Text-mining news on Social Work Profession in Singapore :singapore:

README WORK IN PROGRESS

Data-Search

I searched for news articles in **Factiva**

Note: i previously had done search in Nexis University (previously Lexis Nexis Academic). But this database does not have: (1) articles before year 1994 and (2) The New Paper. See additional notes in Misc about differences in extracting data between Lexis and Factiva

Search date was on 12th Feb 2021

A total of 10427 articles were downloaded in html format.

Search parameters

(1) Date of articles: All date range available

(2) Language: English

(3) Date: Search was done by year

(4) Sources: Five newspapers in Singapore:

  - Channel NewsAsia
  - Today (Singapore)
  - The Straits Times (Singapore)
  - The Business Times (Singapore)
  - The New Paper

(5) Search string: "(social work) OR (social worker*) OR (social-work) OR (social-worker*)"

(6) Search fields: "All fields" (i.e., including title, leading section, body etc)

Results of search: 10427 articles spanning year 1989-2021

Note: Lexis University's policies do not allow me to share the data

Data-extraction

The data extraction involves extracting text content and meta-data information (e.g. title, data, source, author etc) I relied on tm.plugin.factiva package to do this. However, the package requires the files to be in html. Downloading the files from Factiva as html is rather laborious. This website may be helpful. I have a video recording of how I did it. Feel free to ask me for it.

Data-Preparations & Classification

Classification Model

I developed a classfication model using various predictors to classify the articles. Training was developed using 1000 randomly sampled articles. Classification algorithms used included logistic regression, random forest, and GBM. 10-fold cross-validation resampling with 5 repeat were used to test the model. Accuracy was 84% which was significantly better than the no-information rate of 74%. The model was then implemented to predict the rest of the articles. I used a threshold of .80 to classify articles as "Yes - included for the study".

After prediction, I also manually looked through the included articles to further assess articles that should not be included. After further checking, another 500 articles were removed due to several reasons:

(1) not from Singapore (e.g. social work in UK or USA)
(2) voluntary work
(3) Summary of headline (e.g., “What’s news”)
(4) Social workers who are interviewed for non-social work issues (e.g. winning a lucky draw)
(5) International or local Movies,talks, and dramas [often from Life section]
(6) Advertisments or upcoming weekend talks
(7) Duplicates

Final Sample Size: 7848

Data-Preprocessing

WIP

Misc

Extracting from Nexis University database

Before extracting from Factiva, I had extracted data from Nexis. But I switched to Factiva because Factiva offered articles up to as old as 1989 and included articles from “The New Paper” source. In Nexis, the files are downloaded in docx files with each file containing 100 articles. Unlike Factiva which requires frequent CAPCHA authentication, Nexis database did not require. However, Nexis displays 10 articles per web page and multiple clicking (10 clickings) are needed to collect up to 100 articles for each saving. In Factiva, 100 articles are displayed in the website. Thus, although Factiva requires more clicking (because of the CAPCHA authentication), downloading from FACTIVA is much faster than in Lexis.

Extracting data (e.g., title, main body, source, geographic etc) from Lexis’s docx files requires some coding using stringr commands based on patterns in each article. My codes for extracting data is not provided here but can be requested from me.

Task List

Task List

☒ Data-collection
☒ Training model
☒ Cleaning data
☐ Descriptives
☐ Analysis