School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2017)
Workshop exercises: Week 3

1. Finish any remaining questions from last week, if necessary.

&

2. What is a **vector space model**, and why is it useful?

3. Consider the following collection of "documents" $\mathcal{C}$:

   (i) *It is what it is.*

   (ii) *Jean's hat is finer than Karl's hat.*

   (iii) *We are obsessing about gene issues.*

   Build a vector space model for this collection.

4. Use a model to decide which of the sentences in $\mathcal{C}$ is most similar to the following sentence:

   (iv) *Karl is obsessed with genes.*

   Based on the following metrics:

   (a) Jaccard similarity
   (b) the Dice coefficient
   (c) Euclidean distance
   (d) Manhattan distance
   (e) Cosine similarity

&

Consider the following dataset, representing a collection of documents:

ID	*apple*	*ibm*	*lemon*	*sun*	LABEL
A	4	0	1	1	FRUIT
B	5	0	5	2	FRUIT
C	2	5	0	0	COMP
D	1	2	1	7	COMP

5. If we wished to estimate the value of $P(apple, \text{FRUIT})$ based on the data above, we might arrive at the value $\frac{2}{4}$, or the value $\frac{9}{36}$.

   (a) What do each of these probabilities correspond to? Explain the model that underlies each of these interpretations.
   (b) Which of these do you think would be more useful, if we wished to use these probabilities to help predict the labels of unknown documents?

6. Calculate the **entropy** of the distribution of the LABEL attribute, based on the data above.