

Week 2 Tutorial: Preprocessing and APIs

POP77142 Quantitative Text Analysis for Social Scientists

PDF Files

- PDF (**P**ortable **D**ocument **F**ormat) is a file format that captures all the elements of a printable document.
- It is both ubiquitous for digital documents and often contains a large amount of textual data.
- However, extracting text from PDFs can be challenging as a document contains a lot of additional information (e.g. images, layouts, fonts, etc.). That is before we even consider non-machine readable PDFs that store text as images.
- Here we will consider a commonly used data source in political science - electoral manifestos of parties that are often distributed in PDF format.



Extra

Intro to PDF by Leonard Rosenthol

PDF as Text File

```
1 # Extract manifestos from the archive in terminal
2 # Alternatively, you can use GUI for this
3 unzip -o ../data/ireland_ge_2024_manifestos.zip -d ../temp/
```

```
Archive: ../data/ireland_ge_2024_manifestos.zip
  inflating: ../temp/ireland_ge_2024_manifestos/A0.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/FF.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/FG.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/GR.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/II.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/LAB.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/PBP.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/SD.pdf
  inflating: ../temp/ireland_ge_2024_manifestos/SF.pdf
```

```
1 # List files and show their sizes (-s) in human readable format (h)
2 ls -sh ../temp/ireland_ge_2024_manifestos
```

```
total 71M
 9.8M A0.pdf
 14M FF.pdf
 6.5M FG.pdf
 22M GR.pdf
 2.0M II.pdf
592K LAB.pdf
 1.4M PBP.pdf
 11M SD.pdf
 6.1M SF.pdf
```

```
1 manifestos=$(ls ../temp/ireland_ge_2024_manifestos)
```

```
1 echo $manifestos
```

PDF as Text File

- As with other file formats it is a good idea to first try opening a file in a text editor.
- Here a simple `more` command is used to display the content of the first manifesto.

```
1 more "../temp/ireland_ge_2024_manifestos/${manifestos[1]}"
```

- For a more user-friendly and text-focussed you might try `less` command as well (note that it is not directly available in Windows).

```
1 less "../temp/ireland_ge_2024_manifestos/${manifestos[1]}"
```

Exercise 1: Working with PDFs

- In this exercise we will consider party manifestos from the last Irish General Election.
- The ZIP archive with all main party manifestos is available on Blackboard.
- Download the archive and extract all PDF files to a temporary directory.
- Use `pdftools` package (you might try Python's `pypdf` if you are interested in Python approach) to extract text from the first PDF file in the directory.
- Tokenise the text and count the number of tokens.
- What if you stem the tokens?
- Now repeat this for all party manifestos and create a data frame of the following form:

party	year	text	npages	ntokens	ntypes
...	...	...	...	...	...

where:

- `party` is the party name,
- `year` is the year of the election,
- `text` is the text of the manifesto,
- `npages` is the number of pages in the manifesto,
- `ntokens` is the number of tokens in the manifesto,
- `ntypes` is the number of types in the manifesto.

```
1 library("pdftools")
```

```
1 # List all PDF files in the directory
2 manifestos <- list.files(
3   "../temp/ireland_ge_2024_manifestos",
4   pattern = ".pdf$",
5   full.names = TRUE
6 )
7 manifestos
```

```
[1] "../temp/ireland_ge_2024_manifestos/A0.pdf"
[2] "../temp/ireland_ge_2024_manifestos/FF.pdf"
[3] "../temp/ireland_ge_2024_manifestos/FG.pdf"
[4] "../temp/ireland_ge_2024_manifestos/GR.pdf"
[5] "../temp/ireland_ge_2024_manifestos/II.pdf"
[6] "../temp/ireland_ge_2024_manifestos/LAB.pdf"
[7] "../temp/ireland_ge_2024_manifestos/PBP.pdf"
[8] "../temp/ireland_ge_2024_manifestos/SD.pdf"
[9] "../temp/ireland_ge_2024_manifestos/SF.pdf"
```

```
1 # Extract text from the first PDF file
2 txt <- pdftools::pdf_text(manifestos[1])
3
4 # Note that this created a vector whose length
5 # is equal to the number of pages in the PDF
6 length(txt)
```

```
[1] 104
```

```
1 head(txt)
```

```
[1] "      Our\nCommon Sense\n Manifesto 2024\n"
```

```
[2] "      Opening statement\nIn the last year Aontú has come of age. Our vote has surpassed Labour,  
the Soc\nDens and PbP in the last election. We were the only party that stood with the\npeople in the recent  
referendums. We made the common-sense argument and we\nwon.\n\nIndeed, Aontú has, for the last five years, Aontú  
has often stood alone in the Dáil\nfor you. We have held our ground in your interest. Whether it has been  
raising real\nconcerns about the phenomenal government waste, the disastrous government\nimmigration policy, why  
100,000 homes remain empty in the middle of a housing\ncrisis, censorship legislation or carbon taxes on fuel,  
we have been on the right side\nof the argument - your side.\n\nWe are offering a step change in how this  
country is governed. We want to end the\nculture of waste, lack of accountability and indifference shown to many  
across the\nstate who need government to deliver for them.\n\nWe have done that with one TD. Imagine the impact  
we can have with a team of\nAontú TDs fighting your corner and standing firm on issues that matter to you,\nyour  
business and your family. That power is now in your hands. You can change\nthe direction of this country with  
your vote.\n\nWe have proven that we can deliver, and more importantly, that no matter what\nway the wind blows  
or how many other parties follow the latest trend, we will be\nresolute in what we believe in and won't be  
afraid to use our voice for you in the\nbest interests of the country.\n\nI am immensely grateful to all of our  
activists, our newly elected representatives\nand our supporters for standing with us throughout the lifetime of  
the 33rd Dáil. As\nwe look into the 34th, I am confident that 'Team Common Sense' will challenge the\npolitical  
establishment like never before.\n\nThe is a wave of support coming towards Aontú at the doors. Opinion polls  
are\nshowing arise in our support. More and more people are hearing our message. For\nthe first time, there is  
an Aontú candidate running in every constituency, giving\nevery citizen the opportunity to vote Aontú #1.\n\nIn
```

```
1 library("quanteda")
```

```
1 # Tokenise the text  
2 toks <- quanteda::tokens(txt, remove_punct = TRUE)  
3 ntokens <- quanteda::ntoken(toks)  
4 sum(ntokens)
```

```
[1] 28423
```

```
1 ntypes <- quanteda::ntype(toks)  
2 sum(ntypes)
```

```
[1] 16084
```


