*# You need to take the following steps to get prepared for running the code*```  
# STEP 1: install python3.7 and python3.7-dev  
sudo apt-get install python3.7-dev  
  
# STEP 2: install ML packages  
pip3 install pycld2  
pip3 install pyicu  
pip3 install spacy  
  
# STEP 3: install other dependencies  
pip3 install -r requirements.txt  
  
# STEP 4: install spacy packages  
YOUR\_INTERPRETER\_PATH -m spacy download en  
# i.e.  
~/.pyenv/versions/3.7.3/bin/python -m spacy download en  
  
# STEP 5: install nltk packages  
import nltk  
nltk.download('vader\_lexicon')  
nltk.download('punkt')  
  
# STEP 6: install polyglot packages  
polyglot download embeddings2.en  
polyglot download ner2.en  
```  
  
# install more dependencies  
docker container start CONTAINER\_ID  
docker exec -it news\_scraper\_update:latest /bin/bash  
python3 -m spacy download   
  
python3  
import nltk  
nltk.download('vader\_lexicon')  
exit()  
```  
  
```  
pip3 freeze > requirements.txt  
```  
  
```  
# run the scraper  
scrapy crawl nlp\_news  
```

#it may take some time to scrape all the news

We applied a ‘Scrapy’ frame work to run the news scraper.

All the news that we scraped is from the website newsapi.org, which provides easy APIs for users to get large bulk of wanted company news with many parameters, including ticker and time. In our case, we tried to get the news for the stocks that are stored in the file ‘companies.xlsx’ that happened in the past two hours from now.

In ‘nlp\_news.py’, we defined the spider class with parsing methods to extract and parse the need information from the raw news. The ‘parse’ and ‘parse\_article’ function are the core of our parsing procedure. By using pre-trained sentiment analysis model and tagging model, we are able to get information such as sentiment, named entities, locations, genres that can give the users a really easy way to understand what the news is about. All the parsed attributes are then stored in the ‘item’ object and pass into the ‘pipelines.py’. In ‘pipelines.py’, we assign a tag telling whether the news is relevant to the ticker or not, and then store all the information to an Excel file directly.