**Samet Ayhan**

**10/28/2014**

**Lab #5**

**Part #1**

Fields I used and the comparison function:

fields = [

{'field' : 'title', 'type': 'String'},

{'field' : 'price', 'type': 'Custom', 'has missing':True, 'comparator' : customComparator},

{'field' : 'manufacturer', 'type': 'String', 'has missing' : True}

]

For the implementation, please see **product\_dedup.py**

The precision and recall the algorithm achieved:

found duplicate

1286

precision

0.206065318818

recall

0.171521035599

The precision value it attains is relatively low. This says that number of true positives (i.e. the number of items correctly labeled as dedups) divided by the total number of elements labeled as dedups (i.e. the sum of true positives and false positives) is ~21%. Hence, it appears that there are plenty of false positives in the dedup outcome.

The recall value in comparison to precision is not that bad, which is ~17%. This is reflection of the number of true positives divided by the total number of elements that actually belong to the dedup (i.e. the sum of true positives and false negatives). Here it seems that there are quite a bit false negatives also.

For the evaluation, please see **product\_evaluation.py**

**Part #2.1**

For the script, please see **Assignment 5 - Part 2.ipynb**

File: news1.html

Recognized Named Entities:

Reuters , ORGANIZATION

Ralph Nader , PERSON

Apple Inc , PERSON

Tim Cook , PERSON

Wall Street Journal , FACILITY

File: news2.html

Recognized Named Entities:

Reuters , ORGANIZATION

Google Inc , PERSON

Larry Page , PERSON

Sundar Pichai , PERSON

File: news3.html

Recognized Named Entities:

Reuters , ORGANIZATION

Hewlett-Packard Co , ORGANIZATION

Wall Street Journal , FACILITY

File: news4.html

Recognized Named Entities:

Reuters , ORGANIZATION

White House , FACILITY

West Africa , LOCATION

Ebola , PERSON

File: news5.html

Recognized Named Entities:

Reuters , ORGANIZATION

Sevilla , PERSON

Barcelona , PERSON

La Liga , PERSON

File: news6.html

Recognized Named Entities:

Reuters , ORGANIZATION

Gonzalo Higuain , PERSON

Rafael Benitez , PERSON

File: news7.html

Recognized Named Entities:

Reuters , ORGANIZATION

PSV Eindhoven , ORGANIZATION

File: news8.html

Recognized Named Entities:

Reuters , ORGANIZATION

Michael Fallon , PERSON

File: news9.html

Recognized Named Entities:

Reuters , ORGANIZATION

Kobani , PERSON

Islamic , ORGANIZATION

File: news10.html

Recognized Named Entities:

Reuters , ORGANIZATION

Gaza , LOCATION

Sinai , ORGANIZATION

**Part #2.2**

For the script and the results, please see **Assignment 5 - Part 2.ipynb**