Homework 3: Information Extraction
INF 558 BUILDING KNOWLEDGE GRAPH

DUE DATE: Monday, 09/25/2017 @ 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

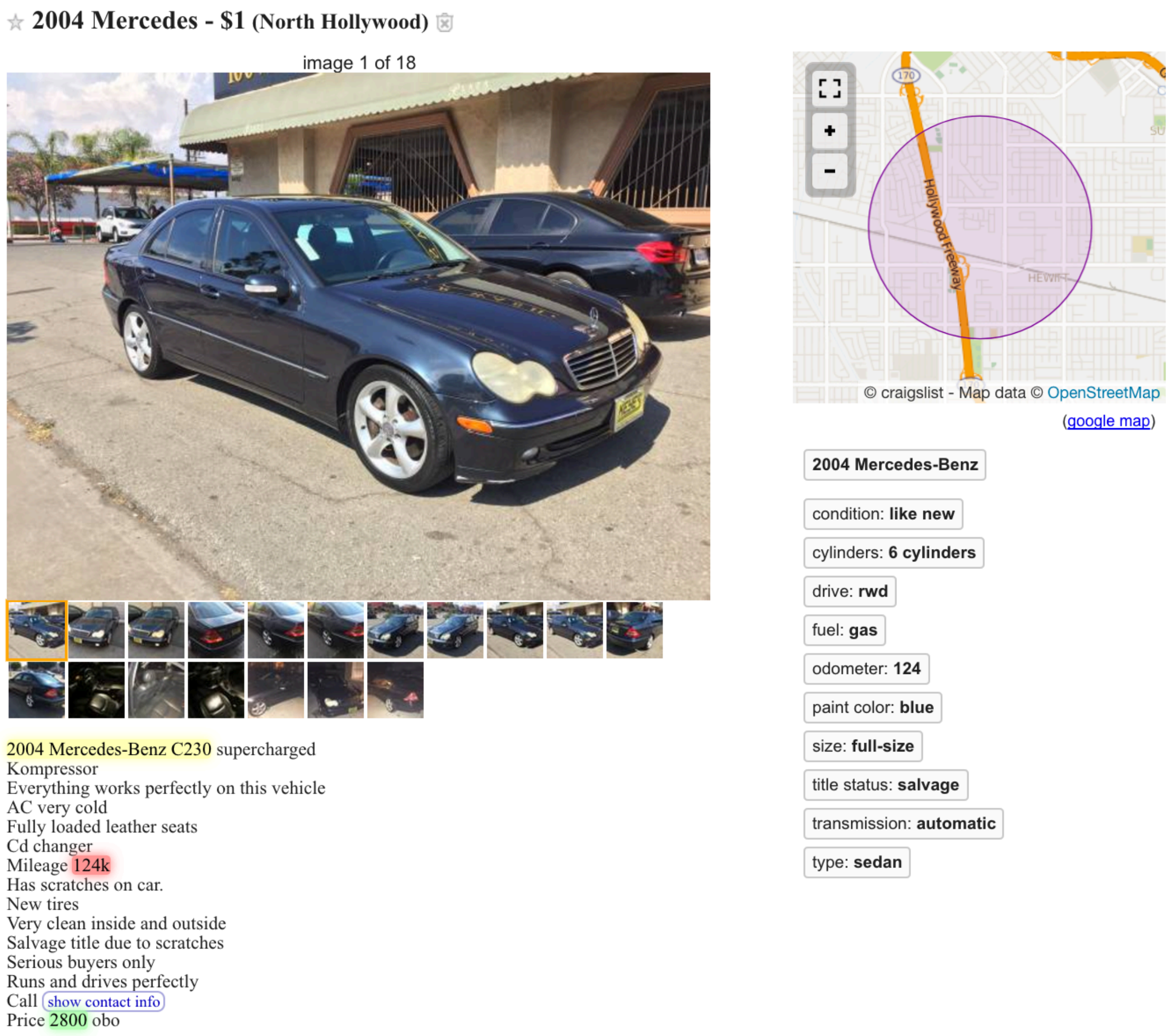
In this homework, you will build a CRF classifier to extract information from unstructured text.

Task 1 – Preliminary Work (1 point)

First, you should identify the source of unstructured text and the key information you want to extract from it. We encourage students to use the webpages they crawled for HW1, but if it is not suitable for this HW (e.g. lack of unstructured text), you may either use the webpages crawled by your project partner (but the two of you should develop your own code separately for this HW) or crawl a new set of webpages for this HW.

We would like you to label each word in the unstructured text with either the key information you want, or an "irrelevant" label. Define the kinds of labels you will use. The information you want should be those that cannot be easily extracted from other parts of the webpage. You should define **at least one other label** besides "irrelevant".

Example:



☆ 2004 Mercedes - $1 (North Hollywood)

2004 Mercedes-Benz C230 supercharged
Kompressor
Everything works perfectly on this vehicle
AC very cold
Fully loaded leather seats
Cd changer
Mileage 124k
Has scratches on car.
New tires
Very clean inside and outside
Salvage title due to scratches
Serious buyers only
Runs and drives perfectly
Call show contact info
Price 2800 obo

2004 Mercedes-Benz

condition: **like new**

cylinders: **6 cylinders**

drive: **rwd**

fuel: **gas**

odometer: **124**

paint color: **blue**

size: **full-size**

title status: **salvage**

transmission: **automatic**

type: **sedan**

Figure 1. Sample post from craiglist

When posting ads on craiglist, sellers can enter different fields of information (boxes on the right) in craiglist which makes it easy to extract directly using wrappers. However, a lot of ads contains wrong or insufficient in these fields, opting to list all the information within the description.

Figure 1 shows one instance of such case, where the odometer field and the price on title are incorrect. Moreover, the model of the car is also missing in the form. These fields of information can only be found in the description (highlighted text). We can use CRF to extract this number from the unstructured text.

Task 2 – Training the CRF

Download a CRF-supported library. Some suggestions are:

- http://www.chokkan.org/software/crfsuite/
- https://python-crfsuite.readthedocs.io/en/latest/
- https://taku910.github.io/crfpp/

Read the tutorial to understand how the library should be used.
Prepare a set of training data for the CRF. Only a portion of the webpages you collected should be used as the training data.

You need to perform the following steps:

- Select one or more types of information that you will extract. You must extract **at least one type of information**, but you can extract more, especially if you need them for your project.
- Decide what kind of tags and features you would like to extract from each word/token. (E.g. part-of-speech, prefixes, suffixes). Write code or use existing tools to extract the features and append them to your data.
- Manually label **at least 50 records** for training. These are to be used as the training data.
- Manually label **at least 20 records** for testing. These are to be used as testing data.
- Write program to train a CRF classifier using the training data and calculate your classifier performance (precision, recall and F-1 measure) on the testing data.

Task 3 – Report

Answer the following questions:

- What was the information you were looking for? Describe the labels you chose and why. Also include at least one screenshot of the webpage you are using and show where is the information you are looking for.
- What kind of tags did you tag your data with? Explain your choices.
- Report your classifier's precision, recall and F-1 measure. Why did your classifier perform well (or not satisfactorily)?

Submission Instructions

You must submit the following files/folders in a single .zip archive named Firstname_Lastname_hw3.zip and submit it via Blackboard:

- Firstname_Lastname_hw3_report.pdf: A pdf file containing your answers to **Task 3**.
- training: This folder contains your training data
- testing: This folder contains your testing data
- source: This folder includes all the code you wrote to accomplish **Task 3**.