

Análisis de la duración de la fama de la página de Nicolás Copérnico en la Wikipedia española

JJ Merelo

10 de enero de 2017

Abstract

A consecuencia de un vídeo que se ha hecho viral en estos días y en los que se menciona a Copérnico, la figura ha recibido una cantidad considerable de atención en estos días, pero fue decayendo rápidamente en los días siguientes hasta desaparecer de los primeros puestos a los tres días. Veamos exactamente cuanta y cómo ha evolucionado a lo largo del tiempo.

Introducción

La wikipedia, a través de un interfaz, permite acceder a las visitas de cada una de las diferentes páginas creadas en ella. Es un interfaz de tipo REST al que se puede acceder directamente. Lo hacemos mediante este *script*, que, como el resto de los datos y programas, son software libre y se pueden descargar de la página que figura en las conclusiones.

```
curl https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/es.wikipedia/all-access/all-agents,
curl https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/es.wikipedia/all-access/all-agents,
```

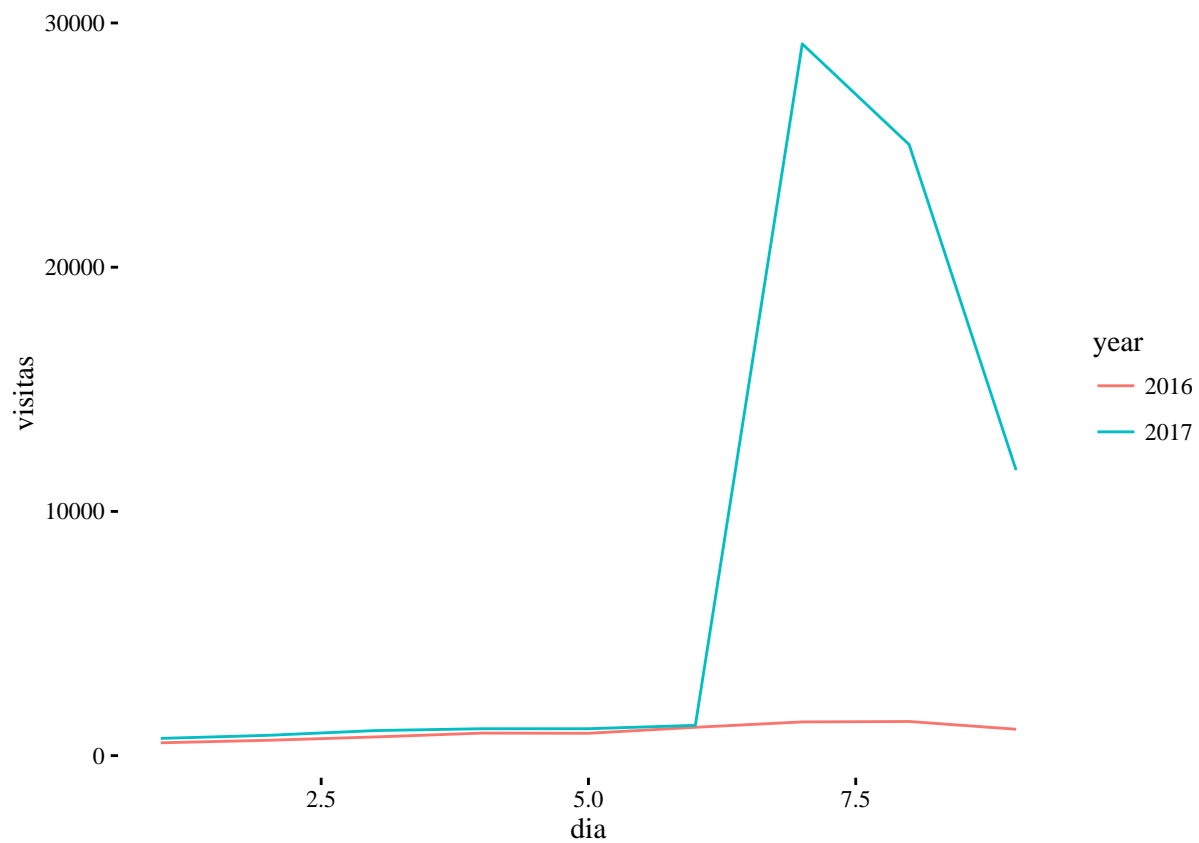
que usa `curl` para descargarse el fichero, sin más complicación que codificar como URI el nombre de la página, justo detrás de `all-agents`, y poner como dos últimos parámetros las dos fechas en las que nos movemos, desde principios de año hasta ahora. Como esto descarga un fichero en JSON, extraemos el dato que nos interesa usando la utilidad `jq`, que permite hacer búsquedas complejas sobre ficheros JSON: las visitas, y lo vertimos en sendos ficheros de datos.

