Interpreting the topics: an example

Salfo Bikienga

April 30, 2018

(scroll down to see the plots at the bottom)

Generally, to interpret the topics generated by topic modeling, we use the list of the most important (in term of word weight) words for that particular topic.

```
library(topicmodels)
library(tm)
library(tidyverse)
options(digits = 3)
load('data/constructed_data/blogs_lda_10.RData')
K = 10
theta_matrix <- posterior(blogs_lda_10)$topics # Extract the theta matrix
theta_matrix <- round(as.data.frame(theta_matrix), digits = 3)
names(theta_matrix) <- paste0("Topic.", 1:K) # Name the columns
terms_matrix <- terms(blogs_lda_10, 15)
terms_matrix
```

```
##       Topic 1   Topic 2  Topic 3    Topic 4   Topic 5    Topic 6
##  [1,] "cost"    "peopl"  "think"    "bank"    "studi"    "human"
##  [2,] "peopl"   "think"  "probabl"  "financi" "predict"  "will"
##  [3,] "care"    "like"   "make"     "market"  "famili"   "futur"
##  [4,] "will"    "seem"   "just"     "money"   "research" "time"
##  [5,] "health"  "person" "reason"   "risk"    "test"     "like"
##  [6,] "money"   "social" "argument" "capit"   "school"   "world"
##  [7,] "insur"   "status" "like"     "asset"   "data"     "much"
##  [8,] "year"    "bias"   "differ"   "will"    "face"     "even"
##  [9,] "work"    "make"   "know"     "credit"  "local"    "seem"
## [10,] "make"    "good"   "will"     "regul"   "html"     "intellig"
## [11,] "incom"   "want"   "mean"     "rate"    "inform"   "system"
## [12,] "spend"   "also"   "question" "loan"    "main"     "year"
## [13,] "compani" "much"   "thing"    "fund"    "paper"    "power"
## [14,] "product" "group"  "point"    "price"   "result"   "chang"
## [15,] "benefit" "even"   "believ"   "invest"  "student"  "make"
##       Topic 7  Topic 8    Topic 9    Topic 10
##  [1,] "think"  "govern"   "debt"     "market"
##  [2,] "like"   "polit"    "will"     "econom"
##  [3,] "just"   "will"     "economi"  "year"
##  [4,] "peopl"  "obama"    "rate"     "technolog"
##  [5,] "know"   "state"    "countri"  "financ"
##  [6,] "will"   "american" "econom"   "post"
##  [7,] "time"   "right"    "polici"   "advertis"
##  [8,] "thing"  "power"    "growth"   "updat"
##  [9,] "make"   "http"     "euro"     "rate"
## [10,] "work"   "vote"     "currenc"  "subscrib"
## [11,] "read"   "system"   "china"    "time"
## [12,] "good"   "public"   "govern"   "five"
## [13,] "want"   "nation"   "crisi"    "price"
## [14,] "realli" "parti"    "european" "next"
```

```
## [15,] "look"   "presid"   "europ"    "growth"
```

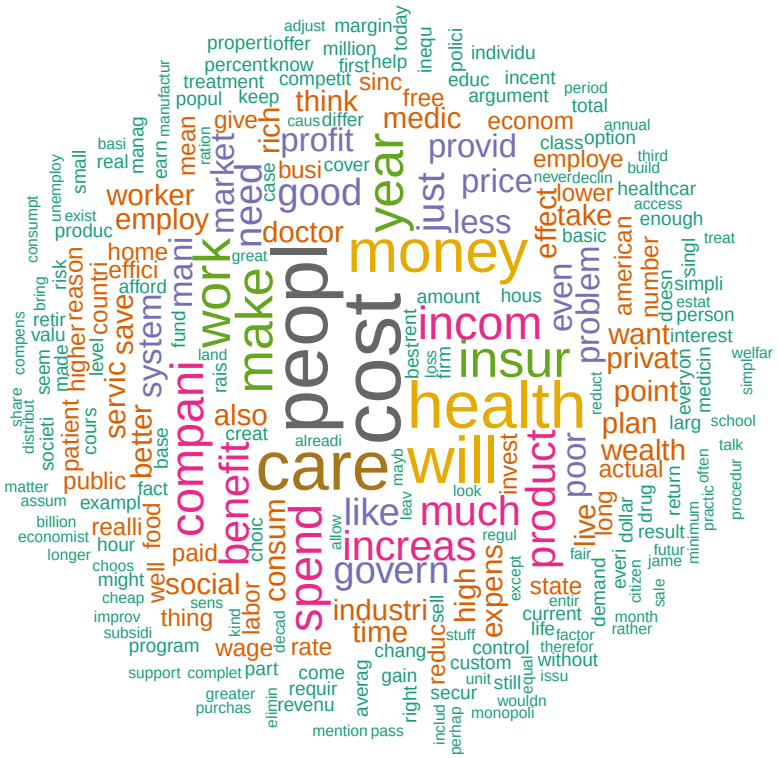
```
phi_matrix <- posterior(blogs_lda_10)$terms # Extract the phi matrix
phi_matrix <- round(phi_matrix, 3) # Round the numbers to 3 decimals
#phi_matrix[, 1:20] # Print out the first 20 words
T_phi_matrix <- as.data.frame(t(phi_matrix))
names(T_phi_matrix) <- paste0("Topic.", 1:K)
```

An alternative to the list of the most important words, we may use a word cloud. Here, we use a word cloud to interpret topic 1.

```
library("wordcloud")
```

```
## Loading required package: RColorBrewer
```

```
library("RColorBrewer")
cloud_df <- data.frame(words = row.names(T_phi_matrix),
                       freq = T_phi_matrix[, 1])
set.seed(314) # needed to reproduce the exact same wordcloud
wordcloud(words = cloud_df$words, freq = cloud_df$freq, min.freq = 1,
          max.words=250, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



The wordcloud clearly suggests that topic.1 is about health care cost. We can see that health care related words are prominent in the plot. Note the size of the words such as: health, cost, doctor, insurance, benefit etc.