لازم به ذکر است تقریبا در انتهای هر بخش دیتای به دست آمده ذخیره شده و در بخش بعد استفاده میشود. به همین دلیل شما نیازی به به ترتیب اجرا کردن ستون ها ندارید بلکه میتوانید با در اختیار داشتن دادگان مناسب از هر بخشی که نیاز دارید استفاده کنید.

**چکیده**

در این پژوهش دادگان به صورت آنلاین از افراد بزرگسال مختلفی جمع آوری شده­اند، که شامل دو وظیفه گفتاری مختلف که اولی سه سوال در مورد گذشته، حال و آینده آزمودنی و دومی توصیف هشت تصویر توسط آزمودنی است. همچنین چهار پرسش‌نامه به نام‌های پرسش‌نامه افسردگی بک، پرسش‌نامه اضطراب بک، پرسش‌نامه اضطراب حالت صفت بک و پرسش‌نامه پاناس نیز از آزمودنی‌ها جمع‌آوری شده است. هدف تحقیق روی بیماری افسردگی است. برای ارزیابی علائم افسردگی، از وظیفه گفتاری توصیف هشت تصویر توسط 119 آزمودنی تصادفی بالای هجده سال سن، از میزان امتیازات افراد در پرسش‌نامه افسردگی بک استفاده شد. این پژوهش با هدف تحلیل رابطه بین نمرات پرسش‌نامه افسردگی بک و ویژگی‌های مربوط به متن پیش برده شده است. با استفاده از تحلیل ماتریس شباهت بردارهای معنایی، نشان داده شد که بعضی از این ویژگی‌ها تأثیری در علائم افسردگی دارد. برای مثال بزرگسالان با علائم افسردگی بیشتر مقدار انحراف معیار فاصله اقلیدسی و کسینوسی بیشتر بین بردارهای معنایی در بین جملات هستند. علاوه بر این، در حالی که مشخص شد برخی ویژگی‌ها در توضیح علائم افسردگی و اضطراب موثرند، برای تعیین اهمیت آن‌ها در پیش‌بینی این وضعیت‌های روانی، از مدل‌ه یادگیری ماشین رگرسیون خطی استفاده شده است. ویژگی‌های بررسی شده گراف گفتار، احساسات متن، میزان پرش فکری، بردارهای معنایی می‌باشند. از بردارهای معنایی برای تحلیل ارتباط بین ویژگی‌ها و امتیازات پرسش‌نامه افسردگی نیز استفاده شده است. بردار‌های معنایی به سه روش بردارهای معنایی به دست آمده به روش توکن‌محور، جمله‌محور و میانگین تعبیه جملات محور به دست آمده‌اند و با هم مقایسه شده‌اند. تمامی‌مراحل(پیدا کردن میزان همبستگی بین ویژگی‌ها با امتیازات پرسش‌نامه افسردگی بک، مشاهده ماتریس شباهت به تفکیک هر تصویر، ‌پیش‌بینی امتیازات پرسش‌نامه افسردگی بک با استفاده از مدل رگرسیون خطی) روی هر سه مورد انجام شده است. از آن‌ها نیز برای تحلیل ماتریس شباهت استفاده شده است تا اگر ویژگی ای روند خاصی طبق امتیازات پرسش‌نامه افسردگی بک با ماتریس شباهت بین بردارهای معنایی داشته باشد از آن ویژگی به عنوان ورودی طبقه بند برای ‌پیش‌بینی امتیازات پرسش‌نامه افسردگی بک استفاده شود.

# 1.Create only image CSV

لازم به ذکر است در صورتی که شما فایل CSV دادگان را در اختیار دارید نیاز به اجرای این بخش ندارید.

دیتاست شامل داده های آزمون صوتی تشکیل شده از دو تسک گفتاری است. یکی از تسک های گفتاری مربوط به سه سوال در مورد گذشته و حال و آینده آزمودنی ها است و تسک صوتی دوم صحبت های آزمودنی در مورد هشت تصویر است. در این پروژه تنها از دادگان تسک صوتی دوم که مربوط به 8 تصویر است استفاده شده است به همین دلیل در ابتدای کد ها Create only image CSV انجام شده است. که در آن ابتدا دادگان با فرمت mp3 به wav تبدیل شده اند(که این کد فقط یک بار نیاز به اجرا شدن دارد).

بعد از آن speech to text با استفاده از vosk انجام شده است(کد های سلولهای این بخش به ترتیب باید اجرا شوند).

پس از یک سری تغییرات روی ستون های دیتاست فعلا به دیتافریم زیر میرسیم:

