Google Trends and Brazilian public opinion surveys

Renato P. dos Santos

11 de maio de 2016

Abstract

People do like Science and Technology but are happy enough not to know very much about it, even at the risk of a huge price tag for the ordinary citizen. Fortunately, interest is not a propensity that a person is born with, and everyone's interest in Science can be triggered and developed. General surveys on public's interest in Science in Brazil intended measuring public interest in predefined topics previously selected by experts using "forced-choice" questions. However, this methodology is subject to a 'socially desirability bias', which may lead respondents to inform a preference for more "socially desirable" responses to certain sensitive issues. Conversely, there is evidence that the same respondents feel at ease in the privacy of their Internet searches. In this paper, we employ Google Trends as a non-survey-based methodology to understand the genuine Brazilian public interests in Science in comparison to those national surveys.

Methods

We made use of the *dplyr*, *ggplot2*, *cowplot*, *lubridate*, and *knitr* R packages.

Data was downloaded from Google Trends site. As it allows for at most five keywords each search (Google, n.d.a), we started by the first five topics from Table 1, namely 'Medicina + Saúde' (Medicine & Health), 'Meio ambiente' (Environment), 'Religião' (Religion), 'Economia' (Economy), and 'Esporte' (Sports). As Google differentiates misspellings as well as spelling variations in the search terms, these were included in the search as 'Medicina+Saúde+Saude' (Google, n.d.b), which generated the myData dataset, and repeated the search to obtain data from all the terms in Table 1.

Besides, as Google Trends results are normalized to the total number of searches done on Google over time (Google, n.d.a), we have included the 'Medicina+Saúde+Saude' (Medicine & Health) search group as the first one to act as a standard to the normalization of the remaining ones, to guarantee the comparabilty of results. Therefore, the second search included the search group terms 'Medicina + Saúde + Saude' (Medicine & Health), 'Ciência+Ciencia+Tecnologia' (Science & Technology), 'Arte+Cultura' (Art & Culture), 'Moda' (Fashion), and 'Política+Politica' (Politics), generating the myData2 dataset.

Datasets: Relative search volume data

The variables included in this dataset are:

- week: Starting and ending dates for the weeks during which the searches in Google occurred.
- Up to 5 search terms and their search volumes relative to the highest point on the chart, taken as 100 (Google, n.d.a).

Each dataset is stored in a comma-separated-value (CSV) file, and there is a total of 522 observations in it.

```
# Read datafile
myData <- read.csv("medicina+saude-meio_ambiente-religiao-economia-esporte-2006-2015.csv",
                   header = TRUE, stringsAsFactors = FALSE,
                   skip = 4, nrows = 522)
myData2 <- read.csv("medicina+saude-ciencia+tecnologia-arte+cultura-moda-politica-2006-2015.csv",
                   header = TRUE, stringsAsFactors = FALSE,
                   skip = 4, nrows = 522)
```

Spelling variations are removed from the variable names, which are then converted to the categories in the survey.

```
# Rename variables according to categories in the  survey
names(myData) <- c("week", "Medicine.Health", "Environment", "Religion", "Economy", "Sports")
names(myData2) <- c("week", "Medicine.Health", "Science.Technology", "Art.Culture", "Fashion", "Politics
```

The datasets are combined into one and the duplicated 'esporte' column is removed.

```
# Merge datasets
myData <- merge(myData, myData2, by.x = "week", by.y = "week")
rm(myData2)
myData <- mutate(myData,
                 Science.Technology = Science.Technology/Medicine.Health.y*Medicine.Health.x,
                 Art.Culture = Art.Culture/Medicine.Health.y*Medicine.Health.x,
                 Fashion = Fashion/Medicine.Health.y*Medicine.Health.x,
                 Politics = Politics/Medicine.Health.y*Medicine.Health.x)
myData$Medicine.Health.y <- NULL
myData <- rename(myData, Medicine.Health = Medicine.Health.x)
```

Only the starting date of the week is kept and it is converted to 'date' type.

```
# Transform 'week' to date
myData$week <- as.Date(substr(myData$week, 1, 10), "%Y-%m-%d")
myData$Year <- year(myData$week)
```

The weekly values are averaged for each of the relevant 2006, 2010, and 2015 years.

```
yearlyAverages <- aggregate(cbind(Medicine.Health, Environment,
                                  Religion, Economy, Sports,
                                  Science.Technology, Art.Culture,
                                  Fashion, Politics) ~ Year,
                            data = subset(myData,
                                          Year %in% c(2006, 2010, 2015)),
                            FUN = mean)

yearlyAverages <- as.data.frame(t(yearlyAverages))
yearlyAverages <- as.data.frame(cbind(row.names(yearlyAverages)
                                      , yearlyAverages))
yearlyAverages <- yearlyAverages[2:nrow(yearlyAverages),]
colnames(yearlyAverages) <- c("Search terms", "2006", "2010", "2015")
rownames(yearlyAverages) <- NULL

yearlyAverages <- arrange(yearlyAverages, desc(`2006`), desc(`2010`), desc(`2015`))

kable(yearlyAverages, digits=1,
      caption = "Yearly averaged volume searches from 2006 to 2015")
```

Table 1: Yearly averaged volume searches from 2006 to 2015

Search terms	2006	2010	2015
Medicine.Health	60.8	47.8	39.2

Search terms	2006	2010	2015
Art.Culture	52.3	32.3	23.3
Sports	20.2	78.6	67.0
Science.Technology	18.4	12.3	8.2
Fashion	15.2	27.6	15.8
Politics	13.6	7.9	5.8
Economy	13.4	7.1	5.3
Environment	9.7	6.4	3.2
Religion	6.2	3.4	2.9

One should be aware, however, that search terms are not the same as categories.

