Generación multilingüe automática de subtitulado de video

Trabajo Final Herramientas y Aplicaciones de la Inteligencia Artificial

Ángel Aso Mollar
José Arias Moncho
Victoria Beltrán Domínguez

Junio 2022

Índice general

Introducción

Hoy en día, en el mundo globalizado en el que vivimos y a pesar de la gran cantidad de información multimedia que existe, hay ocasiones en las que por desconocimiento o por imposibilidad no tenemos a nuestro alcance mecanismos para consumir contenido en otros idiomas.

Dicha imposibilidad se manifiesta de forma más clara dentro del colectivo de personas con discapacidad auditiva, para las cuales ver un vídeo no adaptado se les hace tan o incluso más complicado que comprender conversaciones de habla no signada en la vida real, puesto que en los vídeos aparecen diversos factores como el ruido, interferencias, etcétera, que hacen la tarea todavía más difícil.

Grandes plataformas como Youtube presentan herramientas de subtitulado automático que, pese a ayudar con esta tarea, en ocasiones no están disponibles para todo vídeo. Además, algunos que están en otro idioma nos son imposibles de comprender si no está la opción de transcripción con traducción también presente, por no hablar de todas las plataformas que existen y que no presentan siquiera estas opciones, muchas veces por falta de recursos.

Esta última realidad es la que hace particularmente interesante el enfoque de desarrollo utilizando los servicios de Amazon Web Services (AWS) Translate y Transcribe, puesto que teóricamente una pequeña empresa podría, usando nuestro enfoque, elaborar un servicio de cara al usuario en el que se pueda hacer una transcripción y traducción de un vídeo de forma relativamente barata.

Es por esto que en este trabajo se va a hacer una demostración de cómo se podría integrar un servicio de transcripción y traducción de vídeo utilizando esta tecnología, haciendo además una interfaz de usuario para que la generación sea amigable y sencilla.

En la siguiente sección hablaremos de los objetivos concretos del trabajo para posteriormente comentar diversas plataformas y/o empresas que ya están utilizando estas tecnologías. A continuación, hablaremos del diseño concreto de nuestro sistema, y luego presentaremos la herramienta de forma explícita. Por último, haremos una pequeña evaluación de las prestaciones del transcriptor-traductor, con el fin de sacar unas conclusiones finales.

Objetivos

Nuestro proyecto, en su planteamiento, tuvo los siguientes objetivos específicos, los cuales han ido conformándolo:

- Entender la integración de tecnologías de manipulación de vídeo en AWS.
- Comprender la tecnología de Amazon Transcribe.
- Comprender la tecnología de Amazon Translate.
- Integrar ambas tecnologías para el desarrollo de un software capaz de generar archivos con los subtítulos de un vídeo para posteriormente traducirlos.
- Desarrollo de una interfaz amigable, destinada al usuario, para que él mismo pueda elegir el vídeo a traducir y el idioma en el que quiere los subtítulos.

Casos de uso

Son diversas las empresas que utilizan Amazon Transcribe y Amazon Translate en sus procesos o en sus productos. Por una parte, Transcribe se puede utilizar como conversor general de audio a texto y, específicamente, destacamos tres casos de uso más concretos: análisis de llamadas, transcripción médica y subtitulado. Por otra parte, Translate se puede utilizar como localizador de idioma, análisis de texto y comunicación. En este apartado procedemos a presentar algunas de las funcionalidades provistas por los servicios que vamos a utilizar, así como ejemplos de empresas que hacen uso de ellas.

Funcionalidades Amazon Transcribe

Amazon Transcribe Call Analytics

Los servicios de Call Analytics son utilizados para generar transcripciones textuales de llamadas telefónicas, basándose en modelos entrenados específicamente en estos dominios para proporcionar transcriptores de calidad.

El servicio proporciona a cada empresa mucha información relativa a la experiencia del usuario, como por ejemplo análisis de sentimientos, análisis de interrupciones, enmascaramiento de información sensible, etcétera, con

el objetivo doble de mantener la privacidad del cliente y a la vez extraer valiosa información para mejorar la experiencia.

Amazon Transcribe Medical

Es un servicio de reconocimiento automático del habla especializado en vocabulario médico, capaz de transcribir con precisión medicamentos, procedimientos y estados médicos y enfermedades.

