Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

TRANSCRIPCIÓN AUTOMÁTICA DE VOZ A TEXTO PARA ANÁLISIS DE DISCURSO A PARTIR DE ARCHIVOS DE AUDIO

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.

Sector	**Lenguaje**	**Fuente de datos**
Ciencia, tecnología e innovación	Python API de Microsoft Azure	Archivos de audios provenientes de la plataforma YouTube.

Presentación
Para la Unidad de Científicos de Datos del DNP es de gran importancia estar en capacidad de trabajar con datos e información de diversos tipos y formatos; esto incluye trabajar con fuentes de información no estructurada tales como imágenes, texto, videos y audio, entre otras. Este proyecto es un primer acercamiento de la UCD al análisis de datos de audio, e involucra componentes de carga, adecuación y procesamiento de audios. En este piloto se desarrolló una herramienta de transcripción de voz a texto, apoyada en servicios cognitivos de Microsoft Azure, que permite obtener de manera automatizada la transcripción a texto plano de un discurso, intervención o conversación. Luego, a partir del texto generado se utilizan técnicas de minería de texto para analizar los temas sobre los que trata el discurso, basados en los términos más utilizados y la relación entre ellos. Esta herramienta permite automatizar la labor de análisis de discurso a figuras públicas de interés, y existen numerosos y variados tipos de análisis que se pueden realizar a partir de los textos resultantes.

For the National Planning Department's Data Scientists Unit (UCD) it is of great importance to be able to work with data and information of various types and formats; this includes working with sources of unstructured information such as images, text, videos and audio, among others. This project is a first approach of the UCD to the analysis of audio data, and involves components of loading, adaptation and processing of audios. In this pilot, a voice-to-text transcription tool was developed, supported by Microsoft Azure cognitive services, which allows to obtain automatically the plain text transcription of a speech, intervention or conversation. Then, from the generated text, text mining techniques are used to analyze the topics that the discourse deals with, based on the most used terms and the relationship between them. This tool allows to automate the work of discourse analysis for public figures of interest, and there are numerous and varied types of analyzes that can be performed from the resulting texts.

Objetivo general
Implementar una herramienta que permita automatizar la obtención y adecuación de archivos de audio con personas hablando, su transcripción a texto y el posterior análisis de este, utilizando técnicas de minería de texto y procesamiento de lenguaje natural.

Objetivos específicos
1. Implementar rutina que permita descargar de manera automática audios a partir de videos de la plataforma YouTube.
2. Efectuar el adecuamiento de archivos de audio para que, independientemente de su origen, sus características (formato, tasa de muestreo, número de canales) sean las adecuadas para realizar la transcripción de voz a texto.

3. Utilizar la API de Microsoft Azure para realizar la transcripción de voz a texto, utilizando como insumo los archivos de audio adecuados.
4. Utilizar técnicas de minería de texto y procesamiento de lenguaje natural para analizar componentes tales como temáticas, sentimiento y claridad del discurso.

Metodología

La metodología utilizada para el desarrollo de esta herramienta se puede ver en la Figura 1. Como se puede apreciar en el diagrama, el proyecto se compone de 4 etapas distintas, que serán descritas a continuación.

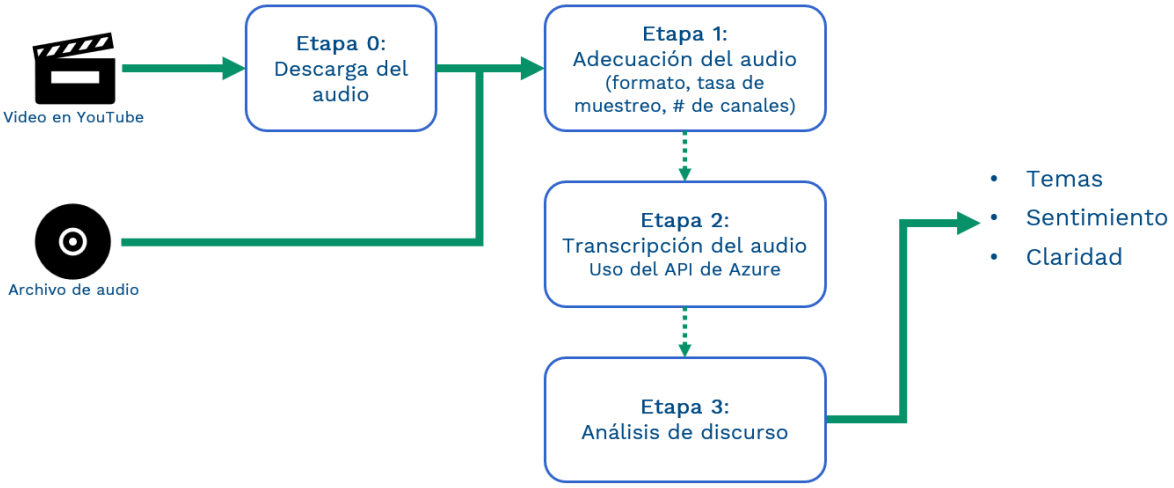

Figura 1: Componentes de la herramienta y flujo del proceso.