Google Trends provides the following 24 different topical categories Arts & Entertainment, Autos & Vehicles, Beauty & Fitness, Books & Literature, Business & Industrial, Computers & Electronics, Finance, Food & Drink, Games, Health, Hobbies & Leisure, Internet & Telecom, Jobs & Education, Law & Government, News, Online Communities, People & Society, Pets & Animals, Property, Reference, Science, Shopping, Sports, and Travel, which divide themselves into sub-categories, which are divided again into third-level subcategories, and so on. Naturally, this categorization tree reflects the observations made on May 2016 and may change at any moment in the future.

Google Trends provides filtering the search results by categories and comparison of the popularity of the search term over time to the popularity of a category as a whole (Google Inc., n.d.c), but, unfortunately, it does not provide aggregated search volume data for categories. To get a sense of it, Segev & Ahituv (2010) made use of the list of the 'Top searches', the terms that are most frequently searched within a chosen category, country, region, and/or time period (Google Inc., n.d.d) and, then, turned to the Open Directory Project (ODP) classification system of content, on which Google itself is based (Dmoz.org., 2014), to attain consistency and accuracy of the categorical classification. However, ODP does not seem to work well for Portuguese as it oft turns to categories in other languages. On the other hand, when only one search term is introduced in Google Trends, it suggests which, from the 24 categories above, two or three are the most probably related to it.

Following this methodology, we firstly collected the 50 most frequently overall searched terms on 2006, 2010, and 2015.

The Top 50 search terms on 2006, 2010, and 2015 dataset is stored in a unique comma-separated-value (CSV) file, and there is a total of 50 observations in it for each year.

```
# Top searches in Brazil, on 2006, 2010, and 2016
topSearches2006 <- read.csv("top searches Brasil 2006-2010-2015.csv",
                    header = FALSE, stringsAsFactors = FALSE,
                    skip = 7, nrows = 50)
names(topSearches2006) <- c("Queries", "Volume")

topSearches2010 <- read.csv("top searches Brasil 2006-2010-2015.csv",
                    header = FALSE, stringsAsFactors = FALSE,
                    skip = 59, nrows = 50)
names(topSearches2010) <- c("Queries", "Volume")

topSearches2015 <- read.csv("top searches Brasil 2006-2010-2015.csv",
                    header = FALSE, stringsAsFactors = FALSE,
                    skip = 111, nrows = 50)
names(topSearches2015) <- c("Queries", "Volume")

topSearches <- cbind(arrange(topSearches2006, desc(Volume)),
                     arrange(topSearches2010, desc(Volume)),
```

```
                    arrange(topSearches2015, desc(Volume)))

kable(head(topSearches, 15), digits=0,
      caption = "Top searches from 2006 to 2015")
```

Table 2: Top searches from 2006 to 2015

Queries	Volume	Queries	Volume	Queries	Volume
orkut	100	jogos	100	facebook	100
brasil	75	orkut	70	youtube	35
fotos	75	youtube	50	google	35
jogos	60	globo	45	hotmail	30
download	60	hotmail	35	globo	20
musicas	40	musicas	35	jogos	15
letras	35	uol	30	tradutor	15
videos	30	msn	30	videos	15
musica	30	tradutor	30	filmes	15
uol	30	google	30	uol	15
receita	30	jogo	25	frases	10
concurso	30	yahoo	25	face	10
mensagens	30	baixaki	20	gmail	10
msn	30	caixa	20	caixa	10
terra	25	terra	20	olx	10

