Homework 2: Building Wrapper
INF 558 BUILDING KNOWLEDGE GRAPH

DUE DATE: Monday, 09/18/2017 @ 11:59pm on Blackboard

Ground Rules
This homework must be done individually. You can ask others for help with the tools, but the submitted homework needs to be your own work.

Summary
In this homework, you will construct a wrapper to extract data from semi-structured sources like webpages.

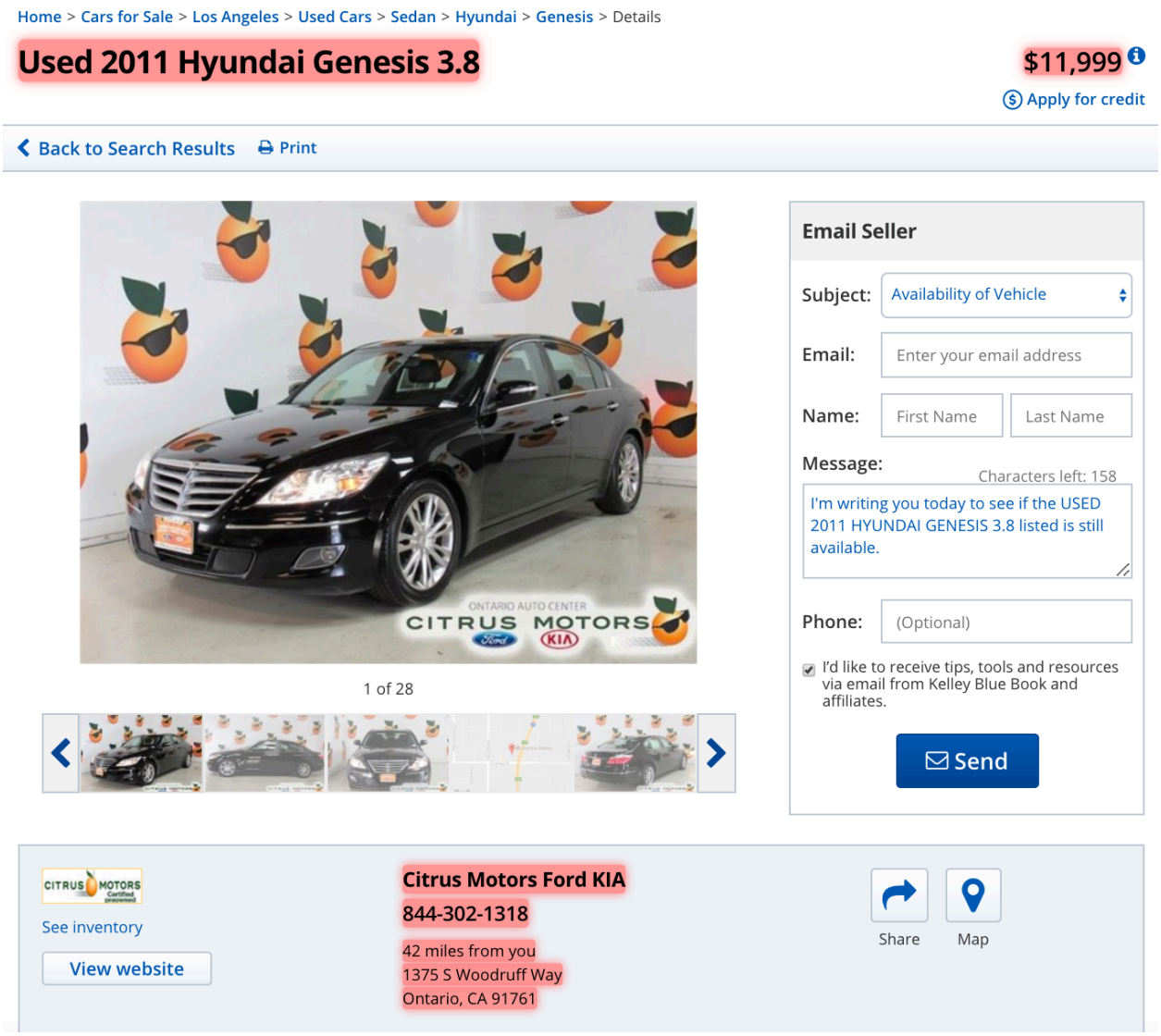
Task 1 (1 point)

Identify your data source and the data you want to extract. You should choose a website and scrape at least 1000 webpages. You can use crawled webpages from Homework 1. Choose one sample web page, list all the interesting fields and highlight them on the webpages.

Examples: (This is just an example. You should list all the fields that you think can be extracted from the web pages)

Interesting fields from Kelly Blue Book webpage:

- Car:
    - Model/Make/Year
    - Price
- Retailer:
    - Name
    - Phone Number
    - Address



Task 2 (4 points)

Use Inferlink extraction tool and try to extract highlighted fields in Task 1 from your webpages. **Put screenshots of your extracted data from Inferlink tool and download the extracted data under csv format on your report.**

Answer the following questions (maximum 2 sentences for each question):

- Can Inferlink tool extract your highlighted fields?
- If not, list up to 3 fields that cannot be extracted and explain why the tool cannot extract these fields. Hint: using lectures about wrapper learning and Inferlink extraction tool.
- Choose one extracted rule and explain how the rules can be used to extract field from webpages.

Task 3 (5 points)
Construct your wrapper using any existing tools available. For example, one library you can use is BeautifulSoup (http://www.crummy.com/software/BeautifulSoup/) for writing your own scaper. Another example is using the service provided byhttp://scrapinghub.com/.
You can either manually construct the wrapper or use a wrapper learner. Extract at least 5 fields from your webpages, and save them into a file in json format.
Take screenshot of your first two json objects and show it on your report.

Describe in your report how your wrapper works. e.g. how you developed it, what library you are using or what is the way you used to find where data is located, etc.

Submission Instructions
You must submit the following files/folders in a single .zip archive named Firstname_Lastname_hw2.zip and submit it via Blackboard:

- **Firstname_Lastname_hw2_report.pdf**: A pdf file containing your answers to the Task 1, Task 2 and Task 3.
- **inferlink.json:** contains your Inferlink extracted data
- **inferlink.**rules: contain your Inferlink rules
- **wrapper.**json: contains your wrapper extracted data
- **source**: This folder includes all the code you wrote to accomplish Task 3.