Capítulo 14: Predicción de los Oscars a través de Twitter

Leslie Karen Castillo Alvarado

Introducción

En el Capítulo 14 del libro "Automated Data Collection with R", que muestra cómo los datos de redes sociales, en particular Twitter, pueden ser utilizados para realizar predicciones sobre eventos del mundo real.

El capítulo se enfoca en el caso específico de los premios Oscars 2014 y plantea la hipótesis de que el volumen de menciones en Twitter podría estar relacionado con los ganadores. Para ello, los autores se enfocaron en tres categorías principales de los Oscars:

- Mejor actor
- Mejor actriz
- Mejor película

En el capítulo original, se utilizaron los paquetes `twitteR` y `streamR` para acceder a la API de Twitter y recolectar tweets de forma directa. Posteriormente, se realizaron búsquedas de menciones de los nominados en las tres categorías, se contaron las menciones y se graficó la cantidad de tweets por hora. Sin embargo, debido a los cambios recientes en las políticas de acceso a la API (actualmente gestionada por la empresa X), este procedimiento ya no es factible sin suscripciones de pago. Por ello, se optó por generar un conjunto de datos simulado que emula el comportamiento de usuarios comentando en tiempo real sobre el evento.

¿Qué es la minería de texto?

La minería de texto (text mining) es una técnica del análisis de datos que permite extraer información útil a partir de textos no estructurados. Esta práctica se basa en identificar patrones, frecuencias, relaciones semánticas o sintácticas dentro de grandes volúmenes de texto, como artículos, redes sociales o comentarios.

En este proyecto, aplicamos minería de texto sobre publicaciones de Twitter para identificar cuántas veces se mencionan ciertos nombres clave, como actores, actrices y películas nominadas a los premios Oscars 2014. Esto nos permite analizar tendencias de conversación, estimar popularidad y comparar los resultados con los ganadores reales del evento.

El uso de minería de texto en redes sociales es especialmente valioso, ya que estas plataformas reflejan en tiempo real las opiniones del público, lo que convierte cada tweet en un dato potencial para análisis social, cultural y comercial.

¿Por qué Twitter?

Twitter es una de las plataformas más utilizadas para comentar eventos en tiempo real. Su relevancia como fuente de datos se debe a:

- Su carácter público y abierto (en comparación con redes privadas como Facebook o WhatsApp).
- Su uso masivo durante eventos de gran interés como los Oscars, elecciones o partidos deportivos.
- La facilidad para analizar sus datos mediante técnicas de minería de texto.

Aplicaciones reales de este tipo de análisis

El uso de minería de texto para eventos en redes sociales ha tenido aplicaciones prácticas en diversas áreas:

- En campañas políticas para evaluar menciones de candidatos.
- En el cine y entretenimiento para medir la expectativa y recepción de películas.
- En el análisis de crisis o desastres naturales, para detectar reacciones tempranas y coordinar respuestas.

En todos estos casos, el análisis de grandes volúmenes de texto permite generar conocimiento a partir de expresiones espontáneas del público, ofreciendo una perspectiva complementaria a las encuestas o análisis tradicionales.

1. Origen de los Datos

Como ya se mencionó, debido a las restricciones actuales de acceso a la API de Twitter, se generó un conjunto de datos simulado que contiene 300 tweets distribuidos aleatoriamente entre el 28 de febrero y el 2 de marzo de 2014. Estos tweets hacen referencia a actores, actrices y películas nominadas en los Oscars 2014.

El dataset se encuentra en formato CSV bajo el nombre `oscars_tweets_big.csv` y contiene las siguientes columnas:

- `created_at`: fecha y hora de publicación del tweet (por ejemplo: "2014-02-28 10:38:00")
- `text`: contenido del tweet (por ejemplo: "Gravity was a masterpiece, definitely deserves Best Picture. #Oscars2014")

Este conjunto de datos permite replicar la estructura del análisis del libro, manteniendo el enfoque en la minería de texto y el comportamiento temporal de las menciones.

Metodología

En base al dataset de 300 tweets simulados, se realizó un análisis de texto buscando las menciones exactas y aproximadas de:

- Actores: Leonardo DiCaprio y Matthew McConaughey
- Actriz: Cate Blanchett
- Películas: Gravity y 12 Years a Slave

Se graficó la cantidad de tweets por hora para visualizar el comportamiento temporal de la conversación.

