School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 1, 2017)
Workshop sample solutions: Week 8

1. In the **PageRank** algorithm:

   (a) What is the mechanism for the "random walk"?

   - We begin at a random page (or probabilistically, we begin at all pages with equal probability).
   - We choose one of the following:
     - We follow one of the outgoing links from this page, with probability $(1-\alpha)$ — evenly distributed amongst all outgoing links on this page
     - We "teleport" to a page entirely at random, with probability α — evenly distributed amongst all pages (Note: if there are no out-going links, we do this with probability 1, evenly distributed.)

   (b) In terms of user behaviour, what is the significance of "teleporting"?

   - Following a link from a page is kind of obvious, based on user behaviour. "Teleporting", on the other hand, corresponds to navigating via entering a URL into the address bar (which doesn't appear to be related the content of this page).

   (c) The lecture example of the PageRank algorithm was given as follows:

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.5 = 0.3$	$0.5 \times 0.2 \times 0.5 + 0.5 \times 0.8 + 0.5 \times 0.5 = 0.7$
2	$0.3 \times 0.2 \times 0.5 + 0.7 \times 0.5 = 0.38$	$0.3 \times 0.2 \times 0.5 + 0.3 \times 0.8 + 0.7 \times 0.5 = 0.62$
3	$0.38 \times 0.2 \times 0.5 + 0.62 \times 0.5 = 0.348$	$0.38 \times 0.2 \times 0.5 + 0.38 \times 0.8 + 0.62 \times 0.5 = 0.652$

   i. Which terms represent "following a link" and which represent "teleporting"?

   - d_1 can only be reached via "teleporting" (there are no in-links), both terms in the sum correspond to that (the first from d_1 itself, the second from d_2)
   - d_2 can be reached by following the only link from d_1, which is given by the second of the three terms in the d_2 calculations; the other two terms are "teleporting" (note that they are the same as the d_1 terms)

   ii. What is the value of α in the above example? Re-do the above with $\alpha = 0.5$

   - We know that following a link from document m to document n gives a term like the following:

$$d_{(n,t+1)} \quad = \quad d_{(m,t)} \times \alpha \times \frac{1}{|m|}$$

     where $|m|$ represents the number of links in document m.
   - In this case, d_1 only has one outgoing link, so we can see that d_2's score is being updated by the weight of d_1 in the previous iteration, times $(1-\alpha)$, which means that, in this case $\alpha = 0.2$
   - Changing to $\alpha = 0.5$ gives us:

t	$\pi(d_{(1,t)})$	$\pi(d_{(2,t)})$
0	0.5	0.5
1	$0.5 \times 0.5 \times 0.5 + 0.5 \times 0.5 = 0.375$	$0.5 \times 0.5 \times 0.5 + 0.5 \times 0.5 + 0.5 \times 0.5 = 0.625$
2	$0.375 \times 0.5 \times 0.5 + 0.625 \times 0.5 \approx 0.406$	$0.375 \times 0.5 \times 0.5 + 0.375 \times 0.5 + 0.625 \times 0.5 \approx 0.594$
3	$0.406 \times 0.5 \times 0.5 + 0.594 \times 0.5 \approx 0.398$	$0.406 \times 0.5 \times 0.5 + 0.406 \times 0.5 + 0.594 \times 0.5 \approx 0.602$

2. What is **Machine Learning** (or **Data Mining**) and why is it a **Knowledge Technology**? Why has this field become important? Contrast the use of data/information/knowledge/wisdom with the way we have used them previously in this subject.

   - "Data mining" is a blanket term referring to a set of knowledge technologies problems where the analysis of quantifiable data leads to inferences about quantifiable (typically statistical) observations or patterns about the data.
   - "Machine learning" is a cover term for a set of algorithms (like classification, clustering, or regression algorithms) for mechanically manipulating our data set to produce certain statistical observations; usually used as part of data mining
   - These are important because of the explosion in the quantity of available data, where useful information is encoded — this has increased far more quickly than our capacity for managing data (as humans)

3. An urn initially contains 10 red balls and 6 black balls. At each trial, a ball is selected from the urn, its colour is noted, and then it is returned to the urn with 2 more balls of the same colour.

   - There are two events: the first trial T_1 and the second trial T_2, each of the trials can show either a red ball R or a black ball B.

   (a) Compute the probability of obtaining a red ball with both the first and second trials

   - First, we find the probability of observing a red ball on the first trial:

$$\begin{aligned} P(T_1 = R) &= \frac{\text{\# of successful instances}}{\text{total \# of instances}} \\ &= \frac{10}{16} = \frac{5}{8} \end{aligned}$$

   - Then, we notice that, if we observe a red ball the first time around, the distribution in the urn for the second trial changes to 12 red balls and 6 black balls. So, the probability of observing a red ball on the second trial, *given* that we have observed a red ball on the first trial, is $\frac{12}{18}$. Now we can find the joint probability (using the **product rule**):

$$\begin{aligned} P(T_1 = R \cap T_2 = R) &= P(T_2 = R \mid T_1 = R) \cdot P(T_1 = R) \\ &= \frac{12}{18} \cdot \frac{10}{16} \\ &= \frac{5}{12} \end{aligned}$$

   (b) Show that the events "red ball on the first trial" and "red ball on the second trial" are **not** independent

   - If two events are independent, then the joint probability that they both occur is equal to the product of the prior probabilities of each event occurring. In this case, that would be:

$$P(T_1 = R \cap T_2 = R) = P(T_1 = R) \cdot P(T_2 = R)$$

   - The left hand side is $\frac{5}{12}$ from (a). We also found that $P(T_1 = R) = \frac{10}{16}$ in (a). As for $P(T_2 = R)$, we need to consider both of the cases: where the first trial was red, and where the first trial was black (the (normalised) **sum rule**):

$$\begin{aligned} P(T_2 = R) &= P(T_2 = R \cap T_1 = R) + P(T_2 = R \cap T_1 = B) \\ P(T_2 = R) &= P(T_2 = R \mid T_1 = R) \cdot P(T_1 = R) + P(T_2 = R \mid T_1 = B) \cdot P(T_1 = B) \\ &= \frac{12}{18} \cdot \frac{10}{16} + \frac{10}{18} \cdot \frac{6}{16} \\ &= \frac{5}{8} \end{aligned}$$

   - So, the right-hand side of the independence equation above is $\frac{5}{8} \cdot \frac{5}{8} = \frac{25}{64}$, and the left hand side is $\frac{5}{12} = \frac{25}{60}$. These are clearly not equal, so the events must not be independent.