README

Scraping the Director's Guild of Canada production list

This is a project to help determine the number of active director's guild of Canada Productions at any given time. The pdf from https://www.dgc.ca/en/british-columbia/avails-and-production-lists/production-list/ is downloaded manually. At this point the file is available at https://www.dgc.ca/assets/Uploads/BritishColumbia/AvailsProductionLists/Documents/Production-List.pdf. This raw data is stored in the "pdf" subfolder of the "data" subfolder in the project. The file is then run through Tabula available at https://tabula.technology/, which is a cross platform open source project for changing pdf data into a tabular format. The output from Tabula is not actually tabular and therefore not all that useful, but at least it can be accessed programmatically and turned into a tabular format. This is done using R and RStudio. After the data is in tabular format it is loaded into Tableau to be visualized. This part might be remade in D3 so as to be usable for free on the web.

Software versions

- Tabula: Version 1.2.1.18052200 (1)
- RStudio: Version 1.2.5033

```
## [1] "R version 3.6.3 (2020-02-29)"
```

```
## [1] "R packages"
```

```
## [1] "tidyverse 1.3.0"
## [1] "data.table 1.12.8"
## [1] "lubridate 1.7.4"
```

Tabularizing the data

In Tabula, The settings are left as the default and the entirety of every page is selected by drawinga rectangular selection around the first page (excluding the very top with the title), selecting the entire second page then clicking on repeat for the rest of the document. The template for this selection was saved as **Production-List-20200906** and can probably be used in the future to make selections. (Playing around with it will be pretty self explanatory). When the **Extract Data** button is pushed it will download the csv to the downloads folder. Move the exported file to the data folder in the project.

Next open RStudio and replace the path of the **INPUT_FILE** to be the path of the tabula output. Next run the script line by line. The "names" variable should be checked to make sure the number of productions adds up with the number of productions. There are usually a couple of short words made up of all caps that end up in the list of names that can be removed with by adding them to the **JUNK_TO_REMOVE** variable.

The **GREP_DATES** shouldn't need to be changed, but there might be some noise at the end of the startAndEndDates list to be removed by adding in more characters to split the last field and remove it from consideration as a possible part of the end date.

Once the end of the **cleanProductions.R** script is reached then the exported csv with a name like **cleandedtabula-Production-List-20200906.csv** can be used in the tableau workbook that creates the GANTT chart to show an anticipated schedule of the shows.

The **getActiveProductionCounts.R** script is not as finicky as the previous and will likely just run without issue (I haven't had one yet). It outputs three CSVs. The first one with a name like **tabula-Production-List-20200906WithActiveProductionbyDate.csv** is not all that useful but the other two that look like **MOWtabula-Production-List-20200906activeProductionsperDay.csv** and **tabula-Production-List-20200906activeProductionsperDay.csv** can be used in the tableau file for creating count of active productions chart.

The tableau files are named: **Individual film job schedules.twbx** and **count of active film productions.twbx**.

Next Steps

- Automate the data download in R
- Remake the dashboards in D3.