2. Código completo

Instalación y carga de librerías necesarias

```
install.packages("lubridate")
install.packages("stringr")
install.packages("plyr")
install.packages("ggplot2")
```

```
library(lubridate)   # Para manejo de fechas y horas
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(stringr)     # Para búsqueda de texto dentro de strings
library(plyr)        # Para convertir listas a data.frame fácilmente
library(ggplot2)     # Para crear gráficas profesionales
```

Carga de datos y vista previa

```r
oscars <- read.csv("oscars_tweets_big.csv") # Cargar dataset de tweets desde archivo CSV
#View(oscars)  # Visualiza el dataset cargado
head(oscars)  # muestra las primeras filas del dataset
```

```
##            created_at
## 1 2014-02-28 10:18:00
## 2 2014-02-28 10:27:00
## 3 2014-02-28 10:38:00
## 4 2014-02-28 10:50:00
## 5 2014-02-28 11:44:00
## 6 2014-02-28 11:46:00
##                                                                    text
## 1                 Can't wait to see who wins Best Picture tonight! #Oscars
## 2                          Blanchett all the way for Best Actress! #Oscars
## 3 Gravity was a masterpiece, definitely deserves Best Picture. #Oscars2014
## 4                   Matthew McConaughey has been incredible lately #Oscars
## 5                     McConaughey killed it in Dallas Buyers Club. #Oscars
## 6                           12 Years a Slave is so powerful. Hope it wins!
```

Preparación de fechas para análisis temporal

```r
oscars$time <- as.POSIXct(oscars$created_at, tz = "UTC")
oscars$round_hour <- round_date(oscars$time, unit = "hour")
```

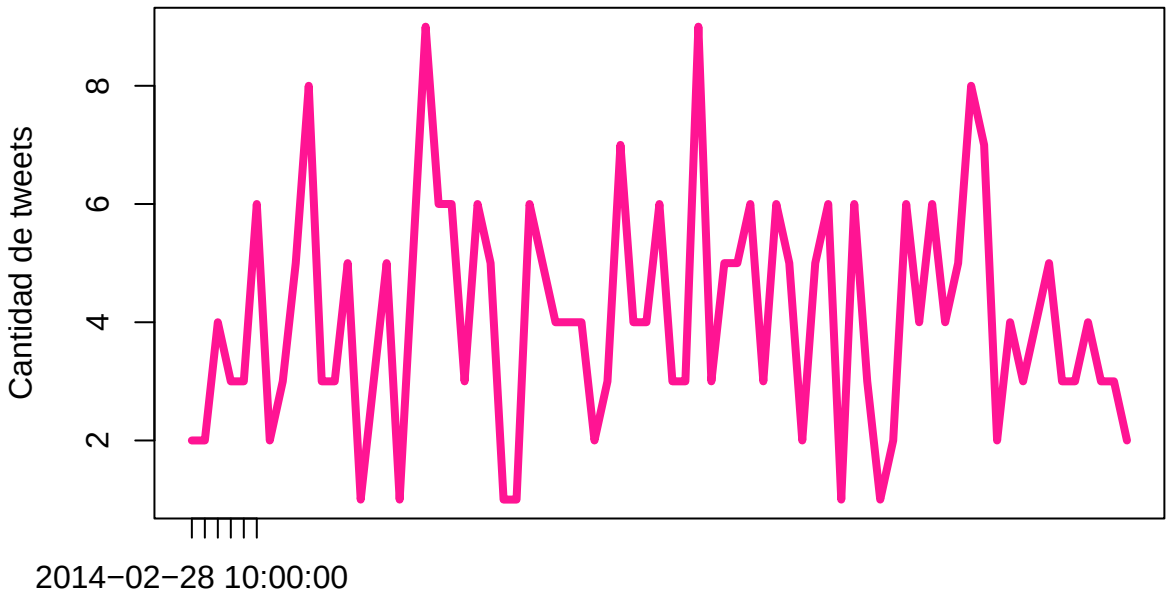
Crear tabla de tweets por hora

```r
plot_time <- as.data.frame(table(oscars$round_hour))
```

Conteo y visualización de tweets por hora

```r
plot(plot_time[,2],
     type = "l",
     xaxt = "n",
     xlab = "Hora",
     ylab = "Cantidad de tweets",
     col = "deeppink",      # <- color rosa
     lwd = 4)               # grosor de línea

axis(1,
     at = c(1, 2, 3, 4, 5, 6),
     labels = plot_time[c(1, 2, 3, 4, 5, 6), 1])
```



Esta gráfica representa la cantidad de tweets publicados por hora relacionados con los Oscars 2014, según mi dataset simulado, así podemos ver cómo fue cambiando la frecuencia de publicaciones en el tiempo.

