POP77142 Assignment 1: Text Preparation

Before Submission

- Make sure that you can run all cells without errors
- You can do it by clicking `Kernel`, `Restart & Run All` in the menu above
- Make sure that you save the output by pressing Command+S / CTRL+S
- Rename the file from `01_assignment.ipynb` to `01_lastname_firstname_studentnumber.ipynb`
- Use Firefox browser for submitting your Jupyter notebook on Blackboard.

Overview

In this assignment you will need to collect and prepare textual data for analysis. As the data source we will debates in the Dáil Éireann (Irish Parliament) for the first 2 months of 2025 (but in practice once you implement a solution for those it should be relatively straightforward to scale up).

There are 2 broad strategies that can be used to obtain Dáil debates:

1. Use the Oireachtas website to scrape the debates using R (e.g. `rvest`) or Python (e.g. `Beautiful Soup`). There can be different strategies to solve this, but, crucially, the website is largely static, so dealing with it as a set of HTML files is quite manageable.
2. Use the Oireachtas API to scrape the debates using R (e.g. `httr2`) or Python (e.g. `requests`). This might be a more advanced option, but it is also a lot more powerful and flexible. Importantly, this API does not require authentication, which makes working with it quite a bit simpler than with many other APIs.

Part 1: Data Acquisition

In this part you will need to write a scraper that collects the data either directly from the Oireachtas website or using the Oireachtas API. The data should be collected for the first 2 months of 2025 (January and February, but the bulk of the debates would be in February).

Depending on how you choose to organise your code, you may choose to build up a usual tabular dataset straightaway or you might find it easier to store the data in a different container (e.g. a list of vectors, a list of lists, a list of dictionaries, etc.) and then convert it to a tabular format in the next part.

You may use generative AI to help you with trialing different approaches. If you do use AI, you need to report the version of the LLM that you are using (e.g. `code-davinci-002`, `meta-llama-3.1-8b-intruct`, etc.). Hardware permitting, I encourage you to use offline models to have better control over the data and the model.

While there maybe also some bindings for the API that are readily available, none of them are officially supported, so you shouldn't be relying on those.

Part 2: Text Preprocessing

In this part you will need to clean up the collected data. Depending on how the previous part was implemented it might take more or fewer steps. The ultimate goal is to have a dataset of the following form:

dail	vol	no	date	speaker	text	ntokens	ntypes

where:

`dail` - is the number of the Dáil (e.g. 34th Dáil)

`vol` - is the volume number of the debates (e.g. 1000)

`no` - is the number of the debate in the volume (e.g. 1)

`date` - is the date of the debate (in YYYY-MM-DD form, e.g. 2025-01-01)

`speaker` - is the name of the speaker

`text` - is the text of the speech

`ntokens` - is the number of tokens in the speech

`ntypes` - is the number of types in the speech

Note that you **don't** need to submit the actual dataset. However, after organising the textual data in this way, you will need to perform the following steps:

- Print out the first and last 5 rows of the data
- Print the dimensionality of the data (number of rows and number of columns)
- Print the total number of unique speakers in the dataset.