Project Stage 2. Extracting Structured Information from the Data

Started on Fri Feb 26. Pls submit by midnight Mon Mar 14 (see how to submit below).

In this stage, you need to extract brand name values from the "Product Name" attribute of the products.

Brand of a product refers to the name assigned by a manufacturer to a product or a range of products. For example, suppose we have a product name "Apple iPhone 6s Plus - Space Gray". Here "Apple" is the manufacturer name, "iPhone" is the brand name, "6s Plus" is the model number, and "Space Gray" is the color.

However, if you closely analyze the brand name values in the product dataset of Stage 1, you will see that the brand names often refer to the manufacturer names. This is because the notion of "brand" is very subjective and different people interpret it differently.

In this stage, your task is to automatically extract brand/manufacturer names from the "Product Name" attribute value in the JSON representation of a product. To do this, you can take a dictionary-based approach. We will provide you with a dictionary of ~8K brand name values that we have extracted by processing a large database of electronic products.

Step 1. You can download the data for this stage from http://pages.cs.wisc.edu/~sanjibkd/784-stage2/ The data contains 3 files:
(a) elec_pairs_stage2.txt that has 10K labeled product pairs (these pairs are different from the pairs of stage 1 because now we have suppressed the "Brand" and "Manufacturer" attributes so that you can extract them from "Product Name").

(b) elec_brand_dic.txt that has 8442 entries, where each entry is of the form:
brand name <TAB> frequency. Here frequency denotes the no. of products (in our database) having that brand name

(c) readme.txt describes file (a)

Step 2. From the set of 10K labeled product pairs, randomly sample 350 products. Let this sample be S. For each product in S, go into the attribute "Product Name" and pull out the correct band name (if exists). This will give you the golden data (that you can use to debug, estimate accuracy, etc.).

Step 3. Randomly split S into a development set I and a testing set J. The set I must have at least 200 products and the set J must have at least 120 products.

Step 4. Now use the development set I to develop your brand name extractor. You can use any method you'd like. If you decide to use the dictionary based approach, we have supplied a dictionary of brand names for you. YOU CAN ONLY USE THE DEVELOPMENT SET I. YOU ARE NOT SUPPOSED TO EVEN LOOK AT THE TESTING SET J. We will discuss more in the class ideas on using the development set I.

Step 5. Once you are satisfied with your brand name extractor, you will apply it to the testing set J to compute precision and recall on J.

What to submit and how

Submit the following by providing links to them on your group homepage. Do not email the instructors.

+ a file storing the set S of golden data that you have created. Give a readme file describing the format of S.

+ a file storing the development set I.

+ a file storing the test set J.

+ link to the code of your brand name extractor.

+ a report describing precision and recall that you obtain on the set J, how your brand name extractor works, and an analysis on why it hasn't been able to reach higher accuracy.
