SOLE 2015 Posters 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.

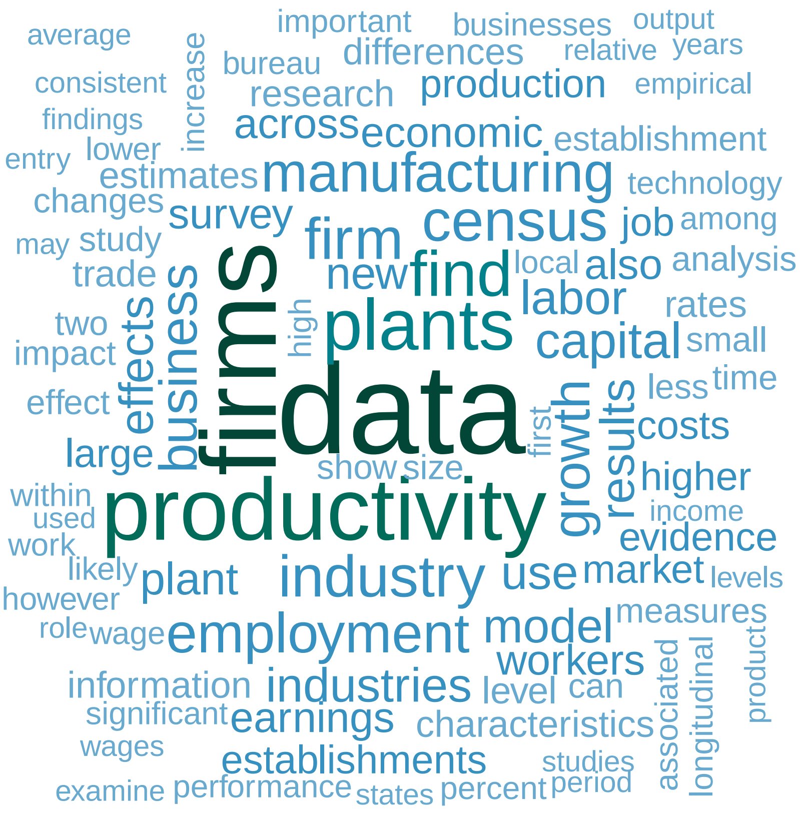
```
## Warning in wordcloud(rownames(as.matrix(TDM)), rowSums(as.matrix(TDM)), :
## relationship could not be fit on page. It will not be plotted.

## Warning in wordcloud(rownames(as.matrix(TDM)), rowSums(as.matrix(TDM)), :
## estimate could not be fit on page. It will not be plotted.

## Warning in wordcloud(rownames(as.matrix(TDM)), rowSums(as.matrix(TDM)), :
## increases could not be fit on page. It will not be plotted.

## Warning in wordcloud(rownames(as.matrix(TDM)), rowSums(as.matrix(TDM)), :
## plantlevel could not be fit on page. It will not be plotted.

## Warning in wordcloud(rownames(as.matrix(TDM)), rowSums(as.matrix(TDM)), :
## provide could not be fit on page. It will not be plotted.
```



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
#tmp2 <- read.delim("ceswp_1988_to_2013.rdf.txt", sep="",header = FALSE)
#tmp <- read.delim("ceswp_1988_to_2013.rdf.txt", sep=":",header = FALSE)
tmp <- fread("iconv -t UTF8 -c < ceswp_1988_to_2013.rdf.txt | grep -e '^Abstract:' | sed 's/^Abstract:/

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

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

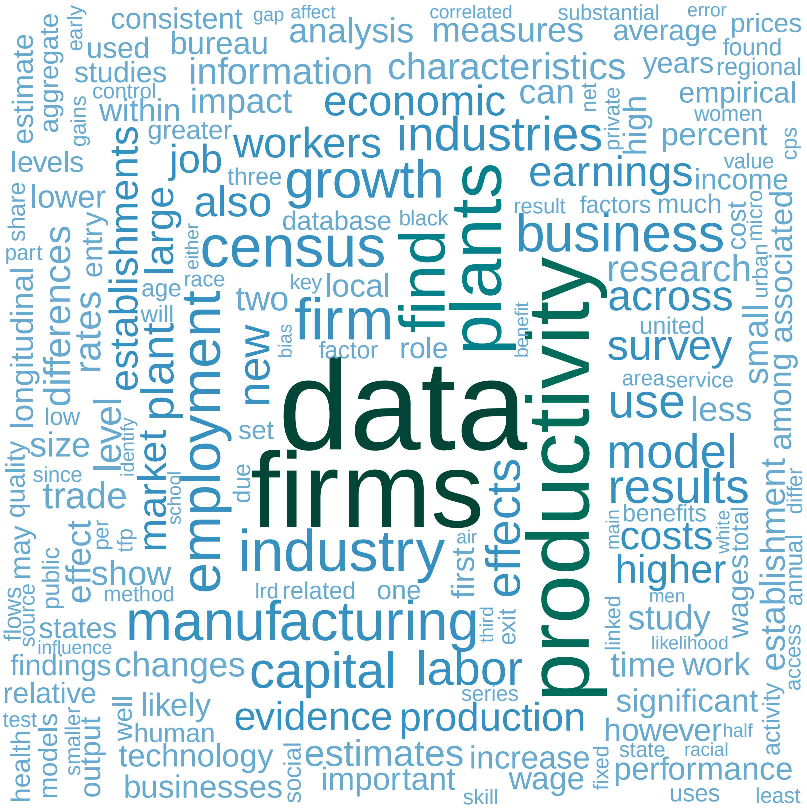
```
doc.vec <- VectorSource(t(accepted))
doc.corpus <- Corpus(doc.vec)
doc.corpus <- tm_map(doc.corpus, content_transformer(function(x) iconv(enc2utf8(x), sub = "byte")))
doc.corpus <- tm_map(doc.corpus, content_transformer(tolower))
#doc.corpus.nw <- tm_map(doc.corpus, stripWhitespace)
doc.corpus <- 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
plants	472
find	397

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

```
R.Version()
```

```
## $platform
## [1] "x86_64-suse-linux-gnu"
##
## $arch
## [1] "x86_64"
##
## $os
## [1] "linux-gnu"
##
## $system
## [1] "x86_64, linux-gnu"
##
## $status
## [1] ""
##
## $major
## [1] "3"
##
## $minor
## [1] "2.3"
##
## $year
## [1] "2015"
##
## $month
## [1] "12"
##
## $day
## [1] "10"
##
## $`svn rev`
## [1] "69752"
##
## $language
## [1] "R"
##
## $version.string
## [1] "R version 3.2.3 (2015-12-10)"
##
## $nickname
## [1] "Wooden Christmas-Tree"
```

```
Sys.info()
```

```
##                                                 sysname 
##                                                 "Linux" 
##                                                 release 
##                                     "3.16.7-29-desktop" 
##                                                 version 
## "#1 SMP PREEMPT Fri Oct 23 00:46:04 UTC 2015 (6be6a97)" 
##                                                nodename 
##                                              "zotique2" 
##                                                 machine 
##                                                "x86_64" 
##                                                   login 
##                                              "vilhuber" 
##                                                    user 
##                                              "vilhuber" 
##                                          effective_user 
##                                              "vilhuber" 
```