P01

June 6, 2019

1 Reporte de práctica 1: Preparación de los datos

1.1 Origen de los datos

Los datos a analizar durante el curso fueron proporcionados por la Dra. Elisa y son de un festival de cine que se lleva a cabo anualmente en Colombia y se está empezando a difundir la idea en México. Se cuenta con la información de las boletas de inscripción al festival de los años 2015 al 2018. La información capturada varía y aumenta de acuerdo al año de su captura, es decir, la información de 2018 contiene más datos que la de 2015. A causa del crecimiento de la información cada año, la limpieza de los mismos me dejará con muchas celdas vacías.

Las preguntas con las que se pudiera trabajar serían:

- £Cuáles son los teléfonos móviles que se utilizan más los productores de acuerdo a la categoría que pertenecen?
- £Cuál es el género que más producciones tiene?
- £Cuál fue el medio más efectivo por el que la mayoría de los concursantes se enteraron del festival?
- £Cuáles son los países extranjeros más activos en el festival, en que categorías participan y que géneros prefieren?

1.2 Datos disponibles

Como los datos disponibles varían con cada año, buscaré describir los datos por cada archivo y posteriormente buscaremos el común de estos.

1.2.1 Año 2015

El primer año del que se tiene registro. En este festival solo se contaba con los siguientes 8 datos: * País: País de procedencia de la producción. * Categoría: Si el video es producido por infantiles, profesionales ó aficionados. * Nombre del corto. * Medio de envío: El método por el cual la producción del corto hizo llegar el video al festival. * Género: Categoría en la que el productor clasificó su cortometraje. * Imágenes: Incomprensible que clase de información querían obtener. En la mayoría de los casos está vacío este campo, en otros dice "video". * Referencia Celular: el dispositivo móvil con el que se grabó el cortometraje. * Sinopsis: Una breve sinopsis del cortometraje. Habrá que revisar si tiene un límite de caracteres.

1.2.2 Año 2016

En este año la información fue clasificada en diferentes hojas de cálculo de acuerdo a la categoría de los participantes, además se agregaron 2 categorías, Juvenil y SmarTIC, donde al parecer los de SmarTIC presentan algún tipo de discapacidad. Los datos capturados son los siguientes: * ID: Supongo que el número de registro. No se cuentan con todos los ID consecutivos. * Cómo se enteró * Nombre del corto * Género * Sinopsis * País * Ciudad: Cuenta con muchos espacios vacíos, si el participante no es de Colombia se deja vacío. * Departamento: Es como el concepto de estado en México. * Referencia celular

1.2.3 Año 2017

En este año se agregó una categoría llamada "Reto al guion" la cual contiene muy poca información y no es compatible con la información proporcionada por las demás categorías, será descartada. A partir ahora obtenemos información del que registro el cortometraje, pero también se agrega una mini encuesta de SI/NO al final de la forma de inscripción. También se decidió categorizar los Celulares utilizados, agregando dos elementos a la clasificación: "Tipo de dispositivo" y "marca". Los datos capturados son: * ID. * Cómo se enteró * Sexo * Edad * Colombiano: Me imagino que es otra forma de decir el país de procedencia. * Departamento: En la cual se capturó la ciudad. * (Campo vacío): En el cual se capturó el departamento.
* Nombre del corto * Género * Sinopsis * (Campo vacío): En el cual se repite el valor guardado en Departamento. * (Campo vacío): En el cual se repite el departamento. * Tipo de dispositivo * Marca: del celular * Referencia: cual celular usó.

- Campaña publicitaria. *A partir de aquí empieza la mini encuesta*
- making of: Si cuenta con detrás de cámaras.
- documental: Es tipo documental?. Lo cual considero innecesario ya que Documental es un género de cortometrajes.
- infantil: Además de que es una categoría de participantes, también es un género de cortometrajes. No entiendo la existencia de este campo.
- afiche: Según Google, es otra forma de decir cartel o póster. Me imagino que si tuvo campaña publicitaria debe tener póster.
- Cantidad de personas que realizaron el corto. Yo creo que es un campo muy ambiguo, ya que puede que la persona que realizó el corto sea solo una, pero en él participaron más de una. Como esta lleno este campo parece acertado que se refiera a las personas que participaron y no que lo realizaron.

