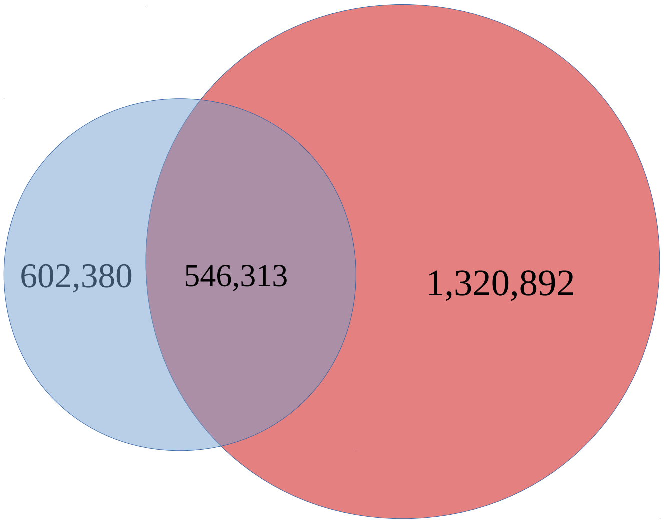
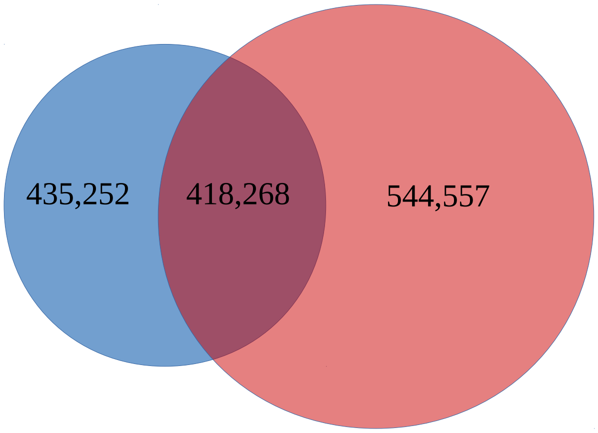
Italian Dataset

Blue: Text + Title scraped for urls.
Red: Total number of urls.



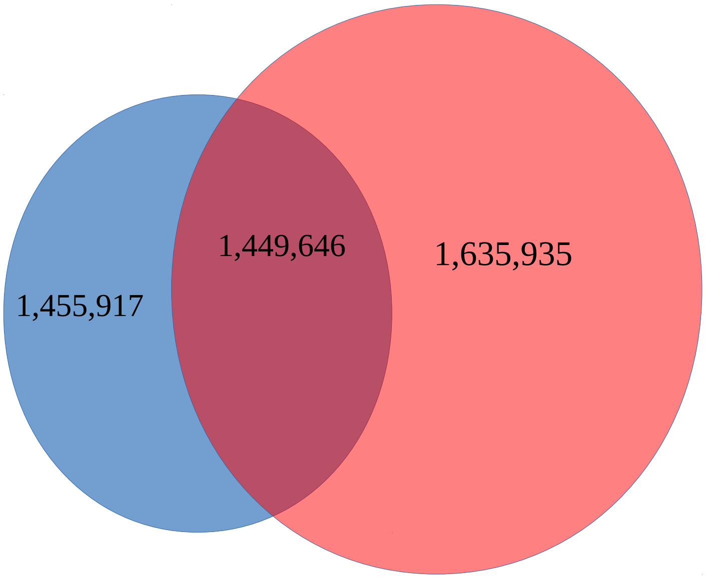
No. of missing articles(un-scraped) = 56,067

PDF 2015



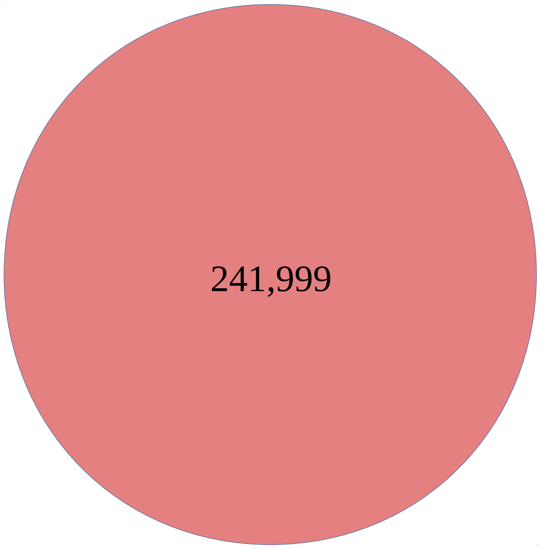
No. of missing articles(un-scraped) = 16,984

