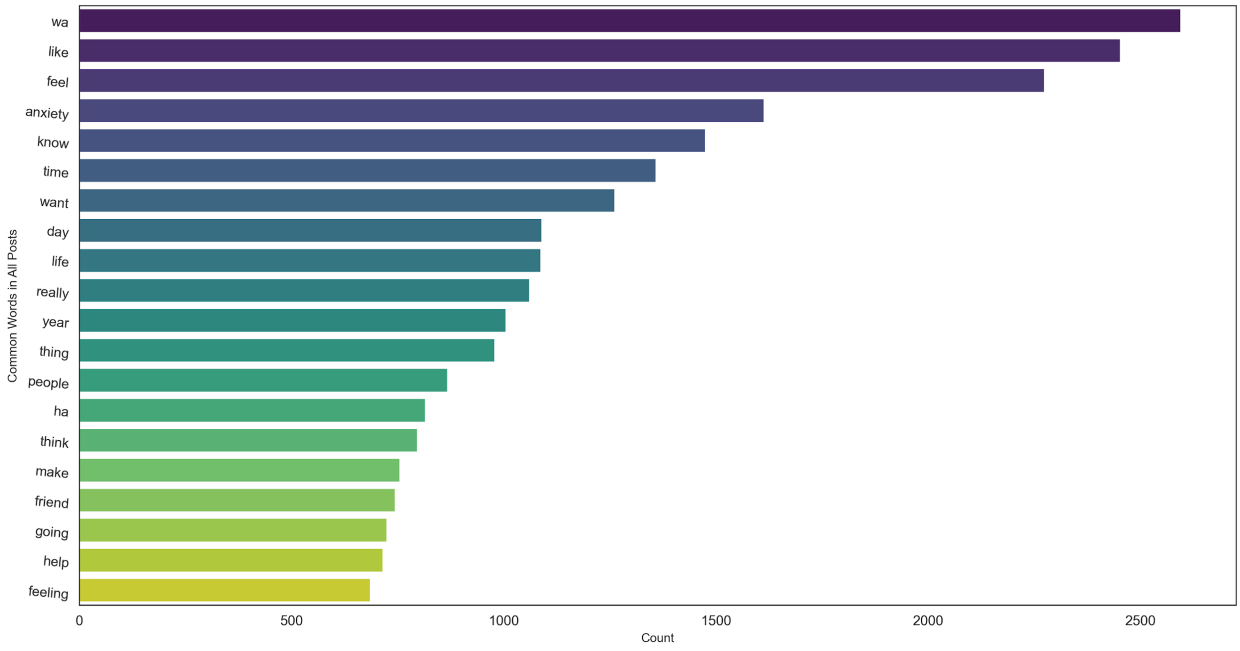
Anxiety top 10 TFIDF words

	word	count
709	anxiety	0.349713
10141	wa	0.328356
5517	like	0.303602
3707	feel	0.268170
4151	get	0.191966
9462	time	0.166726
5334	know	0.158232
7522	really	0.141730
2464	day	0.139545
3438	even	0.121344

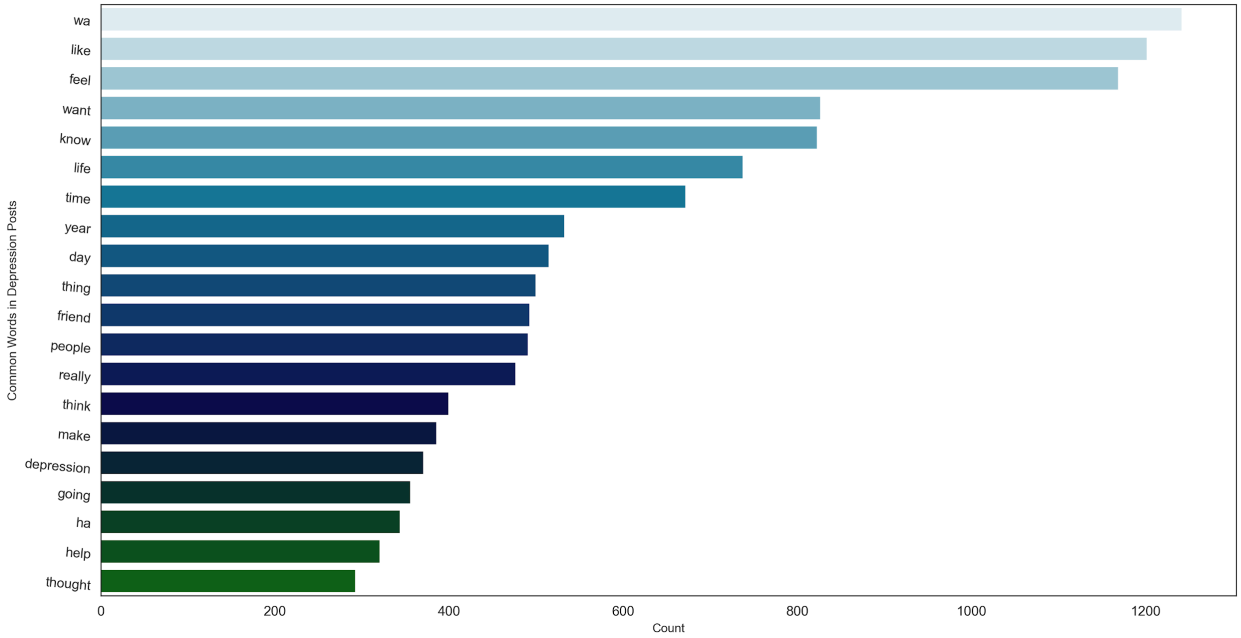
vii. هیستوگرام تعداد تکرار هر کلمه منحصر به فرد به ترتیب از فرکانس بالا به پایین

