

Wikia census: cleaning and analysis

Abel Serrano Juste

28 de diciembre, 2018

Índice general

Introducción	1
Problema a resolver	2
Fuentes de datos	2
Descripción del Dataset	2
Wikia census dataset	2
Wikia page views dataset	7
Limpieza	8
Transformación	10
Integración	11
Transformación	13
Análisis	13
Modelo de predicción para wikis visitadas	19
Visualización	22
Resultados	23
Referencias	23

Introducción

En esta práctica, vamos a realizar un proyecto analítico de ciencia de datos sobre el ecosistema de las wikis.

Según la definición de Wikipedia para wiki:

El término wiki (proviene del hawaiano wiki, «rápido») alude al nombre que recibe una comunidad virtual, cuyas páginas son editadas directamente desde el navegador, donde los mismos usuarios crean, modifican, corrigen o eliminan contenidos que, generalmente, comparten.

Las wikis son un interesante objeto de estudio puesto que permiten investigar la colaboración masiva de usuarios online para crear un contenido común.

Utilizaremos los términos usuario y editor indistintamente, puesto que en el contexto de una wiki se pueden entender como sinónimos.

Problema a resolver

Analizar la actividad y diversidad de las wikis alojadas en el servicio Wikia.

Fuentes de datos

Para esta práctica, vamos a usar dos datasets: **Wikia census** y **Wikia page views**; ambos disponibles en mi cuenta de kaggle.

1. El censo de Wikia. Se trata de un dataset de un conjunto de 300k wikis que corresponde a todas las wikis alojadas en Wikia. Este dataset contiene datos descriptivos de cada wiki como: número de páginas, número de usuarios, número de ediciones, etc. Los métodos de extracción y la información proporcionada en este censo está explicada en el paper: A Wikia census: motives, tools and insights (Jimenez-Diaz, Serrano, y Arroyo 2018).
2. Wikia page views. Se trata de una captura de datos, realizada mediante web scrapping de todas las wikis de Wikia, que contiene el número de visitas para cada una de las wikis de Wikia en las últimas cuatro semanas. Este dataset se obtuvo para la práctica anterior de esta misma asignatura y el código fuente para su obtención está en el este repositorio de Github: https://github.com/Akronix/scrap_wikia_page_views.

Descripción del Dataset

Como hemos explicado previamente, vamos a usar dos datasets: el censo de wikia y los números de páginas visitadas.

Wikia census dataset

El primer paso consistirá en cargar los datos:

```
# Cargamos el juego de datos
wikis<-read.csv("data/20181019-wikia_stats_users_birthdate.csv",header=T,sep=",")
```

A continuación, haremos una breve descripción de los datos, ya que nos interesa tener una idea general de los datos que disponemos. Para ello, primero calcularemos las dimensiones de nuestra base de datos y mostraremos una muestra de los datos para interpretar qué tipos de atributos tenemos.

```
dim(wikis)
```

```
## [1] 277795      32
```

Disponemos de datos de 277795 wikis (filas) con 32 atributos sobre cada uno de ellos (columnas).

```
head(wikis)
```

```
##                                     url
## 1 http://spellmagotm.wikia.com
## 2 http://2017-monster-energy-nascar-cup-series.wikia.com
## 3 http://10low46japreligion.wikia.com
## 4 http://indigo-showdown.wikia.com
## 5 http://animewiki2.wikia.com
## 6 http://elena-ofavalar-fans.wikia.com
##           creation_date                  domain
## 1 2012-05-01 13:58:13 spellmagotm.wikia.com
```

```

## 2 2017-07-24 22:35:31 2017-monster-energy-nascar-cup-series.wikia.com
## 3 2009-09-15 23:21:34 10low46japreligion.wikia.com
## 4 2014-05-30 15:43:23 indigo-showdown.wikia.com
## 5 2011-02-18 23:15:14 animewiki2.wikia.com
## 6 2018-09-12 17:52:40 elena-ofavalor-fans.wikia.com
##   founding_user_id headline      hub      id lang language
## 1          5069110      Games 529058    en    en
## 2          32529801       TV 1601247    en    en
## 3          1602876  Lifestyle 52061    en    en
## 4          25001469      Games 982346    en    en
## 5          1160460       TV 221590    en    en
## 6          36888175       TV 1806604    en    en
##                                     name stats.activeUsers
## 1           Spellmagotm Wiki 0
## 2 2017 Monster Energy NASCAR Cup Series Wiki 0
## 3 Ancient Japanese Religion (Daramalan Assignment) Wiki 0
## 4           Indigo showdown Wiki 0
## 5           Animewiki2 Wiki 2
## 6 Elena ofavalor fans Wiki -1
##   stats.admins stats.articles stats.discussions stats.edits stats.images
## 1          1          8        0      275     11
## 2          1         45        0      230     40
## 3          1         71        1      470     20
## 4          1         17        1      242     11
## 5          1        102        NA     6365    4938
## 6          1          1        0       85      0
##   stats.pages stats.users stats.videos
## 1          229 16246020        0
## 2          153 16555972        0
## 3          310 16117576        4
## 4          156 16387583        2
## 5          5305 15397437        2
## 6          83 15565473        0
##                                     title      topic
## 1           Spellmagotm Wiki Gaming
## 2 2017 Monster Energy NASCAR Cup Series Wiki Entertainment
## 3 Ancient Japanese Religion (Daramalan Assignment) Wiki Philosophy
## 4           Indigo showdown Wiki Gaming
## 5           Animewiki2 Wiki Anime
## 6 Elena ofavalor fans Wiki Entertainment
##   wam_score stats.nonarticles users_1 users_5 users_10 users_20 users_50
## 1 0.0000          221      5      1      1      1      1
## 2 0.0000          108      6      3      1      1      1
## 3 0.0000          239      7      2      1      1      1
## 4 0.0000          139      9      4      3      3      1
## 5 0.0161          5203     33     21     16     14      9
## 6 0.0000          82       3      0      0      0      0
##   users_100 bots      birthDate datetime.birthDate
## 1          0   5 14:14, May 1, 2012 2012-05-01 14:14:00
## 2          1   4 22:35, July 24, 2017 2017-07-24 22:35:00
## 3          1   5 23:21, September 15, 2009 2009-09-15 23:21:00
## 4          0   5 15:43, May 30, 2014 2014-05-30 15:43:00
## 5          6   3 23:15, February 18, 2011 2011-02-18 23:15:00
## 6          0   2 17:52, September 12, 2018 2018-09-12 17:52:00

```

```

str(wikis)

## 'data.frame': 277795 obs. of 32 variables:
## $ url      : Factor w/ 277795 levels "http://0002oifos.wikia.com",...: 225470 532 147 126585 9385 594
## $ creation_date : Factor w/ 276956 levels "2001-01-15 00:00:00",...: 53387 183565 10462 107722 30199 27
## $ domain    : Factor w/ 277794 levels "0002oifos.wikia.com",...: 225469 532 147 126584 9385 59413 232
## $ founding_user_id : num 5069110 32529801 1602876 25001469 1160460 ...
## $ headline   : Factor w/ 8152 levels "", ., ..., 1 1 1 1 1 1 1 1 1 ...
## $ hub       : Factor w/ 8 levels "Books", "Comics", ...: 3 8 4 3 8 8 4 8 8 3 ...
## $ id        : num 529058 1601247 52061 982346 221590 ...
## $ lang      : Factor w/ 79 levels "aa", "af", "am", ...: 20 20 20 20 20 20 20 20 20 20 ...
## $ language   : Factor w/ 79 levels "aa", "af", "am", ...: 20 20 20 20 20 20 20 20 20 20 ...
## $ name      : Factor w/ 258804 levels "", ., ..., 175995 608 8765 90942 10708 56819 174447 217529 1
## $ stats.activeUsers : num 0 0 0 0 2 -1 0 0 0 0 ...
## $ stats.admins   : num 1 1 1 1 1 1 1 1 1 ...
## $ stats.articles  : num 8 45 71 17 102 1 6 7 6 4 ...
## $ stats.discussions : num 0 0 1 1 NA 0 1 0 0 0 ...
## $ stats.edits     : num 275 230 470 242 6365 ...
## $ stats.images    : num 11 40 20 11 4938 ...
## $ stats.pages     : num 229 153 310 156 5305 ...
## $ stats.users     : num 16246020 16555972 16117576 16387583 15397437 ...
## $ stats.videos    : num 0 0 4 2 2 0 0 0 0 0 ...
## $ title         : Factor w/ 258804 levels "", ., ..., 175995 608 8765 90942 10708 56819 174447 217529 1
## $ topic          : Factor w/ 29 levels "", "Anime", "Auto", ...: 12 8 19 12 2 8 16 9 9 1 ...
## $ wam_score      : num 0 0 0 0 0.0161 0 0 0 0 0 ...
## $ stats.nonarticles : num 221 108 239 139 5203 ...
## $ users_1        : int 5 6 7 9 33 3 9 7 4 4 ...
## $ users_5        : int 1 3 2 4 21 0 1 2 0 1 ...
## $ users_10       : int 1 1 1 3 16 0 0 2 0 1 ...
## $ users_20       : int 1 1 1 3 14 0 0 2 0 0 ...
## $ users_50       : int 1 1 1 1 9 0 0 0 0 0 ...
## $ users_100      : int 0 1 1 0 6 0 0 0 0 0 ...
## $ bots           : int 5 4 5 5 3 2 5 6 2 2 ...
## $ birthDate      : Factor w/ 267762 levels "00:00 10 mar 2013", ...: 123272 240694 251221 141386 249826 170
## $ datetime.birthDate: Factor w/ 261494 levels "0001-07-28 20:25:00", ...: 51943 178041 10093 104783 29390 1

