DS 320: Homework 3

Due: Tuesday, October 21, 2019, 11:59pm (EST)

Name: JOSIAH KIM

Question	Points	Score
1	10	
2	20	
3	5	
4	5	
5	10	
Total	50	

1. (10 points) Consider the following two data sources and their mediated schema:

Source S1:
Customers (ID, firstName, lastName, Address)
Products (ID, title, basePrice)
Purchases (cID, pID, date, quantity, totalPrice)

Source S2:
Cust (cID, cName, cAddress)
Items (iID, iTitle, iPrice)
Transactions (tID, cID, tDate, tTotal)
Transactions_Details (tID, iID, quantity, price)

Mediated schema:
Cust (ID, fullName, address)
Products (ID, title, unitPrice)
Sales (cID, pID, date, totalPrice)

Propose a set of view mappings between your data sources and your mediated schema using global-as-view mapping.

Cust (ID, fullName, address) :-
S1. Customers (ID, firstName + lastName, Address)

Products (ID, title, unitPrice) :-
S1. Products (ID, title, basePrice)

Sales (cID, pID, date, totalPrice) :-
S2. Cust (cID, cName, cAddress),
S2. Transactions (tID, cID, tDate, tTotal)

2) (20 points) Consider the following dynamic programming equations for the global alignment algorithm (left) and scoring matrix (right):

$$s(i,j)=\max\begin{cases} s(i-1,j-1)+c(x_i,y_j) \\ s(i-1,j)-c_g \\ s(i,j-1)-c_g \end{cases}$$

	d	a	v	e
d	2	-1	-1	-1
a	-1	2	-1	-1
v	-1	-1	2	-1
e	-1	-1	-1	2

$$s(0,j)=-jc_g$$
$$s(i,0)=-ic_g$$

i) (10 points) Show the dynamic programming matrix between 'daave' and 'dva' using a gab penalty equals 2.

		D	A	A	V	E	
	0	-2	-4	-6	-8	-10	
D	-2	1	-1	-3	-5	-7	
V	-4	-1	0	-2	-2	-4	
A	-6	-3	0	1	-1	-3	

ii) (10 points) Write down the optimal alignment score and the corresponding alignment between the two strings.

```
D A A V E
D V A — —
```

SCORE:

-3

3. (5 points) In matching elements between two schemas explain why we need to employ more than one matcher.

Different matchers employ different matching techniques. Therefore, one technique may be preferred over the other.

4. (5 points) Given two schemas S and T where S has 10 Tables and 45 elements while T has only one Table with 4 attributes. How many machine learning classifiers do you need to train in order to match elements in the two schemas? Explain how to get the training data for training one of these classifiers.

We only need to train one machine learning classifier to match elements in the two schemas. We get the training data by taking all the current data instances of S to be positive examples while the rest of the data instances of S are negative.

5. (10 points) Given three machine learning based matchers for matching an element, e, in a schema S with elements in another schema T. Let's call these matchers, $L_{e,NB}, L_{e,DT}, and\ L_{e,SVM}$. Suggest four ways for combining the outcome of these matchers.

① We can take the average of all the outcomes from the matchers

② We can take the largest outcome from any one of the matchers.

③ We can take the smallest outcome from any one of the matchers.

④ We can prioritize the outcome of some matchers by adding weights to those matchers.