

بسم الله الرحمن الرحيم

گزارش چهارم پروژه "استخراج کلمات
کلیدی از منابع صوتی و ویدئویی
با استفاده از هوش مصنوعی"
فاطمه وحیدیونسی



مرکز نوآوری علوم و فناوری های شناختی

دانشگاه علم و صنعت ایران

تاریخ تدوین: 1403/1/23

❖ تشریح جزئیات مربوط به مراحل مختلف پروژه

با توجه به تست های انجام گرفته بر روی نمونه های صوتی (از سورس تلگرام) و نمونه های ویدیویی (از سورس اینستاگرام)، دقت و زمان اجرا، برای هشت مرحله مختلف اجرای پروژه، در جدول زیر نمایش داده شده است.

***دقت شود که ستون چهارم جدول (زمان تقریبی اجرا) وابسته به سخت افزار استفاده شده است. در اینجا از GPU در اختیار قرار داده شده در Google Colab در پلن رایگان (Tesla T4) استفاده شده است که به علت اشتراکی بودن این سخت افزار، از سرعت بالایی در اجرا بهره مند نیست. بنابراین دقت ها و زمان بندی نمایش داده شده در جدول زیر براساس اجرای کد بر روی این سخت افزار بوده و بر روی سخت افزار های قوی تر نتایج بهتری را نیز می تواند به دست دهد.