Estos datos los vamos a procesar con R para convertirlos en un sólo conjunto:

```
library(ggplot2)
library(ggthemes)
datos.2016 <- read.table('2016.dat')
datos.2017 <- read.table('2017.dat')
datos <- data.frame( dia=c(1:nrow(datos.2016),1:nrow(datos.2017)),
                    year=c(rep('2016',nrow(datos.2016)),rep('2017',nrow(datos.2017))),
                    visitas=c(datos.2016$V1,datos.2017$V1))
```

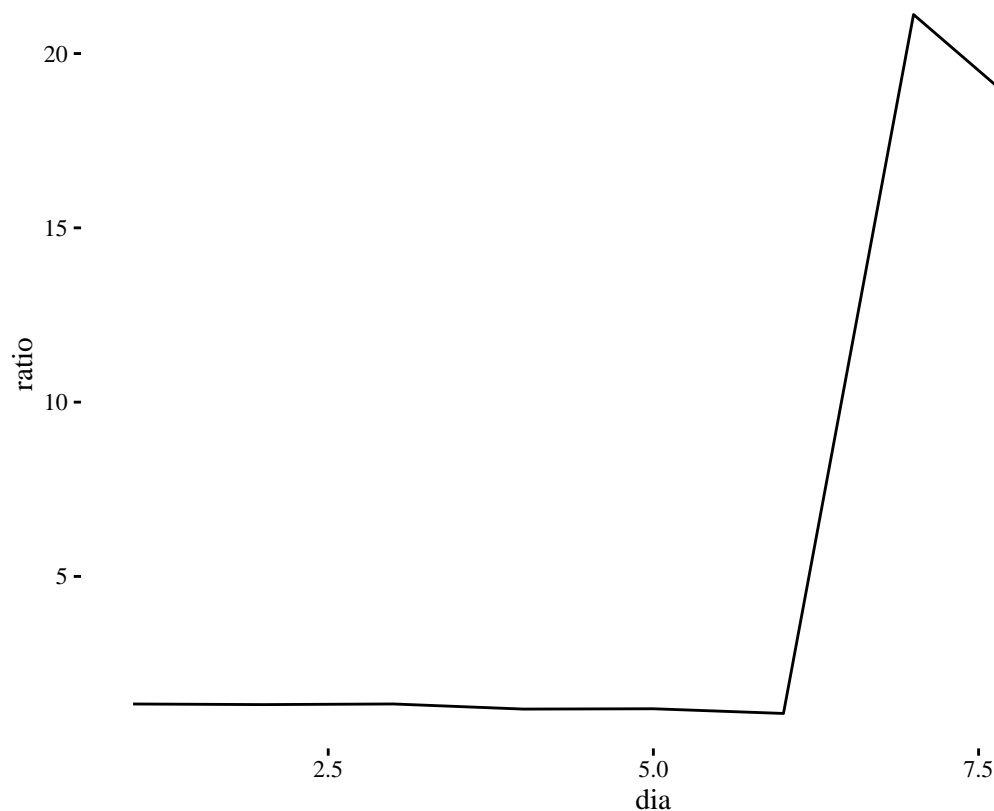
A continuación trazamos los datos de cada día, por año.

```
ggplot(datos,aes(x=dia,y=visitas,group=year,col=year))+geom_line()+theme_tufte()
```



Se nota que, tras alcanzar un pico el primer día, desciende el segundo y mucho más rápidamente el tercero. Vamos a ver la relación de visitas entre el año pasado y el corriente

```
ratio <- data.frame( dia=c(1:nrow(datos.2016)),ratio=datos.2017$V1/datos.2016$V1)
ggplot(ratio,aes(x=dia,y=ratio))+geom_line()+theme_tufte()
```



