1. Create test dataset: shops come from the same city but have different id. All London fish & chips shops are treated as test datasets
2. Use HTMLPaser to filter HTML source code:
3. Filter HTML tags
4. Filter more useless words
5. Filter special symbols

Problems:

1. Because of the encoding problems, some words will appear ASCII special characters or spaces (\xc2\xa0) after they add to the list. This will influence the reduce procedure that the same words cannot match properly. In CSV file, special symbols appear as spaces in CSV. Firstly, the project tries to use regular expression to filter the special symbols before adding words to the list which used for writing to the CSV file. However, the project finds that these special symbols will appear after the words added to the list. Thus, the project tries to find a way to filter these special symbols in the list. The project first use replace() function to filter them when the word is added into list (str(split\_result[j]).replace('\xc2\xa0', '')) and this function succeeds. However, if the project uses this method then the project needs to find all the special symbols. As a consequence, the project tries to find another way which use regular expression to filter the list directly. The project converts the whole list to string and then use regular expression to filter the string. However, after filtering, the project should convert the string to list again. Fortunately, eval() function can help to do this and the project successfully solved encoding problem.
2. Fix bug in reduce.py: Coordinates of different shops in the same area can be added to coordinate sets. Besides, the project uses a simpler method to achieve the reduce procedure and the efficiency of reduce procedure has improved.
3. Use UNIX sort command to sort word.
4. Find more websites: Find more websites in the UK, so that the distribution of shops will not too concentrated. The project has found more menu websites in 19 regions.

Problem:

1. The number of shops in less populated areas is small and many do not have their own menu websites.
2. Some menu websites are PDF and are useless.
3. Some region has few (2 or 3) websites.
4. Test and improvement:
5. Regional testing.
6. All areas tested together.
7. Evaluation: Find some useless words and useless symbols to improve filter procedure.