CS506 Bay State Banner Project
Deliverable 1

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

1. Current Work

1) We have tried web scraper to scrape the website. However, web scraper is suitable for html pages instead of aspx pages.
2) We have tried a web scraping tool called Octoparse to scrape the real estate website.

2. Problems

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

3. Future Plan

1) Get names and address from Team 1 from ZBA, NAIOP, BPDA, etc. (all data sets).
2) 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).
3) Try to solve the scraping wall problem of the website.