Comparing the domains of two Danish broadcast companies

```
# This is for reference for interested parties
# TV stuff

dr <- tbl_df(read.table(“dr.dat”,
           header = FALSE,
           col.names = c("domain",
                    "date",
                    "httpstatus",
                    "size",
                    "uri",
                    "dp",
                    "referrer",
                    "mime",
                    "workerid",
                    "timestamp",
                    "sha1",
                    "sourcetag",
                    "annotations"),
           na.strings = "-",
           strip.white = TRUE,
           comment.char = ""))

# How big is the DR domain
dr %>%
 group_by(httpstatus) %>%
 summarise(sizeSum=sum(as.numeric(size))) %>%
 select(httpstatus,sizeSum) %>%
 arrange(sizeSum) %>%
 filter(httpstatus==200) %>%
 select(sizeSum)

#       sizeSum
# 1 190607816675

# Create a data set with all linked documents external to the domain
dr.outside <- dr %>%
 filter(!grepl("dr.dk",uri), httpstatus == 200, is.numeric(size)) %>%
 extract(uri,"outside.domain","https?://([^/]+)", remove=F) %>%
 select(outside.domain, uri, size)
```

```
# group those URLs according to domain
dr.outside %>%
 group_by(outside.domain) %>%
 summarise(domain.size = sum(as.numeric(size))) %>%
 arrange(desc(domain.size)) %>%
 print(n=50)

# How much data is harvested from outside DR?
dr.size <- sum(as.numeric(dr.outside$size))
# [1] 8,212,917,228

# How big is TV2 domain
tv2 %>%
 group_by(httpstatus) %>%
 summarise(sizeSum=sum(as.numeric(size))) %>%
 select(httpstatus,sizeSum) %>%
 arrange(sizeSum) %>%
 filter(httpstatus==200) %>%
 select(sizeSum)

#       sizeSum
# 1 284906503438

# Create a data set with all linked documents external to the domain
tv2.outside <- tv2 %>%
 filter(!grepl("tv2.dk",uri), httpstatus == 200, is.numeric(size)) %>%
 extract(uri,"outside.domain","https?://([^/]+)", remove=F) %>%
 select(outside.domain, uri, size)

tv2.outside %>%
 group_by(outside.domain) %>%
 summarise(domain.size = sum(as.numeric(size))) %>%
 arrange(desc(domain.size)) %>%
 print(n=50)

# How much data is harvested from outside tv2.dk?

tv2.size <- sum(as.numeric(tv2.outside$size))
[1] 110411610105
```

```
sum(as.numeric(tv2$size))
```

RESULT

DR   4% from outside

TV2 38% from outside!!