

بسم الله الرحمن الرحيم

گزارش فاز اول پروژه پردازش زبان طبیعی

غزاله محمودی

تاریخ: 21 اردیبهشت ماه 1400

داده‌های استفاده شده از پست‌های subreddit استخراج شده‌اند. به کمک api reddit داده‌های مورد نیاز به دست آمد.

→ <https://www.reddit.com/r/success.json>



به کمک api موجود subreddit و request به آن دیتا مورد نظر استخراج شده است. برای استخراج داده‌ها کافیهست فایل data_collection.py اجرا شود (لازم است فیلتر شکن روشن باشد). برای اینکه از بلاک شدن توسط سرور یا هر مشکل دیگری جلوگیری شود بین هر دو ریکوئست زمان تصادفی توقف می‌کنیم. با استفاده از آخرین پست و فیلد after، یک صفحه از پست های reddit را استخراج می‌کنیم.

داده به صورت فایل جیسون استخراج شده است. این داده‌گان شامل اطلاعات زیادی است که ستون‌های موجود به صورت زیر است.

```
Index(['approved_at_utc', 'subreddit', 'selftext', 'author_fullname', 'saved',
      'mod_reason_title', 'gilded', 'clicked', 'title', 'link_flair_richtext',
      ...,
      'permalink', 'parent_whitelist_status', 'stickied', 'url',
      'subreddit_subscribers', 'created_utc', 'num_crossposts', 'media',
      'is_video', 'author_cakeday'],
      dtype='object', length=103)
```

در این پروژه و در قسمت تمیز کردن دیتا قسمت مورد استفاده جدا می‌شود.

	title	author	score	url	selftext
0	Our most-broken and least-understood rules is ...	SQLwitch	2325	https://www.reddit.com/r/depression/comments/d...	We understand that most people who reply immed...
1	Regular Check-In Post, with important reminder...	SQLwitch	334	https://www.reddit.com/r/depression/comments/m...	Welcome to /r/depression's check-in post - a p...
2	I'm becoming a bit confident lately and I want...	Giovanni_Islost	341	https://www.reddit.com/r/depression/comments/n...	Guys lately I feel like I'm a good mood and I'...
3	Isn't it messed up that we're just born withou...	dumbbinch99	143	https://www.reddit.com/r/depression/comments/n...	and now because of circumstances that were com...
4	It sucks when you wake up and the only thing y...	Chonknyster	2654	https://www.reddit.com/r/depression/comments/n...	The first 6 words of the title is accurate too.

در این قسمت تنها ستون‌های title, نویسنده، score و url و متن پست نگه داشته می‌شود.

این داده‌ها شامل دو کلاس افراد دچار استرس-افسردگی در مقابل افراد شاد-موفق می‌باشد.









داده‌های استرس-افسردگی با لیبل '0' مشخص می‌شوند. داده‌های افراد شاد-موفق در با لیبل '1'

مشخص می‌شوند. در آنالیزها و آمار تنها لیبل‌ها با 0 و 1 مشخص شده‌اند. پوشه بندی فایل‌ها

به صورت زیر است. نکته جالب توجه این است که تعداد پست‌ها و اطلاعات هر پست

استرس-افسردگی بیشتر از شاد موفق می‌باشد. به نظر می‌آید متن موجود در پست‌های لیبل '0'

بیشتر از لیبل '1' است.

- ▼ data
 - ▼ cleaned
 - { } data_cleand.json
 - ▼ images
 -  hist-20top-label_0.png
 -  hist-20top-label_1.png
 -  hist-label_1.png
 - ▼ row
 - ▼ 0
 - { } depression.json
 - { } stress.json
 - ▼ 1
 - { } happiness.json
 - { } success.json
 - ▼ sentence_broken
 - ≡ 0_sentences.txt
 - ≡ 1_sentences.txt
 - ▼ word_broken
 - ≡ 0_words.txt
 - ≡ 1_words.txt
 - ▼ src
 -  data_analysis.py
 -  data_cleaning.py
 -  data_collection.py
 - { } analysis-report.json
 -  Phase1-Report.pdf
 -  README.md

data_collection.py : برای دریافت داده.

data_cleaning.py : برای دریافت تمیز کردن داده و شکستن جملات و کلمات.

data_analysis.py : برای آنالیز آماری داده.

