Universidad Nacional Autónoma de México
IIMAS

García Avendaño Martín Jesús

Minería de datos

Tratar de predecir el delito más común basado en los videos de tendencias de YouTube

introducción:

La enorme cantidad de información que depositamos en las plataformas que utilizamos a diario, ha provocado que lleguemos a un punto en el que internet nos conoce mejor que nosotros mismos, hoy en día nuestros perfiles en internet reflejan todos nuestros gustos y son un claro reflejo de lo que somos como individuos, si a muchos individuos les gusta un video este video se vuelve una tendencia, tomando en cuenta los miles de millones de videos de YouTube que miles de personas vean el mismo video significa que hay algo en el que les llama la atención puede ser algo como un video de un perrito o algo triste como un niño llorando sea cual sea el video o la característica que se encuentre en el esta refleja un sentimiento mutuo de interés en un grupo de personas, analizar el contenido de estos videos y observar como este impacta en la sociedad no es una tarea fácil. La ciencia de datos en estos últimos años ha desarrollado algoritmos que poco a poco son mas precisos y robustos para utilizar esta información en temas que ayuden a la sociedad

Propósito del estudio:

El propósito del estudio es hacer un acercamiento al uso de la información que se genera en las plataformas digitales para la solución de problemas de índole social, aplicando técnicas de ML y ciencia de datos en la búsqueda de crear técnicas y soluciones que ayuden al avance del aprovechamiento de la información que se encuentra en internet en problemas de la vida cotidiana.

Descripción del proyecto:

El proyecto busca encontrar si existe una relación entre las tendencias de YouTube y los crimines que se cometen en las ciudades de Nueva York y la CDMX, debido a que fue imposible hallar un registro nacional de crímenes que correspondiera a las fechas de las tendencias en YouTube se optó por tomar los crímenes de las ciudades más densamente pobladas de ambos países bajo la suposición de que esto sería un muestra estadística adecuada ya que estas ciudades albergan a gran parte de la población de ambos países además de ser ciudades cuya cobertura tecnológica y de acceso a internet es casi total, tomando los conjuntos de datos de tendencias de YouTube en estados unidos y México se busca entrenar un modelo de clasificación Naive Bayes multinomial con los títulos y descripciones de los videos que fueron tendencias, para que usando esta información el modelo sea capaz de determinar cuál será el delito más común en ese día, al abordar este problema como un problema de clasificación el modelo tomara como categorías los delitos que se cometieron en las ciudades de México y Nueva York, de esta forma el modelo será capaz de predecir cual será el delito más común en un día basado en las tendencias de YouTube.

Flujo de trabajo:

Exploración de los datos:

Los datos para utilizar en este proyecto son 4 conjuntos de datos. 3 extraídos de Kaggle y uno extraído del sitio de la procuraduría de la ciudad de México, 2 de los conjuntos de datos extraídos de Kaggle corresponden a las tendencias de videos en YouTube en México y Estados Unidos entre 2017 y 2018 los cuales contienen 40949 el set de USA y 40451 el de México cada uno correspondiente a un video que fue tendencia en el periodo de tiempo antes mencionado, dicho video tendencia esta descrito en 16 columnas entre las cuales destacan las columnas de **fecha, título del video, categoría del video y descripción del video**. Ambos conjuntos de datos solo muestran valores faltantes en la columna de categoría 570 para el set de datos de USA y 4224 para el de México. El tercer conjunto de datos de Kaggle corresponde a los delitos efectuados en la ciudad de Nueva York entre el 2017 y 2018 obtenidos del departamento de policía de la ciudad, este conjunto contiene 946583 filas lo que representa a 946583 delitos efectuados en la ciudad de Nueva York, cada delito esta descrito en 35 columnas entre las cuales las más relevantes son **fecha, descripción del delito, grupo de edad de la víctima, rango de edad de la víctima, lugar donde se llevó acabo la infracción.** El set de datos presenta múltiples valores faltante especialmente en las columnas de latitud, longitud, grupo de la víctima, sexo de la víctima, edad de la víctima y fecha de termino de la investigación, por suerte estas columnas no son de mucho interés para nuestro proyecto. El ultimo conjunto de datos se obtuvo de la secretaria de seguridad ciudadana de la CDMX el cual contiene 479923 registros de delitos efectuados en la cdmx entre el 2017 y el 2018 descritos en 18 columnas entre las que destacan **fecha, delito, alcaldía** ya que son las columnas de nuestro interés, existen otras columnas en el set de datos como la fecha en que se concluyó la investigación pero esa columna tiene mas valores NaN que registros, varias columnas repiten ese comportamiento por suerte las columnas de nuestro interés antes mencionadas no presentan valores nulos