Then, to get a sense of the most searched categories, for each term in each of the three 50 top searches lists, a search query was submitted to GT, which returned the main and most frequent classification. This allowed us to categorize each one of those 150 most searched terms from 2006 to 2015. Drilling down the categories tree, we also ascribed the lowest level subcategories to the terms. However, this classification suggested by GT was carefully controlled by the researcher to keep its coherence with other results. Very few cases were too general or vague, leading to a classification process not being straightforward. Since those cases were very rare, it is unlikely that a mistake in an intelligent guess would have adversely affected the results.

```
categorization <- matrix(data=c(c("4shared", "Computers & Electronics",
                                  "File Sharing & Hosting"),
                                c("americanas", "Shopping",
                                  "Shopping Portals & Search Engines"),
                                c("amor", "Society", "Romance"),
                                c("azul", "Business & Industrial",
                                  "Aviation"),
                                c("baixaki", "Computers & Electronics",
                                  "File Sharing & Hosting"),
                                c("bol", "Arts & Entertainment",
                                  "Web Portals"),
                                c("bradesco", "Finance", "Banking"),
                                c("brasil", "Finance", "Banking"),
                                c("caixa", "Finance", "Banking"),
                                c("caixa economica", "Finance",
                                  "Banking"),
                                c("caixa economica federal", "Finance",
                                  "Banking"),
                                c("carros", "Games",
```

```
  "Driving & Racing Games"),
c("casas bahia", "Shopping",
  "Shopping Portals & Search Engines"),
c("celular", "Internet & Telecom",
  "Mobile Phones"),
c("cep", "Reference",
  "Business & Personal Listings"),
c("claro", "Internet & Telecom",
  "Phone Service Providers"),
c("click jogos", "Games", "Online Games"),
c("concurso", "Law & Government",
  "State & Local Government"),
c("concursos", "Law & Government",
  "State & Local Government"),
c("contos", "Books & Literature",
  "Books & Literature"),
c("correios", "Business",
  "Mail & Package Delivery"),
c("detran", "Law & Government",
  "Vehicle Licensing & Registration"),
c("dicionario", "Reference",
  "Dictionaries & Encyclopedias"),
c("download", "Computers & Electronics",
  "Music Streams & Downloads"),
c("enem", "Jobs & Education",
  "Colleges & Universities"),
c("esporte", "Sports", "Sports"),
c("face", "Online Communities",
  "Social Networks"),
c("facebook", "Online Communities",
  "Social Networks"),
c("filmes", "Arts & Entertainment",
  "Online Video"),
c("filmes online", "Arts & Entertainment",
  "Online Video"),
c("fotos", "Arts & Entertainment",
  "Online Image Galleries"),
c("frases", "People & Society",
  "Romance"),
c("friv", "Games", "Online Games"),
c("futebol", "Sports", "Football"),
c("g1", "Arts & Entertainment",
  "Web Portals"),
c("games", "Games", "Games"),
c("globo", "Arts & Entertainment",
  "Web Portals"),
c("globo esporte", "News", "Sports News"),
c("gmail", "Internet & Telecom",
  "Text & Instant Messaging"),
c("google", "Internet & Telecom",
  "Search Engines"),
c("google tradutor", "Reference",
  "Dictionaries & Encyclopedias"),
```

```
                        c("hotmail", "Internet & Telecom",
                          "Text & Instant Messaging"),
                        c("ig", "Arts & Entertainment",
                          "Web Portals"),
                        c("imagens", "Arts & Entertainment",
                          "Online Image Galleries"),
                        c("instagram", "Hobbies & Leisure",
                          "Photo & Image Sharing"),
                        c("itau", "Finance", "Banking"),
                        c("jogo", "Games", "Games"),
                        c("jogos", "Games", "Games"),
                        c("jogos online", "Games", "Online Games"),
                        c("letras", "Arts & Entertainment",
                          "Song Lyrics & Tabs"),
                        c("mapa", "Reference", "Maps"),
                        c("maps", "Reference", "Maps"),
                        c("mega sena", "Hobbies & Leisure",
                          "Lottery & Sweepstakes"),
                        c("mensagem", "People & Society",
                          "Romance"),
                        c("mensagens", "People & Society",
                          "Romance"),
                        c("mercado livre", "Shopping",
                          "Shopping Portals & Search Engines"),
                        c("minecraft", "Games", "Massive Multiplayer"),
                        c("mp3", "Arts & Entertainment",
                          "Music Streams & Downloads"),
                        c("msn", "Internet & Telecom",
                          "Text & Instant Messaging"),
                        c("mulheres", "Arts & Entertainment",
                          "Online Image Galleries"),
                        c("musica", "Arts & Entertainment",
                          "Music Streams & Downloads"),
                        c("musicas", "Arts & Entertainment",
                          "Music Streams & Downloads"),
                        c("naruto", "Arts & Entertainment",
                          "Anime & Manga"),
                        c("net", "Internet & Telecom", "ISPs"),
                        c("noticias", "News", "News"),
                        c("oi", "Internet & Telecom",
                          "Phone Service Providers"),
                        c("olx", "Shopping",
                          "Shopping Portals & Search Engines"),
                        c("orkut", "Online Communities",
                          "Social Networks"),
                        c("orkut login", "Online Communities",
                          "Social Networks"),
                        c("outlook", "Internet & Telecom",
                          "Text & Instant Messaging"),
                        c("previsão do tempo", "News", "Weather"),
                        c("radio", "Arts & Entertainment",
                          "Radio"),
                        c("rbd", "Arts & Entertainment",
```