Exercise 2: Working with APIs

- In this part we will try using Guardian API to collect newspaper articles.
- First, you will need to go through a simple registration process to obtain an API key.
- The main **endpoint** of the Guardian API is the content or search endpoint.
- Study the documentation of the search endpoint.
- Below is a simple example of working with the Guardian API to search for articles related to Ireland.
- Modify the search query to search for articles related to elections in Ireland.
- Set the time frame for around the time of the last GE (e.g. between 1 November and 31 December 2024).
- How many articles are returned?

```
1 library("httr")
```

```
1 # It is a good idea to not hard-code the API key in the script  
2 # and, instead, load it dynamically from a file  
3 api_key <- readLines("../temp/guardian_api_key.txt")
```

```
1 # Define the URL for the API  
2 base_url <- "https://content.guardianapis.com/search"
```

```
1 # Define the query parameters  
2 params <- list(  
3   "api-key" = api_key,  
4   "q" = "ireland",  
5   "page-size" = 10  
6 )
```

```

1 # Make the request and receive the response
2 response <- httr::GET(url = base_url, query = params)
3
4 # Check the status of the response
5 response$status_code

```

```
[1] 200
```

```

1 # Extract the content of the response
2 json <- httr::content(response, as = "text", encoding = "UTF-8")

```

```
1 json_parsed <- jsonlite::fromJSON(json)
```

```
1 str(json_parsed)
```

List of 1

\$ response:List of 9

```

..$ status      : chr "ok"
..$ userTier    : chr "developer"
..$ total       : int 112806
..$ startIndex  : int 1
..$ pageSize    : int 10
..$ currentPage: int 1
..$ pages       : int 11281
..$ orderBy     : chr "relevance"
..$ results     : 'data.frame': 10 obs. of  11 variables:
.. ..$ id       : chr [1:10] "world/2024/dec/02/ireland-election-2024-latest-results"
"sport/live/2025/feb/22/wales-ireland-six-nations-rugby-union-live" "commentisfree/2025/mar/14/ireland-neutral-
europe-rearm-triple-lock-donald-trump" "sport/2025/feb/01/ireland-england-six-nations-rugby-union-player-
ratings-aviva-stadium" ...
.. ..$ type     : chr [1:10] "article" "liveblog" "article" "article" ...
.. ..$ sectionId : chr [1:10] "world" "sport" "commentisfree" "sport" ...
.. ..$ sectionName : chr [1:10] "World news" "Sport" "Opinion" "Sport" ...
.. ..$ webPublicationDate: chr [1:10] "2024-12-03T07:45:44Z" "2025-02-22T16:41:09Z" "2025-03-14T11:40:10Z"
"2025-02-01T20:14:35Z" ...

```

```
1 head(json_parsed$response$results)
```

```

                                                                                                                                               id
1                                world/2024/dec/02/ireland-election-2024-latest-results
2                                sport/live/2025/feb/22/wales-ireland-six-nations-rugby-union-live
3                                commentisfree/2025/mar/14/ireland-neutral-europe-rearm-triple-lock-donald-trump
4                                sport/2025/feb/01/ireland-england-six-nations-rugby-union-player-ratings-aviva-stadium
5                                uk-news/2024/dec/26/police-surveillance-of-journalists-in-northern-ireland
6 sport/2025/mar/08/fabian-galthie-fears-antoine-dupont-suffered-serious-injury-against-ireland
  type      sectionId sectionName  webPublicationDate
1 article        world  World news 2024-12-03T07:45:44Z
2 liveblog       sport      Sport 2025-02-22T16:41:09Z
3 article commentisfree  Opinion 2025-03-14T11:40:10Z
4 article        sport      Sport 2025-02-01T20:14:35Z
5 article        uk-news    UK news 2024-12-26T17:26:05Z
6 article        sport      Sport 2025-03-08T19:18:15Z
                                                                                                                                               webTitle
1                                Ireland election 2024: full results
2                                Wales 18-27 Ireland: Six Nations - as it happened
3 Europe is rapidly rearming. Will that leave neutral Ireland defenceless? | Brigid Laffan
4                                Ireland 27-22 England: Six Nations player ratings
5                                Police surveillance of journalists in Northern Ireland | Letters
-                                - - - - -
```