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

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).
The goal of this assignment is to work with Edgar datasets.

- Deadline: Febraury 16th midnight.
- All submissions are to be done through github
- The repository shouldn't be updated after Febraury 16th midnight or will incur penalty.
- Send the github link to analyticsneu@gmail.com and the TA.

Problem 1: Data wrangling Edgar data from text files (50 points)

Part 1: Parse files
https://datahub.io/dataset/edgar lists how to access data from Edgar. The goal of this exercise is to extract tables from 10Q filings using R/Python

Given a company with CIK (company ID) XXX (omitting leading zeroes) and document accession number YYY (acc-no on search results), programmatically generate the url to get data (http://www.sec.gov/Archives/edgar/data/51143/000005114313000007/0000051143-13-000007-index.html for IBM for example). Parse the file to locate the link to the 10Q file (https://www.sec.gov/Archives/edgar/data/51143/000005114313000007/ibm13q3_10q.htm for the above example). Parse this file to extract "all" tables in this filing and save them as csv files.

Part 2: Dockerize this pipeline
Build a docker image that can automate this task for any CIK and document accession number which could be parameterized in a config file. We should be able to replace IBM's CIK and document accession number with Google's to generate the url https://www.sec.gov/Archives/edgar/data/1288776/000128877615000046/0001288776-15-000046-index.htm and then parse this document to look for the 10Q filing https://www.sec.gov/Archives/edgar/data/1288776/000128877615000046/goog10-qq32015.htm and extract all tables from this filing. The program should log all activities, then zip the tables and upload the log file and the zip file to Amazon S3. Parameterize your cik, accession number and amazon keys so that anyone can put their cik, accession number and amazon keys and locations and reuse your code. (We will do a demo on working with Docker and Amazon S3 next week)

Submission:
Submit the github with the Docker file and source code so that we can rebuild the Docker images. Also register your Docker image on Dockerhub and provide links. Write a report detailing:

- Your design and implementation for both the parts.
- Review your outputs stored on Amazon S3 and discuss outputs. How do you handle exceptions when you don't find the cik/accession number or if the amazon keys aren't valid?

Problem 2:

Missing Data Analysis (50 points)

You are asked to analyze the EDGAR Log File Data Set [https://www.sec.gov/data/edgar-log-file-data-set.html]. The page lists the meta data for the datasets and you are expected to develop a pipeline which does the following. Given a year, your program (In R or Python) should get data for the first day of the month(programmatically generate the url http://www.sec.gov/dera/data/Public-EDGAR-log-file-data/2003/Qtr1/log20030101.zip for Jan 2003 for example) for every month in the year and process the file for the following:

- Handle missing data
- Compute summary metrics (Decide which ones)
- Check for any observable anomalies
- Your program should log all the operations (with time stamps) into a log file.
- Compile all the data and summaries of the 12 files into one file
- Upload this compiled data file and the log file you generated to your Amazon S3 bucket (Google on R/Python packages to use for this.

The code should work for any year on the page. You should create a Docker image which runs the pipeline. Note: Don't put sensitive information like amazon keys in your Docker files. Parameterize it so that anyone can put their specific keys and locations and reuse your code.

Try your Docker image on AWS and run it for 2008 and share the locations for the AWS bucket with the processed data and log file in your report

Submission:
Submit the github with the Docker file and source code so that we can rebuild the Docker images. Also register your Docker image on Dockerhub and provide links.

Write a report detailing:

- Your design and implementation for the Docker part.
- Review your outputs stored on Amazon S3 and discuss results. How does it handle exceptions ? What if amazon keys are invalid? What if data wasn't found?

Reference:

1. https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm
2. https://www.sec.gov/data/edgar-log-file-data-set.html