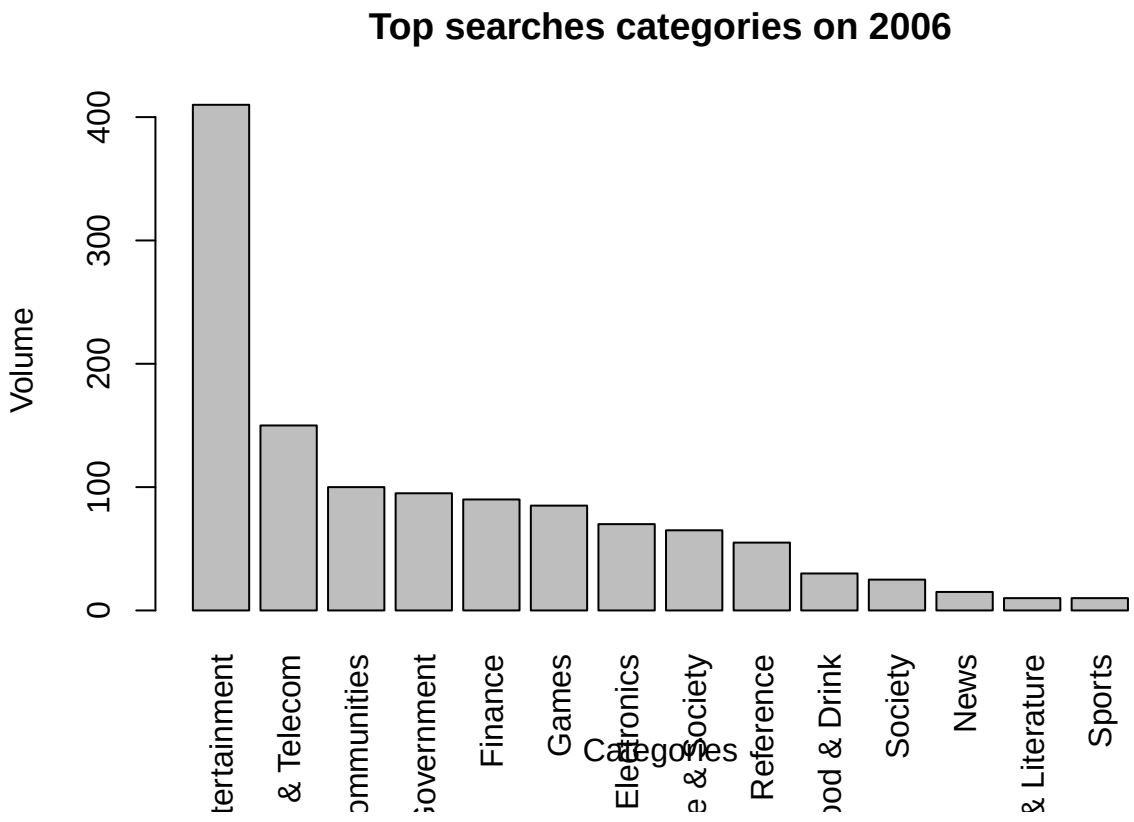
```
                                  "Latin pop"),
                                c("receita", "Food & Drink", "Cooking & Recipes"),
                                c("receita federal", "Law & Government",
                                  "Public Finance"),
                                c("tempo", "News", "Weather"),
                                c("terra", "Arts & Entertainment",
                                  "Web Portals"),
                                c("tradutor", "Reference",
                                  "Dictionaries & Encyclopedias"),
                                c("tradutor google", "Reference",
                                  "Dictionaries & Encyclopedias"),
                                c("twitter", "Internet & Telecom",
                                  "Microblogging"),
                                c("uol", "Arts & Entertainment",
                                  "Web Portals"),
                                c("vagalume", "Arts & Entertainment",
                                  "Song Lyrics & Tabs"),
                                c("videos", "Arts & Entertainment",
                                  "Online Video"),
                                c("videos", "Arts & Entertainment",
                                  "Online Video"),
                                c("vivo", "Internet & Telecom",
                                  "Phone Service Providers"),
                                c("whatsapp", "Internet & Telecom",
                                  "Text & Instant Messaging"),
                                c("yahoo", "Internet & Telecom",
                                  "Search Engines"),
                                c("you", "Arts & Entertainment",
                                  "Video Sharing"),
                                c("you tube", "Arts & Entertainment",
                                  "Video Sharing"),
                                c("youtube", "Arts & Entertainment",
                                  "Video Sharing")),
                          nrow=90, ncol=3, byrow = TRUE)

topSearches2006$category <- as.factor(categorization[match(topSearches2006$Queries, categorization[,1])
topSearches2010$category <- as.factor(categorization[match(topSearches2010$Queries, categorization[,1])
topSearches2015$category <- as.factor(categorization[match(topSearches2015$Queries, categorization[,1])

topSearches2006$subcategory <- as.factor(categorization[match(topSearches2006$Queries, categorization[,
topSearches2010$subcategory <- as.factor(categorization[match(topSearches2010$Queries, categorization[,
topSearches2015$subcategory <- as.factor(categorization[match(topSearches2015$Queries, categorization[,
```

It is, of course, impossible to be completely sure of what kind of information each individual user intended to acquire from a search term. However, it is possible to obtain a category based on the majority of search results and, therefore, to assign the category to the relevant term with a high degree of confidence (Segev & Ahituv, 2010).

Now, we can summarize by categories the results for each of the 2006, 2010, and 2015 years.