1.2.4 Año 2018

En este año tenemos los datos de dos países, Colombia y México.

Colombia Para este año tenemos los datos concentrados y no en diferentes hojas de cálculo. Los datos son los siguientes: * Carpeta: Una cadena de caracteres, parecen cifrados. * Categoría * £Como se enteró? * Edad Participante * Colombiano * País * Departamento * Ciudad * Nombre Corto * Género Corto * Duración: Tal parece que en segundos, por la magnitud de las cifras, pero algunos son tan pequeños que pudieron haberlo puesto en minutos. * Sinopsis * País del corto: decidieron categorizar el cortometraje. * Departamento del corto * Ciudad corto * Tipo de dispositivo * Marca Dispositivo * Referencia Dispositivo * Otro dispositivo: Se abre la posibilidad de usar

más de un dispositivo, antes se usaba una barra para dividir diferentes dispositivos pero ahora se pueden poner en dos campos diferentes. * Corto Días: tiempo de producción. * Corto Marcas: No tengo ni idea que signifique esto. * Personas * Discapacidad: Debido a que se concentraron los datos, en la misma lista están los de SmarTIC y ellos tienen discapacidad, los demás no.

México Se realizó un festival en México con la misma mecánica que en Colombia y aquí se almacenaron los siguientes datos: * Categoría: Parece ser que no se les explicó a los participantes que la categoría era donde iban a inscribir su cortometraje, no el género del mismo. * Edad participante * País participante * Departamento participante * Ciudad participante * Nombre de Corto * Género * Sinopsis * País * Departamento * Ciudad * Tipo de dispositivo * Marca de dispositivo * Referencia dispositivo * Días * Personas

1.3 Preprocesamiento

Los datos me fueron proporcionados en documentos .xlsx (extensión de Office Excel 2013 o superior), al momento de procesarlo en bash, la información proyectada era incomprensible. Googleando un poco, encontré el método para convertir el .xlsx a .csv y tratar la información.

Primero instalé el GNUmeric

```
In [ ]: $ sudo apt-get gnumeric
```

Después pasé a la conversión de los documentos.

```
In [ ]: $ ssconvert 2015.xlsx 2015.csv
        $ ssconvert 2016.xlsx 2016.csv
        $ ssconvert 2017.xlsx 2017.csv
        $ ssconvert 2018.xlsx 2018.csv
        $ ssconvert 2018mx.xlsx 2018mx.csv
```

Al final los archivos ya eran tratables en bash. Para trabajar con los archivos me topé con que estaban muy sucios los datos, debido a la conversión, por eso modifique los parámetros del convertidor y agregué la siguiente instrucción a cada uno.

```
In [ ]: --export-options="separator=ą"
```

Decidí usar el separador "ą" debido a que las comas, y otros símbolos eran utilizados en títulos de los cortos o en las sinopsis de los cortos. Consideré que era el signo que menos se utilizaba. Al final usé el comando *awk* para obtener la información de cada archivo. Además cambie la extensión de salida a .txt

Las columnas que utilicé y que consideré en común fueron: * Año (Columna agregada) * Categoría * País * Género * £Cómo se enteró? * Referencia celular

Los datos requieren limpiarse aún más, ya que al usar el comando:

```
In [ ]: awk -F 'ą'  {' if(NF > 15)  print $2'} 2015.txt | sort | uniq -c
```

```
  8
309 Aficionado
  1 Categoría
 73 Infantil
143 Profesional
```

aún me arroja espacios vacíos.

1.4 Conclusiones

En esta práctica pude aprender a utilizar la plataforma de Jupyter Notebook, el bash de linux y a convertir archivos desde terminar. A pesar del preprocesamiento realizado, creo que tengo que buscar otro símbolo para realizar la separación de columnas. La información obtenida en el año 2015 es prácticamente el cuello de botella de la información que podría procesar.

--04 de junio 2019-- Luis Angel Gutiérrez Rodríguez 1484412