User Manual
sec-10-k-scraper

Creators: Ed Carl, Chloe Dorward, Seth Markarian, Sena Yevenyo

An up-to-date version of this document can always be found on Google Docs.
A video demonstration of the app can be found here.

Table of Contents

Summary

Purpose of Application

This app is a local desktop application for Windows, MacOS, and Linux that obtains companies' 10-K, 10-Q, and 10-F SEC filings, downloads the filings as HTML files, extracts the text of key sections of those documents, analyzes the text for named entities, and gathers that information in a Microsoft Excel spreadsheet. Users can either search for filings based on company names/ CIK numbers, filing type, and date range, or upload a file including similar parameters for batch uploading.

Using a desktop application rather than a web service minimizes the technical and financial requirements for FracTracker to operate it. The application is entirely self-contained on users' computers and there will be no need to spend time on configuring it or money on hosting a server. A user only needs to have an internet connection in order to run the application.

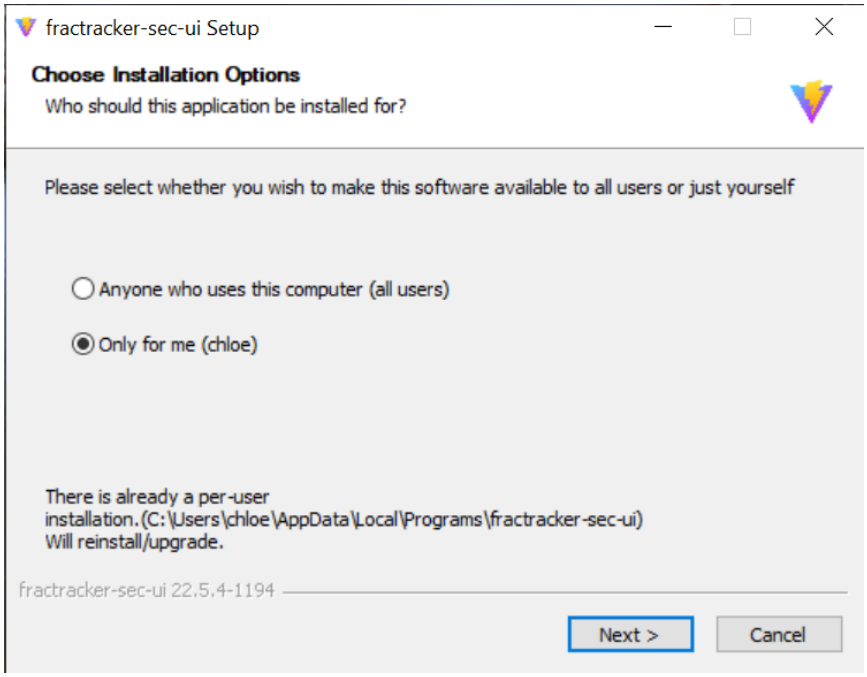
The App

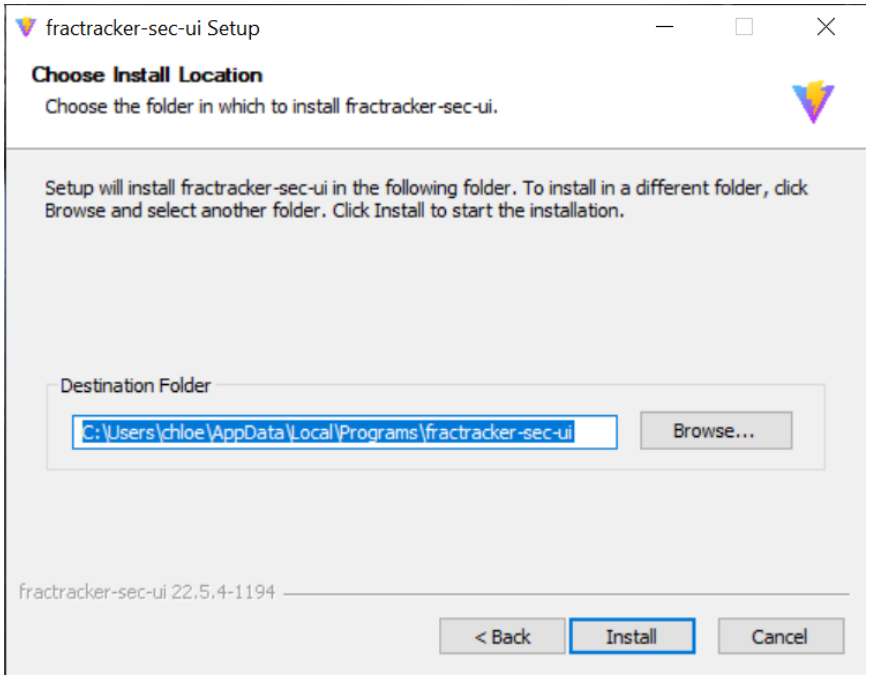
Installing the App

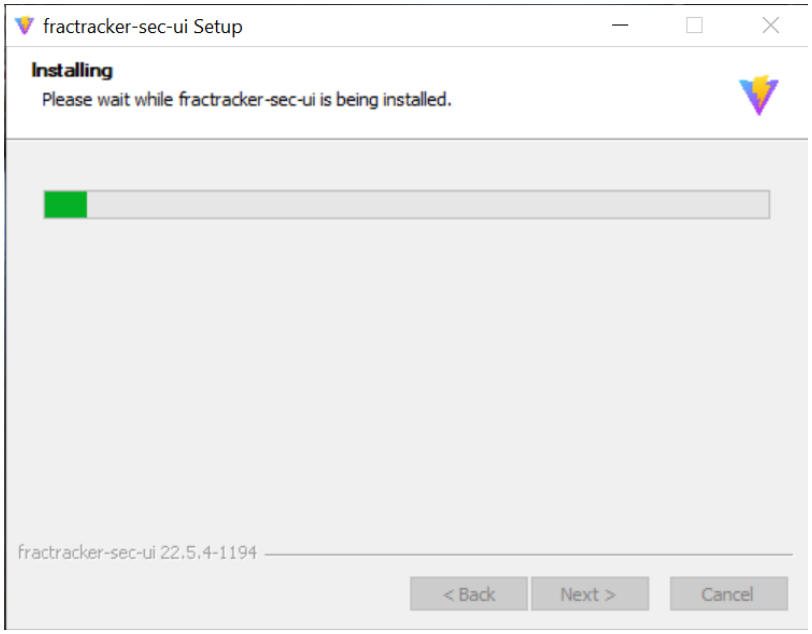
Releases of the application installer can be found under the releases page of the project's github repository. github.com/edwardmcarl/sec-10-k-scraper/releases