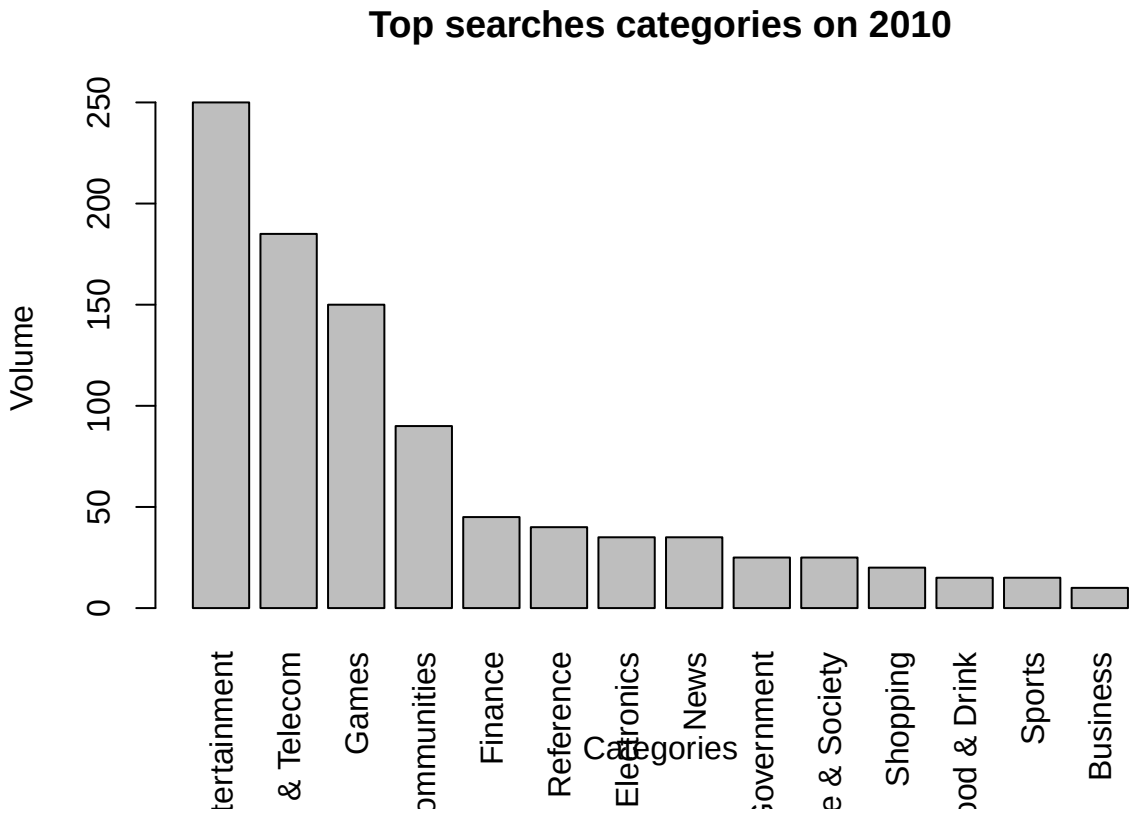
```
totals2006 <- aggregate(Volume ~ category, data = topSearches2006,
                        FUN = sum)
totals2006$category <- factor(totals2006$category ,
                              levels = totals2006[order(totals2006$Volume,
```

```
                                                    decreasing = TRUE), 1])
totals2006 <- arrange(totals2006, desc(Volume))

totals2010 <- aggregate(Volume ~ category, data = topSearches2010,
                        FUN = sum)
totals2010$category <- factor(totals2010$category ,
                          levels = totals2010[order(totals2010$Volume,
                                                    decreasing = TRUE), 1])
totals2010 <- arrange(totals2010, desc(Volume))

totals2015 <- aggregate(Volume ~ category, data = topSearches2015,
                        FUN = sum)
totals2015$category <- factor(totals2015$category ,
                          levels = totals2015[order(totals2015$Volume,
                                                    decreasing = TRUE), 1])
totals2015 <- arrange(totals2015, desc(Volume))

with(totals2006, barplot(Volume, names.arg = category,
                         las = 3,
                         xlab = "Categories",
                         ylab = "Volume",
                         main = "Top searches categories on 2006"))
```

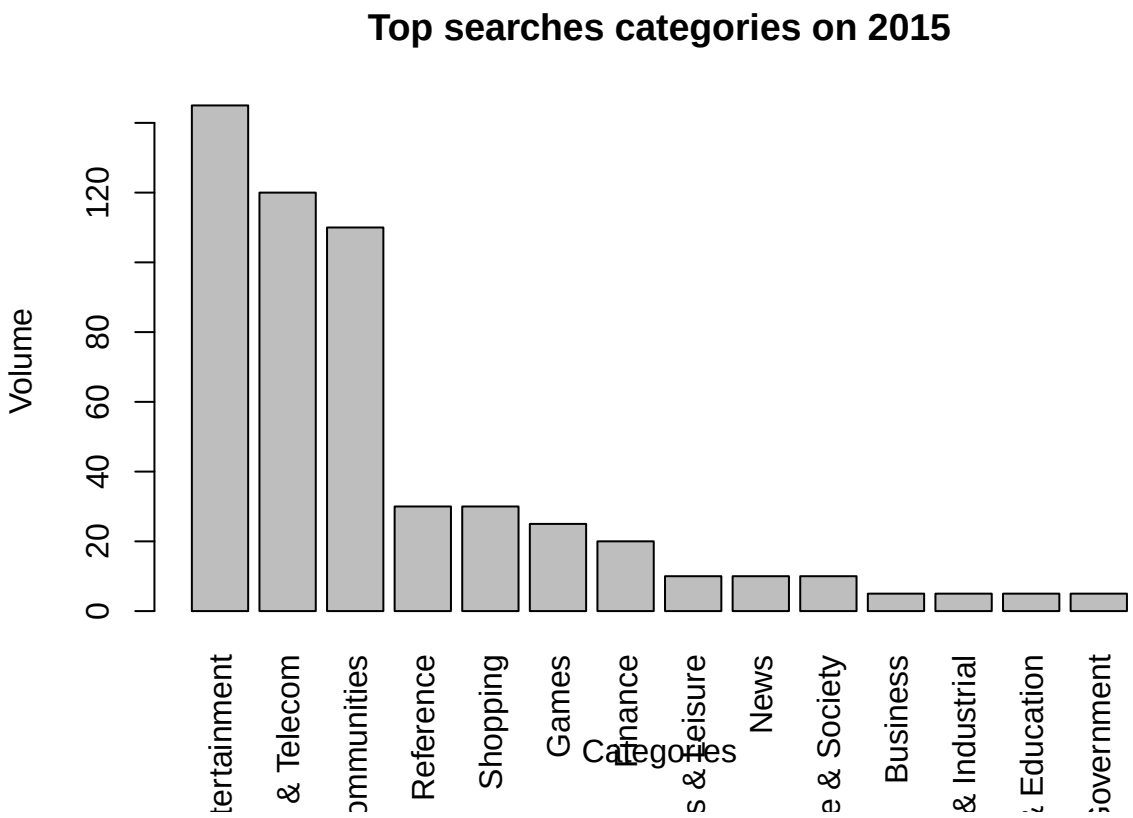



```
with(totals2010, barplot(Volume, names.arg = category,
                         las = 3,
```

```
                        xlab = "Categories",
                        ylab = "Volume",
                        main = "Top searches categories on 2010"))
```




```
with(totals2015, barplot(Volume, names.arg = category,
                        las = 3,
                        xlab = "Categories",
                        ylab = "Volume",
                        main = "Top searches categories on 2015"))
```



