Reading Nexis Uni News Articles into R and Basic Preprocessing

Trevor Burrows

3/13/2021

Nexis Uni allows you to download full-text news articles in PDF, DOC, or RTF formats. This can be a great source of material for text analysis but a number of steps are needed to bring the articles into R in a useful way. The `nexisFunctions.R` file has several functions to help with reading Nexis Uni news articles into R, isolating date and publication metadata, and doing some initial cleaning.

Note: This process works specifically for news articles from Nexis Uni downloaded as PDFs. It does not work on other types of documents downloaded through Nexis Uni or other file formats. See other caveats below. This file and repo will be updated with additional functions later.

Download your articles from Nexis Uni

1. Perform your search and select your articles using the selecting boxes to the left of each item.
2. Download your articles: with the same articles still selected, click "Download" and choose the following options:

- Under Basic Options, choose "Full Documents", "PDF", and "Save as Individual Files"; if you want, you can specify a filename for the zip file.
- Uncheck *all* the boxes on "Formatting Options" tab
- On the content-specific option, uncheck the "include classification information" box When complete, click "Download."

3. Unzip your downloaded files into a folder accessible by your instance of R. If you need to perform multiple downloads, be sure to put all of your articles in the same folder.

Use `processNexisNews` to read your files into a dataframe with some basic preprocessing

Once your files are in a central location, use `source()` to read the nexisFunctions.R file into R.

```
source("nexisFunctions.R")
```

Then use `processNexisNews()` to process the files and return a dataframe.

`processNexisNews()` takes a single argument: the filepath to the directory holding your Nexis Uni. Given that filepath, `processNexisNews()` will:

- Read in the files using `readtext::readtext()` as a dataframe.
- Identify the date of publication and the publication title for each article, storing them in individual columns and converting the date into a R-readable YYYY-MM-DD format.**
- Remove everything except the section identified as the body of the article's text.
- Return the final dataframe as a tibble.

```
processNexisNews("data/")
```

```
## # A tibble: 100 x 4
##    doc_id                   date       publication       text
##    <chr>                    <date>     <chr>              <chr>
##  1 'Adriana Lecouvreur'_ i~ 2021-03-11 Il Resto del Car~ "Il film-opera 'Adrian~
##  2 'I've been full of happ~ 2021-03-10 The Irish Times   "UK-based soprano Anna~
##  3 'In ascolto dell'Innamo~ 2021-03-12 Il Resto del Car~ "SCANDIANO Ultimi vide~
##  4 'It's essentially pure ~ 2021-03-11 standard.co.uk    "If ever there were a ~
##  5 'Nothing more than a so~ 2021-03-11 Express Online    "Alicia Kearns told LB~
##  6 [Herald Interview] Eric~ 2021-03-12 THE KOREA HERALD  "The highly anticipate~
##  7 ___ANSA_ Quasi 70 milio~ 2021-03-11 ANSA Notiziario ~ "(ANSA) - 11 MAR - Un'~
##  8 _Asse di penetrazione, ~ 2021-03-11 La Nazione (Ital~ ""I lavori stradali e ~
##  9 _Darsena Europa, tempi ~ 2021-03-12 La Nazione (Ital~ "LIVORNO Darsena Europ~
## 10 _Erosioni, il Tresinaro~ 2021-03-11 Il Resto del Car~ "SCANDIANO È ufficialm~
## # ... with 90 more rows
```

** In some cases, the date of publication for non-English news sources will not be used because it would not be easily converted to the desired format; instead, the load date provided by Nexis Uni will be used as the date. In most cases the load date is identical or within a day of the listed publication date. Hoping to tweak this in the future for better results!