Se ha creado con el objetivo de facilitar a las empresas diversas tareas como la transcripción de conversaciones entre pacientes y médicos para la documentación clínica, la captura de llamadas telefónicas en farmacovigilancia, el dictado médico o el subtitulado de consultas de telemedicina.

Amazon Transcribe for Subtitling

Servicio de subtitulado de cualquier tipo de media: desde transcripción bajo demanda, video en streaming, etc, y utilizando tecnologías como la posibilidad de añadir un modelo de lenguaje concreto específico del dominio, filtrado de vocabulario sensible o específico, opción de subtítulos multilingüe...

Funcionalidades Amazon Translate

Amazon Translate for Language Localization

Amazon Translate permite la traducción de enormes cantidades de contenido generado por usuarios en tiempo real. Las grandes capacidades de la aplicación permiten a los sitios web y las aplicaciones disponer un gran número de contenidos en una gran variedad de idiomas con un simple click.

Amazon Translate for Text Analysis

La aplicación permite sobrepasar la barrera del idioma en cuanto a análisis textual se refiere: las empresas pueden conocer cualquier tipo de opinión en red, traducirlo al inglés y utilizar otros servicios como Amazon Comprehend para analizar contenido contextual dentro de cada opinión individual.

Amazon Translate for Communication

El servicio puede integrar traducción automática en cualquier sistema de tiempo real, por lo que un agente o persona angloparlante encargada en una empresa puede agregar funcionalidades dentro de un chat, correo electrónico, etcétera, para facilitar la comunicación acerca de su marca o producto, o entre los propios usuarios de una aplicación.

Uso en la empresa

Empresas de todo tipo utilizan tanto Amazon Transcribe como Amazon Translate integrados en sus servicios. Ambos han permitido mejorar con creces la experiencia de usuario dentro de estas empresas, por lo que cada vez más marcas confían en AWS para automatizar partes o la totalidad de su negocio.

Algunas como Deliveroo, conocida empresa de reparto a domicilio de comida, utilizan Amazon Transcribe para convertir las llamadas a texto y, de esa forma, mejorar la experiencia de cliente comprendiendo mejor por qué los clientes llaman a sus centros telefónicos y analizando cualquier inconveniente en la atención al usuario.

El servicio de retransmisión en directo de carreras automovilísticas F1 TV cuenta con subtítulos automáticos en directo en tres idiomas diferentes: inglés, español y francés, lo cual es especialmente desafiante por la combinación de velocidades muy altas y comentarios dinámicos de múltiples colaboradores, así como de una terminología especializada.

Cerner Corporation, proveedor estadounidense de servicios, dispositivos y hardware de tecnología para la información sanitaria, está actualmente en desarrollo de una aplicación que integra Amazon Transcribe Medical con el objetivo de escuchar interacciones médico-paciente y transcribir los conceptos necesarios para ingresar en una BBDD médica.

Otras empresas como BMW, Siemens, CaptionHub, Hotels.com o la Universidad de Sheffield (Inglaterra), utilizan Amazon Translate como complemento para mejorar la comprensibilidad de sus contenidos a un ampio abanico de personas en todo el mundo.

Diseño del sistema

Una vez contextualizado el uso de los principales servicios de Amazon que vamos a utilizar, en este apartado, pasamos a describir de manera precisa de qué forma se ha diseñado el sistema con el fin de generar subtítulos en el idioma especificado.

De esta manera, en primer lugar, vamos a centrarnos en el uso de las herramientas propias de AWS, que son los pilares de nuestro proyecto: Amazon Transcribe y Amazon Translate. Una vez clara la funcionalidad que aporta cada uno de estos módulos, comentaremos cómo se ha utilizado Flask para poder crear un sistema cómodo y sencillo para el usuario final.

Amazon Transcribe

Amazon Transcribe, como se ha comentado anteriormente, es un servicio de reconocimiento automático del habla que permite agregar soporte de voz a texto en cualquier aplicación. Particularmente, en nuestro proyecto, dado un vídeo, necesitamos extraer el audio de este y guardarlo en algún tipo de archivo como texto escrito. Con este propósito, vamos a utilizar la función *start_transcription_job*.

