Scraping tv2 noter

We have decided to scrape tv2 for articles that result when we search for the word ”sygeplejersker”. We noticed in the url, that when we go to next page, the url changes with the page destination. We utilized this by creating a for-loop with an f-string to fill in the url with the page number. We saw that there were 10700 results for ”sygeplejersker” and therefore our for loop went to 1070.

Unfortunately, every page over page 10 resulted in no articles. We discovered that tv2 will only make the 10 first pages available even if it shows that there are more. We were not satisfied with only 100 articles. We discovered that if we included the date in our search, like ”sygeplejersker 2018-05” the search engine would return articles written around that time.

We utilized this by creating a-for loop generating links for the 10 first pages of the search for sygepljejersker and year and month. Eg. ”sygeplejersker 2018-05” page 1, sygeplejersker 2018-05” page 2.

For each of these links, we used a for-loop to locate the links to the articles, and append them to a list. From this, we were left with a list of links to 8300 articles. It’s these articles that we wanted to scrape

To download the html for each of the articles we created a for-loop using the re module to get the html. From there we created a beautifulsoup item and searched for each of the elements we thought relevant. These elements were; the title, the sub title, paragraph headers, the content of the paragraphs, the date, the author, and finally a tag for news topic eg. ”udland” and ”business”.

We created a list for each of these elements with 1 row being 1 article. We made sure to append an empty row if an article didn’t return anything for a particular element

We then joined the lists in a dataframe,