I am looking to scrape the blog articles listed on this site:

http://skytides.com/research

The script should take three input parameters: a start date, an end date, and the output directory where files should be written. It should pull down all of the articles listed on the page that were published between the start and end date. For each article, it should pull down the report’s raw HTML file, and write an extra “metadata” file containing some additional information about the report.

The “metadata” file for each report should be a csv file (with headers). The columns of the csv file should be as follows:

* Report title
* Report link
* Publish Date – taken from the article page
* Tickers – search the article for things like (XXXX), where everything in the parathesis is capitalized, and report back all of the matches here. For instance, on the top article, this would be something like “PODD,MDT, DXCM,TNDM, …”. There will likely be several tickers for every article.

When you write the files to the output directory, put them in a subdirectory corresponding to when the scraper was run. For instance, if the scraper ran at 8:30am on 5/17/2017, then the downloaded reports should go in:

[output\_dir]/2017/05/17/083000/

* insulet-president-provided-critical-misinformation-misled-investors.html
* insulet-president-provided-critical-misinformation-misled-investors.metadata.csv
* …

Do not worry about duplicate output on subsequent reruns – if the script it run twice, it should generate an additional copy of all of the reports in the output folder, timestamped to when it was run.

The script should be implemented in python 2.7, using standard libraries like requests, lxml, beautiful soup, etc. Please use argparse for the argument parsing. If you are going to use any more involved libraries, please let me know ahead of time – it is probably ok, I just need to make sure it will work in my environment. And feel free to ask any questions to help clarify things.