```

En base a la muestra, al conocimiento sobre el campo en el que estamos trabajando (wikis) y a la descripción proporcionada en el paper “*A Wikia census: motives, tools and insights*”, deducimos los siguientes atributos:

- url: url de la wiki
- creation_date: fecha de creación de la wiki en un timestamp
- domain: dominio web de la wiki
- founding_user_id: user id del fundador de la wiki
- headline: ??
- hub: Categoría de la wiki dentro de las definidas por Wikia.
- id: id de la wiki
- lang y language: idioma de la wiki
- name: Nombre propio de la wiki
- stats.activeUsers: número de usuarios activos en el último mes. Los usuarios activos son los usuarios que han hecho al menos una acción (una edición) en los últimos 30 días.
- stats.admins: número de usuarios administradores.
- stats.articles: número de artículos de la wiki.
- stats.discussions: ???
- stats.edits: número de ediciones en la wiki.
- stats.images: número de imágenes subidas.

- stats.pages: número de páginas de la wiki.
- stats.users: número de usuarios registrados en toda Wikia (potencialmente, cualquier usuario de cada wiki porque los usuarios se registran a nivel de toda Wikia).
- stats.videos: número de videos subidos.
- title: título de la wiki
- topic: Temática de la wiki definida por el administrador de la wiki.
- wam_score: Puntuación que le da Wikia a las wikis: <http://community.wikia.com/wiki/WAM/FAQ>
- stats.nonarticles: número de páginas no artículos en la wiki.
- users_{1,5,10,20,50,100}: Número de usuarios con al menos {una, cinco, diez, veinte, cincuenta, cien} edición(es).
- bots: Número de usuarios de tipo bot (no humanos)
- birthdate: fecha de creación de la wiki en formato natural
- datetime.birthDate: fecha de creación de la wiki en formato datetime de Python

Para terminar con el estudio previo de los datos, pedimos a R que nos muestre un resumen de cómo están distribuidos los valores de los atributos:

```
summary(wikis)
```

```
##                               url      creation_date
## http://0002oifos.wikia.com     : 1  2006-03-16 13:42:45: 388
## http://001-game-creator.wikia.com: 1  2004-11-11 23:33:14: 79
## http://001littlebighelp.wikia.com: 1  2006-10-28 11:16:22: 14
## http://007fanon.wikia.com      : 1  2017-10-16 21:02:01: 12
## http://007goldeneye.wikia.com   : 1  2017-10-02 21:37:50: 8
## http://007-james-bond.wikia.com : 1  2018-02-08 13:53:15: 7
## (Other)                      :277789 (Other)           :277287
##                               domain    founding_user_id
## cliff-side.wikia.com       : 2  Min.    : 0
## 0002oifos.wikia.com        : 1  1st Qu.: 4948217
## 001-game-creator.wikia.com: 1  Median  :25524286
## 001littlebighelp.wikia.com: 1  Mean    :20302830
## 007fanon.wikia.com         : 1  3rd Qu.:32701075
## 007goldeneye.wikia.com     : 1  Max.    :36902116
## (Other)                     :277788 NA's    :35
##                               headline      hub      id
## :269610 Games      :104758 Min.    : 1
## Assassin's Creed Wiki : 3 Lifestyle: 65294 1st Qu.: 659216
## The Elder Scrolls Wiki: 3 TV       : 44343 Median  :1278192
## Alice Wiki              : 2 Books     : 22619 Mean    :1137079
## Animal Crossing Wiki   : 2 Comics    : 14545 3rd Qu.:1685929
## Bakugan Wiki             : 2 Movies    : 14174 Max.    :1807291
## (Other)                  : 8173 (Other)   :12062
##                               lang      language      name
## en   :198927 en      :198927 Test Wiki      : 67
## es   : 30372 es     : 30372 Naruto Wiki    : 59
## ru   : 12967 ru     : 12967 Harry Potter Wiki: 54
## de   :  9176 de     :  9176 Pokemon Wiki  : 48
## pl   :  7582 pl     :  7582 Minecraft Wiki: 44
## fr   :  7029 fr     :  7029 Final Frontier Wiki: 39
## (Other): 11742 (Other): 11742 (Other)       :277484
## stats.activeUsers  stats.admins  stats.articles
## Min.    : -1.000  Min.    : 0.000  Min.    : 0.0
## 1st Qu.:  0.000  1st Qu.:  1.000  1st Qu.:  4.0
## Median :  0.000  Median :  1.000  Median : 10.0
```

```

##  Mean   : 0.823  Mean   : 1.481  Mean   : 122.1
##  3rd Qu.: 0.000  3rd Qu.: 1.000  3rd Qu.: 30.0
##  Max.   :7311.000  Max.   :248.000  Max.   :2483023.0
##
##    stats.discussions   stats.edits       stats.images
##  Min.   : 0.00  Min.   :     0  Min.   :-15.0
##  1st Qu.: 0.00  1st Qu.: 116  1st Qu.: 1.0
##  Median : 0.00  Median : 245  Median : 10.0
##  Mean   : 2.97  Mean   : 2755  Mean   : 210.7
##  3rd Qu.: 1.00  3rd Qu.: 470  3rd Qu.: 40.0
##  Max.   :77051.00  Max.   :23302062  Max.   :784356.0
##  NA's   :52295
##    stats.pages      stats.users      stats.videos
##  Min.   :-1   Min.   :-1   Min.   : 0.00
##  1st Qu.: 101  1st Qu.:16140740  1st Qu.: 0.00
##  Median : 172  Median :16374841  Median : 0.00
##  Mean   : 1046  Mean   :16229539  Mean   : 7.45
##  3rd Qu.: 285  3rd Qu.:16555972  3rd Qu.: 0.00
##  Max.   :7727191  Max.   :23922667  Max.   :51799.00
##
##          title           topic        wam_score
##  Test Wiki       : 67  Video Games :55625  Min.   : 0.000
##  Naruto Wiki     : 59  Gaming       :36285  1st Qu.: 0.000
##  Harry Potter Wiki : 54  Entertainment:29969  Median : 0.000
##  Pokemon Wiki    : 48  Creative      :25359  Mean   : 1.064
##  Minecraft Wiki  : 44  Fanon        :23055  3rd Qu.: 0.000
##  Final Frontier Wiki: 39  TV          :14075  Max.   :99.834
##  (Other)          :277484  (Other)      :93427
##    stats.nonarticles   users_1           users_5
##  Min.   :-1   Min.   : 0.00  Min.   : 0.0
##  1st Qu.: 95  1st Qu.: 4.00  1st Qu.: 1.0
##  Median : 150  Median : 5.00  Median : 2.0
##  Mean   : 924  Mean   : 27.48  Mean   : 11.5
##  3rd Qu.: 256  3rd Qu.: 9.00  3rd Qu.: 4.0
##  Max.   :7725359  Max.   :127087.00  Max.   :32279.0
##
##          users_10         users_20        users_50
##  Min.   : 0.00  Min.   : 0.000  Min.   : 0.000
##  1st Qu.: 1.00  1st Qu.: 0.000  1st Qu.: 0.000
##  Median : 1.00  Median : 1.000  Median : 1.000
##  Mean   : 7.67  Mean   : 5.044  Mean   : 2.813
##  3rd Qu.: 3.00  3rd Qu.: 2.000  3rd Qu.: 1.000
##  Max.   :18682.00  Max.   :15113.000  Max.   :11067.000
##
##          users_100        bots            birthDate
##  Min.   : 0.000  Min.   : 0.000  20:00, December 11, 2013 : 379
##  1st Qu.: 0.000  1st Qu.: 2.000  09:37, September 25, 2006: 348
##  Median : 0.000  Median : 4.000  21:24, February 22, 2018 : 224
##  Mean   : 1.756  Mean   : 4.135  15:07, October 4, 2016  : 168
##  3rd Qu.: 1.000  3rd Qu.: 5.000  21:27 22 feb 2018       : 73
##  Max.   :8521.000  Max.   :34.000  22:19 29 mar 2017       : 69
##                                         (Other)                  :276534
##
##          datetime.birthDate
##  2013-12-11 20:00:00: 380

```

```

## 2006-09-25 09:37:00: 348
## 2018-02-22 21:24:00: 224
## 2016-10-04 15:07:00: 168
## 2018-02-22 21:27:00: 73
## 2017-03-29 22:19:00: 69
## (Other)           :276533

```

Los campos url y domain son identificadores de la wiki y deberían ser únicos. Aunque tenemos un repetido en el dominio cliff-side.wikia.com que trataremos más adelante.

Ahora podemos deducir que **headline** se refiere a una especie de subtítulo de la wiki. En cualquier caso, se trata de un campo de texto informativo para los usuarios de la wiki, pero que a nosotros no nos interesa. El campo **stats.discussions** podría corresponderse con el número de páginas de discusión (o talk pages).

