Web Mining

Assignment 4: Analysis of Entities on Wikipedia
(5% weightage)

Date Posted: Oct 21, 2013
Date of submission: Oct 29, 2013. 9pm.

Goal: To make students understand the Wikipedia dataset, especially the entity info boxes.

Task: We will provide you with the Wikipedia dump. Your aim is to extract information about various entity types.
The steps for this task are as follows:
1. Given the Wikipedia dump, gather all the pages from Wikipedia with Info boxes on them.
2. Find the set of all possible entity types on Wikipedia
3. Find the set of all possible attributes that can be associated with any entity type on Wikipedia.
4. From a few values of these attributes, infer the data type of these attributes as one of the following: String, set of strings, duration, number, set of durations, date, other.
5. Find various units that can be used to express the value of a numeric attribute. E.g., for "height" attribute of "person" entities, the units could be "cms, inches"
6. For numeric attributes, find typical ranges (using the most popular unit). E.g., For person entities, the age attribute should have the range as 0-150 years.
7. For attributes which are semantically similar but have different names used across different entities of the same type, merge them. E.g., identify that the attribute "birthdate" is the same as "bdate". Don't do this manually.

Submission Instructions: Create a directory with the name "<rollno>_as4". Within that you need to put

a. One file per entity type and name the file as <entity-type>.txt
b. A readme file README.txt
c. Your code directory zipped as code.zip.

The format of each <entity-type>.txt is as follows.
There is one line per attribute. Each line contains the following fields separated by a tab.

1. Attribute name
2. Inferred data type of the attribute
3. List of units used for the attribute (separated by commas)
4. If the attribute is numeric, put in the min value observed for that attribute, else put -1
5. If the attribute is numeric, put in the max value observed for that attribute, else put -1
6. List of other names for this attribute separated by a comma.
7. Number of entities of type= <entity-type> containing this attribute.

The README.txt should

1. Describe your methodology in brief.
2. What heuristics you used to infer the entity type from the Info boxes.
3. What heuristics you used to infer the data type of the attributes
4. What heuristics you used to infer that 2 different attributes actually mean the same
5. Average number of attributes per entity type
6. Number of entities per entity type.

The code directory should contain all your code. Zip the code directory to create code.zip

Finally, zip the “<rollno>_as4” directory to get <rollno>_as4.zip and submit the zip file.