Etapa 0: Descarga de audios

La presente herramienta tiene dos posibles fuentes de entrada de datos: archivos de audio guardados en una ubicación determinada (en un computador/servidor local o en la nube), o videos de la plataforma YouTube. En el segundo caso, es necesario definir la forma en la que se debe descargar un video en particular como un archivo de audio.

Para lograr este objetivo, se desarrolló un módulo en Python que toma como entrada la dirección url del video de interés y extrae el audio de dicho video junto con su metadata. El resultado de este módulo es un archivo de audio con la mejor calidad disponible, en el formato especificado por el usuario (en este proyecto se utilizaron audios en formato MP3).

Etapa 1: Adecuación de archivos de audio

Una vez se tiene el archivo de audio en el formato deseado, el siguiente paso es adecuar este audio para que pueda ser utilizado en el API de Azure. Este proceso involucra las siguientes operaciones:

1. Normalizar el sonido con respecto a la máxima ganancia del audio, para que todos los audios queden en una escala comparable.
2. Remover el offset DC del audio, para eliminar ruidos de fondo que no son deseados en el análisis.
3. Convertir el audio a monoaural (1 solo canal).

4. Aumentar el volumen del audio por un valor parametrizable (en este caso se aumentó en 3 decibeles) para acentuar las partes de diálogo en el archivo. Dado que el audio ya se encuentra normalizado en este punto, es posible hacer esta adición de manera general con un mismo valor.
5. Realizar un re-muestreo del audio, para que quede en una tasa de bits establecida. En este caso, es necesario que el audio tenga una tasa de 16 Kbits por segundo para que pueda ser utilizado por el API de Azure.
6. Convertir el archivo a formato WAV (el soportado por Azure), y exportar el nuevo archivo de audio con todos los cambios aplicados.

El resultado de ese módulo es un archivo de audio listo para entrar en el API de voz a texto de Azure.

Etapa 2: Transcripción del audio
En esta etapa se realiza la conversión de voz a texto. Para ello, se utilizan los servicios de Microsoft Azure, que cuentan con un API que realiza esta acción. Es importante destacar que es necesario contar con un token válido y activo para poder utilizar este servicio. El resultado de este módulo es un archivo plano de texto con la transcripción del archivo de audio, que es guardado en una ubicación establecida por el usuario.

Etapa 3: Análisis de discurso
Una vez se tiene el texto plano del audio, es posible aplicar técnicas de minería de texto para realizar diferentes tipos de análisis de discurso. Tal y como se muestra en la Figura 1, algunos de estos análisis incluyen estudiar temática, sentimiento y claridad del discurso. En este primer piloto se hizo un análisis de tipo descriptivo, estudiando las palabras y bigramas más frecuentes en el estudio, para dar una idea de la temática del mismo. Estos términos se pueden visualizar en forma de nubes de palabras, o pueden ser representados por medio de tablas que indiquen sus respectivas frecuencias de aparición.

También se hizo un análisis de coocurrencias de términos dentro del texto, para observar qué palabras se relacionan más entre sí. Esta relación entre términos puede ser visualizada por medio de un grafo no dirigido, en donde el grosor de las aristas representa la coocurrencia de un par de términos.

Resultados
En la realización de este piloto se tomó como muestra para pruebas una alocución del presidente Iván Duque realizada el 10 de marzo de 2019, y que tiene aproximadamente 12 minutos de duración. Este audio se eligió por varias razones. En primer lugar, se trata de la alocución de una de las figuras públicas más importantes del país, por lo que puede ser de gran interés hacer análisis automatizado de su discurso. En segundo lugar, está el hecho de que tanto el video como su transcripción oficial son públicos, y el video está disponible en la plataforma Youtube, lo que permitió probar el flujo completo de la herramienta. Finalmente, están otras razones de tipo técnico, como el hecho de que el audio de este video es bastante claro y la alocución tiene buen ritmo y dicción, lo que permite tener un caso en condiciones cercanas a las ideales para un primer aproximamiento a esta herramienta.

El video tiene una duración de 13 minutos y 13 segundos. El proceso completo de audio a texto (que incluye las etapas 0 a 2, descritas en la sección anterior, más la eliminación de archivos temporales generados durante el proceso) tomó un total de 6 minutos y 53 segundos. A partir del archivo de texto generado se realizó el análisis de texto (etapa 3). La Figura 2 muestra la nube de palabras con los términos más frecuentes del discurso.

Figura 2: Términos más frecuentes del discurso estudiado

En la Figura 3 se presenta el grafo de coocurrencias del texto, mostrando la fortaleza de relaciones entre palabras, y los términos que más se relacionan.