|  | **file\_name** | **transcript** | **id** | **TimeStamp** |
| --- | --- | --- | --- | --- |
| **0** | 48xdti2f\_1681027836 | "این دوتا داداش دارم می‌کشد من کوچیکتر دوسال پ... | 48xdti2f | 1681027836 |
| **1** | 48xdti2f\_1681028046 | "شوید خیلی خیلی شرطی جالبی نیست یعنی آنقدر به ... | 48xdti2f | 1681028046 |
| **2** | 48xdti2f\_1681028216 | "در آینده که در قانون زیاده ولی چیزی که الان خ... | 48xdti2f | 1681028216 |
| **3** | 48xdti2f\_1681028409 | "این عکس مال یه دختر یه پسر بچه بودم با بازی م... | 48xdti2f | 1681028409 |
| **4** | 48xdti2f\_1681028467 | "یه دختره که داره از داخل خون به بیرون نگاه می... | 48xdti2f | 1681028467 |
| **...** | ... | ... | ... | ... |
| **1420** | zatfknoo\_1678035079 | "این نرم یا تصویر یادمه ناراحت شاید یا شادی اد... | zatfknoo | 1678035079 |
| **1421** | zatfknoo\_1678035135 | "من زیاد با این تصویر احساس می‌کنم که ارتباط ب... | zatfknoo | 1678035135 |
| **1422** | zatfknoo\_1678035222 | "اما خب چندتا برداشت می‌شود این تصویر داشت نمی... | zatfknoo | 1678035222 |
| **1423** | zatfknoo\_1678035290 | "تصویر از رستوران اینجا که کلا اصلا هیچوقت دلم... | zatfknoo | 1678035290 |
| **1424** | zatfknoo\_1678035325 | "این تصویر منو در کتک خوردن انداخت این که این ... | zatfknoo | 1678035325 |

1425 rows × 4 columns

## preproccess

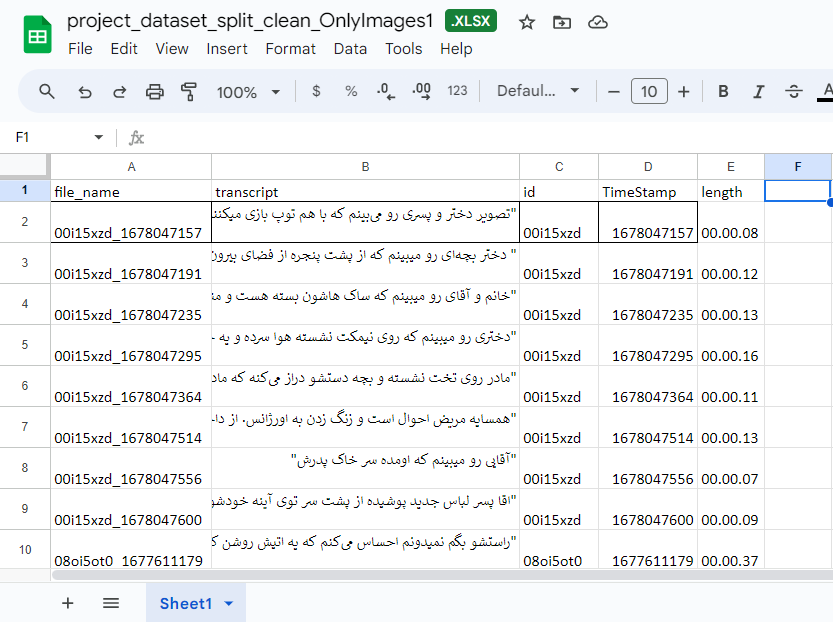
در بخش preproccess دادگانی که ناقص بودند پیدا شده و به صورت دستی آنهایی که ناقص بودند یا حذف و یا مجددا توسط آزمودنی ها کامل شدند. ( دیگران نیازی به این بخش ندارند و این قسمت برای دیتای خاص پروژه اینجانب انجام شده است.)

Transcriptهایی که به صورت اتوماتیک از طریق vosk به دست آمده بودند در این بخش بررسی شدند و دیتا تمیز شد. هشت تصویر در تسک وجود داشت. مشاهده شد کسانی که شمرده شمرده صحبت کرده بودند transcript بسیار خوبی از آنها در دسترس بود. کسانی که بخاطر ضعیف بودن اینترنت یا اشتباهی رد کردن یک عکس در مورد هفت تصویر صحبت کرده بودند و تنها یک ویس از آنها وارد سرور نشده بود هم وارد دیتاست شدند و متن آن یک فایل به صورت دستی وارد دیتاست شد. بعضی از فایلها جابجا شده بودند(ترتیب فایل های صوتی که مربوط به عکسها به هم خورده بود) آن ها هم درست شدند. کسانی که اشتباها فایل صوتی مربوط به یک عکس را ضبط نکرده بودند و در عکس بعدی در مورد عکس قبلی صحبت کرده بودند هم به صورت دستی درست شدند.

طول هر صدا نیز وارد دیتاست شد.

تعداد افراد باقی مانده 123 نفر است. تنها دادگان مربوط به هشت تصویر در یک فایل ذخیره شده و از این پس روی آن تحلیل خواهد شد. دیتاست به تعداد 984 سطر دارد که هر 8 سطر مربوط به یک آزمودنی است.

عکس پیوست شده تصویری از ستون های دیتاست در این مرحله را نشان میدهد.



985 rows × 7 columns

# 2.COMPUTE BDI

در بخش COMPUTE BDI میزان امتیازات پرسشنامه افسردگی بک محاسبه میشود. لازم به ذکر است که میزان امتیازات پرسشنامه های دیگر با فرمول های متفاوتی محاسبه می شوند که خودتان نیاز به پیاده سازی آن ها دارید.

در این بخش امتیازات BDI افراد محاسبه شد و به دیتاست اضافه شد تا بعدا بعنوان لیبل از آنها استفاده شود. به افرادی که پرسشنامه BDI را انجام نداده بودند یا بخاطر سرعت اینترنت پاسخنامه آنها وارد سرور نشده بود پیامی فرستاده شد تا آن را تکمیل کنند.در ادامه تکلیف پرسشنامه هایی که بیش از یک بار پاسخ داده شده اند مشخص خواهد شد.