Tenemos muchísimas wikis sin videos y también muchas sin imágenes.

Wikia page views dataset

Ahora procedemos a cargar los datos de visitas:

```

# Cargamos el juego de datos
wikis_pgv<-read.csv("data/20181113_wikia-page-views.csv", header=T, sep=",")
head(wikis_pgv)

```

	url	visited_pages
## 1	http://spellmagotm.wikia.com	0
## 2	http://2017-monster-energy-nascar-cup-series.wikia.com	0
## 3	http://10low46japreligion.wikia.com	0
## 4	http://de.bibel.wikia.com	0
## 5	http://indigo-showdown.wikia.com	0
## 6	http://animewiki2.wikia.com	41

	total_views
## 1	0
## 2	0
## 3	0
## 4	0
## 5	0
## 6	560

Mostramos información descriptiva de estos datos:

```
str(wikis_pgv)
```

```

## 'data.frame': 278889 obs. of 3 variables:
## $ url      : Factor w/ 278888 levels "http://0002oifos.wikia.com",...: 225669 511 138 40364 122441 9030 ...
## $ visited_pages: int 0 0 0 0 0 41 0 0 5 0 ...
## $ total_views : int 0 0 0 0 0 560 0 0 47 0 ...
summary(wikis_pgv)

```

	url	visited_pages
## http://pl.6bp-6-batalion-pancerny.wikia.com:	2	Min. : 0.00
## http://0002oifos.wikia.com	: 1	1st Qu.: 0.00
## http://001-game-creator.wikia.com	: 1	Median : 0.00
## http://001littlebighelp.wikia.com	: 1	Mean : 12.36
## http://007fanon.wikia.com	: 1	3rd Qu.: 2.00
## http://007goldeneye.wikia.com	: 1	Max. :1000.00

```

##  (Other) :278882
##   total_views
##   Min. : 0
##   1st Qu.: 0
##   Median : 0
##   Mean   : 2892
##   3rd Qu.: 15
##   Max.  :20386167
##

```

Observamos que hay una wiki duplicada: <http://pl.6bp-6-batalion-pancerny.wikia.com/>. También observamos que la mayoría de las wikis (más de la mitad) no han tenido ni una sola visita a sus páginas en las últimas cuatro semanas. Es decir, podríamos considerar que estas wikis están muertas, puesto que ni siquiera usuarios de internet externos a la comunidad las visitan. En el otro extremo tenemos también otras wikis muy populares y así vemos como las medias que obtenemos de tanto páginas visitas como de visitas son mucho mayores que cero. Pero estas wikis que acumulan muchas visitas son escasas y no aparecen hasta más tarde del tercer cuartil de wikis.

Mostramos top 10 wikis con mayor número de visitas:

```
head(wikis_pgv[with(wikis_pgv, order(desc(wikis_pgv$total_views))), ], n = 10)
```

	url	visited_pages	total_views
## 65555	http://oldschoolrunescape.wikia.com	999	20386167
## 156168	http://fallout.wikia.com	996	11916618
## 74539	http://dnd5e.wikia.com	676	11416101
## 28452	http://warframe.wikia.com	998	11165952
## 5484	http://elderscrolls.wikia.com	999	11160352
## 253171	http://naruto.wikia.com	998	9236970
## 246128	http://bokunoheroacademia.wikia.com	996	8660570
## 202431	http://reddead.wikia.com	978	7722284
## 228607	http://onepiece.wikia.com	993	7543366
## 273208	http://yugioh.wikia.com	996	7038509

Y top 10 wikis con mayor número de páginas visitadas:

```
head(wikis_pgv[with(wikis_pgv, order(desc(wikis_pgv$visited_pages))), ], n = 10)
```

	url	visited_pages	total_views
## 24049	http://skylanders.wikia.com	1000	384642
## 24429	http://pokemon-uranium.wikia.com	1000	927726
## 32078	http://unsolvedmysteries.wikia.com	1000	265778
## 58515	http://fable.wikia.com	1000	283271
## 61960	http://terraria.wikia.com	1000	607245
## 73099	http://yokaiwatch.wikia.com	1000	858819
## 75246	http://gamelore.wikia.com	1000	11782
## 96655	http://pl.naruto.wikia.com	1000	336321
## 97769	http://saintsrow.wikia.com	1000	250143
## 104590	http://ru.grimdark.wikia.com	1000	181918

Limpieza

Vamos a hacer limpieza de los datos que muestran valores raros o que no deberían estar: Empezamos por tratar que el número de imágenes sea negativo:

Fijaremos a 0 cuando stats.images sea inferior a 0.

```
wikis$stats.images[wikis$stats.images < 0] = 0
```

Fijaremos a 0 cuando stas.activeUsers sea inferior a 0. (Significa que no tenemos usuarios activos en esa wiki, pero no tiene sentido que tengamos valores menores que 0):

```
wikis$stats.activeUsers[wikis$stats.activeUsers < 0] = 0
```

Eliminamos wikis con stats.users o stats.nonarticles o stats.pages menores que cero, puesto que más bien representan que la wiki no tiene datos válidos (una wiki normal al menos debe tener un usuario registrado o una página):

```
invalid_wikis = wikis$stats.users < 0 | wikis$stats.pages < 0 | wikis$stats.nonarticles < 0  
dim(wikis[invalid_wikis,]) # number of invalid wikis to delete
```

```
## [1] 3 32
```

```
wikis = wikis[-invalid_wikis,]  
dim(wikis)
```

```
## [1] 277794      32
```

Después, vemos qué pasa con los duplicados por dominio:

```
wikis[duplicated(wikis$domain),]
```

```
##                               url      creation_date          domain  
## 128184 http://cliffside.wikia.com 2018-08-18 06:39:50 cliff-side.wikia.com  
##         founding_user_id headline hub      id lang language      name  
## 128184           1824669          TV 1788785   en      en CliffSide Wiki  
##         stats.activeUsers stats.admins stats.articles stats.discussions  
## 128184             1            1        56            0  
##         stats.edits stats.images stats.pages stats.users stats.videos  
## 128184           859           173       393  15621017            10  
##             title      topic wam_score stats.nonarticles users_1  
## 128184 CliffSide Wiki Entertainment            0            337            5  
##         users_5 users_10 users_20 users_50 users_100 bots  
## 128184           1            1            1            1            1    2  
##             birthDate  datetime.birthDate  
## 128184 06:40, August 18, 2018 2018-08-18 06:40:00  
  
#wikis[domain == "cliff-side.wikia.com", ] # Solo hay este dominio duplicado  
# el dominio cliff-side.wikia.com está repetido. Eliminamos el último:  
wikis = wikis[-duplicated(wikis$domain),]  
dim(wikis)
```

```
## [1] 277793      32
```

Eliminamos también el duplicado que hemos visto para los datos de las visitas:

```
wikis_pgv = wikis_pgv[!duplicated(wikis_pgv$url),]  
summary(wikis_pgv)
```

```
##                               url      visited_pages  
## http://0002oifos.wikia.com : 1  Min.   : 0.00  
## http://001-game-creator.wikia.com: 1  1st Qu.: 0.00  
## http://001littlebighelp.wikia.com: 1  Median : 0.00  
## http://007fanon.wikia.com     : 1  Mean    : 12.36  
## http://007goldeneye.wikia.com   : 1  3rd Qu.:  2.00  
## http://007-james-bond.wikia.com : 1  Max.    :1000.00
```

```

##  (Other) :278882
##   total_views
##   Min. : 0
##   1st Qu.: 0
##   Median : 0
##   Mean   : 2892
##   3rd Qu.: 15
##   Max.  :20386167
##

```

Transformación

En lugar de tener la fecha de creación del censo con formato fecha, que es un formato poco comparable y clasificable en intervalos, vamos a convertirlo a una nueva variable `age` que será la edad de la wiki en días:

```

wikis$datetime.birthDate = as.POSIXct(wikis$datetime.birthDate)
#str(wikis$datetime.birthDate)
wikis$age = as.integer(Sys.time() - wikis$datetime.birthDate)
summary(wikis$age)

```

```

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      101     326    1181    1366    2190  736846

```

El valor máximo de 736808 corresponde al año 1 d.C., lo cual es imposible. Miramos cuántos valores de estos anómalos hay para `age`:

```

wikis[wikis$age > 365 * 22, c('birthdate', 'datetime.birthDate', 'age')] # wikis con más de 22 años

##                 birthDate datetime.birthDate     age
## 201614 20:25, July 28, 0001  1-07-28 20:25:00 736846

```

Observamos que hay un error en la wiki con url: <http://jrmime.wikia.com>. La fecha en `birthDate` es incorrecta (año 1), mientras que la fecha de creación en realidad es 2013-03-11.

