گزارش دوم پروژه "استخراج کلمات کلیدی از منابع صوتی و ویدئویی با استفاده از هوش مصنوعی"

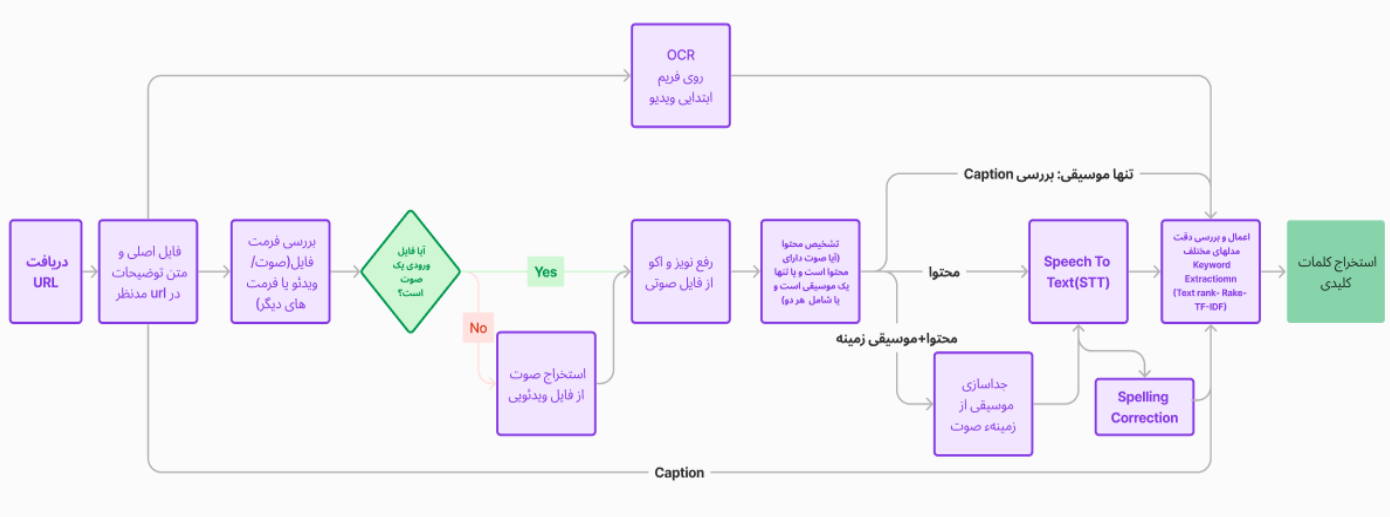
فاطمه وحیدیونسی

مرکز نوآوری علوم و فناوری های شناختی

دانشگاه علم و صنعت ایران

تاریخ تدوین: 1402/11/13

هدف از این پروژه ساخت یک فریمورک از ماژول های مختلف جهت رسیدن به مقصود "استخراج کلمات کلیدی با دقت بالا از فایل های صوتی و ویدئویی" از منابعی مانند اینستاگرام است. فلوچارت جدید(مرتبط با گزارش دوم) قرار داده شده در زیر، مراحل انجام پروژه را نشان می دهد:



نمودار1- فلوچارت جدید مراحل انجام پروژه

جهت درک بهتر، مراحل پروژه برای نمونهء استخراج کلمات کلیدی از "پست های اینستاگرام" در زیر تشریح می شود(مراحل هفتم، دهم و یازدهم در گزارش دوم اضافه شده اند):

1- فایل اصلی پست شده به همراه توضیحات (Caption) موجود در پست دریافت می شود. در این پروژه از هر دو سورس(متن توضیحات کپشن و متن ناشی از تبدیل صوت(STT) جهت استخراج دقیق تر کلمات کلیدی استفاده خواهد شد).
2- فایل ورودی از جهت فرمت، بررسی شده، در صورتی که فایل ورودی ویدئویی داشته باشد، مرحلهء استخراج صوت از ویدئو اجرا می شود.
3- نویز، اکو و صداهای زمینه
4- تشخیص "محتوای" فایل صوتی: درصورتیکه فایل تنها یک موسیقی است و فاقد محتواست، وارد مرحله استخراج کلمات کلیدی از سورس دیگر، متن caption ، می شویم. درصورتیکه صدای یک گوینده به همراه موسیقی بی کلام در بکگراند است، عملیات مرحله 6 ام آغاز می شود.
5- درصورتیکه در پس زمینهء صدا، موسیقی وجود دارد، حذف موسیقی از بکگراند و ذخیره محتوای اصلی جهت آماده سازی برای مرحلهء STT اجرا می شود.
6- استفاده از متد های مختلف STT جهت تبدیل صدای نهایی به متن، ارزیابی دقت آنها و انتخاب روش بهینه
7- اصلاح خطاهای املایی حاصل از STT

8- استفاده از متد های مختلف Keyword Extraction برای متن Caption و متن حاصل از صوت
9- یافتن ضریب بهینهء بتا در استخراج کلیدواژه از دو سورس "متن کپشن"و "متن حاصل از صوت"
10- بررسی افزایش دقت کلمات استخراج شده، با اضافه کردن منبع تصویری(استفاده از تکنیک OCR روی فریم اول هر ویدیو).
11- دررابطه با فایل هایی که دارای دو پارت موسیقیایی و پارت محتوایی، در زمان هایی متفاوت، هستند: تشخیص این پارت ها از یکدیگر، زمان بندی هر یک از آنها و تشخیص بازه زمانی که فایل صوتی دارای محتواست.