پرسشنامه هایی که بیش از یک بار پاسخ داده شده بودند فقط دفعه اولشان نگه داشته شد و سایر پاسخنامه ها حذف شد. کسانی هم که خودشون به صورت دستی در پرسشنامه آیدی خود را تغییر داده بودند هم طبق زمان انجام پرسشنامه و آزمون صوتی شناسایی شدند و به دیتاست اضافه شدند.

ستونهای دیتاست هم اصلاح شدند. از این نظر که هر سطر از دیتاست مربوط به یک فایل صوتی است. هر 8 سطر مربوط به یک نفر بود که در مورد 8 فایل صوتی صحبت کرده بوند. اینها طوری گروه بندی شدند که در دیتاست جدید هر گروه فقط یک آیدی داشته باشد و طول صحبت ها با هم جمع زده شود و متن تمامی فایلهای صوتی با هم ادغام شدند.

دو بخش first block: read images csv and first BDI و second block:read only second BDI csv مربوط به این است که ابتدا محاسبه BDI برای یک سری دیتا انجام شده بود،بعدا که یک سری دیتا اضافه شد دوباره امتیازات BDI دادگان جدید محاسبه شده و به قبلی ها اضافه شده است. در صورتی که شما هم میخواهید به دیتاست خود دادگان جدید اضافه کنید این بخش به درد شما خواهد خورد.

# 3. categorization

این بخش برای زمانی استفاده میشود که شما به جای استفاده از رگرشن بخواهید از کلاسیفیکیشن استفاده کنید. در این صورت 6 کلاس 'normal', 'mild\_dep', 'borderline\_dep', 'moderate\_dep', 'severe\_dep','extreme\_dep' خواهید داشت.

Total Score : Levels of Depression

1-10 : These ups and downs are considered normal

11-16 : Mild mood disturbance

17-20 : Borderline clinical depression

21-30 : Moderate depression

31-40 : Severe depression

over 40 : Extreme depression

## groupby id

در بخش "groupby id" دادگان هر شخص که 8 سطر است در یک سطر ادغام شده اند(همه حرف هایشان در همه تصاویر در کنار هم قرار داده شده است) و طول کل صحبت هایی که هر نفر داشته نیز به عنوان یک ستون "total\_length" به دادگان اضافه شده است. و در بخش "vectorised groupbied data" کل صحبت های هر شخص در مورد هر هشت تصویر تبدیل به یک بردار شده است با پارس برت. دیتاست تبدیل به بردار شده دارای ابعاد (118, 773) است.

لازم به ذکر است که از این قسمت برای امتحان کردن روش های مختلف استفاده شده و به دلیل اینکه نتیجه خوب نبوده دیگر استفاده نشده است و راه حل دیگری استفاده شده است که در ادامه گفته خواهد شد.

# 4. vectorized - analysis

در بخش vectorised with parsbert دادگان اصلی که شامل صحبت های افراد به تفکیک هر تصویر است با پارس برت تبدیل به بردار شده اند.

! لازم به ذکر است که ادامه این بخش فقط برای تحلیل و آنالیز بیشتر و آشنا شدن با دادگان است و شما اصلا نیازی به اجرای آن ندارید. زیرا در یخش های بعد تحلیل هایی که نیاز هستند را انجام خواهید داد.

در بخش Similarit Marix از آن بردار ها استفاده شده برای رسم ماتریس شباهت. با توجه به اینکه ابعاد دادگان وکتورایز شده(768, 952) است ابعاد ماتریس شباهت (952, 952) می باشد.similarity matrix برای وکتورها رسم شد. سپس بر اساس امتیاز BDI ، طول زمانی که افراد صحبت کرده اند، بر اساس ترتیب عکس ها، و بر اساس امتیازات BDI کتگوری شده، sort شد.

یک سری آنالیز داده در این بخش انجام شده که در صورت دلخواه م یتوانید از آنها استفاده کنید.

# 5. similarity matrix

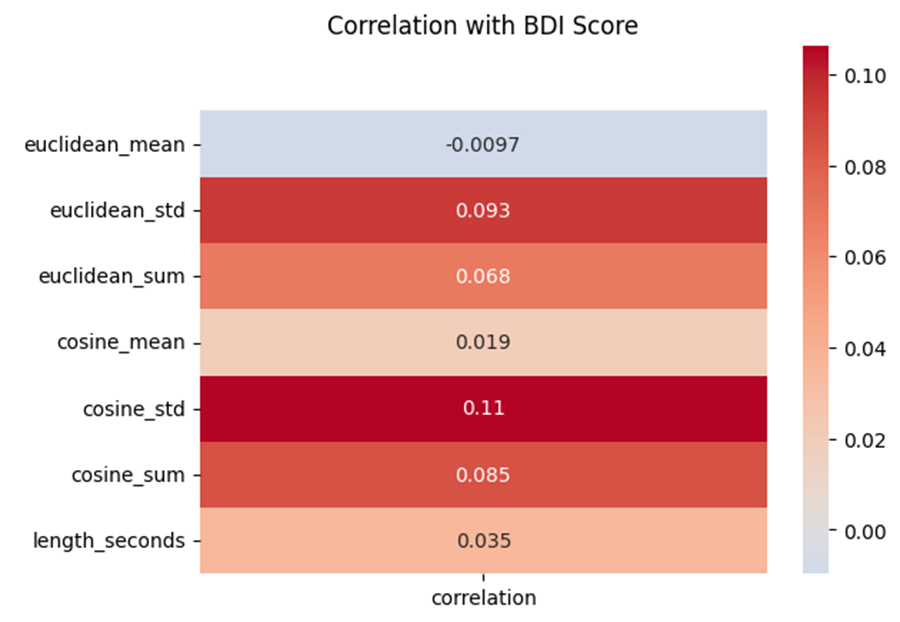
با در دست داشتن دادگان وکتورایز شده که ویژگی های زبانی آن نیز در دیتاست وجود دارد میتوان ماتریس شباهت را به ترتیب ویژگی های مورد نیاز حتی به تفکیک تصاویر تحلیلی و بررسی کرد. این کار در ادامه انجام خواهد شد.

