CES Working Papers as a picture

Lars Vilhuber

Contents

Or: the power of **data**.

A picture is worth a thousand words. Or in this case 7138 words - the number of distinct words in titles of the 726 papers accepted.

To produce this graph, we downloaded the ReDif-formatted metadata for the CES Working Paper archive, and read in the Abstract field.

```
# Source: titles of all published papers
tmp <- fread("iconv -t UTF8 -c < ceswp_20160204.rdf | grep -e '^Abstract:' | sed 's/^Abstract:/Abstract

accepted <- tmp[tmp$V1=="Abstract",]
```

We then used the *R text mining library* to clean and parse the titles:

```
doc.vec <- VectorSource(t(accepted))
doc.corpus <- Corpus(doc.vec)
doc.corpus <- tm_map(doc.corpus, content_transformer(function(x) iconv(enc2utf8(x), sub = "byte")))
doc.corpus <- tm_map(doc.corpus, content_transformer(tolower))
#doc.corpus.nw <- tm_map(doc.corpus, stripWhitespace)
doc.corpus <- tm_map(doc.corpus, removePunctuation)
```

```
doc.corpus <- tm_map(doc.corpus, removeNumbers)
doc.corpus <- tm_map(doc.corpus, removeWords, stopwords("english"))
doc.corpus <- tm_map(doc.corpus, removeWords, c("the","abstract","paper","using"))

TDM <- TermDocumentMatrix(doc.corpus)
# find the most frequent word
m <- as.matrix(TDM)
v <- sort(rowSums(m),decreasing = TRUE)

try_max <- v[[1]]
try_five <- v[[5]]
restrict_num <- 30
top_100 <- v[[100]]
most_freq <- findFreqTerms(TDM,try_max)
#
```

which generated a "corpus" of documents.

In fact, we lied somewhat above: we did not show **7138** words, but rather, for the sake of clarity, restricted ourselves to the top 100 words. If we had instead wanted to show the **476** words with at least 30 mentions in the (cleaned) corpus, we would have obtained the following graph:

For the curious, while the most frequent word is **data**, the top **5** are:

	Frequency
data	944
firms	770
productivity	614

	Frequency
plants	472
find	397

- The code behind this endeavor is available at github.com/larsvilhuber/ceswp-wordart
- This document was produced using R, RStudio.