Lo arreglamos:

```

aux = wikis[wikis$age > 365 * 22,]
aux$age = as.integer(Sys.time() - as.POSIXct(aux$creation_date))
wikis[wikis$age > 365 * 22,] = aux
# Comprobamos de nuevo si hay algún otro caso raro:
wikis[wikis$age > 365 * 22,] # wikis con más de 22 años

```

```

## [1] url           creation_date   domain
## [4] founding_user_id headline       hub
## [7] id             lang          language
## [10] name          stats.activeUsers stats.admins
## [13] stats.articles stats.discussions stats.edits
## [16] stats.images  stats.pages     stats.users
## [19] stats.videos  title          topic
## [22] wam_score     stats.nonarticles users_1
## [25] users_5        users_10      users_20
## [28] users_50       users_100     bots
## [31] birthDate      datetime.birthDate age
## <0 rows> (or 0-length row.names)

```

```

summary(wikis$age)

##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##      101     326   1181    1363    2190    7512

```

Integración

Ahora vamos a unir los dos datasets de los cuales disponemos: Wikia Census y Wikia page views. Es lo podemos hacer añadiendo los datos de visitas al dataframe `wikis` que ya tenemos. Para ello debemos juntar ambos dataframes usando como columna identificadora común la columna `url`.

```

wikis_all = merge(wikis, wikis_pgv, by="url")
dim(wikis_all)

## [1] 266408     35
summary(wikis_all)

##                                     url                      creation_date
## http://0002oifos.wikia.com      : 1  2006-03-16 13:42:45: 370
## http://001-game-creator.wikia.com: 1  2004-11-11 23:33:14: 77
## http://001littlebighelp.wikia.com: 1  2017-10-16 21:02:01: 12
## http://007fanon.wikia.com       : 1  2006-10-28 11:16:22: 11
## http://007goldeneye.wikia.com   : 1  2017-10-02 21:37:50: 8
## http://007-james-bond.wikia.com : 1  2018-02-08 13:53:15: 7
## (Other)                         :266402 (Other)                  :265923
##                                     domain          founding_user_id
## cliff-side.wikia.com      : 2  Min.   : 0
## 0002oifos.wikia.com       : 1  1st Qu.: 4893531
## 001-game-creator.wikia.com: 1  Median :25311196
## 001littlebighelp.wikia.com: 1  Mean   :19946785
## 007fanon.wikia.com        : 1  3rd Qu.:32232648
## 007goldeneye.wikia.com   : 1  Max.   :36902116
## (Other)                     :266401 NA's   :35
##                                     headline          hub
##                               :258992 Games   :100378
## Alice Wiki                   : 2 Lifestyle: 62942
## Assassin's Creed Wiki       : 2 TV       : 42839
## Bakugan Wiki                 : 2 Books    : 21363
## Bienvenidos a PokéMewtwo y PokéFantasy.: 2 Comics   : 13757
## BioShock Wiki                : 2 Movies   : 13436
## (Other)                      : 7406 (Other) : 11693
##                                     id          lang      language
## Min.   : 1 en      :191596 en      :191596
## 1st Qu.: 643464 es     : 28618 es     : 28618
## Median :1240829 ru     :12298 ru     :12298
## Mean   :1121512 de     : 8741 de     : 8741
## 3rd Qu.:1668302 pl     : 7318 pl     : 7318
## Max.   :1807291 fr     : 6558 fr     : 6558
## (Other): 11279 (Other):11279 (Other):11279
##                                     name      stats.activeUsers stats.admins
## Test Wiki      : 60  Min.   : 0.000  Min.   : 0.000
## Naruto Wiki   : 53  1st Qu.: 0.000  1st Qu.: 1.000
## Harry Potter Wiki: 46  Median : 0.000  Median : 1.000

```

```

## Pokemon Wiki      : 46   Mean    : 0.813   Mean    : 1.483
## Minecraft Wiki  : 42   3rd Qu.: 0.000   3rd Qu.: 1.000
## Final Frontier Wiki: 39   Max.    :7311.000   Max.    :248.000
## (Other)          :266122
## stats.articles   stats.discussions   stats.edits
## Min.    : 0.0   Min.    : 0.00   Min.    :       0
## 1st Qu.: 5.0   1st Qu.: 0.00   1st Qu.: 121
## Median  : 10.0  Median  : 0.00   Median  : 250
## Mean    : 104.6 Mean   : 2.56   Mean    : 2482
## 3rd Qu.: 30.0  3rd Qu.: 1.00   3rd Qu.: 478
## Max.    :1145898.0 Max.    :77051.00  Max.    :23302062
## NA's    :51310
## stats.images     stats.pages      stats.users
## Min.    : 0.0   Min.    : -1    Min.    :      -1
## 1st Qu.: 2.0   1st Qu.: 103   1st Qu.:16129209
## Median  : 11.0  Median : 181   Median :16362464
## Mean    : 190.1 Mean   : 949   Mean   :16221110
## 3rd Qu.: 41.0  3rd Qu.: 288   3rd Qu.:16543432
## Max.    :784356.0 Max.    :7727191  Max.    :23922667
##
## stats.videos      title           topic
## Min.    : 0.000  Test Wiki      : 60   Video Games :52417
## 1st Qu.: 0.000  Naruto Wiki    : 53   Gaming      :35770
## Median  : 0.000  Harry Potter Wiki: 46   Entertainment:28205
## Mean    : 6.76   Pokemon Wiki   : 46   Creative    :25021
## 3rd Qu.: 0.000  Minecraft Wiki : 42   Fanon       :22258
## Max.    :51799.00 Final Frontier Wiki: 39   TV          :13509
## (Other)          :266122  (Other)     :89228
## wam_score        stats.nonarticles   users_1
## Min.    : 0.0000 Min.    : -1    Min.    : 0.00
## 1st Qu.: 0.0000 1st Qu.: 96    1st Qu.: 4.00
## Median  : 0.0000 Median : 156   Median : 5.00
## Mean    : 0.9481 Mean   : 844   Mean   : 25.09
## 3rd Qu.: 0.0000 3rd Qu.: 257   3rd Qu.: 9.00
## Max.    :99.8342 Max.    :7725359  Max.    :127087.00
##
## users_5          users_10         users_20
## Min.    : 0.00  Min.    : 0.000  Min.    : 0.0
## 1st Qu.: 1.00  1st Qu.: 1.000  1st Qu.: 0.0
## Median  : 2.00  Median : 1.000  Median : 1.0
## Mean    : 10.44 Mean   : 6.975  Mean   : 4.6
## 3rd Qu.: 4.00  3rd Qu.: 3.000  3rd Qu.: 2.0
## Max.    :32279.00 Max.    :18682.000  Max.    :15113.0
##
## users_50         users_100        bots
## Min.    : 0.000 Min.    : 0.000 Min.    : 0.000
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 2.000
## Median  : 1.000 Median : 0.000 Median : 4.000
## Mean    : 2.573 Mean   : 1.606 Mean   : 4.167
## 3rd Qu.: 1.000 3rd Qu.: 1.000 3rd Qu.: 5.000
## Max.    :11067.000 Max.    :8521.000 Max.    :34.000
##
## birthDate        datetime.birthDate
## 20:00, December 11, 2013 : 375   Min.    :1-07-28 20:25:00

```

```

## 09:37, September 25, 2006: 332 1st Qu.:2012-12-02 14:52:45
## 21:24, February 22, 2018 : 211 Median :2015-08-03 18:29:00
## 15:07, October 4, 2016   : 165 Mean    :2015-03-04 03:23:54
## 21:27 22 feb 2018      : 73 3rd Qu.:2017-12-27 03:25:15
## 22:19 29 mar 2017      : 67 Max.    :2018-09-17 19:34:00
## (Other)                 :265185
##       age      visited_pages      total_views
## Min.   : 101   Min.   : 0.00   Min.   :     0
## 1st Qu.: 366   1st Qu.: 0.00   1st Qu.:     0
## Median :1242   Median : 0.00   Median :     0
## Mean   :1392   Mean   : 12.47   Mean   : 2959
## 3rd Qu.:2217   3rd Qu.: 2.00   3rd Qu.:    15
## Max.   :7512   Max.   :1000.00  Max.   :20386167
##

```

Transformación

Definimos una wiki como inactiva cuando el número de usuarios activos en el último mes (atributo `stats.activeUsers`) es igual a cero. Esto nos resultará útil para el posterior paso de análisis.

Para ello, creamos una nueva columna llamada `active` que será TRUE si `stats.activeUsers > 0` o FALSE en caso contrario.

```

wikis_all$active = wikis_all$stats.activeUsers > 0
str(wikis_all$active)

```

```

## logi [1:266408] FALSE FALSE FALSE TRUE FALSE FALSE ...

```

Similarmente, creamos una columna `visited` que será TRUE si ha habido alguna visita en el último mes en la wiki, o FALSE en caso contrario.

```

wikis_all$visited = wikis_all$total_views > 0
str(wikis_all$visited)

```

```

## logi [1:266408] FALSE TRUE FALSE TRUE TRUE TRUE ...

```

Análisis

Un análisis que me resulta interesante es ver si hay correlación entre ciertas variables y la actividad/inactividad de las wikis.