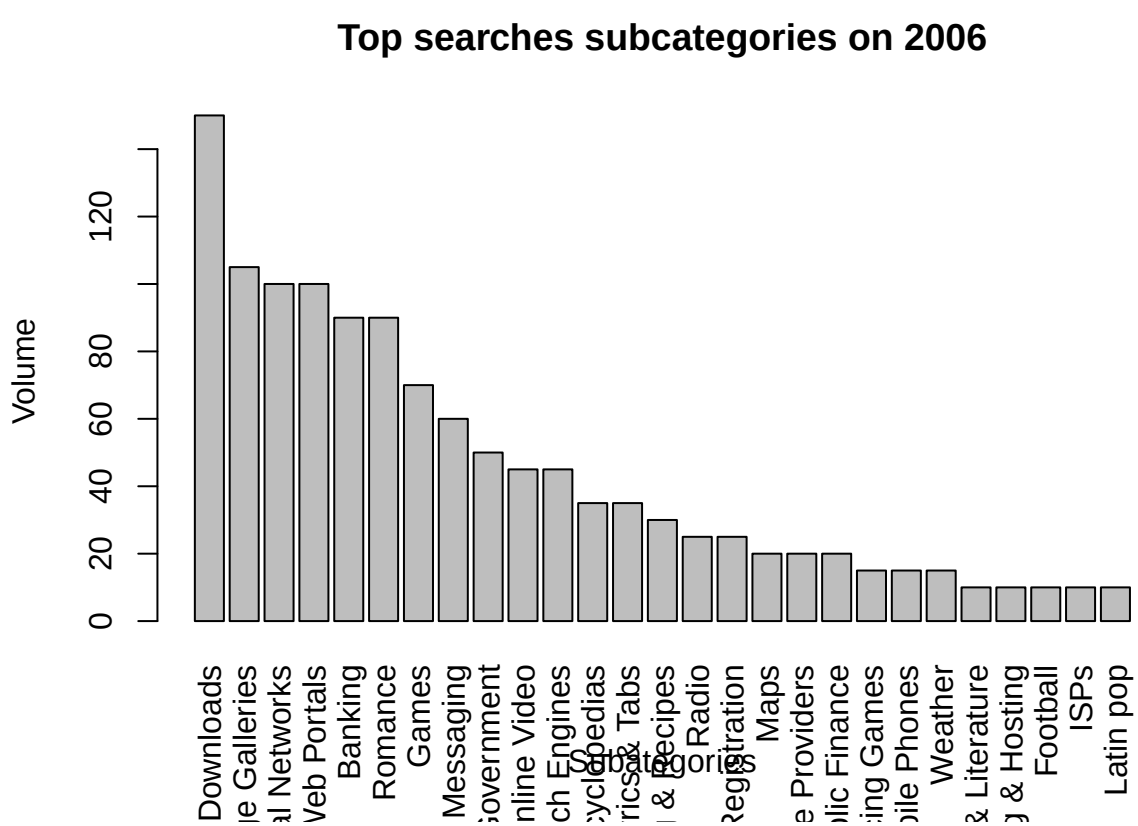

```
totals <- merge(totals2006, totals2010, by.x = "category", by.y = "category")
totals <- rename(totals, `2006` = Volume.x, `2010` = Volume.y)
totals <- merge(totals, totals2015, by.x = "category", by.y = "category")
totals <- rename(totals, `2015` = Volume)
totals <- arrange(totals, desc(`2006`), desc(`2010`), desc(`2015`))

kable(totals, digits=0,
      caption = "Top searches categories from 2006 to 2015")
```

