**ETL Project Report**

**Extraction:**

Our application idea was so people can find developer jobs in the Irvine area that are the easiest to apply to, so we extracted our data from two websites listed below:

***url1*** *https://www.glassdoor.com/Job/jobs.htm?sc.keyword=developer&locT=C&locId=1146798&locKeyword=Irvine,%20CA&jobType=all&fromAge=-1&minSalary=0&includeNoSalaryJobs=true&radius=25&cityId=-1&minRating=0.0&industryId=-1&sgocId=-1&seniorityType=all&companyId=-1&employerSizes=0&applicationType=1&remoteWorkType=0"*

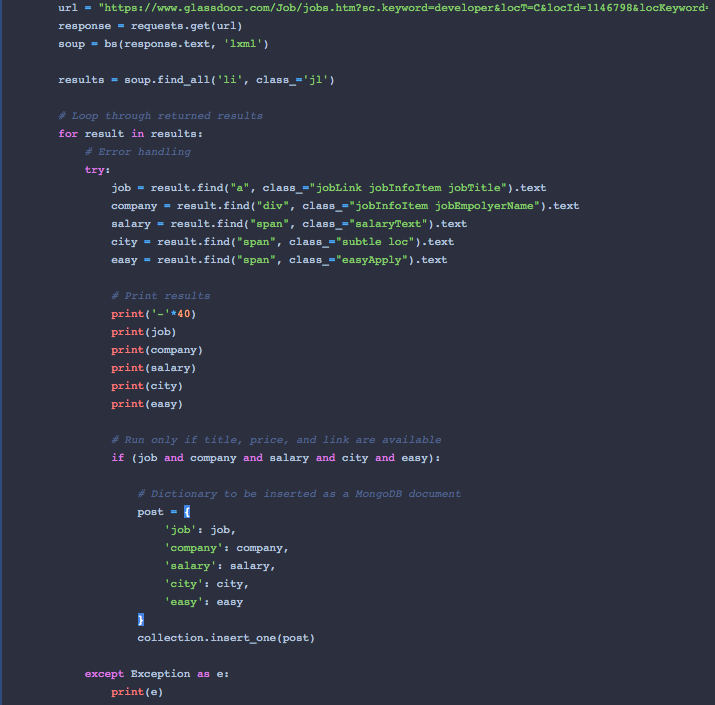
***url2-****https://www.indeed.com/jobs?q=developer&l=Irvine%2C%20CA&start=10&advn=6908518619934192&vjk=9fc0a07016cd1dbf*

They were formatted as html, so we used BeautifulSoup to parse through both html’s and find the specific tags and classes we needed. We also utilized the request function to retrieve the data from the url’s.

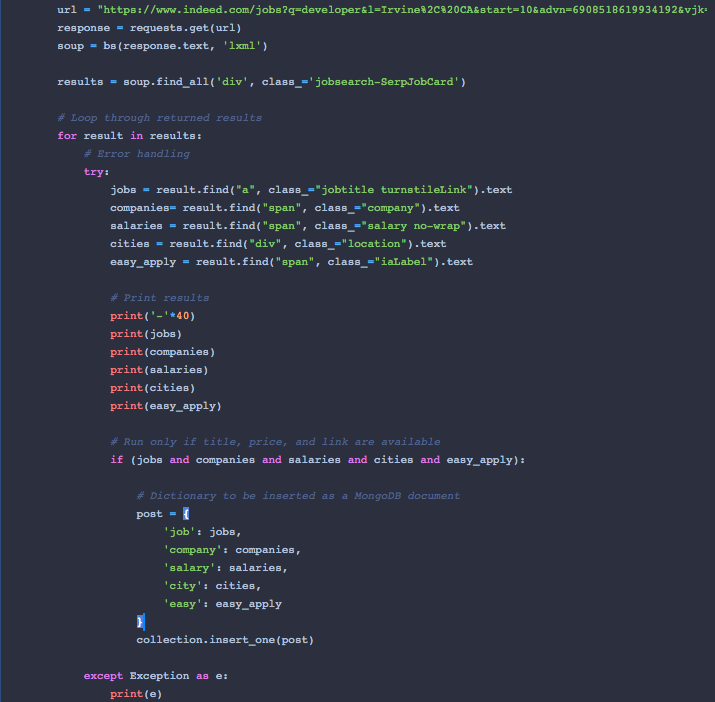
**Transformation:**

\*By importing pymongo, we connected our scrape information to MongoDB using the connection techniques taught in class. We also created two separate databases: glass\_db and indeed\_db

***Tranformation Process for Glassdoor URL:***

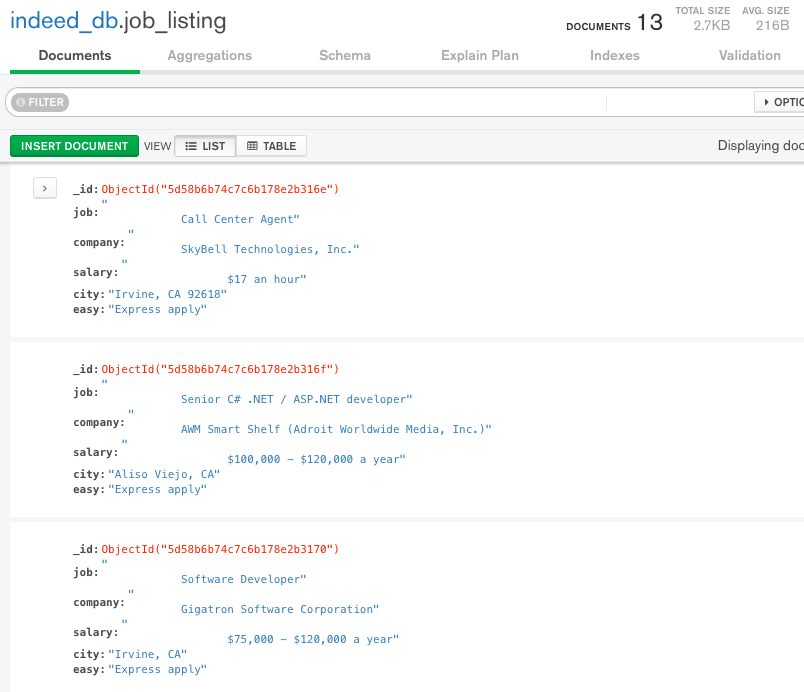
****

***Tranformation process for Indeed:***

****

**Load:**

We used MongoDB Compass instead of SQL because we were not trying to join and extract any information from tables. Instead, we wanted to scrape data from the data sources we chose and transform them into json’s so we can reference them in MongoDB.

****