CS506 Deliverable 1

Xiaoyu An

Qing Han

Qihao Sun

Tingyi Zhang

Ruihong Zhu

Methods for Dataset Collection

1. OCPF

   1. Raw data collection: by searching keywords of Contributor/Occupation/Employer we can find out the raw data including attributes: Date, Contributor, Address, City, State, Zip, Occupation, Employer, Principal Office, Amount, CPF_ID, Recipient, Tender_Type_ID, Tender_Type_Description, Record_Type_ID, Record_Type_Description, Source_Description.

   2. Duplication deleting and abnormal data replacement: remove duplication by judging if Date, Contributor, Address, Amount, CPF_ID are same at the same time; replace empty attributes of entries with NaN.

   3. Data integration: integrate cleansed data together as the OCPF_RealEstate_11/2019~12/2019.

   4. Further data collection: since our clients have provided more searching keywords, we may need to do more OCPF data collection works like stated above.

2. ZBA

   1. ZBA Meeting raw data collection: Use the Python Library Tika to parse PDF files and use regular expressions to search for keywords and get data based on the following attributes: BOA number, Address and applicant name.

   2. Check if the three datasets have the same amount of data and handle some edge cases.

   3.ZBA Decision information can be simply retrieved from website and put into csv files.

3. Approved Building Dataset

   1. Raw data collection: filter the data by mixed and commercial occupancy type and download the csv file.

2. Extract owners: we use the PANDAS package to process the data and extract the all owners of these sites.

3. Deal with duplicates: this process is pretty straight forward, we simply use the “drop_duplicates” method to remove duplicates.

Difficulties

1. For ZBA Meeting data collections, several of the PDF files are scanned-version and cannot be parsed by using Python Library. We addressed this issue with Spark staff members and they agreed that we can skip these files, since most ZBA meeting data will be included in ZBA Decision data.

Is it possible to achieve the final result based on our current progress?

Yes, we are making significant progress toward the final goal. We have successfully completed most of the data collection. Next, we are going to merge and match businesses to development/ real estate projects, and then analyze data to understand the influence of the real estate industry in Boston by looking at their prominence as political contributors and city decisionmaking. The project is developing smoothly, and we believe we could achieve the final goal based on our current progress.