Math 498
Homework #2 – solutions

Chartier/Greenbaum
Winter 2008

1. (a) We begin by forming the normalized document matrix where chocolate, ice cream and sprinkles correspond to rows 1, 2 and 3, respectively and column i corresponds to document Di. Therefore, we find

$$\tilde{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Normalizing the columns, we find:

$$A = \begin{pmatrix} 1 & \sqrt{2}/2 & \sqrt{2}/2 & \sqrt{3}/3 \\ 0 & \sqrt{2}/2 & 0 & \sqrt{3}/3 \\ 0 & 0 & \sqrt{2}/2 & \sqrt{3}/3 \end{pmatrix}.$$

(b) The query `chocolate, ice cream` corresponds to the query vector $q = (\,1 \quad 1 \quad 0\,)$. Therefore, if we find the cosine between the column vector a_i of A and the query vector q, we find:

$$\begin{aligned} \theta_1 &= \cos^{-1}\left(\frac{a_1 \cdot q}{\|q\|}\right) = \cos^{-1}(1/\sqrt{2}) = \pi/4, \\ \theta_2 &= \cos^{-1}\left(\frac{a_2 \cdot q}{\|q\|}\right) = \cos^{-1}(\sqrt{2}/\sqrt{2}) = 0, \\ \theta_3 &= \cos^{-1}\left(\frac{a_3 \cdot q}{\|q\|}\right) = \cos^{-1}((\sqrt{2}/2)/\sqrt{2}) = \cos^{-1}(1/2) = \pi/3, \quad \text{and} \\ \theta_4 &= \cos^{-1}\left(\frac{a_4 \cdot q}{\|q\|}\right) = \cos^{-1}(2/(\sqrt{3}\sqrt{2})) = 0.6155. \end{aligned}$$

Therefore, the documents are related to the query in the following order of decreasing relevance: D2, D4, D1, and D3.

2. See solution code available on the course web page for a sample code.

3. There are a variety of ways to accomplish this. In MATLAB, the `rand` command chooses its random numbers on the closed interval $[2^{-53}, 1-2^{-53}]$. Therefore, one way to construct a discrete random number generator that takes on the values 1 through 6 with equal probability in MATLAB is with the line `x = 6*ceil(rand)`. [If `rand` could generate an exact 0, then there would be some chance (albeit a tiny one, and with a true random number generator the probability of generating any particular number such as 0 is 0) of getting `x = 0`]. Another way to produce such a discrete random number generator that avoids this possibility is with the following algorithm:

```
r = rand;
if (r < 1.0/6)
 k = 1;
elseif (r < 2.0/6)
 k = 2;
elseif (r < 3.0/6)
 k = 3;
elseif (r < 4.0/6)
 k = 4;
elseif (r < 5.0/6)
 k = 5;
else
```

```
  k = 6;
end
```

4. See course web page for solution codes.

   For the simulation of the rolling of a fair die, the expected value is $E(X) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$. Similarly, $E(X^2) = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}$. Therefore, $\text{var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$. When we consider A_{1000} we see a plot that appears to be normally distributed. By the CLT, we know that as we increase the number of rolls and the number of experiments, the resulting histogram more closely approximates a normal distribution. For this experiment, we know from the CLT that the mean will be the same as the experiment above. Therefore, the mean will be 3.5. Further, $\text{var}(A_{1000}) = (35/12) \cdot (1/1000) \approx 0.0029$. Therefore, $\sigma \approx 0.0540$.