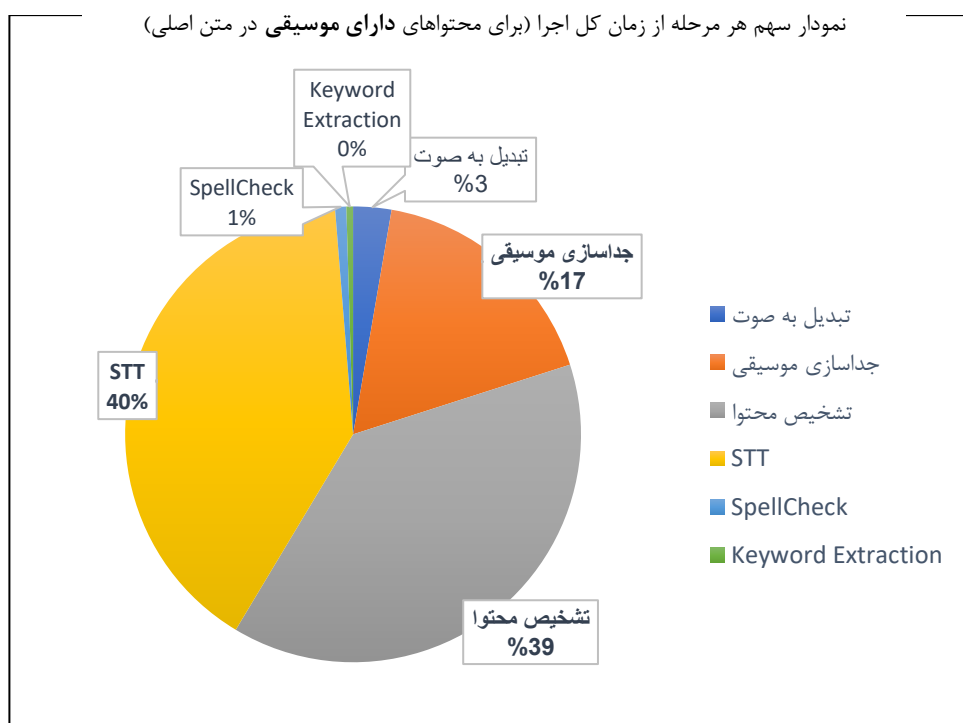
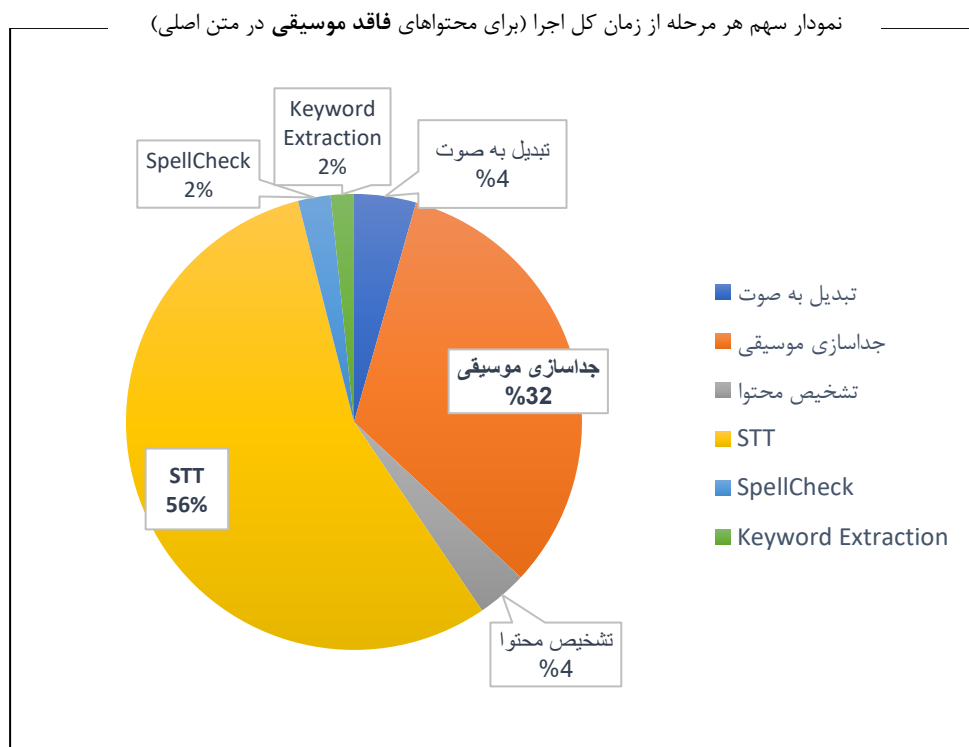
بر اساس اجرا بر روی Tesla T4 GPU			
مرحله	متشکل از	حدود دقت	زمان تقریبی اجرا
1-تشخیص فرمت فایل ورودی	تابع is_video	100%	کسری از ثانیه
2- تبدیل فایل ویدئویی به صوتی	تابع vid2audio	100%	4٪ از کل زمان اجرا همزمان وابسته به حجم و مدت زمان ویدئو. بنابراین با کاهش حجم فایل و فشرده سازی زمان، می توان از زمان اجرای این مرحله کاست. (در صورتیکه تکنیک اجرا شده، زمان بر نباشد). در کل زمان اجرای این مرحله بیشتر وابسته به "طول زمانی" فایل است.
3- جداسازی موسیقی زمینه از صوت	استفاده از مدل HTdemucs(v4) جهت جداسازی محتوای صوت به 2 بخش Vocal و Music (بخش Music خود شامل 3 فایل مربوط به instrument های مختلف موسیقایی است).	این مرحله دارای 2 معیار کیفی است: Mean opinion score (MOS)=4 Overall SDR ^{1*} =9.0	حدود 22% طول زمانی صوت 17٪ از کل زمان اجرا برای "محتوای دارای موسیقی در متن اصلی"
4- تشخیص نوع صوت از نظر وجود محتوای مفید برای مرحله STT (صوت به متن)	- استفاده از dBFS جهت محاسبه تراز شدت صوت موزیک زمینه و vocal - استفاده از تابع music_length جهت یافتن طول زمان موسیقی در صوت - استفاده از تابع music_mode جهت تشخیص محتوای صوت (vocal -vocal و موزیک زمینه- موسیقی)	90٪ در فایل هایی که همزمان حاوی چندین نوع محتوا در کنار هم هستند خطا رخ می دهد. به عنوان مثال، در فایل هایی که بخشی از آن حاوی سخنرانی و بخش دیگر حاوی موسیقی باشد. راه حل: 1- برای مدیاهایی که دارای محتوای مشخص هستند (به عنوان مثال، صوت های سخنرانی در یک کانال تلگرامی) این مرحله قابل حذف است. 2- برای دیگر مدیاهای (اینستاگرام) که معمولاً دارای موسیقی هستند در حال بررسی تکنیک های دیگر اکوستیکی جهت افزایش دقت شناسایی محتوا و کاهش زمان اجرا هستیم.	4٪ از کل زمان اجرا برابر با 2.5٪ طول فایل برای محتوای "فاقد موسیقی" در متن اصلی 39٪ از کل زمان اجرا برابر با 41٪ طول فایل برای محتوای "دارای موسیقی" در متن اصلی

