Text Mining the History of Ideas

Mattias Skipper Rasmussen
filmsr@cas.au.dk

Frontiers in Digital Scholarship
Aarhus University
October 2016

Generic "toy" questions

Some generic questions that may interest intellectual historians:

- Concept prevalence: how prevalent is concept X?
- Concept association: which concepts are typically associated with X?
- Topic discovery: which topics are frequently discussed?
- Topic prevalence: is there a correlation between the prevalence of different topics?

Intellectual historians typically assess such questions qualitatively.

I will illustrate how to approach them quantitatively.

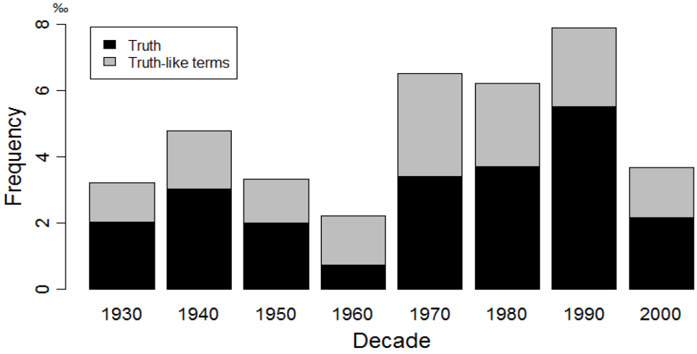
Test corpus

- Sample of 324 philosophical articles from 1930-2010:
  - The Journal of Philosophy: 118 articles
  - The Philosophical Review: 206 articles

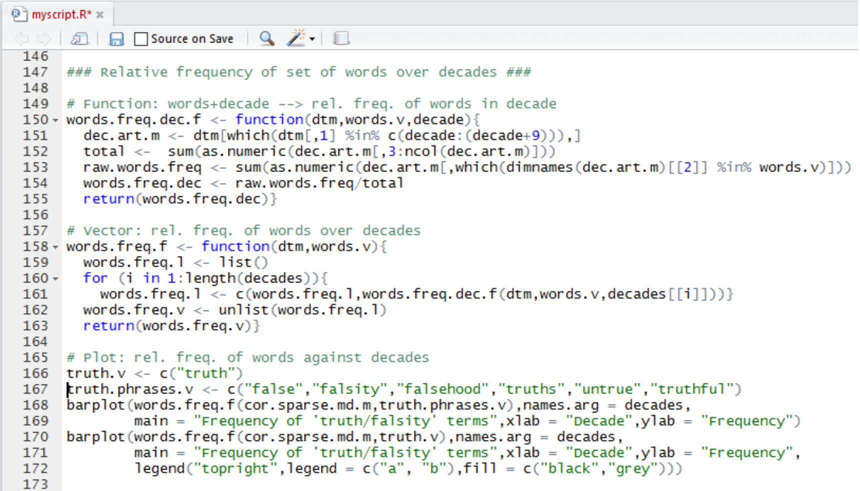
- Subcorpora:
  - 1930s: 38 articles
  - 1940s: 37 articles

    ⋮
  - 2010s: 53 articles

Prevalence of “truth-like” concepts (1)

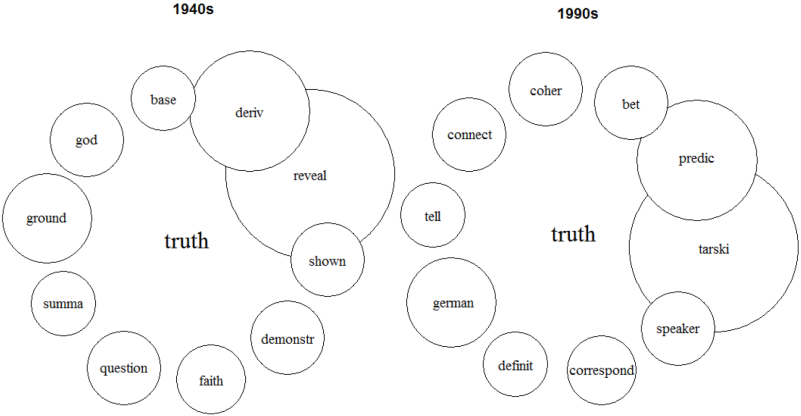


Prevalence of "truth-like" concepts (2)

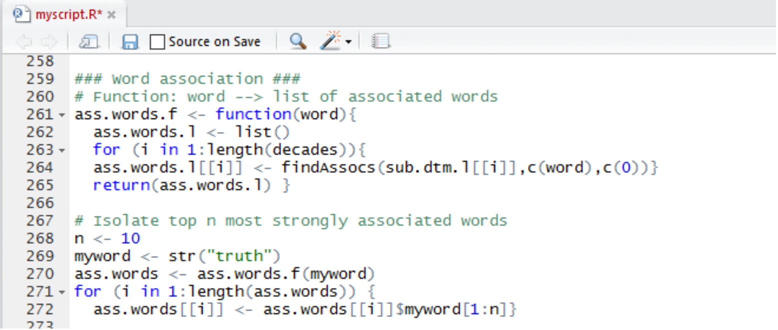


```r

### Relative frequency of set of words over decades ###

# Function: words+decade --> rel. freq. of words in decade
words.freq.dec.f <- function(dtm,words.v,decade){
  dec.art.m <- dtm[which(dtm[,1] %in% c(decade:(decade+9))),]
  total <-  sum(as.numeric(dec.art.m[,3:ncol(dec.art.m)]))
  raw.words.freq <- sum(as.numeric(dec.art.m[,which(dimnames(dec.art.m)[[2]] %in% words.v)]))
  words.freq.dec <- raw.words.freq/total
  return(words.freq.dec)}

# Vector: rel. freq. of words over decades
words.freq.f <- function(dtm,words.v){
  words.freq.l <- list()
  for (i in 1:length(decades)){
    words.freq.l <- c(words.freq.l,words.freq.dec.f(dtm,words.v,decades[[i]]))}
  words.freq.v <- unlist(words.freq.l)
  return(words.freq.v)}

# Plot: rel. freq. of words against decades
truth.v <- c("truth")
truth.phrases.v <- c("false","falsity","falsehood","truths","untrue","truthful")
barplot(words.freq.f(cor.sparse.md.m,truth.phrases.v),names.arg = decades,
        main = "Frequency of 'truth/falsity' terms",xlab = "Decade",ylab = "Frequency")
barplot(words.freq.f(cor.sparse.md.m,truth.v),names.arg = decades,
        main = "Frequency of 'truth/falsity' terms",xlab = "Decade",ylab = "Frequency",
        legend("topright",legend = c("a", "b"),fill = c("black","grey")))