Primero, vamos a ver si hay correlación lineal entre el número de usuarios activos y el número de visitas en el último mes:

```

regmodel <- lm( stats.activeUsers ~ total_views + visited_pages, data = wikis_all)
summary(regmodel)

```

```

##
## Call:
## lm(formula = stats.activeUsers ~ total_views + visited_pages,
##      data = wikis_all)
##
## Residuals:
##      Min      1Q  Median      3Q     Max 
## -969.8   -0.2   -0.2   -0.2  7273.8

```

```

## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.855e-01  3.498e-02   5.304 1.14e-07 ***
## total_views 4.712e-05  3.921e-07 120.175 < 2e-16 ***
## visited_pages 3.914e-02  5.044e-04  77.597 < 2e-16 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.79 on 266405 degrees of freedom
## Multiple R-squared:  0.1005, Adjusted R-squared:  0.1005 
## F-statistic: 1.488e+04 on 2 and 266405 DF,  p-value: < 2.2e-16

```

Los resultados muestran que ambas variables: total_views y visited_pages son variables explicativas para determinar el número de usuarios activos.

Y viceversa: ¿El número de visitas determina el número de usuarios activos?

```

regmodel <- lm( total_views ~ stats.activeUsers, data = wikis_all)
summary(regmodel)

```

```

## 
## Call:
## lm(formula = total_views ~ stats.activeUsers, data = wikis_all)
## 
## Residuals:
##      Min       1Q       Median      3Q      Max  
## -10144756    -1817     -1817     -1813   20342218
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1817.006   173.014   10.5 <2e-16 ***
## stats.activeUsers 1404.409    9.217   152.4 <2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 89220 on 266406 degrees of freedom
## Multiple R-squared:  0.08016, Adjusted R-squared:  0.08016 
## F-statistic: 2.322e+04 on 1 and 266406 DF,  p-value: < 2.2e-16

```

Efectivamente, la variable usuarios activos también determina si la wiki recibe visitas o no, por lo que podemos pensar que la wiki esté activa porque hay editores colaborando es equivalente a que tenga interés / utilidad para el resto del mundo.

A continuación, vamos a ver qué factores determinan que una wiki esté activa/inactiva. Usamos un modelo de regresión logística y seleccionamos un subconjunto de variables que nos resulten relevantes para este análisis:

```

RELEVANT_ATTRS = c("hub", "language", "stats.articles", "stats.admins", "stats.edits", "stats.pages", " ")
formula <- as.formula(paste("active ~ ", paste(RELEVANT_ATTRS, collapse = "+") ))
regmodel.1 <- glm( formula = formula, family = binomial(link = 'logit'), data = wikis_all)

```

```

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(regmodel.1)

```

```

## 
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = wikis_all)

```

```

##
## Deviance Residuals:
##      Min       1Q   Median      3Q     Max
## -8.4904  -0.7120  -0.6713  -0.5095  5.8588
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.323e+01  2.018e+02  -0.066  0.9477
## hubComics            -1.007e-01  2.518e-02  -4.000 6.34e-05 ***
## hubGames             -2.874e-01  1.742e-02 -16.496 < 2e-16 ***
## hubLifestyle         -6.746e-01  1.908e-02 -35.365 < 2e-16 ***
## hubMovies            5.162e-02  2.485e-02   2.078  0.0377 *
## hubMusic             -5.878e-02  2.939e-02  -2.000  0.0455 *
## hubOther              -2.620e+00  1.134e-01 -23.113 < 2e-16 ***
## hubTV                -3.226e-01  1.963e-02 -16.432 < 2e-16 ***
## languageaf            1.352e+01  2.018e+02   0.067  0.9466
## languageam            1.379e+01  2.018e+02   0.068  0.9455
## languagear            1.147e+01  2.018e+02   0.057  0.9547
## languageast           -2.761e-01  5.722e+02   0.000  0.9996
## languageaz            1.306e+01  2.018e+02   0.065  0.9484
## languagebe            1.250e+01  2.018e+02   0.062  0.9506
## languagebg            1.182e+01  2.018e+02   0.059  0.9533
## languagebn            1.224e+01  2.018e+02   0.061  0.9516
## languagebs            1.221e+01  2.018e+02   0.061  0.9517
## languagebxr           1.477e+01  2.018e+02   0.073  0.9417
## languageca            1.177e+01  2.018e+02   0.058  0.9535
## languagecho            2.739e+01  5.722e+02   0.048  0.9618
## languagecs            1.173e+01  2.018e+02   0.058  0.9536
## languagecy            1.337e+01  2.018e+02   0.066  0.9472
## languagecz            -4.954e-01  5.722e+02  -0.001  0.9993
## languageda            1.151e+01  2.018e+02   0.057  0.9545
## languagede            1.172e+01  2.018e+02   0.058  0.9537
## languageel            1.131e+01  2.018e+02   0.056  0.9553
## languageen            1.215e+01  2.018e+02   0.060  0.9520
## languageeo            1.306e+01  2.018e+02   0.065  0.9484
## languagees            1.188e+01  2.018e+02   0.059  0.9530
## languageeu            6.689e-02  2.959e+02   0.000  0.9998
## languagefr            1.221e+01  2.018e+02   0.061  0.9518
## languagegd            -5.460e-01  5.722e+02  -0.001  0.9992
## languagegsw           1.326e+01  2.018e+02   0.066  0.9476
## languagegu            1.087e-01  5.722e+02   0.000  0.9998
## languagegv            2.934e-01  5.722e+02   0.001  0.9996
## languagehe            1.205e+01  2.018e+02   0.060  0.9524
## languagehi            1.282e+01  2.018e+02   0.064  0.9493
## languagehr            1.204e+01  2.018e+02   0.060  0.9524
## languagehu            1.152e+01  2.018e+02   0.057  0.9545
## languagehy            1.259e+01  2.018e+02   0.062  0.9503
## languageit            1.206e+01  2.018e+02   0.060  0.9523
## languageja            1.201e+01  2.018e+02   0.060  0.9525
## languageka            1.160e+01  2.018e+02   0.057  0.9541
## languagekl            2.706e+01  5.722e+02   0.047  0.9623
## languagekn            1.278e+01  2.018e+02   0.063  0.9495
## languagekr            -7.791e-02  5.722e+02   0.000  0.9999
## languageksh           6.389e-02  5.722e+02   0.000  0.9999

```

```

## languagelb      2.664e+01  5.722e+02  0.047  0.9629
## languagelt     1.186e+01  2.018e+02  0.059  0.9531
## languagemk     1.429e-01  3.064e+02  0.000  0.9996
## languageml     -3.370e-02  3.329e+02  0.000  0.9999
## languagemn     -5.485e-02  2.964e+02  0.000  0.9999
## languagemr     -1.223e-01  4.247e+02  0.000  0.9998
## languagems     1.258e+01  2.018e+02  0.062  0.9503
## languagemy     1.483e-01  3.681e+02  0.000  0.9997
## languagene     -3.805e-01  5.722e+02  -0.001 0.9995
## languagenl     1.191e+01  2.018e+02  0.059  0.9529
## languageom     2.664e-01  5.722e+02  0.000  0.9996
## languagepag    1.556e-01  5.722e+02  0.000  0.9998
## languagepie    9.667e-04  5.722e+02  0.000  1.0000
## languagepl     1.220e+01  2.018e+02  0.060  0.9518
## languageps     2.093e-01  3.690e+02  0.001  0.9995
## languagept     1.154e+01  2.018e+02  0.057  0.9544
## languagept-br   1.957e-01  5.722e+02  0.000  0.9997
## languagerm     1.377e+01  2.018e+02  0.068  0.9456
## languagero     1.191e+01  2.018e+02  0.059  0.9529
## languageru     1.245e+01  2.018e+02  0.062  0.9508
## languagesah    2.067e-01  5.722e+02  0.000  0.9997
## languagesco    -8.117e-02  5.722e+02  0.000  0.9999
## languagesi     3.141e-01  4.290e+02  0.001  0.9994
## languagesk     1.158e+01  2.018e+02  0.057  0.9542
## languagesl     1.130e+01  2.018e+02  0.056  0.9553
## languagesq     1.158e+01  2.018e+02  0.057  0.9542
## languagesr     1.262e+01  2.018e+02  0.063  0.9501
## languageata    1.147e+01  2.018e+02  0.057  0.9547
## languageath    1.181e+01  2.018e+02  0.059  0.9533
## languageatl    1.118e+01  2.018e+02  0.055  0.9558
## languageatlh   2.634e+01  5.722e+02  0.046  0.9633
## languageatr    1.182e+01  2.018e+02  0.059  0.9533
## languageatw    1.238e+01  2.018e+02  0.061  0.9511
## languageuk     1.224e+01  2.018e+02  0.061  0.9516
## languageur    -2.581e-01  3.350e+02  -0.001 0.9994
## languagevi     1.247e+01  2.018e+02  0.062  0.9507
## languageyi    -2.074e-01  4.212e+02  0.000  0.9996
## languagezh     1.232e+01  2.018e+02  0.061  0.9513
## stats.articles 1.077e-04  1.070e-05  10.058 < 2e-16 ***
## stats.admins   3.972e-02  3.456e-03  11.493 < 2e-16 ***
## stats.edits    1.511e-05  2.067e-06  7.312 2.64e-13 ***
## stats.pages    -5.904e-05  4.088e-06 -14.443 < 2e-16 ***
## users_1        1.324e-02  5.612e-04  23.597 < 2e-16 ***
## users_5        -2.163e-02  3.206e-03 -6.749 1.49e-11 ***
## users_10       -2.216e-03  5.323e-03 -0.416  0.6772
## users_20       -8.592e-03  5.933e-03 -1.448  0.1476
## users_50       -1.520e-02  7.532e-03 -2.018  0.0435 *
## users_100      1.056e-01  7.120e-03  14.833 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 282008  on 266407  degrees of freedom

```

```

## Residual deviance: 269750  on 266313  degrees of freedom
## AIC: 269940
##
## Number of Fisher Scoring iterations: 12

```

Los atributos que influyen a la hora de determinar que la wiki esté activa o no, son (sin ordenar por importancia): hubComics, hubGames, hubLifestyle, hubMovies, hubOther, hubTV, stats.articles, stats.admins, stats.edits, stats.pages, users_1, users_5, users_50 y users_100.

hubMusic y users_20 influyen muy poco, así que consideraremos que no son relevantes para nuestro modelo.

