G22.2590 Natural Language Processing
Assignment 1 Solutions

1. Document similarity

   a vocabulary of 3 words:
   W = [woof, meow, squeak]

   each document is characterized by a vector of word counts:
   V1 = [2, 1, 0]
   V2 = [2, 0, 1]

   a) $\Sigma a_i \times b_i = [2, 1, 0] \bullet [2, 0, 1]$
      $= 4 + 0 + 0 = 4$
      $\Sigma a_i^2 = 4 + 1 + 0 = 5$
      $\Sigma b_i^2 = 4 + 0 + 1 = 5$

      $\mathrm{Sim}(A, B) = 4/(\sqrt{5}*\sqrt{5}) = 4/5 = \mathbf{0.8}$

   b) $\mathrm{IDF}_{woof} = \log(N/n_{woof}) = \log(2/2) = 0$
      $w_1 = 0$
      $w_2 = 0$
      $\mathrm{IDF}_{meow} = \log(2/1) = .30103$
      $w_1 =.30103$
      $w_2 = 0$
      $\mathrm{IDF}_{squeak} = \log(2/1) = .30103$
      $w_1 = 0$
      $w_2 = .30103$
      V1 = [0, .30103, 0]
      V2 = [0, 0, .30103]

      $\Sigma a_i \times b_i = [0, .30103, 0] \bullet [0, 0, .30103] = 0$

      Sim(A, B) = **0**

      (the only word the documents have in common is "woof", but "woof" appears in every document and so gets an IDF weight of 0)

   c) Word counts for the third document: V3 = [0, 1, 1]

      $\mathrm{IDF}_{woof} = \log(3/2) = .17609$
      $w_1 = .35218$
      $w_2 = .35218$
      $w_3 = 0$
      $\mathrm{IDF}_{meow} = \log(3/2) = .17609$

$w_1 = .17609$
$w_2 = 0$
$w_3 = .17609$
$IDF_{squeak} = \log(3/2) = .17609$
$w_1 = 0$
$w_2 = .17609$
$w_3 = .17609$
$V1 = [.35218, .17609, 0]$
$V2 = [.35218, 0, .17609]$
$V3 = [0, .17609, .17609]$

$\Sigma a_i \times b_i = [.35218, .17609, 0] \bullet [.35218, 0, .17609]$
$= .12403 + 0 + 0 = .12403$

$\Sigma a_i^2 = .35218^2 + .17609^2 + 0 = .12403 + .03101$
$= .15504$
$\Sigma b_i^2 = .35218^2 + 0 + .17609^2$
$= .15504$

$Sim[A, B] = .12403/(\sqrt{}.15504 * \sqrt{}.15504) = .12403/.15504$
$=$ **.7999**

("woof" no longer appears in every document, and so has non-zero IDF in the larger document collection)