میزان پرش فکری که آزمودنی­ها داشته‌اند نیز با استفاده از فاصله اقلیدسی و با استفاده از کسینوس زاویه بین بردار‌های معنایی تعبیه شده هر جمله در متون محاسبه شده و میانگین، انحراف معیار و مجموع فاصله‌ها به عنوان ویژگی استفاده شده‌اند. با داشتن n جمله در یک متن، n-1 معیار در آن متن خواهیم داشت. این به آن معنا است که برای متونی که فقط یک جمله دارند نمی‌توان این ویژگی را محاسبه کرد و مقدار nall به آن تعلق میگیرد.

در بخش sort by euclidean\_mean feature and then show the similarity matrix همانطور که از اسمش پیداست ابتدا بر اساس میانگین اقلیدسی دادگان مرتب شده اند بعد میزان cosine\_similarity بین بردارهای محاسبه شده است سپس ماتریس شباهت به دست آمده و رسم شده است.(دلیل کم شدن دیتا به 686در 686 این است که در دادگانی که تنها حاوی یک جمله بودند نمیتوان ویژگی های فاصله پرش فکری را محاسبه کرد و آنها حذف شده اند)

این عمل برای شش ویژگی فاصله پرش فکری انجام شده است. این 6 حالت شامل میانگین و انحراف معیار و مجموع فواصل در هر دو حالت فاصله اقلیدسی و فاصله کسینوسی می باشد.

از میزان هم‌بستگی هر یک از ویژگی‌های به دست آمده با امتیازات پرسش‌نامه افسردگی بک می‌توان به ویژگی‌های مهم دست یافت و در ادامه به آن‌ها پرداخت. حتی امتیازهای منفی ولی بزرگ نیز می‌توانند مهم باشند. این همبستگی زمانی که مقدارش از دو تقسیم بر رادیکال n بیشتر باشد قابل قبول است. ویژگی‌های زیر روی 119 نفر که هر یک دارای صحبت در مورد توصیف هشت تصویر هستند می‌باشند. پس در نتیجه ویژگی‌های زیر برای 119\*8=952 گفتار محاسبه شده‌اند. این به آن معنی است که همبستگی‌های بالای دو تقسیم بر رادیکال 952 که معادل است با مقدار 0.06 ، ارتباط معناداری با امتیازات پرسش‌نامه افسردگی بک دارند. مشاهده می‌شود که مقادیر میزان انحراف معیار و مجموع فاصله اقلیدسی و کسینوسی مقادیر قابل قبولی هستند. در بخش correlation میزان همبستگی هر یک از این ویژگی ها با میزان امتیازات پرسشنامه افسردگی بک محاسبه شده است که به شرح زیر است:



# 7. scatter plot

در این بخش برای بیشتر آشنا شدن با دادگان همان ماتریس شباهت زمانی که به دو بعد کاهش بعد داده میشود رسم شده است.

# 8. unsupervised

این بخش زمانی استفاده میشود که شما بخواهید بدون استفاده از دادگان پرسشنامه فقط روی دادگان صوتی تحلیل و نتیجه گیری انجام دهید. به دلیل اینکه نتیجه خوبی در نهایت از این بخش گرفته نشد در نتیجه ی نهایی از آن استفاده نشده است.

در این بخش پس از تبدیل داده های گروه بندی شده هر شخص(هر سطر مربوط به همه ی صحبت های یک شخص است و به تفکیک تصاویر جدا نشده اند) به بردار با استفاده از روش TF-IDF vectorization با روش kmeans و Hierarchical دادگان کلاستر میشوند.

# 9. feature extraction

پردازش زبان طبیعی یا NLP استخراج اطلاعات معنی دار از داده‌های زبان طبیعی و دستکاری ‌‌آن‌ها با استفاده از رایانه است. مدل‌های پردازش زبان طبیعی معمولا مبتنی بر زبان انگلیسی هستند. برای زبان‌های دیگر از مدل‌های چندزبانه با منابع محدود استفاده می‌شود. ‌پارس‌برت یک برت تک زبانه را برای زبان فارسی پیشنهاد می‌دهد. ‌پارس‌برت روی یک دیتای عمومی‌حجیم (Persian Wikidumps, MirasText) و داده‌های وبسایت‌های زیادی آموزش دیده شده. بخشی از متودولوژی ‌پارس‌برت این است که پیش‌پردازشی گسترده برای آوردن مجموعه داده‌ها به یک قالب مناسب انجام داده است. مدل ‌پارس‌برت به‌طور خاص بر روی واژگان خاص زبان با مقادیر انبوهی از داده‌های متن فارسی آموزش داده شده است.