Observamos que el idioma en el que esté la wiki no es relevante para determinar que esté activa o no.

Ordenamos los atributos por importancia:

```

idx <- order(coef(summary(regmodel.1))[,4]) # sort out the p-values
out <- coef(summary(regmodel.1))[idx,]      # reorder coef, SE, etc. by increasing p
print(xtable(out, caption = "Atributos relevantes para que una wiki esté activa", auto = TRUE))

## % latex table generated in R 3.4.4 by xtable 1.8-3 package
## % Fri Dec 28 16:58:46 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{lrrrr}
##   \hline
##   & Estimate & Std. Error & z value & Pr(>$$|z$|$) \\
##   \hline
##   hubLifestyle & -0.6745896 & 0.0190751 & -35.3649964 & 0.0000000 \\
##   users\_1 & 0.0132433 & 0.0005612 & 23.5973697 & 0.0000000 \\
##   hubOther & -2.6201704 & 0.1133654 & -23.1126146 & 0.0000000 \\
##   hubGames & -0.2874306 & 0.0174244 & -16.4958355 & 0.0000000 \\
##   hubTV & -0.3226256 & 0.0196344 & -16.4316578 & 0.0000000 \\
##   users\_100 & 0.1056070 & 0.0071197 & 14.8331651 & 0.0000000 \\
##   stats.pages & -0.0000590 & 0.0000041 & -14.4427072 & 0.0000000 \\
##   stats.admins & 0.0397204 & 0.0034561 & 11.4926748 & 0.0000000 \\
##   stats.articles & 0.0001077 & 0.0000107 & 10.0578704 & 0.0000000 \\
##   stats.edits & 0.0000151 & 0.0000021 & 7.3118275 & 0.0000000 \\
##   users\_5 & -0.0216337 & 0.0032057 & -6.7485987 & 0.0000000 \\
##   hubComics & -0.1007007 & 0.0251762 & -3.9998431 & 0.0000634 \\
##   hubMovies & 0.0516201 & 0.0248464 & 2.0775665 & 0.0377493 \\
##   users\_50 & -0.0152032 & 0.0075322 & -2.0184286 & 0.0435466 \\
##   hubMusic & -0.0587823 & 0.0293918 & -1.9999566 & 0.0455049 \\
##   users\_20 & -0.0085916 & 0.0059331 & -1.4480687 & 0.1475978 \\
##   users\_10 & -0.0022159 & 0.0053233 & -0.4162741 & 0.6772095 \\
##   languagebxr & 14.7673944 & 201.7725499 & 0.0731883 & 0.9416563 \\
##   languageam & 13.7872593 & 201.7712758 & 0.0683311 & 0.9455220 \\
##   languagerm & 13.7680205 & 201.7712836 & 0.0682358 & 0.9455979 \\
##   languageaf & 13.5222401 & 201.7665372 & 0.0670192 & 0.9465664 \\
##   languagecy & 13.3705069 & 201.7675603 & 0.0662669 & 0.9471654 \\
##   languagegsw & 13.2619401 & 201.7713213 & 0.0657276 & 0.9475947 \\
##   (Intercept) & -13.2307210 & 201.7663178 & -0.0655745 & 0.9477166 \\
##   languageeo & 13.0594319 & 201.7666821 & 0.0647254 & 0.9483926 \\
##   languageaz & 13.0582166 & 201.7676622 & 0.0647191 & 0.9483977 \\
##   languagehi & 12.8209651 & 201.7670817 & 0.0635434 & 0.9493338 \\
##   languagekn & 12.7807586 & 201.7700876 & 0.0633432 & 0.9494932 \\
##   languagesr & 12.6197159 & 201.7665292 & 0.0625461 & 0.9501279 \\
##   languagehy & 12.5885349 & 201.7680668 & 0.0623911 & 0.9502514 \\

```

```

## languagems & 12.5818490 & 201.7665779 & 0.0623584 & 0.9502774 \\
## languagebe & 12.4953896 & 201.7670272 & 0.0619298 & 0.9506187 \\
## languagevi & 12.4683050 & 201.7663343 & 0.0617958 & 0.9507255 \\
## languageru & 12.4466037 & 201.7663181 & 0.0616882 & 0.9508111 \\
## languagetw & 12.3805334 & 201.7694363 & 0.0613598 & 0.9510727 \\
## languagezh & 12.3163688 & 201.7663389 & 0.0610427 & 0.9513252 \\
## languagebn & 12.2419812 & 201.7671549 & 0.0606738 & 0.9516190 \\
## languageuk & 12.2410700 & 201.7663675 & 0.0606695 & 0.9516224 \\
## languagebs & 12.2092644 & 201.7679203 & 0.0605114 & 0.9517483 \\
## languagefr & 12.2069324 & 201.7663193 & 0.0605003 & 0.9517571 \\
## languagepl & 12.2029189 & 201.7663191 & 0.0604805 & 0.9517730 \\
## languageen & 12.1527293 & 201.7663172 & 0.0602317 & 0.9519711 \\
## languageit & 12.0578902 & 201.7663232 & 0.0597617 & 0.9523455 \\
## languagehe & 12.0526038 & 201.7663918 & 0.0597354 & 0.9523663 \\
## languagehr & 12.0382448 & 201.7665549 & 0.0596642 & 0.9524231 \\
## languageja & 12.0101621 & 201.7663309 & 0.0595251 & 0.9525339 \\
## languagenl & 11.9079301 & 201.7663314 & 0.0590184 & 0.9529374 \\
## languagero & 11.9069685 & 201.7663904 & 0.0590136 & 0.9529412 \\
## languagees & 11.8807538 & 201.7663178 & 0.0588837 & 0.9530447 \\
## languagelt & 11.8635294 & 201.7666179 & 0.0587983 & 0.9531128 \\
## languagetr & 11.8178407 & 201.7663381 & 0.0585719 & 0.9532931 \\
## languagebg & 11.8150433 & 201.7665168 & 0.0585580 & 0.9533042 \\
## languageth & 11.8103829 & 201.7663655 & 0.0585349 & 0.9533225 \\
## languageca & 11.7723058 & 201.7664919 & 0.0583462 & 0.9534729 \\
## languagecs & 11.7278428 & 201.7663557 & 0.0581259 & 0.9536484 \\
## languagede & 11.7155951 & 201.7663193 & 0.0580652 & 0.9536967 \\
## languageka & 11.6008399 & 201.7668050 & 0.0574963 & 0.9541499 \\
## languagesq & 11.5782903 & 201.7692414 & 0.0573838 & 0.9542394 \\
## languagesk & 11.5766039 & 201.7666370 & 0.0573762 & 0.9542455 \\
## languagept & 11.5436753 & 201.7663590 & 0.0572131 & 0.9543755 \\
## languagehu & 11.5202277 & 201.7663600 & 0.0570969 & 0.9544680 \\
## languageda & 11.5129150 & 201.7664336 & 0.0570606 & 0.9544969 \\
## languagefa & 11.4714977 & 201.7691718 & 0.0568546 & 0.9546611 \\
## languagear & 11.4681217 & 201.7664959 & 0.0568386 & 0.9546738 \\
## languageel & 11.3146543 & 201.7664012 & 0.0560780 & 0.9552797 \\
## languagesl & 11.3004768 & 201.7677018 & 0.0560074 & 0.9553359 \\
## languagegl & 11.1775912 & 201.7667122 & 0.0553986 & 0.9558209 \\
## languagecho & 27.3924569 & 572.1667293 & 0.0478750 & 0.9618159 \\
## languagekl & 27.0621219 & 572.1667293 & 0.0472976 & 0.9622760 \\
## languagelb & 26.6395085 & 572.1667296 & 0.0465590 & 0.9628647 \\
## languagetlh & 26.3397367 & 572.1667297 & 0.0460351 & 0.9632823 \\
## languagegd & -0.5459810 & 572.1667296 & -0.0009542 & 0.9992386 \\
## languagecz & -0.4954070 & 572.1667297 & -0.0008658 & 0.9993092 \\
## languageur & -0.2581363 & 335.0286486 & -0.0007705 & 0.9993852 \\
## languagesi & 0.3140756 & 429.0014053 & 0.0007321 & 0.9994159 \\
## languageen & -0.3805423 & 572.1667295 & -0.0006651 & 0.9994693 \\
## languageps & 0.2092766 & 369.0460257 & 0.0005671 & 0.9995475 \\
## languagegv & 0.2934355 & 572.1667294 & 0.0005128 & 0.9995908 \\
## languageyi & -0.2074206 & 421.2382828 & -0.0004924 & 0.9996071 \\
## languageast & -0.2760598 & 572.1667294 & -0.0004825 & 0.9996150 \\
## languagemk & 0.1429115 & 306.3785995 & 0.0004665 & 0.9996278 \\
## languageom & 0.2663622 & 572.1667293 & 0.0004655 & 0.9996286 \\
## languagemy & 0.1483238 & 368.0521855 & 0.0004030 & 0.9996785 \\
## languagesah & 0.2066543 & 572.1667293 & 0.0003612 & 0.9997118