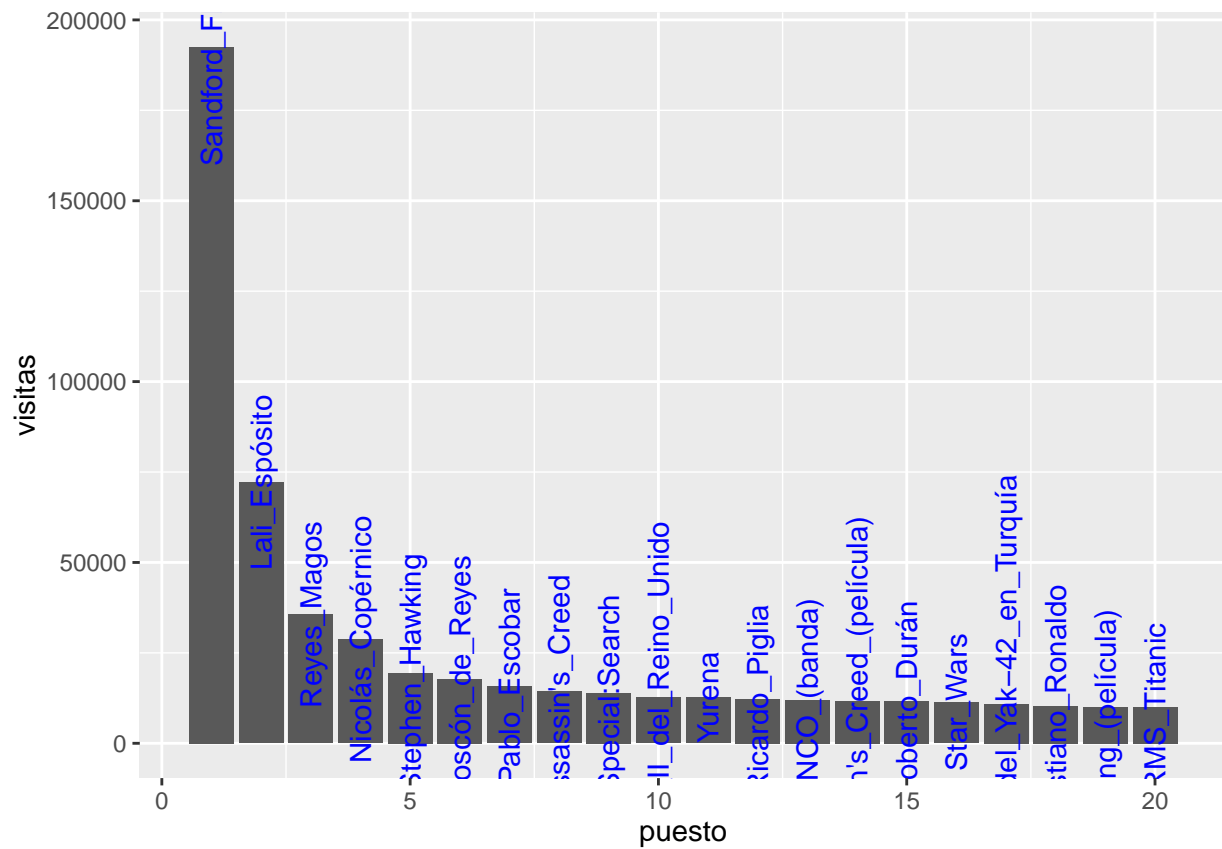
De tener un número de visitas parecido, aunque ligeramente superior, se pasa a tener 20 veces más visitas, número que va descendiendo hasta 10 veces a los tres días del comienzo de la notoriedad.