در صورت اجرا هر یک از فایل‌های بالا با دستور `python *file_name` کار بیان شده انجام می‌شود.

4. پیش‌پردازش‌های انجام شده

کلیه مراحل این مرحله به صورت متوالی در فایل `data_cleaning.py` نوشته شده و پس از اتمام اجرا جملات در `data/sentence_broken` در فایل `txt`، کلمات/توکن‌ها در `data/word_broken` و داده تمیز در `data/cleaned` در قالب فایل `json` ذخیره می‌شود.

4.1. روش/ابزار تفکیک جملات

برای تفکیک جملات از `nltk.tokenize.punkt.PunktSentenceTokenizer` استفاده شده است. در نهایت خروجی تفکیک جملات در `data/sentence_broken` در دو فایل `txt` بر حسب کلاس داده‌ها موجود می‌باشد.

4.2. روش/ابزار تمیزکردن

در این مرحله ابتدا `row` های `null` دیتا با `"IS NULL"` جایگزین شدند. سپس برای یکدست شدن کلمات همه حروف انگلیسی به `lower` تبدیل شدند. در قدم بعدی `hyper link` های موجود در متن به کمک `re` حذف شدند. در این مرحله با استفاده از `nltk.tokenize.punkt.PunktSentenceTokenizer` جملات استخراج می‌شوند.

در ادامه فرایند تمیز کردن داده‌ها با `nlTK.tokenize.RegexpTokenizer` متن به توکن‌هایی تقسیم می‌شود. مرحله بعد اعداد موجود در متن حذف شدند چرا که این طور به نظر می‌آید توکن اعداد کمکی در تشخیص کلاس و تمایز دو کلاس نمی‌کند. سپس توکن‌های موجود را `Lemmatize` کرده و در نهایت `stop words` را از میان توکن‌ها حذف می‌کنیم. در نهایت تنها توکن‌هایی که `alphabet` باشند را نگه داشته و بقیه را حذف کردیم به این امید که مواردی چون `emoji` نیز حذف شود. در این مرحله دیتا تمیز شده را با در نظر گرفتن لیبل داده‌ها در `data/cleaned_data` در قالب فایل `json` ذخیره کردم. در این مرحله تنها ستون‌های مورد استفاده داده از جمله `cleaned_text`, `self_text`, `score`, `author`, `url` را نگه داشته شد. نمونه‌ای از مقایسه داده تمیز شده به صورت زیر است.

	selftext	selftext_clean
0	<p>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...</p> <p>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.</p>	<p>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...</p> <p>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</p>
1		

4.3. روش/معیارهای تفکیک توکن‌ها/کلمات

همان‌طور که در قسمت قبل با جزئیات بیان شد از `nlTK` برای تفکیک داده‌ها استفاده می‌شود. پس از اتمام مراحل تمیز کردن دیتا توکن/کلمات هم استخراج می‌شوند.

4.4. اندازه داده‌ها قبل و بعد از تمیز کردن داده

قبل از تمیز کردن داده‌ها مجموع حجم کل داده‌ها 30M بود. بعد از تمیز کردن به حوالی 5 مگابایت کاهش پیدا کرد که دلیل اصلی آن پاک کردن ستون‌های اضافه بود.

5. واحد برچسب گذاری

داده‌های استخراج شده از [stress.json](#)، [depression.json](#) لیبل '0' و [happiness.json](#) لیبل '1' دارند که به صورت واضح دو کلاس از یکدیگر تمایز دارند.

6. آمار

کلیه آمار به دست آمده در analysis.json قابل مشاهده می‌باشد.

