Web Scraping

PROJECT REPORT

ANWESHA PAUL : MDS202213

DCBD

REPORT

1. **OBJECTIVE :**
   To extract addresses from web pages.

2. **BEST SAVINGS :**
   The best savings we could achieve on the given input is 98.09%.

3. **DATA PREPROCESSING :**

- Used the nltk library (a NLP library) for the following processess.
- Removed HTML tags using BeautifulSoup.
- Tokenized the list into separate words for further processing. (Using word tokenize from nltk library)
- Tokenizing every word of the file/Splitting a sentence into list of words. Giving a regular expression. we are seperating the stopwords which have alphabets and letters and storing it in a string to remove all the stopwords of english
- Looked for certain keywords in the string that occur mainly near an address, then extracted all words after and before these keywords upto a certain length from the string.
- We removed the parts of speech from the string which mostly does not occur in the address.

4. **IMPROVING SAVINGS SCORE :**
   We could have made the code more efficient in order to get the exact address using NLP techniques only.

5. **CHALLENGES FACED :**
   We would rate the difficulty level of the assignment as moderate to difficult since without any prior knowledge, working with ”beautifulsoup” and ”nltk” was challenging.

KEY QUESTIONS

(1) What is the best savings you could achieve on the given input?

(2) What data processing steps did you perform?

(3) If provided more time, what more could you have done to improve your savings score?

(4) How easy/difficult was this task? What challenges did you come across?