Preparación de los datos:

Tendencias en YouTube USA:

Lo primero que se hizo para este conjunto de datos fue seleccionar las columnas de interés para la realización de nuestro proyecto en este caso las columnas seleccionadas fueron las de **titulo, descripción, id categoría y fecha.** Estas fechas fueron seleccionadas ya que se busca entrenar el modelo con las palabras almacenadas en el titulo y la descripción además la fecha en que un video o conjunto de videos y sus características fueron tendencias, la categoría por su parte nos dice de que tipo de video se trata y de esa forma las palabras en el titulo y la descripción del video tienen mas sentido y no son palabras vacías.

Una vez que se obtuvieron las columnas de interés lo que se hizo fue tomar las columnas de **titulo del video** y **descripción** del video para limpiar el texto que se encuentra en ellas para de esta forma aumentar la precisión del modelo ya que al limpiar el texto se eliminan palabras vacías (the, and,etc.) palabras que sabemos que se repetirán mucho en el texto y que no aportan nada al contenido de este pero que si afectan a la extracción de características cuando se desea entrenar el modelo, otra cosa que se hace a la hora de limpiar el texto es lematizar y reducir todo el texto a palabras minúsculas de esta forma reducimos variaciones de palabras

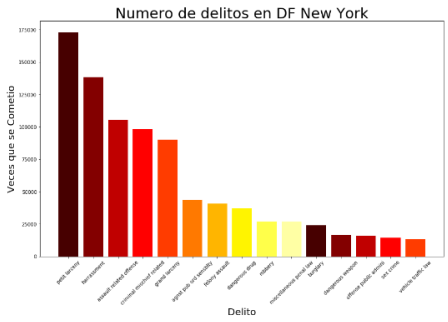
con un significado igual que para los modelos pueden ser tomadas como diferentes pero en realidad son lo mismo como ejemplo: brincar y brincando que son palabras cuyo significado es el mismo y que el modelo tomaría como distintas por lo que se procede a convertir ambas a brincar.

Para la columna de **id-categoría** esta se encontraba con valores que expresaban una categoría como ejemplo 20 = deportes, 34 = cine, etc. Sin embargo, había varios valores para una misma categoría por lo que mediante un diccionario estos valores distintos para una misma categoría se unificaron.

La columna de **Fecha** se encontraba en un formato que incluía la hora, el minuto y el segundo en que fue extraído el dato además de el año fecha y mes, pero esta fue convertida a un formato más amigable y fácil de trabajar de día mes y fecha.

Crimen en Nueva York:

Para el conjunto de datos de delitos en la ciudad de Nueva York primero se seleccionaron las columnas de **hora, descripción del delito, fecha, sexo, grupo y edad** la selección de estas columnas esta basada en que de este conjunto de datos solo necesitamos los delitos y su fecha ya que este solo servirá como un referente categórico para el modelo, sim embargo se seleccionaron mas columnas ya que estas se usaron para obtener algunos datos que pudieran resultar interesantes acerca de los delitos en la ciudad de nueva york por ejemplo:



Los delitos mas comunes en la ciudad de Nueva York en la que podemos observas que la mayoría de los delitos en esta ciudad están concentrados en robos pequeños y acoso, entre algunas otras relaciones implementadas en el notebook.