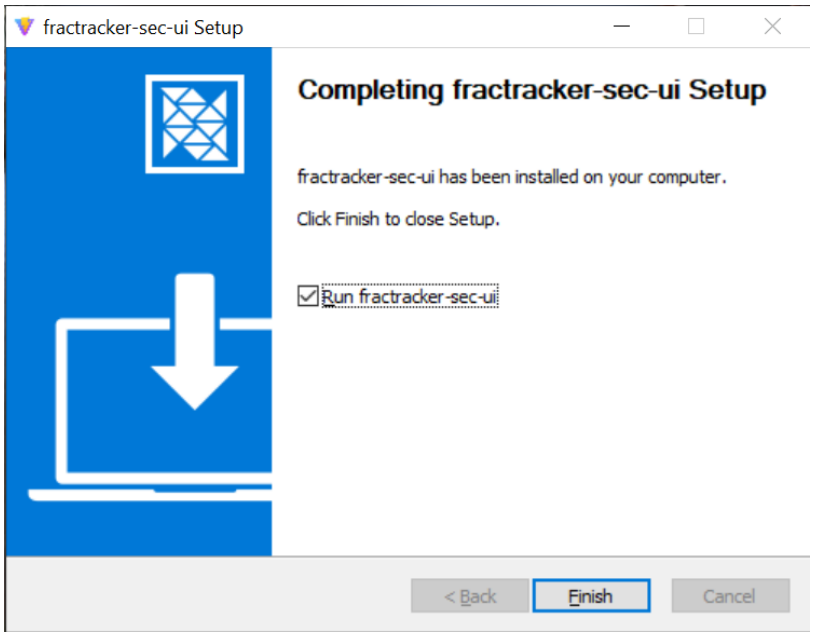
Windows

Find the installer in your file explorer (fractracker-sec-ui Setup 22.5.4-1194.exe), and run it. Navigate through the prompts to install.



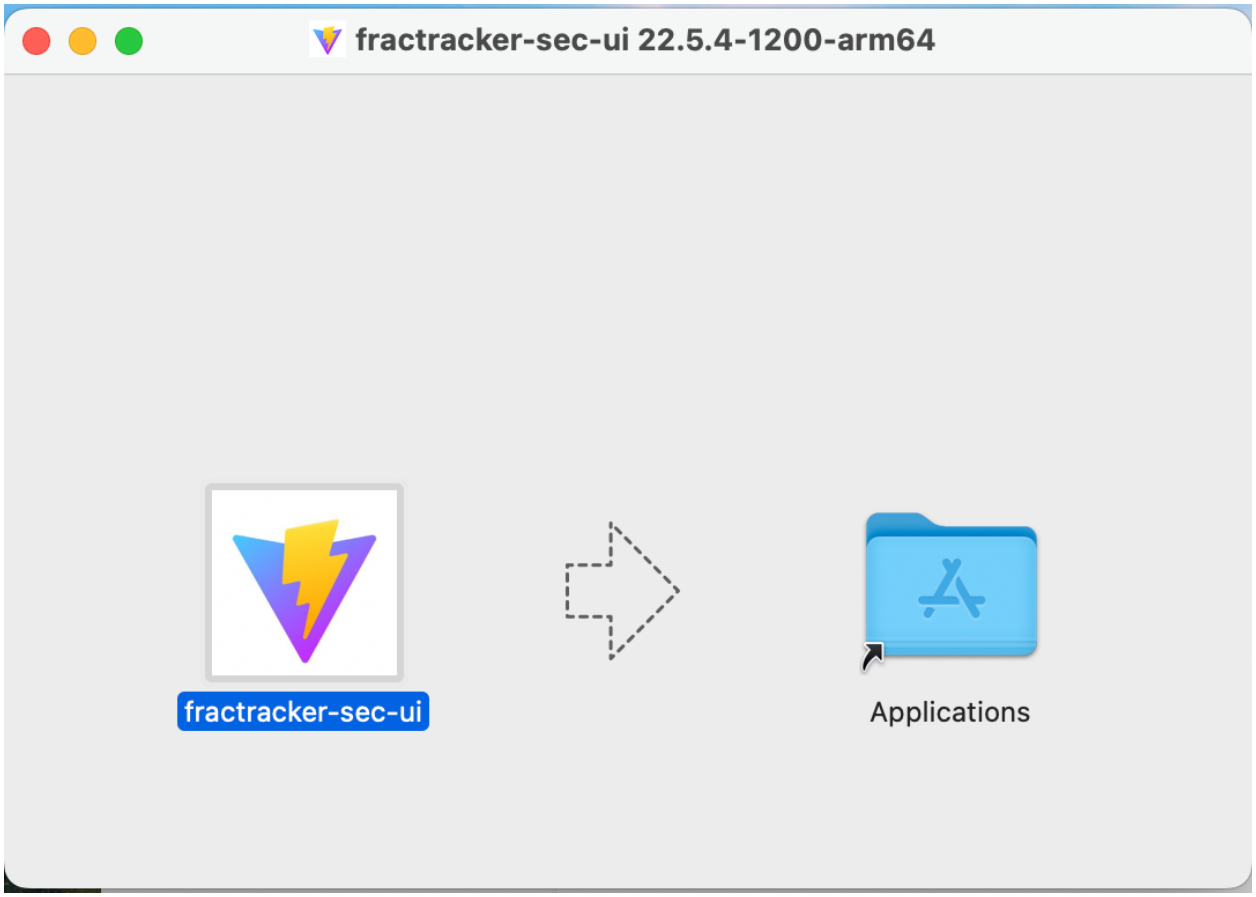






An icon will appear on your desktop.

Mac



Simply mount the installer .dmg and drag the app into the Applications folder.

Linux

Just double-click the installer AppImage.

