

Report capstone project 2022

Erica Ravarelli

Research question

This short research is aimed at investigating the relation between gender and tendency to talk about gender equality in Italian parliament. An insight on the relation between party ideology and tendency to talk about women rights will be given, too. The analysis focuses on the XVII legislature, with a specific attention to 2015.

Previous research have pointed out the fact that political representation is influenced by personal identity traits of the politician, specifically by his/her gender and race. Evidences about the relation between belonging political party and tendency to pay attention to gender equality have been found, too, with a strong association between left parties and sensitivity on gender equality issues. However, there is a lack of studies exploring these relations in the Italian context. My research aims at filling this gap.

The selection concerning time period has been driven by the fact that the highest number of women deputies has been reached in the context of the last legislature, if we exclude the current one which is still ongoing. Moreover, the year 2015 has been the only one in which no change of government took place, suggesting that parliamentary activity hasn't been altered by external interferences.

Research structure and methods

This project is structured as follows: in the first section, I will briefly describe the data collection process, in the second one I will go through the descriptive part of the analysis, whereas in the last section I will illustrate the consistency of the relations through the application of a chi-squared test.

Technical overview

The whole R scripts are available in this Github repository. Therefore, in this short report I will only include and comment the most important passages of the analysis and the related lines of code. However, this does not affect the accessibility and reproducibility of the workflow.

R packages

In the following lines I list all the packages used in my research.

```
library(SPARQL)
library(here)
library(vroom)
library(stringr)
library(tidyverse)
library(lubridate)
library(tidytext)
```

```
library(tidyr)
library(tm)
library(dplyr)
library(tables)
library(forcats)
library(ggplot2)
```

Data collection

Data have been collected through the Open Data portal of Camera dei Deputati, in particular the Virtuoso endpoint has been queried using the SPARQL language; the returned data have been imported into R thanks to the SPARQL package. In both queries, the scraped url is the following one

```
url="http://dati.camera.it/sparql"
```

The first query allowed to import data for the whole XVII legislature, in particular I asked for discussion topic, personal details of the deputy, belonging parliamentary group and office beginning and ending dates, along with intervention date. These last three variables were used in order to apply a filter: indeed, specifying that I was interested only in interventions which had taken place between the starting and the ending office dates have been fundamental in order to avoid including unwanted results. The `intervento` and the `discussione` variables allowed to check that no duplicates were included. Finally, only discussions which took place during the XVII legislature have been selected.

```
"SELECT DISTINCT
?argomento ?nome ?cognome ?genere ?dataNascita
?gruppo_parlamentare ?background ?inizioIncarico ?fineIncarico ?data_intervento
?intervento ?discussione

WHERE {
##define the variable 'dibattito' and select all those belonging to the XVII legislature.
##define the variable 'discussione' starting from the variable 'dibattito'
    ?dibattito a ocd:dibattito;
                ocd:rif_leg <http://dati.camera.it/ocd/legislatura.rdf/repubblica_17>;
                ocd:rif_discussione ?discussione.

##define the variable 'argomento' starting from 'discussione'. Ask for 'data_intervento',
##which will be useful to fix a bug
##select only discussions which have the word 'donna' or 'donne' in their title

    ?discussione ocd:rif_intervento ?intervento;
                rdfs:label ?argomento.
    FILTER(regex(?argomento, 'donn(e|a)', 'i')).
    OPTIONAL{?discussione dc:date ?data_intervento.}

##define the variable 'deputato' (intervento) starting from the variable 'intervento'
##ask for name, surname, gender and profesional background of the deputy
##define the variable 'intervento' and 'background'

    ?intervento ocd:rif_deputato ?deputato.
    ?deputato foaf:firstName ?nome; foaf:surname ?cognome; foaf:gender ?genere.
    ?deputato ocd:rif_ufficioParlamentare ?ufficioparl; rdfs:label ?incarico.
    OPTIONAL {?deputato dc:description ?background.}
```

```

##define the variable 'gruppo_parlamentare' (passing through 'aderisce')

    ?deputato ocd:aderisce ?aderisce.
    ?aderisce rdfs:label ?gruppo_parlamentare.

##define the variables 'inizioIncarico' and 'fineIncarico', which will be useful
to fix a bug

    OPTIONAL{?aderisce ocd:endDate ?fineIncarico.}
    OPTIONAL{?aderisce ocd:startDate ?inizioIncarico.}

##ask for 'data di nascita', starting from 'persona'

    ?persona foaf:surname ?cognome; foaf:firstName ?nome; foaf:gender ?genere.
    ?persona <http://purl.org/vocab/bio/0.1/Birth> ?nascita.
    ?nascita <http://purl.org/vocab/bio/0.1/date> ?dataNascita.

##specify that I only want deputies who interveined during their mandate
##(don't take the same deputy more than once only because they
changed office/parliamentary group)

    FILTER(?data_intervento >= ?inizioIncarico && ?data_intervento < ?fineIncarico).

}
ORDER BY ?cognome"

