Week 4 Tutorial: Dictionaries and Text Classification

POP77032 Quantitative Text Analysis for Social Scientists

Dictionaries

- A list of **key-value pairs** where:
  - **Keys** - labels for equivalence classes for the concept of interest.
  - **Values** - terms or patterns that are declared equivalent occurrences of the key class
- In practice might resemble Python dictionary objects:

```python
# In Python
ideo_dict = {
  "liberal": ["benefits", "worker", "trade union"],
  "conservative": ["restriction", "immigration", "reduction"]
  }
```

```r
# In R
ideo_dict <- list(
  liberal = c("benefits", "worker", "trade union"),
  conservative = c("restriction", "immigration", "reduction")
)
```

- But, ultimately, are independent of programming language (usually, some form of text file).
- Dictionaries are, however, highly dependent on natural language!

Exercise 1: Dictionary Application

- In this exercise, we will apply a dictionary to a set of Irish party manifestos for 2020 and 2024 General Elections.
- This time we will skip the parsing part and go straight to text.
- You can find the manifestos in the `ireland_ge_2020-24_manifestos.csv` file available on Blackboard.
- Try to apply the LexiCoder and Laver & Garry dictionaries to the manifestos.
- Calculate the overall sentiment in each of them trying difference scaling formulae.
- Try plotting the changes in the estimated measures (ideology or sentiment) over time across parties.

```r
manifestos <- readr::read_csv(
  "../data/ireland_ge_2020-24_manifestos.csv"
)
```

```r
str(manifestos)
```

```
spc_tbl_ [17 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ party: chr [1:17] "AO" "FF" "FG" "GR" ...
 $ year : num [1:17] 2024 2024 2024 2024 2024 ...
 $ text : chr [1:17] "Our\nCommon Sense\n  Manifesto 2024\n\n                Opening statement\nIn the last year
Aontú has come of ag"| __truncated__ "MOVING FORWARD. TOGETHER.\n\nAG BOGADH AR AGHAIDH. LE CHÉILE.\nGeneral
Election Manifesto 2024\n\n\n\n\n       "| __truncated__ "General Election 2024\n   M A N I F E S TO\n
1\n\nFINE GAEL | GENERAL ELECTION MANIFESTO"| __truncated__ "towards\n2030\na decade of change\nvolume II\nGreen
Party Manifesto 2024\n\n\n\n\n              greens\n       "| __truncated__ ...
 - attr(*, "spec")=
  .. cols(
  ..   party = col_character(),
  ..   year = col_double(),
  ..   text = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

Exercise 2: Dictionary Creation with LLM

- Try using an LLM model to create your own dictionary.
- You can apply to a set of party manifestos or another dataset of your choosing.
- Start by writing a comprehensive prompt for a relatively unambiguous concept.
- Experiment with asking a generative AI model to produce a simple list of terms as opposed to data containers that can de directly integrated into R or Python code.
- Try breaking down a manifesto (or other document used) into sentences and hand-coding a few examples for the defined concept.
- Then apply a created dictionary.
- How do the results compare?