Table 3: Top searches categories from 2006 to 2015

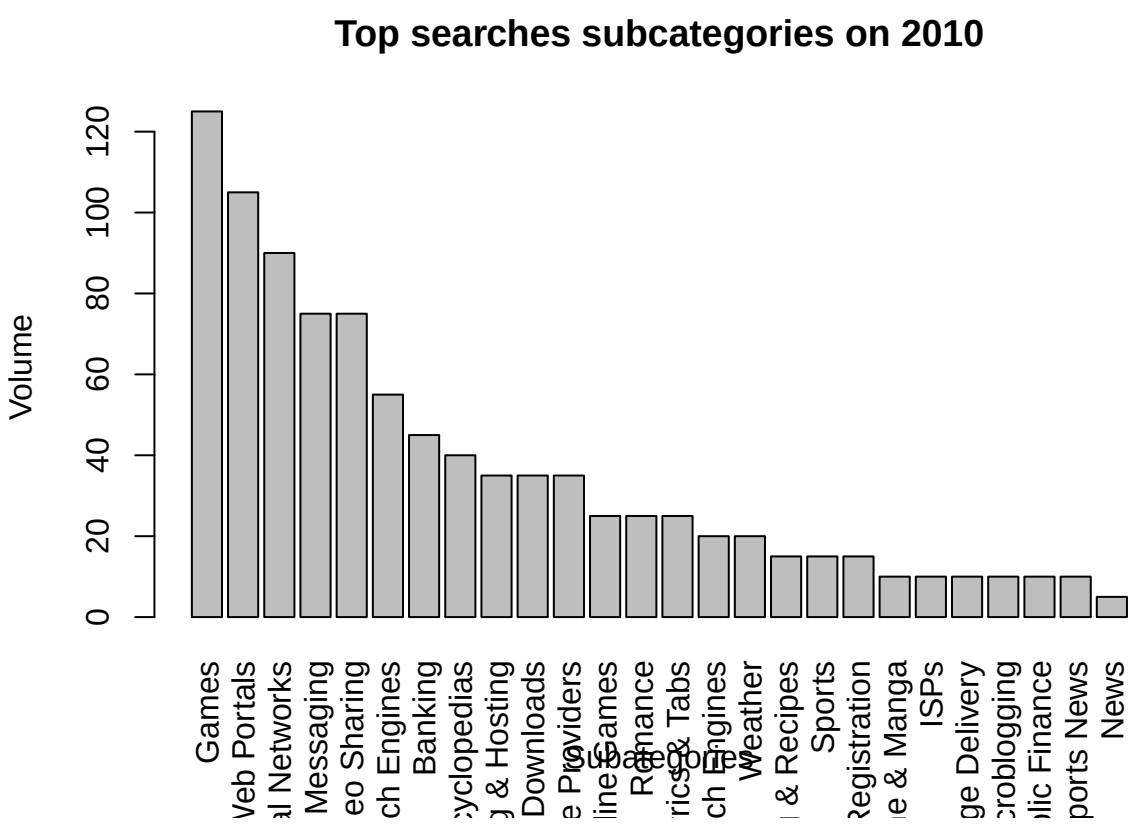
category	2006	2010	2015
Arts & Entertainment	410	250	145
Internet & Telecom	150	185	120
Online Communities	100	90	110
Law & Government	95	25	5
Finance	90	45	20
Games	85	150	25
People & Society	65	25	10
Reference	55	40	30
News	15	35	10

Now, we can summarize by subcategories the results for each of the 2006, 2010, and 2015 years.