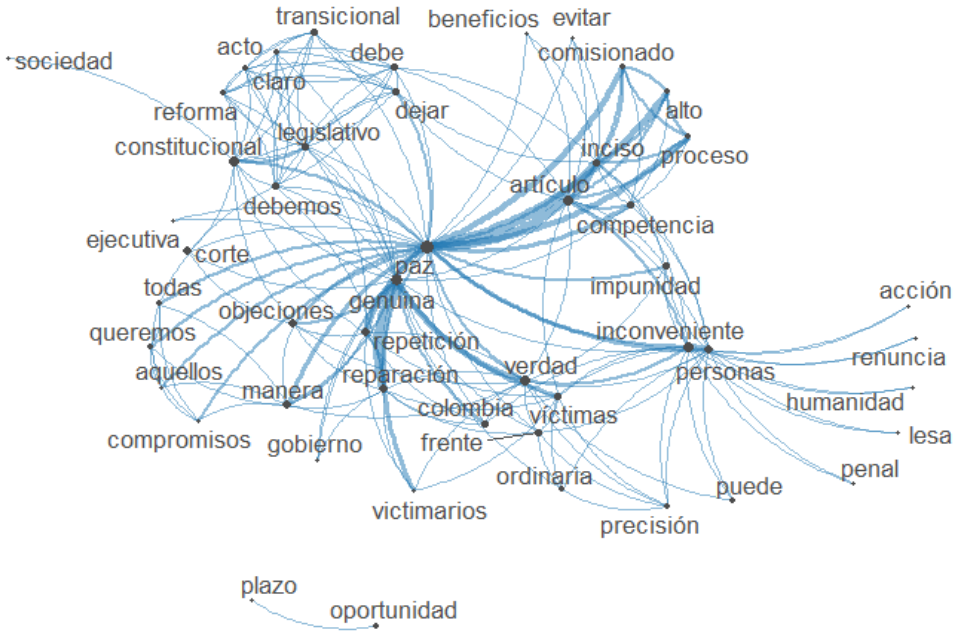

Figura 3: Términos más frecuentes del discurso estudiado

Adicionalmente, al tener una transcripción oficial del video de prueba, se procedió a comparar este texto con el texto obtenido de manera automatizada, para validar el desempeño de la herramienta. Para realizar esta comparación se realizaron dos acciones. En primer lugar, se hizo el análisis descriptivo para ambos textos y se compararon los resultados, siendo estos muy similares. Las nubes de palabras de bigramas, tanto para la transcripción oficial como para la obtenida a través de la herramienta, se muestran en la figura 4.

Figura 4: Comparación de nubes de palabras para las dos transcripciones del audio

En segundo lugar, y con el deseo de tener una métrica más puntual y objetiva de la semejanza entre las dos transcripciones, se procedió a calcular vectorizar ambos textos por medio de la técnica de *Bag of Words*, y calcular la similitud coseno entre ambos vectores. Esta medida dio un valor de 0.9908668 (en una escala de 0 a 1), lo que indica que ambas transcripciones son muy similares, y que la herramienta tuvo un buen desempeño.

Es necesario destacar de nuevo que este es un caso en condiciones bastante ideales, en donde el audio es de buena calidad y no tiene ruidos de fondo que obstaculicen el proceso de voz a texto, y el discurso se presenta de manera clara y neutral. En otras condiciones del audio es posible que el desempeño de la herramienta empeore.

Conclusiones

1. Apoyados en servicios externos (Azure), se ha desarrollado una herramienta *end to end* (es decir, que incluye todos los pasos intermedios necesarios) para realizar transcripción de audios a textos planos que posibiliten su posterior análisis por medio de técnicas de minería y procesamiento de lenguaje natural.
2. Un producto secundario de esta herramienta es una serie de programas y funciones que permiten leer, editar y exportar archivos de audio. Esto será de gran utilidad para el desarrollo de otros proyectos que involucren trabajar con archivos de este tipo.
3. Como complemento a la herramienta de transcripción, y aprovechando la experiencia de la UCD en analítica de textos, se han incorporado análisis de temática a las transcripciones generadas, para determinar de qué habla un audio/discurso en particular. En el caso de este piloto el análisis de texto se delimitó a frecuencia de términos e interacción entre ellos. Sin embargo, para una segunda fase de este proyecto es posible introducir otros elementos que permitan identificar temáticas con mayor complejidad, analizar sentimiento a lo largo del discurso y analizar la claridad de los términos empleados, entre otros.
4. La primera validación de la herramienta fue positiva, pues el texto generado es altamente similar a la transcripción similar. Sin embargo, es necesario estudiar el comportamiento de la herramienta en casos menos ideales, en el que hay ruido externo en los audios, o las personas hablan con diferentes acentos, ritmos y pronunciaciones. También es un campo para explorar el analizar conversaciones (audios en los que 2 o más personas participan), identificando quién está hablando en cada ocasión.
5. Finalmente, algo deseable para la futura aplicación de esta herramienta sería remover su dependencia de servicios externos (Azure), que nos limitan a un proveedor en particular y representan costos adicionales. Una alternativa es la de entrenar un modelo de voz a texto por medio de herramientas de software libre y técnicas de aprendizaje de máquina.

Socialización

Este piloto ha sido socializado con la Dirección de Desarrollo Digital y con la Subdirección General Sectorial del DNP, destacando el potencial de la herramienta e identificando posibles aplicaciones para esta dentro de la entidad.