Advance Data Science

Edgar Data

Assignment – 1

Report by:

Abhinav Tiwari

Nilesh Nerkar

Dharit Shah

**Goal:**

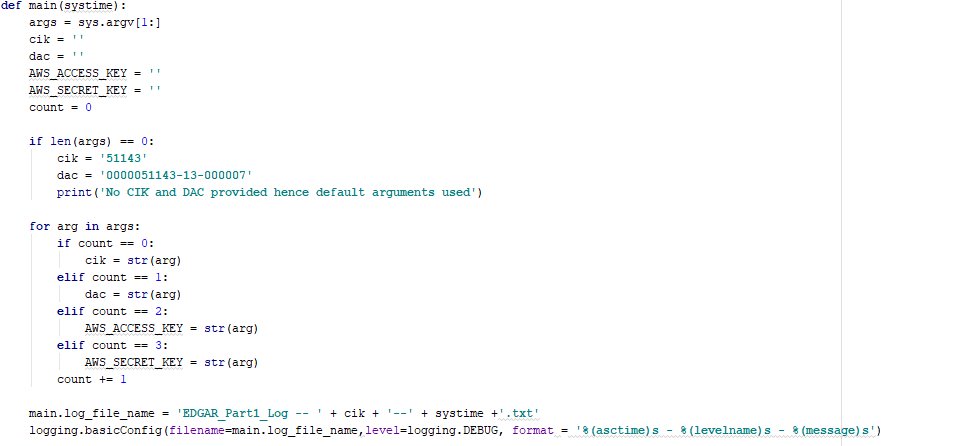
Working with Edgar datasets: Wrangling, Pre-processing and exploratory data analysis

Summery

EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (the "SEC"). The database is freely available to the public via the Internet (Web or FTP).

**Problem 1: Data wrangling Edgar data from text files**

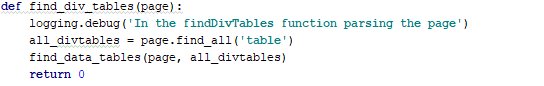
**We are taking default value because no CIK and DAC provided.**



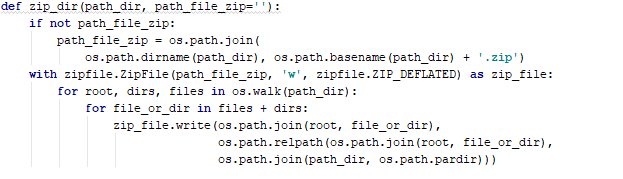
**In the URL we are passing landing page of the CIK provided by user and we are web scrapping and removing <a> and <href> tag using beautifulsoup library.**



We are scrapping table data using <div> tag.