1 **Overall SDR**: the mean of the SDR for each of the 4 sources. SDR(Source-to-Distortion Ratio) is usually considered to be an overall measure of how good a source sounds.

2 **WER**: Word Error Rate , **CER**: Character Error Rate

<p>41٪ از طول زمانی فایل، برابر با 56٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی و 40٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی</p>	<p>دقت برای مدل Large Whisper: به طور میانگین 36٪ "خطای" WER² و 12٪ "خطای" CER (88٪ دقت تشخیص حروف و 64٪ دقت تشخیص کلمات)</p> <p>***بررسی دقت در این بخش نیازمند تست بر روی داده های بیشتر بر روی سخت افزاری با سرعت بالاتر است. گزارش زیر، براساس بررسی های انجام شده "تاکنون" است:</p> <table border="1"> <thead> <tr> <th>Merging</th><th>Medium</th><th>Large</th><th></th></tr> </thead> <tbody> <tr> <td>Wer 35% Cer 12%</td><td>Wer 50% Cer 16%</td><td>Wer 36% Cer 12.8%</td><td>Without spell checking</td></tr> <tr> <td>Wer 26.9% Cer 11.7 %</td><td>Wer 39% Cer 13.9%</td><td>Wer 29% Cer 12.3%</td><td>With spell checking</td></tr> </tbody> </table> <p>همانگونه که در جدول فوق نشان داده شده است، افزایش دقت در حالت ادغام نتایج دو مدل Large Whisper , Medium اندک بوده اما درصد زمان اجرای مرحله STT را تا 30٪ افزایش می دهد. در نتیجه، در صورت عدم محدودیت زمانی می توان از تابع Integrated_stt و ادغام دو مدل استفاده نمود. در غیر این صورت، در صورت وجود محدودیت زمانی در اجرا، مدل Large به تنهایی دقت مناسبی را ارائه می دهد.</p>	Merging	Medium	Large		Wer 35% Cer 12%	Wer 50% Cer 16%	Wer 36% Cer 12.8%	Without spell checking	Wer 26.9% Cer 11.7 %	Wer 39% Cer 13.9%	Wer 29% Cer 12.3%	With spell checking	<p>- Predict با استفاده از مدل های Medium و Large Whisper - استفاده از تابع Integrated_stt جهت ادغام نتایج دو مدل و یافتن بهترین پیش بینی</p> <p>حجم مدل Medium: 1.4Gb حجم مدل Large: 2.9 Gb</p>	<p>STT+ Merging -5 Large & Medium Models</p>
Merging	Medium	Large													
Wer 35% Cer 12%	Wer 50% Cer 16%	Wer 36% Cer 12.8%	Without spell checking												
Wer 26.9% Cer 11.7 %	Wer 39% Cer 13.9%	Wer 29% Cer 12.3%	With spell checking												
<p>1.7٪ از طول زمانی فایل برابر با 2٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی و 1٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی</p>	<p>به طور میانگین دقت تشخیص کلمات را 7٪ افزایش می دهد. (خطای WER را 7٪ کاهش می دهد).</p>	<p>اصلاح خطای املائی با استفاده از پکیج Parsivar حجم مدل های اصلاح خطاهای املائی Onegram و My_bigram به ترتیب 10 و 157 مگابایت</p>	<p>Spelling -6 Correction</p>												
<p>کسری از ثانیه</p>	<p>100٪</p>	<p>برای محتوای دارای کپشن در کنار ویدیو: - حذف ایموجی از متن کپشن - یافتن هشتگ ها، ذخیره سازی و حذف underline - حذف آدرس های آیدی از کپشن</p>	<p>Text -7 (Caption) Modification</p>												
<p>1.2٪ از زمان فایل برابر با 1.6٪ از کل زمان اجرا برای محتواهای "فاقد موسیقی" در متن اصلی و 0.5٪ از کل زمان اجرا برای محتواهای "دارای موسیقی" در متن اصلی</p>	<p>ارزیابی دقت بر روی دیتاست Thesis Abstract شامل 450 مقاله از پایگاه Irاندو در زمینه علوم انسانی انجام گرفته است. دقت شود که با توجه به هدف پروژه، مرحله STT در اولویت بالاتر قرار داشته و تاکنون کمتر بر روی مرحله هشتم اقدام توسعه ای انجام گرفته، در صورت نیاز می توان دقت این مرحله را بالاتر برد:</p>	<p>- ادغام متن حاصل از STT و کپشن (در صورت وجود) - استفاده از مدل TopicRank با استفاده از پکیج Perke جهت استخراج 20 کلمه کلیدی اصلی</p>	<p>Keyword -8 Extraction</p>												

