**Installation**

The data from snopes, emergent and politifact was scraped using the python library, [Newspape](http://newspaper.readthedocs.io/en/latest/)r. It only works with python3 and does not have any prerequisites.

**Crawling**

Iterate over the CSV and download each article link in the page\_url column, parse the link returned and extract the required text. The exception handling is done so that the scraping doesn’t stop if a server or 404 error is encountered.

**Problems**

Some of the problems I faced while working were:

While saving to csv, if the text of the article was too long, it overflowed to the other cells and became a little hard to read. This happened because excel has a limit of 3200 words per cell. For this purpose, I had to remove the extra spaces and lines in the text and add a check to mark a flag if the text got too long.

The library scraped data even from those websites that did not have article data, for example: reddit, facebook pages, twitter pages, etc. I had to add a check in the code to prevent crawling from these sources.