# به نام خدا

**عنوان محتوا : راهنمای ساخت و جمع آوری داده (DataSet) برای پروژه های صوت به متن (Speech To Text)**

**گردآورنده : سید محمد مسعود پرپنچی**

**تاریخ گردآوری : بهار ۱۳۹۹**

# فهرست مطالب

## مقدمه :

**پروژه های یادگیری عمیق نیاز به مجموعه داده با حجم بالا دارند. از این رو برای شروع یک پروژه یادگیری عمیق و ساخت یک محصول باید داده مناسب جمع آوری شود.**

**برای این منظور این مطلب گردآوری شد. دراین فایل سعی بر این شده است منابع موجود در سطح اینترنت بررسی شوند و نکات مناسب ساخت داده یا جمع آموری داده تا حد امکان ذکر شود.**

**توجه : این فایل نهایی نیست و همواره تلاش بر بروزرسانی آن هست.**

## جمع بندی مطالب ارائه شده در منابع مختلف  برای زبان فارسی( به زودی ):

**با بررسی تمام نکات موجود ( که در صفحات بعدی آنهارا مشاهده خواهید کرد) جمع بندی نهایی برای ساخت داده برای زبان فارسی با نکات مهم آن را برای  شما در این بخش ذکر خواهیم کرد.**

# نکات ارائه شده برای ساخت دیتاست در microsoft :

**ابتدا ذکر منبع :** https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-test-and-train

https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-human-labeled-transcriptions

**در منبع اول از نکات مایکروسافت, در مورد داده های صوتی صحبت شده است و جدول پایین به ادامه نکات کمک میکند :**

| نوع داده | استفاده برای تست ؟ | حجم مناسب برای تست | استفاده برای train؟ | حجم مناسب برای train |
|---|---|---|---|---|
| صوت برای تست | بله | حداقل ۵ فایل صوتی | خیر | --- |

| | | | | |
|---|---|---|---|---|
| صوت + متن | بله برای سنجیدن دقت مدل | از نیم ساعت تا پنج ساعت | بله | از ده ساعت تا هزار ساعت |
| متن مرتبط | خیر | --- | بله | از یک مگابایت تا دویست مگابایات متن مرتبط |

**حال به توضیح مطالب جدول میپردازیم :**

**- داده صوتی برای تست یک نیاز حیاتی برای سنجیدن مدل میباشد جدول پایین ویژگی های مناسب برای صوت تستی را میگوید**

| ویژگی | مقدار |
|---|---|
| فرمت فایل | Wav |
| Sample rate | ۸۰۰۰ – ۱۶۰۰۰ هرتز |
| کانال | ۱ (mono) |
| حداکثر طول صوت | ۲ ساعت |
| Sample format | Pcm – 16bit |

**- متن + صوت :**

**صوت به همراه متن مورد نظر حیاتی ترین داده برای سیستم ما میباشد. ساختن متن مناسب برای صوت ها ممکن است وقت گیر باشد ولی دقت مدل به کیفیت این داده وابسطه است.**

| ویژگی | مقدار |
|---|---|
| فرمت فایل | Wav |
| Sample rate | ۸۰۰۰ – ۱۶۰۰۰ هرتز |
| کانال | ۱ (mono) |
| حداکثر طول صوت | ۲ ساعت برای تست / ۱ دقیقه برای یادگیری |
| Sample format | Pcm – 16bit |

**میزان ده ساعت الی هزار ساعت برای یادگیری عددی کاملا حیاتی هست که باید رعایت شود.**

**متن هایی که برای صوت ها ساخته میشود. اصلی ترین داده ما هستند. باید با دقت و کلمه به کلمه ساخته شوند. بعضی پیش پردازش ها برای آنها حیاتی هست که در بخش بعدی آنهارا ذکر میکنیم.**

**- متن مرتبط : ( نیاز به مطالعه بیشتر)**

**با توجه به چیزی که از مستندات یافت میشود, برای آنکه مدل ما نتیجه بهتری بر روی اسامی شرکت ها, نام ها , اسامی خاص, و کلمات کم تکرار داشته باشد. میبایست ما متن هایی که این نوع کلمات در آنها میباشد را در جملات ذکر کنیم. برای شخصی**

**سازی مدل و محصول برای صنایع مختلف و بازار هیا هدف متنوع این فایل متفاوت میشود.**

**برای ساخت این متن ها دو نوع متن خواهیم ساخت.**

**– مورد اول : باید متن هایی از سخنرای ها داشته باشیم. نیازی نیست که از نظر قواعد دستور زبانی درست باشند, اما باید دقیقا مطابق چیزی که بیان میشود باشند.**

**کلمات و اصطلاحات خاصی که در بازار هدف هستند باید در این متن جای بگیرند.**

**نکات ساخت این متن ها :**

– **هیچ حرفی را بیشتر از ۴ بار تکرار نکنید. مثلا :** "دیروز در راه مدرسه آآآآ بودم "