## sentiment analysis

یکی از ویژگی‌های مورد استفاده، احساسات است. احساسات متن به دو روش محاسبه شده و از آن‌ها به عنوان ویژگی در ‌پیش‌بینی استفاده شده است. در روش اول یک تابع تعریف شده که با گرفتن یک رشته از کلمات به زبان فارسی با استفاده از کتابخانه hazm عملیات توکنیزه کردن و نشاندار کردن یک رشته کاراکتر به کوچکترین واحدی که در یک زبان قابل اهمیت است انجام می‌شود. بعد با استفاده از مدل طبقه بندی متن میزان امتیاز احساسات هر متن ‌پیش‌بینی می‌شود. در روش دوم تابع خط لوله از کتابخانه‌ترنسفورماتورها یک مدل یادگیری ماشین از پیش آموزش دیده شده را بارگزاری کرده و از آن برای تسک پردازش طبیعی تحلیل احساسات استفاده کرده است. احساسات عددی بین صفرو یک است که احتمال مثبت بودن یک متن را نشان میدهد. جایی که امتیاز احساسات نزدیک به یک باشد نشان دهنده احساسات مثبت‌تر و نزدیک به صفر نشان دهنده احساسات منفی‌تر است. اگر این میزان نزدیک به نیم باشد به این معنی است که مدل نمی‌تواند با اطمینان در مورد احساسات متن نظر دهد.

## Speeh Graph Analysis

با استفاده از گراف گفتار که در آن هر کلمه به عنوان یک گره و ارتباط بین کلمات متوالی به عنوان یک یال نمایش داده می‌شوند یک سری ویژگی از متون استخراج شده است. ویژگی­های استخراج ­شده به شرح زیر هستند:

ویژگیهای زبانی شامل ویژگی‌های عمومی[1] مانند تعداد کل کلمات[2]، تعداد گره‌های گراف[3] و تعداد ارتباطات بین کلمات[4] می‌باشند. ویژگی‌های اجزاء متصل[5] مانند تعداد گره‌های بزرگترین جزئیات متصل یعنی بزرگترین زیرگراف که در آن همه کلمات با هم مرتبطند در گراف بدون جهت[6] و تعداد گره‌های بزرگترین جزئیات قوی متصل در گراف جهت دار[7] است.

ویژگی‌های بازگشتی[8] مانند مجموع تمام یالهای ساده یعنی لینک‌های تکرار شده[9] و لینک‌های موازی[10] که یک جفت گره را به هم مرتبط می‌کنند و ماتریس مجاورت بین کلمات[11] و نصف مجذور ماتریس مجاورت[12] هستند.

همچنین ویژگی‌های جهانی[13] مانند ، چگالی گراف[14]، قطر یا طول بزرگترین کوتاهترین مسیر بین در دو نود در گراف[15]، ضریب خوشه بندی[16]، متوسط کوتاه‌ترین مسیر بین هر دو نود در شبکه[17]، میانگین کل درجات[18] و .. محاسبه شده است.

[1] general attributes

[2] Word count

[3] nodes

[4] edges

[5] connected components

[6] Largest connected components

[7] Largest strongly connected components

[8] recurrence attributes

[9] Related edges

[10] Parallel edges

[11] L1

[12] L2

[13] global attributes

[14] density

[15] diameter

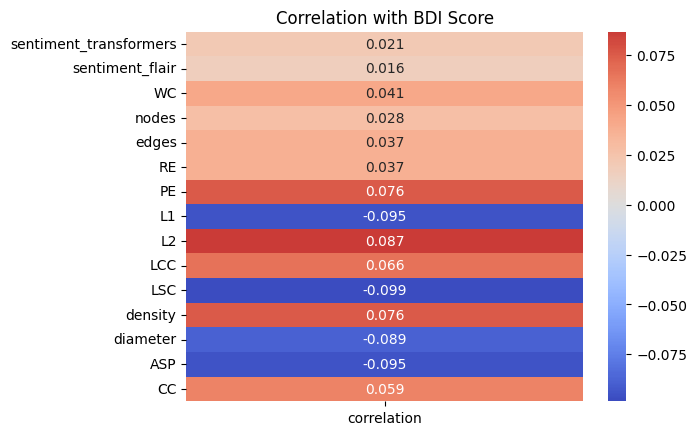
[16] clustering coefficient

[17] average shortest path

[18] average total degree

## correlation

در این بخش میزان همبستگی هر یک از ویژگی های استخراج شده با BDI-score محاسبه شده است.از میزان هم‌بستگی هر یک از ویژگی‌های به دست آمده با امتیازات پرسش‌نامه افسردگی بک می‌توان به ویژگی‌های مهم دست یافت و در ادامه به آن‌ها پرداخت. حتی امتیازهای منفی ولی بزرگ نیز می‌توانند مهم باشند. این همبستگی زمانی که مقدارش از دو تقسیم بر رادیکال n بیشتر باشد قابل قبول است. ویژگی‌های زیر روی 119 نفر که هر یک دارای صحبت در مورد توصیف هشت تصویر هستند می‌باشند. پس در نتیجه ویژگی‌های زیر برای 119\*8=952 گفتار محاسبه شده‌اند. این به آن معنی است که همبستگی‌های بالای دو تقسیم بر رادیکال 952 که معادل است با مقدار 0.06 ، ارتباط معناداری با امتیازات پرسش‌نامه افسردگی بک دارند. شکل زیر میزان همبستگی هر یک از سایر ویژگی‌ها با امتیازات پرسش‌نامه افسردگی بک را نشان میدهد. مشاهده می‌شود که ویژگی‌های PE، L1، L2، LCC، LSC، چگالی و قطر گراف گفتار و ASP ویژگی‌های قابل قبولی هستند که مقدار همبستگی آن‌ها با امتیازات پرسش‌نامه افسردگی بک از یک حدی بیشتر است.