Ejes de la gráfica:

- Eje X: representa el tiempo, específicamente las horas entre el 28 de febrero y los días siguientes (hasta el 2 de marzo aprox.).
- Eje Y: representa la frecuencia, es decir, el número de tweets publicados en cada hora.

En este caso, notamos lo siguiente:

- Variabilidad natural: A diferencia de una distribución uniforme, esta gráfica muestra subidas y bajadas, refleja cómo la gente comenta más o menos dependiendo de lo que está pasando en el evento.
- Picos de actividad: Hay momentos en los que la línea alcanza valores cercanos a 8 o 9 tweets por hora. Eso podría coincidir (en un caso real) con momentos clave del evento, como la entrega de un premio o la presentación de un actor famoso.
- Momentos bajos: También hay horas donde apenas hubo 1 o 2 tweets. Esto podría ser de madrugada o cuando el evento aún no empezaba.
- Tendencia dispersa: No hay una curva suave sino que hay subidas y bajadas, lo cual es normal en redes sociales, donde el flujo de mensajes cambia rápido según lo que esté ocurriendo en tiempo real.

Es importante mencionar que la gráfica no nos dice quién va a ganar el Oscar, pero sí cuándo la conversación estuvo más activa, lo que puede ser útil para cruzar con menciones específicas de actores o películas.

3. Análisis de menciones por palabra clave

```
oscars$lotext <- tolower(oscars$text)  # Convertir texto a minúsculas

actor <- c("leonardo dicaprio", "matthew mcconaughey")
actress <- c("cate blanchett")
film <- c("gravity", "12 years a slave")

dat_actor <- ldply(lapply(oscars$lotext, str_detect, actor))
```

```r
dat_actress <- ldply(lapply(oscars$lotext, str_detect, actress))
dat_film <- ldply(lapply(oscars$lotext, str_detect, film))

colnames(dat_actor) <- c("dicaprio", "mcconaughey")
colnames(dat_actress) <- c("blanchett")
colnames(dat_film) <- c("gravity", "twelve_years_slave")
```

```r
apply(dat_actor, 2, sum)
```

```
##    dicaprio mcconaughey
##          25          30
```

```r
apply(dat_actress, 2, sum)
```

```
## blanchett
##        37
```

```r
apply(dat_film, 2, sum)
```

```
##            gravity twelve_years_slave
##                 53                 49
```

Búsqueda aproximada (para errores ortográficos)

```r
length_actor <- unlist(lapply(lapply(actor, agrep, oscars$lotext), length))
length_actress <- unlist(lapply(lapply(actress, agrep, oscars$lotext), length))
length_film <- unlist(lapply(lapply(film, agrep, oscars$lotext), length))

names(length_actor) <- c("dicaprio", "mcconaughey")
names(length_actress) <- c("blanchett")
names(length_film) <- c("gravity", "twelve_years_slave")

length_actor
```

```
##    dicaprio mcconaughey
##          25          30
```

```r
length_actress
```

```
## blanchett
##        37
```

```r
length_film
```

```
##            gravity twelve_years_slave
##                 53                 49
```

Esta sección busca menciones que podrían estar mal escritas o con ligeras variaciones (como “dicaprio” sin mayúsculas). Esto mejora nuestro análisis.

4. Conclusiones

El análisis permite replicar la metodología planteada en el libro, adaptándolo a un entorno moderno, superando las limitaciones actuales del acceso a Twitter. El uso de un dataset simulado permitió desarrollar los objetivos principales del estudio: analizar el comportamiento del público en redes sociales mediante la minería de texto, detectar tendencias y representar gráficamente la conversación.

Se observó que actores como Leonardo DiCaprio y Matthew McConaughey fueron mencionados frecuentemente, así como las películas “Gravity” y “12 Years a Slave”. Estos patrones son coherentes con la popularidad de los nominados en ese año.

Con los datos obtenidos, concluimos que el análisis de menciones en Twitter puede proporcionar indicios sobre los ganadores de un evento, pero no garantiza una predicción exacta, ya que el volumen de menciones no siempre determina el resultado final. A pesar de esto, pudimos ver que este tipo de análisis nos permite entender mejor la relación entre la conversación pública y los resultados de eventos de interés masivo.