```

```

## languagept-br & 0.1956767 & 572.1667293 & 0.0003420 & 0.9997271 \\
## languagemr & -0.1223275 & 424.7265600 & -0.0002880 & 0.9997702 \\
## languagepag & 0.1555981 & 572.1667294 & 0.0002719 & 0.9997830 \\
## languageeu & 0.0668853 & 295.8625537 & 0.0002261 & 0.9998196 \\
## languagegu & 0.1086794 & 572.1667294 & 0.0001899 & 0.9998484 \\
## languagemn & -0.0548505 & 296.3766633 & -0.0001851 & 0.9998523 \\
## languagesco & -0.0811697 & 572.1667294 & -0.0001419 & 0.9998868 \\
## languagekr & -0.0779089 & 572.1667292 & -0.0001362 & 0.9998914 \\
## languageksh & 0.0638890 & 572.1667294 & 0.0001117 & 0.9999109 \\
## languagengl & -0.0337017 & 332.8981870 & -0.0001012 & 0.9999192 \\
## languagepie & 0.0009667 & 572.1667294 & 0.0000017 & 0.9999987 \\
## \hline
## \end{tabular}
## \caption{Atributos relevantes para que una wiki esté activa}
## \end{table}

```

Respecto a los hubs, según este modelo “todos”, aunque los hubs de Movies y Music lo son en menor nivel, son algo relevantes para determinar la actividad o no de la wiki. Esto es como no decir nada puesto que si cualquier hub es relevante y toda wiki pertenece a uno y a un solo hub, nunca podemos distinguir cuando es un hecho relevante y cuando no.

Dicho todo lo anterior, observamos que los cinco atributos más importantes son:

1. Que haya una base de usuarios registrados (users_1)
2. Que la wiki tenga pocas páginas (el estimador es negativo, por tanto la relación es inversa)
3. Que la wiki tenga usuarios asiduos y con experiencia (users_100)
4. Que la wiki tenga pocos administradores (stats.admins, negativo)
5. Que la wiki tenga páginas de contenido

Modelo de predicción para wikis visitadas

Repetimos el análisis pero ahora usando como indicador de actividad que la wiki haya tenido alguna visita en el último mes:

```

formula <- as.formula(paste("visited ~ ", paste(RELEVANT_ATTRS, collapse = "+") ))
regmodel.2 <- glm( formula = formula, family = binomial(link = 'logit'), data = wikis_all)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(regmodel.2)

## 
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = wikis_all)
## 

## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -8.4904   -0.8993   -0.7436    1.1382    4.5640
## 

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.458e+01  3.322e+02  -0.044   0.9650
## hubComics   2.782e-01  2.425e-02   11.473 < 2e-16 ***
## hubGames    4.102e-02  1.684e-02    2.436   0.0149 *
## hubLifestyle -2.021e-01  1.770e-02  -11.420 < 2e-16 ***
## hubMovies   3.329e-01  2.419e-02   13.760 < 2e-16 ***

```

## hubMusic	2.112e-01	2.801e-02	7.539	4.75e-14	***
## hubOther	-1.471e+00	5.877e-02	-25.021	< 2e-16	***
## hubTV	2.840e-01	1.866e-02	15.224	< 2e-16	***
## languageaf	1.479e+01	3.322e+02	0.045	0.9645	
## languageam	2.833e+01	7.011e+02	0.040	0.9678	
## languagear	1.451e+01	3.322e+02	0.044	0.9652	
## languageast	2.139e+01	9.432e+02	0.023	0.9819	
## languageaz	1.246e+01	3.322e+02	0.037	0.9701	
## languagebe	1.444e+01	3.322e+02	0.043	0.9653	
## languagebg	1.451e+01	3.322e+02	0.044	0.9652	
## languagebn	1.415e+01	3.322e+02	0.043	0.9660	
## languagebs	1.278e+01	3.322e+02	0.038	0.9693	
## languagebxr	1.507e+01	3.322e+02	0.045	0.9638	
## languageca	1.367e+01	3.322e+02	0.041	0.9672	
## languagecho	2.772e-01	9.432e+02	0.000	0.9998	
## languagecs	1.349e+01	3.322e+02	0.041	0.9676	
## languagecy	1.500e+01	3.322e+02	0.045	0.9640	
## languagecz	2.723e+01	9.432e+02	0.029	0.9770	
## languageda	1.452e+01	3.322e+02	0.044	0.9651	
## languagede	1.400e+01	3.322e+02	0.042	0.9664	
## languageel	1.342e+01	3.322e+02	0.040	0.9678	
## languageen	1.351e+01	3.322e+02	0.041	0.9676	
## languageeo	1.413e+01	3.322e+02	0.043	0.9661	
## languagees	1.348e+01	3.322e+02	0.041	0.9676	
## languageeu	1.341e+01	3.322e+02	0.040	0.9678	
## languagefr	1.376e+01	3.322e+02	0.041	0.9670	
## languagedg	-7.548e-01	9.432e+02	-0.001	0.9994	
## languagegs	1.380e+01	3.322e+02	0.042	0.9669	
## languagegu	-5.295e-01	9.432e+02	-0.001	0.9996	
## languagegv	-4.672e-01	9.432e+02	0.000	0.9996	
## languagehe	1.443e+01	3.322e+02	0.043	0.9654	
## languagehi	1.461e+01	3.322e+02	0.044	0.9649	
## languagehr	1.428e+01	3.322e+02	0.043	0.9657	
## languagehu	1.387e+01	3.322e+02	0.042	0.9667	
## languagehy	1.449e+01	3.322e+02	0.044	0.9652	
## languageit	1.396e+01	3.322e+02	0.042	0.9665	
## languageja	1.417e+01	3.322e+02	0.043	0.9660	
## languageka	1.394e+01	3.322e+02	0.042	0.9665	
## languagekl	-3.229e-01	9.432e+02	0.000	0.9997	
## languagekn	1.348e+01	3.322e+02	0.041	0.9676	
## languagekr	-9.619e-02	9.432e+02	0.000	0.9999	
## languageksh	2.737e+01	9.432e+02	0.029	0.9769	
## languagelb	-8.884e-01	9.432e+02	-0.001	0.9992	
## languagelt	1.408e+01	3.322e+02	0.042	0.9662	
## languagemk	1.475e+01	3.322e+02	0.044	0.9646	
## languageml	1.479e+01	3.322e+02	0.045	0.9645	
## languagemn	1.168e+01	3.322e+02	0.035	0.9719	
## languagemr	-1.810e-01	7.019e+02	0.000	0.9998	
## languagems	1.412e+01	3.322e+02	0.042	0.9661	
## languagemy	-2.448e-01	6.018e+02	0.000	0.9997	
## languagene	-1.273e-01	9.432e+02	0.000	0.9999	
## languagenl	1.360e+01	3.322e+02	0.041	0.9674	
## languageom	2.922e+01	9.432e+02	0.031	0.9753	
## languagepag	3.135e-01	9.432e+02	0.000	0.9997	

```

## languagepie    2.722e+01  9.432e+02  0.029  0.9770
## languagepl    1.369e+01  3.322e+02  0.041  0.9671
## languageps    1.344e+01  3.322e+02  0.040  0.9677
## languagept    1.388e+01  3.322e+02  0.042  0.9667
## languagept-br -3.015e-01  9.432e+02  0.000  0.9997
## languagerm    1.424e+01  3.322e+02  0.043  0.9658
## languagero    1.481e+01  3.322e+02  0.045  0.9644
## languageru    1.432e+01  3.322e+02  0.043  0.9656
## languagesah   -7.279e-01  9.432e+02 -0.001  0.9994
## languagesco   -3.452e-01  9.432e+02  0.000  0.9997
## languagesi     1.465e+01  3.322e+02  0.044  0.9648
## languagesk     1.407e+01  3.322e+02  0.042  0.9662
## languagesl     1.390e+01  3.322e+02  0.042  0.9666
## languagesq     1.498e+01  3.322e+02  0.045  0.9640
## languagesr     1.400e+01  3.322e+02  0.042  0.9664
## languageta    1.269e+01  3.322e+02  0.038  0.9695
## languageth    1.410e+01  3.322e+02  0.042  0.9661
## languagetl    1.489e+01  3.322e+02  0.045  0.9643
## languagetlh   1.724e+01  9.432e+02  0.018  0.9854
## languagetr    1.403e+01  3.322e+02  0.042  0.9663
## languagetw    1.305e+01  3.322e+02  0.039  0.9687
## languageuk    1.478e+01  3.322e+02  0.044  0.9645
## languageur    2.892e+01  5.446e+02  0.053  0.9577
## languagevi    1.443e+01  3.322e+02  0.043  0.9653
## languageyi    1.291e+01  3.322e+02  0.039  0.9690
## languagezh    1.500e+01  3.322e+02  0.045  0.9640
## stats.articles 1.361e-03  9.372e-05  14.518 < 2e-16 ***
## stats.admins   -1.971e-01  4.945e-03 -39.867 < 2e-16 ***
## stats.edits    1.419e-04  1.368e-05  10.369 < 2e-16 ***
## stats.pages    -5.278e-05  3.118e-05 -1.693  0.0905 .
## users_1        4.481e-02  1.745e-03  25.675 < 2e-16 ***
## users_5        1.126e-01  5.103e-03  22.059 < 2e-16 ***
## users_10       -2.239e-02  7.311e-03 -3.062  0.0022 **
## users_20       -1.948e-02  8.072e-03 -2.413  0.0158 *
## users_50       4.530e-02  9.937e-03  4.558  5.16e-06 ***
## users_100      3.074e-01  1.114e-02  27.600 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 364973  on 266407  degrees of freedom
## Residual deviance: 306559  on 266313  degrees of freedom
## AIC: 306749
##
## Number of Fisher Scoring iterations: 13