ولی ما به این بسنده نکردیم و پس از آن به تفکیک هر تصویر میزان همبستگی ویژگی‌های پرش ذهنی با امتیازات پرسش‌نامه افسردگی بک محاسبه شده است. در این صورت می‌توان همسبتگی‌های جالبی بین ویژگی‌های درون هر تصویر به صورت خاص مشاهده کرد. در بخش‌های بعد این عمل با سه روش توکن‌محور، جمله‌محور و میانگین تعبیه جملات محور به تفکیک هر تصویر انجام شده و مقدار همبستگی و میزان قابل قبول بودن آن نیز به دست آورده ایم و در جداولی در ادامه نوشته شده است.

# 10. SVM - RF- DT - balance - LOO

در این بخش سه روش svm و random forest و decision tree پیاده سازی شده و

دادگان بالانس شده

از روش leave one out استفاه شده.

## TFIDF with 2 class

ابتدا در این بخش دادگان به دو کلاس نرمال و افسرده تقسیم شده که بین دو کلاس بالانس برقرار باشد و تعداد دادگان در یک کلاس خیلی کم یا خیلی زیاد نباشد. البته با این کار انگار یک سری اطلاعات دور ریخته میشوند زیرا میزان افسردگی دیگر در نظر گرفته نشده است. بعد روی بردارهایی که با TFIDF به دست آمده اند کاهش بعد انجام شده و بعد از پیاده سازی LeaveOneOut روی مدل svm میزان معیارهای ارزیابی و Confusion Matrix و ROC Curve و Precision-Recall Curve نمایش داده شده اند.

## Scatter Plot between features:

در این بخش اسکتر پلات بین ویژگی های مختلف به دست آمده تا ملموس تر بشد و بتوان تحلیل کرد تا ویژگی های نامناسب را حذف کرد.

# TFIDF with 63 class \_ balance

بالانس کردن در این بخش با تقسیم دادگان به دو کلاس انجام نمیشود بلکه با حذف دادگان از کلاسهایی که تعداد زیادی داده دارند انجام میشود تا در هر کلاس میزان یکسانی داده قرار داشته باشد. البته مشخص است که این روش ممکن است ممناسب نباشد و این مراحل فقط برای آزمایش راه حل های مختلف پیاده سازی شده بودند. که نتایج خوبی نداشتند.

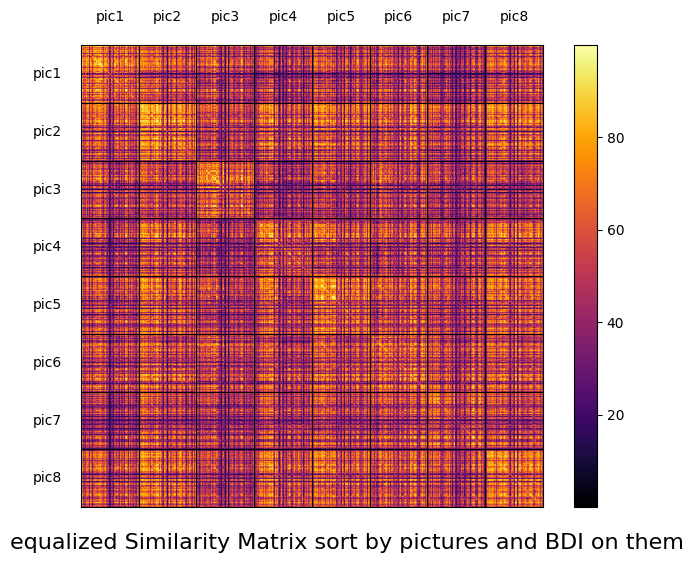
# visualized similarity matrix

در این بخش رسم ماتریس شباهت انجام شده است. درون صحبت های هر تصویر نیز به ترتیب BDI-score مرتب شده است. با استفاده از percentile میتوان ویژوالیزیشن بهتری از کاتریس شباهت داشت.

به منظور تحلیل بیشتر از ماتریس شباهت استفاده شده تا تحلیل انجام شود. تحلیل و تفسیر دادگان در فضای ماتریس عدم تشابه بازنمایی[ representational dissimilarity matrix] انجام شده است تا شباهت تصاویر و ارتباط با امتیازات پرسش‌نامه افسردگی بک ملموس‌تر باشد. ماتریس شباهت از روی بردارهای به دست آمده برای هر متن بین گفتارها محاسبه شده است و بر اساس امتیازات پرسش‌نامه افسردگی بک مرتب سازی شده است تا بتوان ارتباطات معنادار را مشاهده کرد. این ماتریس بر اساس تصاویر به صورت مجزا نیز رسم شده است. رسم آن نیز بر حسب درصد[ percentile] می‌باشد تا بتوان درک بهتری از آن داشت. تبدیل یک ماتریس شباهت به یک ماتریس صدک می‌تواند برای درک بهتر و مقایسه شباهت بین مجموعه‌های مختلف داده مفید باشد. درصدها راهی برای رتبه‌بندی مقادیر نسبت به یکدیگر در یک توزیع فراهم می‌کنند که می‌تواند به شناسایی نقاط پرت یا الگوهایی در داده‌ها کمک کند که ممکن است بلافاصله با مشاهده نمرات شباهت خام آشکار نشوند.

