**CS 839, Stage 4**

Lan Bai, Chaoqun Mei, Yuzhe Ma

May 8, 2018

**Abstract** In this project stage, we merged the two tables extracted from Amazon and Barnes & Noble and then did analysis (clustering on the merged table) .

**1. Merge of A and B**

To merge A and B, we first apply the matcher we obtained in Stage 3 to all pairs after blocking to drop the unmatched pairs, so we get a table with all matched pairs, which contains the ID of matched tuples in A and B. Then we scan A, check the ID of the tuple, if the ID does not exist in the table of matched pairs, we append that tuple to the end to B.

We do not have other tables.

ID: Number ID of book in the table, from 1 to N (number of tuples).

Publisher: The book's publisher.

Time: The book's publish date.

Author: The book's author's name.

Title: The book's title (name).

**2. Statistics of E**

We use an AttrEquivalenceBlocker and an OverlapBlocker. We assume that the same books should have same publish date, so we use AttrEquivalenceBlocker to block tuple pairs whose Time attributes are not the same (We let the Time to be the same format in preprocessing). We also assume that the same books should have overlap of at least one word in their title, so we use OverlapBlocker to block tuples whose Title attributes have less than 1-word overlap.

Schema of E is the same as schema of A and B, which has 5 attributes:

ID: Number ID of book in the table, from 1 to N (number of tuples).

Publisher: The book's publisher.

Time: The book's publish date.

Author: The book's author's name.

Title: The book's title (name).

There are 6317 tuples in total.

Example:

6524,Touchstone,1/2/2018,Rob Barnett,The Economists' Diet: The Surprising Formula for Losing Weight and Keeping It Off

6525,Wiley; 1 edition,4/14/2014,Than Merrill,"The Real Estate Wholesaling Bible: The Fastest, Easiest Way to Get Started in Real Estate Investing"

6528,Harper Wave,3/13/2018,Leah Weiss PhD,"How We Work: Live Your Purpose, Reclaim Your Sanity, and Embrace the Daily Grind"

6529,Keys to the Vault; First edition,12/15/2017,Keith J. Cunningham,The Road Less Stupid

6530,Zondervan,4/25/2017,Bill High,Giving It All Away\_\_nd Getting It All Back Again: The Way of Living Generously

**Appendix**

Code of Merge

Feature = em.extract\_feature\_vecs(C,

feature\_table=F,

show\_progress=False);

# In[23]:

Feature = em.impute\_table(Feature,

exclude\_attrs=['\_id', 'ltable\_ID', 'rtable\_ID'],

strategy='mean');

# In[24]:

pred = Y.predict(table=Feature, exclude\_attrs=['\_id', 'ltable\_ID', 'rtable\_ID'],

append=True, target\_attr='predicted', inplace=False);

# In[34]:

Matched = [0]\*(len(A) + 1)

drop\_unmatch = []

for index, row in pred.iterrows():

if row['predicted'] == 1:

Matched[int(row['ltable\_ID'])] = 1;

else:

drop\_unmatch.append(index)

all\_match = C.drop(drop\_unmatch)

em.to\_csv\_metadata(all\_match, './Matches.csv')

# In[27]:

Result = B

id = 3862

for index, row in A.iterrows():

if Matched[row['ID']] == 0:

row['ID'] = id

id = id + 1

Result.loc[row['ID']-1]=row

em.to\_csv\_metadata(Result, './E.csv')