**– مورد دوم : ساخت متن هایی که نوع نشوتن کلمات آنها با خواندن آنها متفاوت است مانند :**

| Recognized/displayed form | Spoken form |
|---|---|
| 3CPO | three c p o |
| CNTK | c n t k |
| IEEE | i triple e |

**حال به بررسی منبع دوم از مستندات مایکروسافت میپردازیم که به پیش پردازش های مناسب برای ساخت متن هست :**

**توجه : این نکات ذکر شده در وبسایت مایکروسافت برای زبان های آلمانی و انگلیسی ذکر شده است.**

**چهار نکته ذکر شده در این بخش :**

**۱- اعداد با فرمت غیر استاندارد را به صورت متنی بنویسید. ( در کل اعداد را به صورت متنی بنویسید)**

Determine if each of the following numeric data standard or nonstandard data

| | |
|---|---|
| 345.12 | Standard |
| $345.12 | Nonstandard |
| 3,456.12 | Nonstandard |
| 20DEC2010 | Nonstandard date |
| 12/20/2010 | Nonstandard date |

**۲- حروف و عبرات خاصی که جزو الفبا نیستند باید به همان صورتی که خوانده میشوند نوشته شوند : مانند : $ باید بشود دلار.**

**۳- برای کلمات مخفف انگلیسی با حروف بزرگ به همان انگلیسی نوشته شوند.**RAM, NASA

**۴- کلمات مخفف انگلیسی که به صورت تک حرف تک حرف خوانده میشوند باید به همان صورت تک حرف با space از هم جدا شوند.** CPUبشود C P U

**چند مثال به زبان انگلیسی :**

| Original text | Text after normalization |
|---|---|
| Dr. Bruce Banner | Doctor Bruce Banner |
| James Bond, 007 | James Bond, double oh seven |
| Ke$ha | Kesha |
| How long is the 2x4 | How long is the two by four |
| The meeting goes from 1-3pm | The meeting goes from one to three pm |
| My blood type is O+ | My blood type is O positive |
| Water is H20 | Water is H 2 O |
| Play OU812 by Van Halen | Play O U 8 1 2 by Van Halen |
| UTF-8 with BOM | U T F 8 with BOM |

**۵- تمام علایم نگارشی حذف شوند**

**۶- اعداد را به همان صورتی که میخوانیم بنویسیم.**

**مثال**

| Original text | Text after normalization |
|---|---|
| "Holy cow!" said Batman. | holy cow said batman |
| "What?" said Batman's sidekick, Robin. | what said batman's sidekick robin |
| Go get -em! | go get em |
| I'm double-jointed | I'm double jointed |
| 104 Elm Street | one oh four Elm street |
| Tune to 102.7 | tune to one oh two point seven |
| Pi is about 3.14 | pi is about three point one four |
| It costs $3.14 | it costs three fourteen |

**حال بعضی از نکات که برای زبان آلمانی ذکر شده است را بررسی میکنیم :**

**۱- برای ذکر اعداد از حروف استفاده کنید. هر صورت از اعداد کسری اعشاری ...**

**۲- برای ذکر علایم ریاضیاتی مثل + – = < > از کلمات معادل آنها استفاده کنید.**

## نکات ارائه شده برای ساخت دیتاست در IBM Watson :

**منبع :** https://cloud.ibm.com/docs/speech-to-text?topic=speech-to-text-corporaWords#workingCorpora

**در مستندات IBM بیشتر به نحوه ساخت متن ها اشاره شده است. لازم به ذکر است در این مستندات نیز به زبان های لاتین و شرق آسیا تمرکز شده است.**

**۱- برای نمایش اعداد از حروف استفاده کنید.**

**۲- عباراتی که معنی خاصی را منتقل میکنند را به صورت کلمه در بیاورید.**
**مانند سمبل دلار, یورو, درصد و غیره..**

**۳- عباراتی که داخل پرانتز گیومه و .. هستند و خوانده نمیشوند را داخل متن نیاورید.**

**۴- علائم نگارشی را حذف کنید. مانند نقطه ویرگول و ...**

## نکات ارائه شده برای ساخت دیتاست در Mozilla Deep Speech

## (این بخش همواره در حال بروزرسانی میباشد).

**این سامانه به دلیل** open source **بودن یک فروم بسیار بزرگ دارد که تمام نکات از آنجا استخراج شده است. منبع :** https://discourse.mozilla.org/c/deep-speech