Para poder ejecutar esta función, deberemos tener subido en un bucket el vídeo objetivo y saber de antemano su idioma. Una vez cumplidos esos prerequisitos, la llamada sólo necesita el nombre del trabajo, el enlace del vídeo a traducir en s3 (especificando también el bucket donde reside), el idioma del vídeo, y en qué ruta queremos guardarlo. Además, también se puede especificar qué formato de salida queremos para los subtítulos (vtt o srt). En nuestro caso, solamente necesitamos el vtt.

Esta llamada es asíncrona. Así, con el fin de poder gestionarla en nuestra aplicación, hemos ido comprobando en un bucle infinito el estado del job o trabajo, saliendo del bucle o bien cuando el estado coincide con *COMPLETED* o *FAILED*.

Amazon Translate

Amazon Translate es un servicio de traducción automática neuronal que ofrece traducciones a diferentes idiomas de manera rápida, accesible y personalizable. En este contexto, necesitamos Amazon Translate para poder traducir los subtítulos generados por Transcribe al lenguaje objetivo seleccionado. En este servicio, se nos plantearon dos alternativas:

- *start_text_translation_job*: Esta alternativa, dada la uri de una carpeta localizada en un bucket de s3, traduce todos los documentos dentro de esa carpeta en la uri de salida especificada. De esta manera, debemos especificar la uri de la carpeta fuente y destino, el nombre del job o trabajo, los idiomas fuente y destino y finalmente el rol que tenemos. Específicamente, para utilizar este servicio es preciso identificarse con un AWS Identity and Access Management (IAM), un sevicio que ayuda a administrar de manera segura el acceso a los recursos de AWS. Después de una pequeña búsqueda, en nuestro caso solamente hemos tenido que especificar el rol de *arn:aws:iam::ID_CUENTA_AMAZON:role/LabRole*, puesto que somos estudiantes. Al probar esta alternativa, nos dimos cuenta de que tardaba 20 minutos en generar una traducción de un archivo bastante pequeño. Por ello, esta opción quedó descartada y utilizamos la siguiente alternativa.

- *translate_text*: En esta segunda alternativa, simplemente dada una cadena, el idioma original y el idioma destino, devuelve la traducción de la cadena. Con el fin de poder utilizarla, se ha preprocesado el archivo vtt, obteniendo solo aquellas frases que tenían que traducirse y convirtiéndolas en una sola cadena con saltos de línea y pasándola como entrada. Esta solución podría no servir para vídeos de larga duración, pues quizás el texto transcrito podría no caber en la llamada a la API, pero es mejor que realizar una llamada por cada línea. En efecto, al realizar una prueba con un vídeo de gran duración (30 minutos), se ha comprobado que no funciona. Por ello, se ha decidido realizar una segmentación de las frases, yendo de 50 en 50 y así se ha solucionado este posible fallo de implementación.

Integración de los servicios en Flask

Flask es un framework que permite crear aplicaciones web en Python, y que se ha utilizado para cargar las vistas HTML y para crear el backend de la herramienta. Este framework tiene una estructura de trabajo muy marcada y sencilla, de forma que solo se ha necesitado un fichero Python para implementar la página web. En este fichero, hemos definido diferentes rutas, que son vistas web accesibles a partir de una URL en el navegador. En nuestra aplicación, tenemos 4 rutas diferentes, 2 de ellas siendo vistas y las otras 2 siendo rutas utilitarias.

- La ruta raíz o /: En ella, se puede observar el formulario que el usuario tendrá que rellenar para generar los subtítulos de su vídeo. Además, esta ruta se encargará también de recibir el formulario, procesar los campos y, tras comprobar que sean correctos, llamar al código encargado de trabajar con AWS.

- La ruta /show_video: En esta segunda ruta y última vista de la aplicación es donde el usuario podrá visualizar el vídeo que ha seleccionado subtitulado. Esta ruta necesita dos parámetros, el nombre del vídeo y el de los subtítulos. Si ambos son correctos, se visualizará este vídeo subtitulado.

- Dentro de las rutas utilitarias tenemos /subtitles y /videos, rutas que nos devolverán el fichero indicado en un parámetro, buscando este fichero en las carpetas correspondientes.

Con el fin de esquematizar toda la información previamente descrita, se adjunta la Figura 1 en la que se puede observar todo el flujo del sistema dentro de la aplicación.