Evolución en el ranking de páginas visitadas

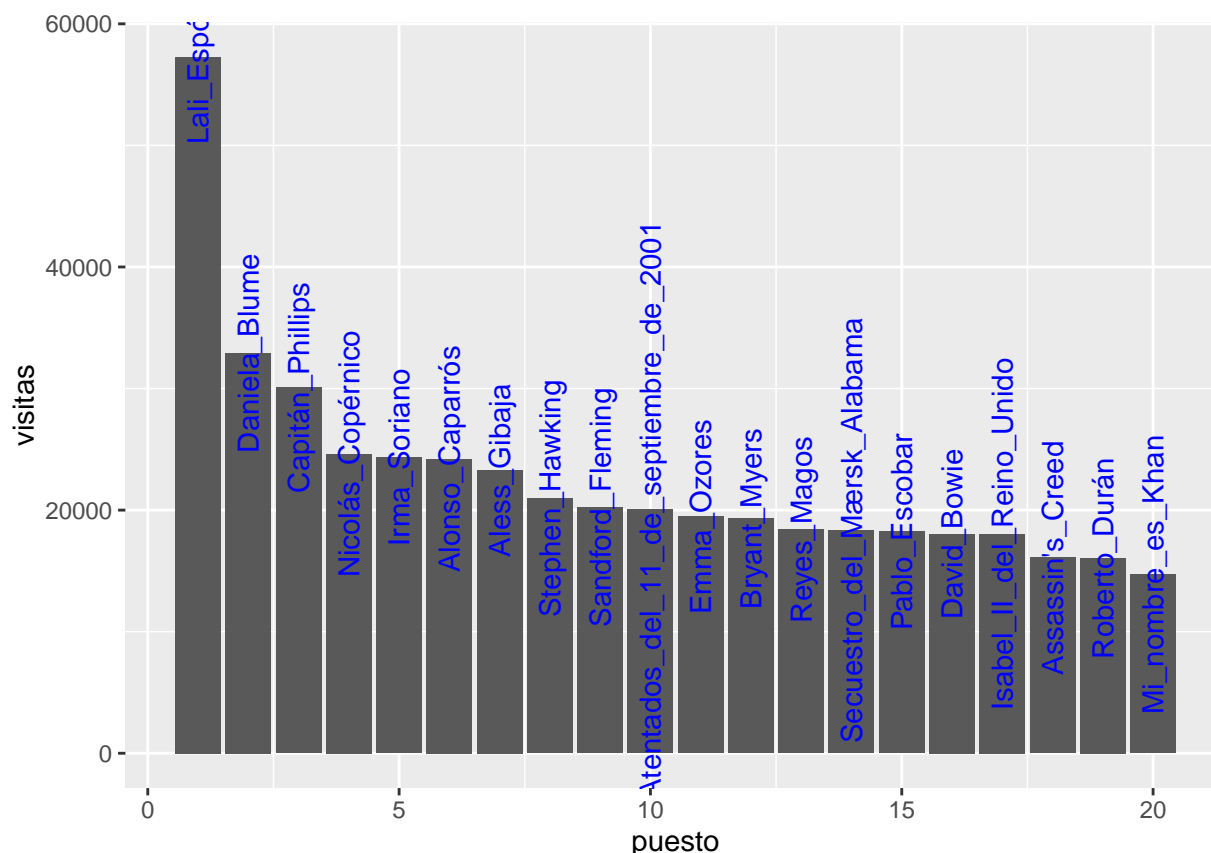
La wikipedia también ofrece en su API el ranking de las mil primeras páginas vistas, que se extraen con este script: `~ curl https://wikimedia.org/api/rest_v1/metrics/pageviews/top/es.wikipedia/all-access/2017/01/07 | jq '[.items[] | .articles[] | {articulo:.article,puesto: .rank,visitas:.views}]'` > ranking-070117.json ~ que genera, en formato JSON, un fichero con las visitas, el puesto que ocupa cada artículo y el artículo en sí. Esto se repite para el resto de los días y a continuación el fichero se filtra dejando solamente los artículos de la Wikipedia y eliminando páginas especiales como la entrada o la página de búsqueda.

En el *script* siguiente lo leemos y lo trazamos para dos días seguidos comenzando en el día 7, además, filtramos las páginas *especiales* de la wikipedia dejando solamente los artículos.

```
library(jsonlite)
ranking.07 <- fromJSON("ranking-070117-filtrado.json")
ggplot(ranking.07[1:20,],aes(x=puesto,y=visitas)) + geom_bar(stat="identity")+ geom_text(aes(label=arti
```



```
ranking.08 <- fromJSON("ranking-080117-filtrado.json")
ggplot(ranking.08[1:20,],aes(x=puesto,y=visitas)) + geom_bar(stat="identity")+ geom_text(aes(label=artículo))
```



De no estar entre las 1000 páginas con más visitas el día anterior, ha entrado directamente en el número 4, ligeramente detrás de los Reyes Magos el día 7 y de la película que, al parecer, estaban poniendo en alguna televisión, Capitán Phillips. Al día siguiente, día 9, el artículo de Copérnico cayó hasta el puesto 49 del ránking filtrado.

Conclusión

Las visitas a día equivalente se han multiplicado por 20, más o menos, el primer día. Los días siguientes se va perdiendo interés.

Todos los datos y ficheros necesarios están en <http://github.com/JJ/Cop-rnico-visitas>, con una licencia libre. Este artículo extiende (Merelo 2017a) y (Merelo 2017b) añadiendo los datos para el tercer día, filtrando los rankings de forma que solamente aparezcan los datos de páginas y añadiendo un gráfico con la diferencia de visitas entre el año anterior y este.

Como trabajo por hacer, habría que continuar viendo la evolución temporal y también encontrar correlación con otras páginas. Sería interesante ver si esta evolución se corresponde también con la de otras páginas cuya notoriedad ha aparecido en las redes sociales.

Bibliografía

- Merelo, Juan J. 2017a. “Copérnico Famoso Por Un Día.” 1. GeNeura team, Universidad de Granada.
- . 2017b. “Evolución de La Notoriedad de La Página de Nicolás Copérnico En La Wikipedia Española.” 2. GeNeura team, Universidad de Granada.