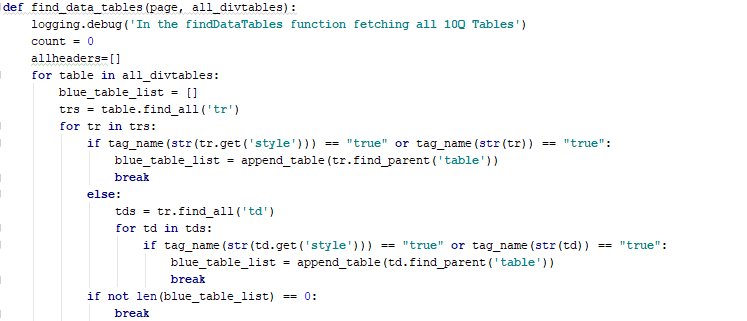


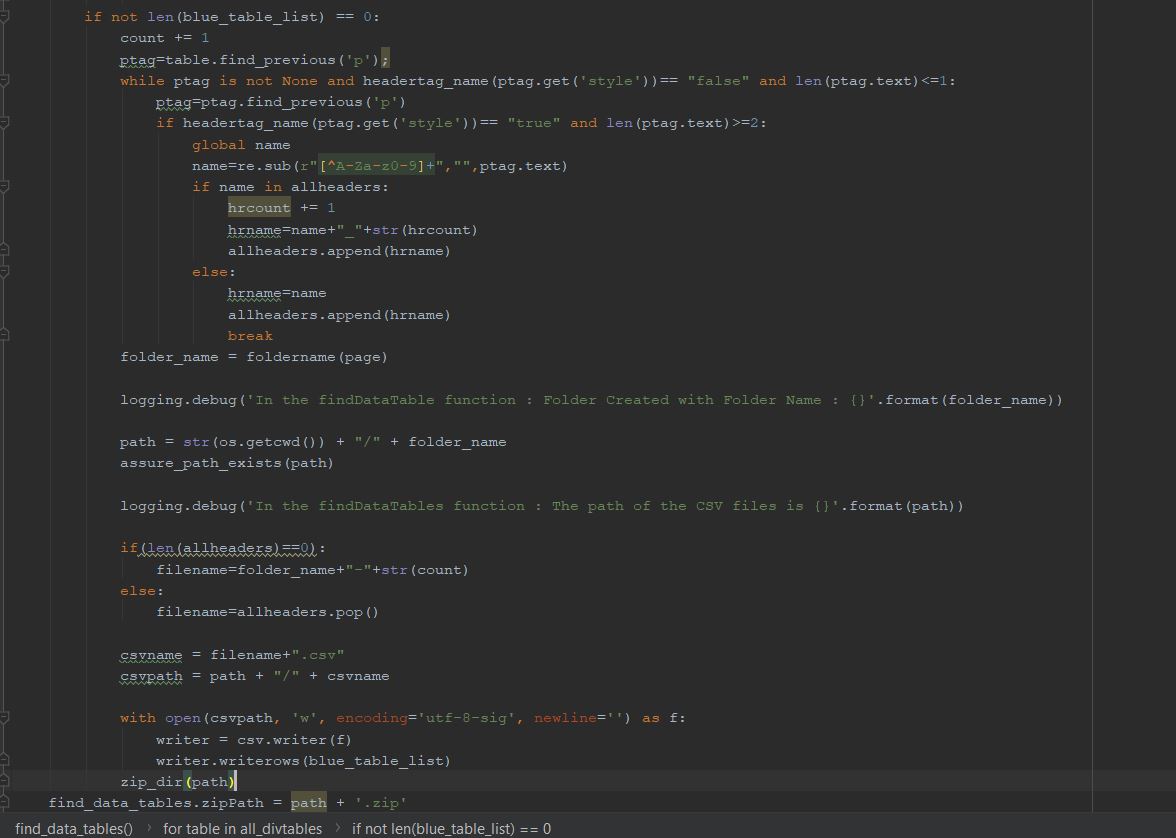
We are zipping a folder which contains all the csv file using zip\_dir function.



Using find\_data\_tables fetching all 10Q tables. We are scrapping using <tr> and <td>

And scrapping data using <style>





PART 2

**Goal:**

The report summarizes the design and implementation of the analysis performed on the Edger

log files. This report is divided into two sections:

**1.** Fetching and Analysis of Edgar logs file Data Set

**2.** Handling Missing Data

**3.** Compute summary metric for Edgar Log files

**4.** Log all the operations (with time stamps) into a log file

**5.** Dockerize the process using Docker file and Docker Hub

**Fetching and analysis of EDGAR log file Data set:**

The EDGAR Log File Data Set contains information in CSV format extracted from

Apache log files that record and store user access statistics for the SEC.gov website. These

log are captured on a daily basis and are stored in a zip format on Edgar website.