با داشتن ماتریس تشابهی که نشان دهنده شباهت بین بردار‌های مختلف است، تبدیل آن به صدک می‌تواند کمک کند تا تشخیص دهید بردار‌های مربوط به کدام تصاویر به طور خاص به یکدیگر مرتبط هستند (یعنی مقادیر صدک بالایی دارند) و کدامیک به هم نامرتبط‌تر هستند. (یعنی مقادیر صدک کمتری دارند). این می‌تواند برای اهداف طبقه بندی یا برای شناسایی روابط تکاملی بین تصویر‌های مختلف مفید باشد.

در ماتریس مجاورت هر سطر و هر ستون نشان دهنده یک گفتار هستند. شباهت بردارهای معنایی سطر‌های ماتریس مجاورت با ستونهای ماتریس مجاورت، با رنگ درون هر سلول مشخص شده‌اند. قطر اصلی ماتریس مجاورت تماما یک می‌باشد به این معنی که هر گفتار با خودش بیشترین شباهت را دارد. ماتریس مجاورت یک ماتریس مربعی است که ابعاد آن به تعداد گفتارها در تعداد گفتار‌ها می‌باشد. میزان همبستگی بین ویژگی‌ها و امتیازات پرسش‌نامه افسردگی بک به صورت عادی و مرتب نشده در شکل های رسم شده (مانند شکل زیر) نمایش داده شده است.



هر یک از این بخش ها نیز میتواند از نزدیک تر نمایش داده شود که کد آن در بخش sub plots نوشته شده است.

# 11. jump of idees & correlation:

هر شخص در مورد هشت تصویر صحبت کرده است. پس در ماتریس شباهت اطلاعات هر آزمودنی در واقع یک ماتریس هشت در هشت میباشد. عناصر قطر اصلی که شباهت هر گفتار با خودش را نشان میدهد همه یک هستند و عناصر بالا و پایین قطر اصلی نیز یکسان هستند. به همین دلیل تنها از 28 مقداری که در بالای قطر اصلی است استفاده شده است.ماتریس شباهت برای هر شخص به صورت مجزا که فقط از عناصر بالای قطر اصلی استفاده می‌شود چون عناصر زیر قطر اصلی تکراری هستند و عناصر روی قطر اصلی تماما یک هستند.

وقتی که با اطلاعات ماتریس‌های 8 در 8 افراد با ویژگی‌های مربوط به فاصله اقلیدسی و کسینوسی هر یک به صورت مجزا میزان امتیازات پرسش‌نامه افسردگی بک ‌پیش‌بینی می‌شود سه حالت داریم:

1. حالت توکن‌محورtocken based embedding

2. حالت جمله‌محور sentence based embedding

3. حالت میانگین تعبیه جملات محور mean word based embedding

در هر یک از این سه حالت کارهای یکسانی انجام شده فقط نحوه تبدیل به بردار متفاوت است. در هر بخش بعد از به دست آوردن فواصل سینوسی و کسینوسی بردارهای معنایی ویژگی های مجموع و انحراف معیار و میانگین به دست آمده است، سپس میزان همبستگی هر ویژگی با BDI-score محاسبه شده است. مقدار significant در آن تعداد داده نیز محاسبه شده است و نمایش داده شده است.

# 12. percentile similarity matrix :

در این بخش برای سه حالت زیر رسم ماتریس مجاورت بر حسب درصد انجام شده است:

1. حالت توکن‌محورtocken based embedding

2. حالت جمله‌محور sentence based embedding

3. حالت میانگین تعبیه جملات محور mean word based embedding

به منظور تحلیل بیشتر از ماتریس شباهت استفاده شده تا تحلیل انجام شود. تحلیل و تفسیر دادگان در فضای ماتریس عدم تشابه بازنمایی[ representational dissimilarity matrix] انجام شده است تا شباهت تصاویر و ارتباط با امتیازات پرسش‌نامه افسردگی بک ملموس‌تر باشد. ماتریس شباهت از روی بردارهای به دست آمده برای هر متن بین گفتارها محاسبه شده است و بر اساس امتیازات پرسش‌نامه افسردگی بک مرتب سازی شده است تا بتوان ارتباطات معنادار را مشاهده کرد. این ماتریس بر اساس تصاویر به صورت مجزا نیز رسم شده است. رسم آن نیز بر حسب درصد[ percentile] می‌باشد تا بتوان درک بهتری از آن داشت. تبدیل یک ماتریس شباهت به یک ماتریس صدک می‌تواند برای درک بهتر و مقایسه شباهت بین مجموعه‌های مختلف داده مفید باشد. درصدها راهی برای رتبه‌بندی مقادیر نسبت به یکدیگر در یک توزیع فراهم می‌کنند که می‌تواند به شناسایی نقاط پرت یا الگوهایی در داده‌ها کمک کند که ممکن است بلافاصله با مشاهده نمرات شباهت خام آشکار نشوند.

با داشتن ماتریس تشابهی که نشان دهنده شباهت بین بردار‌های مختلف است، تبدیل آن به صدک می‌تواند کمک کند تا تشخیص دهید بردار‌های مربوط به کدام تصاویر به طور خاص به یکدیگر مرتبط هستند (یعنی مقادیر صدک بالایی دارند) و کدامیک به هم نامرتبط‌تر هستند. (یعنی مقادیر صدک کمتری دارند). این می‌تواند برای اهداف طبقه بندی یا برای شناسایی روابط تکاملی بین تصویر‌های مختلف مفید باشد.