```

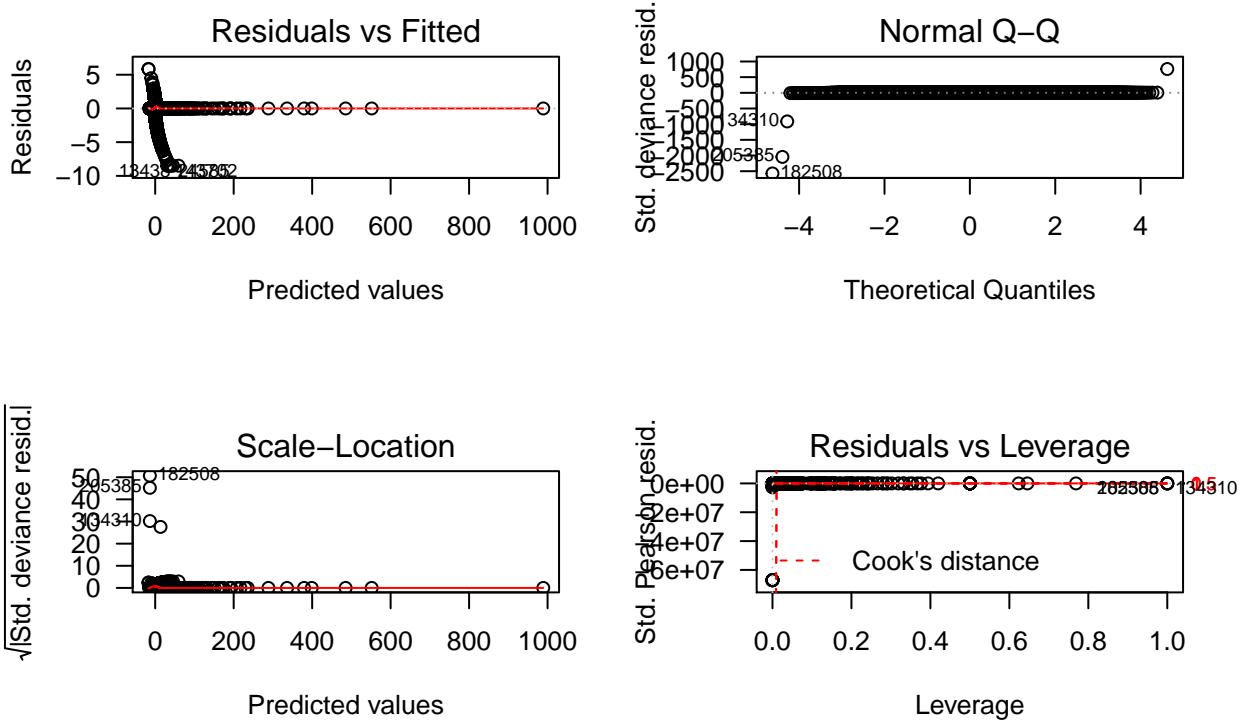
Los resultados son muy parecidos a los obtenidos anteriormente para la variable active. Las diferencias son: 1) para este modelo sí es realmente relevante que la wiki sea del hub Music y sin embargo, no es demasiado relevante que sea del hubGames; los valores de users_10 y users_20 cobran mayor importancia que en el modelo anterior, aunque siguen siendo menos importantes que los demás del modelo anterior (users_1, users_5, users_50 y users_100); la variable stats.pages no es demasiado relevante.

Visualización

Mostramos cómo son las correlaciones que hemos realizado en el paso anterior:

```
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(regmodel.1, las = 1)      # Residuals, Fitted, ...
```

```
## Warning: not plotting observations with leverage one:
## 28577, 36302, 106551, 111879, 133385, 134370, 135962, 156352, 161435, 163426, 166053, 182140, 207148, 21
## Warning: not plotting observations with leverage one:
## 28577, 36302, 106551, 111879, 133385, 134370, 135962, 156352, 161435, 163426, 166053, 182140, 207148, 21
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
glm(formula)
```



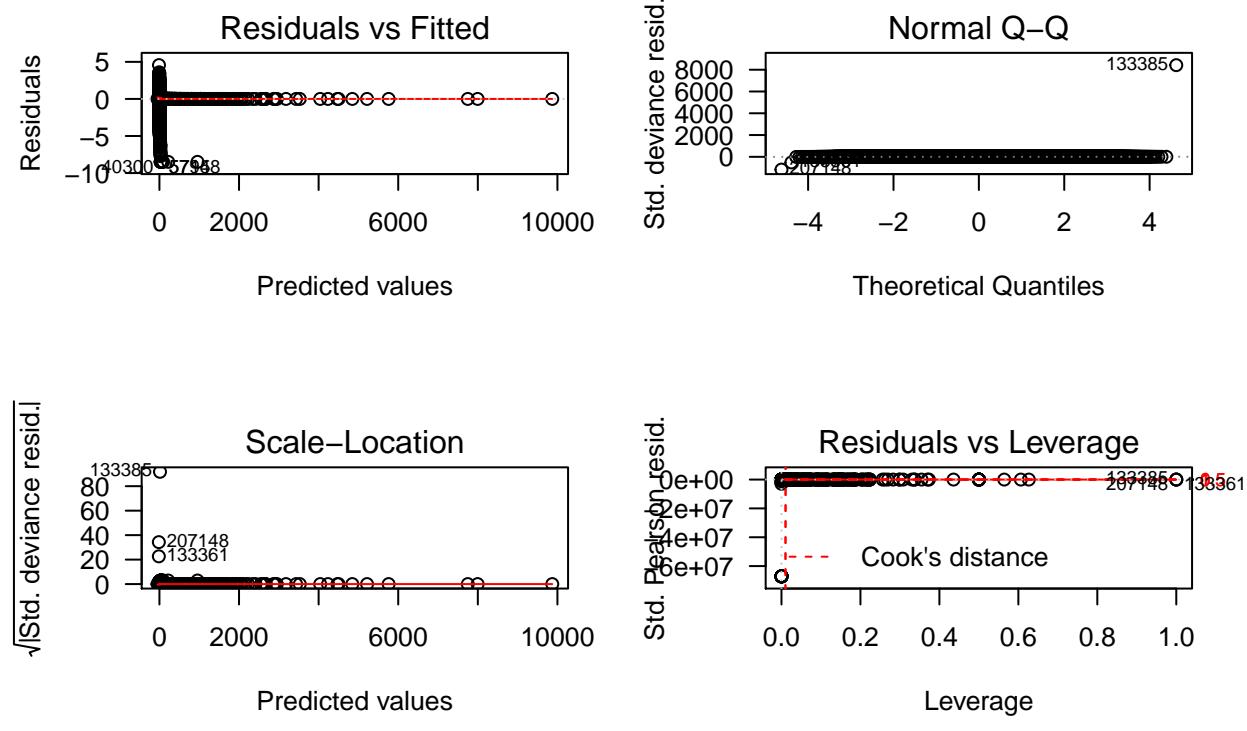
```
par(opar)

opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(regmodel.2, las = 1)      # Residuals, Fitted, ...

## Warning: not plotting observations with leverage one:
## 28577, 36302, 106551, 111879, 134310, 134370, 135962, 156352, 161435, 163426, 166053, 182140, 182508, 20
## Warning: not plotting observations with leverage one:
## 28577, 36302, 106551, 111879, 134310, 134370, 135962, 156352, 161435, 163426, 166053, 182140, 182508, 20
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
glm(formula)
```



```
par(opar)
```

Resultados

Finalmente, vamos a producir un fichero con los datos ya limpiados, transformados e integrados, y con los nuevos campos que hemos añadido durante todo el trabajo:

```
write.csv(wikis_all, "output_data/20181019-wikia_census_and_page_views-clean.csv")
```

Referencias

Jimenez-Diaz, Guillermo, Abel Serrano, y Javier Arroyo. 2018. «A Wikia Census: Motives, Tools and Insights». En *Proceedings of the 14th International Symposium on Open Collaboration*, 2:1-2:6. OpenSym '18. New York, NY, USA: ACM. doi:10.1145/3233391.3233526.