6.1. تعداد «واحد» داده

```
"number of data units":
  "0": 2495,
  "1": 2490
```

6.2. تعداد جملات

```
"sentences count":
  "0": 22664,
  "1": 8957
```

6.3. تعداد کلمات

```
"words count": {
  "0": 181314,
  "1": 65329
```


6.4. تعداد کلمات منحصر به فرد

```
"intersection words count": 4059,
```

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

```
"uncommon words count": 12062,
```

6.6. 10 کلمه پرتکرار غیر مشترک هر برچسب

```
"most frequency uncommon words": {
  "0": {
    "cant": 145,
    "suicide": 127,
    "treatment": 111,
    "suicidal": 101,
    "diagnosed": 87,
    "severe": 82,
    "sleeping": 80,
    "breathing": 72,
    "dying": 68,
    "chest": 67
  },
  "1": {
    "drey": 188,
    "ourself": 31,
    "happinessnow": 24,
    "swept": 17,
    "attain": 17,
    "alice": 17,
    "jeremy": 16,
    "winner": 15,
    "entrepreneur": 15,
    "dedication": 14
  }
}
```

6.7. 10 کلمه مشترک برتر هر برچسب نسبت به برچسب‌های دیگر بر اساس معیار RNF

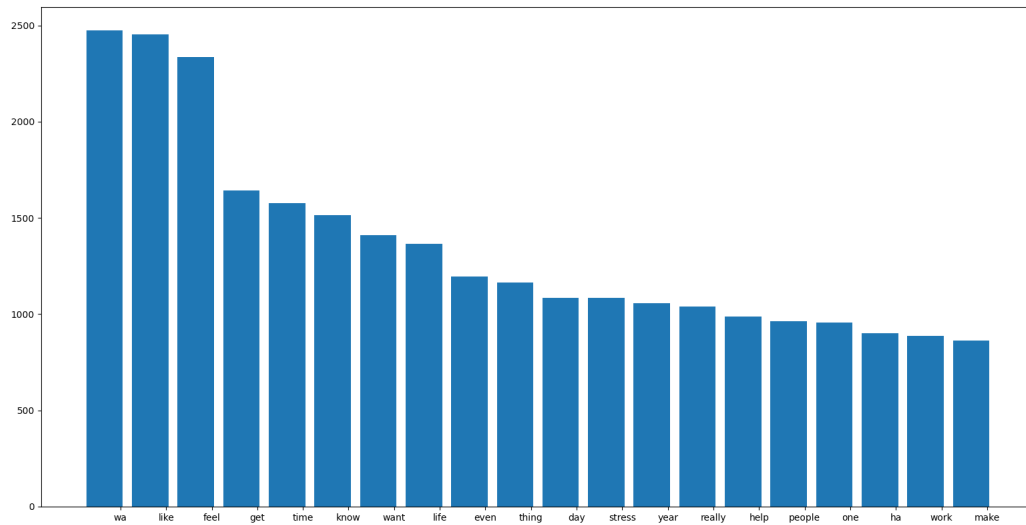
```
"Relative Normalize Frequency": {  
  "0": {  
    "worried": 89.75048258821712,  
    "idk": 80.86429619334415,  
    "mentally": 67.53501660103467,  
    "medication": 66.64639796154736,  
    "kinda": 66.64639796154736,  
    "sorry": 62.64761408385453,  
    "cut": 62.20330476411088,  
    "barely": 60.42606748513628,  
    "symptom": 59.093139525905336,  
    "participant": 56.87159292718709  
  },  
  "1": {  
    "teaching": 54.016422644132604,  
    "divine": 37.13629056784116,  
    "embracing": 28.133553460485725,  
    "lt": 27.148879089368727,  
    "imposter": 27.008211322066302,  
    "abundance": 25.882869183646868,  
    "famous": 23.63218490680801,  
    "accountability": 21.381500629969153,  
    "craft": 21.381500629969153,  
    "john": 20.256158491549723  
  }  
}
```

6.8. 10 کلمه برتر هر برجسب بر اساس $TF-IDF(w)$ در اینجا یک داکيومنت برابر است با تمام داده های متناظر با یک برجسب)