در ماتریس مجاورت هر سطر و هر ستون نشان دهنده یک گفتار هستند. شباهت بردارهای معنایی سطر‌های ماتریس مجاورت با ستونهای ماتریس مجاورت، با رنگ درون هر سلول مشخص شده‌اند. قطر اصلی ماتریس مجاورت تماما یک می‌باشد به این معنی که هر گفتار با خودش بیشترین شباهت را دارد. ماتریس مجاورت یک ماتریس مربعی است که ابعاد آن به تعداد گفتارها در تعداد گفتار‌ها می‌باشد.

# 13. prediction SVM:

لازم به ذکر است که این بخش برای تحلیل و آزمایش بیشتر انجام شده و نتایج اصلی در بخش بعد به دست آمده اند.

هر شخص در مورد هشت تصویر صحبت کرده است. پس در ماتریس شباهت اطلاعات هر آزمودنی در واقع یک ماتریس هشت در هشت میباشد و از همین به عنوان ویژگی برای مدل استفاده شده است. عناصر قطر اصلی که شباهت هر گفتار با خودش را نشان میدهد همه یک هستند و عناصر بالا و پایین قطر اصلی نیز یکسان هستند. به همین دلیل تنها از 28 مقداری که در بالای قطر اصلی است استفاده شده است.ماتریس شباهت برای هر شخص به صورت مجزا که فقط از عناصر بالای قطر اصلی استفاده می‌شود چون عناصر زیر قطر اصلی تکراری هستند و عناصر روی قطر اصلی تماما یک هستند.

وقتی که با اطلاعات ماتریس‌های 8 در 8 افراد با ویژگی‌های مربوط به فاصله اقلیدسی و کسینوسی هر یک به صورت مجزا میزان امتیازات پرسش‌نامه افسردگی بک ‌پیش‌بینی می‌شود سه حالت داریم:

1. حالت توکن‌محورtocken based embedding

2. حالت جمله‌محور sentence based embedding

3. حالت میانگین تعبیه جملات محور mean word based embedding

در هر یک از این سه حالت کارهای یکسانی انجام شده فقط نحوه تبدیل به بردار متفاوت است

از ویژگی های فاصله کسینوسی و اقلیدسی به صورت مجزا برای پیش بینی استفاده شده است.(euclidean و cosine). هر یک از بخش های این قسمت بر اساس نامی که دارند مشخص است که چه پیاده سازی ای در آنها انجام شده است.

# 14. prediction : Linear regression

در این بخش از مدل رگرسیون خطی، برای ‌پیش‌بینی استفاده شده است و عملکرد ‌پیش‌بینی شدت افسردگی گزارش شده بر حسب ضریب تعیین[R-squared] و ریشه میانگین مربعات خطا[Root Mean Squared Error] ارزیابی شده است.

رگرسیون خطی یک روش آماری است که برای مدل‌سازی رابطه بین یک متغیر وابسته و یک یا چند متغیر مستقل استفاده می‌شود. هدف رگرسیون خطی یافتن بهترین خطی است که این رابطه را توصیف می‌کند و به ما امکان می‌دهد تا بر اساس مقادیر متغیرهای مستقل، در مورد متغیر وابسته پیش‌بینی کنیم. ایده اصلی پشت رگرسیون خطی این است که نقاط داده را روی یک نمودار پراکنده رسم کنیم و سپس یک خط مستقیم را در میان آن‌ها قرار دهیم که فاصله بین خط و نقاط داده را به حداقل برساند. سپس می‌توان از این خط برای پیش‌بینی مقادیر آینده متغیر وابسته بر اساس مقادیر جدید متغیرهای مستقل استفاده کرد. مدل‌های رگرسیون خطی در طیف وسیعی از کاربردها از جمله برای شناسایی روندها و روابط بین متغیرها استفاده می‌شوند که می‌تواند به تصمیم‌گیری در زمینه‌های مختلف کمک کند.

همانطور که در بخش های قبل گفته شده هر شخص در مورد هشت تصویر صحبت کرده است. پس در ماتریس شباهت اطلاعات هر آزمودنی در واقع یک ماتریس هشت در هشت میباشد و از همین به عنوان ویژگی برای مدل استفاده شده است. عناصر قطر اصلی که شباهت هر گفتار با خودش را نشان میدهد همه یک هستند و عناصر بالا و پایین قطر اصلی نیز یکسان هستند. به همین دلیل تنها از 28 مقداری که در بالای قطر اصلی است استفاده شده است.ماتریس شباهت برای هر شخص به صورت مجزا که فقط از عناصر بالای قطر اصلی استفاده می‌شود چون عناصر زیر قطر اصلی تکراری هستند و عناصر روی قطر اصلی تماما یک هستند.

وقتی که با اطلاعات ماتریس‌های 8 در 8 افراد با ویژگی‌های مربوط به فاصله اقلیدسی و کسینوسی هر یک به صورت مجزا میزان امتیازات پرسش‌نامه افسردگی بک ‌پیش‌بینی می‌شود سه حالت داریم:

1. حالت توکن‌محورtocken based embedding از حالت tf-idf برای embedding استفاده شده است.

2. حالت جمله‌محور sentence based embedding از حالت labse برای embedding استفاده شده است.

3. حالت میانگین تعبیه جملات محور mean word based embedding از حالت parsbert برای embedding استفاده شده است.

برای هر یک از این سه حالت پیش بینی با روش Linear regression به تفکیک دو ویژگی (euclidean و cosine) انجام شده است.