به نام خدا

گزارش دو هفتگی پروژه تبدیل صوتی به متن

محمدمهدی برقی

فعالیت اصلی بنده در این دو هفته، امکان سنجی استقاده از کتاب‌های صوتی به عنوان دیتاست مورد استفاده برای پروژه بود

کتاب های صوتی زیادی وجود دارند که می توانند دیتاست بسیار مناسبی برای پروژه باشند اما این صوت ها اکثرا چهار مشکل بزرگ برای تیم ما دارند.

اولین و بزرگترین مشکل استخراج متن این کتاب‌های صوتی و یافتن متن آنها بود زیرا:

- اکثر اپلیکیشن‌ها و انتشارات مربوط به کتاب‌های صوتی، به دلیل حفظ حقوق ناشران کتاب ها از در اختیار قرار دادن محتوای متنی این کتب با فرمت مناسب و به صورت مستقیم اجتناب می کردند. و قابلیت دانلود و دسترسی مستقیم به این کتب با هر فرمت دشوار هست
- کتاب‌ها اکثرا با فرمت pdf. و ebook. هستند و همچنین در تمامی این فرمت ها اکثرا در واقع به صورت عکس هستند و نمی توان به سادگی از آنان متن استخراج کرد

که در این راستا من به بررسی سه کتاب های صوتی با خانمان، طلای درون و عطر خوش مرگ با صدای که هر سه با صدای احسان کرمی هستند در اپلیکیشن‌ها و مراجع مختلف پرداختم. تا ابتدا به صدا و ویس مناسب دست پیدا کنم و سپس مهم تر از آن مشکلات مطرح شده بابت متن را حل کنم. به صورتی که با بررسی اپلیکیشن های مطرح شده، راهی برای خروجی گرفتن از کتاب ها با هر فرمتی پیدا شود.

در گام بعدی از آنجایی که با توجه به دیتا های دیده شده به احتمال زیادی فایل از فرمت pdf. هست. برای برنامه نویسی و یا پیدا کردن کدی که بتواند متن فارسی را از pdf. استخراج کند تلاش کردم و در طی آن شروع به بررسی مخازن و پروژه های مختلفی کردم از جمله:

- https://github.com/hooshvare/pdf2word
- https://pypi.org/project/farsi-tools/

پرداختم.

چالش بعدی فایل های ebook. و pdf. ای بودند که بعضا تصاویر اسکن شده از کتاب‌ها بودند و بایستی متن ها از آن کتاب ها استخراج می شدند در همین راستا تحقیقات اولیه ای در مورد کتابخانه‌هایی که برای تبدیل عکس به متن فارسی بودند نیز شروع کردم از جمله:

- https://github.com/tesseract-ocr/tessdata
- https://github.com/amirmgh1375/TextRecognitionDataGenerator

از دیگر مشکلات مهم برای استفاده از کتاب های صوتی به عنوان دیتاست نبود فایل های صوتی با فرمت wav. هست (فرمت درخواست از سمت تیم نرم افزار) که به شدت این فرمت برای کتاب های صوتی کمیاب بوده بنابراین تحقیقاتی مبنی بر بررسی امکان تبدیل mp3. به wav. مناسب را شروع کردم.

پردازش بعدی که باید برای قابل استفاده کردن کتاب‌های صوتی به عنوان دیتاست مورد استفاده قرار بگیرد. بخش بندی داده ها و تبدیل آن کتاب ها (مطابق با آن بخش بندی صدا های مربوط به کتاب) به بخش بسیار کوچک تر و در حد چند ثانیه ای بود.

که هنوز به آن مرحله نرسیدیم و با جلو تر رفتن پروژه امکان رسیدن به آن بررسی می شود.

کار بعدی که نیاز است تا انجام شود تا کتاب های صوتی به دیتا قابل قبول تبدیل شوند تبدیل آنها به فینگلیش هست (دیتای درخواستی از طرف تیم نرم افزار) در راستای این کار، فعالیت های زیادی انجام شده، برنامه ای توسط من نوشته شده که با بررسی ده ساعت دیتایی که در فاز های قبلی توسط تیم شکل گرفته واژه نامه ای شامل ۳۰,۰۰۰ کلمه فارسی و مطابق با آنها کلمات فینگلیش تشکیل شده تا بتوان فرایند تبدیل متن فارسی به داده فینگلیش را انجام داد.