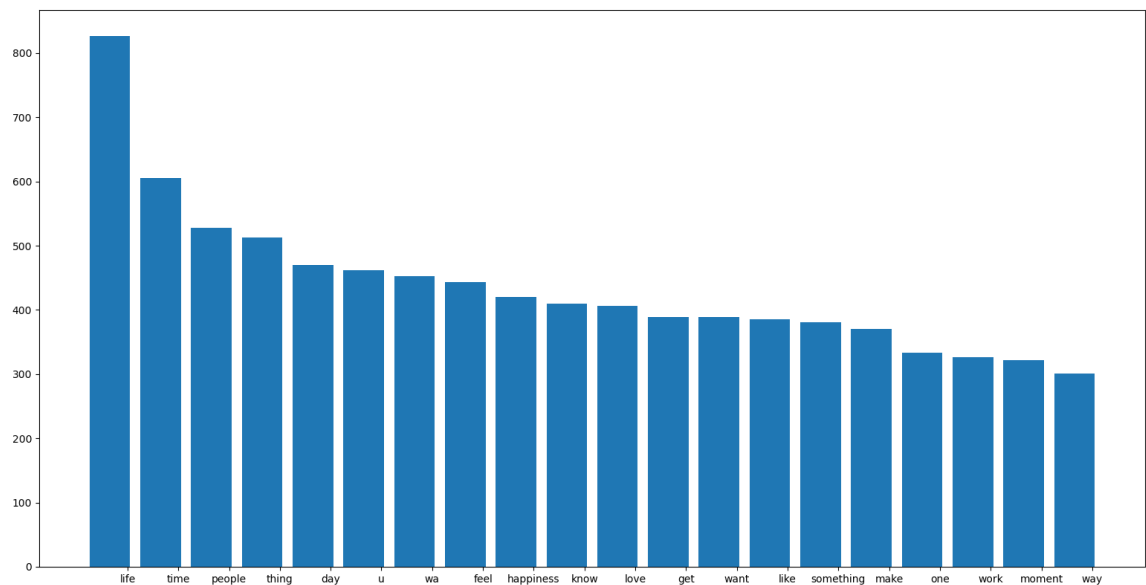
```
"TF IDF": {  
  "0": {  
    "wa": 0.29643031258185404,  
    "like": 0.29415284556242205,  
    "feel": 0.27988871001966403,  
    "get": 0.196701230467781,  
    "time": 0.18902976261285231,  
    "know": 0.1817178948136234,  
    "want": 0.16913189286413102,  
    "life": 0.16361802534340103,  
    "even": 0.14312082216851343,  
    "thing": 0.13976455498198212  
  },  
  "1": {  
    "life": 0.2942159401479946,  
    "time": 0.21549714744495974,  
    "people": 0.18807023777014667,  
    "thing": 0.18272733328804025,  
    "day": 0.1674110071060018,  
    "wa": 0.1613557153596145,  
    "feel": 0.15779377903821018,  
    "happiness": 0.1496013254989803,  
    "know": 0.14603938917757603,  
    "love": 0.1446146146490143  
  }  
}
```

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

نتایج این قسمت در data/images قابل مشاهده می‌باشد.



هیستوگرام 20 کلمه برتر افراد دچار استرس-افسردگی



هیستوگرام 20 کلمه برتر افراد دچار شاد-موفق

[How to Convert Pandas DataFrame to a Dictionary](#)

[Python sort dictionary by Value/Key | Ascending & Descending order](#)

[TF IDF | TFIDF Python Example. An example of how to implement TFIDF... | by Cory Maklin](#)

[Stemming vs Lemmatization. Truncate a word to its root or base... | by Aditya Beri](#)

https://github.com/hesamuel/goodbye_world/tree/master/code

<https://ivyproschool.com/blog/text-preprocessing-using-nltk-in-python/>

[Term Frequency and Inverse Document Frequency with scikit-learn | by Rohan Paul | Analytics Vidhya](#)

[NLP Snippets #1: Clean and Tokenize Text With Python](#)