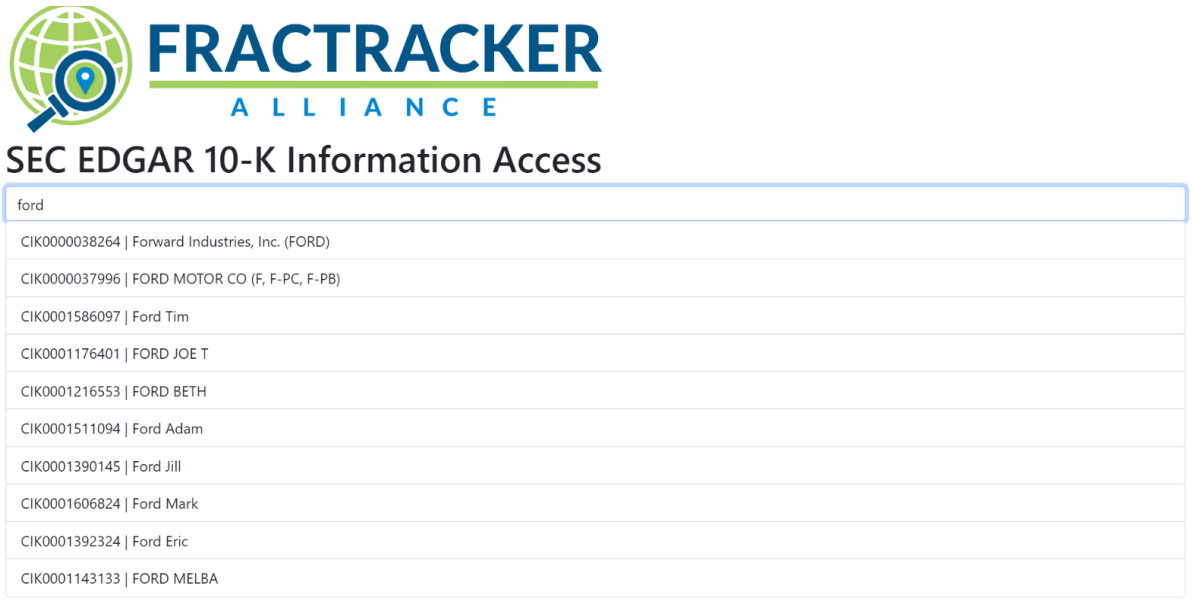
Running The App

Step 1: Selecting Filings

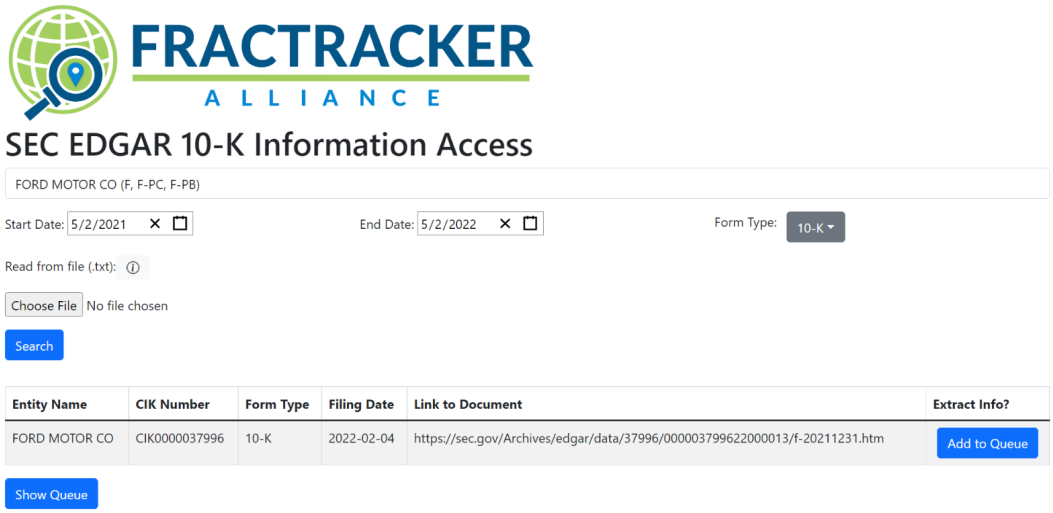
Option 1: Searching

Searching

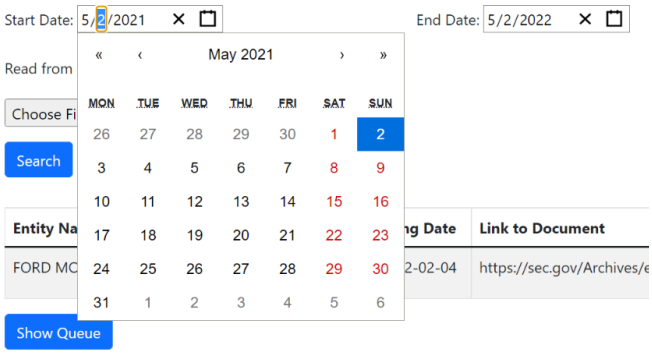
The first option is to select an entity, and search for filings based on a date range.



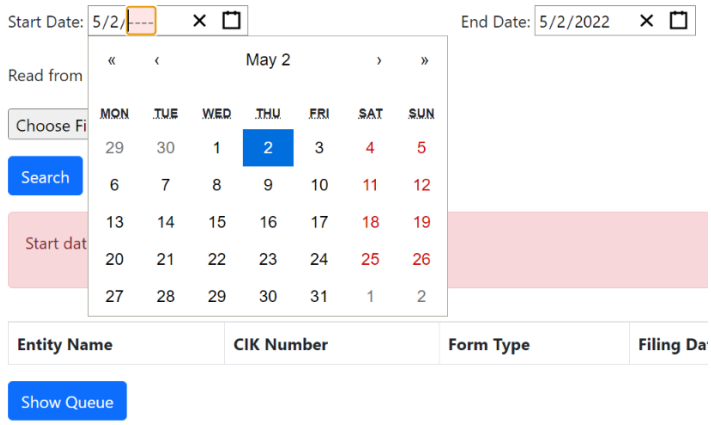
Click on the search bar, and begin to input the appropriate entity name or CIK number. A list of predicted entities will appear in a list below the bar.



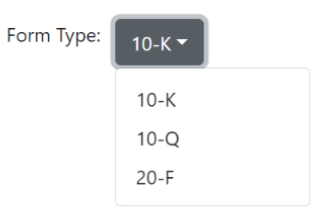
Select an entity by clicking on it (for example, FORD MOTOR CO). This will populate the results table with the default search parameters - 10-K filings from the past year.



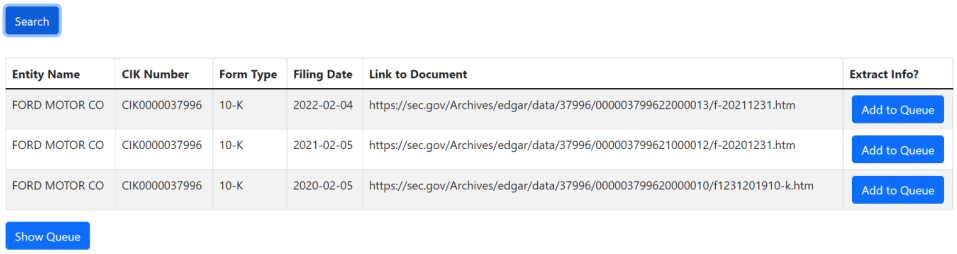
You can adjust your date range by clicking on the date selector, and changing the date accordingly.



You can also type in the date rather than using the provided selector.