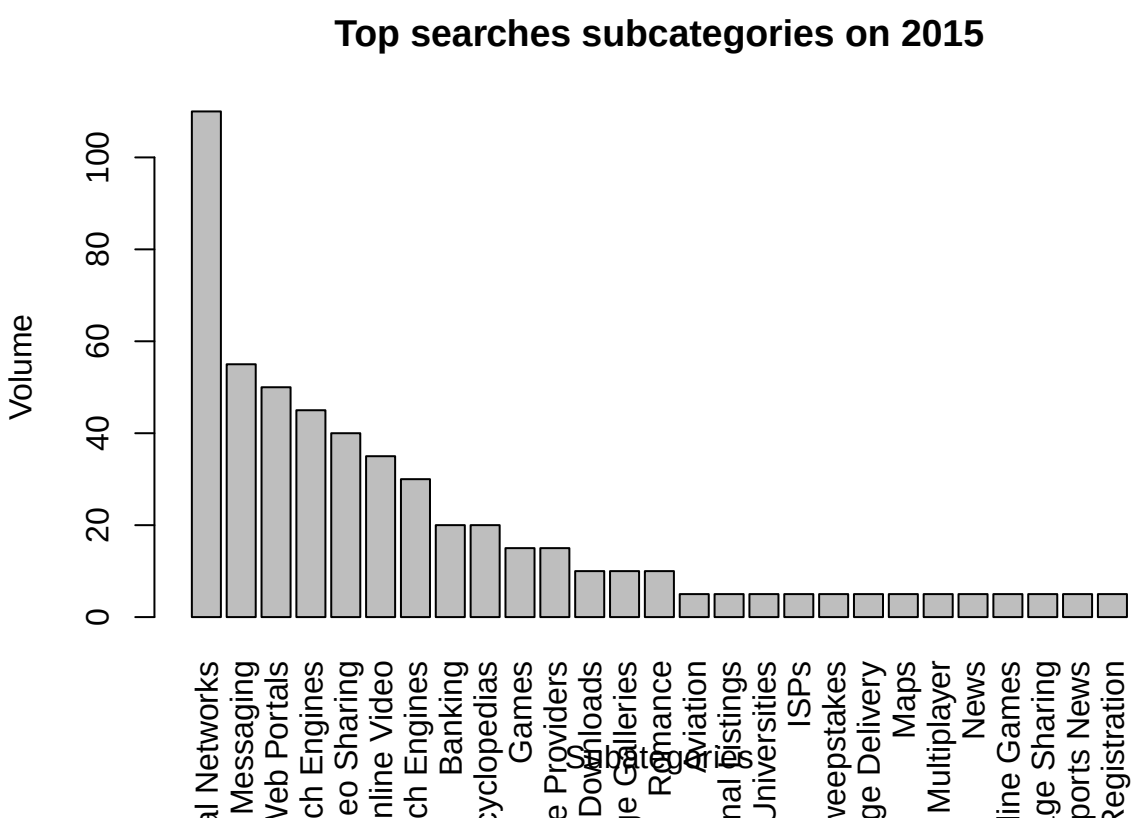
```
subtotals2006 <- aggregate(Volume ~ subcategory, data = topSearches2006,
                           FUN = sum)
subtotals2006$subcategory <- factor(subtotals2006$subcategory ,
                            levels = subtotals2006[order(subtotals2006$Volume,
                                                         decreasing = TRUE), 1])
subtotals2006 <- arrange(subtotals2006, desc(Volume))

subtotals2010 <- aggregate(Volume ~ subcategory, data = topSearches2010,
                           FUN = sum)
subtotals2010$subcategory <- factor(subtotals2010$subcategory ,
                            levels = subtotals2010[order(subtotals2010$Volume,
                                                         decreasing = TRUE), 1])
subtotals2010 <- arrange(subtotals2010, desc(Volume))

subtotals2015 <- aggregate(Volume ~ subcategory, data = topSearches2015,
                           FUN = sum)
subtotals2015$subcategory <- factor(subtotals2015$subcategory ,
                            levels = subtotals2015[order(subtotals2015$Volume,
                                                         decreasing = TRUE), 1])
subtotals2015 <- arrange(subtotals2015, desc(Volume))

with(subtotals2006, barplot(Volume, names.arg = subcategory,
                           las = 3,
                           xlab = "Subategories",
                           ylab = "Volume",
                           main = "Top searches subcategories on 2006"))
```




```
with(subtotals2010, barplot(Volume, names.arg = subcategory,
                            las = 3,
                            xlab = "Subategories",
                            ylab = "Volume",
                            main = "Top searches subcategories on 2010"))
```




```
with(subtotals2015, barplot(Volume, names.arg = subcategory,
                            las = 3,
                            xlab = "Subategories",
                            ylab = "Volume",
                            main = "Top searches subcategories on 2015"))
```




```
subtotals <- merge(subtotals2006, subtotals2010, by.x = "subcategory", by.y = "subcategory")
subtotals <- rename(subtotals, `2006` = Volume.x, `2010` = Volume.y)
subtotals <- merge(subtotals, subtotals2015, by.x = "subcategory", by.y = "subcategory")
subtotals <- rename(subtotals, `2015` = Volume)
subtotals <- arrange(subtotals, desc(`2006`), desc(`2010`), desc(`2015`))

kable(subtotals, digits=0,
      caption = "Top searches subcategories from 2006 to 2015")
```

Table 4: Top searches subcategories from 2006 to 2015

subcategory	2006	2010	2015
Music Streams & Downloads	150	35	10
Web Portals	100	105	50
Social Networks	100	90	110
Banking	90	45	20
Romance	90	25	10
Games	70	125	15
Text & Instant Messaging	60	75	55
Search Engines	45	55	45
Dictionaries & Encyclopedias	35	40	20
Vehicle Licensing & Registration	25	15	5
Phone Service Providers	20	35	15
ISPs	10	10	5

References

- Dmoz.org. (2014, March 14). About DMOZ. Retrieved May 1, 2016, from http://www.dmoz.org/about.html
- Google Inc. (n.d.a). Trends graphs and forecasts. Retrieved April 29, 2016, from https://support.google.com/trends/answer/4355164?hl=en-GB&rd=1
- Google Inc. (n.d.b). Search tips for Trends. Retrieved May 1, 2016, from https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530
- Google Inc. (n.d.c). Refine Trends results by category. Retrieved May 2, 2016, from https://support.google.com/trends/answer/4359597?hl=en&ref_topic=4365530
- Google Inc. (n.d.d). Find related searches. Retrieved May 2, 2016, from https://support.google.com/trends/answer/4355000?hl=en-GB&rd=1
- Segev, E., & Ahituv, N. (2010). Popular Searches in Google and Yahoo!: A "Digital Divide" in Information Uses? The Information Society: An International Journal, 26(1), 17-37. http://doi.org/10.1080/01972240903423477