Tutorial 2 Worksheet AY 22/23 Sem 1

DSA2101

Practice with regular expressions

This tutorial focuses on regular expressions, and cleaning and extracting information from text data. *Try to answer each question in more than one way.* That's a good way to verify that our answer is correct, and also to challenge ourselves.

The zip file `ieer.zip` contains collections of news documents from the NLTK (natural language toolkit). Unzip it and store the files in your data folder. Run the following command to read the filenames into R as a vector.

```r
fnames <- list.files("../data/ieer", pattern = "^[^R]", full.names = TRUE)
```

Each file contains a set of news documents, tagged with information such as the date, time and source of each document. The text of each document contains tags that indicate the type of entity, e.g. PERSON, ORGANIZATION, etc. (The dataset is from a Named Entity Recognition training task).

1. Extract
   - the basename of the files.
   - the directory immediately above, along with the filename.
2. Which files contain documents written in April?
3. Read the contents of the first file APW_19980314 into R using `readLines()`, and find the start and end positions of each document within it.
4. From the first file, extract
   - all document numbers,
   - all document datetimes. Store these as a vector `date_strings`.
5. Swap the month and day positions in the `date_strings`.
6. Extract the unique set of entity types found across all files.
7. Find and return only the filenames that contain documents with the word "Israel" in them.
8. Extract all *sentences* in the `TEXT` sections in APW_19980314, containing the key-word `police`.