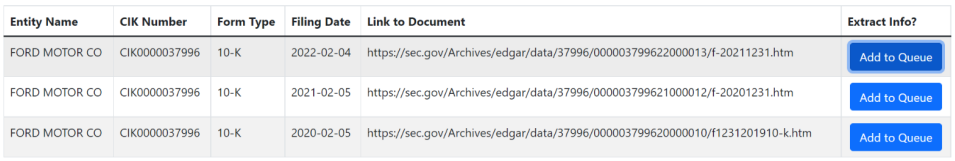


You can change the selected form type by clicking on the gray dropdown on the right hand side, and clicking on the appropriate form.

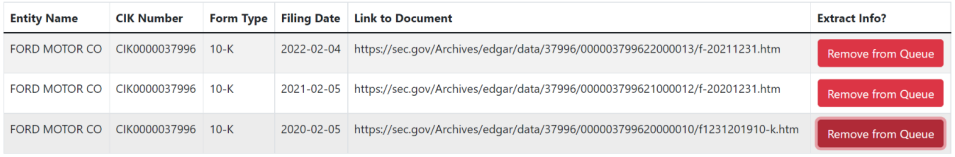


Clicking on the blue Search button will repopulate the results table based on your selected parameters. Changing the date, form type, or search components also repopulates the results.

Selecting Filings



Individual filings from different years can be selected by clicking on the Add to Queue button.



Individual filings can be removed by clicking on the Remove from Queue button.
Repeat the process outlined above to add filings from different entities to the same Queue.

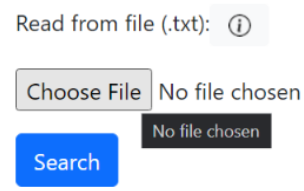
Option 2: Uploading

The second option is to upload a text file with different CIK numbers and appropriate date ranges. This option will add every filing within a CIK's date range to the Queue.

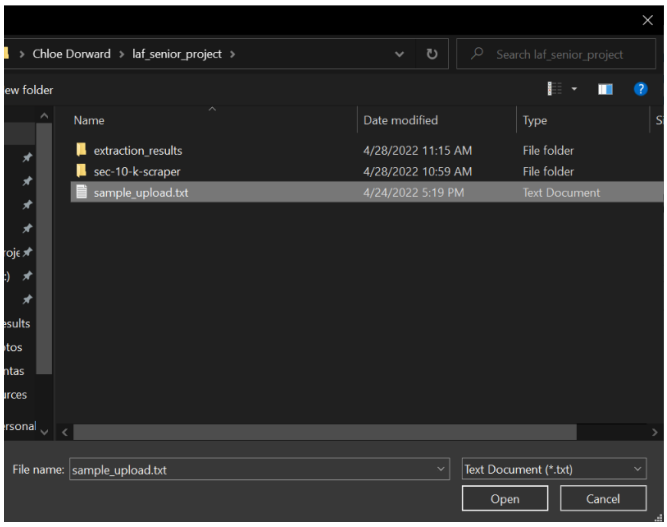
File Format

The file must be a .txt file. Your computer will have a program to create text files, for example, Notepad, TextEdit, etc.
Each line represents a new entity, and should contain the CIK number, the start of the date range, and the end of the date range:
`CIKXXXXXXXXXX YYYY-MM-DD YYYY-MM-DD where X is 0-9 inclusive.`

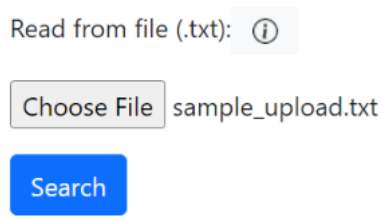
Uploading the File



Click on the Choose File button.

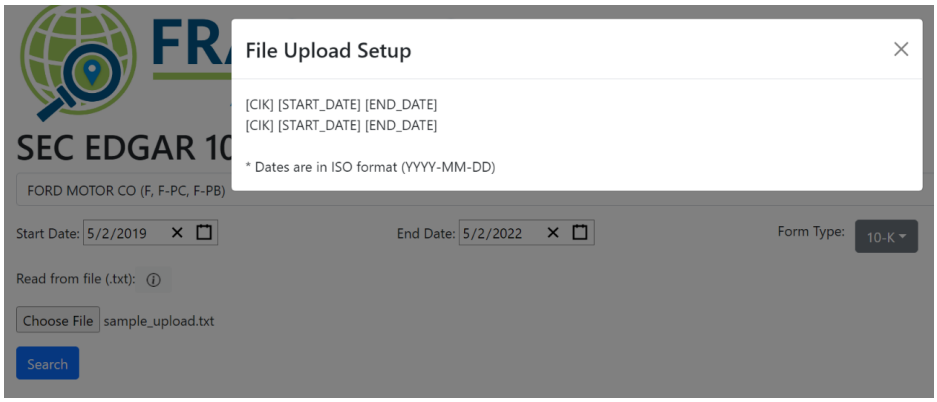


Navigate through your file system to select your file.



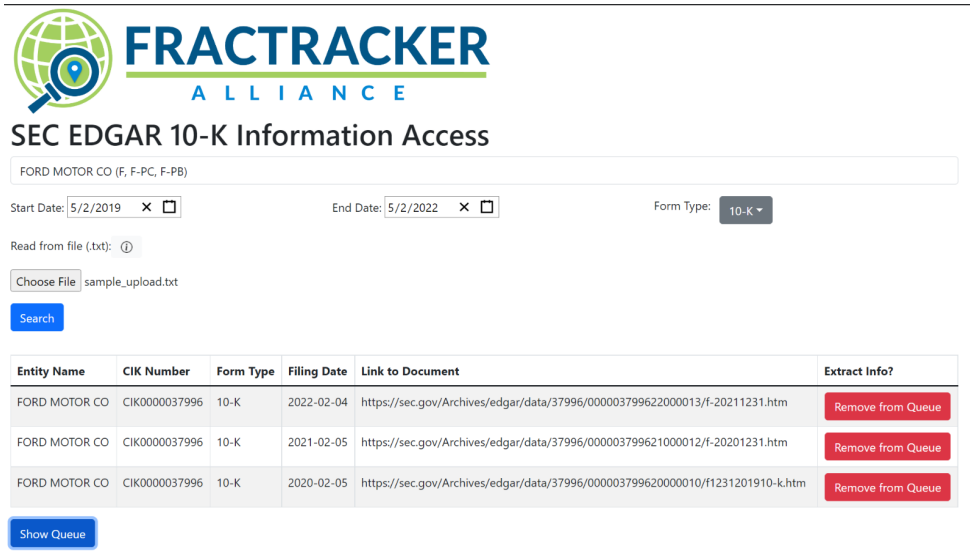
The uploaded file name will appear next to the Choose File button.

This option can be used in combination with Option 1. Any results that have already been selected through the File Upload option will appear as such in the search results table.

