Rent Scraping

Author: Himi Yadav

Overview of Issues and their Solutions

1. Site code for the landing page is different across years
    a. Extract all the text from the various landing pages.

```
def get_contents(soup, content_text):
  try:
    parents_blacklist=['[document]','html','head',
                       'style','script','body',
                       'section','tr',
                       'td','label','ul','header',
                       'aside']
    content=''
    text=soup.find_all(text=True)
    for t in text:
        if t.parent.name not in parents_blacklist and len(t) > 5:
            content=content+t+' '
    content_text.append(content)
  except Exception:
    content_text.append('')
    pass
```

    b. Write functions for filtering the fields (eg: names, owners, addresses, prices) that are needed for the final scraped output
    c. How to filter?
        i. *Regular Expressions*: For the more straight forward fields like addresses, phone numbers, and prices, regular expressions that fit all possible ways to write any of the above fields is enough and achieves 100% accuracy. Use `regex` and `spacy` packages for this. Below has example regular expressions.

```
contacts = re.findall(r'\(?[0-9]{3}\)?\/?\.?-?\s?[0-9]{3}-?\.?-?\s?[0-9]{4}',
text)
```

```
address = regex.findall(r'[\s\n]*[1-9][0-9]{,4} (?:[A-Z][a-zA-Z]*,?\s*){,5}
Chicago,?[\w\s,]{,3}\s*[0-9]{,6}?', text)
```

```
prices =
re.findall(r'\$[1-9]?[0-9]{,3},?[0-9]{,3}\s*-\s*\$?[1-9]?[0-9]{,3},?[0-9]{,3}',t
ext)
```

ii. *Training a Custom Named Entity Recognition Network*: For the names of the properties or the names of the owners, it’s not possible to find patterns by hand. This is where a neural network needs to be trained to achieve high accuracies. Named Entity Recognition comes to play and the custom entities for this particular project are property name and owner name.

   1. Create training database using a web scraping technique where you manually adapt to the changes in the structure of the landing pages. The training data size should be considered in depth and sufficient data should be given as training data to the neural network while training.
   2. Label the entities that the neural network is being trained for. Here, the entities of interest are property name and owner name. Now, the training data is ready and the next step is to train the transformer.
   3. Train a transformer that can classify the text to entities.
   4. Use the trained network to filter text for future years or future data (test data).

Useful Links and Tutorials:

1. https://towardsdatascience.com/something-from-nothing-use-nlp-and-ml-to-extract-and-structure-web-data-3f49b2f72b13
2. https://docs.python.org/3/howto/regex.html
3. https://newscatcherapi.com/blog/train-custom-named-entity-recognition-ner-model-with-spacy-v3
4. https://towardsdatascience.com/named-entity-recognition-ner-using-spacy-nlp-part-4-28da2ece57c6