	Partial F1 score at 10 extracted keywords (pF1@10)= 12.7% Partial Precision(pP@10)=10% Partial Recall(pR@10)=17.35%	- رتبه بندی براساس کلمات موجود در هشتگ ها استفاده از مدل pos_tagger با حجم 19 مگابایت	
--	---	---	--



❖ پیشبرد و توسعه آتی در پروژه:

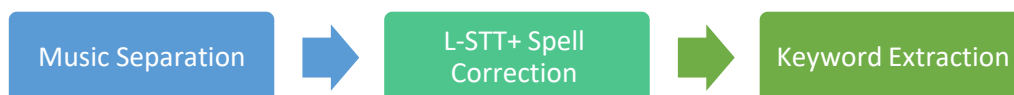
جهت افزایش دقت، بررسی تغییر تنظیمات مدل "کاهش نویز" بر روی دقت خروجی مدل STT جزو پیشبردهای آتی قرار دارد.

جهت افزایش سرعت اجرا، روش های زیر توصیه می شود و در دست بررسی است:

- همانگونه که گفته شد زمان بندی های فوق براساس اجرا بر روی پلن رایگان GPU گوگل کولب است. برای افزایش سرعت اجرا، می بایستی سخت افزار (GPU) قوی تری را به کار برد. (متعاقب آن، امکان تست بر روی مجموعه دادگان بیشتر را نیز ممکن می سازد).
- بررسی تاثیر کاهش حجم فایل و یا فشرده سازی محتوا (افزایش سرعت صوت) بر روی زمان اجرای هر یک از مراحل. درمورد مرحله تبدیل صوت به متن (STT) باید بررسی شود که این عملیات باعث کاهش کیفیت و دقت این مرحله نگردد.
- بررسی تکنیک های اکوستیکی جهت افزایش سرعت مرحله "تشخیص محتوا"
- حذف فرآیندهای کم تاثیر روی دقت (مانند ادغام مدل های STT). همچنین درمورد مدیاهایی با محتوای مشخص (مانند یک کانال سخنرانی در تلگرام) مرحله جداسازی و تشخیص محتوای موسیقایی لازم نبوده و قابل حذف است. به مثال زیر توجه کنید، با توجه به تفاوت محتوا در شبکه های اجتماعی و مدیاهای مختلف، معماری زیر پیشنهاد می شود:
- حالت اول (بررسی محتوای اینستاگرام/آپارات: محتوای دارای caption):



- حالت دوم (بررسی محتوای یک کانال سخنرانی در تلگرام، محتوای صوتی و فاقد caption):



همانگونه که در نمودار بالا نمایش داده شده است، بسته به نوع محتوا و سورس آن، می توان برخی مراحل اجرا را حذف نمود و سرعت اجرا را بالاتر برد.

❖ شایان ذکر است که در کنار فایل این گزارش، مجموعه دادگان استفاده شده در مرحله Keyword Extraction و مفید و قابل استفاده برای پروژه های دیگر مجموعه در زمینه پردازش متن، ارسال شده و در اختیار قرار می گیرد.