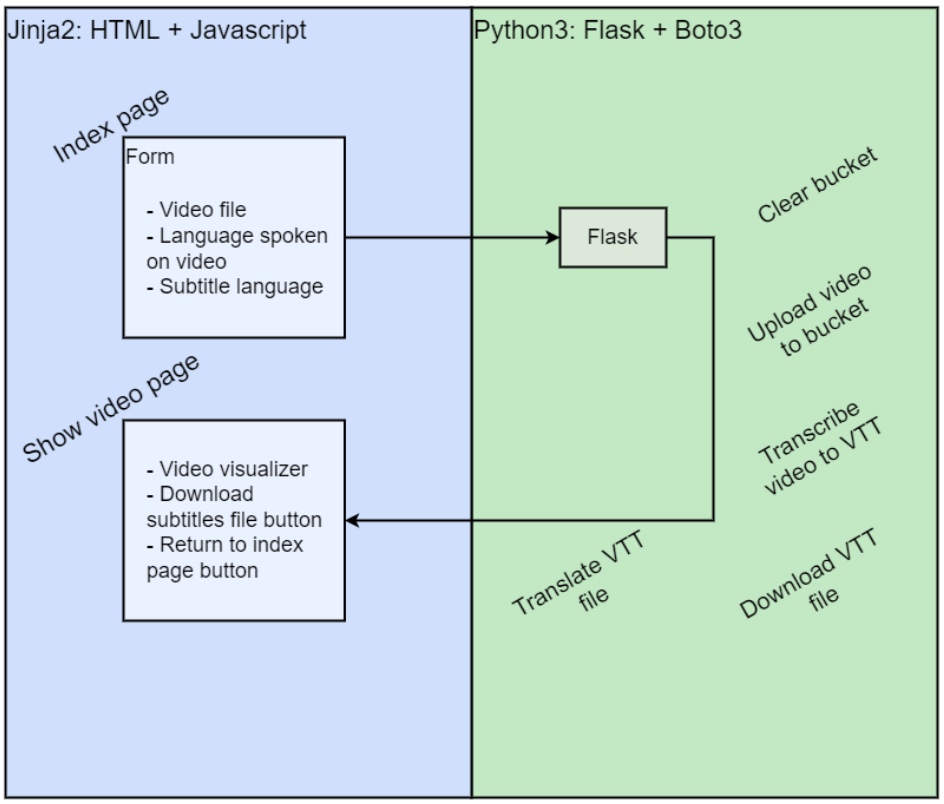
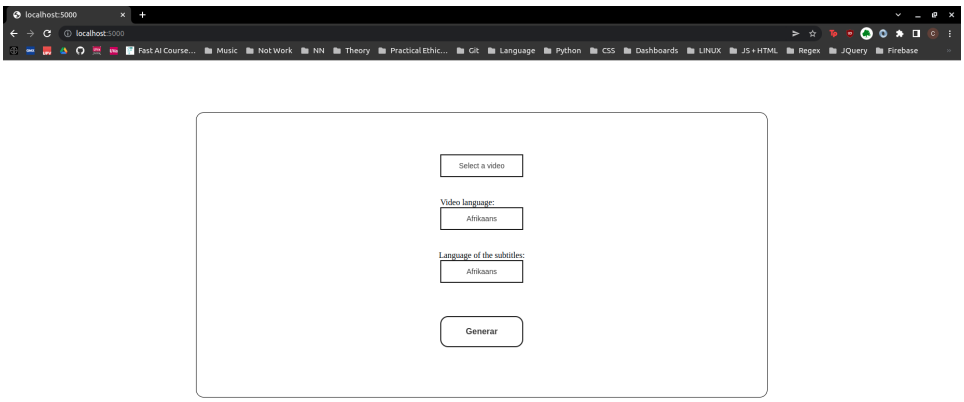