PDF 2016



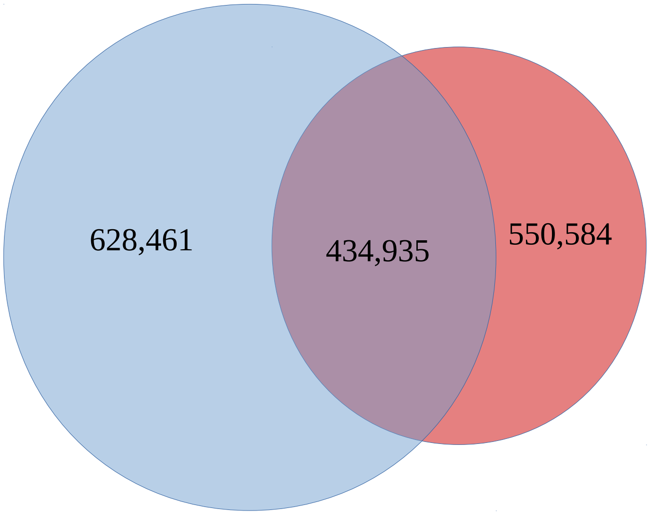
No. of missing urls(un-scraped) = 6271

POLI 2015



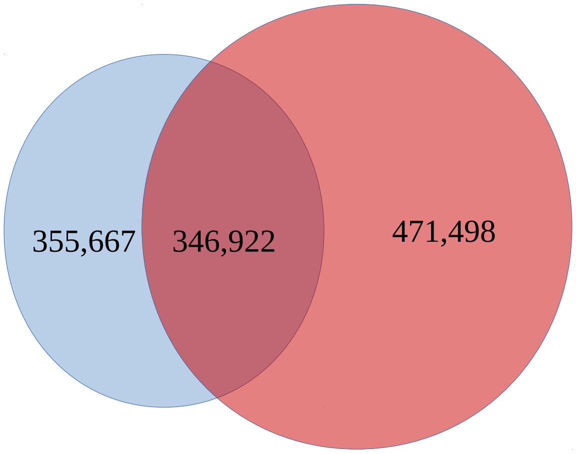
No. of missing urls(un-scraped) = 241,999