**ابتدا از منبع نکاتی را در مورد تجربه این فرد در ساخت مدلی میبنیم :**

**مشخصات صوت ثبت شده : ۱۶۰۰۰ کیلوهرتز / mono/ ۱۶ بیت**

**فرمت متن ها :** utf-8

**توجه : نحوه قرار دادن فایل ها و دایرکتوری ها حتما از مستندات** Mozillaبررسی شود.

**رعایت نکات برای رعایت تنوع و استاندارد بودن ضبط صدا :**

**- ضبط صدا با داشتن فاصله های متنوع از دستگاه ضبط صوت**

**- ضبط صدا در هنگام حرکت**

**- ضبط صدا در زمان های مختلف روز . بعد از خواب / بعد از غذا / هنگام غدا / قبل خواب/ به هنگام هیجان شادی ناراحتی/ داشتن نویز /**

**- برای رعایت فاصله های مختلف از نوقعیت های مختلف نسبت به دستگاه صدا را ضبط کنید**

**- حالت صدای خود را تغییر دهید و صدا های شاد و ناراحت بسازید.**

**- رعایت تنوع صدا. نباید صدا ها را تماما یک نفر بگوید**

**- بررسی اینکه مدل در چه حالت هایی با مشکل موجه میشود و ضبط صداهای بیشتر برای آن موقعیت خاص.**

**- شما میتوانید از https://github.com/mozilla/voice-corpus-tool برای تغییر صوت ها و فایل های csv استفاده کنید.**

**- سایر منابع :**

**با مراجعه به فروم موزیلا میتوانیم نکات دیگری را بیابیم :**

**۱- میزان صوت ها باید زیاد باشد. مثلا ۷۰ ساعت صوت کافی نیست.**

## نکات ارائه شده  برای ساخت دیتاست در مقالات Medium

**در ابتدا خواندن نوشته زیر حتمی و حیاتی هست , در این نوشته نحوه ساختن دیتاست از کتاب های صوتی را یاد داده است :** https://medium.com/@klintcho/creating-an-open-speech-recognition-dataset-for-almost-any-language-c532fb2bc0cf

**در این بخش مقالاتی که استفاده کنندگان سرویس های مختلف و توسعه دهندگان به اشتراک گذاشته اند را میبینیم نکاتشان را میبینیم.**

https://medium.com/ibm-watson/watson-speech-to-text-how-to-train-your-own-speech-dragon-part-1-data-collection-and-fdd8cea4f4b8

**در نوشته بالا نکات بسیار مهمی ذکر شده است :**

**۱- داده های ما نیاز هست که شبیه بازار هدف باشد. به عنوان مثال اگر کاربران ما از گوشی همراه در کارخانه از این سامانه استفاده میکنند مهم هست که داده های آموزشی ما شبیه شرایط هدف باشد.**

**۲-  پیشنهاد این هست که حداقل ۱۰ تا  ۵۰ ساعت صوت مشابه بازار هدف داشته باشیم.**

**۳- ویژگی های بسیار مهم در بازار هدف عبارتند از :**

a. **جغرافیای مورد نظر. ایران تاجیکستان افغانستان ...**
b. **لهجه های متفاوت**
c. **دستگاه ها : گوشی موبایل / لپتاپ/ ...**
d. **محیط اطراف : کارخانه/ مدرسه / بیمارستان / ...**

e. طول زمان صحبت

۴- از داده های مکالمه های تلفنی نیز میتوانید استفاده کنید. ( البته من خودم فکر میکنم به خاطر تفاوت sample rate مشکلاتی بوجود بیاورد. )

۵- جدول زیر بسیار مهم هست. نحوه رعایت نکات بالا در آن ذکر شده است. نکات زیر بسته به نیاز هر پروژه تغییر میکند.

| | |
|---|---|
| Accents:<br>• 75% Plain English<br>• 15% Hispanic English<br>• 5% Indian English<br>• 5% Other English | Environment:<br>• 70% no background noise<br>• 20% traffic background noise<br>• 10% TV background noise |
| Devices:<br>• 70% Land line<br>• 30% Cell phone | Jargons:<br>• LNG<br>• Wildcat well |

برای جمع آوری داده چه کنیم :

شما باید گوینده های متنوعی داشته باشید. هرچه متنوع تر بهتر. شرایط آنها نیز باید با هم تفاوت کند مانند : لهجه آنها/ مخیط آنها/ دستگاه آنها/ ...

نکته مهم : تمام اصوات جمع آوری شده باید بصورت نظارت انسانی transcribe بشوند.جتی اگر متن آنها را داریم. یعنی متن آنها نوشته بشود به دلایل زیر :

۱- گوینده ها بسیار از عباراتی مانند 'اممم' استفاده میکنند

۲- گوینده ها بخش هایی از یک متن را ممکن است دو بار بخوانند

۳- گوینده ها ممکن است بعضی از بخش ها را نخوانند.

۴- گوینده ها ممکن است از خودشان چیزی به متن اضافه کنند.

۵- گوینده ها ممکن است متن را کوتاه تر کنند

**در مرحله آخر خم بسیار حیاتی است که توزیع مناسبی از هر لهجه/ محیط/ دستگاه و...**

**در هر فایل** train **و** test **باشد.**