**Assignment # 1**

Name: Min Guo

NetID: mg4517

NID: N10971010

**Problem1:**

The document collection W = [woof, meow, squeak]

D1 = [woof, woof, meow]

D2 = [woof, woof, squeak]

**(a)** Vector V1 = [2, 1, 0]; Vector V2 = [2, 0, 1]

sim(D1, D2) = (4 + 0 + 0)/ = 4/5 = 0.8

**(b)** The inverse document frequency idfwoof of “woof” is log2(2/2) = 0;

The inverse document frequency idfmeow of “meow” is log22 = 1;

The inverse document frequency idfsqueak of “squeak” is log22 = 1;

Document D1: tfwoof = 2; tfmeow = 1; tfsqueak = 0, so

Wwoof = tfwoof \* idfwoof = 0; Wmeow = tfmeow \* idfmeow = 1; Wsqueak = tfsqueak \* idfsqueak = 0.

Document D2: tfwoof = 2; tfmeow = 0; tfsqueak = 1, so

Wwoof = tfwoof \* idfwoof = 0; Wmeow = tfmeow \* idfmeow = 0; Wsqueak = tfsqueak \* idfsqueak = 1.

sim(D1, D2) = (0 + 0 + 0) / 1 = 0

**(c)** D3 = [meow, squeak]; Vector V3 = [0, 1, 1]

The inverse document frequency idfwoof of “woof” is log2(3/2);

The inverse document frequency idfmeow of “meow” is log2(3/2);

The inverse document frequency idfsqueak of “squeak” is log2(3/2);

Document D1:

tfwoof = 2; tfmeow = 1; tfsqueak = 0, so

Wwoof = tfwoof \* idfwoof = 2\* log2(3/2);

Wmeow = tfmeow \* idfmeow = log2(3/2);

Wsqueak = tfsqueak \* idfsqueak = 0.

Document D2:

tfwoof = 2; tfmeow = 0; tfsqueak = 1, so

Wwoof = tfwoof \* idfwoof = 2\* log2(3/2);

Wmeow = tfmeow \* idfmeow = 0;

Wsqueak = tfsqueak \* idfsqueak = log2(3/2);

Document D3:

tfwoof = 1; tfmeow = 1; tfsqueak = 1, so

Wwoof = tfwoof \* idfwoof = 0;

Wmeow = tfmeow \* idfmeow = log2(3/2);

Wsqueak = tfsqueak \* idfsqueak = log2(3/2);

sim(D1, D2) = [2\* log2(3/2)]2/ = 4/5;

sim(D1, D3) = [log2(3/2)]2/ = 1/;

sim(D2, D3) = [log2(3/2)]2/ = 1/;

**Problem2:**

P(+ | "great food served") = P("great food served" | +) \* P(+);

P(- | "great food served") = P("great food served" | -) \* P(-);

P(+) = 5/10 = 0.5; P(-) = 5/10 = 0.5

**(a)** P("great" | +) = 5/10 = 0.5; P("food" | +) = 5/10 = 0.5; P("served" | +) = 0/10 = 0

P("great food served" | +) = P("great" | +) \* P("food" | +) \* P("served" | +) = 0;

P(+ | "great food served") = P("great food served" | +) \* P(+) = 0.

P("great" | -) = 0/11 = 0; P("food" | +) = 5/11; P("served" | +) = 1/11

P("great food served" | +) = P("great" | +) \* P("food" | +) \* P("served" | +) = 0;

P(+ | "great food served") = P("great food served" | +) \* P(+) = 0.

**(b)** P("great" | +) = (5 + 1)/(10 + 4) = 6/14; P("food" | +) = (5 + 1)/(10 + 4) = 6/14; P("served" | +) = 1/(10 + 4) = 1/14

P("great food served" | +) = P("great" | +) \* P("food" | +) \* P("served" | +) = 36/143;

P(+ | "great food served") = P("great food served" | +) \* P(+) = 18/143.

P("great" | -) = 1/(11 + 4) = 1/15; P("food" | -) = (5 + 1)/(11 + 4) = 6/15; P("served" | -) = 2/(11 + 4) = 2/15

P("great food served" | -) = P("great" | -) \* P("food" | -) \* P("served" | -) = 12/153 ;

P(- | "great food served") = P("great food served" | - ) \* P(-) = 6/153.

P(+ | "great food served") is larger than P(- | "great food served").