

EDGAR-CRAWLER: Automating Financial Data Harvesting and Preprocessing for NLP



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Lefteris Loukas | Manos Fergadiotis | Ilias Stogiannidis | Prodromos Malakasiotis

💰 What's the problem? 🤔

Most **NLP datasets** are often behind paywalls. **EDGAR**, a notable free source, hosts comprehensive annual reports (10-K reports) from US publicly traded companies. However, these reports, stored as **complex PDF/HTML/TXT files** with numerous sections and pages, pose **challenges for researchers**. Extracting specific information becomes laborious, requiring downloading a vast number of reports for manual text extraction. This is an **impractical and time-consuming** process.

💰 Our Solution:

EDGAR-CRAWLER, a free, open-source package that downloads and extracts information from annual reports (EDGAR 10-K documents) into an easy-to-manage JSON format.

Our software, **EDGAR-CRAWLER**, is made up of two modules:

```
1. python edgar_crawler.py
2. python extract_items.py
```

1. Responsible for crawling and downloading financial reports.
Supports multiple input arguments.

2. Cleans and extracts the text of all or particular items from downloaded 10-K reports and saves them as JSON files.


💰 Scientific Contributions in ML & NLP:


- Trusted by the community (**160+ stars** on Github!)
- **Multiple citations** in relevant literature.




💰 Future Work:

Looking for contributors for these, send us a message if interested 😊


Support more types of documents like quarterly reports.


Create a GUI for more user-friendly configuration.


Deploy a live demo to increase accessibility.



edgar-crawler
Turn unstructured financial documents into clean JSON files.

