

COMP 370 Homework 6 – Data Collection

Assigned Oct 17, 2023
Due Oct 23, 2023 @ 11:59 PM

In this homework, we'll work on doing some data collection using web APIs, along the lines of what we saw in class. We'll be using the News API available at newsapi.org.

Let's consider that we're working on a project called "newscover" where we're looking at the coverage that different topics get in the news.

✓ Task 1: Setup access to News API

You'll need to create a developer account with newsapi.org (it won't cost anything). This will give you an API key that you can use to call the API. This will give you 100 API calls per day – this should be plenty for this assignment, though you should be economical and get started on the assignment early – just in case you run out of API calls on any given day.

✓ Task 2: Build your newsapi utility

In your project, create a top-level python package called `newscover`. Inside that create the `newsapi.py` module. In this, create a function:

```
fetch_latest_news(api_key, news_keywords, lookback_days=10)
```

which queries the NewsAPI and returns a python list of english news articles (represented as dictionaries) containing those news keywords and published within the last `<lookback_days>`.

Test 3: Write unit tests

Write three unit tests (i.e., one `TestCase` class, two functions) that test your `NewsAPI` module. Put them in the `newscover.tests.newsapi` module.

- ✓ One test should ensure that `fetch_latest_news` fails when no `news_keywords` are provided.
- Another test should ensure that when `lookback_days` is set, it doesn't produce articles outside that timeframe
- The last test should ensure that `fetch_latest_news` fails when a keyword contains a non-alphabetic character.

✓ Task 4: Write a data collection tool

Write a CLI tool sitting in the `newscover.collector` that has the following behavior

```
Python -m newscover.collector -k <api_key> [-b <# days to lookback>] -i <input_file> -o <output_dir>
```

The input file is a json file containing a dictionary of named keyword lists. Like this

```
{ "trump_fiasco": ["trump", "trial"], "swift": ["taylor", "swift", "movie"] }
```

For each keyword set with name `N` and keyword list `X`, the collector will execute a query for the keywords `X` and write the results to the `<output_dir>/N.json`.

Submission Instructions

- Your `newscover` package directory and all `*.py` files