Edgar log files consist of the following columns:

Fetching Dataset:

The program takes year as the parameter from the user. It then outputs all the log files from the

1st day of every month. The years range from 2003 to 2017 and if user enters anything not valid,

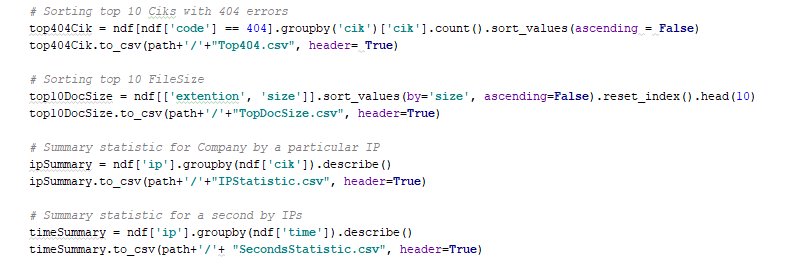
the program as for valid data.

On valid data, the program will fetch the url from Edgar dataset which consists of all the log files.

The log file consists of:

* CIK No: Every company has a CIK No for filing purpose
* Accession No: Every file has an accession no
* Extension: consist of the file that is requested by supplying CIK and accession number
* Code: implies the response code from the server
* Date: log file creation date

**Analysis**



**Missing Value Analysis**



We are parsing the html page to get csv’s of first day of all months for particular year

def websrape(page, year):  
 soup = BeautifulSoup(page,'html.parser')  
 name\_box = soup.findAll('div', attrs={'id':'asyncAccordion'})  
  
 for aTag in name\_box:  
 aTagList = aTag.findAll("a")  
 asd=[]  
 for aTag in aTagList:  
 hrefTagList = aTag.get('href')  
 asd.append("https://www.sec.gov" + hrefTagList)  
 zipFinalListAll = []  
  
 for zipList in asd:  
 if str(year) in zipList:  
 linkhtml = u.urlopen(zipList)  
 allzipfiles = BeautifulSoup(linkhtml, "html.parser")  
 zipListAll = allzipfiles.find\_all('a')  
 zipFinalListAll.append(zipListAll)  
 # else:  
 # print("No data available for " + year + " on edgar")  
  
 z=zipFinalListAll[0]  
 all\_days\_links = []  
  
 for aTag in z:  
 hrefTagList = aTag.get('href')  
 all\_days\_links.append(hrefTagList)  
 first\_day\_of\_month = []  
  
 for i in all\_days\_links:  
 if '01.' in i:  
 first\_day\_of\_month.append(i)  
  
 downloadZipFilesToSystem(first\_day\_of\_month)  
  
def downloadZipFilesToSystem(first\_day\_of\_month):  
 path = str(os.getcwd()) + "\\Downloaded"  
 for first in first\_day\_of\_month:  
 with u.urlopen(first) as zipFirstMonth:  
 with ZipFile(BytesIO(zipFirstMonth.read())) as zipFirstMonthFile:  
 zipFirstMonthFile.extractall(path)  
 getCSVFiles(path)  
  
def getCSVFiles(path):  
 allFiles = glob.glob(path + "/\*.csv")  
 folder\_path = allFiles[0][-12:].split('.')[0][:-4]  
 # list\_ = []  
 for file\_ in allFiles:  
 df0 = pd.read\_csv(file\_,index\_col=None, header=0, low\_memory= False)  
 # list\_.append(df)  
 if df0.empty == False:  
 df1 = change\_dataTypes(df0)  
 df2 = missingValueAnalysis(df1)  
 evaluateFile(str(file\_), df2)  
 else:  
 logging.debug('No data for '+file\_)  
  
 zip\_dir(path+'/'+folder\_path)

**Summery Metrices**

We have created summery metrices which has all 4 csv of given year’s 12 months.

**LINKS**

**GitHub: https://github.com/tiwari-abhi/ADS\_FALL2018/tree/master/Edgar%20Assignment**

**Docker : https://hub.docker.com/r/tiwariabhi/edgar-part2/**

**https://hub.docker.com/r/tiwariabhi/edgar-part1/**

**Reference:**

1. <https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm>

2. https://www.sec.gov/data/edgar-log-file-data-set.html