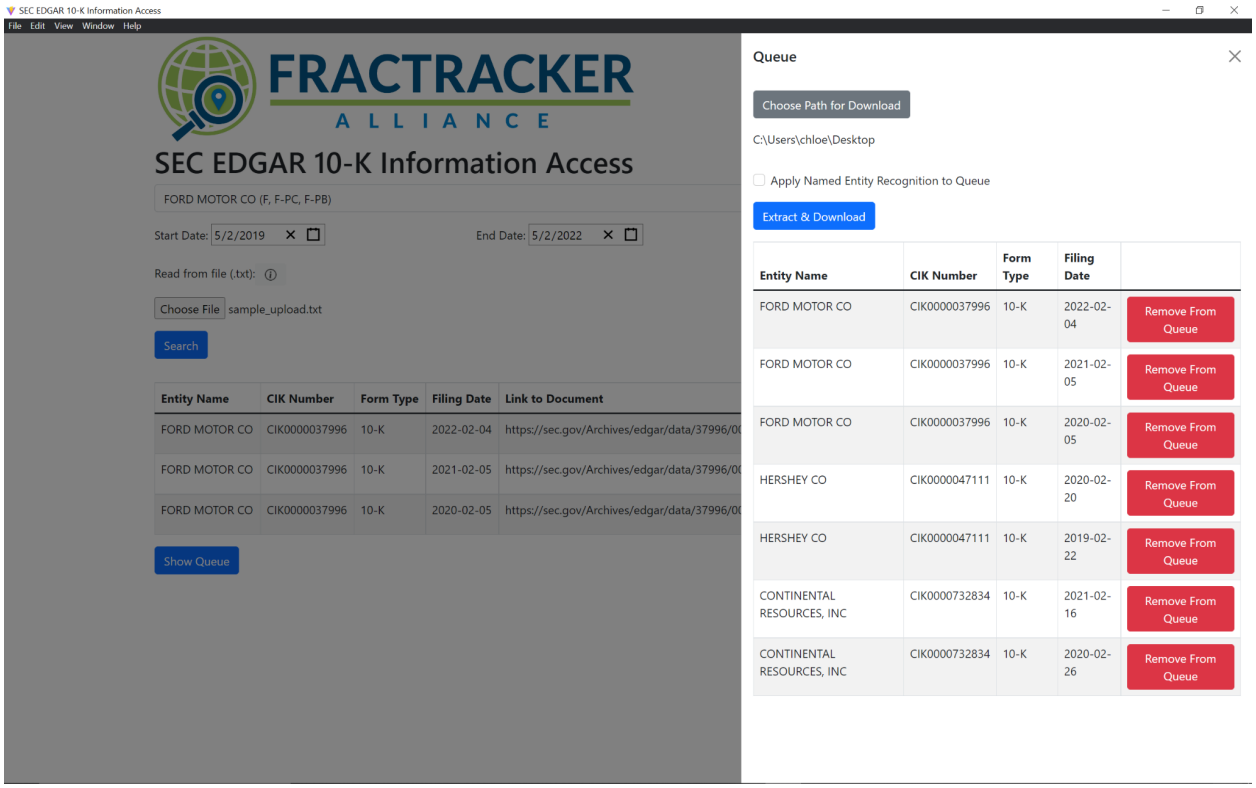


If you forget the format for file uploading, it can be found by clicking on the i button.

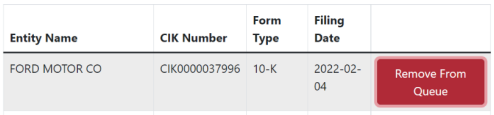
Step 2: Viewing Queue



View the Queue by clicking on the blue Show Queue button on the bottom of the page.

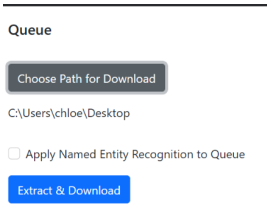


This will slide open the Queue tab.

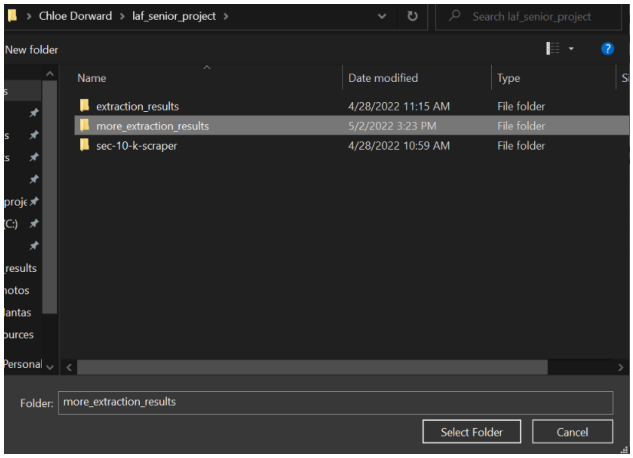


From here, you can remove filings from the queue.

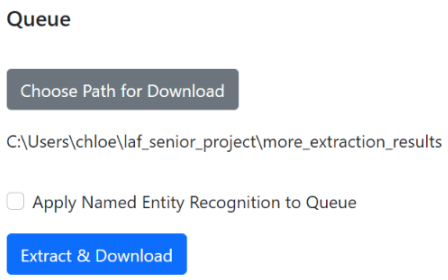
Step 3: Downloading Files



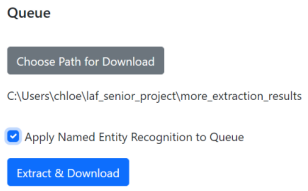
At the top of the Queue tab, click the Choose Path button. This allows you to select a place to save your extracted files. The default path is your Desktop folder.



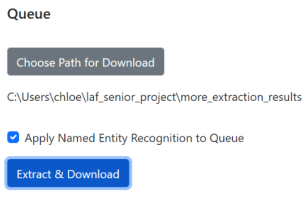
Navigate through the file dialog, and select a folder to save the files to.



