Assignment 2

1.

a. c

b $I(X;Y) = h(X) - h(X|Y)$

because X and Y are independent $h(X) = h(X|Y)$
$I(X;Y) = 0$

2. a.) $P(X_A = 0, X_B = 1) = \frac{(N_b - N_{Ab})}{N}$

$P(X_A = 0, X_B = 0) = \frac{N - (N_A + N_B + N_{AB})}{N}$

b.) $I(X_{computer}; X_{program}) =$

$$\sum_{u \in (0,1)} \sum_{(0,1)} P(X_{computer} = u, X_{program} = v) \log \frac{P(X_{computer}, X_{program})}{P(X_{computer}) P(X_{program})}$$

$P(X_{computer} = 0, X_{program} = 0) = \frac{23,981}{26,394} = .8708$

$P(X_{computer} = 0, X_{program} = 1) = \frac{2021}{26394} = .07657$

$p(X_{computer}=1;\ X_{program}=0) = \frac{1041}{26394} = .03944$

$p(X_{computer}=1,\ X_{program}=1) = \frac{349}{26394} = .01322$

$p(X_{computer}=0) = \frac{25004}{26394} = .94734$

$p(X_{computer}=1) = \frac{1390}{26394} = .05266$

$p(X_{program}=0) = \frac{24024}{26394} = .91021$

$p(X_{program}=1) = \frac{2370}{26394} = .08979$

$I(X_{computer};\ X_{program}) =$

$$(.87078) \cdot \log\left[\frac{.87078}{(.91021)(.94734)}\right] + .07657 \cdot \log\left[\frac{.07657}{(.94734)(.08979)}\right]$$

$$+ .03944 \cdot \log\left[\frac{.03944}{(.05266)(.91021)}\right] + .01322 \cdot \log\left(\frac{.01322}{(.05266)(.08979)}\right)$$

$$= .01239 + -.01162 + .01109 + .02002$$

$$\boxed{.0097}$$

$I(X_{computer}; X_{baseball})$

$p(X_{computer}=0, X_{baseball}=0) = .86698$

$p(X_{computer}=0, X_{baseball}=1) = .07875$

$p(X_{computer}=1, X_{baseball}=0) = .05179$

$p(X_{computer}=1, X_{baseball}=1) = = .008714$

$p(X_{computer}=1) = .05266$

$p(X_{computer}=0) = .94734$

$p(X_{baseball}=1) = .08123$

$p(X_{baseball}=0) = .91877$

$I(X_{computer}; X_{baseball}) =$

$$.86698 \cdot \log\left(\frac{(.86698)}{(.94734)(.91877)}\right) + .07875 \cdot \log\left(\frac{.07875}{(.94734)(.08123)}\right)$$

$$+ .05179 \cdot \log\left(\frac{.05179}{(.05266)(.91877)}\right) + .008714 \cdot \log\left(\frac{.008714}{(.08123)(...)}\right)$$

$$= -.004906 + .002623 + .005015 + -.002$$

$$= \boxed{.000827}$$

2.C $I(X_{computer}; X_{program}) > I(X_{computer}; X_{basketball})$

which makes sense because program is more related to computer than basketball

2.D top 10 pair counts

(January, paper) = 161
(language, programming) = 153
(January, time) = 150
(January, program) = 149
(January, system) = 149
(data, January) = 142
(January, presented) = 141
(January, programming) = 139
(program, programs) = 133
(January, method) = 125

2.E) best mutual information

(input, output) = .094
(information, retrieval) = .0624
(language, languages) = .068
(languages, programming) = .0637
(jr, thatcher) = .05996
(differential, equation) = .05945
(memory, virtual) = .05579

These top 10 show words that are more related to each other, than the other top 10, because it normalizes number of occurences.

top 5 words programming:

(programming, languages) = .064
(programming, language) = .05576
(programming, program) = .0198
(programming, programs) = .0197
(programming, paper) = .0162

Yes this list seems to be reasonable

3.) a. KL divergence ≥ 0 if two distributions are equal, the KL divergence will be 0

b.)

x	$p(X=x)$
1	.10
2	.40
3	.50
4	
5	

y	$p(X=y)$
1	.80
2	.15
3	.05
4	
5	
6	

p

Q

$$D(p\|q) = \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right) = 1.237$$

$$D(Q\|P) = \sum_x q(x)\log\left(\frac{q(x)}{p(x)}\right) = 2.023$$

4)

when $q(x) = 0$

the denominator becomes 0 in

$$\sum p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

we can fix this issue by using a smoothing formula to make sure probabilities can't be 0