Figura 1: Estructura de la herramienta diseñada

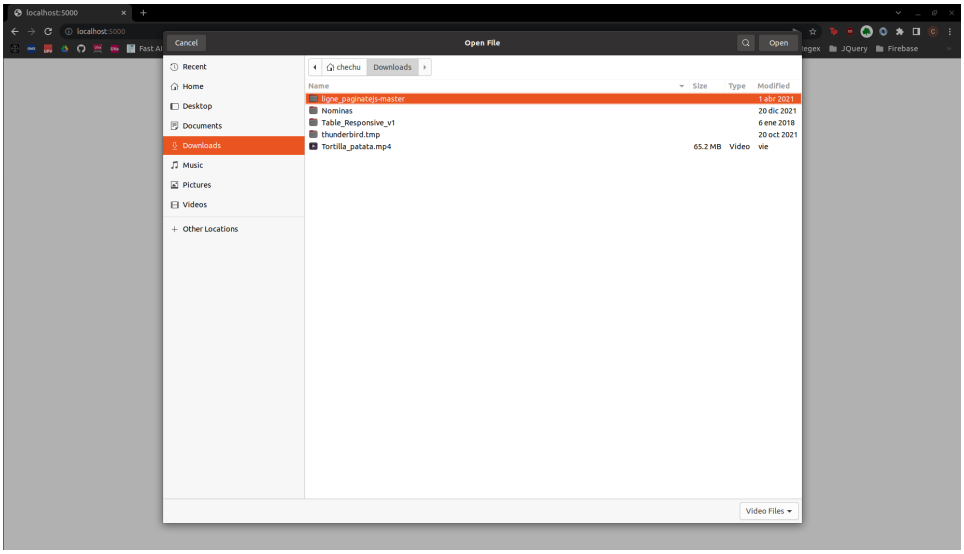
Presentación de la herramienta desarrollada

En esta sección se va a presentar de forma visual la herramienta con su interfaz gráfica así como comentar qué acciones pueden realizarse dentro de ella:

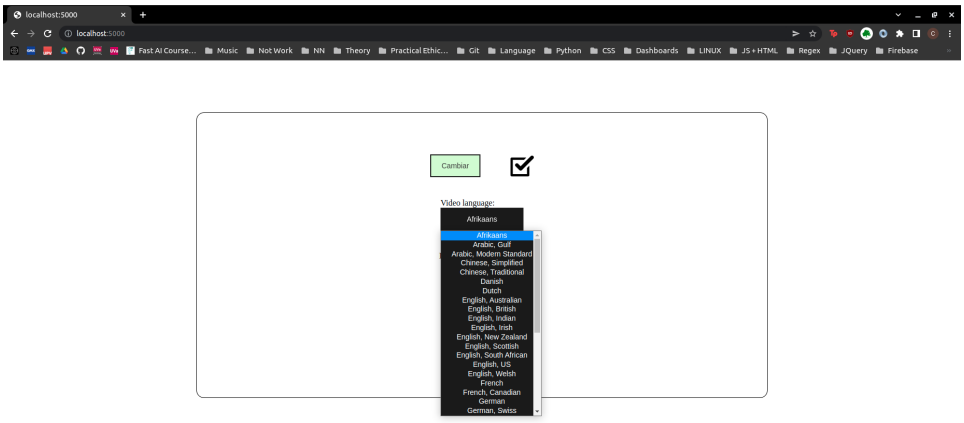
En primer lugar tenemos la ventana principal:



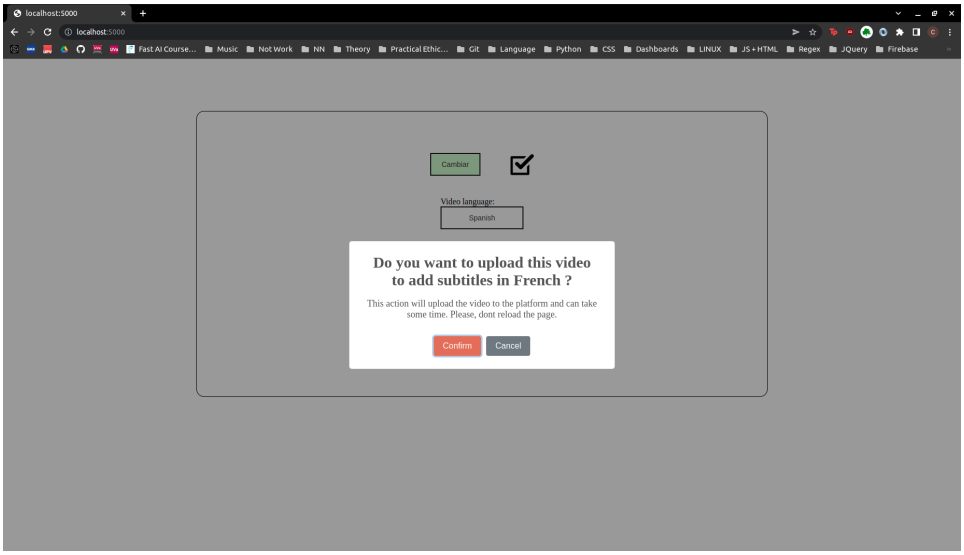
En ella, podremos rellenar el formulario indicado para realizar el subtitulado del vídeo. Si clicamos en la opción de select a video, podremos observar la siguiente ventana.