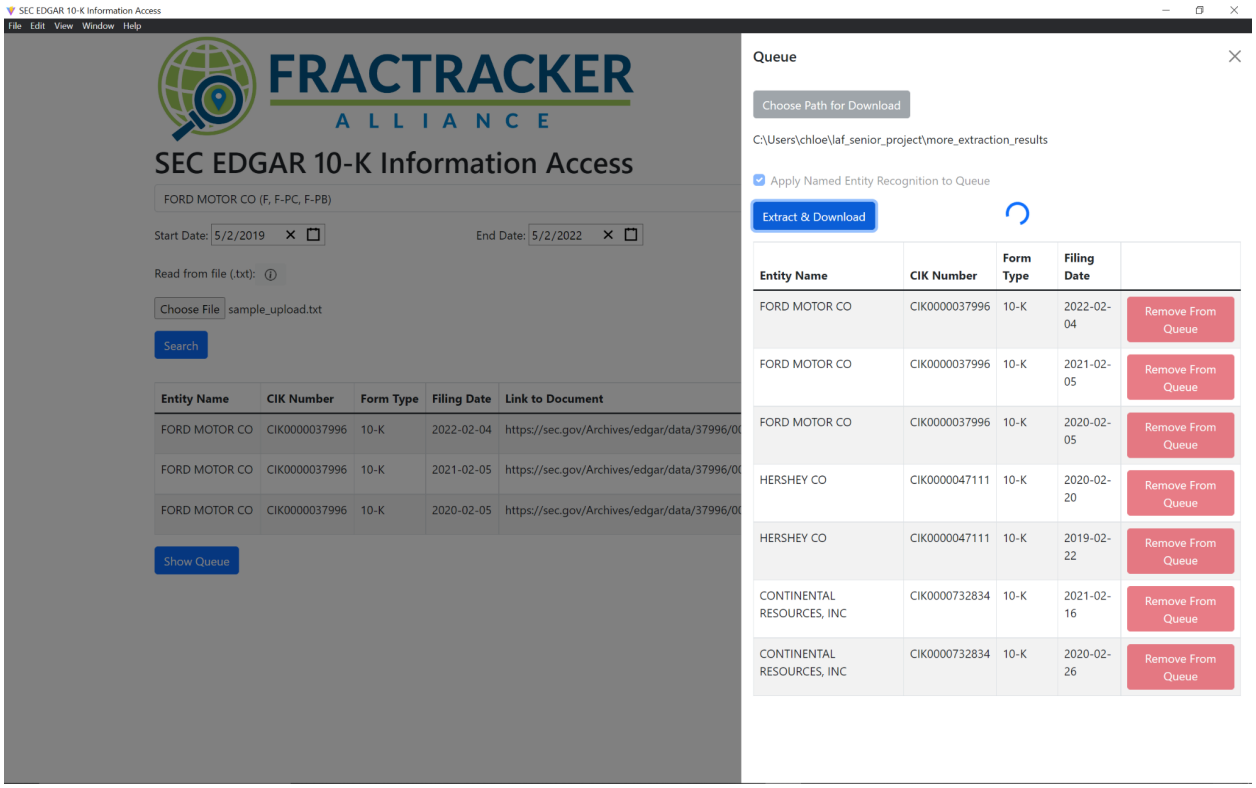
Once a location is selected, the name of the file path will appear below the Choose File button.



Next, you have the option to either apply Named Entity Recognition during the file extraction process, or not. You can select the checkbox to do this.



Finally, press the Extract & Download button.

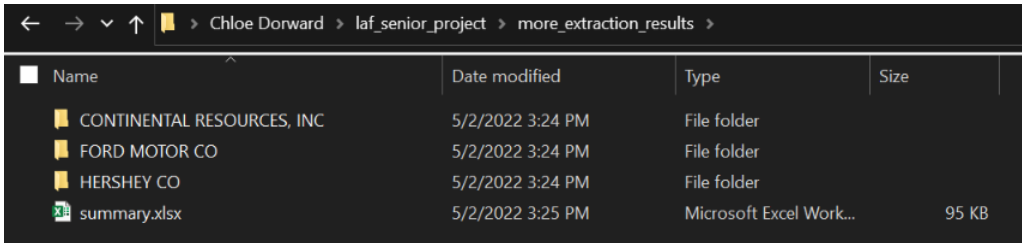


While the system is extracting the files, it will indicate that it is loading. The buttons next to the selected filings will gray out, meaning that they are processing. You cannot deselect an entity once the extraction process has started.
Once the extraction process is complete, the loading spinner will disappear, and the selected filings will no longer be grayed out.

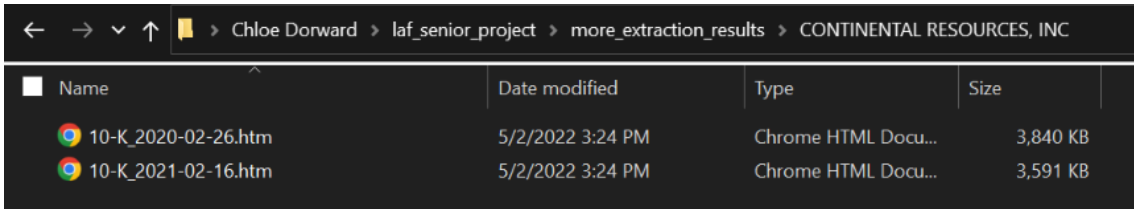
Step 4: Finding/Using Your Files

File Structure

Files will be downloaded to the folder that you selected in Step 3.

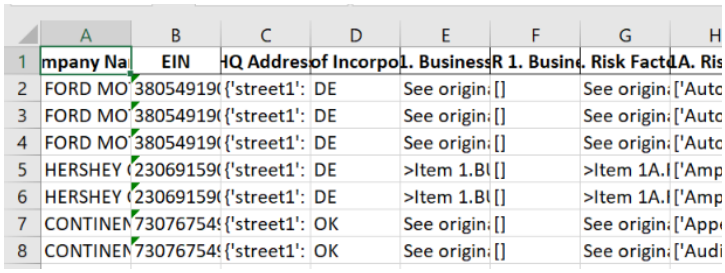


Within your selected folder, there will be a folder for each entity, and an excel spreadsheet.



Each entity folder will contain htm files for each filing. The name of the file will be the filing type and the filing date.

Excel File



	A	B	C	D	E	F	G	H
1	mpany Na	EIN	HQ Addres	of Incorpo	1. Business	R 1. Busine	Risk Facto	1A. Ris
2	FORD MO	38054919(	{'street1':	DE	See origin:	[]	See origin:	['Autc
3	FORD MO	38054919(	{'street1':	DE	See origin:	[]	See origin:	['Autc
4	FORD MO	38054919(	{'street1':	DE	See origin:	[]	See origin:	['Autc
5	HERSHEY (	23069159(	{'street1':	DE	>Item 1.Bl	[]	>Item 1A.I	['Amp
6	HERSHEY (	23069159(	{'street1':	DE	>Item 1.Bl	[]	>Item 1A.I	['Amp
7	CONTINEN	73076754(	{'street1':	OK	See origin:	[]	See origin:	['App
8	CONTINEN	73076754(	{'street1':	OK	See origin:	[]	See origin:	['Audi

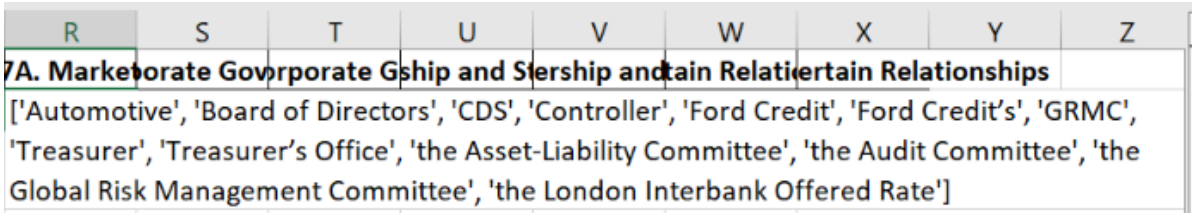
Within the summary.xlsx excel file, each new filing will add a new row to the spreadsheet. If you choose to download your files to a folder which already has the summary.xlsx file, all additional filings will be appended to the end of the spreadsheet; our tool does not look through to check if there are duplicates.