```

In the second query an additional filter was added, in order to select only debates which took place during 2015. Moreover, a loop has been needed because the Virtuoso endpoint allows to scrape no more than 10000 observations within the same call. Therefore, we firstly have to generate an empty tibble which will then be filled-in with scraped data. Each call allowed to scrape 5000 results, then system was set to sleep for 1 second in order to ask for data in a polite way.

```

"SELECT DISTINCT ?argomento ?nome ?cognome ?genere ?gruppo_parlamentare
?background ?inizioIncarico ?fineIncarico ?data_intervento ?intervento ?discussione
WHERE {
  {
    SELECT DISTINCT ?argomento ?nome ?cognome ?genere ?gruppo_parlamentare
    ?background ?inizioIncarico ?fineIncarico ?data_intervento ?intervento ?discussione

          WHERE {
##define the variable 'dibattito' and select all those belonging to the XVII legislature.
##define the variable 'discussione' starting from the variable 'dibattito'
    ?dibattito a ocd:dibattito;
        ocd:rif_leg <http://dati.camera.it/ocd/legislatura.rdf/repubblica_17>.
    ?dibattito ocd:rif_discussione ?discussione.

##define the variable 'argomento' starting from 'discussione'. Ask for 'data_intervento',
##which will be useful to fix a bug

    ?discussione ocd:rif_intervento ?intervento.
    ?discussione rdfs:label ?argomento.
    ?discussione dc:date ?data_intervento.

```

```

##define the variable 'deputato' (intervenuto) starting from the variable 'intervento'
##ask for name, surname, gender and professional background of the deputy
##define the variable 'intervento' and 'background'

?intervento ocd:rif_deputato ?deputato.
?deputato foaf:firstName ?nome; foaf:surname ?cognome; foaf:gender ?genere.
OPTIONAL {?deputato dc:description ?background.}

##define the variable 'gruppo_parlamentare' (passing through 'aderisce')

?deputato ocd:aderisce ?aderisce.
?aderisce rdfs:label ?gruppo_parlamentare.

##define the variables 'inizioIncarico' and 'fineIncarico',
which will be useful to fix a bug

OPTIONAL{?aderisce ocd:endDate ?fineIncarico.}
OPTIONAL{?aderisce ocd:startDate ?inizioIncarico.}

##specify that I only want deputies who intervened during their mandate
##(don't take the same deputy more than once only because
##they changed office/parliamentary group)

FILTER(?data_intervento >= ?inizioIncarico && ?data_intervento < ?fineIncarico).
FILTER(REGEX(?data_intervento, '^2015', 'i')).
}
ORDER BY ?cognome}
}
LIMIT 10000
OFFSET"

tot2015 <- tibble()

#offset to ask for more than 10 000 results-----
query_offset <- c("0", "5000", "10000", "20000", "30000", "40000",
                  "50000", "60000")

i <- 0
for (i in 1:length(query_offset)) {
  interventi_deputati <- str_c(query,
                              query_offset[i],
                              sep = " ")
  final_dataset <- SPARQL(url, interventi_deputati)

  tot2015 <- rbind(tot2015, final_dataset$results)
  Sys.sleep(1)
}

```

After collecting them, data were saved in a csv format through the `write.csv()` command, in order to avoid having to query the endpoint more than needed.

```
write.csv(geip17, here::here("Data.csv/geip17.csv"))
write.csv(tot2015, here::here("Data.csv/tot2015.csv"))
```

Data cleaning

Once collected and stored, data have been cleaned with the aim of preparing the analysis.

XVII legislature

First of all, dates have been converted into a readable format through the `lubridate` package.

```
geip17$dataNascita <- ymd(geip17$dataNascita)
geip17$data_intervento <- ymd(geip17$data_intervento)
geip17$inizioIncarico <- ymd(geip17$inizioIncarico)
geip17$fineIncarico <- ymd(geip17$fineIncarico)
```

After that, regular expressions were used in order to recode the variable concerning the deputies' belonging parliamentary group. In particular:

- The Five Star Movement has been considered as a stand-alone party, taking as valid the narration about not being neither a left nor a right part which was loudly upheld by the Movement members.
- Analogously, but for a different and quite self-explanatory reason, deputies belonging to the mixed group were not assigned to a different category.
- The `left` category includes: “movimento democratico e progressista”, “partito democratico”, “sinistra ecologia libertà”, “sinistra italiana”.
- The `centre` category includes: “alternativa popolare-centristi per l'europa, per l'Italia”.
- The `right` category includes: “forza italia-il popolo della libertà-Berlusconi presidente”, “fratelli d'Italia-alleanza nazionale”, “il popolo della libertà-Berlusconi presidente”, “lega nord e autonomie”, “nuovo centrodestra”.

```
m5s <- str_detect(geip17$gruppo_parlamentare, "STELLE")
mis <- str_detect(geip17$gruppo_parlamentare, "MISTO")
sx <- str_detect(geip17$gruppo_parlamentare, "SINISTRA|DEMOCRATICO")
cen <- str_detect(geip17$gruppo_parlamentare, "CIVIC|POPOLARE|PER")
dx <- str_detect(geip17$gruppo_parlamentare, "POPOLO|FRATELLI|LEGA|DESTRA")
```

After making these decisions in terms of ideological positioning of the parliamentary group, the variable `gruppo_parl` has been generated through as `ifelse` command.

```
gruppo_parl <- ifelse(m5s == TRUE, "M5S",
  ifelse(mis == TRUE, "Misto",
    ifelse(sx == TRUE, "Sinistra",
      ifelse(cen==TRUE, "Centro",
        ifelse(dx==TRUE, "Destra",
          "altro")))))
```

Tokenisation