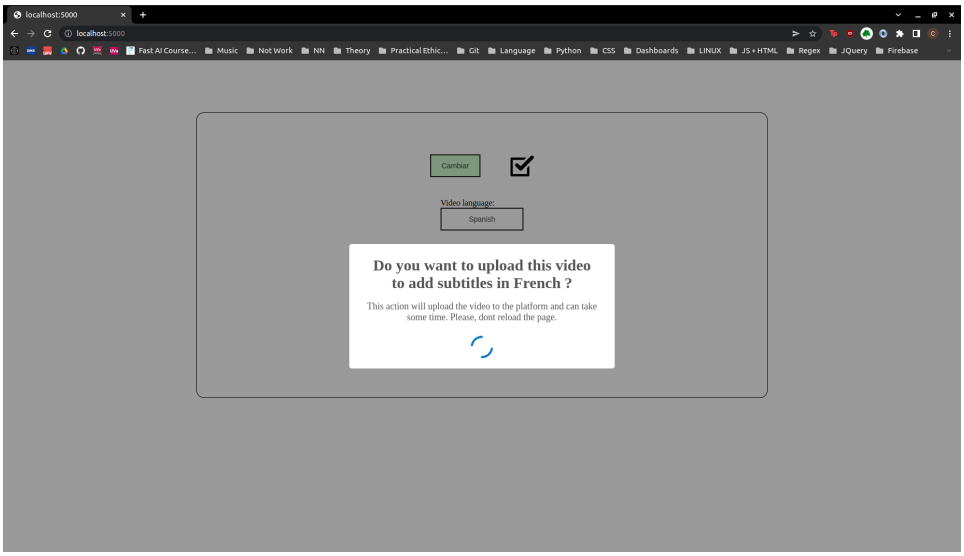
Una vez seleccionado el vídeo que deseamos subtitular, el botón cambiará y podremos pasar a seleccionar el idioma del vídeo y de los subtítulos:



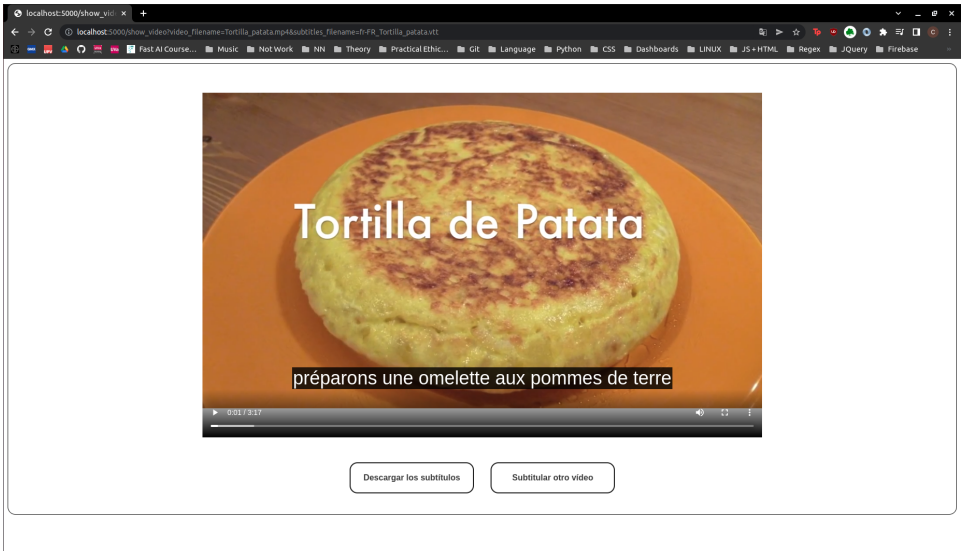
Al clicar en el botón de generar, nos aparece el siguiente diálogo:



Podremos o cancelar esta acción o confirmar. En caso afirmativo, el diálogo pasará a este nuevo estado que se muestra a continuación:

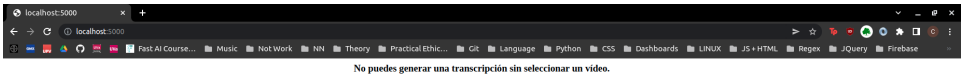


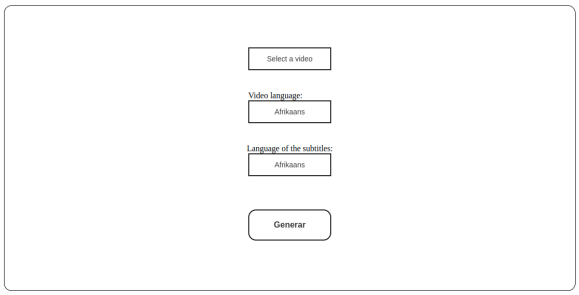
Una vez finalizado este proceso de generación de subtítulos, se nos redireccionará a la pantalla, donde tendremos el vídeo subtítulado:



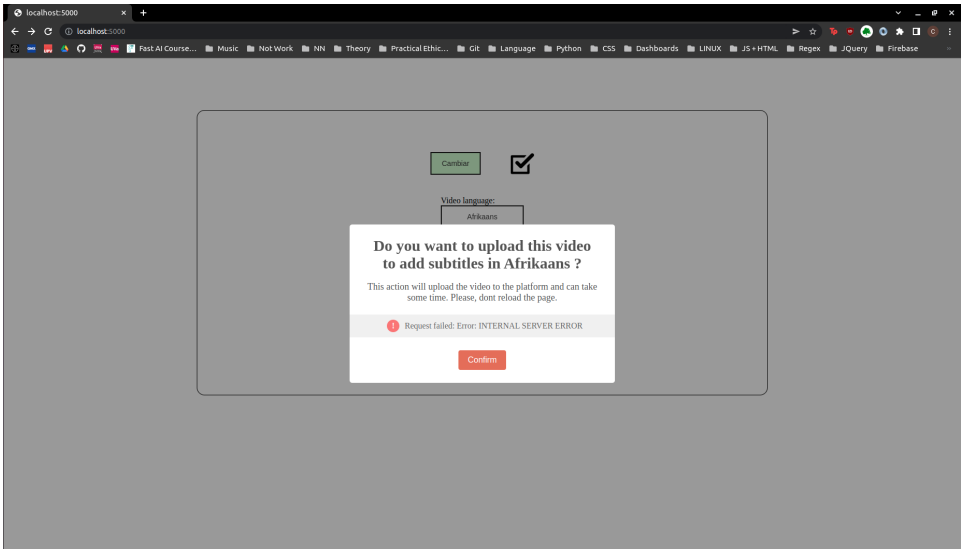
En esta pantalla, podremos clicar en descargar los subtítulos o simplemente ir a la generación de subtítulos otra vez.

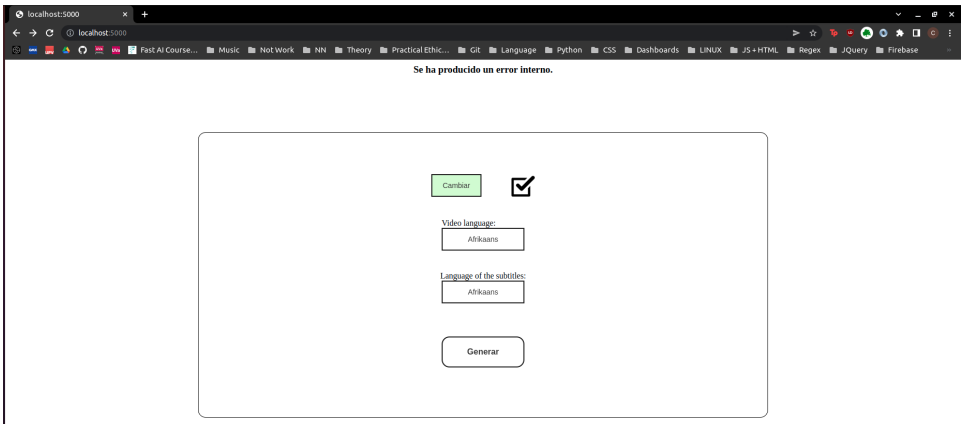
Este sería el flujo de la aplicación sí todos los pasos se siguen correctamente, por otro lado tenemos otras pantallas en función de si se producen fallos:





Esta pantalla se mostrará cuando no hayamos seleccionado un vídeo para subir a la aplicación.





