CS506 Bay State Banner Project
Deliverable 1

Zilin Zhang (U87038789)
Mingyan Yang (U30215865)
Yujing Chen (U70567267)
Yicheng Li (U29503597)

1) Current Work

a) We have tried web scraper to scrape the website. However, web scraper is suitable for html pages instead of aspx pages.
b) We have tried a web scraping tool called Octoparse to scrape the real estate website.
c) Get names and addresses from Team 1 from ZBA, NAIOP, BPDA, etc. (all data sets).
d) Searching names and addresses (unique employee name or name + zip code) of people and companies listed in ZBA etc. data within OCPF (get from campaign finance project).

2) Problems

a) The website has rejected our request after we scraped only about 300 datas.
b) The website contains millions of data, which would take a long time to scrape. Using OpenCV to extract key words would also cost a long time.
c) It's a dynamic website, so we cannot save the aspx page as PDF format.
d) At least four types of annual report, including one kind of hand-written report, can cause serious trouble for text extracting.

3) Future Work

a) Try to solve the scraping wall problem of the website.
b) Extract names and addresses from PDF files downloaded from the website.

4) Goals for this week

a) aGet names and address from Team 1 from ZBA, NAIOP, BPDA, etc. (all data sets)
b) Searching names and addresses (unique employee name or name + zip code) of people and companies listed in ZBA etc. data within OCPF (get from campaign finance project)

5) Data description

a) All datasets are stored in .csv or .xlsx files. Filename implies the source of the datasets.
b) Datasets from ZBA from 2017 to 2019 are stored separately in different files while datasets from other sources are stored in one file.

c) First row contains the names of attributes like Date, First name, Last name, etc.
d) Not all datasets are complete while some datasets are missing names and some are missing addresses.

6) Questions

a) Problems from missing attributes. How should we deal with data without the attributes we want?
b) Relationships between data. Is there any relationships hidden between these datasets? How should we figure that out?
c) Missing datasets. When we try to extract information from OCPF datasets, we find Team 1 only has the OCPF datasets from Nov. to Dec. 2019 and 2018. And we can only get processed data from the campaign_finance_city's google drive or campaign_finance_state group. So how can we get the origin "2019_ocpf.xlsx" file?
d) Attributes description. We know that we need to extract names and addresses from datasets but what exactly we want? Do we need state addresses? Which is our target, contributors or employers? Should names be in "First Name Last Name" form or in other forms?