The Columns in the spreadsheet are:

- Company Name
- EIN
- HQ Address
- State of Incorporation
- 1. Business
- NER 1. Business
- 1A. Risk Factors
- NER 1A. Risk Factors
- 2. Properties
- NER 2. Properties
- 3. Legal Proceedings
- NER 3. Legal Proceedings
- 6. [Reserved]
- NER 6. [Reserved]
- 7. Financial Condition
- NER 7. Financial Condition
- 7A. Market Risk
- NER 7A. Market Risk
- 10. Corporate Governance
- NER 10. Corporate Governance
- 12. Security Ownership and Stockholder Matters
- NER 12. Security Ownership and Stockholder Matters
- 13. Certain Relationships
- NER 13. Certain Relationships

Cells pertaining to entity information should always be filled in.

Cells pertaining to subsections of the filing will contain the text within that section if possible. If the number of characters in a section happens to be too long for a single Excel cell (32,767 characters) the statement ‘See original document; Section too long for Excel’ will be placed in the cell.



Cells pertaining to Named Entity Recognition will contain a bracketed, comma-separated list of named entities. If you did not select the Named Entity Recognition checkbox in Step 3, cells under the NER columns will be blank.

Errors

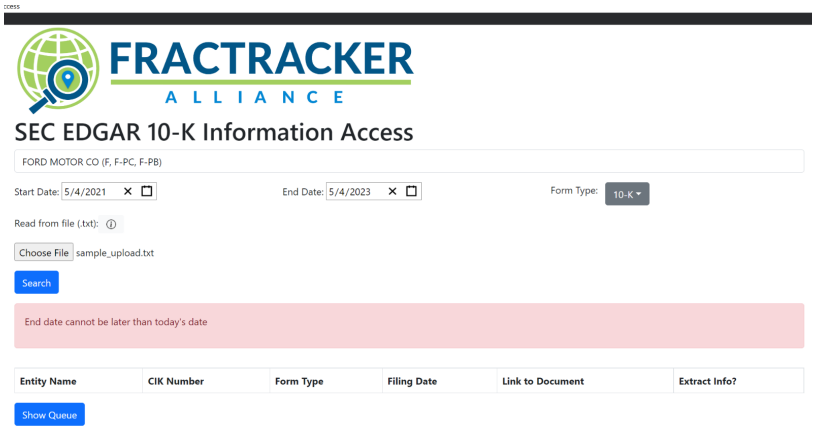
Error Dialog

There are two error dialogs in the application:

- Search Error Dialog
- Queue Error Dialog

Search Error Dialog

The search error dialog box is found in the main page of the application. This light red box with error text in deep red pops up during search for companies’ documents that are to be added to the queue for extraction. An image of the error box is found below:



Explanation of the errors of the application are listed below:

- **The SEC EDGAR server could not process the request**
    - This means that the SEC server could not process your search query. The server might be down, blocking all requests or there have been too many requests to the server (occasionally happens). Simply try again.
- **The application failed to reach the server**
    - This means that there might be no internet connection. Check your internet connection.
- **CIK number input not in correct format**
    - Happens often during batch processing. This means that the CIK number inputted is not in the structure stated under Option 2: File Format under Step 1: Selecting Files.
- **Start date input not in ISO format**
    - Happens often during batch processing. The start date given in the file is not in ISO 8601 format. ISO format for the dates are: `YYYY-MM-DD`
- **End date input not in ISO format**
    - Happens often during batch processing. The end date given in the file is not in ISO 8601 format. ISO format for the dates are: `YYYY-MM-DD`
- **Start date cannot be earlier than 1994-01-01**
    - This is the earliest date of records in the SEC EDGAR database server
- **End date cannot be later than today's date**
    - A user cannot query documents from future dates
- **Start date cannot be later than end date**
    - This error means exactly what it is. Please adjust the start state to be earlier than the end date
- **Something occured when decompressing and/or decoding response from SEC EDGAR server**
    - Rarely would this happen. If it does there is a change in the way the server is compressing data. Please reach out to Lafayette College Computer Science Department if this happens.
- **The CIK number input does not exist in SEC EDGAR database**
    - Happens often during batch processing. The CIK number is in a correct format but does not exist in the SEC server
- **The form type inputted is not supported by this application**
    - Rarely happens. This means that the form input is no longer supported by the SEC server for querying.

Note that when you submit a text file for batch processing to add companies filing to the queue, if there are errors with processing the text file, any error that occurs when processing each line of the text file will be in the search dialog box. An example of the error would be like:

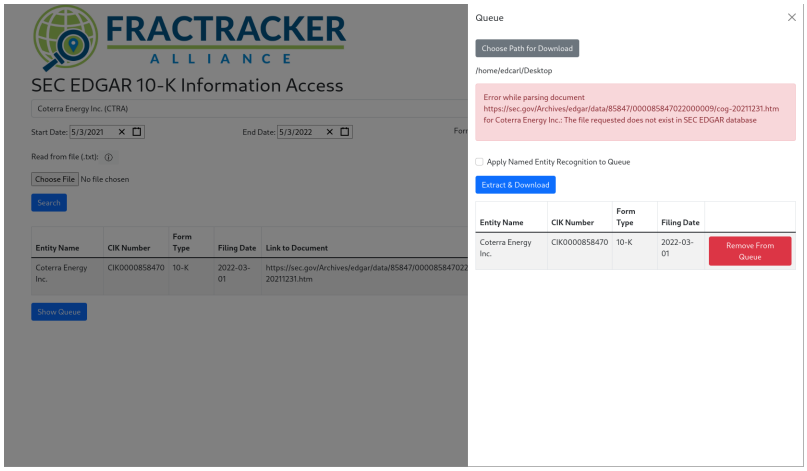
Line 14: CIK number input not in correct format*

*given that Line 14 in your text file for batch processing/file upload does not have the correct CIK number

If multiple errors are encountered, they are all placed in the error box.

