**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, please see **Assignment 5 - Part 2.ipynb**

[ORG: u'NPR'] u"'s" [LOC: u'Washington']

[ORG: u'WHYY'] u'in' [LOC: u'Philadelphia']

[ORG: u'McGlashan &AMP; Sarrail'] u'firm in' [LOC: u'San Mateo']

[ORG: u'Freedom Forum'] u'in' [LOC: u'Arlington']

[ORG: u'Imagine Media'] u'of' [LOC: u'Brisbane']

[ORG: u'Audio Net'] u', a' [LOC: u'Dallas']

[ORG: u'White House'] u"support and facing only tepid opposition from the nation's governors," [LOC: u'Silicon Valley']

[ORG: u'Brookings Institution'] u', the research group in' [LOC: u'Washington']

[ORG: u'Idealab'] u', a self-described business incubator based in' [LOC: u'Los Angeles']

[ORG: u'Open Text'] u', based in' [LOC: u'Waterloo']

[ORG: u'WGBH'] u'in' [LOC: u'Boston']

[ORG: u'Congress'] u'and a noted expert on' [LOC: u'Russia']

[ORG: u'Home Box Office'] u'documentary about a' [LOC: u'New York']

[ORG: u'Bastille Opera'] u'in' [LOC: u'Paris']

[ORG: u'Met'] u'. After returning to' [LOC: u'New York']

[ORG: u'Omnicom'] u'in' [LOC: u'New York']

[ORG: u'DDB Needham'] u'in' [LOC: u'New York']

[ORG: u'DDB Needham'] '' [LOC: u'U.S.']

[ORG: u'Kaplan Thaler Group'] u'in' [LOC: u'New York']

[ORG: u'BBDO South'] u'in' [LOC: u'Atlanta']

[ORG: u'Georgia-Pacific'] u'in' [LOC: u'Atlanta']

AT(u'bob\_edwards', u'national\_public\_radio')

AT(u'wertheimer', u'npr')

AT(u'mike\_godwin', u'electronic\_frontier\_foundation')

AT(u'cohn', u'mcglashan\_&\_sarrail')

AT(u'lillian\_r\_berkman', u'center\_for\_internet\_and\_society')

AT(u'james\_boyle', u'american\_university')

AT(u'robert\_mergess', u'berkeley\_center\_for\_law\_and\_technology')

AT(u'jack\_balkin', u'yale')

AT(u'pam\_samuelson', u'berkeley')

AT(u'clinton', u'berkeley')

AT(u'samuelson', u'berkeley')

AT(u'samuelson', u'harvard')

AT(u'michael\_froomkin', u'university\_of\_miami')

AT(u'dan\_burk', u'seton\_hall\_university')

AT(u'david\_post', u'cyberspace\_law\_institute')

AT(u'frank\_easterbrook', u'7th\_us\_circuit\_court\_of\_appeals')

AT(u'jones', u'qradio')

AT(u'jones', u'qradio')

AT(u'vern\_fotheringham', u'qradio')

AT(u'william\_gale', u'brookings\_institution')

AT(u'joel\_slemrod', u'university\_of\_michigan')

AT(u'alan\_braverman', u'credit\_suisse\_first\_boston')

AT(u'bill\_gross', u'idealab')

AT(u'abe\_kleinfield', u'open\_text')

AT(u'braverman', u'yahoo')

AT(u'kaufman', u'tv\_books\_llc')

AT(u'kaufman', u'tv\_books')

AT(u'michael\_coffey', u'publishers\_weekly')

AT(u'lorne\_michaels', u'tv\_books')

AT(u'james\_billington', u'congress')

AT(u'sherry\_lansing', u'paramount\_motion\_picture\_group')

AT(u'lindsay\_doran', u'united\_artists')

AT(u'baz\_luhrmann', u'mtv')

AT(u'laura\_ziskin', u'fox\_2000')

AT(u'rick\_yorn', u'addis-wechsler\_&\_associates')

AT(u'tom\_rothman', u'20th\_century\_fox')

AT(u'charlotte\_forest', u'homestead\_editorial')

AT(u'richard\_strauss', u'la\_scala')

AT(u'wilson', u'met')

AT(u'philip\_glass', u'met')

AT(u'john\_wren', u'omnicom')

AT(u'steve\_davis', u'wells')

AT(u'wren', u'wells')

AT(u'charlie\_moss', u'moss/dragoti')

AT(u'ken\_kaess', u'ddb\_needham')

AT(u'lawrence', u'wells')

AT(u'linda\_kaplan\_thaler', u'wells')

AT(u'kaplan\_thaler', u'wells')

AT(u'kaplan\_thaler', u"toys\_``r''\_us")

AT(u'kaplan\_thaler', u'wells')

AT(u'ken\_haldin', u'georgia-pacific')

AT(u'davis', u'wells')

AT(u'nan\_a\_talese', u'doubleday')

[PER: u'Robert Mergess'] u', the co-director of the' [ORG: u'Berkeley Center for Law and Technology']

[PER: u'Jack Balkin'] u", director of the school's program. ``What happened at" [ORG: u'Yale']

[PER: u'David Post'] u', co-founder of the' [ORG: u'Cyberspace Law Institute']

[PER: u'Frank Easterbrook'] u'of the' [ORG: u'7th U.S. Circuit Court of Appeals']

[PER: u'Vern Fotheringham'] u', a founding partner of' [ORG: u'Qradio']

[PER: u'Bill Gross'] u', president of' [ORG: u'Idealab']

[PER: u'Kaufman'] u', president of the privately held' [ORG: u'TV Books LLC']

[PER: u'Kaufman'] u'would not disclose financial details of the deal.' [ORG: u'TV Books']

[PER: u'Michael Coffey'] u', managing editor of' [ORG: u'Publishers Weekly']

[PER: u'Lorne Michaels'] u", the executive producer of ``Saturday Night Live.''" [ORG: u'TV Books']

[PER: u'James Billington'] u', the librarian of' [ORG: u'Congress']

[PER: u'Sherry Lansing'] u', chairwoman of the' [ORG: u'Paramount Motion Picture Group']

[PER: u'Lindsay Doran'] u', president of' [ORG: u'United Artists']

[PER: u'Laura Ziskin'] u', president of' [ORG: u'Fox 2000']

[PER: u'Tom Rothman'] u', president of production at' [ORG: u'20th Century Fox']

[PER: u'Philip Glass'] u'. This was not part of the' [ORG: u'Met']

[PER: u'Charlie Moss'] u', the chairman of' [ORG: u'Moss/Dragoti']

[PER: u'Ken Kaess'] u', president of the' [ORG: u'DDB Needham']