نمودار برای بیشترین ۲۰ مورد اول رسم شده است.

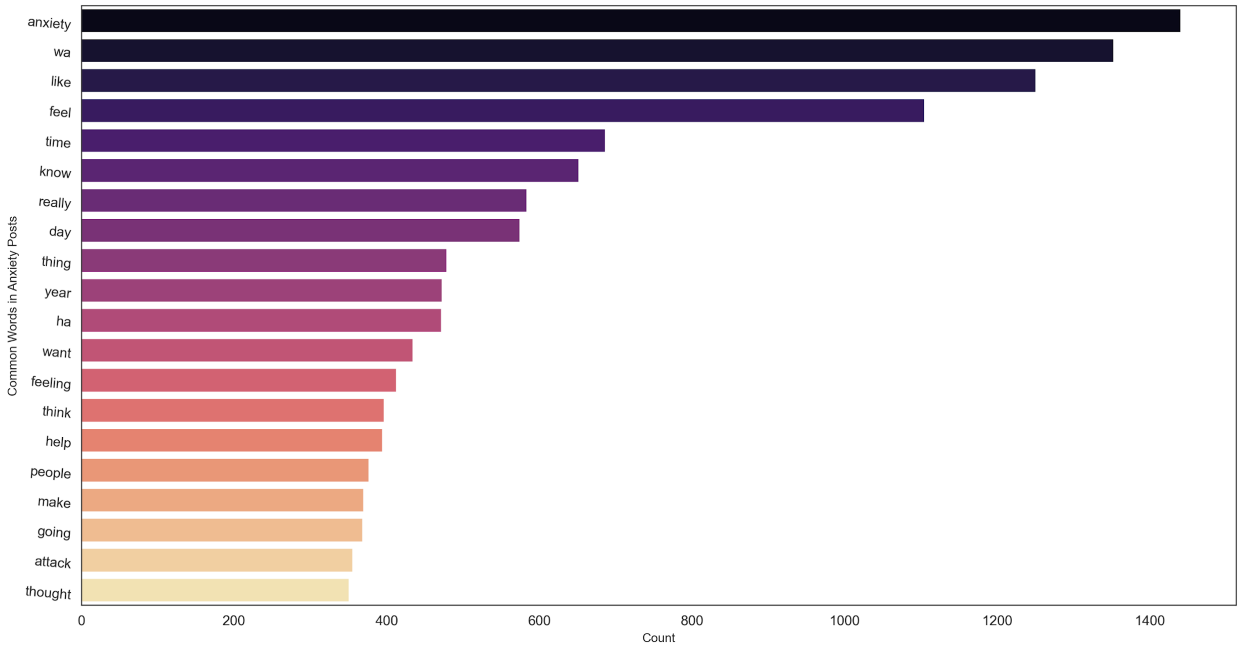
تمام پست ها:


wa
like
feel
anxiety
know
time
want
day
life
really
year
thing
people
ha
think
make
friend
going
help
feeling
Common Words in All Posts
0
500
1000
1500
2000
2500
Count

پست های depression:



پست های anxiety:



Word cloud .viii

تمام پست ها:

پست های depression:

Top Words used in Depression Posts

پست های anxiety:

Top Words used in Anxiety Posts