```

Words associated with "truth" (1)



Words associated with "truth" (2)

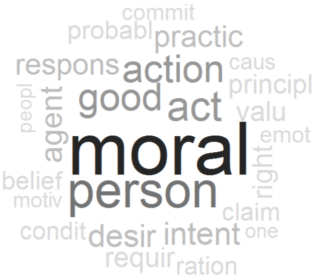


```
### Word association ###
# Function: word --> list of associated words
ass.words.f <- function(word){
  ass.words.l <- list()
  for (i in 1:length(decades)){
  ass.words.l[[i]] <- findAssocs(sub.dtm.l[[i]],c(word),c(0))}
  return(ass.words.l) }

# Isolate top n most strongly associated words
n <- 10
myword <- str("truth")
ass.words <- ass.words.f(myword)
for (i in 1:length(ass.words)) {
  ass.words[[i]] <- ass.words[[i]]$myword[1:n]}
```

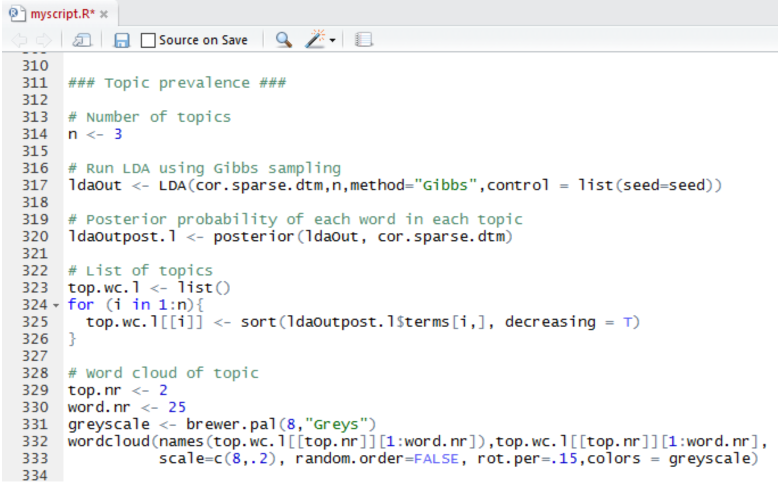
Topic discovery (1)

Topic models are algorithms that allow us to identify hidden thematic patterns in (sets of) texts.

- A topic is treated as a distribution over words.
- A document is treated as a distribution over topics.

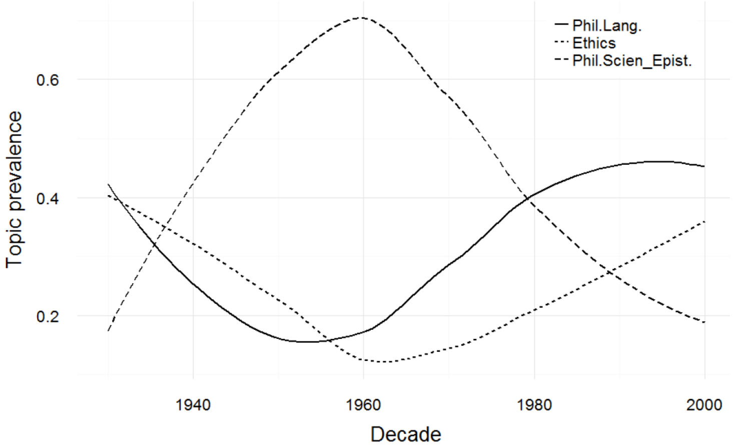


Topic discovery (2)



```r

### Topic prevalence ###

# Number of topics
n <- 3

# Run LDA using Gibbs sampling
ldaOut <- LDA(cor.sparse.dtm,n,method="Gibbs",control = list(seed=seed))

# Posterior probability of each word in each topic
ldaOutpost.l <- posterior(ldaOut, cor.sparse.dtm)

# List of topics
top.wc.l <- list()
for (i in 1:n){
  top.wc.l[[i]] <- sort(ldaOutpost.l$terms[i,], decreasing = T)
}

# Word cloud of topic
top.nr <- 2
word.nr <- 25
greyscale <- brewer.pal(8,"Greys")
wordcloud(names(top.wc.l[[top.nr]][1:word.nr]),top.wc.l[[top.nr]][1:word.nr],
          scale=c(8,.2), random.order=FALSE, rot.per=.15,colors = greyscale)

```

Topic prevalence



Thanks for your attention!