در زیر، جدول وضعیت اجرای مراحل فوق، تا کنون، قرار داده شده است:

** ردیف هایی که با رنگ سبز متمایز شده اند، پیشرفت پروژه را نشان می دهند.

مرحله	وضعیت
1- دریافت فایل اصلی پست شده به همراه توضیحات (Caption) موجود در پست	UI آپلود فایل انجام شده.
2- بررسی فرمت فایل ورودی، در صورتی که فایل ورودی ویدئویی داشته باشد، استخراج صوت	انجام شده
3- کاهش نویز، اکو و صداهای زمینه	انجام شده در این مرحله با استفاده از ماژول voicefixer نویز و reverb در زمینه برخی فایل های ورودی حذف شد.
4- تشخیص "محتوای" فایل صوتی(موسیقی باکلام است یا موسیقی بی کلام در زمینه به همراه محتوای اصلی توسط یک گوینده است.)	انجام شده بررسی منابع و مدل ها، در زمینه تشخیص محتوای موسیقیایی از محتوای عادی انجام شد، مدلی برای تشخیص یک "آهنگ" از یک "موسیقی بی کلام که صداگذاری" شده است تاکنون طراحی نشده است. کدی توسعه داده شد که ازطریق ویژگی های صوتی موجود در فایل های ورودی، با دقت نزدیک به 90٪ می تواند نوع محتوای فایل را شناسایی کند.
5- درصورتیکه در پس زمینهء صدا، موسیقی وجود دارد، حذف موسیقی از بکگراند و ذخیره محتوای اصلی جهت STT	انجام شده این عملیات در دو مرحله انجام می پذیرد: ابتدا با استفاده از مدل Htdemucs نت های موسیقیایی از زمینه صدای اصلی(vocal) حذف شده و سپس از طریق تابع طراحی شده، محتوای فایل ویدیویی/صوتی بررسی می شود. خروجی این فرآیند تشخیص 4 حالت در صداست:

	1-صوت فاقد هرگونه موسیقی است و تماما دارای محتواست. 2- صوت دارای موسیقی در زمینه و محتوای گفتاری یک راوی است. 3- صوت تماما موسیقی و فاقد محتوای گفتار انسان است. 4- صوت یک آهنگ با صدای یک خواننده است. این مرحله در ادامهء مراحل پروژه بسیار تعیین کننده است چرا که می توان اینگونه تعریف کرد که درصورتیکه صوت یکی از حالات 3 یا 4 را داشته باشد، فاقد محتوا تشخیص داده شود و نیازی به طی مراحل STT و اصلاح نگارش و استخراج کلمات کلیدی از صوت نباشد و یا محتوای آن دارای ارزش پایین تری به نسبت محتوای Caption باشد.
6- استفاده از متد های مختلف STT جهت تبدیل صدای نهایی به متن و ارزیابی دقت آنها	انجام شده مدل های STT موجود در زبان فارسی با استفاده از کتابخانه های Whisper ، Vosk، Deep speech و Hezar(مدل Whisper Fine-tuned) بررسی شد، مدل whisper medium در برخی متون و مدل whisper large نیز در برخی متون نتایج بهتری را خروجی می دهند. به همین سبب کدی برنامه نویسی شد که بهترین خروجی را به ازای هر بازه زمانی، از هر دو مدل دریافت کرده و بهترین آن را به عنوان خروجی نشان دهد.
7- اصلاح خطاهای املایی حاصل از STT	با استفاده از کتابخانه های Parsivar، Kenlm و Faspell خطاهای احتمالی موجود در خروجی مدل STT شناسایی و رفع می شود و متن نهایی آماده ورود به مرحلهء استخراج کلمات کلیدی می گردد.
8- استفاده از متد های مختلف Keyword Extraction برای متن Caption و متن حاصل از صوت	API استخراج کلمات کلیدی بررسی شد. چند متد نیز(Text-Rank, Topic Rank) شخصا پیاده سازی شد. دقت خروجی روش های پیاده سازی شده بیشتر است.
9- یافتن ضریب بهینهء بتا در استخراج کلیدواژه از دو سورس "متن کپشن"و "متن حاصل از صوت"	در دست انجام
10- بررسی افزایش دقت کلمات استخراج شده، با اضافه کردن منبع تصویری(استفاده از تکنیک OCR روی فریم اول هر ویدیو)	انجام شده از طریق چند متد مختلف با استفاده از کتابخانه های Persian-OCR ، Hezar، video_ocr و

	Tesseract ، عملیات OCR روی متن فارسی موجود در فریم اول هر ویدیو انجام شد. با توجه به دقت تشخیص متدهای OCR در زبان فارسی و همچنین احتمال تشخیص کلمات نامرتبط با محتوای پست، اضافه کردن این منبع، باعث افزایش کیفیت نتایج خروجی نمی شود.
11- دررابطه با فایل هایی که دارای دو پارت موسیقیایی و پارت محتوایی، در زمان هایی متفاوت، هستند: تشخیص این پارت ها از یکدیگر، زمان بندی هر یک از آنها و تشخیص بازه زمانی که فایل صوتی دارای محتواست.	در دست انجام

همانگونه در جدول فوق توضیح داده شده است، تا تاریخ 13 ام بهمن ماه (17روز پس از استارت پروژه)، پروژه، از مرحله ی دریافت فایل تا نمایش کلمات کلیدی حاصل از ویدیو/صوت ورودی به پیش رفته است و مرحله ای که بعدها، پس از تعریف پروژه اضافه گردید، یعنی مرحلهء استخراج کلمات کلیدی از متن caption یا توضیحات فایل و ادغام آن با نتایج حاصل از فایل صوتی/ویدیویی(مرحله9) و همچنین تشخیص بازه زمانی وجود محتوا در برخی فایل های ورودی(مرحله 11) باقی مانده است که در دست انجام است.