Unir dataframes:

Una vez que los 4 conjuntos de datos estuvieron limpios y con los formatos correctos había que unirlos de alguna forma de unirlos, ya que teníamos muchos delitos para cada día de y a su vez teníamos muchos videos para cada día era imposible realizar un análisis con tantas variables por lo que la solución fue tomar le delito más común por día y asignarlo a cada video que tuviera esa fecha. El primer paso fue contar los delitos ocurridos en una misma fecha del set de datos tanto de delitos en la CDMX como de delitos en la ciudad de Nueva York para después obtener el delito que presentaba la mayor cantidad de repeticiones y así asignar a ese día ese delito más común eso se hizo para 730 fechas de ambos conjuntos de datos.

Concatenar los conjuntos de datos:

Una vez que ya tuvimos el delito más común por día se procedió a la creación de un diccionario donde cada fecha era la llave y el valor era el delito más común para esa fecha de esta forma se tomaron los conjuntos de datos de tendencias en YouTube tanto de México como de USA y a cada video en base a la fecha en que fue tendencia se le asigno un delito más común estuvo fue bastante eficaz ya que de esta forma se encontraron valores de videos que se salían del rango establecido de 2017 a 2018, de esta forma se obtuvieron 211 videos en el caso de tendencias en USA y 11 en el de tendencias de México que se salían del rango establecido y por lo tanto no tenían un delito asignado, posteriormente se procedió a eliminar esos registros ya que al ser muy pocos la información del set de datos no se veía comprometida.

Al unir los datos se obtuvo esto:

	titulo	descripcion	fecha	categoria	Titulo + Descripcion	delito
	want talk marriage	shantells channel	11/13/2017	People and blogs	want talk marriage shantells channel	petit larceny
	trump presidency last week tonight john oliver...	one year presidential election john oliver dis...	11/13/2017	Entertainment	trump presidency last week tonight john oliver...	petit larceny
	racist superman rudy mancuso king bach lele pons	watch previous video nnsubscribe	11/12/2017	Comedy	racist superman rudy mancuso king bach lele po...	petit larceny

Evaluación de Modelos:

Crimen y YouTube USA:

Una vez que se limpiaron los datos y se unieron los conjuntos de datos se realizó una prueba para validar que tan bueno era el modelo para clasificar para esta prueba se tomo a las columnas de **texto, descripción** y se creó una nueva llamada **texto + descripción** la cual solo era una suma de los valores de texto más título, se entrenaron tres modelos de Nave Bayes Multinomial cada uno con una columna diferente como marco de referencia para la extracción de características y se evaluó su eficacia para clasificar el texto en una determinada categoría en la que cada categoría era la categoría que por defecto traían los videos de YouTube(deportes, música, etc.),obteniendo los siguientes resultados:

Titulo	Descripcion	Titulo + Descripción
96.08625	93.80531	94.87723

Después de comprobar que el modelo es bastante bueno para clasificar según el texto el siguiente paso fue probar su eficacia a la hora de clasificar el texto, pero ahora para los delitos mas comunes, de igual manera como se hizo en el ejemplo de prueba se evaluó el modelo con 3 diferentes columnas **título, descripción, titulo + descripción** arrojando los siguientes resultados:

Titulo	Descripcion	Titulo + Descripción
96.08625	93.80531	94.87723

Crimen y YouTube México:

Al igual que con el conjunto de datos anterior después de limpiar los datos se procedió a probar la eficacia del modelo para utilizar la información de textos en la clasificación, de igual forma se probó con **título, descripción y titulo + descripción** obteniendo los siguientes resultados:

Titulo	Descripcion	Titulo + Descripción
66.87371	62.05659	65.87992

Como se puede observar los resultados son peores que con el set de USA esto puede ser causado a que la lematización de las palabras en español es muy mala y pocas veces se obtiene un resultado correcto, es por ello que decidí aplicar la función de traducción la cual traducía todo el texto de las columnas **título, descripción** a inglés y comprobar los resultados obtenidos después de efectuar este cambio.

Titulo	Descripcion	Titulo + Descripción
41.66667	41.66667	39.58333

Los resultados son peores que si no se hubiera implementado la traducción por lo que decidí entrenar el modelo de Naive Bayes con el texto en español al comprobar que los resultados son mejores.