POLI 2016



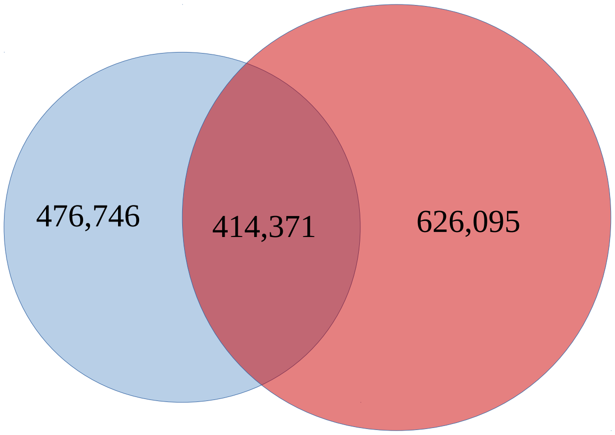
*Extra urls scraped = 193526

PUL 2015



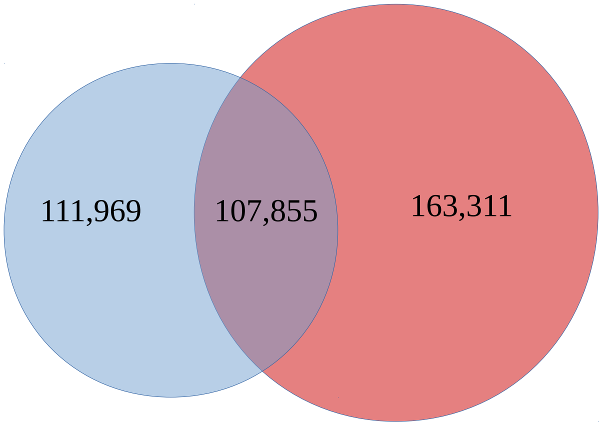
No. of missing urls(un-scraped) = 8745

PUL 2016



No. of missing urls(un-scraped) = 62,375

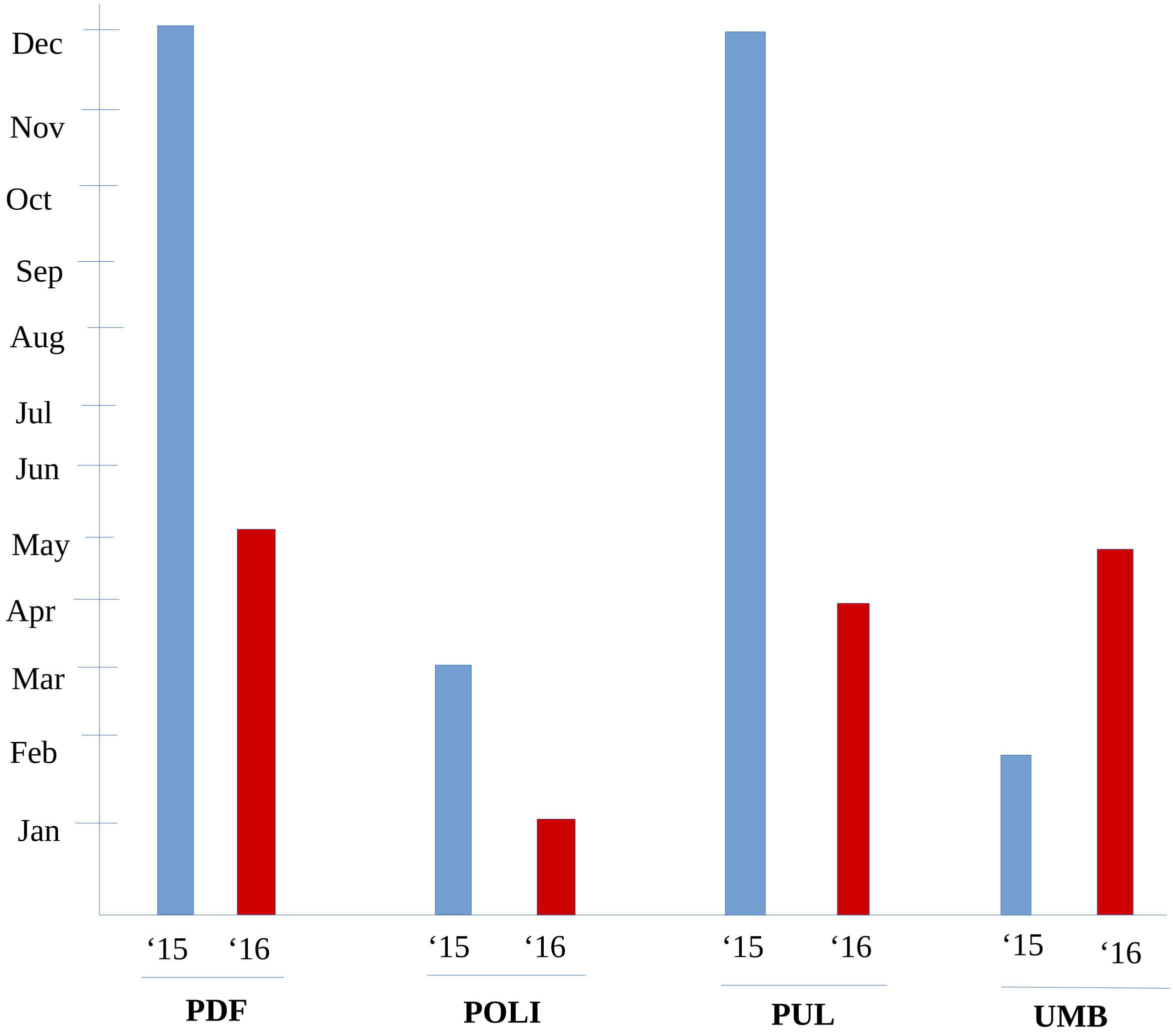
UMB 2015



No. of missing urls(un-scraped) = 4114

UMB 2016

Data comparison statistics:


Dec
Nov
Oct
Sep
Aug
Jul
Jun
May
Apr
Mar
Feb
Jan
‘15
‘16
PDF
‘15
‘16
POLI
‘15
‘16
PUL
‘15
‘16
UMB