Por último, estas serán las pantallas sucesivas que se visualizarán cuando haya algún problema en el servicio, como en este caso, en el que las credenciales de AWS no estaban correctamente actualizadas en el servidor.

Todo el código desarrollado se encuentra disponible en el siguiente link.

Evaluación

En este apartado vamos a centrarnos en evaluar de forma objetiva el funcionamiento del sistema, comentando los resultados generales al trabajar con AWS así como aquellos aspectos que podrían mejorarse:

- En primer lugar hemos decidido centrarnos en Amazon Transcribe, donde se ha observado su rendimiento al trabajar con vídeos largos. Específicamente, se ha observado que con un vídeo de 30 minutos, se tarda un total de 4 minutos y 20 segundos (260 segundos) en realizar la transcripción, lo cual es poco tiempo dada la duración del vídeo, pero sigue siendo bastante para realizarse online. Esta solución no serviría para herramientas que quieran subir un vídeo y en el momento hacer una transcripción, pero podría explorarse el uso de esta herramienta de manera online. Por otro lado, los resultados son buenos, pero tienen la importante limitación de que tienes que indicar el idioma que se habla en el vídeo, algo que impide que se reconozcan diferentes idiomas o variantes dentro de uno mismo. Además, los resultados no son perfectos cuando únicamente se trabaja con dos locutores. Estas dos limitaciones son bastante importantes, pero sabemos que son aspectos

que están en desarrollo dentro de los trabajos realizados en el campo del habla.

- En segundo lugar tenemos Amazon Translate, que funciona muy bien. Específicamente, al trabajar con el vídeo de 30 minutos, se producía un fichero VTT con un total de 2305 líneas, de las cuales solo había que traducir 576. Esto se ha realizado trabajando de 50 en 50 frases y ha tomado muy poco tiempo, 7 segundos para ser exactos. El trabajo de Amazon Translate es muy bueno, consiguiendo unos muy buenos resultados de forma rápida, a partir del resultado que Amazon Transcribe producía anteriormente.

- En tercer lugar queremos expresar nuestra perplejidad en cuanto al coste de estos dos sistemas, puesto que sorprendentemente después de todas las pruebas realizadas ninguna de las tres cuentas ha gastado ni un sólo dolar. Tras investigar esto, hemos observado que tienes un período gratuito de prueba tanto de Amazon Transcribe como de Amazon Translate. Finalmente, AWS S3 no ha supuesto un problema, ya que cada vez que se hace esta tarea, se limpia todo el bucket, con lo que solo tendríamos un vídeo y dos ficheros de subtítulos almacenados. Con esto, queremos indicar que no podemos sacar conclusiones acerca de la calidad de nuestro proyecto en base al coste, aunque vistos los precios, consideramos que podría ser una opción asequible, siempre que se estudiase el caso concreto en cuanto a rentabilidad y beneficios asociados.

- Por último, tenemos que comentar el desempeño de nuestro sistema de forma global, dentro del cual solo queremos indicar que se podría mejorar algunos detalles de la gestión de memoria del servidor, ya que los vídeos subidos por los usuarios no se borran, cosa que a la larga podría llevar a una gran cantidad de disco ocupada por vídeos que son prácticamente inaccesibles. Otro problema sería que si dos usuarios suben vídeos con el mismo nombre, actualmente esto no se solventa de ninguna manera, solo se quedará el más reciente de ambos.

Conclusiones

En este proyecto, hemos explorado los casos de uso y funcionalidades de Amazon Translate y Transcribe, creando un proyecto que nos permite subir un vídeo y o bien visualizarlo con los subtítulos indicados o bien descargar los subtítulos.

Como ya se ha comentado, consideramos que nuestro proyecto tiene muchísima utilidad, ya que existe una gran cantidad de idiomas con los que podemos trabajar y podría ser una opción bastante útil para pequeñas plataformas como plataformas de tutoriales con vídeos o pequeñas plataformas de consumo de contenido.

Consideramos que la herramienta ha quedado bastante pulida y funcional y que, a pesar de las desventajas comentadas previamente, Amazon consigue proveer de un servicio cuya calidad va de la mano con el precio.