Una vez realizado el intento de mejorar el texto mediante traducción y fallar se procedió a intentar la traducción del delito mas frecuente en la cdmx basado en las tendencias de los videos de YouTube, el cual tomara como características las palabras que de los títulos y descripciones de los videos de YouTube y buscara predecir el delito basado en estas características, los resultados son los siguientes:

Titulo	Descripcion	Titulo + Descripción
41.8807	29.09417	37.03933

Análisis de resultados:

Modelo	Titulo	Descripcion	Titulo + Descripción
CDMX_Crimen	41.8807	29.09417	37.03933
YT_MX_Traducido	41.66667	41.66667	39.58333
YT_MX_Categoria	66.87371	62.05659	65.87992
YT_USA_Categoria	96.08625	93.80531	94.87723
NY_Crimen	84.9737132	80.6700086	85.1693361

Todos los resultados obtenidos con los modelos fueron simplificados en esta tabla como podemos observar se obtienen muy buenos resultados en especial para los modelos que fueron entrenados con palabras en inglés.

En los modelos que buscaban predecir la categoría de YouTube el entrenado con palabras en ingles nativas supero al que usaba palabras en español esto se debe principalmente al tema de la lematización y la variabilidad de la misma palabra que tienen las palabras en español y que afecta a la hora de extraer características, también podemos destacar que el titulo por si solo es el que presenta mejores resultados y esto se debe a que aunque el numero de palabras es menor los títulos son un poco mas generales comparados con la descripción por lo cual las características que se logran extraer son mejores.

Traducir el texto de las columnas de titulo y descripción resulto contra producente ya que se buscaba mejorar el modelo sin embargo sucedió todo lo contrario y la mala traducción obtenida por la función de traducción causo que el modelo se viese afectado ya que el texto resultado de la traducción es una combinación de palabras en inglés, español y otras tantas que no existen lo que hizo que el modelo se viese muy mal.

Comparando los resultados del clasificador para encontrar el crimen de la CDMX y New York podemos observar que la diferencia en cuanto a resultados es enorme y esto se puede argumentar tomando en cuenta que la cantidad de clases (delitos) que existen en Nueva York están casi nulamente distribuidas ya que mas del 50% son delito menor los cual causo que el modelo al ser una variante de naive bayes siempre apuntara al que mas repite y con una afectividad enorme ya que las posibilidades de acertar suponiendo delito menor son casi absolutos. Por el contrario los delitos en la cdmx se encuentran distribuidos de forma mas uniforme sin tener delitos que abarquen todos los números, también los delitos en la cdmx tienden a mas categorías ya que en el crimen de Nueva York solo 5 tenían un índice de ocurrencia no despreciable pero que aclaro era mucho menor que el que tenia robo pequeño, en la ciudad de México se tienen casi el doble de categorías con un numero de incidencia mas uniforme lo que causo la enorme diferencia entre los modelos

Respuesta a las hipótesis planteadas.

Tanto los resultados bajos como los resultados altos de dos de los clasificadores nos dan como resultado que la hipótesis no pudo ser cumplida y esto se debe a que los resultados altos en el conjunto de datos de USA muestran un claro sesgo del modelo ante patrones repetitivos lo que denota falta de robustez en el modelo, lo cual se complementa con los

resultados del set de datos de YouTube de México donde el aumento en el numero de variable provoca que el modelo se confunda.

En conclusión, si bien este proyecto fue un acercamiento a la respuesta el proyecto no arrojo los resultados esperados.

Conclusiones y recomendaciones

Aunque no se obtuvieron los resultados esperados ni se pudo cumplir la hipótesis planteada el proyecto sirve como una base sobre lo que se debe y no hacer al tratar de predecir un comportamiento social, en un futuro al abordar temas de razón social se deben incluir mas variables y trabajar con datos en tiempo real produciría un cambio además de intentar la predicción con redes neuronales puede mejorar los resultados.

Bibliografía:

https://www.kaggle.com/datasnaek/youtube-new

indice de criminalidad

https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva

indice de criminalidad

https://www.kaggle.com/manjeetsingh/retaildataset

diccionarios es-en:

https://github.com/facebookresearch/MUSE