CFCS Tutorial One: Solutions

Miles Osborne

January 30, 2008

1. Given the following two vectors: $\mathbf{a} = (1, 2, 3)$ and $\mathbf{b} = (4, 1, 2)$. Compute:

   - The length (norm) of each vector. Answer:3.7 and 4.6
   - The dot product of these two vectors. Answer: 12
   - The length of **a** - **b** and **b** - **a**. Answer: 3.3 and 3.3 (they are the same)

2. Suppose all documents consist of just sentences using the following words: *the dog cat sat on mat barked meowed*. Represent the following documents:

   (a) *the dog barked*

   (b) *cat meowed*

   (c) *the cat sat on the mat*

   using vectors. You should also explain how your representation works.

   Answer: Assuming we enumerate all the words and use a 0/1 notation for presence/absence, we have a = (1 1 0 0 0 0 1 0), b = (0 0 1 0 0 0 0 1) and c = (1 0 1 1 1 1 0 0)

3. Work out the lengths (norms) of your vectors. 1.7, 1.4 and 2.2

4. For each vector, work-out the corresponding unit vector. norm(a) = (0.58, 0.58, 0.0, 0.0, 0.0, 0.0, 0.58, 0.0)
   norm(b) = (0.0, 0.0, 0.7, 0.0, 0.0, 0.0, 0.0, 0.7)
   norm(c) = (0.44, 0.0, 0.44, 0.44, 0.44, 0.44, 0.0, 0.0)

5. Now, work out the following distances between each vector:

   - Absolute distance. d(a,b) = 2.2; d(a,c) = 2.4, d(b, c) = 2.2
   - Cosine angle. cos(a,b) = 0.0, cos(a,c) = 0.26, cos(b,c) = 0.3

6. Which documents are most similar to each other? Does this vary according to the distance metric? Obvious answer.

7. If you changed your document representation, how would it affect our distances? Answer: Yes, eg stop words could be 0.5 etc.

8. In reality, vector representations of documents can deal with millions of possible words. What would happen to your represention? Any ideas how you can make it more space efficient? Answer: This is a question about sparse representations.