Queue Error Dialog Box

The queue error dialog box is found in the side draw of the application which is shown when the user clicks the ‘Show Queue’ button. This error message pops up below the text of the selected path as shown below:



A list of possible errors are in the possible structure:

Error while *{stage where error occured}* document *{document URL}* for *{company}*: *{Error message}*

Error sections explained below:

{stage where error occured}
The stage where error has occurred possible messages are:

- parsing
    - This means that the error happened during parsing of the document for the itemized sections
- downloading
    - This means that the error occurred when downloading data from the SEC server
- applying NER to
    - This means that the error occurred when applying named entity recognition to the extracted itemized fields
- adding spreadsheet row for
    - This means the error occurred when appending the new data to the output Excel file

{document URL}
This is just the document URL being handled for which the error occurred.

{company}
This is the company's name whose document is currently handled.

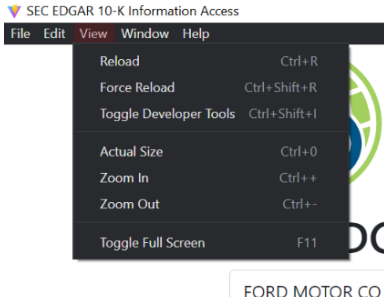
{Error message}
Why was the error encountered. The possible messages are:

- **Parsing for this document is not supported**
  - Rarely happens. This means that the document is not an HTML (htm. or .html) file and thus cannot be parsed.
- **The SEC EDGAR server could not process the request**
  - This means that the SEC server could not process your document query. The server might be down, blocking all requests or there have been too many requests to the server (occasionally happens). Simply try again.
- **The application failed to reach the server**
  - This means that there might be no internet connection. Check your internet connection.
- **Something occured when decompressing and/or decoding response from SEC EDGAR server**
  - Rarely would this happen. If it does there is a change in the way the server is compressing data. Please reach out to Lafayette College Computer Science Department if this happens.
- **The file requested does not exist in SEC EDGAR database**
  - Occasionally happens. This means that the file does not exist in the SEC EDGAR database.

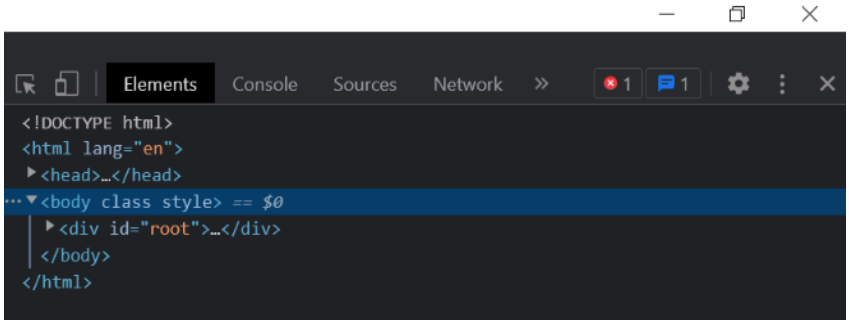
A possible message then is:
Error while downloading document https://example.com/file-does-not-exist.html for Example Inc: The SEC EDGAR server could not process the request.

Developer Tools

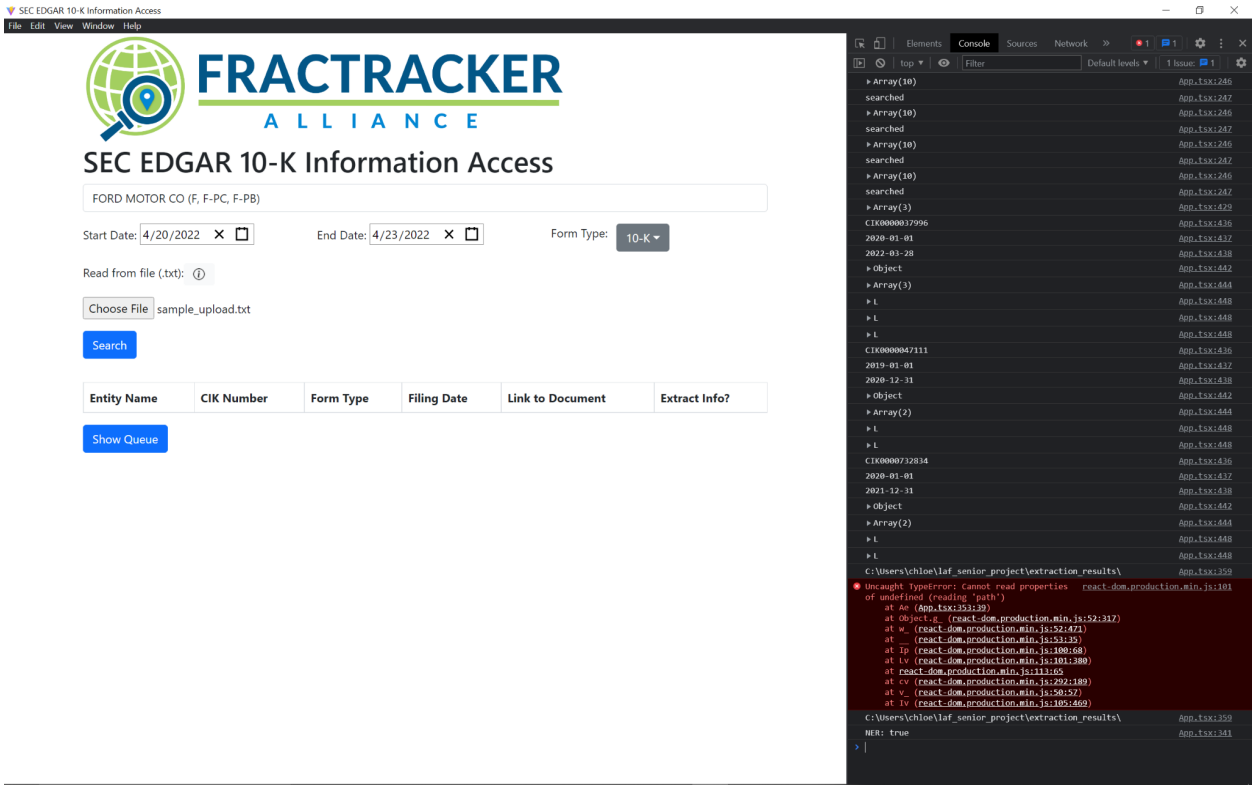
If you are experiencing errors and cannot tell why, you may be able to find out what is happening by viewing the developer tools.



On the top menu, click on the View tab.
Click the Toggle Developer Tools option.



Click the Console tab.



This will display statements that get written out to the application's console, including uncaught error messages.